Ontology-Regularized Multimodal Neural Networks for Explainable Mental Health Assessment

Bhavya Jain[1], Sumit Dalal[2], and Bharat Bhargava[3]

[1] Indian Institute of Technology Bhilai
[2] Bennett University
[3] Purdue University

Abstract. Detecting depression in early stages is critical for effective intervention, but hard in traditional healthcare systems. Moreover, recent AI-based systems are black-boxes, which limits their usability and user trust in such sophisticated domains. While prior work has provided a semantic framework for text-based depression indicators, a comprehensive ontology covering audio and video modalities remains unexplored. This work proposes an ontology-regularized neural network (OR-NN) that incorporates structured multimodal depression features into model training via SHAP-informed constraints. The extended ontology models acoustic and visual descriptors, semantically mapped to psychological states, to provide better explanations. Neural Network is regularised by SHAP-informed constraints by introducing a custom regularisation loss that penalises deviations between model weight distributions and feature importances, and dropout bias that increases the probability of dropping out less important features during training. Experiments on the EDAIC dataset demonstrate that the proposed approach outperforms baseline neural models in both predictive accuracy and other metrics

Keywords: Multimodal Depression Detection · Explainable Artificial Intelligence (XAI) · Knowledge-Infused Neural Networks · Depression Feature Ontology · Neuro-Symbolic AI · Ontology-Guided Learning

1 Introduction

Depression is one of the most widespread mental health disorders, affecting more than 280 million individuals globally and ranking as a leading cause of disability according to the World Health Organization. Early detection and monitoring are essential for effective intervention, yet traditional diagnosis methods rely heavily on self-reporting and clinical interviews, which are often subjective, time-consuming, and limited in scalability. With the rapid growth of digital data streams (ranging from social media posts to voice recordings and video interactions), smart healthcare systems are increasingly leveraging machine learning (ML) and deep learning (DL) techniques for automated depression detection.

A growing body of research has explored depression recognition through textual data (e.g., social media posts, clinical notes), acoustic features (e.g., pitch, loudness), and visual behavior (e.g., facial expressions, gaze, head pose). Among these, speech and facial cues are particularly valuable, as they capture subtle changes in prosody, vocal energy, affective expressivity, and social engagement. These multimodal features provide strong predictive signals for detecting depression.

Existing deep learning models achieve promising performance but remain opaque black boxes, limiting their trust and thus usability in a high-stakes domain like healthcare. Clinicians require not only accurate predictions but also interpretable reasoning consistent with psychological knowledge. Ontologies provide a structured way to encode domain expertise by mapping low-level computational features to higher-level constructs such as social withdrawal and exhaustion. Prior work [2], structures text-based features, but a comprehensive multimodal ontology remains unexplored.

Additionally, while explainability techniques such as SHAP have become popular for post-hoc feature attribution, their integration into training remains limited. We argue that SHAP global importances provide a natural bridge between ontology knowledge and neural learning, guiding models to align with clinically meaningful features during training.

To address these gaps, we propose an Ontology-Regularized Neural Network (OR-NN) framework that fuses multimodal depression ontology with neural training. This work has the following contributions:

1. Extending DFO ontology to include audio and video descriptors linked to psychological constructs.
2. Designing methods for applying SHAP informed constraints on neural network and evaluating the proposed OR-NN against baseline neural network models, demonstrating improved accuracy and explainability
3. Generating and systematically comparing SHAP-based model explanations with and without ontology integration to assess the impact on interpretability.

2 Related Work

Research on automatic depression detection has advanced along two partly-overlapping axes: (1) development of robust *multimodal* feature extraction and fusion methods and (2) methods that improve *interpretability* and *knowledge alignment* by infusing domain knowledge or producing clinically-meaningful explanations. We summarize these strands below and situate the scope of the current study.

2.1 Multimodal Fusion Approaches

Recent work has explored multimodal depression detection to combine diverse signals for stronger prediction. For example, a Transformer-based feature en-

hancement network [5] integrated video, audio, and rPPG signals, capturing intra- and inter-modal relationships via stacked Transformers and graph fusion networks, achieving high performance on AVEC2013 and AVEC2014. Another approach, MFM-Att [6], leveraged audio-visual-text data with multi-level attention and LSTM/Bi-LSTM networks, learning intra- and inter-modal correlations and outperforming prior baselines on DAIC-WOZ. These studies demonstrate the effectiveness of multimodal feature extraction and fusion for depression detection. Similarly, text-driven LLM models have been combined with different type of features to predict depression severity. A BiLSTM-based tri-modal fusion model incorporating audio (MFCCs), visual (AUs), and textual features with GPT-4 embeddings achieved strong performance on the DAIC-WOZ dataset, surpassing several state-of-the-art baselines [11] . Some studies[16]. [13] extended LLMs by integrating acoustic landmarks into textual transcripts, showing that speech-aware LLM framework improves multimodal depression detection beyond text-only systems .

2.2 Explainability Oriented Systems

While multimodal methods improve performance, interpretability has also become a focus. The EMDRC task explicitly framed multimodal depression recognition as both prediction and explanation, generating PHQ-8–grounded symptom summaries alongside severity scores [18]. Similarly, the PSAT framework [3] embedded clinical practice guidelines (CPG) into Transformer attention, producing clinician-interpretable outputs grounded in PHQ-9 and SNOMED-CT. Additionally, different recent studies have integrated SHAP as a post-hoc explanation method [15] [14]. This reflects a shift toward clinically meaningful explanations rather than purely statistical predictions.

2.3 SHAP guided models

Several studies have integrated SHAP beyond post-hoc explanation. SHAP-guided regularization has been proposed to impose entropy-based penalties, encouraging sparse and stable feature attributions during training [12]. Other work such as LassoNet [8] focuses on feature selection, enforcing hierarchical constraints so that only relevant features contribute to hidden units. More recently, SHAP-informed optimization methods [7]use SHAP values to guide gradient updates and learning rates, with efficient variants like C-SHAP and FastSHAP reducing computational cost.

2.4 Ontology Based and Neuro Symbolic Approaches

Parallel to these methods, ontology-guided frameworks have emerged to align models with clinical knowledge. The DepressionFeature Ontology (DFO) was extended to social media concepts and validated on text-based depression datasets [2]. Building on DFO, the KiNN model [1] infused both domain knowledge and

commonsense reasoning, offering user-level explainability. Neuro-symbolic approaches such as TAM-SenticNet [4] combined neural sentiment analysis with symbolic reasoning from SenticNet for structured interpretation of emotional signals.

Together, these advances highlight two converging trends: performance gains from multimodal fusion and explainability through knowledge alignment. However, most prior systems remain text-focused, or they integrate symbolic knowledge in limited or post-hoc ways. In contrast, our study introduces a multimodal OR-NN framework that embeds SHAP-derived global feature importances into the learning objective, driving prediction and explicitly aligning model explanations with an ontology of depression symptoms.

3 Designing multimodal ontology

The methodology followed in this work is organised into three main phases: Feature identification,extraction and preprocessing , Ontology Construction via Neural network and SHAP, Ontology Knowledge infusion in Neural Network.

3.1 Feature Identification

We use the EDAIC dataset, a widely used benchmark for multimodal depression detection. The dataset contains clinical interviews with audio, video, and transcribed text, enabling multimodal feature extraction. We employ standardised, interpretable descriptors for audio and video, such as pitch, jitter, loudness, gaze, and facial action units (AUs). Unlike features extracted via Wav2Vec or MFCCs as in [17], these descriptors have direct semantic mappings to psychological constructs, making them inherently suitable for ontology-based modelling. We chose statistical aggregation over methods like Fisher vector encoding as in [17] or temporal stacking as in [9]) to ensure clean, ontology-aligned representations, avoiding issues related to frame-wise concatenation and variable-length sequences.

Audio Features: For the acoustic analysis, we extracted a comprehensive set of frequency, energy, spectral, temporal, and cepstral features, which have previously been associated with depression-related speech patterns.

- **Frequency-related Parameters** Pitch, or the fundamental frequency of vocal fold vibration, was captured through several statistics including the mean, normalized standard deviation, the 20th, 50th, and 80th percentiles, the range between the 20th and 80th percentiles, and the slopes of rising and falling signal parts. Formant-related measures included the mean and normalized deviation of the first three formant frequencies and their corresponding bandwidths.
- **Energy and Amplitude-related Parameters** Shimmer was extracted to capture short-term variations in sound wave amplitude Loudness was represented by the mean and normalized deviation, percentiles, the range between

percentiles, and the slopes of rising and falling parts. The Harmonics-to-Noise Ratio (HNR) quantified harmonic content relative to noise.

- **Spectral Parameters** Spectral balance was captured using the alpha ratio, and the Hammarberg Index, which compares energy between peaks. Spectral slopes were computed for the 0–500 Hz and 500–1500 Hz ranges. Other measures included formant-relative energies, harmonic differences, and their normalized deviations.
- **Temporal Features** Temporal aspects included the rate of loudness peaks per second, the mean and standard deviation of continuously voiced regions, the mean and standard deviation of continuously unvoiced regions, and the pseudo-syllable rate, a rough measure of speaking tempo.
- **Extended Set of Features** Additional extended features included measures over unvoiced regions such as the alpha ratio, Hammarberg Index, and spectral slopes. Mel-frequency cepstral coefficients (first four coefficients, with mean and coefficient of variation computed for both the full signal and voiced-only regions) were also extracted. Spectral flux was measured through the mean and coefficient of variation across voiced and unvoiced regions, along with bandwidth measures of the second and third formants. Finally, the equivalent sound level was computed as the average sound pressure level across the entire recording.

Video Features: Visual descriptors were extracted from frame-level recordings and aggregated into statistical functionals. These cover:

- **Pose-related Features** Head translation was measured along the three spatial axes. For each axis(x,y,z), the average position, variability, range of motion, and velocity was calculated . Head rotation was similarly represented along three axes: pitch (nodding), yaw (turning the head side to side), and roll (tilting).
- **Gaze-related Features** Eye gaze features included horizontal (x-axis) and vertical (y-axis) angles, for which the mean, variability, and range were computed. In addition, the fraction of frames with downward gaze was quantified.
- **Facial Action Unit (AU) Intensity Features** We extracted intensity-based features for 17 action units (AUs), which represent muscle activations in the face. For each AU (all AUs described in [10]), we computed five functionals: mean intensity, standard deviation, maximum value, fraction of active frames, and average duration of continuous activation. These measures capture both the strength and stability of facial expressions.
- **Facial Action Unit (AU) Occurrence Features** Occurrence-based features were also extracted for each of the 17 AUs. These included the proportion of frames where the AU occurred, the average duration of activation and the number of transitions representing how often the AU switched on and off. Such occurrence statistics reflect the frequency and persistence of facial expressions.

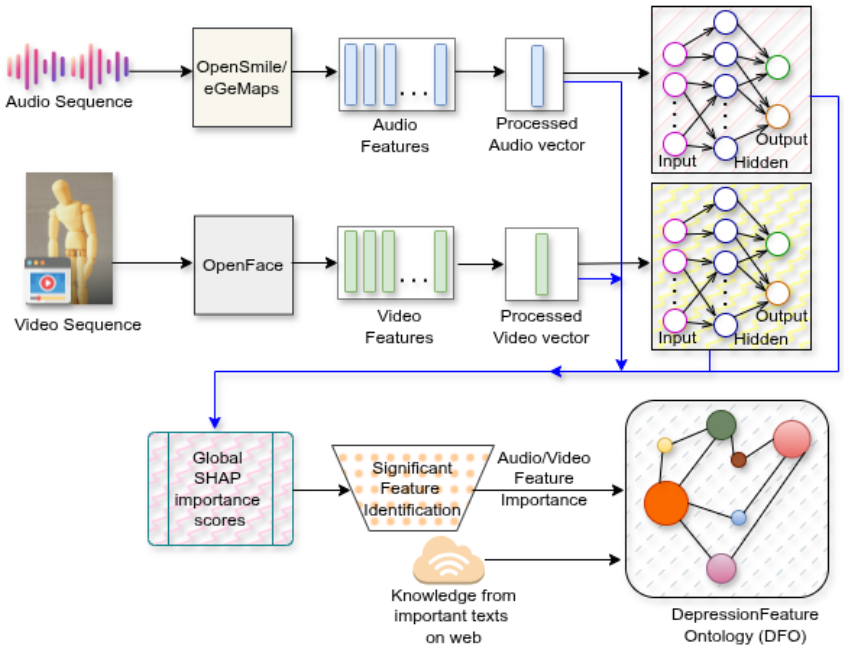

Fig. 1. Depression Feature Ontology (DFO) Extension Process. Multimodal features extracted from audio (eGeMAPS) and video (OpenFace) are modeled through modality-specific networks. SHAP gives significant features with importances which are integrated into the DFO with relation to semantic mappings from the web.

3.2 Ontology Construction via Neural network and SHAP

The extracted audio and video descriptors, along with their statistical functionals, form the low-level computational features that serve as input to our SHAP-based global importance estimation. To guide ontology construction, we trained a simple feedforward neural network on the extracted audio (81) and video (170) features. The network consisted of two fully connected layers(128 and 64 units) with ReLU activations, dropout (default=0.3), and a sigmoid output for binary depression prediction. After training for 300 epochs, we applied SHAP (SHapley Additive exPlanations) to compute global feature importances. Input features were standardized, a random seed of 42 ensured reproducibility, and 100 background samples were used for SHAP output. Refer Figure 1

For each feature, the mean absolute SHAP value across test samples was calculated, producing a ranked list of features by relevance. We also derived semantic mappings between low-level multimodal descriptors and psychological constructs using SHAP-based global importances(refer Figure 3). This procedure ensures that the ontology captures both low-level descriptors and clinically meaningful patterns, forming the basis for ontology-regularized model training.

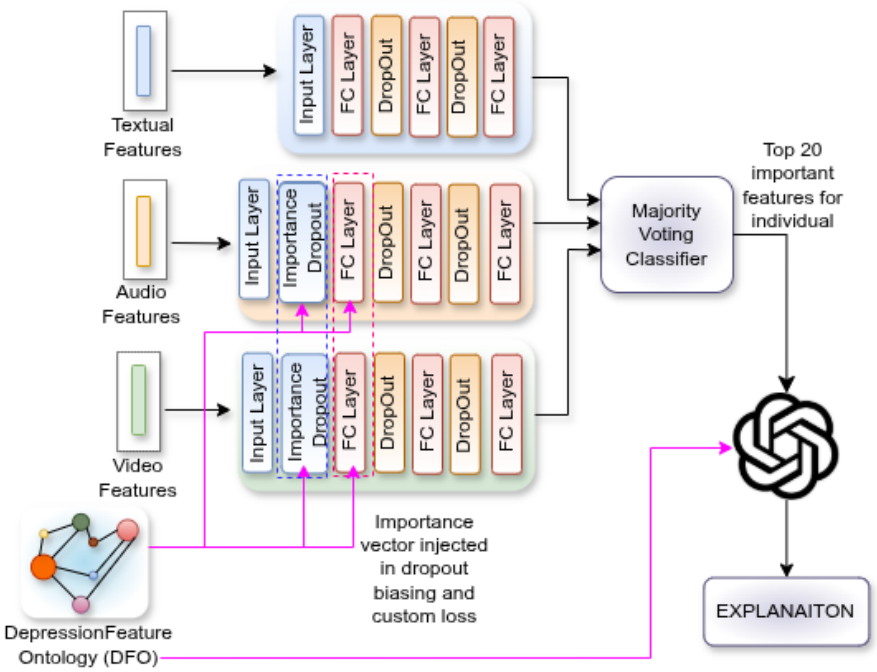

Fig. 2. Textual, audio, and video features are independently processed through modality-specific neural networks. The Depression Feature Ontology (DFO) provides domain knowledge through an importance vector, which is injected into the dropout mechanism and custom loss. Outputs from all modalities are fused using a majority voting classifier to generate final predictions. The framework also identifies the most influential features for each individual and leverages the ontology to provide interpretable explanations of model decisions.

4 Ontology Knowledge Infusion in Neural Network

This work implements a **hybrid neural network** with two key mechanisms, *custom loss* and *importance-biased dropout*, to leverage feature importance for better interpretability and generalization (See Figure 2).

Custom Loss Function. The custom loss incorporates SHAP-derived feature importances into the training objective. Here, $\mathbf{w}_1 \in R^{d\times h}$ is the first-layer weights, and $\mathbf{s} \in R^d$ is the normalized importance vector. The regularization term penalizes deviations between the mean weight per input feature and its importance:

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

The total loss is calculated as:

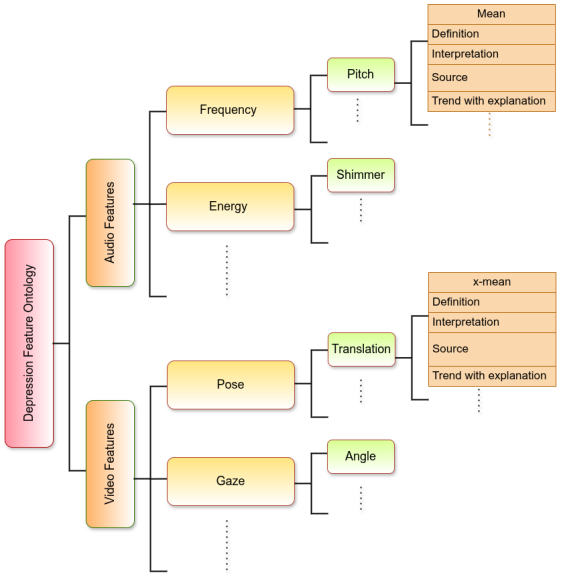

Fig. 3. Hierarchy level diagram for O?ntology

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \lambda \mathcal{L}_{\text{reg}}, \tag{2}$$

where $\mathcal{L}_{\text{BCE}}$ is the standard binary cross-entropy, and λ is a hyperparameter controlling the regularization strength. Features with higher SHAP importance are encouraged to maintain stronger connections in the first layer, while less important features are suppressed. Hence, the ontology only controls how the model starts "seeing" the data, not how it decides later. This step is followed to avoid over-constraining the model and losing its generalization.

Importance-Biased Dropout. To further leverage feature relevance, A custom dropout mechanism is introduce that scales the dropout probability inversely with the SHAP-derived importance scores. Specifically, for each feature i, let s_i denote its SHAP importance, the dropout probability is defined as:

$$p_i^{\text{drop}} = p_{\text{base}} \left(1 - \frac{s_i}{\max(s)}\right), \quad i = 1, \ldots, d \tag{3}$$

During training, each input feature x_i is independently zeroed out with probability p_i^{drop}. A binary mask of the same dimensionality as the input is generated prior to the first linear layer, ensuring that dropout is applied directly to input features rather than hidden activations. This allows the network to focus more on important features and improves robustness to noisy or less relevant inputs.

Reproducibility: All hyperparameters (learning rate, optimiser, number of epochs, layer sizes, dropout rates) were kept the same as in the baseline experiments described in Section 3. Random seeds for PyTorch, NumPy, and Python `random` module were fixed at 42. Input features were standardised and missing values imputed with median values. SHAP importance vectors were normalised for use in both custom loss and dropout biasing. Ontology regularisation adds conceptual constraints, not extra layers, so the runtime and memory overhead is negligible or even slightly reduced. Our method uses fewer parameters than the baseline (0.16M vs 0.28M) and achieves faster training per epoch (2.06 s vs 3.52 s). It also demonstrates lower inference latency, meaning predictions can be made almost instantly (0.02 ms vs 0.05 ms). These results indicate that ontology regularisation provides meaningful conceptual constraints without increasing computational overhead, and that it scales efficiently for both training and real-time inference.

5 Results and Discussion

5.1 Dataset and Evaluation Metrics

Due to the high class imbalance in the dataset (approximately 75% non-depressed participants), accuracy alone can be misleading. Therefore, we also report precision, recall, F1-score, Matthews Correlation Coefficient (MCC) to provide a balanced evaluation of model performance.

5.2 Performance Comparison Setup

To isolate the effect of knowledge infusion we have (a) included unimodal (audio-only, video-only) baselines with the multimodal baseline, (b) trained the same neural architecture without ontology regularization. These choices directly measure the contribution of each modality and the ontology regularizer on the same dataset and setup. Reproducing a large number of external, multimodal baselines proved impractical due to differences in dataset preprocessing and modality coverage; instead, we selected internal baselines that control for architecture and dataset, thereby directly measuring the impact of knowledge infusion. We majorly want to answer the following questions.

- Does ontology help over plain baselines?
- Which technique (DB, CL, or both) works better?
- How do unimodal vs multimodal setups compare?

5.3 Results

Table 1 presents the performance of different modality combinations under two experimental settings: without ontology and with ontology regularization. Text models are highly conservative, yielding strong precision but almost negligible

Table 1. Performance comparison of modalities with and without ontology infusion

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**

recall, which reflects a tendency to predict only the majority class. Audio and video features perform moderately, with video outperforming audio in F1-score and MCC. The ontology-guided models achieve higher F1-score and MCC, indicating a better balance between precision and recall. The improvements are especially notable for audio and video modalities, reflecting the richer feature sets and structured mappings provided by the ontology. SHAP-guided custom loss reduces false negatives, enabling the model to better identify depressed individuals, while importance-biased dropout mitigates overfitting on less informative features.

Table 2. Video modality with and without ontology

Model	Acc	Prec	Recall	F1	MCC
Simple w/o ontology	0.63	0.29	0.12	0.17	-0.02
Dropout biasing	**0.72**	**0.58**	**0.41**	**0.48**	**0.31**
Custom loss	0.70	0.57	0.24	0.33	0.21
DB + CL	0.67	0.45	0.29	0.36	0.15

Table 3. Audio modality with and without ontology

Model	Acc	Prec	Recall	F1	MCC
Simple w/o ontology	**0.67**	0.43	0.18	0.25	0.09
Dropout biasing	0.61	0.39	0.41	0.40	0.11
Custom loss	0.63	0.38	0.29	0.33	0.08
DB + CL	0.65	**0.44**	**0.47**	**0.46**	**0.20**

In unimodal settings, video networks (Table 2) collapse without ontology support, with poor recall, while dropout biasing provides the most robust improvements. By contrast, audio networks(Table 3) achieve their best performance only when both custom loss and dropout biasing are applied together, reflecting a modality-dependent effect: video is ontology-dependent, while audio is ontology-enhanced.

Table 4. Majority voting for multimodal fusion.
Note: Model labels follow the format *Audio-Video*. CL = Custom Loss, DB = Dropout Biasing. For example, CL-DB means the audio network uses Custom Loss while the video network uses Dropout Biasing. If both appear together (e.g., CL,DB-CL), it means both techniques were applied to that modality. Loss uses $\lambda = 1$

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**

Table 5. Probability average for multimodal fusion.
Note: Model labels follow the format *Audio-Video*. CL = Custom Loss, DB = Dropout Biasing. Loss uses $\lambda = 1$

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**

A similar contrast is observed in multimodal fusion (see Table 1). Without the use of an ontology, the combination of audio, video, and text demonstrates superficially high precision but significantly low recall. This indicates that the model predominantly favours the majority class. While both probability averaging and majority voting (refer to Tables 5 and 4) yield similar outcomes, probability averaging tends to be slightly more robust due to its consideration of prediction confidence.

Table 6 shows that both fusion strategies achieve their highest recall at $\lambda = 0.50$, indicating that moderate regularisation enhances sensitivity to positive cases. The consistent performance across nearby λ values further confirms the stability and robustness of our method.

5.4 Discussion

To summarise, ontology-based multimodal configurations substantially improve recall and deliver more balanced performance. The most effective setup, where

Table 6. Performance metrics for different λ values using two ensemble strategies: Majority Voting and Probability Averaging for CL, DB-CL, DB.

Maj Vote (CL, DB-CL, DB)					
λ	**Acc**	**Prec**	**Recall**	**F1**	**MCC**
0.01	0.72	1.00	0.12	0.21	0.29
0.05	**0.74**	**0.80**	0.24	0.36	**0.33**
0.10	0.69	0.50	0.18	0.26	0.14
0.50	0.72	0.63	**0.29**	**0.40**	0.28
1.00	0.72	0.67	0.24	0.35	0.27
Prob Avg (CL, DB-CL, DB)					
λ	**Acc**	**Prec**	**Recall**	**F1**	**MCC**
0.01	0.72	0.75	0.18	0.29	0.27
0.05	0.70	0.60	0.18	0.27	0.20
0.10	0.69	0.50	0.24	0.32	0.17
0.50	0.72	0.63	**0.29**	**0.40**	0.28
1.00	**0.74**	**0.80**	0.24	0.36	**0.33**

both audio and video networks are regularised with custom loss and dropout biasing, achieves the highest MCC and F1, with recall rising nearly fourfold compared to ontology-free fusion. This demonstrates that ontology not only stabilises noisy modalities such as video but also counteracts imbalance-driven bias in multimodal integration, enabling the detection of minority-class (depressed) cases that would otherwise be missed.

Fairness: Fairness is a central concern for multimodal mental health systems. Models trained on limited or homogeneous samples risk producing biased outcomes when deployed in diverse real-world populations. Future extensions of this work should prioritize evaluating system performance across different demographic groups and incorporating balanced datasets to ensure equitable access to accurate depression detection technologies.

Bias Mitigation: Our ontology-regularized loss contributes to bias mitigation by guiding the model to attend to clinically meaningful concepts rather than spurious statistical correlations in the training data. By grounding feature importance in established depression ontologies, we reduce the likelihood of overfitting to noise or incidental cues in audio or visual features.

Privacy: Given the sensitivity of mental health data, privacy is a fundamental requirement. The dataset used in this study was provided in anonymised form, with no personally identifiable information about participants. This safeguards confidentiality while enabling research.

6 Ontology vs Non-Ontology Interpretations

Figure 2 shows the procedure for generating explanations. When the model is trained we extract top 20 features for audio and video modalities using SHAP for an individual. Afterwards, we feed them with the ontology and prompt the LLM to generate explanation useful for the clinicians. Below is the prompt used

: *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. Do not create explanations for features that are missing from the ontology. Produce separate explanations for audio, video, and text modalities. Each explanation should be written in a paragraph only, emphasising why these features indicate depression in this specific individual, rather than simply listing feature names. 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. Ensure that the explanations are personalized to the individual based on their transcript and the top features, rather than generic descriptions of depression symptoms.*
The bold sections represent the betterment of explanations in contrast with the explanations generated without using ontology. The explanations correspond to different symptoms of depression and thus could help the clinician get a better idea of the patient's condition.

7 Limitations and Future Work

We also experimented with assessing the importance of different words in the transcripts using SHAP. However, we observed that the results contained much more noise than meaningful signals, making it unsuitable to directly map these words to the existing Depression Feature Ontology (DFO). Therefore, in future work, we plan to enrich the DFO with words related to depression derived from the EDAIC dataset. Additionally, it is important to note that the transcripts do not include the interviewer's questions, which makes them somewhat incomplete, as participants' answers may not fully reflect the context of the questions asked. For future work, we aim to temporally connect audio and video features with text, enabling cross-modal reasoning. This integration is expected to generate richer and more accurate explanations, while also improving the overall performance of depression detection models. Future work should also explore larger, balanced datasets and advanced augmentation techniques to improve recall without compromising overall performance.

Ontology-based Audio explanation

Their voice is consistently monotone and low in pitch, with minimal rising or falling contours, suggesting reduced emotional expression and a flat mood, **feeling down and loss of interest**. The narrow formant bandwidths and reduced spectral variation indicate slower and less dynamic articulation, reflecting **psychomotor slowing and low energy** . Additionally, the voice exhibits higher instability in pitch and amplitude, producing a slightly breathy or irregular tone, which mirrors the **muscle tension and fatigue** commonly seen in individuals with depression. Longer pauses, fewer energy peaks, and diminished vocal intensity contribute to speech that sounds hesitant and weak, reinforcing the impression of **low motivation and reduced engagement**.

Non-Ontology Based Audio Explanation

Their vocal pitch tends to vary less than typical, with a narrower overall range, and moments of pitch decline occur more gradually, suggesting monotone or subdued intonation. The rhythm and timing of voiced segments are irregular, with shorter and less consistent syllable durations, reflecting reduced speech energy or slowed verbal production. The loudness of the voice is generally lower and less dynamic, with muted peaks and slower changes, which can indicate reduced emotional expressivity. At the spectral level, variations in the resonance of different vocal frequencies—captured by shifts in formants and spectral energy—are less pronounced, pointing to less dynamic articulation and potential tension in the vocal tract. Micro-level voice qualities, such as slight pitch jitter and amplitude irregularities, further suggest subtle instability or **fatigue** in vocal control.

Ontology-based Video Explanation

This individual shows blunted facial expressions and minimal head and eye movements, reflecting **low emotional engagement and psychomotor slowing**. The reduced smiling and limited cheek and lip activity suggest a **loss of interest and pleasure and a depressed mood**. Downward gaze and limited head shifts indicate **social withdrawal and low attentional engagement**, while subtle reductions in eye and mouth movements point to **fatigue and low energy**.

Non-Ontology Based Video Explanation

This individual shows clear signs of emotional withdrawal and reduced expressivity. Movements around the mouth and cheeks—typically associated with smiling or engagement—are minimal, suggesting a blunted affect. In contrast, subtle tension around the lips and chin hints at suppressed emotion or internal strain. The eyes and eyebrows show limited movement, reflecting low energy and diminished responsiveness. Head and gaze patterns are also irregular, indicating **difficulty maintaining focus** or connection during interaction.

References

1. Dalal, S., Jain, S., Dave, M.: Deep knowledge-infusion for explainable depression detection (2024), https://arxiv.org/abs/2409.02122
2. Dalal, S., Jain, S., Dave, M.: Depressionfeature: Underlying ontology for user-specific depression analysis. J. Supercomput. **81**(1) (Oct 2024). https://doi.org/10.1007/s11227-024-06585-w, https://doi.org/10.1007/s11227-024-06585-w
3. Dalal, S., Tilwani, D., Roy, K., Gaur, M., Jain, S., Shalin, V., Sheth, A.: A cross attention approach to diagnostic explainability using clinical practice guidelines for depression (2024), https://arxiv.org/abs/2311.13852
4. Dou, R., Kang, X.: Tam-senticnet: A neuro-symbolic ai approach for early depression detection via social media analysis. Computers and Electrical Engineering **114**, 109071 (2024). https://doi.org/https://doi.org/10.1016/j.compeleceng.2023.109071, https://www.sciencedirect.com/science/article/pii/S0045790623004950
5. Fan, H., Zhang, X., Xu, Y., Fang, J., Zhang, S., Zhao, X., Yu, J.: Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. Information Fusion **104**, 102161 (2024). https://doi.org/https://doi.org/10.1016/j.inffus.2023.102161, https://www.sciencedirect.com/science/article/pii/S1566253523004773
6. Fang, M., Peng, S., Liang, Y., Hung, C.C., Liu, S.: A multimodal fusion model with multi-level attention mechanism for depression detection. Biomedical Signal Processing and Control **82**, 104561 (2023). https://doi.org/https://doi.

org/10.1016/j.bspc.2022.104561, https://www.sciencedirect.com/science/article/pii/S1746809422010151
7. Graham, J., Sheng, V.S.: Shap informed neural network. Mathematics (2025), https://api.semanticscholar.org/CorpusID:276806868
8. Lemhadri, I., Ruan, F., Abraham, L., Tibshirani, R.: Lassonet: A neural network with feature sparsity (2021), https://arxiv.org/abs/1907.12207
9. Li, Y., Yang, X., Zhao, M., Wang, Z., Yao, Y., Qian, W., Qi, S., El Kafhali, S.: Fpt-former: A flexible parallel transformer of recognizing depression by using audiovisual expert-knowledge-based multimodal measures. Int. J. Intell. Syst. **2024** (Jan 2024). https://doi.org/10.1155/2024/1564574, https://doi.org/10.1155/2024/1564574
10. Mahayossanunt, Y., Nupairoj, N., Hemrungrojn, S., Vateekul, P.: Explainable depression detection based on facial expression using lstm on attentional intermediate feature fusion with label smoothing. Sensors **23**(23) (2023). https://doi.org/10.3390/s23239402, https://www.mdpi.com/1424-8220/23/23/9402
11. Patapati, S.V.: Integrating large language models into a tri-modal architecture for automated depression classification on the daic-woz (2024), https://arxiv.org/abs/2407.19340
12. Saadallah, A.: Shap-guided regularization in machine learning models (2025), https://arxiv.org/abs/2507.23665
13. Sadeghi, M., Richer, R., Egger, B., Schindler-Gmelch, L., Rupp, L., Rahimi, F., Berking, M., Eskofier, B.: Harnessing multimodal approaches for depression detection using large language models and facial expressions. npj Mental Health Research **3** (12 2024). https://doi.org/10.1038/s44184-024-00112-8
14. Shen, A., Sun, J., Chen, X., Gao, X.: A data-centric and interpretable eeg framework for depression severity grading using shap-based insights. Journal of Neuroengineering and Rehabilitation **22**(1), 116 (2025). https://doi.org/10.1186/s12984-025-01645-5, https://doi.org/10.1186/s12984-025-01645-5
15. de Sousa Balbino, M., Santana, R., Teodoro, M.L.M., Song, M., Zárate, L.E., Nobre, C.: Predicting depression in children and adolescents using the shap approach. In: International Conference on Health Informatics (2022), https://api.semanticscholar.org/CorpusID:247125367
16. Zhang, X., Liu, H., Xu, K., Zhang, Q., Liu, D., Ahmed, B., Epps, J.: When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection (2024), https://arxiv.org/abs/2402.13276
17. Zhang, Z., Lin, W., Liu, M., Mahmoud, M.: Multimodal deep learning framework for mental disorder recognition. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). pp. 344–350 (2020). https://doi.org/10.1109/FG47880.2020.00033
18. Zheng, W., Xie, Q., Wang, Z., Yu, J., Xia, R.: Towards explainable multimodal depression recognition for clinical interviews (2025), https://arxiv.org/abs/2501.16106