Abstract

This study explores the application of the Qwen 2 Audio model for detecting depression through speech patterns, aiming to develop a more objective diagnostic tool compared to traditional self-reported assessments. We fine-tuned the model using the DAIC-WOZ dataset, comprising 189 patient interviews and over 9,000 processed audio samples. Despite achieving improvements in overall accuracy from 43.03% to 59.26%, further analysis revealed that the model's performance was primarily driven by majority class bias rather than learning meaningful depression indicators. Our findings highlight both the challenges and potential opportunities in developing audio-based depression detection systems, while emphasizing the need for larger, more diverse datasets.

1. Introduction

Depression is a significant mental health concern affecting over 264 million people globally[1]. Traditional diagnostic methods primarily rely on subjective assessments, such as patient-reported questionnaires like the Patient Health Questionnaire (PHQ), which depend on individuals' ability to accurately report their feelings and behaviors. This reliance on self-reporting can complicate the diagnostic process, due to varying self-perception and awareness in individuals, leading to variability in diagnosis and treatment outcomes[2]. Misdiagnosis or delayed diagnosis can result in individuals not receiving appropriate care, potentially worsening the severity of their condition[2].

In contrast, AI models offer the potential for more objective and accurate diagnoses. Patients with depression frequently exhibit reduced speech fluency and altered prosody, such as monotonous speech or diminished expressiveness[3]. Clinicians often pay close attention to changes in speech during the diagnostic process, although this subjective evaluation is susceptible to bias and requires substantial clinical expertise[3]. Previous studies have shown that vocal characteristics that may signal depressive states can be used in modeling depression diagnosis[4], so we are building on top of these studies by leveraging pretrained multimodal LLMs. More specifically, we will be fine tuning the Qwen 2 Audio model to take in 10 second audio clips of patient speech, and classify whether the patient is depressed. The goal is to develop a model that can accurately classify depression solely through short audio clips of patient speech, in order to replace or supplement the current self reported standard.

2. Background

2.1 Related Work

Advancements in artificial intelligence and machine learning have opened new avenues for diagnosing depression. In 2019, Zhang et al. [4] developed a multi-modal framework for depression assessment that processes audio, video, and textual features. Their approach combines

Deep Denoising Auto-Encoder technology with various neural network architectures to analyze multiple types of data simultaneously. More recently, researchers have begun exploring Large Language Models for depression assessment. In 2023, Sadeghi et al. [5] proposed a text-based architecture that leverages modern language models like GPT-3.5 Turbo and fine-tuned DepRoBERTa, along with Support Vector Regression for depression severity analysis. In 2024, Li et al. [6] introduced the Flexible Parallel Transformer (FPT) model, which specifically focuses on integrating video and audio features. Their testing on the E-DAIC dataset demonstrated improvements over simpler models, further supporting the value of multi-modal approaches.

While these studies demonstrate promising results, it's important to note that their reliance on multiple data types (audio, video, and text) may limit their practical applicability. In many real-world clinical settings, collecting high-quality multimodal data can be challenging due to privacy concerns or resource constraints. This motivated our approach to solve the problem, e.g. only using short audio samples of patient speech.

2.2 Qwen 2 Audio

In August 2024, Alibaba released the Qwen 2 Audio model[7], which outperformed previous state-of-the-art models, including Gemini-1.5-pro, in the audio analysis benchmarks without the need for any task-specific fine-tuning. The Qwen 2 audio model employs a dual-tower architecture consisting of an audio encoder and a language model. The audio tower processes raw audio features through convolutional layers and a stack of 32 transformer layers, ultimately encoding audio into a 1280-dimensional space. This encoded audio representation is then projected to match the language model's 4096-dimensional space via a linear projection layer, allowing the 32-layer language model to generate text responses conditioned on the audio input. Qwen 2 Audio has shown to perform well in picking up patterns between acoustic features like pitch, tone, and intonation, so we will be leveraging this new powerful open source model as the base for our developments.

3. Methodology

3.1 Dataset

We utilized a dataset (DAIC-WOZ[8]) obtained from the University of Southern California's Institute of Creative Technologies, comprising 189 distinct patient interview sessions. Each session was conducted between a participant and a virtual interviewer, designated as "Ellie", which was operated by a human interviewer in a separate location. The transcripts revealed dialog similar to those in a therapist session. The sessions averaged 16 minutes in duration, yielding approximately 50.4 hours of audio data. Prior to the experimental sessions, participants completed the Patient Health Questionnaire (PHQ-8), which provided a binary classification of depression status (depressed/not depressed) that served as the ground truth for our analysis. It is noteworthy that while this dataset provides valuable insights, the relatively limited number of

participants represents a methodological constraint that will be addressed in detail in the limitations section of this study.

3.2 Data Preprocessing

When examining the audio data, only around half of the audio was the participant speaking. We removed the remaining audio, which corresponds to the interviewer speaking and silence, and split the audio into 10 second clips to standardize the length of the samples. After this cleaning, there were 9,044 total samples.

Regarding class distribution, 31.62% of the samples were labeled “Depressed” and 68.38% were labeled “Not Depressed”. We were not particularly worried about the class imbalance as there were sufficient samples from the minority class, but one bias to note is the imbalance in the number of samples per speaker. The interview durations ranged from 7-33 minutes and a larger volume of signal from an individual may emphasize some characteristics that are person specific. We had considered addressing this potential source of bias by sampling an equal number of samples from each speaker, but this would drastically reduce the number of samples, from 9,044 to 2,481. This would likely not be enough samples to train a large model on, so we felt it would be better to address this potential bias in future work, when we have the necessary amount of data to do so.

We then employed roughly an 80/10/10 split for the train, validation, and test set respectively. The split is not exact, as we sampled based on participants in order to prevent spillover. We were concerned with the model learning to identify the speaker, rather than general patterns indicating depression, so we processed the datasets to include unique speakers only. Figure 1 shows the final label distribution of each set.

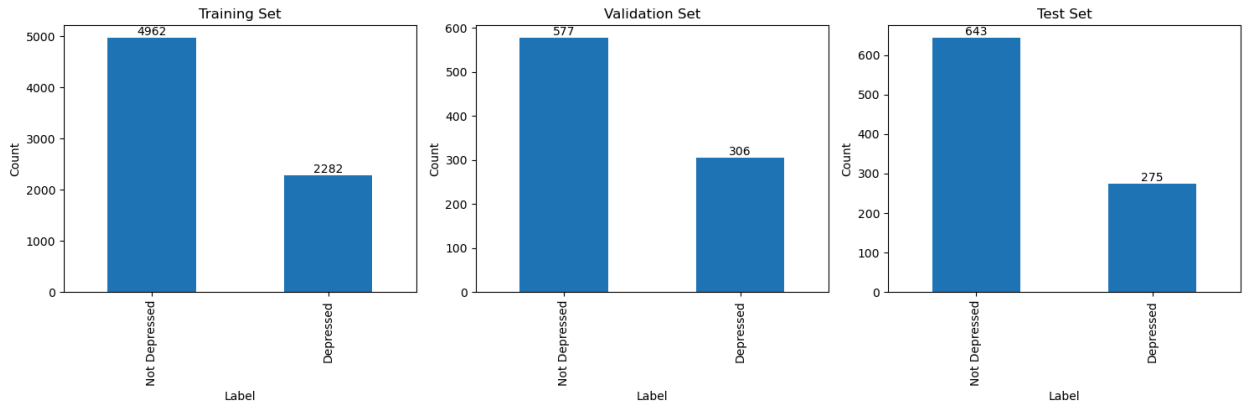


Figure 1: Label distribution for each dataset

Lastly, the input data was processed into a conversation format suitable for the Qwen 2 Audio model, structuring each interaction with a system prompt, user input containing the audio and text components, and an assistant response indicating the depression classification. The user

prompt asked for a binary classification of "Depressed" or "Not Depressed" based on the provided audio sample. The full template is shown in the Appendix.

3.3 Fine Tuning Methods

3.3.1 Lora Config

Our study employed the Qwen2-Audio-7B-Instruct model as the foundation for depression detection from audio samples. We utilized Low-Rank Adaptation (LoRA) for efficient fine-tuning, implementing a two-phase training approach with different learning rate schedules. The base model was augmented with LoRA adapters configured with a rank of 8, alpha value of 32, and a dropout rate of 0.1. For the target modules, we specifically wanted to target the audio encoder layers that process the speech features, cross-attention and projection layers that connect the audio understanding to text generation, and the language model layers that handle the final prediction. This ends up being roughly all the linear layers so we set the target modules to all linear layers for simplicity. Additionally, we employed RsLoRA, along with a causal language modeling task type. This configuration resulted in 27,210,752 trainable parameters, representing approximately 0.323% of the model's total 8,424,305,664 parameters.

3.3.2 Training Process

The training process was conducted in two phases, each with different learning rate scheduling approaches. In the first phase, we employed a linear learning rate schedule with an initial learning rate of 5e-5 for one epoch. The training data consisted of 7,244 total samples, with 2,282 samples labeled as depressed, while the validation set consisted of 883 total samples, with 306 samples labeled as depressed. We utilized a batch size of 1 with gradient accumulation steps of 8 to manage memory constraints. In the attempt to further improve the model performance and stability, we initiated a second training phase using a cosine learning rate schedule, with the same training parameters but allowing the learning rate to decay following a cosine curve over an additional epoch.

To address memory constraints and improve training efficiency, we implemented several optimization strategies. These included gradient checkpointing with disabled reentrant acceleration, evaluation on random subsets of 150 samples with an evaluation accumulation step of 16, and memory clearing between training steps. We also employed the half precision (fp16) version of the base model. With these settings, each epoch took a little less than 2 hours to complete, when training on one H100 SXM GPU.

Model performance was evaluated on the test set using multiple metrics including accuracy, AUC score, F1 score, and recall. Recall was specifically a metric of interest, as undiagnosed cases of depression are especially consequential. Each training phase's model was evaluated independently to assess the impact of the training on an additional epoch with a decaying learning rate schedule.

4. Results and Discussion

Model	Accuracy	AUC Score	F1 Score	Recall
Base Qwen 2 Audio	0.430	0.510	0.427	0.709
First phase model	0.593	0.473	0.204	0.175
Second phase model	0.590	0.455	0.145	0.116

Table 1: Evaluation Metrics of base and fine tuned models

Our experiments with fine-tuning the Qwen 2 Audio model produced disappointing results. As a baseline comparison, we also evaluated the performance of the base Qwen 2 Audio model with no fine tuning on the test set. The base model achieved an accuracy of 43.03%, while our fine-tuned models showed modest improvements. The first fine-tuning phase trained on one epoch achieved an accuracy of 59.26% on the test set, and the second phase trained on 2 total epochs maintained similar performance at 59.04%.

However, further analysis of metrics revealed concerning patterns in the model's behavior. While the accuracy improved, other metrics indicate that this improvement does not reflect learning of depression indicators. The AUC score decreased from 0.510 in the base model to 0.473 in our first fine-tuned version and 0.455 in the second version. Similarly, the F1 score decreased from 0.427 to 0.204 and then to 0.145 across the base, first, and second fine-tuned versions respectively. This is suggesting that the model's ability to distinguish between classes actually deteriorated. The model appears to have simply developed a stronger bias toward predicting "Not Depressed", the majority class in our training data, and that the fine tuned models achieved better accuracy by over-predicting the majority class, rather than from learning predictive indicators.

Additionally, the model's performance on detecting depression cases was extremely weak, with a recall score of 0.175 for the first phase and 0.116 for the second phase. This further supports that while overall accuracy improved, the model's ability to identify actual depression cases has not. This is particularly problematic given that missing depression cases (false negatives) is more consequential than false positives in clinical applications.

When comparing the confusion matrices of the evaluation results (shown in the Appendix), we see a similar story as we've discussed thus far. With the base Qwen 2 Audio model, the model predicted "Not Depressed" 30.5% of the time, while the first fine tuned model predicted "Not Depressed" 78.8% of the time, and the second fine tuned model predicted "Not Depressed" 82.0% of the time. Comparing specifically between the first and second phase models, it appears that the additional training led to further overfitting to the majority class rather than learning more nuanced features. This could indicate that our training strategy needs refinement, possibly

through techniques like balanced sampling or weighted loss functions to address the class imbalance issue.

5. Conclusion

Our investigation into using the Qwen 2 Audio model for depression detection through speech patterns yielded valuable insights despite not achieving the desired performance metrics. While the fine-tuned model showed improved accuracy over the baseline (59.26% vs 43.03%), deeper analysis revealed that this improvement was due to overfitting to the majority class rather than learning meaningful indicators of depression. This is particularly evident in the deteriorating AUC scores and the recall for depression cases, and is a good example of the importance of evaluating models using multiple metrics beyond accuracy.

5.1 Limitations and Future Work

A critical limitation of our study was the relatively small number of unique participants in the dataset (189 individuals). This limited participant pool likely hindered the model's ability to learn generalizable depression indicators, as the patterns it detected were likely more reflective of individual speaking characteristics rather than universal markers of depression. The challenge was further compounded by computational constraints, which restricted our ability to experiment more with different training configurations.

Despite these challenges, our work provides valuable insights for future research directions. The results suggest that depression detection through short clips of patient speech requires larger datasets with more diverse participants to enable learning of generalizable patterns. Additionally, our findings may suggest the need to shift from binary classification to more nuanced approaches that better reflect the spectrum nature of depression, particularly given the complexity of how depression manifests in speech patterns.

References

1. World Health Organization. (2017). Depression and other common mental disorders: global health estimates.

2. Z. Li, M. Ruan, J. Chen, and Y. Fang, “Major Depressive Disorder: Advances in Neuroscience Research and Translational Applications,” Neurosci. Bull., vol. 37, no. 6, pp. 863–880, Jun. 2021, doi: 10.1007/s12264-021-00638-3.

3. König, A., Tröger, J., Mallick, E. et al. Detecting subtle signs of depression with automated speech analysis in a non-clinical sample. BMC Psychiatry 22, 830 (2022). https://doi.org/10.1186/s12888-022-04475-0

4. L. Zhang, J. Driscol, X. Chen, and R. Hosseini Ghomi, “Evaluating Acoustic and Linguistic Features of Detecting Depression Sub-Challenge Dataset,” in Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, in AVEC ’19. New York, NY, USA: Association for Computing Machinery, Oct. 2019, pp. 47–53. doi: 10.1145/3347320.3357693.

5. M. Sadeghi et al., “Exploring the Capabilities of a Language Model-Only Approach for Depression Detection in Text Data,” in 2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Oct. 2023, pp. 1–5. doi: 10.1109/BHI58575.2023.10313367.

6. Li, Yifu, Yang, Xueping, Zhao, Meng, Wang, Zihao, Yao, Yudong, Qian, Wei, Qi, Shouliang, FPT-Former: A Flexible Parallel Transformer of Recognizing Depression by Using Audiovisual Expert-Knowledge-Based Multimodal Measures, International Journal of Intelligent Systems, 2024, 1564574, 13 pages, 2024. https://doi.org/10.1155/2024/1564574

7. Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., ... & Zhou, J. (2024). Qwen2-audio technical report. arXiv preprint arXiv:2407.10759.

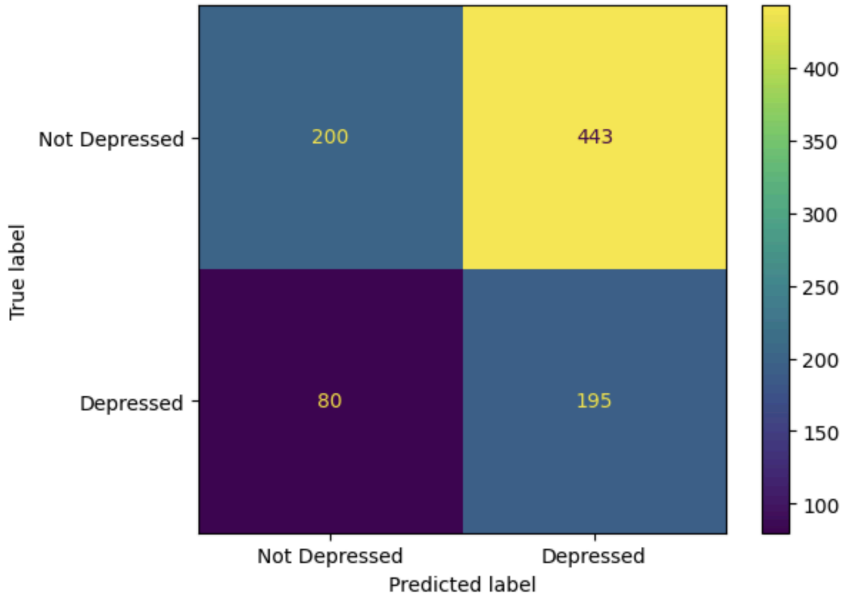
8. Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The Distress Analysis Interview Corpus of human and computer interviews. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Appendix

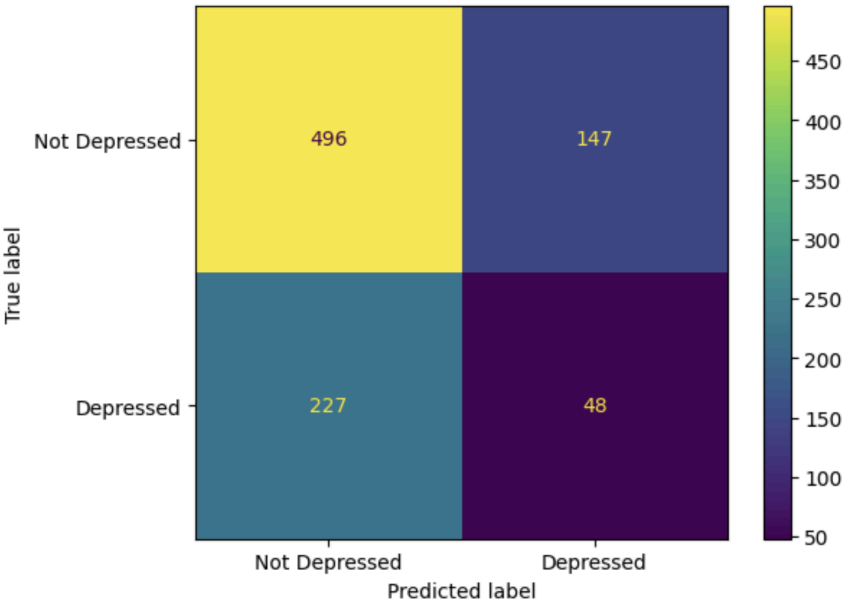
Conservation Template for Qwen 2 Audio Input:

conversation = [{'role': 'system', 'content': 'You are a helpful assistant.'}, {"role": "user", "content": [{"type": "audio", "audio_url": audio_path}, {"type": "text", "text": "Based on the audio, is the person depressed? Answer with either \"Depressed\" or \"Not Depressed\"."},]}, {'role': 'assistant', 'content': "Depressed" if label == 1 else "Not Depressed"},]

Confusion Matrix for Base Qwen 2 Audio Model:



Confusion Matrix for First Phase Fine Tuned Model:



Confusion Matrix for Second Phase Fine Tuned Model: