CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Autofusion: Fusing multi-modalities with interactions
Authors	Nguyen, Thuy-Trinh;Deligianni, Fani;Nguyen, Hoang D.
Publication date	2024
Original Citation	Nguyen, T.-T., Deligianni, F. and Nguyen, H. D. (2024) 'Autofusion: Fusing multi-modalities with interactions', 28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024), Seville, Spain, 11-13 September. Procedia Computer Science, 246, pp. 1029-1038. https://doi.org/10.1016/j.procs.2024.09.522
Type of publication	Conference item
Link to publisher's version	10.1016/j.procs.2024.09.522
Rights	© 2024, the Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-11-10 14:27:59
Item downloaded from	https://hdl.handle.net/10468/15950



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 246 (2024) 1029–1038

Procedia Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

Autofusion: Fusing Multi-Modalities with Interactions

Thuy-Trinh (Chloe) Nguyen[a], Fani Deligianni[a], Hoang D. Nguyen[b,*]

[a]School of Computing Sciences, University of Glasgow, Glasgow, United Kingdom
[b]School of Computer Science and Information Technology, University College Cork, Cork, Ireland

Abstract

In the context of escalating data flows through diverse channels, multimodal machine learning holds the potential to simultaneously process varied data formats from multiple sources, offering robust solutions for uncertainty in various applications. The study highlights the often under-explored correlation information among data modalities and emphasizes the importance of disentangling enriched interactions for more informed decision-making. This paper navigates the burgeoning field of multimodal artificial intelligence (AI) by proposing Autofusion, a pioneering framework addressing representation learning and fusion challenges. Our proposed approach integrates autoencoder structures to address overfitting issues in unimodal machine learning, simultaneously tackling information balancing challenges. The framework's application in Alzheimer's disease detection, using DementiaBank's Pitt corpus, demonstrates promising results, outperforming unimodal methods and showcasing a substantial advantage over traditional fusion techniques. This research significantly contributes by introducing Autofusion as a comprehensive multimodal machine learning solution, demonstrating its efficacy through DementiaBank's Pitt corpus to detect Alzheimer's disease, and shedding light on the influential role of cross-modality interaction for enhanced performance in complex applications.

© 2024 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems

Keywords: multimodal interactions; multimodal fusion; multimodal machine learning; autofusion

1. Introduction

As technologies evolve and data flows through multiple channels, the enormous amount of data from various modalities provides an excellent opportunity for further analysis to support multimodal artificial intelligence (AI) tasks [30]. Multimodal AI has the advantage of simultaneously processing data coming from multiple sources in different formats to combine useful information for making final predictions to tackle uncertainty in real-world applications, such as autonomous vehicles [27] and multi-sensory healthcare monitoring [11]. However, the correlation information among data modalities is often omitted in existing multimodal machine learning approaches; hence, it is valuable to disentangle enriched interactions among the data points.

* Corresponding author.
E-mail address: hn@cs.ucc.ie

1877-0509 © 2024 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems
10.1016/j.procs.2024.09.522

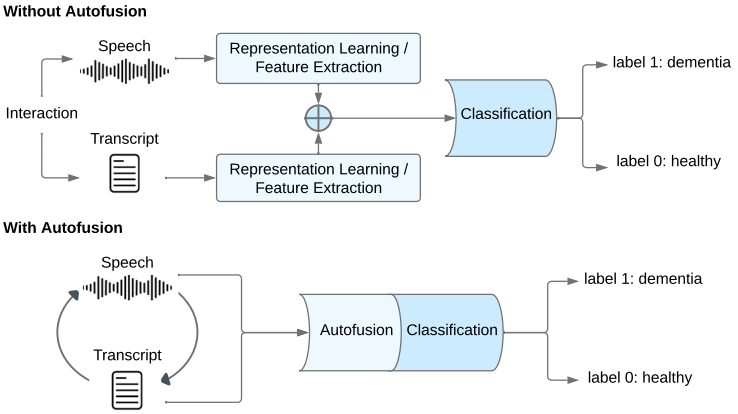


Fig. 1. Autofusion Motivating Example

In unimodal machine learning, autoencoder has been widely adopted as a common class of representation learning techniques [25]. Nevertheless, autoencoder methods have the tendency to overfit towards the learning of identity-function between the input and the output that could slow down the training and testing process [26]. On the other hand, information balancing refers to the challenge of loss control between the autoencoder loss and that of the downstream task, which significantly decides the learning pattern and progress of the model. Due to the heterogeneity of multimodal data streams, the multimodal data fusion mechanism has been one of the main issues with multimodal representation learning and downstream AI tasks. The cross-modality interaction has been considered a promising factor to leverage fusion among multimodalities [6, 28]. Hence, this paper delves into the realm of multimodal machine learning and proposes Autofusion, an autoencoder-infused fusion architecture that harnesses the cross-modality interaction among different modalities to overcome the discussed issues of representation learning and fusion in multimodal machine learning. Figure 1 illustrates a motivating example that compares common autoencoder-based models and Autofusion architecture.

On a worldwide scale, around 55 million people suffer from dementia, which is increasing by 10% annually. Accounting for 60-80% of dementia cases, Alzheimer's disease (AD) is a degenerative cognitive condition that not only interferes with the patient's daily activities, work routine, and social interaction, but it can result in immobility and organ failure in later stages [1]. To help with early screening of this disease, this study demonstrates the use of Autofusion architecture to detect AD using DementiaBank's Pitt corpus.

This research makes meaningful contributions on three fronts. Firstly, the paper introduces Autofusion, an end-to-end multimodal machine learning framework that leverages autoencoder structures for learning representation and cross-modality interactions to achieve a highly effective fusion of modalities. Secondly, the paper benchmarks Autofusion architecture on DementiaBank's Pitt corpus, a real-world dataset for Alzheimer's disease. Autofusion achieves a promising performance of 79.89% in accuracy, 83.85% in recall, 81.72% in precision, and 82.47% in F1. The technique consistently outperforms all unimodal methods by an average of 5.24% across all metrics. Especially against the late fusion technique, our method shows a significant difference of 20% on average across metrics. Finally, this study captures the impact of the cross-modality interaction term as its incorporation enhances the model performance by 2-3% across metrics. The proposed approach harnesses the power of the interaction effect as a leveraging element to achieve superior results over unimodal baselines and common fusion alternatives. The paper is structured with a review of related literature, methodology, experiments, results discussion, and a conclusion summarising the findings.

2. Related Works

2.1. Multimodal Machine Learning

The world is designed for humans to experience with different senses. Seeing, hearing, touching, smelling, and tasting contribute to our perception of life. Fascinatingly, while these senses are often connected to generate a complete understanding of something, even its fragmented pieces can infer the whole picture. Different ways of experiencing are forms of *modalities*. The concept of integrating multiple facets of a subject transcends human natural senses. This idea has been investigated and adopted in machine learning, now known as multimodal machine learning. Joining

information from two or more modalities to predict an outcome measure has been one of the original focuses of multimodal machine learning.

Model-agnostic approaches do not directly rely on the architecture of a specific machine learning model. Early fusion methods gained popularity because (1) fusion at an initial stage gives an opportunity for low-level features from modalities to interact and potentially provide cross-modality correlation, and (2) only one model is required for fine-tuning. However, modality-specific features need to be represented in a similar format prior to fusion, which could put a toll on the multimodal representation step [19]. In other words, time synchronisation is a challenge for early fusion. Since features from various modalities often vary significantly, they must be standardised before fusion can occur. On the contrary, late fusion integrates the outcomes of modality-wise predictions after model training. Late fusion allows for more flexibility and impacts training each modality separately and offers scalability regarding the number of modalities involved [2]. Nevertheless, this method necessitates the training of multiple models, which can render the learning process computationally intensive and protracted. Additionally, this technique might overlook potential correlations at the feature level within individual modalities. For a comprehensive analysis of these methods, refer to [3], which provides a detailed examination of the advantages and disadvantages of each approach.

Model-based. In terms of the fusion level, this group of methods is model-level. In contrast to model-agnostic approaches, model-based fusion incorporates the fusing step into the architecture of the models, which allows for more involvement during the fusion process. Popular models used in hybrid fusion often employ classic methods such as Support Vector Machine (SVM) [7, 17], graphical models [31, 23], and neural networks [15, 14]. Some works also explore the use of autoencoders in the fusion process and achieve promising results because of their advantages over early and late fusion [24, 10]. To utilise the flexibility and scalability of model-based fusion, this paper introduces autoencoder-based fusion into the Autofusion architecture.

2.2. Multimodal Machine Learning in Dementia Detection

In recent years, machine learning and deep learning techniques have been proven reliable in predicting degenerative conditions [5]. Utilising automated detection schemes to analyse EHR data has the potential to improve the quality of life for those with dementia and their families. Despite multimodal machine learning achieving great performance 88% F1 for Pitt Corpus [21], this group of techniques is still under-explored for AD detection. Even for those multimodal papers whose performances are promising, their fusion techniques are limited at hard voting (i.e., late fusion) [21] and simple concatenation early fusion [18]. Achieving a potential result using model-based fusion, [16] explores joint fusion using a simple feed-forward neural network combining text and image modalities. All considered papers use classification models. The majority of papers focus on traditional machine learning techniques, including Support Vector Machine (SVM) [13, 20]. [21] opted for a designed neural network to classify AD patients. Based on the analysis of performance evaluation for AD detection, it is apparent that there is a significant need for automated detection models. The initial results indicate ample opportunities for growth and improvement in this area.

3. Approach

This section describes the proposed framework to detect dementia using multiple modalities: audio and transcripts. Regarding model features, the proposed framework employs those engineered by [21] to be valid for benchmarking. The framework is inspired by the potential cross-modality interaction and denoising effects of autoencoders in multimodal machine learning; thus, the input features will be compressed using an autoencoder, and its bottleneck will be used for the primary classification. Figure 2 illustrates the detailed architecture of the network.

3.1. Feature Processing

The processing of the audio and transcript files from the DementiaBank Pitt dataset results in three distinct feature types. Table 1 presents general statistics on the generated feature sets which are inspired by [21].

- *Acoustic*: We applied ComParE 2013 feature set [9] using the OpenSMILE toolkit to extract 6,373 features from audio files. These features include MFCC, low-level descriptors (LLDs), and other statistical features. We apply a linear network to reduce the dimensionalities of acoustic data.

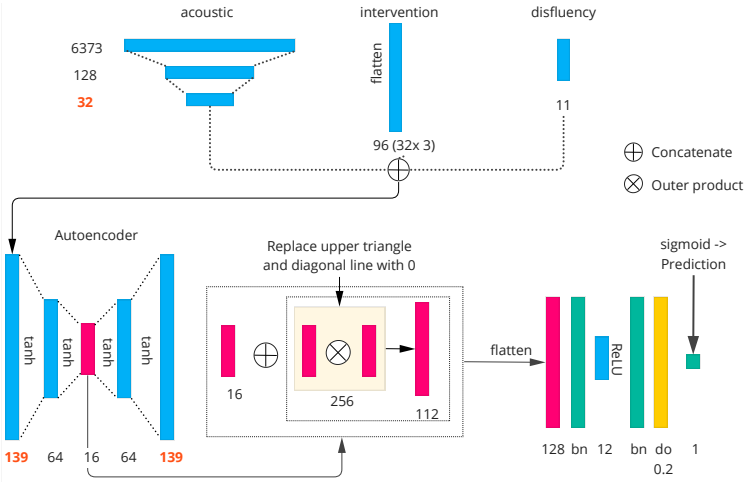


Fig. 2. Autofusion Architecture

Feature	Extraction	Dimension
Acoustic (Audio)	ComParE 2013 feature set using OpenSmile toolkit	[1, 6373]
Disfluency (Transcript)	Handcrafted features including word rate, pause rates of various kinds, and intervention rate	[1, 11]
Intervention (Transcript)	One-hot encoded sequence to indicate the speaker (i.e., interviewer or participant)	[1, 32, 3]

Table 1. Statistics of generated features

- *Disfluency*: Disfluency is a set of 11 hand-crafted features from the transcript, including word rate and pause rates, to reflect any discontinuity in the speech.
- *Intervention*: The intervention feature set provides more context for the transcript by extracting a sequence of speakers to identify whether it is the interviewer or the participant speaking. The intervention feature set is padded or truncated to a fixed length of 32 with 3 channels of one-hot values (i.e., 1 or 0).

As different feature sets have different value ranges, some normalisation operation is conducted to ensure homogeneity in the input scale. Hence, a standard scaler is applied to address the varied ranges. Specifically, we implemented the *StandardScaler* function from the scikit-learn library to achieve this transformation, which standardises features by removing the mean and scaling to unit variance.

3.2. Autofusion Mechanism

Autoencoders have been studied for their reconstructive, denoising qualities and their potential in multimodal fusion that can encourage interactions among modalities [3]. We use a modified version of the bottleneck's outer product with itself as the input of the classification network. This component shall be referred to as the cross-modality interaction or the interaction term interchangeably in the subsequent sections.

After feature transformations, three feature sets become inputs of the Autofusion network for multimodal fusion. Different from traditional early and late fusion, the Autofusion mechanism focuses on studying the interactions among

modalities. First, three feature sets are concatenated and fed through an autoencoder with an undercomplete architecture. The bottleneck of the Autofusion network is computed as follows:

$$h_0 = x_a \oplus x_d \oplus x_i \tag{1}$$

$$h_1 = \sigma(W_0 h_0) \tag{2}$$

$$x_{bottleneck} = \sigma(W_i h_i) \tag{3}$$

where:

- x_a, x_d, x_i: Input feature acoustic, disfluency, and intervention respectively
- $\oplus$: Concatenation
- h: Input
- W: Weight
- σ: Activation function
- $x_{bottleneck}$: Bottleneck vector

3.2.1. Classification Network

The classification network comprises two main components: the Autofusion bottleneck and its linearly transformed version. The two inputs are combined using the outer product function, creating the interaction term. The calculation of the interaction term can be found in Equation 4.

$$h_{interaction} = x_{bottleneck} \otimes x_{bottleneck} \tag{4}$$

Equation 5 and 6 describe the calculation of the network inputs of the first layer h_0 and the following layers h_l.

$$h_0 = h_{interaction} + x_{bottleneck} \tag{5}$$

$$h_l = \sigma(W_{(l-1)} h_{(l-1)}) \tag{6}$$

where:

- $x_{bottleneck}$: Bottleneck vector from Autofusion
- $h_{interaction}$: Interaction term among feature sets
- $\otimes$: Outer product
- h: Input
- W: Weight
- σ: Activation function
- l: Layer number [0, number of layers]

3.2.2. Objective Function and Evaluation

We used Mean Square Error (MSE) loss to monitor the autoencoder training process and Binary Cross Entropy (BCE) loss for classification. The formulas of MSE and BCE losses are summarised in Equation 7 and 8, respectively.

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{7}$$

$$\mathcal{L}_{BCE} = -(y \log(\hat{y}) + (1-y)\log(1-\hat{y})) \tag{8}$$

- $\mathcal{L}_{MSE}$: Mean Square Error (MSE) loss
- N: Total number of samples
- y: Label
- $\hat{y}$: Prediction

- $\mathcal{L}_{BCE}$: Binary Cross Entropy (BCE) loss
- y: Label
- $\hat{y}$: Prediction

To calculate the total loss of the network, we applied the loss formula in Equation 9 with a 0.1 loss factor for autoencoder loss and a 0.9 loss factor for classification loss.

$$\mathcal{L}_{total} = (\lambda_{MSE} \times \mathcal{L}_{MSE}) + (\lambda_{BCE} \times \mathcal{L}_{BCE}) \tag{9}$$

where:

- $\mathcal{L}_{total}$: Total loss function
- λ_{MSE}: Loss factor of the Autofusion network ranged [0.0, 1.0]
- λ_{BCE}: Loss factor of classification network ranged [0.0, 1.0]
- $\mathcal{L}_{MSE}$: MSE loss from autoencoder network
- $\mathcal{L}_{BCE}$: BCE loss from classification network

4. Experiments

4.1. Dataset

We select DementiaBank's Pitt corpus [4] as our benchmarking dataset. The dataset has audio recordings and text transcripts of an image description task in which the participants are required to describe the activities in the Cookie Theft image. Some highlighted variables in the dataset include demographic data (i.e., gender, age, education level), psychological test scores, and diagnosis codes. The complete list of variables is referenced in [4]. The Pitt corpus is selected for benchmarking since it is one of the largest corpra there is for AD detection tasks. Additionally, due to the gradual degradation of AD, longitudinal studies are impactful; hence, the Pitt corpus is our dataset choice.

To identify labels for the experiment, we selected individuals with a primary diagnosis code of 800 as the control group and those with a diagnostic code of 100 as patients with AD dementia. Diagnostic codes were assigned via clinical assessment. For our experiments, in total, we used 552 samples, which included 243 speech samples from 99 control healthy subjects and 309 speech samples from 194 AD subjects. Table 2 provides essential statistics.

	Sample	Individual	Sex	
			male (1)	female (0)
Control	243	99	41	58
Dementia	309	194	68	126
Total	552	293	109	184

Table 2. Dataset statistics

Of the 293 individuals included in the dataset, their entry age range varies between 46 and 88 years old. It is noteworthy that age data is approximate because (1) the study only records the age upon entry of an individual but not updated age for each participating year, and (2) no date of birth is given. On average, the mean age is 67.4 years old. In terms of participants' genders, 37.2% (109 out of 293) of the participants are male, while 62.8% (184 out of 293) are female.

4.2. Resources

This section describes the computational power used for the experiments and the framework for the experiment setup. Our proposed model is implemented using PyTorch (1.13.1). The experimental setup is running on Ubuntu 22.04 with Intel Core i9 10900X 10 Cores, 256 GB RAM and Nvidia 3090 24 GB.

4.3. Running Parameters

In our experiments, we ran 300 epochs and utilised the Adam optimizer for all networks. K-fold cross-validation method was selected with $k = 10$. The autoencoder training network was implemented with a *tanh* activation, while the classification network used $ReLU$. Additionally, we applied a learning rate decay mechanism with an initial learning rate of $lr_{init} = 5 \times 10^{-4}$ that decreases by 20% every 100 epochs, as shown in Equation 10, to fine-tune training and

prevent local minima. This approach has been proven effective in other studies and allowed us to achieve accurate results while avoiding common machine learning issues, like overfitting.

$$lr_e = lr_{init} \times 0.2^{\frac{e}{100}} \tag{10}$$

where:

- lr_{init}: Initial learning rate ($= 5 \times 10^{-4}$)
- e: Current epoch
- $\frac{e}{100}$: Integer quotient of current epoch number divide by 100

We run our experiments in 10 folds with the Adam optimizer, as well as *tanh* (for autoencoder) and *ReLU* (for classification) as activation functions. The initial learning rate is 5×10^{-4}, with a decay rate of 20% per 100 epochs. The loss factors are 0.1 for MSE loss and 0.9 for classification loss, respectively. We used a random seed of 0 and a batch size of 24 for all experiments. For fold indexes, we employed the stratified K-fold method with 10 folds, a random state of 999, and shuffling set to True.

4.4. Benchmarking Baselines

We compare the results of our approach to the benchmarks set out in [21]. This paper was submitted for the INTERSPEECH ADReSS Challenge 2020. Three factors led to the selection of this paper. First, the experiment's source code is made public, allowing for the potential reproducibility of the findings. Second, this article compares results using the DementiaBank Pitt Corpus. Finally, this paper has the highest reported performance on DementiaBank's Pitt Corpus among those meeting the above requirements.

The source code of the paper is accessible on Github [1]. While the article is also assessed using the ADReSS Dataset and DementiaBank's Pitt Corpus, the Pitt corpus is our main focus. [21] presents a baseline employing late fusion (hard voting) with accuracy, recall, precision, and F1 scores of 0.88, 0.82, 0.92, and 0.88, respectively. We made an effort to replicate the experiments using the source code; however, the outcome is different from what was originally published. There are a few theories as to why this occurs. Firstly, the random elements used in the experiments, such as the random seed number, are not specified in the source code. Secondly, although both papers perform 10-fold cross-validation on the dataset, there is no provided information on fold index (i.e., sample indexes in each fold), which implies different fold splits and contributes to the performance disparity. Last but not least, while all experiments in the benchmarking study are trained and assessed on the ADReSS Dataset before being retrained on DementiaBank's Pitt Corpus, only the Pitt corpus is used in our paper for training and evaluation. Consequently, this enhances the possibility of a data leak. Therefore, we re-baselined the published solutions for reproducibility and reliability.

5. Results and Discussion

5.1. Results

Table 5.1 summarises the results of both unimodal and multimodal experiments. These results are reported as the average performance of 10-fold cross-validation.

(1) Comparison with Unimodal Baselines

We conducted experiments to compare our proposed technique against three single-input models, including acoustic, intervention, and disfluency models similar to the unimodal baselines in [21]. Autofusion achieved a performance of 79.89% in accuracy, 83.85% in recall, 81.72% in precision, and 82.47% in F1. It consistently outperformed all unimodal methods by an average of 5.24%. Specifically, our proposed method surpassed acoustic, intervention, and disfluency unimodal baselines with an average improvement of 7.4%, 6.2%, and 2.1% across all metrics, respectively.

[1] https://github.com/ wazeerzulfikar/alzheimers-dementia

Type	Experiment	Accuracy	Recall	Precision	F1
Unimodal	Acoustic	72.01 (±0.07)	76.61 (±0.08)	74.33 (±0.06)	75.35 (±0.06)
	Intervention	70.82 (±0.07)	82.26 (±0.15)	74.08 (±0.12)	75.92 (±0.04)
	Disfluency	77.70 (±0.07)	80.58 (±0.07)	80.93 (±0.09)	80.30 (±0.05)
Multimodal	Late fusion	56.87 (±0.03)	60.84 (±0.06)	61.56 (±0.02)	61.09 (±0.04)
	Early fusion	78.07 (±0.06)	81.22 (±0.07)	80.19 (±0.06)	80.53 (±0.05)
	Autofusion	**79.89** (±0.08)	83.85 (±0.07)	**81.72** (±0.09)	**82.47** (±0.06)
	Autofusion (No interaction)	77.91 (±0.08)	**84.15** (±0.06)	78.62 (±0.08)	81.12 (±0.06)

Table 3. Experiment results

In terms of accuracy, autofusion achieves a higher accuracy of 79.9% compared to the accuracy of 72%, 70.8%, and 77.7% achieved by the acoustic, intervention, and disfluency models correspondingly. Regarding F1 performance, autofusion achieves an F1 score of 82.5%, surpassing the acoustic model by 7.1%, intervention model by 6.6%, and disfluency model by 2.2%. Given the imbalanced nature of the Pitt corpus with 309 AD and 243 control samples, F1 is a more robust and reasonable metric compared to other evaluation metrics in this case.

In addition to the evaluation metrics, the stability of the models is another noteworthy factor. Stability refers to the consistency of a model's performance across different fold splits (i.e., standard deviation across folds). Most experiments report a variation of 0.03 - 0.09% across 10 folds, except for intervention models, which have a variation of 0.15% in recall and 0.12% in precision. Autofusion shows relative robustness compared to other single-input models as it deviates marginally from the unimodal averages. This demonstrates that multimodal machine learning models are a highly effective method for automated medical screenings.

(2) Comparison with Other Fusion Baselines

We compare the autofusion method against the other fusion alternative methods. For the early fusion technique, three input streams are concatenated and fed into the same classification network architecture as the autofusion technique. Autofusion outperforms early fusion with an average of 2% improvement in overall metrics. Specifically, the performance increases from 78.07% to 79.89% in accuracy, from 81.22 to 83.85% in recall, from 80.19% to 81.72% in precision, and from 80.63% to 82.47% in F1 with the addition of autofusion mechanism prior to classification. Regarding the stability of the two fusion models, autofusion shows a greater standard deviation of 0.02% in accuracy, 0.03% in precision, and 0.01% in F1 compared to early fusion. As early fusion and autofusion share a similar classification architecture, the main difference lies in the fusion mechanism of concatenation versus autoencoder-incorporated fusion to capture the interaction effect.

As per the guidelines in [21], we use the hard-voting late fusion technique as another baseline. The result presents a notable improvement of 20% average across all metrics using autofusion against late fusion. Although the three single-input models perform well independently, their combination through late fusion does not work as effectively. Late fusion likely ignores the relationships between modalities and their representations in earlier stages, resulting in a potential disconnect between predictions. Where the models are unstable, non-overlapping predictions may occur, resulting in a lower-performing majority vote outcome. Regarding the stability of the two models, the late fusion model establishes a smaller variation across folds than the autofusion model. The standard deviations of the late fusion model vary from 0.02 to 0.06%, while those of autofusion range from 0.06 to 0.09%. Based on the experiment results, the proposed autofusion method outperforms both early and late fusion.

(3) Additional Ablation Studies

The inputs of the classification network comprise two elements: (1) the autoencoder latent representation (i.e., bottleneck) and (2) an interaction term, which is a form of outer product between the bottleneck and itself. The use of the outer product is inspired by the Tensor Fusion Network in which the outer product contributes to the model acquiring both intra-modality and inter-modality dynamics [29]. When we omit the interaction term, the results reduce by roughly 2% from 79.89% to 77.91% in accuracy, 3% from 81.72% to 78.62% in precision, and 1.4% from 82.47% to 81.12% in F1. One exception is that by removing the interaction term, the model presents a 0.03% increase from

83.85% to 84.15% in recall. In terms of stability, as there is a minimal difference of 0.01% in recall and precision standard deviations between experiments across 10 folds.

5.2. Discussion

5.2.1. Domain Explainability

The empirical results indicate that the disfluency model outperforms the other two models by a considerable margin, suggesting that speech discontinuity has the potential to serve as a valuable predictor of AD. Semantic characteristics have been proven to be potential indicators of other disorders. Speech disfluency has been identified as a promising indicator for identifying individuals with social anxiety disorder, particularly in terms of speech duration, jitter, and shimmer [22], in line with observations made by psychologists. This finding from the study not only highlights the effectiveness of the disfluency model in detecting AD, but also provides a basis for further research into the relationship between speech patterns and other disorders.

5.2.2. Impact of Cross-modality Interaction in Multimodal Fusion

This finding corresponds to the effectiveness of previous literature on model-based fusion techniques. Empirical results show that the interaction term improves the overall performance by 2-3% of the model. Although modest, the improvement provided by the interaction term suggests that when used in conjunction with multimodal fusion, autoencoders can enhance the correlation among participating modalities. Therefore, refinement in this area could lead to significant advancements in multimodal fusion techniques. Autofusion model stability, however, can be improved as the standard deviations across folds exhibit a slight increase compared to early and late fusion.

5.2.3. Implications

The recent research study has substantial implications for machine learning and medical research. The study introduces a new approach to multimodal fusion by incorporating autoencoders, which can be further explored to improve the current state-of-the-art. The network architecture proposed in this study can serve as a valuable reference and benchmark for developing solutions for DementiaBanks's Pitt corpus. Furthermore, the study confirms that automated medical screening using multimodal machine learning can effectively leverage the complementary effect of different data streams. This has the potential to transform chronic disease detection, as the condition is notoriously difficult to diagnose and often takes years to identify.

6. Conclusion

This paper explores the application of cross-modality interaction in the context of multimodal fusion, focusing on its validation in the domain of mental disorder detection. The study delves into the intricacies of multimodal fusion through the introduction of the Autofusion mechanism, which is then assessed on DementiaBank's Pitt corpus for Alzheimer's disease detection. Autofusion demonstrates promising performance. Notably, it consistently outperforms all unimodal methods by an average of 5.24% across all metrics. To summarise, this paper contributes novel insights in multiple respects. Firstly, it elucidates the predictive capabilities of speech disfluency characteristics for Alzheimer's Disease, advancing speech analysis within clinical diagnostics through the demonstration of specific speech markers' effectiveness. Secondly, it introduces an innovative use of autoencoders in the Autofusion technique to improve cross-modality interaction in multimodal fusion, achieving important performance improvements. This paper outlines potential avenues for future research. To further validate the technique, it could be benchmarked against additional datasets, such as DAIC-WOZ [12] for depression and Bipolar Disorder Corpus [8] for bipolar disorder, which would provide insights into the performance across various mental conditions. Secondly, while this study focuses on audio and transcript samples, expanding the range of modalities by incorporating medical data, such as physiological features, could advance the model's architecture and broaden its applicability.

Acknowledgement. This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant 12/RC/2289-P2.

References

[1] Association, A., et al., 2021. 2021 alzheimer's disease facts and figures. Alzheimer's & Dementia 17, 327–406.

[2] Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S., 2010. Multimodal fusion for multimedia analysis: a survey. Multimedia systems 16, 345–379.

[3] Baltruaitis, T., Ahuja, C., Morency, L.P., 2019. Multimodal machine learning: A survey and taxonomy. IEEE TPAMI 41, 423–443.

[4] Becker, J.T., Boiler, F., Lopez, O.L., Saxton, J., McGonigle, K.L., 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. Archives of neurology 51, 585–594.

[5] Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., Litjens, G., 2020. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology 21.

[6] Chen, H., Li, Y., Su, D., 2019. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. Pattern Recognition 86, 376–385.

[7] Chen, J., Chen, Z., Chi, Z., Fu, H., 2014. Emotion recognition in the wild with feature fusion and multiple kernel learning, in: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 508–513.

[8] Ciftçi, E., Kaya, H., Güleç, H., Salah, A.A., 2018. The turkish audio-visual bipolar disorder corpus. ACII Asia , 1–6.

[9] Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor, in: Proceedings of the 21st ACM international conference on Multimedia, pp. 835–838.

[10] Gao, J., Li, P., Chen, Z., Zhang, J., 2020. A survey on deep learning for multimodal data fusion. Neural Computation 32, 829–864.

[11] Garcia-Ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K.J., Tørresen, J., 2018. Mental health monitoring with multimodal sensing and machine learning: A survey. Pervasive and Mobile Computing 51, 1–26.

[12] Gratch, J., Artstein, R., Lucas, G.M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D.R., Rizzo, A.A., Morency, L.P., 2014. The distress analysis interview corpus of human and computer interviews, in: LREC.

[13] Lopez-de Ipiña, K., Alonso, J.B., Solé-Casals, J., Barroso, N., Henriquez, P., Faundez-Zanuy, M., Travieso, C.M., Ecay-Torres, M., Martinez-Lage, P., Eguiraun, H., 2015. On automatic diagnosis of alzheimer's disease based on spontaneous speech analysis and emotional temperature. Cognitive Computation 7, 44–55.

[14] Jaafar, N., Lachiri, Z., 2023. Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. Expert Systems with Applications 211, 118523.

[15] Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J., 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: Proceedings of the 25th ACM international conference on Multimedia, pp. 795–816.

[16] Krstev, I., Pavikjevikj, M., Toshevska, M., Gievska, S., 2022. Multimodal data fusion for automatic detection of alzheimer's disease, in: Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Health, Operations Management, and Design, Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part II, Springer. pp. 79–94.

[17] Liao, Z., Gao, L., Zhou, T., Fan, X., Zhang, Y., Wu, J., 2019. An oil painters recognition method based on cluster multiple kernel learning algorithm. IEEE Access 7, 26842–26854.

[18] Nasreen, S., Hough, J., Purver, M., et al., 2021. Detecting alzheimer's disease using interactional and acoustic features from spontaneous speech, Interspeech.

[19] Poria, S., Cambria, E., Bajpai, R., Hussain, A., 2017. A review of affective computing: From unimodal analysis to multimodal fusion. Information fusion 37, 98–125.

[20] Santander-Cruz, Y., Salazar-Colores, S., Paredes-García, W.J., Guendulain-Arenas, H., Tovar-Arriaga, S., 2022. Semantic feature extraction using sbert for dementia detection. Brain Sciences 12, 270.

[21] Sarawgi, U., Zulfikar, W., Soliman, N., Maes, P., 2020. Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity. arXiv preprint arXiv:2009.00700 .

[22] Silber-Varod, V., Kreiner, H., Lovett, R., Levi-Belz, Y., Amir, N., 2016. Do social anxiety individuals hesitate more? the prosodic profile of hesitation disfluencies in social anxiety disorder individuals. Speech Prosody 2016 , 1211–1215.

[23] Singh, P., Srivastava, R., Rana, K., Kumar, V., 2021. A multimodal hierarchical approach to speech emotion recognition from audio and text. Knowledge-Based Systems 229, 107316.

[24] Singh, V., Verma, N.K., Ul Islam, Z., Cui, Y., 2019. Feature learning using stacked autoencoder for shared and multimodal fusion of medical images, in: Computational Intelligence: Theories, Applications and Future Directions-Volume I: ICCI-2017, Springer. pp. 53–66.

[25] Somepalli, G., Wu, Y., Balaji, Y., Vinzamuri, B., Feizi, S., 2021. Unsupervised anomaly detection with adversarial mirrored autoencoders, in: Uncertainty in Artificial Intelligence, PMLR. pp. 365–375.

[26] Steck, H., 2020. Autoencoders that don't overfit towards the identity. Advances in Neural Information Processing Systems 33, 19598–19608.

[27] Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., López, A.M., 2020. Multimodal end-to-end autonomous driving. IEEE Transactions on Intelligent Transportation Systems 23, 537–547.

[28] Yan, X., Xue, H., Jiang, S., Liu, Z., 2022. Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling. Applied Artificial Intelligence 36, 2000688.

[29] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P., 2017. Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250 .

[30] Zhang, J., Tran, H., Lu, D., Zhang, G., 2021. Enabling long-range exploration in minimization of multimodal functions, in: Uncertainty in Artificial Intelligence, PMLR. pp. 1639–1649.

[31] Zheng, W., Yan, L., Gou, C., Wang, F.Y., 2020. Graph attention model embedded with multi-modal knowledge for depression detection, in: ICME, pp. 1–6.