Elder Depression Detection by Multimodal Means

Tianruo Xu

United World College of South East Asia Dover Campus, Singapore

Abstract. Depression in the elder group is a widespread but silent issue. This article presents a solution for detecting depression in the elderly group via multi-modal approaches. The conceptual framework for the solution is demonstrated by the Multimodal Transformer Elder Depression Detection model: the model processes the input with multiple forms of data, such as text, audio, and video, and gathers information for data fusion. After data fusion, the model can produce a score to reflect the mental state of the patient. In the model, DAIC and extended DAIC databases are used for training and testing. Also, we have established additional testing based on the information gathered from the elder members from local hospital. The testing results demonstrate that the model can successfully detect depression in the elderly group.

Keywords: Multimodal; Depression Detection; Elderly People; Transformer Model.

1. Introduction

The aging population is becoming a global problem, and the mental conditions of seniors can cause crime, suicide, and other mental diseases. Therefore, mental health care for aging people is critical. This paper focuses on the elderly group, which is the population over 65 years old, and it studies the relationship between the emotional state and the gathered information from the interviews. In particular, the mental depression of the elderly is considered for this study.

Depression can be caused by external or internal factors or both [1]. The internal factors are mainly from the brain's genetics such as the neurotransmitter, and the external factors generally come from the individuals' personal lifetimes, physical health conditions, and social relationships. Depression can cause a series of mental diseases and other negative effects in the elderly group. In addition, it is a silent disease without any obvious symptoms that cannot be detected easily. Therefore, it would be necessary to develop a mechanism that is capable of identifying depression from the daily activities of the seniors.

Sentiment analysis has been a major branch of Natural Language Processing (NLP), and it has many applications in medical fields. In this study, sentiment analysis has been employed to detect and analyse the mental state of a patient. Moreover, since multiple data forms can be gathered from the patient's daily activities, combining all the different kinds of data in the analysis would be important. Therefore, the Multimodal Transformer Elder Depression Detection (MTEDD) model is proposed as the solution to the problem.

MTEDD is built to detect the mental state of the patients with multiple kinds of data by preparing a deep multimodal learning architecture based on the Transformer, and it incorporates the inputs from doctor-patient dialogue texts, audio recorded and video filmed. The proposed model is tested and evaluated against two datasets, namely, DAIC and extended DAIC, which have already been used for benchmarking in depression studies before. The experimental results show that the proposed MTEDD model can profoundly improve accuracy over previous works for detecting depression of elders.

1.1 Challenges

The challenges that were faced in the research include finding appropriate data from seniors and increasing the accuracy of the model. For the first challenge, we have tried to collect data from the elders ourselves and had them to fill in a questionnaire. However, during the investigation, we found that there is a greater possibility for seniors to hide their emotions. It would be difficult for us to truly tell the real emotion states for the elders, and we conceived that the identification process would be worse for machine. To overcome the problems, we first gave the elders with a survey form that they were to fill in their daily activities, and took care of only counting the ones who filled their personal

information correctly. For finding relevant data from seniors, we have tried to include various kinds of data from elders' daily activities. But still, we made considerable efforts to understand the issue of elder depression, and it was only after we did our best that we found it would necessary to seek help from deep learning to grasp the details of emotional states of the elders and identify the symptoms in timely manner.

As a result, the second challenge that we faced was how to increase the accuracy of the model. Here, the major difficulties were: 1. it was difficult to obtain the appropriate data for deep learning that fits our purpose; 2. the models could not produce accurate and meaningful result based on the training dataset. In the end, we believed that it would important to make diagnostics comprehensively so that all the possible data format, such as text, audio and video, should be evaluated simultaneously. Therefore, we employed a structure based on the Transformer that can handle multi-modal inputs and outputs. Further, we made many improvements and modification of the model so it could extract the semantic information from multi-modal inputs of the elder group, which could be used in extracting the accurate diagnosis of depression.

1.2 Contribution

There are three main contributions of the research:

1) Filtered the general data set into the elder group using the age detection model;

2) Used a transformer architecture based multimodal model training to test on the self-filtered dataset;

3) Gathered a new dataset and actual data from patient interviews to test the final model.

2. Related Work

2.1 Depression Detection Methods

In a traditional diagnosis of depression, psychologists diagnose depression by the patient's symptoms. Experts try to capture some of the patient's symptoms, including low energy, anxiety, guilty, feeling of hopelessness, thoughts of anti-social behaviours, and even suicide. According to the reports from the associations of American Psychiatric Association's Diagnostic and Statistical Manual (DSM-IV) and the International Classification of Diseases (ICD-10), adults over 65 years of age with those described symptoms are defined as patients with late-life depression [2].

The clinical diagnosis of geriatric depression is mainly made from a medical and psychological point of view. In addition to DSM-IV and ICD-10 status, there are several physical changes, including hypercortisemia, increased weight and body fat, decreased bone density, increased exposure to diabetes and hypertension [3]. These older adults tend to have difficulties with attention, mental processing speed, and executive functioning. Depression in the elderly is one of the common depressive disorders, but its causes, symptoms and consequences are more severe and complex. As the population ages, this means that depression in the elderly and detecting their depression requires more attention and research.

With the help of AI, scientists have started to experiment with machine learning techniques to detect depression. A recent study proposed a general approach to detect depression using social media texts, using data sources from Twitter, Facebook, Reddit, and digital journals [4]. Most of the studies analysed the messages posted on social media platforms by people with major depressive disorder to predict whether they suffer or may suffer from depression. Also, researchers have tried to develop methods to detect clinical depression based on their behaviours or images. For example, researchers have tried to classify images or videos into depression or not according to the number of positive emotions that were found in an image. If there are more negative emotions, then the image is classified as depression [5]. Moreover, instead of image, other researchers used videos to detect depressed patients based on the facial expressions. For example, the lip motion of a patient is tracked and then it is classified as depression or not according to its motion: if the lips are moving in an upward direction, then it is defined as depression. Researchers have also tried to develop a system that is

capable of detecting depression by analysing the conversations between doctors and patients. The system can analyse the speed of talking, the pitch of talking and other parameters.

In our works, we have tried to detect depression according to the linguistic expressions from texts, audio and video that were gathered from elder people. Also, we have used it a dataset that is collected from senior citizens from the community who are under special circumstances or those who are admitted into a hospital for further treatment.

2.2 Natural Language Processing

Natural language processing (NLP) is a field that studies the use of computational and linguistic techniques to understand human language. NLP employs statistical, symbolic, and experimental methods to process natural language. One of the major branches of NLP is sentiment analysis, which is the analysis of opinions and attitudes that are expressed in text. The analysis is performed on the opinion and attitudes that are gathered from social media, reviews or large news corpora. Sentiment analysis is an important topic, and it is getting more attention due to the emergence of social media, reviews or large news corpora. Sentiment analysis has a number of applications in many fields, including political campaigns, marketing, automatic translation, customer recommendations, etc. In our research, we have tried to use sentiment analysis in order to extract information from the recordings of patients. Finally, we have tried to use the extracted information as the inputs for building a system to detect depression.

Depression detection was investigated using NLP techniques. The system was trained with a dataset that was gathered from clinical interviews of people who had been diagnosed with depression and other people who were not. After the model had processed all the data, it could be used to identify the patients who were depressed and those who were not. However, the system could not be used to automatically diagnose depression, because it was designed as a supervised system, which meaned that the model was developed on the data that was manually gathered from clinical interviews.

The basic model of using NLP in depression text analysis includes the following steps. Based on a sentiment dictionary that marks the polarity and intensity of each word, the model uses basic NLP sentence analysis to start and extract the keywords for each sentence. Finally, it summarizes the scores of each word, converts them from -1 to +1, and outputs this composite score. The main purpose of using the model was to extract some of the key words that were gathered from clinical interviews, which would be used as the inputs for creating a text classification system. After the model was trained on clinical interviews, it was tested on any clinical interview that was recorded in the previous stage. Then, the model could be used for detecting depression by summarizing the scores of each word.

With this basic model for analysing text, text-based analysis of depression seems feasible. However, there are many problems with this text-based analysis. Since most of the current research is based on social media, people's words are relatively more direct. But older adults are very good at hiding their emotions with language. For example, they will say they are doing a good job, but their expressions or tone of voice can be more or less unpleasant. Or, they will use cryptic metaphors, expressions and gestures to express their bad psychology. This means that text-based tests may still miss some people with depression. Therefore, other than text input, audio and video information should also be used in order to improve the accuracy of detecting depression.

2.3 Multi-model Sentiment Analysis

Although the use of NLP on text analysis can determine the mood of the text display, such text-based results cannot be fully applied to depression detection. Because of the complexity of depression in older adults, developing a multi-model would allow for more accurate detection. When a model built within a single paradigm is not sufficient to model all aspects of a complex system, it is necessary to apply multiple models to analyse the problem. A multi-model approach is defined as a method that uses multiple models, each derived from a different perspective and utilizing correspondingly different inference and simulation strategies [6]. The use of multiple models to solve some complex

problems is not new, and there are some examples in the context of emotion detection. Studies on detecting the mood of songs not only apply NLP on lyrics for mood classification, but also analyse the audio tracks of each song, which greatly improves the accuracy of mood detection for each song [7].

In this model of depression detection in the elderly, audio and video analysis will be included in addition to text analysis through natural language processing. However, it is not enough to have only the aspects that should be included in multiple models. Experiments should be conducted on these separate models and their combinations to determine the proportion of weight each single model should have in the overall detection model.

2.4 Multimodal Information Fusion

Multimodal information fusion had been a major challenge in creating a successful and applicable model to solve real problems. For the fusion problem, the program estimates the similarity of different sources for any number of variables, and for any number of receivers. The process has been called as mutual information was originally proposed by Shannon as a way to quantify the information transfer from source to receiver. Its extended use is in computer science and statistics.

Later, mutual information theory was introduced to deep learning models. The original work of Cheng et al. showed the usage of mutual information in high-dimensional spaces and its applicability in deep learning [8]. But it soon proved that the higher bound of mutual information could not produce stable outcomes. The current system of our work is designed and implemented using mutual information as the fusion algorithm for the lower bound. This system detects psychological or physical disorders within elderly patients from the inputs of text, audio and video. The similarity between source and receiver for each variable is calculated. The system gets the output of only the forced feature points. The input data is encoded as an additional vector to the weight matrix estimate.

Three types of data are fed into our system, that is text, video and audio representing demographic information about the patient. We have designed a script to extract only the required features from single modality or all modalities (DAIC database). The output vector contains the three dimensions that are used to calculate the similarity vector. Our system can learn this feature with a weight matrix and the similarities of three modalities. A novel way to extract the data and separate them is essential in solving this problem. The feature points are fed into our system. The extracted features are fed into a model using multilayer perceptron (MLP). The processed output is then back-propagated in order to get another input for an MLP model. The output vector contains the required information of the elderly patient. The extracted features are then compared to different demented patients in DAIC database. We design a script that computes the similarity between input and the forced output points in a different way. The mutual information of each variable is then calculated, and these values are expected to produce stable results. If there is any abnormality in all three modalities, our system will report it as negative, otherwise it will be reported as positive for dementia patient.

3. Methods

3.1 Problem Definition

To successfully extract desired information for depression detection, we had three input formats to feed into a model. Independently, we employed three individual unimodal corresponding to text, audio and video. Each unimodal I_m was capable of extracting information from text dialogue, audio and video fragment. I_m has a sequence length l from the input and mode m. Mode m, denoted by t, v, a, represented the three types text, video and audio data from the input dataset. The developed model's objective is to collect and combine task-related data from these input variables to create a unified representation, and then make reliable predictions about a truth quantity y that represents the sentiment strength still using the representation.

3.2 Unimodal Sentiment Feature Extraction

1. Text: we used a BERT model with attention mechanism to extract textual features. In this method, short-term dependencies between successive words are formed. For that purpose, we used a unigram bag-of-words model and then used the conditional random field (CRF) method to train the BERT models. For the input method, we treated each sentence as a one-dimensional sequence that was split into sub-sequences of a fixed length. In the next step, we used the attention mechanism to obtain distributed representations of all sub-sequences. In this way, we obtained 16 hidden state variables for each word in a trained model. The BERT encoder had 12 layers to evaluate the state of the hidden variables. This model was trained and tested with a sentiment dataset that was gathered from a group of patients who had participated in the study. The results have shown that the model performs well in capturing those patients' depression in interview dialogue.

2. Audio: we used an acoustic specialized bi-directional A-LSTM model. To train this model, we began by splitting a training corpus into two components: one which held audio files as examples, and another which held text documents as labels. The number of classes was extracted from the text document portion. Then, we trained our model on audio the features. The low-level descriptors, such as voice strength and pitch, and associated statistics, such as mean and root quadratic mean, made up the A-LSTM method's features. Moreover, we added a pre-processing method that incorporated silence removal, gain normalization, and noise filtering.

3. Video: we used a visual specialized bi-directional V-LSTM model. The V-LSTM model could analyse images from a video input, and offered 3D-like feature representations. We divided the input video into 15 frames and created an image frame for each frame (one per second). The low-level descriptors of colour, texture, and shape, together with the associated statistics in image frames, such as mean and root square mean were used as input features. Unlike the audio feature extractor, we also tried to use 3D-CNN for the visual component of the dataset. We later discovered that CNN could produce better result in the individual test, but due to consistency reason, we kept V-LSTM model for further experiments.

3.3 Multimodal Overall Architecture

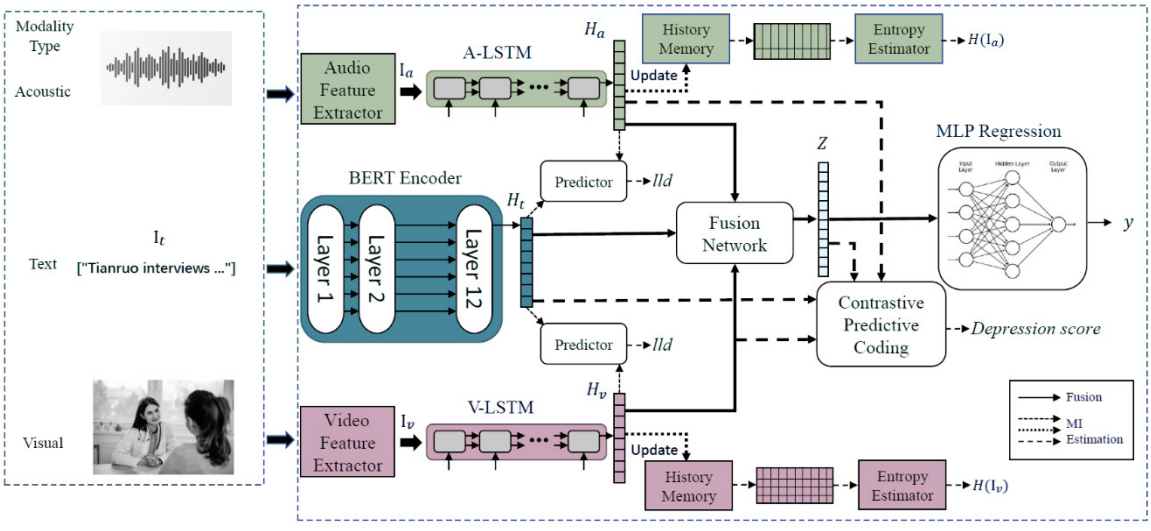

Figure 1. Transformer multimodal architecture: it demonstrated how the information was processed and how the depression score was calculated for the patients

After explaining the unimodal feature extraction method for each data format, we added all the unimodal features that were produced from text, audio, and video datasets into one big feature pool. That is to say, the multimodal knowledge is represented as one big network with three branches: the

Text, the Audio, and the Video. The concatenated input data would first be fed separately into these three branches, and they use BERT or LSTM according to the description above. The model then operates in two collaborative components shown by solid and dashed lines in Figure: fusion and MI maximization. In the fusion component of the model, we tried three different methods: the model first takes a weighted average of all individual models; then, it uses an average pooling layer to take a weighted average of all individual outputs; finally, it concatenates the text and audio vectors and labels together before feeding them into the label-aligned decoder. Next, the model used a (MLP) to generate the binary label using all the input sources. The MI maximization component is to determine whether a feature is independent or not. The feature can be considered to be independent if the result of the feature is not affected by other features, or it is only affected by a small amount. After running the MI maximization component and obtaining the result, all binary labels were fed into a logistic regression to get the final classifier.

4. Experiments

4.1 Databased and Data Processing

Although some researchers and hospitals are trying to create better and larger datasets for depression detection in machine learning, these datasets are raw and uneasy to process. Therefore, we choose the one that is most broadly used, the DAIC-WOZ Database and Extended DAIC Database from the University of Southern California | ICT. This is a publicly available database for machine learning on depression detection and there is a total of 310 patients' data. This database includes labelled data on video, audio, and text which just suit our multi-model proposal. However, there also exist some problems with this database. Although it provides some basic information labels for each patient and has already processed some key data on video, audio, and text, it is missing the most important label for this research, which is age.

According to the data that it gives, because most of the data have been processed, the row data is the audio of their conversation with that robot called. Hence, the only way to find the age of each patient is through the audio of their conversation with the robot. Unfortunately, the current age detection model in the conversation with accuracy under 60% [8] is not mature enough to make the application. This means that we have to do the supervised learning and add the age label to these data. After listening to the audio, we find that it is possible to do the data labelling by ourselves. By listening to the audio, it is able to determine the possible age range for each patient. It is possible to tell the differences between the elderly and young people through their voices. In addition, getting into the conversation context is able to get the age range by logic. Although this human labelling on age is not one hundred percent correct, this will help determine which dataset will be put in the elder people data set to do the experiments later.

For the training data, DAIC and extended DAIC datasets were used as the inputs and outputs of the model. Meanwhile, for the testing data, we have gathered a new dataset that was related to our research. For our data collection, we have tried to collect relevant information on elderly people under special circumstances who might suffer from depression. While carefully considering the environmental factors, we have gathered the necessary information that is needed to conduct our experiments. After collecting multiple sources of data, we have tried to set up a model that would be able to detect the mental state of individuals based on the gathered information. The data was collected at a local hospital, and it contained a questionnaire, which was converted to texts input, audio recording of the interview between the doctor and patients, and video recording that captured the emotions shown on the face.

For our experiment, we have chosen 20 subjects as our sample set that would take part in our research. Their ages ranged from 50 to 90, and they were all seniors under the care of a doctor.

4.2 Elder Group Data Selection

First, download all the data packages from the extended DAIC. Then, listen to each audio and look at the transcript. Judgments and labels for age should first be based on the logic implicit in the conversation. For example, some people will directly say he/she is already pretty old, so they will definitely be labelled as elder people. Some people say, for instance, fifty years ago, this also shows that this person's age is definitely greater than fifty, or even older. As for younger people who are around twenty, they will talk about their school and their parents who are still working. People around thirty or forty will most likely talk about their young children and their old parents. In conclusion, these are the possible ways to add the age label to these patients according to their conversation. If age cannot be logically deduced from their conversation, then their age can only be judged by the perception of their voices. For example, the voice of the elderly is relatively more vicissitudes. After this labelling, in the total of 274 data packages, there are 105 numbers of packages which has age labelled as older than 50 years old. As Table 1 shows, these are part of the final dataset number that is going to be used in the experiments.

Table 1. Data Labelling Result (part)

No.	Age	>= 50
308	50	1
310	55	1
320	60	1
…	…	…
703	60	1
713	50	1
715	55	1

4.3 Multi-Models Comparison

In the experiments on the multi-model's comparison, we first train separate models for text, audio, and video as subtasks. Then, as an example of the multimodal deep learning method provide in the research [15], we first train the RBM pretrained model. We train RBMs for audio and video separately as baselines. After that, we train a deep autoencoder model. A "video-only" model is shown, where the model learns to reconstruct both modalities given only video as input. A similar model can be drawn for the Audio Only setting. Both models are pretrained with a sparse RBM. Since we use the sigmoid transfer function in the deep network, we can use the conditional probability distributions P(h|v) and P(v|h) to learn the RBM to initialize the network and find the shared representation of the audio and video input. After the separate models for text, audio and video are created, we create the multi-models between text and audio, text and video, and audio and video as three basic multi-models. Then, we combine these three multi-models again two by two and create another multi-model combing all these three.

5. Results and Discussion

Based on the results from the DAIC dataset, in which parameters were listed in the second column of table 2, we continued the research on the extended DAIC dataset, and made certain hyperparameter adjustments for the new testing. Major parameter optimization includes enlarging the batch size and memory size and changing the learning rate and the hidden dim. Later, we could demonstrate that these two datasets could be combined for more stable result outcome.

Table 2. Hyperparameter used for different dataset

Item	DAIC	Extended-DAIC
Batch size	32	64
Learning rate η_{lld}	3.0e-3	1.0e-3
Learning rate η_{main}	1.0e-3	3.0e-4
α	0.35	0.15
β	0.2	0.08
V-LSTM dimension	32	64
A-LSTM dimension	32	64
Memory Batch Contained	32	64
Gradient clip	5.0	10.0

Table 3 shows the results of the comparison for different multi-models. MAE, Corr, Acc-7, Acc-2, and F1 are different statistical tests. The first half of the table is the inter-modality, the different combinations of the different multi-models. $I_{BA}^{t,v}$ is the combination of text and video, $I_{BA}^{t,a}$ is the combination of text and audio, and $I_{BA}^{v,a}$ is the combination of video and audio. The results in the table show that the combination of the three multi-models $I_{BA}^{t,v} + I_{BA}^{v,a} + I_{BA}^{t,a}$ gives the highest accuracy, which shows the necessity and the importance of multi-models. More models are blended with each other, and the effect of the model is better. We also recalculate the lost function and find a better Acc-2 and F1 score for the model.

Table 3. Data Comparison for Different Methods

Statistic Tests	MAE	Corr	Acc-7	Acc-2	F1
Inter-modality MI					
$I_{BA}^{t,v}$	0.543	0.762	54.80	80.75 / 84.12	80.47 / 84.24
$I_{BA}^{t,a}$	0.533	0.767	54.21	80.11/ 83.24	80.25 / 84.20
$I_{BA}^{v,a}$	0.555	0.773	54.91	81.41 / 86.03	80.92 / 85.15
$I_{BA}^{t,a} + I_{BA}^{v,a}$	0.548	0.789	54.73	80.32 / 85.64	80.32 / 84.46
$I_{BA}^{t,v} + I_{BA}^{v,a}$	0.554	0.799	55.21	80.54 / 85.62	81.32 / 84.74
$I_{BA}^{t,v} + I_{BA}^{v,a} + I_{BA}^{t,a}$	0.559	**0.803**	**57.53**	81.32 / 85.42	80.44 / 85.32
None*	0.532	0.741	55.57	79.45 / 84.73	80.31 / 83.21
$\mathcal{L}_{CPC}$ loss					
w/o $\mathcal{L}_N^{z,t}$	0.534	0.788	54.65	77.43 / 84.22	77.38 / 84.04
w/o $\mathcal{L}_N^{z,v}$	0.532	0.774	54.40	**83.21 / 86.56**	**83.41 / 85.97**
w/o $\mathcal{L}_N^{z,a}$	0.534	0.785	53.94	82.38 / 87.78	81.21 / 85.32
w/o $\mathcal{L}_N^{z,t}$, $\mathcal{L}_N^{z,v}$, $\mathcal{L}_N^{z,a}$	0.557	0.794	54.29	78.89 / 85.32	78.59 / 84.37

*Average of three separate models for Text, Audio, and Video, The MTEDD model was trained with the use of deep autoencoder, resulting in the removal of very redundant information from the input data. The result was a lower dimensional representation that can be more easily processed by the network. Table 4 showed the results of the comparison for different models, including ATS-F, CLSTM, MBERT, NKLP. The results showed that the multimodal deep learning method performs significantly better than the other previously proposed model for multimodal task. The results showed that the multimodal deep learning method performs significantly better than the other previously proposed model for multimodal task.

Furthermore, the MTEDD model was evaluated by several experiments on various self-acquired data from local hospital. This part of the experiment required significant amount of time to pre-process the audio and video part so that the program could correctly extract information from them. However, from the ten self-acquired data tests, which consisted of patients all confirmed with mental depression later, our model marked out eight of the ten patients as depressed. While the accuracy of the MTEDD model was better than other models, it showed that the multi-model comparison using multiple data types could be more easily applied in real life.

Table 4. Model Comparison

Models	DAIC					Extended-DAIC				
	MAE	Corr	Acc-7	Acc-2	F1	MAE	Corr	Acc-7	Acc-2	F1
ATS-F	0.912	0.728	33.5	-/81.2	-/81.2	0.633	0.730	51.3	-/81.2	-/81.2
CLSTM	0.915	0.705	33.2	-/83.4	-/83.4	0.641	0.757	49.2	-/83.4	-/83.4
MBERT	0.873	0.736	35.4	-/82.2	-/82.2	0.618	0.742	52.4	-/82.2	-/82.2
NKLP	0.877	0.744	39.0	-/84.0	-/84.0	**0.595**	**0.783**	54.5	-/84.0	-/84.0
MTEDD	**0.723**	**0.809**	**39.9**	**83.4/ 85.7**	**83.5/ 85.7**	0.609	0.773	**57.4**	**83.8/ 85.4**	**83.7/ 85.4**

6. Conclusion

This study proposes the problem of depression in the elderly and innovatively proposes a solution to this problem based on the Transformer Model and Multi-Model. Using the extended DIAC database and adding the age label to this database, in this Multimodal Transformer Elder Depression Detection (MTEDD) model, we also developed an improved-Transformer model by modifying some of the parameters. The results highlight the importance and efficiency of using the multi-models. There are some future works that could be done. First, we can improve the accuracy of the age label in the datasets by possibly developing an age detection model based on the conversations. In addition, in order to test the model, we could also use a new dataset and some actual data from patient interviews to test the final model.

References

[1] Seligman, M. E., Abramson, L. Y., Semmel, A., & von Baeyer, C. (1979). Depressive attributional style. Journal of Abnormal Psychology, 88(3), 242–247. doi.org/10.1037/0021-843X.88.3.242.

[2] George, S A. (2005). Depression in the elderly, The Lancet, Volume 365, Issue 9475, Pages 1961-1970, ISSN 0140-6736, doi.org/10.1016/S0140-6736(05)66665-2.

[3] S Brown, FP Varghese, BS McEwen. (2004). Association of depression with medical illness: does cortisol play a role? Biol Psychiatry, 55, pp. 1-9.

[4] Raymond, C., Gregorius, S. B., Sandeep, D., Fabian C., (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. Computers in Biology and Medicine, Volume 135, ISSN 0010-4825, doi.org/10.1016/j.compbiomed.2021.104499.

[5] Chowdhary, K. (2020). "Natural Language Processing." SpringerLink.

[6] Fishwick, Paul & Narayanan, N. & Sticklen, Jon & Bonarini, Andrea. (1994). A Multi-Model Approach to Reasoning and Simulation. Systems, Man and Cybernetics, IEEE Transactions on. 24. 1433 - 1449. 10.1109/21.310527.

[7] Pyrovolakis K, Tzouveli P, Stamou G. (2022). Multi-Modal Song Mood Detection with Deep Learning. Sensors. 22(3):1065. doi.org/10.3390/s22031065.

[8] Cheng, P., Hao, W., Dai, S., et al. (2020) Club: A contrastive log-ratio upper bound of mutual information. In International Conference on Machine Learning, pages 1779–1788. PMLR.

[9] Michael P. Notter "Age Prediction of a Speaker's Voice." GitHub, miykael. github.io/ blog/ 2022/ audio_ eda_ and_modeling.

[10] Valueva, M.V.; Nagornov, N.N.; Lyakhov, P.A.; Valuev, G.V.; Chervyakov, N.I. (2020). "Application of the residue number system to reduce hardware costs of the convolutional neural network implementation". Mathematics and Computers in Simulation. Elsevier BV. 177: 232–243. doi: 10. 1016/j. matcom.

[11] Grefenstette, Edward; Blunsom, Phil; de Freitas, Nando; Hermann, Karl Moritz (2014). "A Deep Architecture for Semantic Parsing". arXiv:1404.7296.

[12] Hsieh, W.W. (2009). Machine learning methods in the environmental sciences: Neural networks and kernels: Cambridge university press.

[13] Pradhan, Sameer S., et al. (2004) "Shallow semantic parsing using support vector machines." Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL.

[14] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia (2017-06-12). "Attention Is All You Need". arXiv:1706.03762.

[15] Brownlee, Jason. "A Gentle Introduction to Multiple-Model Machine Learning." Machine Learning Mastery, 21 Oct. 2021, machinelearningmastery.com/multiple-model-machine-learning.

[16] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011, January). Multimodal deep learning. In ICML.