ONTOLOGY-REGULARIZED MULTIMODAL NEURAL NETWORK FOR EXPLAINABLE MENTAL HEALTH ASSESSMENT (OR-NN)

Bhavya Jain, Sumit Dalal, Bharat Bhargava

ISIC 2025, 06 October 2025

Introduction

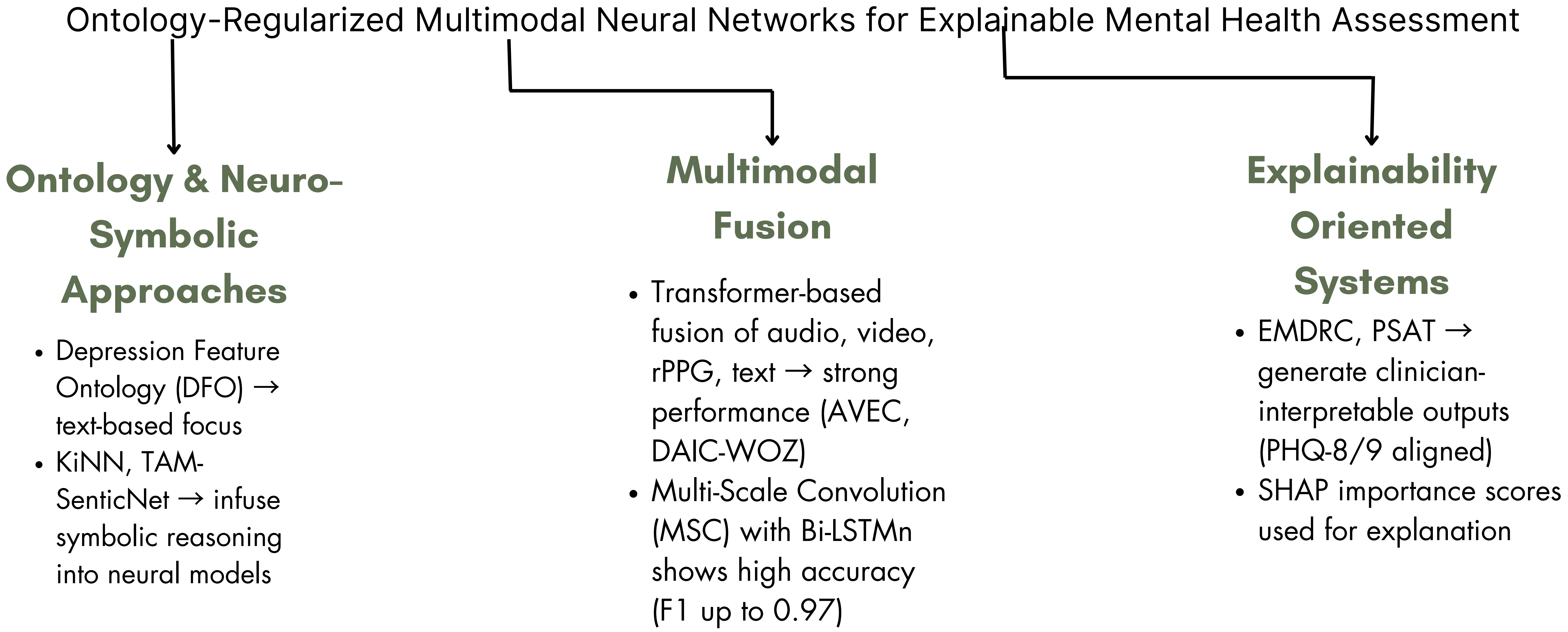
Motivation

- Depression affects 280M+ people worldwide (WHO)
- Traditional diagnosis is subjective, time-consuming and not scalable so we need objective and scalable detection methods
- AI models exist, but they are black-box, thus less trustworthy for clinicians.

Our Aim

- Provide clinicians interpretable predictions with explanations

Related work



Ontology-Regularized Multimodal Neural Networks for Explainable Mental Health Assessment

Ontology & Neuro-Symbolic Approaches

- Depression Feature Ontology (DFO) → text-based focus
- KiNN, TAM-SenticNet → infuse symbolic reasoning into neural models

Multimodal Fusion

- Transformer-based fusion of audio, video, rPPG, text → strong performance (AVEC, DAIC-WOZ)
- Multi-Scale Convolution (MSC) with Bi-LSTMn shows high accuracy (F1 up to 0.97)

Explainability Oriented Systems

- EMDRC, PSAT → generate clinician-interpretable outputs (PHQ-8/9 aligned)
- SHAP importance scores used for explanation

Gaps and Hypothesis

Gaps

- Multimodal models give good accuracy but poor interpretability
- Ontologies have been constructed only from text
- Explanations use cues from text modality only

Hypothesis

- Multimodal ontology (audio+video) could be infused in training
- Ontology-regularised neural network would result in better accuracy
- Using ontology would give clinically meaningful explanations

Dataset & Preprocessing

- DAIC/EDAIC multimodal dataset (audio + video interviews)

PreProcessing →

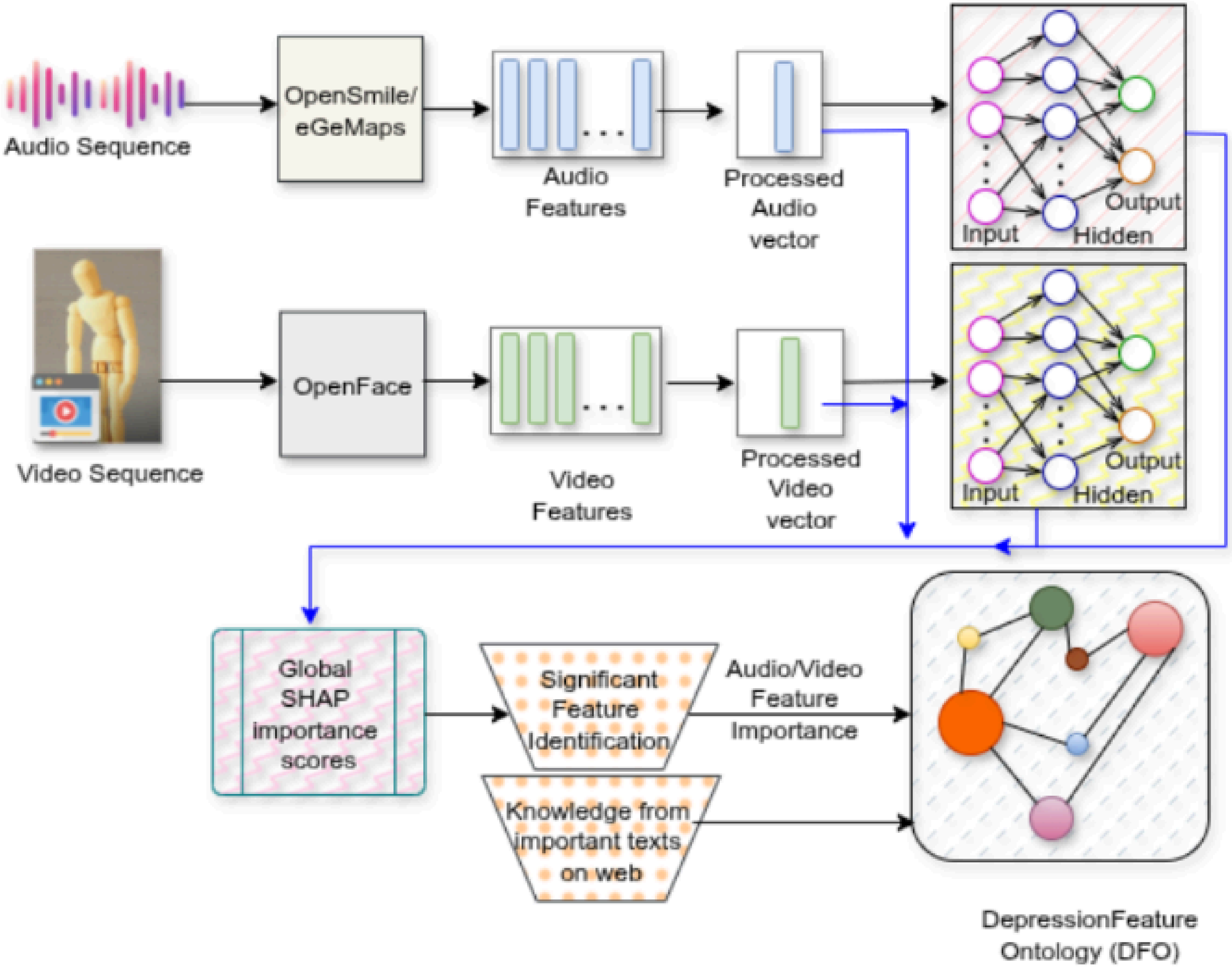
- Summarised frame-wise eGeMAPS and OpenFace features using different statistical methods to form fixed-length interpretable vectors for both audio and video
- Chose aggregation (mean/std) instead of concatenation across frames (like Fisher / temporal stacking) to ensure fixed lenght vectors

Why eGeMAPS and OpenFace? →

- Standardised, interpretable descriptors
- Feature names directly map to psychological constructs

Methodology

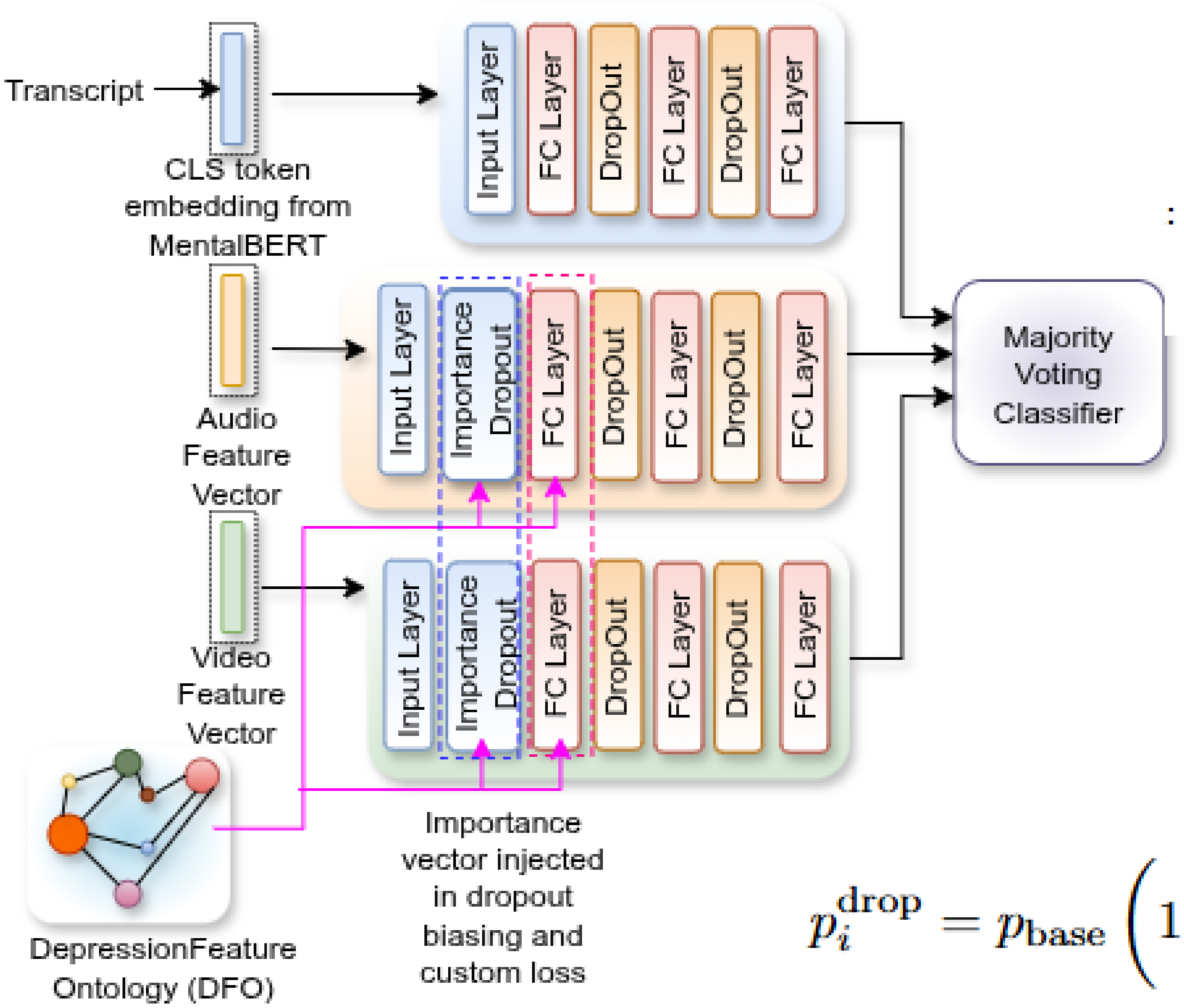
Ontology Construction (NN + SHAP and texts from web)



Audio Sequence
OpenSmile/ eGeMaps
Audio Features
Processed Audio vector
Input
Hidden
Output
Video Sequence
OpenFace
Video Features
Processed Video vector
Input
Hidden
Output
Global SHAP importance scores
Significant Feature Identification
Audio/Video Feature Importance
Knowledge from important texts on web
DepressionFeature Ontology (DFO)

Methodology

Ontology Knowledge Infusion



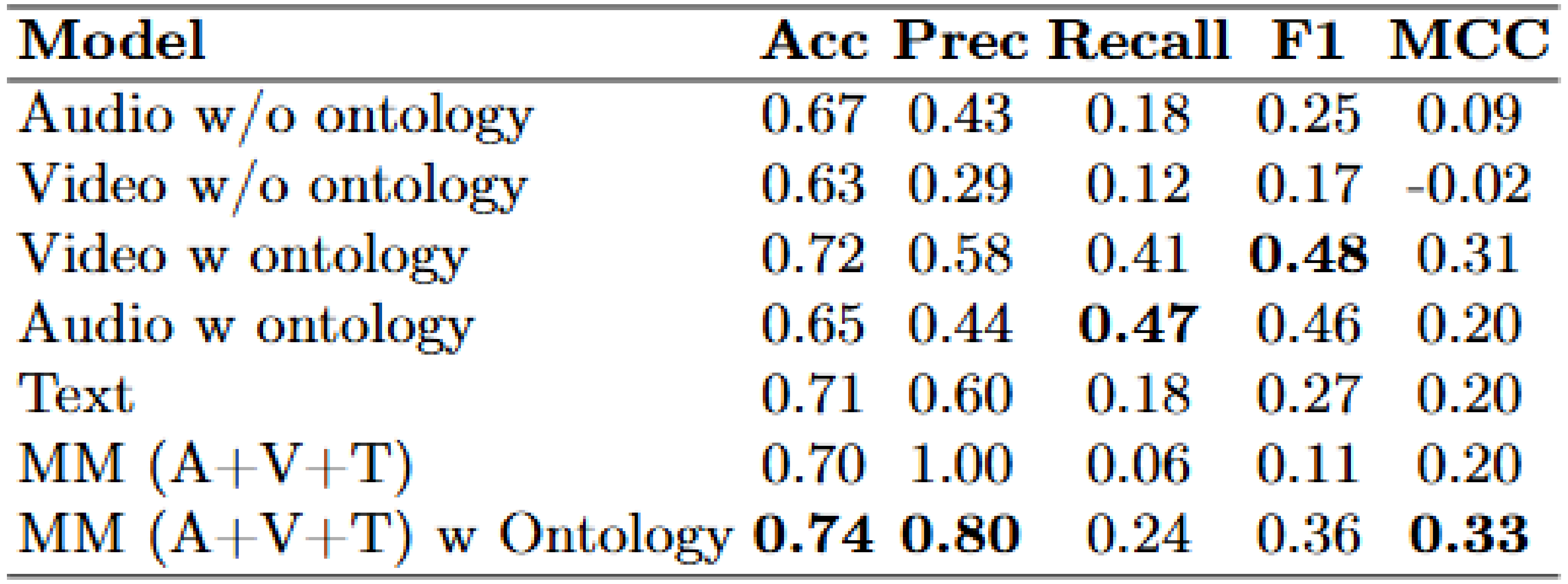
$$\mathcal{L}_{\text{reg}} = \frac{1}{d} \sum_{i=1}^{d} \left(\bar{w}_{1,i} - s_i \right)^2$$

:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \lambda \, \mathcal{L}_{\text{reg}},$$

$$p_i^{\text{drop}} = p_{\text{base}} \left(1 - \frac{s_i}{\max(s)} \right), \quad i = 1, \ldots$$

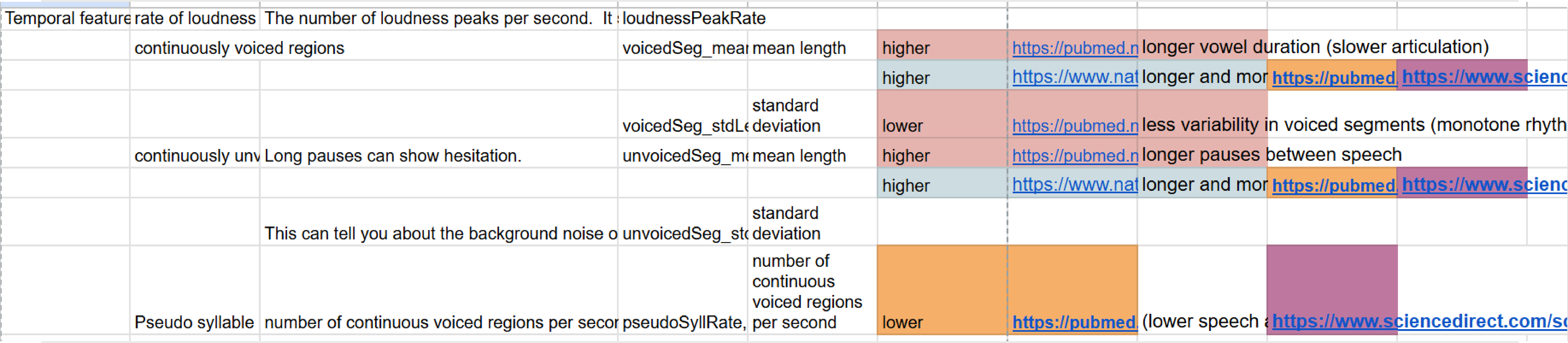
Analysis



Model	**Acc**	**Prec**	**Recall**	**F1**	**MCC**
Audio w/o ontology	0.67	0.43	0.18	0.25	0.09
Video w/o ontology	0.63	0.29	0.12	0.17	-0.02
Video w ontology	0.72	0.58	0.41	**0.48**	0.31
Audio w ontology	0.65	0.44	**0.47**	0.46	0.20
Text	0.71	0.60	0.18	0.27	0.20
MM (A+V+T)	0.70	1.00	0.06	0.11	0.20
MM (A+V+T) w Ontology	**0.74**	**0.80**	0.24	0.36	**0.33**

- Ontology guidance significantly improves recall for both audio and video, with video showing the largest F1 gain
- Ontology guidance boosts recall nearly fourfold compared to non-ontology fusion in multimodal setting
- Text: conservative (high precision, very low recall)
- Best configuration for multimodal uses both custom and dropout biasing in both audio and video models with better recall in probability averaging method for fusion

Preview of Ontology



Temporal feature	rate of loudness	The number of loudness peaks per second. It :	loudnessPeakRate						
	continuously voiced regions		voicedSeg_mear	mean length	higher	https://pubmed.n	longer vowel duration (slower articulation)		
					higher	https://www.nat	longer and mor	https://pubmed.	https://www.scienc
			voicedSeg_stdLe	standard deviation	lower	https://pubmed.n	less variability in voiced segments (monotone rhyth		
	continuously unv	Long pauses can show hesitation.	unvoicedSeg_me	mean length	higher	https://pubmed.n	longer pauses between speech		
					higher	https://www.nat	longer and mor	https://pubmed.	https://www.scienc
		This can tell you about the background noise o	unvoicedSeg_stc	standard deviation					
	Pseudo syllable	number of continuous voiced regions per secor	pseudoSyllRate,	number of continuous voiced regions per second	lower	https://pubmed.	(lower speech :	https://www.sciencedirect.com/sc	

Explanations without ontology and with ontology

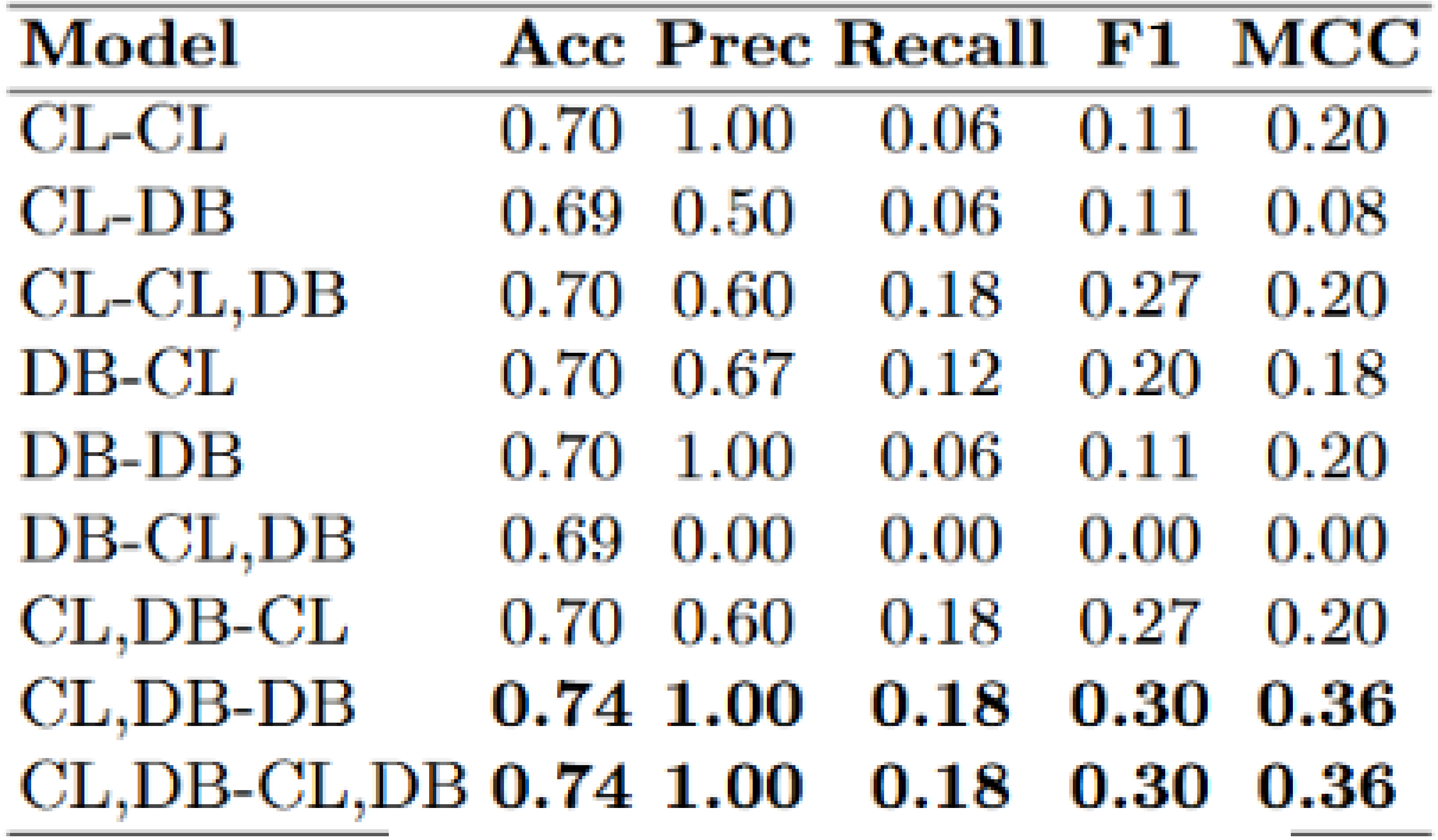
The person’s voice likely sounds flatter and less lively: their pitch stays within a small range so sentences don’t rise or fall much, and the spectral shape of the voice shows fewer rapid high-frequency changes, giving a duller, less bright timbre. At the same time there are signs of instability in the voice source — small, irregular perturbations and amplitude fluctuations — that make the voice occasionally breathy or wavering rather than clean and energized. The speech rhythm also appears compressed: there is reduced variation in how long voiced parts last from one utterance to the next (so vowels and syllables are more uniform and dragged), together with altered timing in pitch fall patterns, which produces long, steady stretches of speech with fewer lively bursts. Clinically, that cluster maps to **psychomotor slowing, low activation/fatigue, and blunted affect** — the kind of voice a clinician would describe as “soft, monotone, and tired,”

The individual’s voice shows a pattern of flatness and low emotional range. The narrow pitch range and limited pitch variation suggest a monotonous tone, where highs and lows in the voice barely shift — a cue often tied to emotional disengagement. Variations in spectral features (like the MFCC coefficients) indicate instability in articulation and timbre, hinting at vocal tension or effort. The uniform speech rhythm (low variability in voiced segments) further supports the idea that their speech may have lacked natural pacing and expressiveness. Subtle irregularities in jitter and shimmer reveal micro-tremors in the voice, often associated with fatigue or emotional strain. Meanwhile, loudness features point toward generally low but steady vocal energy, showing that the person likely spoke softly and with little dynamic range. Altogether, these patterns describe a **tired, effortful, and emotionally flat vocal delivery,** consistent with depressive affect and low engagement.

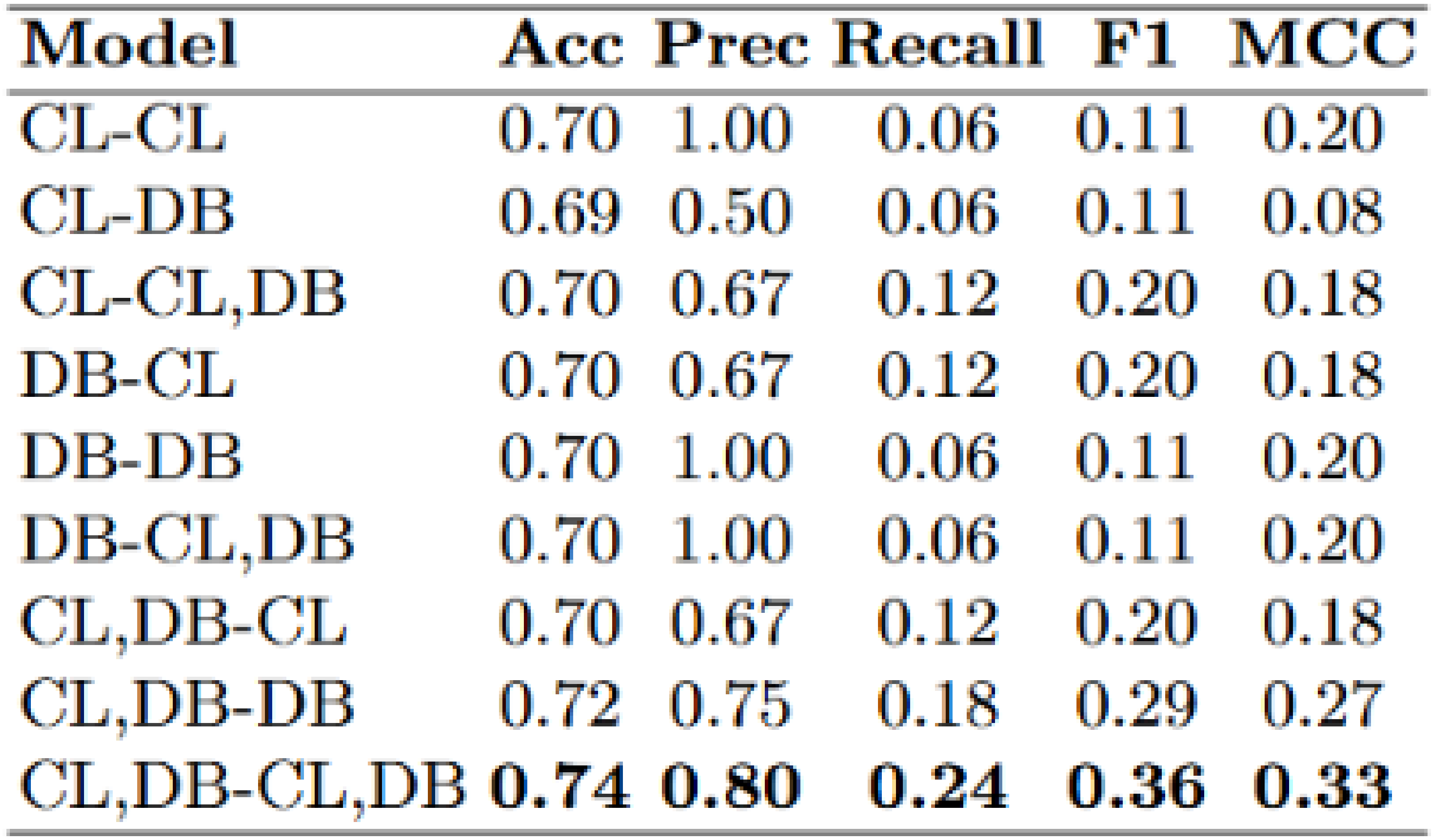
Future Directions

- SHAP analysis on transcripts was noisy so not suitable for direct DFO mapping so we will expand DFO carefully extracting depression-related words from the EDAIC dataset
- Transcripts lack interviewer questions so we plan to add that context in the transcript
- Temporally link audio, video, and text for richer, cross-modal reasoning.

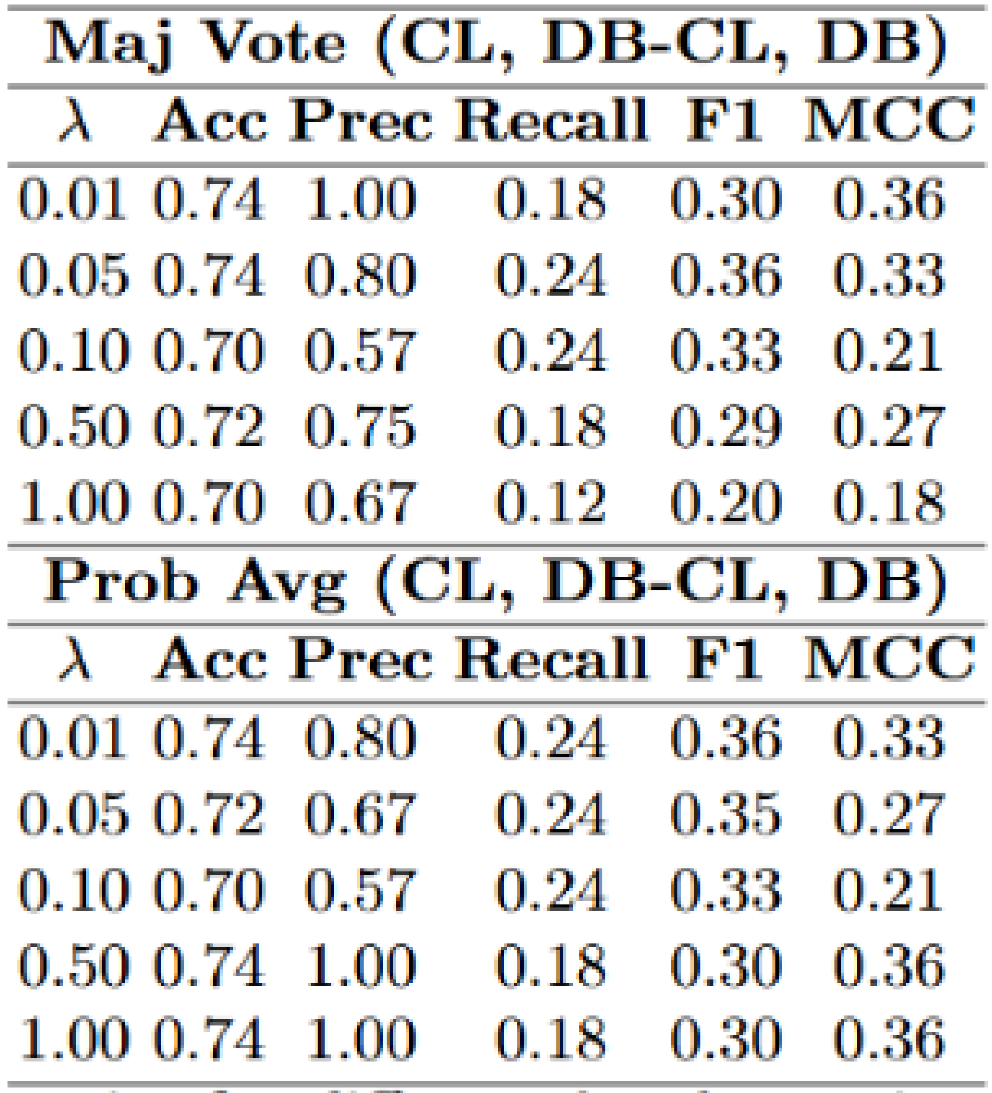
THANK YOU



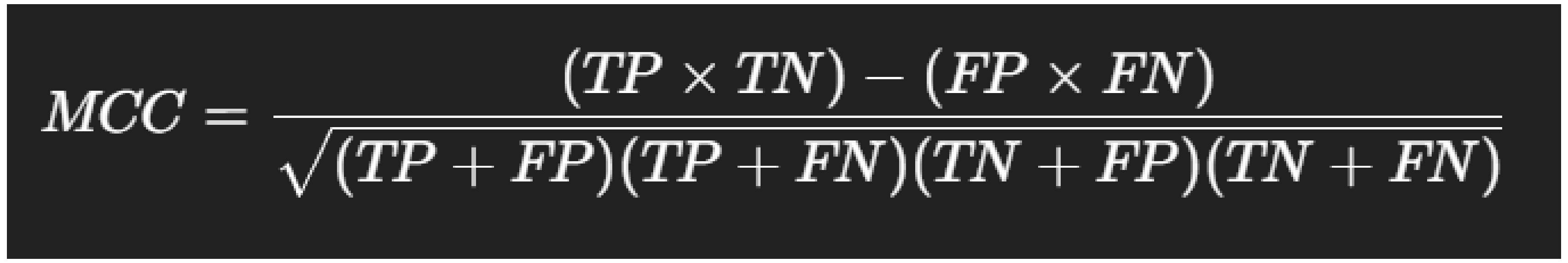
Model	Acc	Prec	Recall	F1	MCC
CL-CL	0.70	1.00	0.06	0.11	0.20
CL-DB	0.69	0.50	0.06	0.11	0.08
CL-CL,DB	0.70	0.60	0.18	0.27	0.20
DB-CL	0.70	0.67	0.12	0.20	0.18
DB-DB	0.70	1.00	0.06	0.11	0.20
DB-CL,DB	0.69	0.00	0.00	0.00	0.00
CL,DB-CL	0.70	0.60	0.18	0.27	0.20
CL,DB-DB	**0.74**	**1.00**	**0.18**	**0.30**	**0.36**
CL,DB-CL,DB	**0.74**	**1.00**	**0.18**	**0.30**	**0.36**



Model	Acc	Prec	Recall	F1	MCC
CL-CL	0.70	1.00	0.06	0.11	0.20
CL-DB	0.69	0.50	0.06	0.11	0.08
CL-CL,DB	0.70	0.67	0.12	0.20	0.18
DB-CL	0.70	0.67	0.12	0.20	0.18
DB-DB	0.70	1.00	0.06	0.11	0.20
DB-CL,DB	0.70	1.00	0.06	0.11	0.20
CL,DB-CL	0.70	0.67	0.12	0.20	0.18
CL,DB-DB	0.72	0.75	0.18	0.29	0.27
CL,DB-CL,DB	**0.74**	**0.80**	**0.24**	**0.36**	**0.33**



Maj Vote (CL, DB-CL, DB)					
λ	**Acc**	**Prec**	**Recall**	**F1**	**MCC**
0.01	0.74	1.00	0.18	0.30	0.36
0.05	0.74	0.80	0.24	0.36	0.33
0.10	0.70	0.57	0.24	0.33	0.21
0.50	0.72	0.75	0.18	0.29	0.27
1.00	0.70	0.67	0.12	0.20	0.18
Prob Avg (CL, DB-CL, DB)					
λ	**Acc**	**Prec**	**Recall**	**F1**	**MCC**
0.01	0.74	0.80	0.24	0.36	0.33
0.05	0.72	0.67	0.24	0.35	0.27
0.10	0.70	0.57	0.24	0.33	0.21
0.50	0.74	1.00	0.18	0.30	0.36
1.00	0.74	1.00	0.18	0.30	0.36



$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

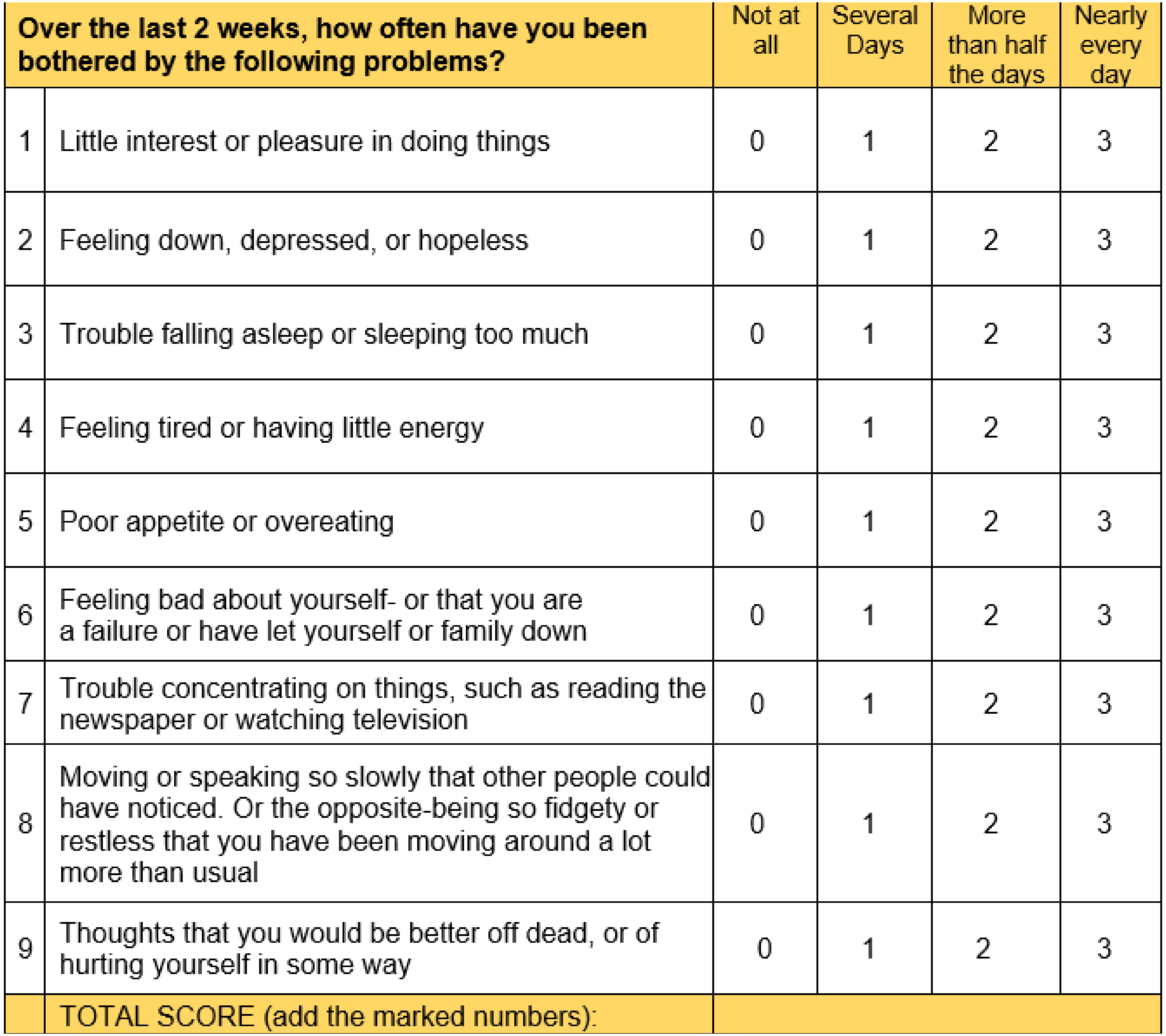
Prompt for GPT

1) You will be provided with the top feature importance scores and the transcript of an individual predicted as depressed. Using the provided ontology, generate explanations only for features that have interpretations available in the ontology.

2) Do not create explanations for features that are missing from the ontology.

3) Produce separate explanations for audio, video, and text modalities. Each explanation should be written in paragraph form, emphasizing why these features indicate depression in this specific individual, rather than simply listing feature names. For audio and video explanations, avoid using raw technical feature names. Instead, describe them in interpretable, intuitive terms, while retaining enough technical detail to clarify how the features relate to the prediction.

4) Ensure that the explanations are personalized to the individual based on their transcript and the top features, rather than generic descriptions of depression symptoms



Over the last 2 weeks, how often have you been bothered by the following problems?		Not at all	Several Days	More than half the days	Nearly every day
1	Little interest or pleasure in doing things	0	1	2	3
2	Feeling down, depressed, or hopeless	0	1	2	3
3	Trouble falling asleep or sleeping too much	0	1	2	3
4	Feeling tired or having little energy	0	1	2	3
5	Poor appetite or overeating	0	1	2	3
6	Feeling bad about yourself- or that you are a failure or have let yourself or family down	0	1	2	3
7	Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8	Moving or speaking so slowly that other people could have noticed. Or the opposite-being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9	Thoughts that you would be better off dead, or of hurting yourself in some way	0	1	2	3
	TOTAL SCORE (add the marked numbers):				