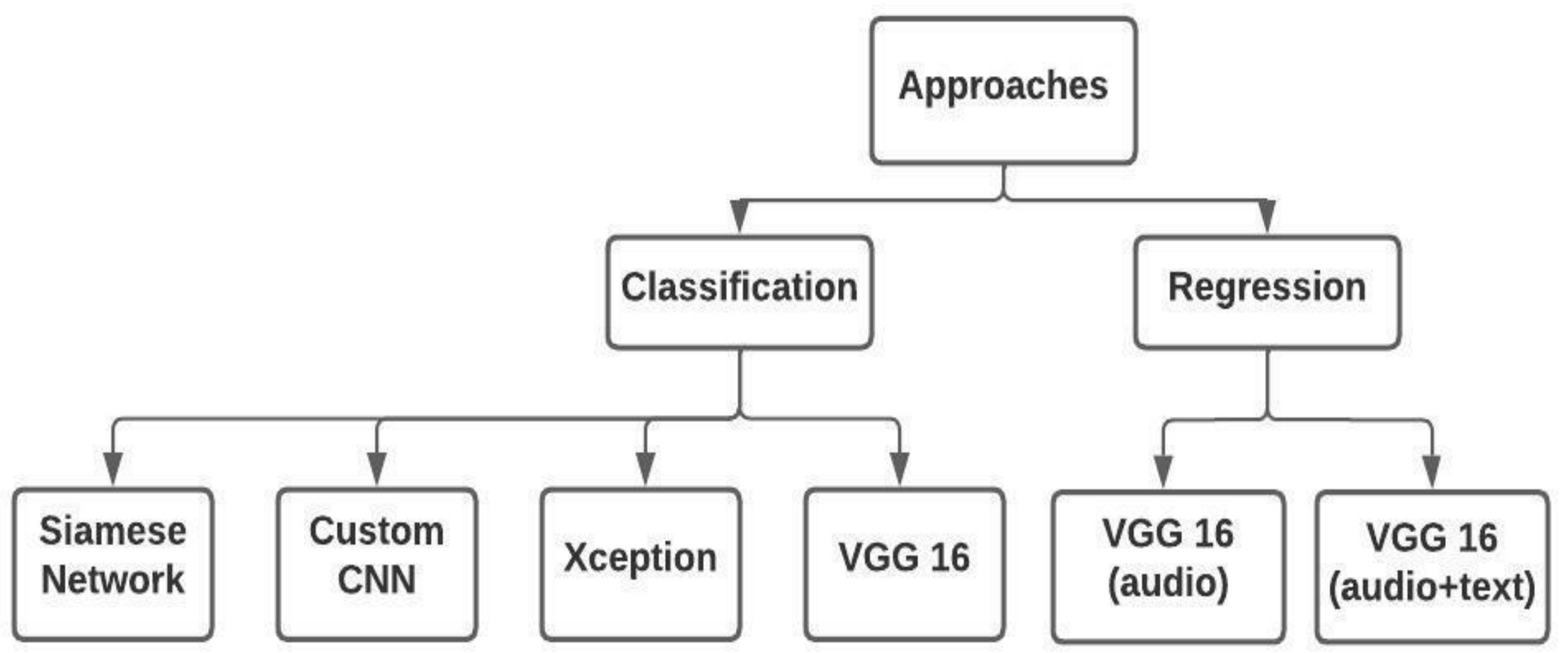
EmpathNet: Exploring Depression Detection from Semi-Clinical Interviews using Deep Learning

Pranav Vijay Chakilam, Dharani Doppalapudi, Harsha Renkila, School of Informatics and Computing, Indiana University

1. Introduction

- We developed a **Depression detection mode**l to aid **frequent and non-intrusive monitoring of depression levels**, to assist individuals take proactive steps to prevent further deterioration of their **mental health**.
- The **Extended Distress Analysis Interview Corpus (EDAIC)** dataset containing 219 sessions with durations from 7 to 33 minute was utilized.
- Built **spectrogram images** from the audio samples and trained **CNNs** to predict depression levels and **PHQ-8 scores**.



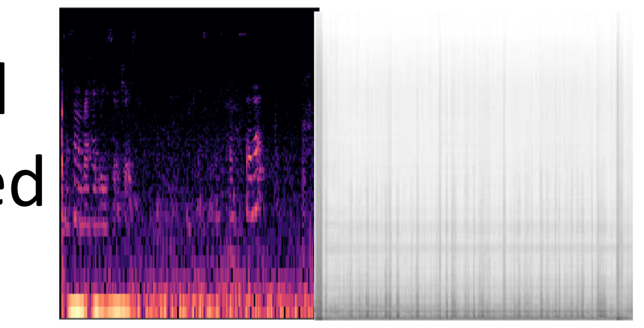
- *A **literature survey** of **35** articles was conducted before settling on the best model to utilize.* We' saw that various **CNNs, RNNs** and **Machine Learning** techniques were utilized to process and model audio-visual data.

2. Challenges

- Difficult to distinguish depression levels due to **common symptoms of mental diseases**.
- **Data augmentation** was challenging as even a minor error could result in loss of underlying data.
- Acquiring access to the dataset, and huge size of datasets created computational challenges.

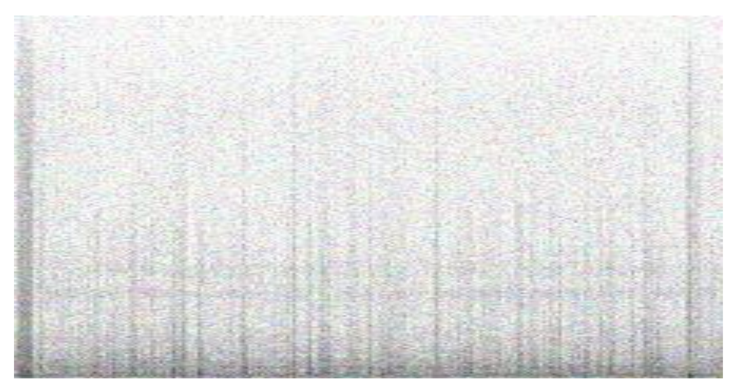
3. Feature Extraction and Data Augmentation

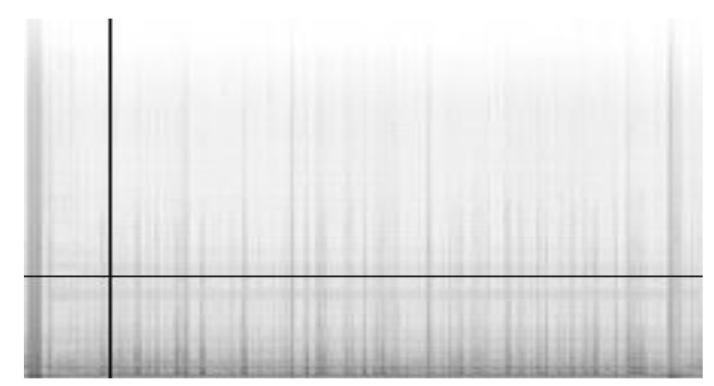
- **Mel-frequency and MFCC** based feature extraction was performed to transform audio samples to spectral images.



Audio Spectrograms

- Google's BERT (language model) was used to encode text descriptions. **Gaussian noise injection** and Google's **SpecAugment strategy** were used for data Augmentation as shown below.



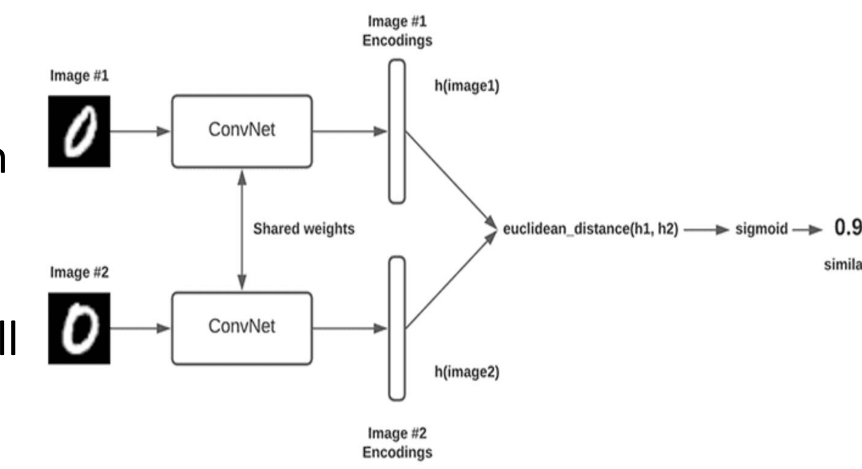


4. Models and Experiments

- For classification, in order to deal with the class imbalance, the **positive (depressed) data samples** were augmented using **gaussian noise injection** and **Spec Augment** strategy.
- We trained 34 models present in the Keras Applications library, pretrained on the ImageNet dataset, for 10 epochs each and selected the **Xception model** as the best performing model with an acceptable number of parameters and fine- tuned it.
- Additionally, we experimented with a **6-layer DNN with 3 convolutional layers** and **2 Fully Connected Layers** to perform classification without overfitting.
- Experimented with combining audio and text features. We utilized the **Vgg16 custom model** for audio and the **DNN** model for text characteristics. We then concatenated both features to improve classification.
- For regression of PHQ-8 score prediction, we used pretrained **Vgg16 model** and custom Vgg16 model with 5x5 conv filters. This was applied to all modules and fine-tuned them.
- In addition to only audio features, we have trained model with combination of audio and text features using **VGG16 for audio and simple DNN** for text.

Siamese network:

- When the data is **limited**, the Siamese network will compute similarity between the two images. As a result, the dataset will grow by a factor of **n^2**.
- For both input networks, the Siamese will use the same base network.



5. Results

Experiments are carried out in Google colab pro + using tensorflow

Model	Accuracy	Recall
Custom CNN	61%	84%
Xception	81%	68%
Siamese Network	69%	67%
Vgg16 (audio + text)	71%	69%

Model	Type of network	Text features	RMSE	MAE
Vgg16	Pretrained	No	76.79	6.74
Vgg16	Custom	No	43.74	5.72
Vgg16 (audio + text)	Custom	Yes	57.65	6.54

file and corresponding text transcript.

***Below:** PHQ-8 score prediction evaluation for a given audio fiie and text transcript using deferent models.*

Observations:

- The Mae score for regression is very high for custom vgg16 model but the outputs are biased towards PHQ-8 between 3-8 range but, the Vgg16 + text model is giving better results.
- The accuracy of the custom CNN produces better recall although it's accuracy is relatively low corresponding to Xception, suggesting that a model with a complexity between these 2 models might yield better results.

6. Future Work

- Create a multi-modal model which also takes **facial videos** as input along with audio and text.
- Explore time-domain based methods for data augmentation.
- Train the Siamese network on a larger dataset.

***Acknowledgments:** This work was funded by the US Department of Defense (Contract W52P1J2093009).*