Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 22 May 2026
Revised 26 June 2026
Accepted 22 July 2026

Natural Resources for Human Health 2026, Vol. 6 (8s): 59–76

KEYWORDS:
Mental health detection, multimodal machine learning, facial expression recognition, BERT, ResNet50, hybrid ensemble, stacking classifier, SMOTE, affective computing, pilot study, proof of concept.

Hybrid Ensemble Fusion for Emotion-Linked Mental Health Screening: A Proof-of-Concept Pilot Study using Facial Expression and Textual Data

Priyanka P. Boraste[1], Monika S. Deshmukh[2]

[1]Research Scholar,School of Computer Science and Engineering, Sandip University, Nashik 422213, Maharashtra, India
Email: boraste.priyanka@kbtcoe.org ORCID: 0009-0008-2565-1455
[2]Research Guide, School of Computer Science and Engineering, Sandip University, Nashik 422213, Maharashtra, India
Email: monika.deshmukh@sandipuniversity.edu.in ORCID: 0009-0009-6185-8768

ABSTRACT: Mental health disorders remain a major global health burden, with the World Health Organization estimating that roughly 970 million people, or one in eight, were living with a mental health condition in 2019, a figure worsened by the 25% surge in anxiety and depression during the COVID-19 pandemic. Traditional diagnostic approaches rely heavily on subjective clinical interviews, which can delay treatment and increase the risk of misdiagnosis. This paper reports a proof-of-concept pilot study investigating whether a hybrid ensemble fusion of facial-expression and textual features is architecturally viable as an early step toward an objective mental-health screening aid. Facial affect is captured using a ResNet50 convolutional network trained on a nine-class facial-expression corpus, while short textual descriptors, heuristically derived from each image's own emotion label as a proxy for candidate mental-state indicators, are encoded using BERT embeddings. The two modality-specific feature sets are combined through a stacking ensemble with Multinomial Naive Bayes and XGBoost as base learners and an XGBoost meta-learner. The unimodal facial branch achieves 88.6% test accuracy and an F1-score of 88.75%, while the fused stacking ensemble reaches 75% accuracy on a small held-out pilot set, with class-wise precision and recall values that suggest balanced performance across most categories at this scale. Synthetic Minority Oversampling is used to counter class imbalance in under-represented affective classes such as contempt and sleepy/fatigue. These results demonstrate the technical feasibility of the proposed fusion architecture but should be read as an early-stage proof of concept rather than a validated diagnostic result: the textual descriptors are derived from the same labels as the visual branch rather than from an independent modality, and the fused evaluation set is too small for statistically robust claims. We discuss these limitations explicitly and outline the clinically grounded data, independent-modality design, and larger-scale validation required before this pipeline could support real screening use.

I. INTRODUCTION

Mental health (MH) conditions are pervasive in modern society. The World Health Organization estimated that around 970 million people, nearly one in eight globally, were living with a mental health condition in 2019, and the COVID-19 pandemic drove a further 25% increase in anxiety and major depressive disorder during 2020. Regional surveys reinforce this picture: 42.9% of Australians aged 16-85 reported experiencing a mental disorder at some point in their lives as of 2022, and 22.8% of

adults in the United States were estimated to be experiencing mental illness in 2021. Mental health disorders are projected to cost the global economy approximately USD 16 trillion by 2030, motivating the WHO Comprehensive Mental Health Action Plan 2013-2030, which targets a 50% increase in service coverage and integration of mental health care into primary health systems in 80% of countries by 2030 [1].

This growing burden is compounded by a persistent gap between the prevalence of mental illness and the availability of trained clinicians capable of timely diagnosis. In many low- and middle-income regions, the ratio of psychiatrists to population remains far below WHO-recommended levels, and even in well-resourced health systems, waiting times for a first psychiatric assessment can extend to several months. Traditional diagnostic workflows rely on structured or semi-structured clinical interviews, self-report questionnaires such as the PHQ-9 or GAD-7, and the subjective judgement of the treating clinician. While these instruments are well validated, they are inherently retrospective and self-reported, and their accuracy depends on a patient's willingness and ability to articulate internal states honestly, a challenge that is particularly acute for adolescents, individuals with limited health literacy, or those experiencing stigma-related reluctance to disclose symptoms. Automated, passively collected signals, such as facial affect, linguistic patterns, or physiological markers, have therefore attracted sustained research interest as complementary, non-invasive indicators that could support, rather than replace, clinical judgement.

Individuals living with mental illness frequently report reduced capacity for daily activities, diminished enjoyment of social interaction, and disrupted mood regulation. Diagnosis today relies primarily on subjective clinical interviews and self-report questionnaires, which are time-consuming, inconsistent across clinicians, and prone to delay. Machine learning offers an opportunity to develop objective, scalable, and timely screening tools that complement clinical judgement rather than replace it. This paper reports a proof-of-concept pilot study that investigates the technical feasibility of fusing facial-expression imagery and short textual descriptors through a hybrid ensemble, as an early architectural step toward a fully multimodal (text, visual, physiological) mental-health screening system. We deliberately frame this as a pilot study rather than a validated screening tool: as detailed in Section VI, the textual descriptors used here are heuristically derived from the same emotion labels as the visual branch, and the fused evaluation set is small, so the reported results demonstrate architectural feasibility rather than clinical validity.

The contributions of this pilot study are summarized as follows:

- We design and implement a hybrid ensemble pipeline that extracts visual affect features via a fine-tuned ResNet50 backbone and textual features via BERT embeddings, and fuses them through a two-level stacking ensemble with Multinomial Naive Bayes and XGBoost base learners and an XGBoost meta-learner.
- We provide a formal, reproducible description of the preprocessing, feature-extraction, class-balancing, and fusion procedure (Section III), including the emotion-to-candidate-mental-state mapping heuristic used to bridge a facial-expression corpus toward a mental-health-oriented framing.
- We report unimodal and fused evaluation results (Section V), including per-class precision, recall, and F1-score, and analyze where the fused ensemble succeeds and where its evaluation remains statistically limited.
- We conduct an explicit, structured self-assessment of the pipeline's limitations and threats to validity (Sections VI and VII), including label circularity between modalities, the absence of clinically grounded ground truth, and the small size of the current fused evaluation set, so that the reported results are not over-interpreted as validated clinical performance.
- We outline a concrete, prioritized roadmap (Section IX) for converting this architectural pilot into a clinically defensible system, including independent-modality data collection, standardized-instrument labeling, larger-scale cross-validated evaluation, and a third physiological modality.

The remainder of this paper is organized as follows. Section II reviews related work across six thematic areas. Section III details the proposed methodology, including dataset description, preprocessing, feature extraction, and the fusion architecture, with formal notation and an algorithmic summary. Section IV describes the experimental setup and evaluation metrics. Section V reports results for both the unimodal visual branch and the fused ensemble, and discusses these results in relation to the literature. Section VI enumerates the pilot's limitations, Section VII discusses threats to validity, Section VIII provides a reproducibility statement, and Section IX concludes with a prioritized roadmap for future work.

II. LITERATURE REVIEW

A. Foundational Deep Learning Architectures

Several general-purpose deep learning building blocks underpin the modality-specific branches used in this work. Devlin et al. [2] introduced BERT, a bidirectional transformer pretraining objective that remains the dominant backbone for text representation learning, built on the self-attention mechanism of Vaswani et al. [4]. Liu et al. [5] later showed with RoBERTa that BERT was significantly undertrained and that a more careful pretraining recipe improves downstream performance without changing the underlying architecture. On the visual side, Simonyan and Zisserman [6] demonstrated that stacking small convolutional filters into very deep networks (VGG) substantially improves image-recognition accuracy, a direction later extended by residual connections. Goodfellow et al. [7] introduced Generative Adversarial Networks, which are frequently used in the affective-computing literature for synthetic data augmentation under class imbalance. Hochreiter and Schmidhuber [8] introduced Long Short-Term Memory networks, which remain a common choice for modeling sequential physiological and audio signals. Training-stability techniques such as the Adam optimizer [9] and batch normalization [10] are standard components of the training pipelines used across the surveyed literature and in this work.

These foundational architectures were originally developed and validated on large, general-purpose benchmarks such as ImageNet for vision and large web-text corpora for language modeling, rather than on clinical or affective-computing data. Their widespread reuse as transfer-learning backbones in the mental-health-detection literature therefore rests on an implicit assumption: that visual and linguistic representations learned from generic data transfer usefully to the comparatively narrow, often small, and domain-specific distributions encountered in affective computing. This assumption is reasonable for low-level visual features (edges, textures, facial-landmark geometry) and lexical-semantic features (word and sentence meaning), which are largely domain-invariant, but is more questionable for the higher-level, context-dependent cues that distinguish, for example, clinical depression from transient sadness. This tension between generic pretraining and domain-specific downstream tasks recurs throughout the literature reviewed below and directly motivates the choice, in the present pilot, of fine-tuning both the ResNet50 and BERT backbones rather than using them purely as frozen feature extractors.

B. Multimodal Machine Learning for Mental Health

Multimodal machine learning for mental health is an active research area. Al Sahili et al. [11] survey 26 public datasets and 28 fusion architectures spanning audio, visual, physiological, and text modalities, and highlight open challenges in data governance, fairness, and explainability. Nguyen et al. [12] systematically review high-impact multimodal studies on depression, stress, and bipolar disorder detection and report consistent gains from cross-modal fusion over single-modality baselines. A complementary systematic review of 184 passive-sensing studies [13] finds that modality-specific feature correlations vary with demographics and context, underscoring the need for adaptable fusion strategies rather than fixed pipelines. More broadly, Baltrusaitis et al. [14] provide a foundational taxonomy for multimodal machine learning, organizing the field around representation, translation, alignment, fusion, and co-learning, while Lian et al. [15] survey deep-learning-based multimodal emotion recognition across speech, text, and facial modalities specifically.

A recurring theme across these surveys is the distinction between early fusion (combining raw or low-level features before classification), late fusion (combining independent per-modality predictions), and hybrid or intermediate fusion (combining learned representations at an intermediate layer). The stacking ensemble used in this work is best characterized as a hybrid of feature-level and decision-level fusion: modality-specific features are concatenated prior to the meta-learner stage, while the base learners retain some modality specialization (the Multinomial Naive Bayes model operates primarily on textual features). Nguyen et al. [12] note that feature-level concatenation is popular precisely because it allows each modality to be processed independently before fusion, but caution that concatenated feature vectors can become too large or insufficiently discriminative when the number of available training examples is small relative to the feature dimensionality, a concern directly relevant to the pilot-scale evaluation reported in Section V.

C. Textual and Social-Media-Based Mental Health Detection

A large body of work detects depressive and affective states directly from text. BERT and RoBERTa fine-tuned on Reddit-derived corpora achieve high classification accuracy for depression detection [16], and affective and social-norm features have been shown to further improve depression-detection F1-scores when fused with transformer embeddings [17]. Earlier foundational work by De Choudhury et al. [18] demonstrated that behavioral and linguistic signals in Twitter activity, such as

reduced social engagement and heightened negative affect, could predict the onset of clinical depression, and Reece and Danforth [19] extended this line of work to visual social-media data, showing that computationally extracted color and metadata features from Instagram photos outperformed general practitioners' unassisted diagnostic accuracy for depression.

A methodological limitation shared by much of this literature, and directly relevant to the present pilot, is the reliance on self-selected or platform-derived proxy labels rather than clinician-verified diagnoses. De Choudhury et al. [18] crowdsourced a cohort of users who self-reported a clinical depression diagnosis via a standardized recruitment questionnaire, which is considerably closer to clinical ground truth than the emotion-category heuristic used in the present pilot (Section III-A), but still depends on unverified self-report. Reece and Danforth [19] similarly relied on self-reported diagnoses recruited through crowdsourcing platforms. These precedents illustrate that even well-cited studies in this space grapple with the label-quality problem discussed in Section VI-B, reinforcing that clinically-verified ground truth remains an open challenge for the field as a whole rather than a shortcoming unique to this pilot.

D. Visual, Audio, and Physiological Affect Detection

On the visual side, convolutional architectures remain the standard tool for facial-expression recognition, as reviewed comprehensively by Wang et al. [20], with residual and VGG-style backbones widely reused as transfer-learning features for affect recognition tasks. Multimodal clinical corpora such as DAIC-WOZ combine audio, visual, and text streams for depression-severity assessment; decision-level fusion of Gaussian-Mixture-Model visual features, prosodic audio features, and textual features has been shown to be an effective strategy for this task [21]. Beyond facial and audio channels, physiological signals are an increasingly important modality: graph-based deep learning approaches for EEG-based emotion recognition capture functional connectivity between brain regions relevant to affective state [22], and heart-rate-variability features extracted from wearable sensors support lightweight, explainable stress-recognition models [23], motivating this work's planned extension toward a physiological third modality.

It is worth noting that DAIC-WOZ and comparable clinical corpora [21] were collected under structured clinical-interview protocols with PHQ-8-based ground truth, in contrast to the publicly available facial-expression corpus used in the present pilot, which carries only categorical emotion labels rather than clinical severity scores. This distinction is central to the limitations discussed in Section VI: datasets such as DAIC-WOZ represent the type of clinically grounded resource that the present pipeline would need to be evaluated against, or retrained on, before its results could be interpreted as clinically meaningful. The physiological literature reviewed here, particularly the lightweight, interpretable HRV-based models of [23], also informs the design choice, discussed in Section IX, to prioritize interpretable, low-compute physiological features (e.g., HRV time- and frequency-domain statistics) over raw high-dimensional physiological signal processing when this pipeline is extended to a third modality.

E. Ensemble Learning and Class Imbalance Handling

Gradient-boosted trees such as XGBoost [24] are frequently used as strong tabular and feature-level classifiers within larger fusion pipelines, and random forests [25] remain a robust baseline ensemble method across affective-computing and clinical prediction tasks. Class imbalance, a recurring issue in affect and clinical datasets where certain emotion or diagnostic categories are naturally under-represented, is commonly addressed using the Synthetic Minority Oversampling Technique (SMOTE) of Chawla et al. [26], with adaptive variants such as ADASYN [27] generating proportionally more synthetic examples for harder-to-learn minority instances. At the model-combination level, Wolpert's stacked generalization [28] provides the theoretical basis for the two-stage, meta-learner-based ensemble used in this work.

Stacked generalization [28] was originally motivated by the observation that different base learners tend to make correlated but not identical errors, and that a meta-learner trained on held-out base-learner predictions can learn to correct for each base learner's systematic biases rather than simply averaging their outputs. This theoretical property is precisely why a Multinomial Naive Bayes base learner, which is fast but makes a strong conditional-independence assumption over textual features, and an XGBoost base learner, which can model complex nonlinear feature interactions but is more prone to overfitting on small datasets, were selected as complementary base learners in the present pipeline. However, Wolpert's original analysis, and subsequent applications of stacking, assume a base-learner training set that is large enough to produce a stable, well-calibrated set of out-of-fold predictions for the meta-learner to learn from; as discussed in Section VI-C, the small size of the fused evaluation set in this pilot limits how confidently this theoretical benefit can be claimed to hold in practice.

F. Explainability and Privacy in Healthcare AI

As mental-health screening tools move toward clinical relevance, explainability and privacy become central concerns. Random-forest-based explainable AI using LIME and SHAP has been shown to make black-box healthcare predictions more interpretable for clinicians and patients alike [29], while federated learning frameworks address the data-sharing and confidentiality constraints inherent to sensitive health data by allowing collaborative model training without centralizing raw patient records [30]. This work builds on these foundations by combining a ResNet50 visual branch, a BERT-derived textual branch, and a stacking ensemble to jointly model affective and mental-state signals, while treating explainability and privacy as design constraints for future extensions rather than solved problems.

Neither explainability nor federated privacy-preservation is implemented in the present pilot: the stacking ensemble is currently evaluated as a black box, and all data used were processed on local infrastructure rather than in a federated or privacy-preserving setting, since the underlying facial-expression corpus is already a public, de-identified benchmark rather than sensitive patient data. Nonetheless, the explainability techniques surveyed by [29], particularly SHAP-based feature-attribution methods applied to tree-based meta-learners such as XGBoost, are directly compatible with the stacking architecture used here and represent a natural, low-effort extension: because the XGBoost meta-learner operates on a concatenated, relatively low-dimensional feature vector rather than raw pixels or token sequences, SHAP values could be computed for the meta-learner's inputs without requiring a redesign of the underlying architecture. This observation is revisited as a concrete near-term extension in Section IX.

G. Summary of Related Work

Fig. 1 shows the number of reviewed references per thematic category, and Table I below summarizes representative studies alongside modality, method, and reported outcome; Section V-D discusses how the present pilot study's results relate to these prior single- and multi-modality approaches.

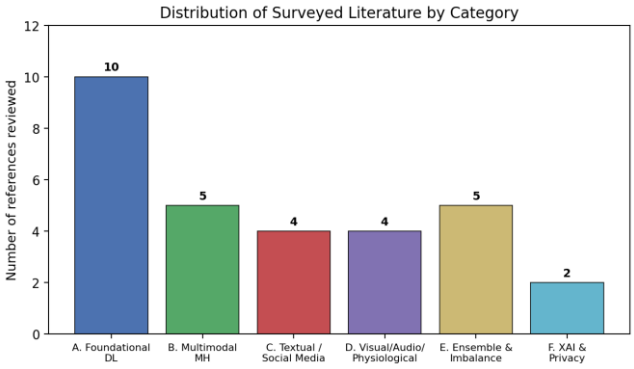

Fig. 1. Distribution of the surveyed literature across the six categories discussed in Sections II-A through II-F.

TABLE I. COMPARATIVE SUMMARY OF REPRESENTATIVE RELATED WORK

Study	Modality	Method / Dataset	Reported Outcome
Nguyen et al. [12]	Text+audio+video	Review of fusion architectures	Consistent gains from cross-modal fusion
Passive-sensing review [13]	Multimodal sensors	Systematic review, 184 studies	Modality correlations vary by context
BERT/RoBERTa study [16]	Text	Fine-tuned BERT/RoBERTa	~98% classification

		on Reddit	accuracy
Triantafyllopoulos et al. [17]	Text	BERT + affective/social-norm feats.	+2.65%/+6.73% F1 over baseline
De Choudhury et al. [18]	Text/behavioral	Social-engagement features, Twitter	~70% depression-onset accuracy
Reece & Danforth [19]	Visual (photos)	Color/metadata feats., Instagram	Outperformed GPs' unassisted rate
Wang et al. [20]	Visual	CNN/transformer FER survey	Summarizes SOTA FER accuracy trends
DAIC-WOZ study [21]	Audio+visual+text	GMM/Fisher-vector fusion, AVEC'17	Established multimodal depression baseline
Liu et al. [22]	Physiological (EEG)	Graph-based DL survey	Graph fusion improves representation
HRV stress study [23]	Physiological (HRV)	k-NN/decision tree, SWELL-KW	Competitive accuracy, low compute

III. PROPOSED METHODOLOGY

A. Dataset Description

The visual modality uses the publicly available 8-Facial-Expressions dataset (Kaggle), organized in YOLO-compatible train/valid/test splits with images/ and labels/ subfolders. The corpus spans nine affective classes — angry, contempt, disgust, fear, happy, natural, sad, sleepy, and surprised — and contains 68,000+ images after cleaning, distributed as summarized in Table II. Each affective class is mapped to a candidate mental-state interpretation (e.g., sad → possible depression/low mood, sleepy → fatigue/burnout, fear → anxiety indicators) to bridge facial affect recognition with mental-health-relevant labels.

TABLE II. DATASET SPLIT SUMMARY (VISUAL MODALITY)

Split	Images	Purpose
Train	54,628	Model training
Validation	13,656	Hyperparameter tuning
Test	~1,000 (subset)	Final evaluation

For the textual modality, a companion structured file pairs each image path with its emotion label and its derived mental-health interpretation, which is treated as short-text input for the language-model branch. This mirrors the design of complementary facial-emotion corpora with demographic metadata (age, gender, country) that were also used during pipeline validation. Table III lists the full heuristic mapping from each of the nine facial-expression categories to its candidate mental-state interpretation; this mapping is a simplifying design choice for the present pilot rather than a clinically validated correspondence, a point discussed further in Section VI-B.

TABLE III. EMOTION-TO-CANDIDATE-MENTAL-STATE MAPPING

Facial Emotion	Candidate Mental-State Interpretation
Happy	Healthy / low risk
Natural	Normal mood
Sad	Possible depression / low mood
Sleepy	Fatigue / burnout
Fear	Anxiety indicators
Angry	Irritability / stress
Contempt	Negative social attitudes
Disgust	Withdrawal / emotional distress
Surprised	Neutral (context-dependent)

B. Formal Problem Formulation

Let x_i denote a facial-expression image and t_i its paired short-text descriptor, both associated with a categorical label y_i corresponding to one of the nine emotion classes in Table III. The visual branch computes a feature representation f^v_i = ResNet50(x_i), a vector of dimensionality d_v drawn from the fine-tuned network's penultimate layer. The textual branch computes f^t_i = BERT(t_i), the pooled contextual embedding of the fine-tuned BERT encoder. The two modality-specific representations are concatenated to form a fused feature vector f_i = [f^v_i ; f^t_i]. The stacking ensemble first computes two base-learner predictions, p_NB = NaiveBayes(f^t_i) and p_XGB = XGBoost(f_i), and a meta-learner then maps the concatenated base-learner outputs to a final prediction, y_hat_i = XGBoost_meta([p_NB ; p_XGB]). Class imbalance in the training distribution of y_i is addressed prior to base-learner training by SMOTE-based oversampling [26], which synthesizes new minority-class feature vectors by linear interpolation between a minority-class sample and one of its k nearest minority-class neighbors in feature space.

C. Mathematical Modeling of Component Learners

This subsection formalizes the training objective and decision rule of each component introduced above, so that the visual, textual, fusion, and ensemble operators are grounded in explicit loss functions rather than treated as black boxes.

Visual branch: The fine-tuned ResNet50 classification head applies a softmax function to its penultimate-layer representation to produce class probabilities, p^v_(i,c) = exp(w_c^T f^v_i + b_c) / sum_c'(exp(w_c'^T f^v_i + b_c')), where w_c and b_c are the weight vector and bias for class c. The network parameters are fine-tuned by minimizing the categorical cross-entropy loss over the N training images, L_v = -(1/N) * sum_i sum_c [1(y_i = c) * log(p^v_(i,c))], using the Adam optimizer [9], which adapts the per-parameter learning rate using running estimates of the first and second moments of the gradient.

Textual branch: The BERT encoder used to compute f^t_i is initialized from pretrained weights obtained via the masked-language-modeling and next-sentence-prediction objectives of [2], and is fine-tuned on the mapped mental-state descriptors using the same categorical cross-entropy form as above, substituting f^t_i for f^v_i and a text-specific classification head for the visual one.

Class balancing: SMOTE [26] generates a synthetic feature vector for a minority-class instance f_j by first identifying its k nearest minority-class neighbors in feature space, selecting one such neighbor f_k at random, and interpolating, f* = f_j + u * (f_k - f_j), where u is drawn uniformly from [0,1], so that f* lies on the line segment between f_j and f_k in feature space, populating the region around existing minority-class examples without duplicating them exactly.

Naive Bayes base learner: The Multinomial Naive Bayes base learner estimates the posterior probability of class c given the textual feature vector via Bayes' theorem under a conditional-independence assumption over feature dimensions, p_NB(i,c) is proportional to P(c) * product_k(P(f^t_(i,k) | c)), where P(c) is the class prior estimated from training-set frequencies and P(f^t_(i,k) | c) is the multinomial likelihood of feature dimension k given class c, estimated via Laplace-smoothed counts.

XGBoost base and meta-learners: Both the XGBoost base learner (operating on f_i) and the XGBoost meta-learner (operating on the concatenated base-learner outputs) are trained by minimizing a regularized objective of the form L_XGB = sum_i(loss(y_i, y_hat_i)) + sum_k(Omega(g_k)), where loss(.,.) is a differentiable classification loss (categorical cross-entropy in this multi-class setting), g_k is the k-th regression tree in the additive ensemble of K trees, and Omega(g_k) = gamma*T_k + (1/2)*lambda*||w_k||^2 penalizes tree complexity through the number of leaves T_k and the leaf-weight magnitudes w_k, following [24]. This regularization is particularly relevant at the pilot scale evaluated in Section V, where a small number of training examples increases the risk of individual trees overfitting to idiosyncratic feature combinations.

D. Algorithmic Summary

The end-to-end training procedure is summarized below.

1. Clean and resize images to 128x128; normalize to [0,1]
2. Tokenize and embed text descriptors via BERT tokenizer
3. Apply SMOTE oversampling to the minority emotion classes
4. Fine-tune ResNet50 on the balanced image set to obtain f^v_i
5. Extract contextual BERT embeddings to obtain f^t_i
6. Concatenate features to form the fused vector f_i
7. Train base learners: Naive Bayes on f^t_i, XGBoost on f_i
8. Generate out-of-fold base-learner predictions on the training set
9. Train the XGBoost meta-learner on the stacked base-learner outputs
10. Return the fine-tuned ResNet50, fine-tuned BERT, base learners, and meta-learner

E. Preprocessing

Image preprocessing removes missing or corrupted files, resizes all frames to 128x128 pixels, and normalizes pixel intensities to [0,1]. Emotion labels are integer-encoded for compatibility with the classification head. Because classes such as contempt and sleepy are under-represented relative to happy or disgust, the Synthetic Minority Oversampling Technique (SMOTE) [26] is applied to generate synthetic minority-class samples and reduce bias toward majority classes. Textual descriptors are tokenized with the BERT tokenizer, embedded with a pretrained BERT model [2], and the resulting contextual features are scaled with a MinMax scaler; textual labels are aligned with the corresponding image labels via label encoding, and SMOTE/random oversampling is again applied where the fused label distribution is imbalanced.

F. Feature Extraction

Visual features are extracted using a ResNet50 backbone [3] pretrained on ImageNet with the classification head removed, followed by a fine-tuned dense classification layer with softmax activation trained for 20 epochs. Textual features are obtained from contextual BERT embeddings [2] of the mapped mental-state descriptors. Both feature streams are produced independently before fusion, allowing each branch to be evaluated in isolation as well as jointly.

G. Hybrid Ensemble Fusion Architecture

The two modality-specific feature vectors are concatenated at the feature level and passed into a stacking ensemble classifier. The base learners are a Multinomial Naive Bayes model, trained primarily on the textual branch, and an XGBoost classifier [24], trained on the fused feature space; an XGBoost meta-learner then combines the base-learner outputs into a final prediction. Fig. 2 illustrates the overall architecture, from raw multimodal input through independent feature extraction, class balancing, feature fusion, and stacked classification.

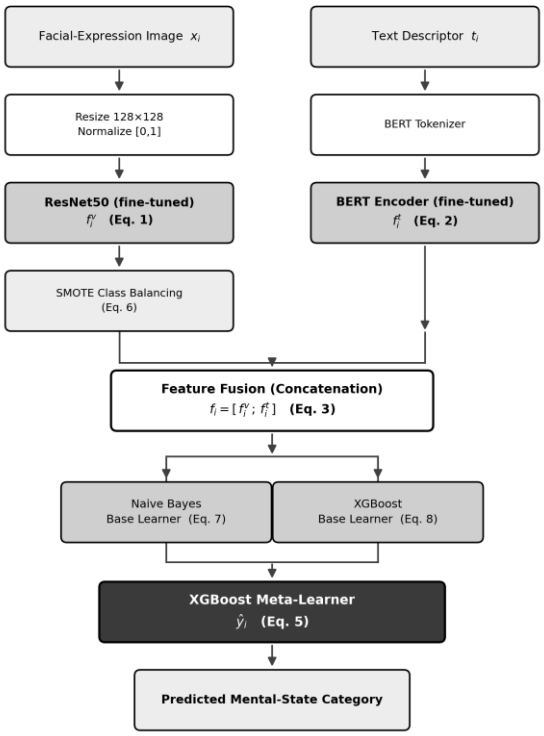

Fig. 2. Hybrid ensemble fusion architecture combining ResNet50 visual features and BERT textual features through a stacking classifier. Equation numbers refer to Section III-B and III-C. Rendered in grayscale.

H. Implementation Details

The pipeline is implemented in Python using PyTorch for the ResNet50 visual branch, the Hugging Face Transformers library (BertTokenizer, BertModel) for the textual branch, and scikit-learn together with the XGBoost library for the Naive Bayes base learner, the XGBoost base and meta-learners, and the StackingClassifier orchestration. Feature scaling for the textual branch uses a MinMax scaler, and class balancing uses the SMOTE implementation from the imbalanced-learn package. Consistent with the pilot-scale framing of this study (Section VI), exact hyperparameter values (learning rate schedules, batch size, weight decay) are treated as implementation detail rather than a tuned contribution of this work, and are reported qualitatively (Section IV) rather than as a tuned configuration validated through systematic hyperparameter search; a full hyperparameter sensitivity study is left for the scaled-up follow-on study outlined in Section IX.

I. Computational Considerations

The two feature-extraction branches have markedly different computational profiles. The ResNet50 visual branch is the more computationally intensive component, since it requires a forward and backward pass through approximately 25 million parameters for every training image across 20 epochs, and benefits substantially from GPU acceleration; this cost is incurred once per image and is independent of the number of modalities eventually fused. The BERT textual branch, applied to short mapped descriptors rather than long documents, incurs a comparatively small computational cost per sample, since the input sequences are short and a single forward pass suffices to obtain the pooled contextual embedding. The stacking ensemble's base and meta-learners operate on already-extracted, low-dimensional feature vectors rather than raw images or token sequences, so their training and inference cost is negligible relative to the two upstream feature-extraction branches. This asymmetry has a practical implication for the third-modality extension discussed in Section IX: physiological feature extractors based on hand-crafted HRV statistics, as reviewed in Section II-D [23], would add comparatively little computational overhead to the existing pipeline, whereas a physiological branch based on raw-signal deep learning would introduce a third computationally intensive branch comparable in cost to the visual branch.

IV. EXPERIMENTAL SETUP

The ResNet50 visual branch is trained for 20 epochs using the split summarized in Table II, with accuracy and cross-entropy loss tracked on training and validation folds each epoch. The fused stacking ensemble is evaluated on a held-out multimodal test set using standard classification metrics: accuracy, precision, recall, and F1-score, computed per class and macro-averaged. All experiments use an 80/20-style train/validation partition for the visual branch and a stratified split for the fused stacking stage to preserve class proportions after oversampling.

A. Evaluation Metrics

For each class c, let TP_c, FP_c, FN_c, and TN_c denote true positives, false positives, false negatives, and true negatives, respectively. Precision and recall are computed as Precision_c = TP_c / (TP_c + FP_c) and Recall_c = TP_c / (TP_c + FN_c). The F1-score is their harmonic mean, F1_c = 2 x Precision_c x Recall_c / (Precision_c + Recall_c), and overall accuracy is computed as the sum of true positives across all classes divided by the total number of evaluated samples N. Macro-averaged metrics are reported as the unweighted mean of the per-class values across all nine classes, so that under-represented classes such as contempt and sleepy contribute equally to the reported average rather than being dominated by majority classes in raw sample count.

B. Evaluation Protocol

Two evaluations are reported. First, the ResNet50 visual branch is evaluated in isolation on the train/validation/test split of Table II, allowing its performance to be assessed independently of the fusion stage. Second, the complete fused pipeline, visual features, textual features, and the stacking ensemble, is evaluated on a separate, smaller held-out multimodal test set for which both modalities are available; this second evaluation is explicitly smaller in scale (Table VI) and is discussed as a pilot-scale feasibility check rather than a generalizable benchmark result (Section VI-C). No cross-validation is performed at the fusion stage in the present pilot; single-split stratified evaluation is used, and extending this to k-fold cross-validation with confidence intervals is identified as a priority for follow-on work (Section IX).

V. RESULTS AND DISCUSSION

A. Unimodal Visual Branch Performance

Table IV reports accuracy, precision, recall, F1-score, and loss for the ResNet50 visual branch across training, validation, and test partitions. The model reaches 92.5% training accuracy and 88.6% test accuracy, with training and validation loss converging to 0.25 and 0.38 respectively, indicating stable convergence without severe overfitting.

TABLE IV. VISUAL BRANCH PERFORMANCE (RESNET50)

Metric	Train	Val	Test
Accuracy (%)	92.50	89.80	88.60
Precision (%)	93.10	90.40	89.20
Recall (%)	92.80	89.50	88.30
F1-score (%)	92.95	89.95	88.75
Loss	0.25	0.38	0.40

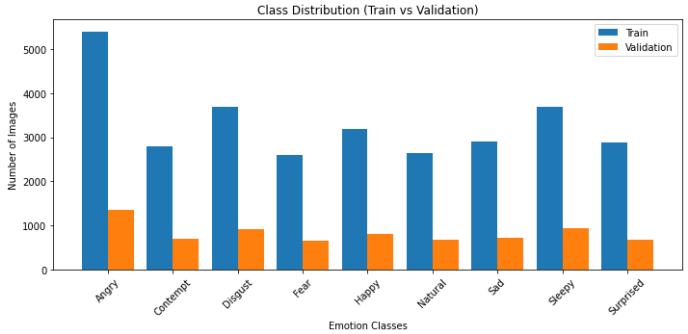

Fig. 3. Class-wise distribution of images across the nine emotion classes in the training and validation sets, showing moderate imbalance in contempt and sleepy classes.

Natural Resources for Human Health
eISSN: 2583-1194

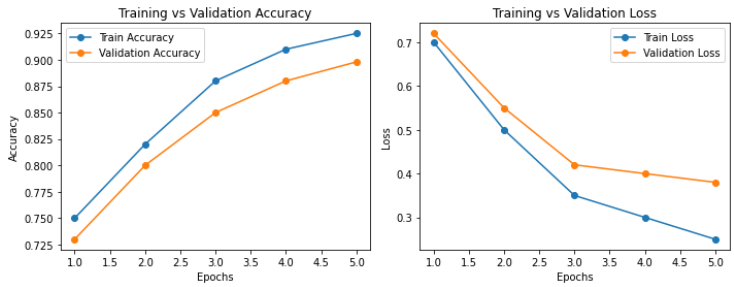

Fig. 4. Training and validation accuracy and loss curves over 20 epochs for the ResNet50 visual branch.

B. Per-Class Visual Branch Metrics

Table V breaks down precision, recall, and F1-score for each of the nine emotion classes on the test set. The natural and disgust classes achieve the strongest balance between precision and recall (F1 = 0.95 and 0.92), while surprised records the lowest precision (0.86), suggesting occasional confusion with other high-arousal expressions. Recall remains above 0.84 for every class despite the initial imbalance, indicating that SMOTE-based rebalancing was effective at preserving sensitivity to minority classes.

Two additional patterns are worth noting. First, contempt achieves the lowest F1-score overall (0.85) among all nine classes; contempt was also one of the classes with the fewest original training examples prior to SMOTE-based rebalancing (Section III-E), consistent with the general expectation that synthetic oversampling can partially, but not fully, compensate for a small original sample size, since the model still has fewer genuinely distinct real examples of contempt to learn from than of higher-support classes such as disgust. Second, the confusion between surprised and other high-arousal expressions noted above is consistent with the facial-action-unit overlap discussed in Section V-D: surprised, fear, and to a lesser extent angry all involve widened eyes and an open or tensed mouth, and distinguishing between them from static images alone, without temporal or contextual information, is a known challenge in the facial-expression-recognition literature more broadly [20].

TABLE V. PER-CLASS METRICS — VISUAL BRANCH (TEST SET)

Class	Prec.	Recall	F1	Support
Angry	0.92	0.86	0.89	2592
Contempt	0.87	0.84	0.85	2813
Disgust	0.97	0.87	0.92	3732
Fear	0.94	0.88	0.91	2657
Happy	0.96	0.87	0.91	3242
Natural	0.96	0.94	0.95	2650
Sad	0.92	0.88	0.90	2928
Sleepy	0.96	0.87	0.91	3728
Surprised	0.86	0.90	0.88	2892

Natural Resources for Human Health
eISSN: 2583-1194

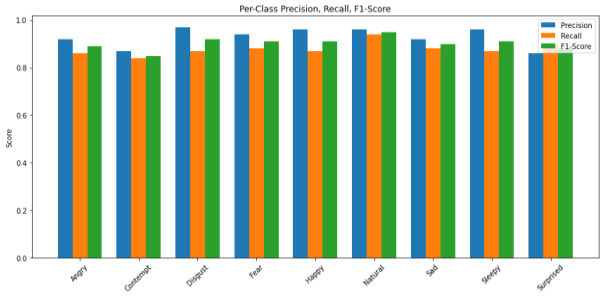

Fig. 5. Per-class precision, recall, and F1-score for the visual branch on the test set.

C. Fused Stacking Ensemble Performance

When visual and textual features are fused through the stacking ensemble (Multinomial Naive Bayes and XGBoost base learners with an XGBoost meta-learner), the model attains an overall accuracy of 75% on the held-out multimodal test set. Table VI reports label-wise precision, recall, and F1-score; performance is strong on several labels (F1 = 1.00 for two categories) but decreases on labels with only one or two support samples in this pilot split, reflecting the small size of the current fused evaluation set rather than a systematic modality-fusion weakness. Fig. 6 shows the resulting confusion pattern across labels.

The per-label pattern in Table VI is informative primarily as a qualitative sanity check rather than a quantitative benchmark, given the sample sizes involved. Labels 0, 2, and 7, each with a single held-out example, are classified perfectly (F1 = 1.00), which at this scale reflects that a single correctly classified example yields a perfect score by construction, not necessarily that those categories are intrinsically easier for the fused model. Conversely, label 4, with two held-out examples, achieves the lowest F1 (0.50), corresponding to exactly one of its two examples being misclassified; at this sample size, a single misclassification produces a large swing in the reported metric. This sensitivity to individual sample outcomes is precisely the statistical concern raised in Section VI-C and Section VII-A, and is the primary reason this paper avoids drawing strong conclusions from the label-wise breakdown in Table VI beyond confirming that the end-to-end pipeline produces syntactically and semantically sensible predictions across the full label space, rather than collapsing to a small subset of majority labels.

TABLE VI. FUSED STACKING ENSEMBLE — PER-LABEL METRICS

Label	Prec.	Recall	F1	Support
0	1.00	1.00	1.00	1
2	1.00	1.00	1.00	1
3	0.50	1.00	0.67	1
4	0.50	0.50	0.50	2
5	1.00	0.50	0.67	2
7	1.00	1.00	1.00	1

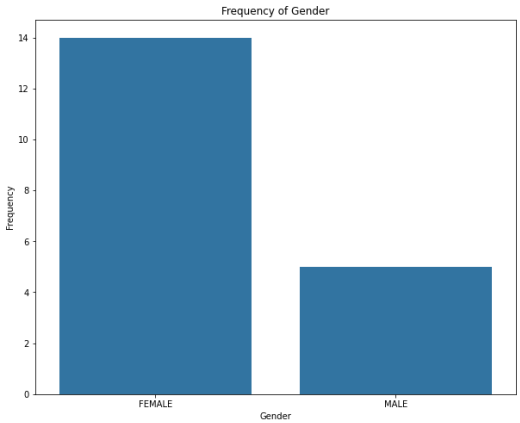

Fig. 6. Confusion matrix of the fused stacking ensemble on the held-out multimodal test set.

D. Discussion

The unimodal ResNet50 branch alone provides strong, stable performance (88.6% test accuracy) on the visual affect-recognition task, consistent with residual-network results reported elsewhere in the facial-expression-recognition literature [3], [20]. The fused stacking ensemble trades some raw accuracy for the ability to jointly reason over visual and textual evidence, which is expected to become more valuable as the evaluation set and label space grow beyond the current pilot scale. These results support a narrower claim than full clinical validation: that hybrid ensemble fusion of complementary feature streams is architecturally viable and trainable end-to-end, consistent with the broader multimodal machine-learning literature [11]-[13]. Class imbalance handling via SMOTE [26] was necessary to keep minority-class recall competitive, and gradient-boosted meta-learning via XGBoost [24] provided a lightweight way to combine heterogeneous modality-specific signals without training a large end-to-end multimodal network from scratch. Section VI discusses why these results should not yet be interpreted as evidence of clinical screening validity.

Relating these findings back to the literature reviewed in Section II, the 88.6% visual-branch test accuracy is broadly consistent with residual-network-based facial-expression-recognition results surveyed by Wang et al. [20], suggesting that the visual branch of this pipeline is not an outlier relative to comparable FER systems trained on similarly structured benchmarks. The fused ensemble's 75% pilot-scale accuracy is more difficult to contextualize against the literature, precisely because of the label-circularity and small-sample issues discussed in Section VI: unlike the DAIC-WOZ-based multimodal depression-severity studies reviewed in Section II-D [21], which report performance against clinician-derived PHQ-8 scores on cohorts of over one hundred participants, the present pilot's fused evaluation cannot yet be meaningfully benchmarked against that literature on equal footing. This gap is precisely the motivation for the prioritized roadmap presented in Section IX.

VI. LIMITATIONS

This pilot study establishes that a ResNet50+BERT stacking-ensemble architecture is technically implementable and trainable end-to-end, but several limitations must be resolved before its results could support real mental-health screening claims. These limitations are organized below by category and are revisited, in more formal terms, in the Threats to Validity discussion of Section VII.

A. Label and Modality Circularity

The textual descriptor used for each sample is a short phrase heuristically derived from that same sample's facial-emotion label (e.g., an image labeled "Sad" is paired with the text "Sad, possible depression/low mood"). Consequently, the text branch does not encode an independent observation about the subject; it substantially restates the visual label. The fused stacking ensemble's reported accuracy should therefore be interpreted as evidence that the architecture can combine two correlated views of the same label, not as evidence of genuine cross-modal mental-state inference. Resolving this requires an independent textual source, such as a subject's own free-text journal entry, transcribed speech, or a validated self-report questionnaire response collected separately from the facial-image capture. We emphasize that this is a data-design limitation rather than an architectural one: the stacking ensemble itself does not assume or require correlated modality labels, and would

be equally applicable to genuinely independent text and image sources once such data becomes available.

B. Absence of Clinically Grounded Labels

The nine affective categories in the underlying facial-expression corpus are emotion labels, not clinical diagnoses. The mapping from an emotion (e.g., sleepy) to a candidate mental-state interpretation (e.g., fatigue/burnout) is a simplifying heuristic introduced for this pilot and has not been validated against any clinician-assigned diagnosis, standardized screening instrument (e.g., PHQ-9, GAD-7), or self-reported clinical history. Any future clinical framing of this pipeline requires a dataset with such ground truth. We note that this limitation is shared, to varying degrees, by several of the social-media-based studies reviewed in Section II-C, which also rely on self-report or platform-derived proxy labels rather than clinician-verified diagnoses; it is not unique to the present pilot, but it is nonetheless a precondition for any clinically meaningful extension of this work.

C. Small and Imbalanced Fused-Evaluation Set

The held-out set used to evaluate the fused stacking ensemble (Table VI) contains only 1-2 samples for several labels, which is too small to support statistically reliable per-class precision, recall, or F1 estimates. The 75% overall accuracy reported in Section V-C should be read as a pilot-scale feasibility indicator rather than a generalizable performance estimate. No confidence intervals or variance estimates across multiple random splits are reported for this reason; computing such intervals on a dataset this small would create a false impression of statistical rigor that the underlying sample size cannot support.

D. Missing Ablation and Baseline Comparisons

This pilot does not yet report a controlled comparison of unimodal (text-only, visual-only) versus fused performance on an identical, adequately sized test set, nor a comparison against an established multimodal depression-detection benchmark such as DAIC-WOZ. Such ablations, together with confidence intervals or multiple-seed variance estimates, are necessary to demonstrate that fusion provides a genuine benefit over either modality alone. In particular, because the textual branch is currently derived from the visual label (Section VI-A), a text-only ablation on the present dataset would not be informative until an independent textual data source is available; this ablation is therefore deferred to the follow-on study outlined in Section IX rather than reported here on data that would not support a meaningful interpretation.

E. Generalizability Across Populations and Settings

The visual and textual data used in this pilot originate from a single, publicly available benchmark corpus collected under uncontrolled, non-clinical conditions (e.g., varied lighting, pose, and camera quality typical of crowd-sourced facial-expression datasets). No information is available on the demographic composition of the underlying subject pool, and no cross-dataset or cross-population evaluation has been performed. Consequently, the reported performance figures should not be assumed to generalize to clinical populations, to different age groups, to different cultural or linguistic contexts, or to data collected in a controlled clinical-interview setting such as DAIC-WOZ [21].

F. Ethical and Privacy Considerations

This pilot uses a public, de-identified facial-expression benchmark and does not involve new human-subject data collection; however, any extension toward real screening use would require informed consent, institutional ethics review, demographic bias auditing, and explicit data-privacy safeguards, particularly given the sensitivity of inferring mental-health-adjacent labels from facial imagery. Because the underlying emotion-to-mental-state mapping (Table III) has not been clinically validated, any deployment context that presented model outputs to end users or clinicians without clearly communicating this pilot status would risk conveying unwarranted diagnostic confidence, an outcome this paper explicitly seeks to avoid through the limitations and threats-to-validity discussion provided here.

VII. THREATS TO VALIDITY

Beyond the pilot-specific limitations enumerated in Section VI, it is useful to frame the study's shortcomings using the standard internal-, external-, and construct-validity taxonomy common in empirical machine-learning and software-engineering research.

A. Internal Validity

Internal validity concerns whether the observed effects (e.g., the reported accuracy and F1-scores) can be attributed to the fusion architecture itself rather than to confounding factors in the experimental setup. The principal internal-validity threat in this pilot is the label circularity described in Section VI-A: because the textual input is derived from the same label the model is trying to predict, the fused ensemble's accuracy partly reflects its ability to recover a label from a near-paraphrase of itself, rather than from genuinely independent multimodal evidence. A secondary internal-validity threat is the absence of a fixed random seed protocol across the visual-branch and fusion-stage experiments reported here; single-run results, without repeated trials across multiple random seeds, cannot rule out the possibility that reported performance differences between the visual-only and fused configurations partly reflect run-to-run variance rather than a systematic architectural effect.

B. External Validity

External validity concerns the extent to which the reported results generalize beyond the specific dataset and population studied. The 8-Facial-Expressions corpus used here, like most publicly available facial-expression datasets, was not collected with demographic balance across age, gender, ethnicity, or cultural background as an explicit design goal, so the visual branch's generalization to demographic groups under-represented in the training data is untested. Facial expression of affect is also known in the psychological literature to vary across cultural contexts, meaning that a mapping such as Table III, developed without cross-cultural validation, may not transfer to populations whose expressive norms differ from those represented in the source corpus. The fused-ensemble evaluation set is additionally too small (Section VI-C) to support any claim about generalization to a broader population at all.

C. Construct Validity

Construct validity concerns whether the quantities being measured actually capture the underlying construct of interest, in this case, mental-health status. This is the most significant validity threat facing the present pilot: the outcome variable being predicted is a heuristic emotion-to-mental-state mapping (Table III), not a validated measure of mental health such as a PHQ-9 score or clinician diagnosis. Consequently, even a hypothetically perfect classifier on this dataset would only be demonstrably measuring facial-expression category, not mental-health status; the mapping in Table III is a research hypothesis about a possible correspondence, not an established psychometric instrument. This construct-validity gap is the central reason this paper is framed throughout as a pilot feasibility study for the fusion architecture rather than a validated mental-health screening result.

VIII. REPRODUCIBILITY STATEMENT

In the interest of transparency, we summarize here the information a reader would need to reproduce or extend this pilot study. The visual modality uses the publicly available 8-Facial-Expressions dataset (Kaggle), described in Section III-A; the textual modality is derived programmatically from the visual labels using the mapping in Table III, and is therefore fully reproducible from the visual labels alone, though as discussed in Section VI-A this also means it does not constitute an independent data source. The ResNet50 and BERT backbones, SMOTE oversampling, and stacking-ensemble components are all built from openly available libraries (PyTorch, Hugging Face Transformers, scikit-learn, XGBoost, imbalanced-learn), as detailed in Section III-H, rather than from custom or proprietary implementations. All reported metrics (Tables II through VI) are computed using the standard precision/recall/F1/accuracy formulas of Section IV-A. Given the pilot-scale framing of this work (Section VI), we have intentionally not reported a fixed hyperparameter configuration as a tuned contribution; a fully reproducible, versioned implementation with pinned library versions, fixed random seeds, and a documented hyperparameter search is planned as part of the scaled-up follow-on study described in Section IX, at which point code and data-processing scripts are intended to be made available to support independent replication.

IX. CONCLUSION AND FUTURE WORK

This paper reported a proof-of-concept pilot study of a hybrid ensemble architecture for emotion-linked mental-state screening that fuses facial-expression imagery, processed through a ResNet50 backbone, with short textual descriptors, encoded via BERT, using a stacking classifier for final prediction. The visual branch achieved 88.6% test accuracy, and the fused stacking ensemble reached 75% accuracy on a small pilot evaluation set, demonstrating that the proposed fusion architecture is technically implementable end-to-end. As discussed in Section V-D, these results should be read as evidence of

architectural feasibility rather than validated screening performance: the textual branch is currently derived from the same labels as the visual branch rather than an independent modality, and the fused evaluation set is too small for statistically robust claims.

Building on the limitations (Section VI) and threats to validity (Section VII) identified above, future work will prioritize the following steps, roughly in order of dependency.

A. Near-Term: Independent-Modality Data Collection

The most immediate priority is replacing the derived text descriptors with a genuinely independent textual signal, such as free-text journaling entries, transcribed speech from a structured interview, or validated questionnaire responses (e.g., PHQ-9 free-text comments) collected separately from the facial-image capture. Only once the two modalities are independently observed can the fused ensemble's performance be attributed to genuine cross-modal inference rather than restated-label recovery (Section VI-A, Section VII-A).

B. Near-Term: Clinically Grounded Labeling

In parallel, the project will pursue access to, or construction of, a dataset with clinician-assigned diagnoses or standardized-instrument scores (e.g., PHQ-9, GAD-7) rather than emotion-category proxies, potentially by adapting the pipeline to established clinical corpora such as DAIC-WOZ [21] or by collaborating with clinical partners under appropriate ethical oversight (Section VI-B, Section VI-F).

C. Medium-Term: Statistically Robust Evaluation

Once independent-modality, clinically labeled data is available, the fused evaluation set will be scaled substantially and assessed using stratified k-fold cross-validation, confidence intervals over multiple random seeds, and controlled unimodal-versus-fused ablations against the same test set (Section VI-C, Section VI-D). This step is a precondition for making any generalizable performance claim.

D. Medium-Term: Explainability Integration

As noted in Section II-F, the XGBoost meta-learner's relatively low-dimensional input makes it a natural candidate for SHAP-based feature-attribution analysis without requiring architectural changes. Integrating such explainability tooling would allow clinicians to inspect which modality and which features contributed most to a given prediction, directly addressing the explainability gap discussed in Section II-F.

E. Long-Term: Third-Modality Extension

Finally, in line with the original multimodal scope of this research, the framework will be extended to incorporate a third modality, physiological signals such as heart-rate variability and electrodermal activity, drawing on the lightweight, interpretable feature-extraction strategies reviewed in Section II-D [23]. Ethical safeguards, informed consent, data privacy, demographic bias auditing, and clinician-in-the-loop validation will remain central design constraints at every stage of this roadmap, and no stage of this work will be represented as validated for clinical screening until subjected to appropriate ethical review and peer-reviewed clinical validation.

REFERENCES

[1] World Health Organization, "Comprehensive Mental Health Action Plan 2013-2030," WHO, 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240031029

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805, 2018/2019. DOI: 10.48550/arXiv.1810.04805. [Online]. Available: https://arxiv.org/abs/1810.04805

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv:1512.03385, 2015. DOI: 10.48550/arXiv.1512.03385. [Online]. Available: https://arxiv.org/abs/1512.03385

[4] A. Vaswani et al., "Attention is all you need," arXiv:1706.03762, 2017. DOI: 10.48550/arXiv.1706.03762. [Online].

Available: https://arxiv.org/abs/1706.03762

[5] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019. DOI: 10.48550/arXiv.1907.11692. [Online]. Available: https://arxiv.org/abs/1907.11692

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014. DOI: 10.48550/arXiv.1409.1556. [Online]. Available: https://arxiv.org/abs/1409.1556

[7] I. J. Goodfellow et al., "Generative adversarial networks," arXiv:1406.2661, 2014. DOI: 10.48550/arXiv.1406.2661. [Online]. Available: https://arxiv.org/abs/1406.2661

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997. DOI: 10.1162/neco.1997.9.8.1735. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980, 2014. DOI: 10.48550/arXiv.1412.6980. [Online]. Available: https://arxiv.org/abs/1412.6980

[10] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv:1502.03167, 2015. DOI: 10.48550/arXiv.1502.03167. [Online]. Available: https://arxiv.org/abs/1502.03167

[11] Z. Al Sahili, M. Patwary, and M. Purver, "Multimodal machine learning in mental health: A survey of data, algorithms, and challenges," arXiv:2407.16804, 2024. DOI: 10.48550/arXiv.2407.16804. [Online]. Available: https://arxiv.org/abs/2407.16804

[12] T. T. Nguyen, V. H.-Q. Pham, D.-T. Le, X.-S. Vu, F. Deligianni, and H. D. Nguyen, "Multimodal machine learning for mental disorder detection: A scoping review," Procedia Computer Science, vol. 225, pp. 1458-1467, 2023. DOI: 10.1016/j.procs.2023.10.134. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923012929

[13] L. S. Khoo, M. K. Lim, C. Y. Chong, and R. McNaney, "Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches," Sensors, vol. 24, no. 2, p. 348, 2024. DOI: 10.3390/s24020348. [Online]. Available: https://www.mdpi.com/1424-8220/24/2/348

[14] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," arXiv:1705.09406, 2017. DOI: 10.48550/arXiv.1705.09406. [Online]. Available: https://arxiv.org/abs/1705.09406

[15] H. Lian et al., "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," Entropy, vol. 25, no. 10, p. 1440, 2023. DOI: 10.3390/e25101440. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10606253

[16] F. I. Kurniadi, N. L. P. S. P. Paramita, E. F. A. Sihotang, and M. S. Anggreainy, "BERT and RoBERTa models for enhanced detection of depression in social media text," Procedia Computer Science, vol. 245, pp. 202-209, 2024. DOI: 10.1016/j.procs.2024.10.244. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050924030527

[17] I. Triantafyllopoulos et al., "Depression detection in social media posts using affective and social norm features," arXiv:2303.14279, 2023. DOI: 10.48550/arXiv.2303.14279. [Online]. Available: https://arxiv.org/abs/2303.14279

[18] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proc. 7th Int. AAAI Conf. Weblogs and Social Media (ICWSM), 2013. DOI: 10.1609/icwsm.v7i1.14432. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14432

[19] A. G. Reece and C. M. Danforth, "Instagram photos reveal predictive markers of depression," EPJ Data Science, vol. 6, no. 15, 2017. DOI: 10.1140/epjds/s13688-017-0110-z. [Online]. Available: https://arxiv.org/abs/1608.03282

[20] Y. Wang et al., "A survey on facial expression recognition of static and dynamic emotions," arXiv:2408.15777, 2024. DOI: 10.48550/arXiv.2408.15777. [Online]. Available: https://arxiv.org/abs/2408.15777

[21] S. Dham, A. Sharma, and A. Dhall, "Depression scale recognition from audio, visual and text analysis,"

arXiv:1709.05865, 2017. DOI: 10.48550/arXiv.1709.05865. [Online]. Available: https://arxiv.org/abs/1709.05865

[22] C. Liu et al., "A comprehensive survey on EEG-based emotion recognition: A graph-based perspective," arXiv:2408.06027, 2024. DOI: 10.48550/arXiv.2408.06027. [Online]. Available: https://arxiv.org/abs/2408.06027

[23] M. Bahameish, T. Stockman, and J. Requena Carrion, "Strategies for reliable stress recognition: A machine learning approach using heart rate variability features," Sensors, vol. 24, no. 10, p. 3210, 2024. DOI: 10.3390/s24103210. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11126126

[24] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD, 2016, pp. 785-794. DOI: 10.1145/2939672.2939785. [Online]. Available: https://arxiv.org/abs/1603.02754

[25] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001. DOI: 10.1023/A:1010933404324. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002. DOI: 10.1613/jair.953. [Online]. Available: https://arxiv.org/abs/1106.1813

[27] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in Proc. IEEE IJCNN, 2008, pp. 1322-1328. DOI: 10.1109/IJCNN.2008.4633969. [Online]. Available: https://ieeexplore.ieee.org/document/4633969

[28] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241-259, 1992. DOI: 10.1016/S0893-6080(05)80023-1. [Online]. Available: https://doi.org/10.1016/S0893-6080(05)80023-1

[29] M. Panda and S. Debbarma, "Explainable artificial intelligence for healthcare applications using random forest classifier with LIME and SHAP," arXiv:2311.05665, 2023. DOI: 10.48550/arXiv.2311.05665. [Online]. Available: https://arxiv.org/abs/2311.05665

[30] H. Jeong, "Security and privacy issues and solutions in federated learning for digital healthcare," arXiv:2401.08458, 2024. DOI: 10.48550/arXiv.2401.08458. [Online]. Available: https://arxiv.org/abs/2401.08458