Interspeech 2022
18-22 September 2022, Incheon, Korea

Detecting Heart Failure Through Voice Analysis using Self-Supervised Mode-Based Memory Fusion

Darshana Priyasad[1 2], *Andi Partovi*[2], *Sridha Sridharan*[1], *Maryam Kashefpoor*[2], *Tharindu Fernando*[1], *Simon Denman*[1], *Clinton Fookes*[1], *Jia Tang*[3], *David Kaye*[3]

[1]SAIVT, Queensland University of Technology, Brisbane, Australia
[2]KeyLead Health, Melbourne, Australia, [3]Alfred Hospital, Melbourne, Australia

{dp.don, s.sridharan, t.warnakulasuriya, s.denman, c.fookes}@qut.edu.au, {andi, mary}@keyleadhealth.com, {j.tang, d.kaye}@alfred.org.au

Abstract

Congestive Heart Failure (CHF) is a progressive disease that affects millions of people worldwide, severely impacting their quality of life. Missed detection of CHF and its progression affects life expectancy, thus it is critical to develop applications to continuously monitor CHF symptoms and disease progression in a patient-centric and cost-effective manner. This paper focuses on a novel non-invasive technique to identify CHF using patients' speech traits. Pulmonary congestion and breathlessness is the most common symptom of heart failure and one of the major contributors to hospitalisation. Since pulmonary congestion results in impairment of a patient's voice, we propose a novel, non invasive method for monitoring CHF through analysis of the patient's speech. We also introduce a new balanced dataset, containing voice recordings from both healthy participants and participants diagnosed with CHF, which contains voice alterations reflective of CHF status. We propose a novel deep machine learning architecture based on mode driven memory fusion for CHF recognition from audio recordings of subject's speech. We have achieved 90% accuracy under a subject-independent evaluation setting, highlighting the applicability of such methods for tele-health and home monitoring applications.

Index Terms: Chronic Heart Failure, Data Fusion, Geometric Deep Learning, Mode Based Explicit Memory

1. Introduction

Recent advances in deep learning have seen acoustic signal processing applied to medical applications [1, 2, 3]. Various mobile applications using acoustic data, and smart devices have been developed to identify and monitor diseases including Parkinson's disease [4, 5] and lung diseases [6, 7] in non-clinical and non-invasive settings.

Congestive Heart Failure (CHF) is a progressive condition, which lowers quality of life, and can ultimately lead to a complete heart failure. This occurs when the heart is unable to pump a sufficient volume of blood to essential organs around the body. Patients need to be knowledgeable about the different stages of the disease and monitor their progress closely after diagnosis to improve their life expectancy. The lack of resources and/or knowledge can reduce an individual's chances of a CHF diagnosis, which can be fatal, and the risk of a missed diagnosis increases for people in rural areas, or under strict restrictions during a pandemic. Thus, it is critical to develop applications that can be used to self-assess CHF symptoms and stages in a patient-centric and cost-effective way, and which can direct patients for immediate professional medical assistance.

Several studies have outlined the relationship between hydration and vocal cord vibration as a result of changes in the viscosity of the vocal cord tissue [8, 9], and have shown that this change of viscosity directly manifests in a person's voice. Based on this principle, voice-based bio-markers have been used to measure pulmonary congestion or breathlessness [10]. Since pulmonary congestion is the most common symptom of heart failure and one of the major contributors to hospitalisations [11, 12], our research utilises impairments observed in a patient's speech as a result of breathlessness, as a bio-marker to monitor CHF. In this paper, we report on a novel deep learning algorithm based on the principles of self-supervised transfer learning to analyse voice recordings as a bio-marker for heart failure.

Disease diagnosis and monitoring using voice bio-markers has precedence in some therapeutic areas, like Parkinson's disease ([13, 14]). Harimoorthy *et al.* proposed a cloud-based system to identify Parkinson's disease by applying machine learning to voice data [15]. Similarly, Asmea *et al.* proposed a neural network to identify Parkinson's disease that sought to identify traits of the voice disorder, dysphonia [16]. Acoustic features used for the above applications are hand-crafted and thus domain expertise is needed. Karaman *et al.* proposed a deep learning-based model for rapid diagnosis using bio-marker derived voice signals [17]. This method was integrated into a smart-electronic device and demonstrated success as a pre-diagnosis tool. These studies do not examine the potential uses of this technology for CHF monitoring. Our aim is to non-invasively identify CHF from speech recordings. Most promisingly, our technology has implications for tele-health and home monitoring applications, as already seen in other therapeutic areas.

2. Methodology

This paper presents a first-of-its-kind study using speech analysis to identify CHF. We report on the development of a novel dataset and have benchmarked this dataset on selected machine learning and transfer learning models. Furthermore, we propose a novel mode based memory fusion architecture (detailed in Section 2.1), where robust acoustic features are extracted from different audio representations, and fused via a neural memory module to improve CHF detection performance. A block diagram of our proposed method is presented in Figure 1.

In this study, we use two different representations of human speech (raw-audio and a 2-D representation of MFCC spectrograms) as inputs to the network to learn informative and robust features. We have utilised SoundNet [18] as the acoustic feature encoder, and VGG-16 [19] as the MFCC spectro-

Copyright © 2022 ISCA

10.21437/Interspeech.2022-643

gram image-based feature extractor. Feature encoders produce a low-dimensional vector representation of high-dimensional data, called *embeddings*.

For our first mode, we use a pre-trained SoundNet to extract acoustic features. Due to differences between the SoundNet training domain and the target domain, we extract our output feature embedding from after the 5^{th} convolution block instead of the last convolution block (8^{th}). This choice was made as the deeper layers (i.e. 8^{th}) tend to learn more objective specific features, while shallower layers (5^{th}) learn audio-specific (i.e. low-level) features, resulting in better performance when there is a significant domain difference. The output embedding is flattened to obtain a 512-D embedding which forms one input (x_s) to the fusion network. Our second mode uses a VGG-16 model pre-trained on ImageNet [20] to extract features from MFCC spectrogram images, obtained by stacking MFCC, delta and delta-delta features of an audio segment (refer to Section 3). We replaced all the fully connected layers in the original implementation with an Average Pooling layer and the resultant flattened embedding (512-D) is used as the second input (x_v) to the fusion network.

Fusion of different modes is proven to improve the performance of a neural network, if carefully designed [21, 22, 23]. However, naive fusion using feature concatenation led to degradation overall, compared to transfer learning on SoundNet alone (see Table 1). To achieve effective fusion of our two modes, we propose a novel mode-based memory fusion. We propose a novel memory network that receives features from different encoder networks and returns a fused representation that incorporates both current features and historical data stored in the memory.

2.1. Mode Based Memory Fusion

As both the MFCC and audio data corresponding to the same sample, the embeddings carry complementary information. To capture the complementary information in the embeddings of the two modes, we use a Graph Convolution Network with multi-head attention (GAT) [24, 25] (see Figure 1). Each embedding is a node in the graph, and we create a complete graph with self cycles. We obtain the mean of the graph structure (to match the graph output dimension with that of the input embedding) and use it as a third mode (x_{sv}) in the proposed architecture, helping the network learn from changes in temporal patterns in the data. Then, we pass each input embedding through separate explicit memory modules (NMN) as illustrated in Figure 1.

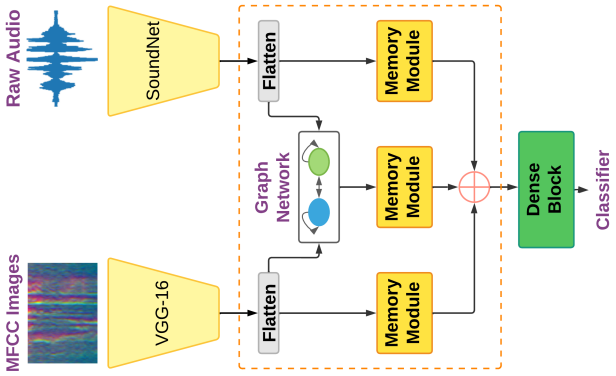

Figure 1: *High-level system architecture. Within the fusion network, the output embedding of the two encoders is fused using a Graph Convolution Network, obtaining a third embedding that captures the relationship between the two inputs. Then, all three embeddings are passed through an NMN to learn long-term relationships within the data. The resultant outputs of the memory modules are then fused and passed to a block of fully connected layers to perform classification.*

Memory networks consist of an explicit storage block, and components to handle read, write, and compose operations [26, 27]. They were proposed to better learn long-term dependencies in sequential data, and have been used in applications including trajectory prediction, autonomous driving and emotion analysis [28, 29, 30]. Memory networks, unlike LSTMs, greatly reduce the chance of forgetting important information due to their explicit consideration of the entire history. However, the performance of such a system may lead to inconsistencies in any common representation learned from fused data when modalities are vastly different in terms of their representation, necessitating specific memory modules for different modalities or feature representations.

The proposed memory layer consists of three major submodules: read, compose and write; along with an explicit memory block and controller (see Figure 2). x_i is the input to the memory layer (where $i \in (s, v, sv)$) of dimension d (set to 512), and the layer outputs a vector, o_i, of dimension d. The memory block ($M \in R^{n,d}$) consists of n independent memory slots containing feature vectors of dimension d, capturing mode-specific information (see Figure 2). All memory slots are randomly initialised using a Xavier uniform distribution [31], and are updated during training.

First, the feature vector, x_i, is used to generate the key vector, z_i, which the controller uses to determine the set of memory locations relevant to x_i (see Equation 1). z_i captures the degree of association between x_i and M_{t-1} (M_{t-1} represents the state of the memory before the forward pass in a mini-batch during training). The read module uses z_i and M_{t-1} to retrieve

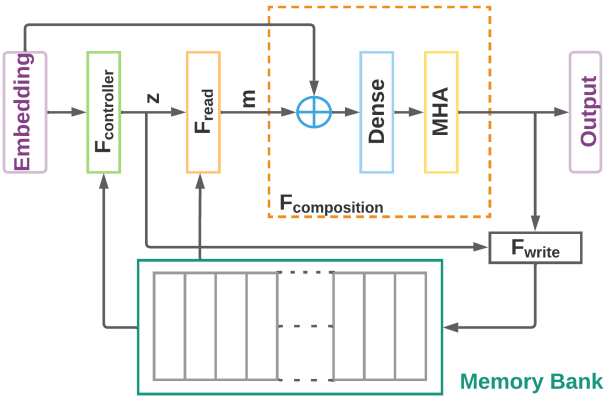

Figure 2: *Proposed memory structure: The explicit memory contains n memory slots, each of size d, to capture long-term relationships within the data. The memory controller calculates the key vector (z_i) which is used to read from and write to the appropriate memory slot. The memory vector corresponding to the input feature vector is obtained through the weighted average of all memory slots, where weights are determined using the key vector. The retrieved memory and the input feature are fused together and passed to the memory composition sub-module, and multi-head attention is applied to focus on salient features. The output of the composition is the output of the memory module and is used to update the best matching memory slot.*

a weighted feature vector from the memory, $m_{i,t}$, which is the weighted sum of all slots as defined in Equation 1, where f_r refers to the read function in the memory module,

$$z_i = softmax(f_r(x_i\, M_{t-1})) \text{ and } m_{(i,t)} = {z_i}^\top\, M_{t-1}. \quad (1)$$

The retrieved memory vector ($m_{i,t}$) is vertically concatenated with the input feature and passed to the memory composition sub-module. The resultant vector is averaged and passed through a fully connected layer to make the composition learnable [28], and to reduce the dimensionality of the composed vector to that of the input. Via this, the composition component learns semantic features from a fused vector space, combining present and historical information. Multi-head attention is applied to individual compositions to focus on salient data in the composed features in different feature spaces, allowing the model to jointly attend to features from different representation sub-spaces [32]. The composition output, c_i, is obtained using Equation 2, where f_m, f_{mha}, FC_i, f_{Re} and $\oplus$ represent the mean function, multi-head attention, fully connected layer for input mode i, activation function, and vertical concatenation respectively,

$$c_i \;=\; f_{Re}(f_m(f_{mha}(f_{Re}(FC_i(f_m(x_i \oplus m_{i,t})))))). \quad (2)$$

The composition output is written back to the memory through the write module using the key vector, z_i. Although we consider the weighted average of all the memory slots during the retrieval of $m_{i,t}$, only the memory slot with the highest importance (slot l, which has the highest softmax value in z_t) is updated. We obtain M_t (the updated memory block state) by replacing slot l with the average of slot l's feature vector and c_i. The memory layer output (o_i) is the composed vector c_i and represents the current input augmented by relevant long-term knowledge. We calculate the output of the three memory modules simultaneously and fuse them through horizontal feature concatenation, to obtain the output of the fusion layer (1536-D). Concatenation of the memory outputs helps the classifier to focus on important features among them as a whole. The fusion output is finally passed through two fully connected layers of size 512 and 2, where the latter is used as the classifier.

3. Dataset and Experimental Setup

The collected dataset contains speech recordings from 74 participants (37 healthy and 37 subjects diagnosed with CHF), recorded under different settings. The Male/Female ratio was kept at 50/50. All unhealthy participants have been clinically diagnosed with CHF by medical professionals. Each participant was asked to read aloud the text in 5 selected paragraphs that evoke distinguishable speech traits associated with CHF. Each recording was annotated with the participant's health condition (healthy or diagnosed with CHF), and the NYHA breathlessness score (not used in this study). Although the scripts were identical, due to variation in each reader's pace, the length of each voice recording differed. All audio recordings were re-sampled at $22,050Hz$, converted to mono-channel, and re-scaled to a [-256,256] range. Each recording was segmented into $5s$ chunks, with a 25% overlap (used to increase the dataset size). No padding is used on the signals, thus any remaining samples that cannot fill a $5s$ window are clipped. This results in a total of $5,640$ samples. Each sample is assigned the label of the original recording (1 if the recording is from a healthy participant, 0 otherwise).

The proposed model uses two encoder networks (transfer learning) as feature extractors (SoundNet and VGG-16, see Section 2). For machine learning models and SoundNet, we use raw-audio as inputs. However, only in SoundNet each chunk is reshaped to $(1, 110250, 1)$. For VGG-16, we use MFCC spectrogram images. We calculate 32 MFCC features (the minimum input dimension for VGG is 32) for each chunk using Librosa's [33] default settings, resulting in a feature of shape $(32, 216)$. Delta and delta-delta features are calculated using the above MFCC features. All features (MFCC, delta and delta-delta) are separately re-scaled to [0-255] (to match the RGB colour scale) and stacked to obtain an image representation of an audio chunk using MFCC spectrograms.

We evaluate the proposed model and other benchmark methods under two different subject-independent evaluation protocols: Leave-One-Subject-Out (LOSO) Cross-Validation, and Leave-One-Subject-Group-Out (LOSGO) Cross-Validation. In the LOSO protocol, samples are split into 74 folds. Each fold contains samples for one subject, and the number of samples in each fold varies. We split folds in a 72:1:1 ratio. 72 folds are used for training, and one fold each is used for validation and testing. As such, 74 trials are conducted such that each fold is considered as the test fold once. Validation and test folds both belong to the same subject class (healthy or diagnosed with CHF). During LOSO cross-validation experiments, we observed that model achieves the highest validation performance during the first few epochs, prior to convergence. Therefore, we followed two settings when selecting the best validation model for LOSO cross-validation: 1) the first occurrence of the highest validation accuracy (denoted LOSO-1), and 2) the last occurrence of the highest validation accuracy (denoted LOSO-2).

For the LOSGO evaluation, samples are divided into 10 folds, each containing a unique set of subjects. Folds are manually crafted to ensure similar proportions of healthy subjects and subjects diagnosed with CHF, resulting in 6 folds with 7 subjects and 4 folds with 8 subjects. Folds are split following a 8:1:1 ratio (8 folds for training, 1 for validation, and 1 for testing); and 10 trails are conducted such that each fold is the test fold once.

In all settings, accuracy is chosen as the evaluation metric as the classes are balanced. Inference is carried out using the best performing validation model. All models were trained with the Adam optimizer [34] with a learning rate of 0.001, a batch size of 64, and categorical cross-entropy loss. Networks are trained for 20 epochs in the LOSO setting, and for 80 epochs with LOSGO setting with early stopping (if the validation accuracy doesn't increase for 15 consecutive epochs after the 20^{th} epoch). The learning rate was reduced to 0.0001 after 50 epochs. Experiments are completed on a high performance computing cluster with an NVidia M40 GPU, $30GB$ of memory, and 10 CPU cores.

4. Results

In this section, we present the evaluation of benchmark models and the proposed memory-based fusion for CHF recognition using speech processing. Table 1 presents the performance of each model when evaluated with the subject independent evaluation protocols (see Section 3). For conventional machine learning approaches, we only use the LOSO-1 evaluation as there is no analogue to training epochs for these methods. Conventional machine learning models achieved performance above random chance (50%), indicating that voice recordings have

Table 1: *Performance comparison (accuracy) of benchmark methods for subject-independent CHF recognition.*

Model	LOSO-1	LOSO-2	LOSGO
SVM	76.17%	–	78.01%
Naive-Gaussian	50.47%	–	68.15%
Decision-Tree	73.54%	–	73.44%
Random Forest	77.04%	–	78.72%
Extra Trees	76.39%	–	78.95%
Voting	76.86%	–	78.37%
VGG-16	85.93%	-	74.86%
SoundNet	87.74%	88.72%	79.46%
Naive Fusion	82.59%	85.50%	84.86%
Memory Fusion	**89.52%**	**90.38%**	**84.87%**

the potential to identify the presence of CHF. A significant performance difference between LOSO and LOSGO was not observed with the above methods, except for the Naive-Gaussian classifier. Transfer learning on SoundNet resulted in an 11% gain in LOSO settings over the best conventional machine learning model, due to the robust features learnt from the data and the ability to learn a complex decision boundary. However, similar gains are not observed for the LOSGO protocol, likely due to the reduced subjects/samples available during training, and domain differences across samples. Applying transfer learning to VGG-16 with MFCC images achieved similar performance to SoundNet for LOSO evaluations, yet performance decreased by 5% in the LOSGO setting. This drop is likely due to domain differences (calculated MFCC images are very different from the ImageNet dataset) and the reduced data available in the LOSGO setting. Although the VGG-16 model recorded decreased performance for LOSGO, we observe that this model offers value when fused with SoundNet. Considering the Naive Fusion method, we observe a 5% performance drop compared to SoundNet alone in the LOSO evaluation; yet with LOSGO we observed a 5% improvement in performance compared to SoundNet (10% improvement over VGG alone), indicating that well designed fusion can improve performance.

The proposed memory fusion approach results in improvements of 7% and 5% for LOSO-1 and LOSO-2 over the naive fusion approach, and achieves similar results during LOSGO cross-validation. This suggests that the memory-enabled fusion can capture and store long-term dependencies in the data to improve accuracy. Furthermore, self-attention applied to the memory augmented input features helps focus on salient elements of the feature vector, filtering out uninformative elements. Results suggest that the use of modality (or representation) specific memory banks along with the fusion layer has allowed the model to leverage useful past features, which provides sta-

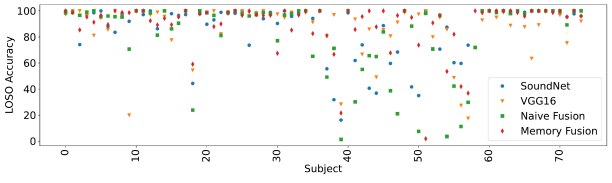

Figure 3: *Subject-wise performance First 37 subjects are diagnosed with CHF and the rest are healthy.: It is observed that memory based fusion is achieves comparatively higher performance and exceeds the chance level prediction (50%) for the majority of subjects, highlighting the superiority of the method.*

Table 2: *Detailed analysis of subject independent (LOSO-1) performance of benchmark methods for CHF recognition.*

Model	Healthy Acc.	With CHF Acc.
SVM	60.88%	78.02%
Naive-Gaussian	75.78%	68.16%
Decision-Tree	71.50%	73.42%
Random Forest	67.01%	78.72%
Extra Trees	63.78%	78.95%
Voting	60.92%	78.37%
SoundNet	82.36%	93.11%
VGG-16	78.09%	**93.77%**
Naive Fusion	72.57%	92.61%
Memory Fusion	**85.36%**	93.68%

ble inference. Figure 3 illustrates the subject-wise performance of the deep learning models, where the proposed memory fusion has resulted in higher performance for the majority of subjects. Similar results with and without the memory-based fusion across the LOSGO evaluation are likely a result of the dataset size, where approximately 16 subjects were kept out during the training with LOSGO cross-validation resulting in reduced subject variations in the data compared to LOSO, where only two subjects were held out.

Table 2 illustrates the class-wise variation in performance, where "Healthy" indicates the average performance when healthy patients were used as the test set and vice versa. It is observed that most of the machine learning models achieved a higher average accuracy for subjects diagnosed with CHF. Similar behaviour is observed with deep learning-based models. Even though the average accuracy of subjects diagnosed with CHF is slightly lower compared to the VGG16 model (0.09%), the average accuracy of healthy patients has increased (3% compared to the next best), resulting in higher overall performance with the memory fusion highlighting the importance of an explicit memory.

5. Conclusion

This paper proposes a novel method for screening of Congestive Heart Failure (CHF) through speech analysis. The proposed approach can be used in tele-health and home monitoring applications as a screening tool, to assist medical professionals and provide better patient care. We have collected and annotated a speech dataset consisting of healthy subjects and subjects diagnosed with CHF. The dataset will be valuable for developing and validating machine learning algorithms to automatically detect CHF from the speech impairments resulting from pulmonary congestion and will be made publicly available to researchers after acquiring the required clearances. We have benchmarked the novel dataset against conventional machine learning methods and demonstrated the applicability of speech traits for recognition of CHF symptoms. A major contribution of the paper is novel neural memory network-based (NMN) fusion architecture to improve recognition performance. As opposed to naive fusion (horizontal concatenation) that only considers the current input features, the proposed memory-based attentive fusion incorporates both the current input and long-term dependencies to improve performance. The proposed NMN-based fusion architecture provides substantial improvement in performance compared to standard machine learning approaches.

6. References

[1] T. Fernando, S. Denman, D. Ahmedt-Aristizabal, S. Sridharan, K. R. Laurens, P. Johnston, and C. Fookes, "Neural memory plasticity for medical anomaly detection," *Neural Networks*, vol. 127, pp. 67–81, 2020.

[2] T. Fernando, H. Ghaemmaghami, S. Denman, S. Sridharan, N. Hussain, and C. Fookes, "Heart sound segmentation using bidirectional lstms with attention," *IEEE journal of biomedical and health informatics*, vol. 24, no. 6, pp. 1601–1609, 2019.

[3] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, H. Ghaemmaghami, and C. Fookes, "A robust interpretable deep learning classifier for heart anomaly detection without segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 6, pp. 2162–2171, 2020.

[4] S. Arora, V. Venkataraman, S. Donohue, K. M. Biglan, E. R. Dorsey, and M. A. Little, "High accuracy discrimination of parkinson's disease participants from healthy controls using smartphones," in *2014 ICASSP*. IEEE, 2014, pp. 3641–3644.

[5] T. Arias-Vergara, J. C. Vasquez-Correa, J. R. Orozco-Arroyave, P. Klumpp, and E. Nöth, "Unobtrusive monitoring of speech impairments of parkinson's disease patients through mobile devices," in *2018 ICASSP*. IEEE, 2018, pp. 6004–6008.

[6] D. B. Chamberlain, R. Kodgule, and R. R. Fletcher, "A mobile platform for automated screening of asthma and chronic obstructive pulmonary disease," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 5192–5195.

[7] F. Zubaydi, A. Sagahyroon, F. Aloul, and H. Mir, "Mobspiro: Mobile based spirometry for detecting copd," in *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2017, pp. 1–4.

[8] K. Verdolini, Y. Min, I. R. Titze, J. Lemke, K. Brown, M. van Mersbergen, J. Jiang, and K. Fisher, "Biological mechanisms underlying voice changes due to dehydration," 2002.

[9] N. P. Solomon, L. E. Glaze, R. R. Arnold, and M. van Mersbergen, "Effects of a vocally fatiguing task and systemic hydration on men's voices," *Journal of Voice*, vol. 17, no. 1, pp. 31–46, 2003.

[10] O. M. Murton, R. E. Hillman, D. D. Mehta, M. Semigran, M. Daher, T. Cunningham, K. Verkouw, S. Tabtabai, J. Steiner, G. W. Dec *et al.*, "Acoustic speech analysis of patients with decompensated heart failure: a pilot study," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. EL401–EL407, 2017.

[11] G. C. Fonarow *et al.*, "The acute decompensated heart failure national registry (adhere): opportunities to improve care of patients hospitalized with acute decompensated heart failure." *Reviews in cardiovascular medicine*, vol. 4, pp. S21–30, 2003.

[12] P. B. Adamson, W. T. Abraham, M. Aaron, J. M. Aranda Jr, R. C. Bourge, A. Smith, L. W. Stevenson, J. G. Bauman, and J. S. Yadav, "Champion trial rationale and design: the long-term safety and clinical efficacy of a wireless pulmonary artery pressure monitoring system," *Journal of cardiac failure*, vol. 17, no. 1, pp. 3–10, 2011.

[13] B. Karan, S. S. Sahu, and K. Mahto, "Parkinson disease prediction using intrinsic mode function based features from speech signal," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 249–264, 2020.

[14] K. Polat, "A hybrid approach to parkinson disease classification using speech signal: The combination of smote and random forests," in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*. IEEE, 2019, pp. 1–3.

[15] K. Harimoorthy and M. Thangavelu, "Cloud-assisted parkinson disease identification system for remote patient monitoring and diagnosis in the smart healthcare applications," *Concurrency and Computation: Practice and Experience*, p. e6419, 2021.

[16] O. Asmae, R. Abdelhadi, C. Bouchaib, S. Sara, and K. Tajeddine, "Parkinson's disease identification using knn and ann algorithms based on voice disorder," in *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*. IEEE, 2020, pp. 1–6.

[17] O. Karaman, H. Çakın, A. Alhudhaif, and K. Polat, "Robust automated parkinson disease detection based on voice signals with transfer learning," *Expert Systems with Applications*, vol. 178, p. 115013, 2021.

[18] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," *Advances in neural information processing systems*, vol. 29, pp. 892–900, 2016.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[21] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of alzheimer's disease stage," *Scientific reports*, vol. 11, no. 1, pp. 1–13, 2021.

[22] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Computation*, vol. 32, no. 5, pp. 829–864, 2020.

[23] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Interpretable seizure classification using unprocessed eeg with multi-channel attentive feature fusion," *IEEE Sensors Journal*, vol. 21, no. 17, pp. 19 186–19 197, 2021.

[24] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[25] P. Lippe, "Graph neural networks," https://uvadlc.github.io/, 2021.

[26] J. Weston, S. Chopra, and A. Bordes, "Memory networks," *arXiv preprint arXiv:1410.3916*, 2014.

[27] Y. Ma and J. C. Principe, "A taxonomy for neural memory networks," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 6, pp. 1780–1793, 2019.

[28] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Going deeper: Autonomous steering with neural memory networks," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 214–221.

[29] T. Fernando, S. Denman, A. McFadyen, S. Sridharan, and C. Fookes, "Tree memory networks for modelling long-term temporal dependencies," *Neurocomputing*, vol. 304, pp. 64–81, 2018.

[30] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Memory based fusion for multi-modal deep learning," *Information Fusion*, vol. 67, pp. 136–146, 2021.

[31] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[33] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.