Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/200418

Multimodal Deep Neural Network for Suicide Risk Classification Using Audio and Facial Features

A. B. Mukhametzhanov[1,*], *A. Shoiynbek*[2], *A. Nurzhas*[3], *B. Meraliyev*[1], *D. Kuanyshbay*[1], *S. Sklyar*[4], *and P. Menezes*[5]

[1]Faculty of Engineering and Natural Sciences, SDU University, Kaskelen, 040900, Kazakhstan
[2]School of Digital Technology, Narxoz Univeristy, Almaty, 050035, Kazakhstan
[3]Faculty of Computer Science and Engineering Astana IT University, Astana, 010000, Kazakhstan
[4]General and Applied Psychology Department, Faculty of Philosophy and Political Science, Al-Farabi KazNU, Almaty, 050040, Kazakhstan
[5]Institute of Systems and Robotics, University of Coimbra, Coimbra, 3030-788, Portugal

Received: 20 Mar. 2026, Revised: 30 Apr. 2026, Accepted: 20 Jun. 2026
Published online: 1 Jul. 2026

Abstract: Suicide-risk assessment remains a major public-health priority, yet current clinical evaluations rely heavily on subjective judgment and self-disclosure. This study investigates the effectiveness of combining low-level audio features and deep visual embeddings for suicide classification using interview-style recordings. The dataset consists of 103 clinical interviews stratified into control and at-risk groups, from which Mel-Frequency Cepstral Coefficients (MFCCs) and VGG16-derived facial embeddings were extracted through a structured preprocessing and alignment pipeline. Unimodal and multimodal models were evaluated using classical machine learning algorithms and a shallow neural network. The best-performing multimodal model, a two-layer neural network, achieved 0.82 accuracy and 0.82 macro F1-score, outperforming logistic regression, gradient boosting, XGBoost, and naive Bayes baselines. The results demonstrate the predictive potential of facial embeddings for suicide-risk detection and provide a reproducible feature-alignment pipeline for future multimodal mental-health classification studies.

Keywords: deep learning, MFCC, multimodal fusion, suicide classification, VGG16

1 Introduction

Suicide remains a critical global public-health challenge, claiming more than 720,000 lives every year according to the World Health Organization (WHO) and ranking among the leading causes of death for individuals aged 15–29 [1].

Early detection of suicidal ideation is widely recognized as a critical component of effective suicide prevention. Prior research indicates that traditional suicide-risk assessment predominantly relies on clinical methods such as self-report questionnaires, clinician-administered scales, and face-to-face interviews, which require substantial expertise and direct patient–clinician interaction and may therefore be subjective, time-consuming, and difficult to scale to large populations [2,3]. To address these limitations, a growing body of work has explored automated suicide detection methods using machine learning techniques applied to naturally occurring data sources. Existing studies report that behavioral and physiological signals observable in audio and visual recordings—such as vocal prosody, speech energy, facial expressions, and other subtle behavioral cues—carry informative markers related to suicidal ideation and depressive symptoms [4]. These findings have motivated the use of speech, affective, and multimodal features as promising signals for objective suicide-risk assessment, complementing traditional clinical evaluation.

Despite recent advances in machine-learning-based suicidal ideation detection, several challenges remain. Prior work highlights that suicidal signals are often subtle, context-dependent, and temporally evolving, which limits the effectiveness of unimodal or purely statistical approaches [5,6]. A persistent barrier identified in the literature is the scarcity of well-annotated, ethically collected datasets, which constrains model generalization across domains and populations and hampers further progress in automatic suicide-risk assessment [4].

* Corresponding author e-mail: asyl.512000@gmail.com

The present study aims to address this gap by evaluating MFCC-based audio features and VGG16-based facial representations for suicide classification using an interview-style dataset. A series of experiments is conducted to compare unimodal models trained separately on audio or video features with multimodal fusion approaches.

The main contributions of this study are as follows:

- A structured preprocessing pipeline is developed for extracting and aligning MFCC audio features and VGG16 facial embeddings from raw interview recordings.
- Baseline experimental results are reported for suicide-risk classification across unimodal and fused models, offering insights into the relative importance of each modality.
- An analysis of fusion strategies is presented, highlighting the potential gains from combining complementary cues and providing guidance for future multimodal mental-health detection systems.

The remainder of this paper is organized as follows. Section 2 presents the materials and methods, including the dataset, preprocessing, feature extraction, and multimodal modeling procedure. Section 3 reports the experimental results and discusses their implications. Finally, the concluding remarks are given in the last part of the paper.

2 Materials and Methods

2.1 Related works

Min et al. (2023) examined the potential of acoustic voice analysis for assessing suicidality among psychiatric patients with mood disorders using machine learning techniques [7]. The study utilized 348 voice recordings collected from 104 patients over multiple time points (baseline and up to 12 months). Two complementary models were developed: a between-person classifier, which detected high-risk individuals based on cross-sectional vocal characteristics, and a within-person model, which identified increased suicidality over time through intra-individual changes in acoustic patterns. Acoustic features such as temporal, formal, and spectral descriptors were extracted, and models were trained using artificial neural networks and extreme gradient boosting (XGBoost). The between-person model achieved 69% accuracy (AUC = 0.62), while the within-person model reached 79% accuracy (AUC = 0.67) in predicting worsening suicidality. The findings highlight that longitudinal variations in speech acoustics can serve as effective indicators of suicide risk, suggesting the feasibility of automated voice-based monitoring tools for clinical and telehealth applications.

There is a private speech dataset gathered by Wang et al., 2025 from the Beijing Huilongguan Hospital psychological support hotline, consisting of 20,630 segmented audio sentences from 105 subjects (total length 80 hours) [8]. Each segment was annotated by experts into negative and non-negative emotions and linked with 12-month follow-up suicide risk outcomes (22 suicidal vs. 24 non-suicidal cases). The study also used the public Vietnamese ViSEC dataset, which contains 5,280 samples from 147 speakers, labeled for emotion, gender, dialect, and speaker identity. Both datasets were resampled to 16 kHz and standardized to 10-second audio clips. Experiments used Wav2Vec2, MFCC, and pitch features fused with cross- and self-attention mechanisms for emotion classification and suicide risk differentiation. On the private dataset, the proposed fusion model achieved an F1-score of 79.13%, outperforming Wav2Vec2 and Whisper baselines. Emotion trend analysis revealed higher emotional change rates and more negative speech segments among suicidal participants, suggesting that emotional instability correlates with suicide risk.

There is a dataset of facial images gathered in study by Kong et al., 2022, where 102 patients with depression and 1132 healthy participants were recruited at the Affiliated Hospital of Shandong University of Traditional Chinese Medicine [9]. Each participant contributed 10 facial photos, captured under standardized clinical conditions using an MI PAD3 camera. After face cropping and normalization (pixel values scaled from 0–255 to 0–1), the dataset was split into training (70%), testing (20%), and validation (10%) subsets. Depression diagnosis followed DSM-5 and Hamilton Depression Scale criteria (scores 17–30), with exclusion of individuals having comorbid or substance-related disorders. In this study, five deep convolutional neural network architectures—FCN with an attention mechanism, VGG11, VGG19, ResNet50, and Inception-V3—were implemented in PyTorch with a batch size of 16, learning rate of 1e-3, and 30 training epochs. Classification was performed using Softmax activation with cross-entropy loss. Model performance was evaluated through accuracy, precision, recall, F1-score, and ROC curves. The FCN model achieved the best overall results, with 98.23% accuracy, 98.11% precision, 98.16% recall, and 98.27% F1-score, outperforming all other tested CNN architectures.

There is a multimodal depression risk detection dataset gathered by Zhang et al., 2024, where researchers at Shenzhen Kangning Hospital recorded synchronized audio, video, and text data from 1911 participants during a structured emotion elicitation paradigm [10]. The paradigm included a reading task (“The Little Match Girl” passage) to induce initial emotions, followed by an interview task with guided personal questions to elicit deeper emotional responses. Each participant’s depression level was labeled using PHQ-9 scores, with 621 classified as at-risk and 1290 as healthy. Video data were preprocessed by cropping face regions using MediaPipe and resizing to 224×224 pixels; audio was segmented and cleaned for interviewee speech, from which MFCC

features (40 filter banks, 16 kHz) were extracted; and text transcripts were generated via automatic speech recognition and manually proofread. A balanced subset of 240 subjects (120 depressed, 120 healthy) was used for experiments. In evaluation, the proposed AVTF-TBN model, integrating audio (GRU + MHA), video (Swin-Transformer + BiLSTM + MHA), and text (BERT + MHA) through an attention-based multimodal fusion (MMF) module, achieved strong detection results — reaching an F1-score = 0.78, Precision = 0.76, and Recall = 0.81 when combining data from both reading and interview tasks. The results demonstrated that multimodal integration and emotional task design substantially improve depression risk detection accuracy.

There is a study by Nykoniuk et al. (2025) that explored multimodal approaches for depression detection using synchronized audio and text from clinical interviews between participants and a virtual agent [11]. The researchers combined data from the DAIC-WOZ and EDAIC-WOZ corpora to address class imbalance and applied preprocessing steps such as noise reduction, tone-shift augmentation, and TF-IDF-based text cleaning. Two neural fusion architectures were compared—an early fusion network integrating features at input level and a late fusion model merging outputs from modality-specific branches. The early fusion approach outperformed the late fusion one, achieving 94% validation accuracy and 85% test accuracy, demonstrating that early integration of multimodal features improves the reliability of depression detection systems.

There is a study by Pigoni et al. (2025) that explored the use of multimodal machine learning to predict 12-month suicide attempts in patients with bipolar disorder [12]. The research analyzed data from 163 hospitalized subjects using clinical information and MRI-derived features such as gray matter (GM) volumes and cortical thickness (CT). Several unimodal SVM classifiers were built separately for clinical and MRI data, followed by a stacking-based fusion model integrating the modalities. The clinical model with time-variant features achieved an AUC of 0.83, while the GM-based model reached AUC = 0.86. When combined, the multimodal classifier improved performance to AUC = 0.88, BAC = 83.4%, and sensitivity = 80%, correctly identifying 8 of 10 suicide attempters. Key predictors included prior suicide attempts, suicidal ideation in the past year, and gray matter alterations in regions such as the frontal pole, thalamus, and cerebellum.

There is a paper by Alghowinem et al. (2023) that presents a comprehensive multimodal framework for suicide-risk assessment using behavioral cues from audio, facial expression, head movement, and full-body gestures [13]. The authors recorded 14 interview videos (10 participants retained after anonymization), aged 15-25, with suicide-risk levels labeled as high, medium, or low. To compensate for the small sample size, interviews were segmented into 2-, 5-, and 10-second windows, yielding multiple samples per participant. Another research effort in this work focused on evaluating multiple fusion strategies across handcrafted and deep representations. Handcrafted multimodal fusion (combining face expression, body gestures, and speech prosody) achieved sample-level accuracies between 0.59 and 0.64, with subject-level accuracies up to 0.80 when fusing all facial and body-region cues. Deep-feature models using I3D showed substantially higher performance, reaching 0.92–0.96 sample-level accuracy and up to 1.0 subject-level accuracy for nose–mouth and multi-region facial fusion configurations. Late-fusion multimodal systems combining handcrafted features with VGGish audio embeddings further improved performance, achieving sample-level accuracies of 0.94-0.96 and perfect subject-level classification (1.00) in several configurations. Overall, the results demonstrated that deep multimodal fusion, particularly combining facial micro-regions with audio - yielded the strongest predictive power, significantly outperforming single-modality models.

There is a study by Shaikh et al. (2025) that proposes a multimodal deep learning framework integrating text, images, and audio for suicide ideation detection [14]. The dataset includes 42,000 text posts, 12,000 images, and 5,000 audio clips from ethically sourced online forums. Modality-specific preprocessing was applied: text was tokenized and cleaned, images were normalized and augmented, and audio signals were converted into MFCCs. The model uses BERT for text embeddings, ResNet-50 for image features, and a BiLSTM for speech patterns, with an attention-based fusion mechanism projecting all modalities into a shared latent space. Experimental results show that the multimodal model achieves the highest performance compared to unimodal baselines, with precision = 91.2%, recall = 89.8%, F1-score = 90.5%, and accuracy = 90.9%, significantly outperforming text-only or text+audio models. The system's explainability tools (SHAP and LIME) improved clinician trust, while a federated learning setup preserved user privacy, confirming both practical effectiveness and ethical feasibility.

Another study by Hu et al.(2025) introduces an integrated framework combining multimodal learning, multitask learning, and transfer learning for joint prediction of depression severity (DS) and suicide risk (SR) using Chinese clinical data [15]. The authors collected a dataset of 200 patients (100 depressed, 100 controls) that includes audio descriptions, transcripts, and clinical questionnaire scores, labeled into three DS levels (HAMD-17) and two SR levels (SAD PERSONS). The proposed model extracts features using state-of-the-art pretrained models - wav2vec 2.0, HuBERT for audio and Longformer, ERNIE-health for text, and fuses them through early fusion before multitask prediction. Experimental results show that multimodal and multitask models consistently outperform single-modal and single-task baselines, with the best combinations achieving AUCs above 0.87 for DS and above 0.8 for SR.

Overall, the study demonstrates that integrating audio and text with MTL substantially improves automated mental-health screening in non-English clinical contexts.

A summary of the representative studies discussed in this subsection, including the target task, modalities, methods, and key performance results, is presented in Table 1.

2.2 Data collection and description

The study utilized a previously established dataset of 103 video-recorded clinical interviews, originally collected in collaboration with a partner clinic in Kazakhstan. Participants were stratified into two groups according to clinical evaluations and Beck Depression Inventory (BDI) scores [16]:

- Control group ($n = 69$): individuals without significant depressive symptoms or suicide-risk indicators.
- Experimental group ($n = 34$): individuals with a history of suicide attempts or elevated suicide-risk markers identified through clinical assessment.

The dataset is imbalanced toward the control group. Participants ranged in age from 18 to 60 years (overall mean = 25.7, SD = 9.6). Females represented the majority in both groups (76% in the experimental group and 74% in the control group). Specifically, the control group had a mean age of 25.8 years (range 19–60; 51 female, 18 male), while the experimental group had a mean age of 25.6 years (range 18–49; 26 female, 8 male).

Prior to participation, all individuals underwent standardized psychological screening that included the Standardized Multifactorial Personality Inventory (SMIL) [17] and the BDI. Each session consisted of a 20–30-minute video-recorded interview with synchronized audio and video streams. The recordings were later segmented into structured transcripts containing utterance-level start and end times, speaker identifiers, and textual content. The prediction task was formulated as binary classification distinguishing between control and at-risk participants.

2.3 Face detection and frame extraction

The video preprocessing pipeline was designed to extract frame-level facial regions from interview recordings while maintaining the hierarchical dataset organization (group - subject - segment - video). The raw recordings were processed iteratively across all directories to ensure structured and consistent output. Each video file was read frame by frame using OpenCV [18], with frame sampling performed at approximately one frame per second to balance temporal coverage and computational efficiency. Face localization was achieved using the RetinaFace detector, a state-of-the-art facial detection model capable of identifying faces under variable lighting and head-pose conditions [19]. For each detected face, bounding box coordinates were extracted and used to crop facial regions from the original frame. Cropped face images were then saved into a dedicated folder structure that preserved the group (control/experimental), subject ID, and segment label. Invalid or corrupted frames were automatically skipped, and any detection errors were logged for later inspection. This step resulted in a large corpus of cropped facial images, standardized for subsequent feature extraction. The final dataset contained thousands of images per group, covering a range of emotional and behavioral expressions during the interview sessions.

2.4 Subject-independent evaluation protocol

To ensure subject-independent evaluation and avoid identity leakage, the dataset was split at the subject level rather than the frame or chunk level. For each class group, three subjects were reserved exclusively for testing in the visual-feature study, while the remaining subjects were used for training. In the aligned multimodal experiments, subjects were stratified by class and split into training (80%) and testing (20%) sets using a fixed random seed of 42. This produced 54 training subjects and 14 held-out test subjects. All chunks from a given participant were assigned to only one split. Thus, the reported chunk-level predictions were computed on chunks from unseen subjects rather than on randomly mixed chunks from the same individuals.

2.5 Visual feature representation

For visual representation learning, features were extracted from the VGG16 convolutional neural network, pre-trained on the ImageNet dataset [20]. The fully connected layers fc1 and fc2 were selected as high-level embedding sources, as they capture both mid-level appearance information and abstract semantic cues relevant to facial expression analysis. Each cropped face image was resized to 224×224 pixels and normalized according to the VGG16 preprocessing scheme. The activations from the fc1 and fc2 layers were computed using a forward pass through the network, concatenated into a single feature vector, and flattened into a 1D representation for downstream analysis.

2.6 Visual feature-set selection

To investigate the impact of feature representation on facial-based classification, a systematic exploration of feature vectors extracted from multiple layers of a pretrained VGG16 network was performed [21,22,23]. The following layer groups were considered:

Table 1: Summary of representative related studies

Study	Task / Population	Modalities	Methods	Key Result
Min et al. (2023)	Suicide risk in mood disorders	Audio features	ANN, XGBoost	69% / 79% accuracy
Wang et al. (2025)	Hotline suicide-risk differentiation	Audio fusion	Attention fusion	F1 = 79.13%
Kong et al. (2022)	Depression detection	Facial images	CNNs	98.23% accuracy
Zhang et al. (2024)	Depression-risk detection	Audio, video, text	Multimodal fusion	F1 = 0.78
Nykoniuk et al. (2025)	Depression detection	Audio, text	Early/late fusion	85% test accuracy
Pigoni et al. (2025)	Suicide-attempt prediction	Clinical + MRI	Stacking fusion	AUC = 0.88
Alghowinem et al. (2023)	Suicide-risk screening	Audio + behavior	Deep fusion	Up to 1.00 subject-level
Shaikh et al. (2025)	Suicide ideation detection	Text, image, audio	Attention fusion	F1 = 90.5%
Hu et al. (2025)	Depression severity and suicide risk	Audio, text, clinical	Multitask fusion	AUC $>$ 0.80

–block4_pool and block5_pool, which preserve spatially localized mid- and high-level convolutional representations;
–fc1 and fc2, which encode high-level compact semantic representations.

Combinations of the four selected VGG16 layers were evaluated, resulting in 15 distinct feature sets including single-layer, pairwise, three-layer, and full concatenation variants. Each feature set was evaluated using three classifiers with complementary inductive biases:

–Random Forest;
–XGBoost;
–Multilayer Perceptron (MLP).

The results indicated that fully connected layers generally outperformed isolated convolutional layers and that the fc1 + fc2 combination produced consistently strong and stable performance. This feature set was therefore selected as the final visual representation.

2.7 Audio segmentation and chunk extraction

The audio preprocessing pipeline was designed to segment interview recordings into fixed-duration speech units while preserving the hierarchical dataset organization (group – subject – segment – audio chunk). This structured approach is commonly adopted in computational mental health research to enable reproducible analysis of acoustic patterns associated with psychological states, including depression and suicide risk [7,8]. Raw audio streams were segmented into contiguous speech chunks of 15 seconds, providing sufficient temporal context for modeling suprasegmental speech characteristics such as prosody and articulation. To avoid unreliable representations, segments shorter than 4 seconds were excluded from further analysis. This segmentation strategy reflects established practices in speech-based mental health assessment, balancing temporal coverage with computational efficiency and model stability [24,25]. All extracted audio segments were resampled to 16 kHz, converted to mono-channel format, and stored in a standardized waveform representation. The resulting chunks were organized into a directory hierarchy that retained class labels (control/experimental), subject identifiers, and segment indices, facilitating downstream unimodal and multimodal analysis. Corrupted or invalid audio files were automatically skipped during processing, and extraction errors were logged for later inspection. This stage yielded a large corpus of standardized speech segments, forming the foundation for subsequent acoustic feature extraction. In the following section, these audio chunks are further processed to derive Mel-Frequency Cepstral Coefficients (MFCCs), which capture short-term spectral properties relevant to affective and clinical speech analysis.

2.8 Acoustic feature extraction

The selection of acoustic features in this study was guided by the characteristics of the dataset and established practices in speech-based mental health analysis. Given the moderate dataset size and the need for robust and computationally efficient representations, Mel-Frequency Cepstral Coefficients (MFCCs) were adopted as the primary acoustic feature set. MFCCs have been extensively validated in prior research on depression and suicide risk detection and remain a widely accepted standard for speech analysis in small- to medium-scale datasets [24,25,26]. MFCCs represent the short-term spectral envelope of speech using a perceptually motivated frequency scale, capturing vocal attributes such as energy distribution, articulation patterns, and spectral dynamics. These characteristics have been shown to correlate with affective states and indicators of psychological distress, making MFCCs particularly suitable for clinical and paralinguistic speech analysis [7, 24]. In this study, MFCCs were extracted using a 25 ms analysis window with a 10 ms frame shift, resulting in 13 coefficients per frame, following common practice in speech processing literature [25]. To obtain fixed-length representations suitable for machine learning models, frame-level MFCCs were aggregated across time using statistical functionals. Mean and variance normalization was applied to reduce inter-speaker variability and recording-condition effects, which has been shown to improve robustness and generalization in speech-based classification tasks [26,27]. Based on their strong theoretical grounding, empirical effectiveness in prior studies, and suitability for the available dataset, MFCCs were selected as the final acoustic feature representation in this work. This choice enabled stable model training

and ensured methodological consistency with existing research in speech-based suicide risk and depression detection.

Let $s_i(t)$ denote the audio signal associated with chunk i. After framing and windowing, the short-time spectrum of frame r is passed through a bank of Mel filters. If $E_{i,r,m}$ is the energy of Mel filter m, the k-th MFCC coefficient is computed as

$$c_{i,r,k} = \sum_{m=1}^{M} \log(E_{i,r,m}) \cos\left[\frac{\pi k}{M}\left(m - \frac{1}{2}\right)\right], \quad k = 1, \ldots, 13.$$

The frame-level coefficients were summarized over all frames in the chunk to obtain the acoustic vector

$$\mathbf{a}_i = [\bar{c}_{i,1}, \bar{c}_{i,2}, \ldots, \bar{c}_{i,13}]^\top \in \mathbb{R}^{13},$$

where $\bar{c}_{i,k}$ denotes the temporal average of coefficient k for chunk i.

2.9 Dataset organization and multimodal alignment

The extracted features were grouped by participant and segment, producing a structured dictionary where each key corresponded to a subject ID and each value stored an array of frame-level embeddings. Audio and video feature vectors were aligned both subject-wise and segment-wise to produce multimodal entries of equal length across modalities:

- Audio features: 13-dimensional MFCC-based vectors per chunk.
- Video features: 8,192-dimensional embeddings derived from concatenated fc1 and fc2 activations of VGG16.

Each aligned sample was associated with a binary label: control (0) or experimental (1). In total, 1,036 multimodal samples were successfully aligned across 68 unique subjects. Quality checks excluded segments lacking valid video passages or having zero-length face arrays. Audio and video features were standardized separately using z-score normalization to prevent information leakage between sets [27]. The resulting multimodal matrices were concatenated along the feature dimension, yielding

$$X_{\text{train}} \in \mathbb{R}^{829 \times 8205}, \qquad X_{\text{test}} \in \mathbb{R}^{207 \times 8205}.$$

2.10 Mathematical formulation of multimodal fusion

For each aligned chunk i, the acoustic representation is denoted by $\mathbf{a}_i \in \mathbb{R}^{13}$ and the visual representation by $\mathbf{v}_i \in \mathbb{R}^{8192}$. The visual vector is obtained by averaging the VGG16 activations over the face frames associated with the same chunk. If $\mathbf{f}_{i,r}^{(1)}$ and $\mathbf{f}_{i,r}^{(2)}$ are the fc1 and fc2 activations for visual frame r, then

$$\mathbf{v}_i = \frac{1}{R_i} \sum_{r=1}^{R_i} \left[\mathbf{f}_{i,r}^{(1)}, \mathbf{f}_{i,r}^{(2)}\right], \qquad \mathbf{v}_i \in \mathbb{R}^{8192},$$

where R_i is the number of valid face frames in chunk i and $[\cdot;\cdot]$ denotes vector concatenation. Early multimodal fusion was then performed by concatenating the standardized audio and visual vectors:

$$\mathbf{x}_i = [\tilde{\mathbf{a}}_i; \tilde{\mathbf{v}}_i] \in \mathbb{R}^{8205}.$$

The binary label is $y_i \in \{0, 1\}$, where 0 denotes the control class and 1 denotes the experimental at-risk class.

Table 2: Notation used in the proposed multimodal classification pipeline

Symbol	Definition
N	Number of aligned multimodal chunks
S	Number of unique subjects
i	Chunk index
$s_i(t)$	Audio signal of chunk i
$\mathbf{a}_i$	13-dimensional MFCC acoustic feature vector
$\mathbf{v}_i$	8,192-dimensional VGG16 visual feature vector
$\mathbf{x}_i$	8,205-dimensional fused feature vector $[\tilde{\mathbf{a}}_i; \tilde{\mathbf{v}}_i]$
y_i	Binary class label, control = 0, experimental = 1
X_{train}	Training feature matrix
X_{test}	Held-out test feature matrix
θ	Trainable model parameters
$\hat{p}_i$	Predicted probability of the experimental class

2.11 Multimodal classification approach

To evaluate the discriminative capacity of extracted features, a series of classical and neural models were trained using the aligned multimodal (audio + video) dataset. The classification task was binary, distinguishing participants in the experimental group (potentially suicidal or depressed) from those in the control group (healthy). Five models with complementary inductive biases were selected to ensure a balanced comparison between linear, probabilistic, ensemble-based, and neural architectures:

- Logistic Regression (LR): a linear baseline trained using the saga solver with L_2 regularization;
- Gradient Boosting (GB): an ensemble of shallow decision trees with 50 estimators;
- XGBoost (XGB): a regularized gradient-boosted tree model optimized with the logloss objective;
- Naive Bayes (NB): a Gaussian probabilistic baseline;

–Deep Neural Network (DNN): a two-layer feedforward model with 64 and 32 hidden units trained with the Adam optimizer.

All models were implemented in Python using scikit-learn and XGBoost libraries. The neural model was trained using iterative updates, while classical models were fitted under the same subject-independent train-test protocol for comparability. For models supporting warm starts, training progress was monitored over 20 fitting iterations using log loss and accuracy. Early stopping and warm-starting were applied where supported.

For each epoch, model parameters were updated using the current training data, and predictions were evaluated on both training and validation sets. The log loss was computed for models with probabilistic outputs, while accuracy was used as a universal metric across all classifiers. To prevent overfitting, early stopping and warm-starting were applied where supported, ensuring gradual refinement of model parameters without reinitialization between epochs.

For the DNN, the forward propagation is defined as

$$\mathbf{h}_{1,i} = \phi(W_1\mathbf{x}_i + \mathbf{b}_1), \qquad \mathbf{h}_{2,i} = \phi(W_2\mathbf{h}_{1,i} + \mathbf{b}_2),$$

$$\hat{p}_i = \sigma(W_3\mathbf{h}_{2,i} + b_3),$$

where $\phi(\cdot)$ is the ReLU activation and $\sigma(u) = 1/(1+\exp(-u))$ is the sigmoid function. The binary cross-entropy objective is

$$\mathscr{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[y_i\log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)].$$

Because the dataset contains more control than experimental subjects, class imbalance was considered during evaluation by reporting macro-averaged precision, recall, and F1-score, which assign equal weight to both classes. In the subject-wise cross-validation analysis, imbalance-aware linear models used balanced class weights

$$w_c = \frac{N}{KN_c},$$

where N is the number of training samples, $K = 2$ is the number of classes, and N_c is the number of training samples in class c. This weighting increases the contribution of the minority class during optimization without duplicating samples or introducing synthetic data.

The Adam optimizer updates the model parameters θ from gradient $g_t = \nabla_\theta \mathscr{L}_t$ using

$$\mathbf{m}_t = \beta_1\mathbf{m}_{t-1} + (1-\beta_1)g_t, \qquad \mathbf{v}_t = \beta_2\mathbf{v}_{t-1} + (1-\beta_2)g_t^2,$$

$$\hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1-\beta_1^t}, \qquad \hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1-\beta_2^t},$$

$$\theta_t = \theta_{t-1} - \eta\frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \varepsilon}.$$

The final class prediction is obtained as $\hat{y}_i = 1$ if $\hat{p}_i \geq 0.5$ and $\hat{y}_i = 0$ otherwise.

2.12 Evaluation metrics

Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The main evaluation metrics are defined as

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}},$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$

$$F_1 = \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}.$$

For macro-averaged metrics, precision, recall, and F1-score were computed separately for the control and experimental classes and then averaged. Probabilistic calibration during training was monitored using log loss,

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i\log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)].$$

The subject-wise cross-validation analysis also reports ROC-AUC and Precision–Recall AUC (PR-AUC). ROC-AUC summarizes discrimination across all decision thresholds, while PR-AUC is especially informative for imbalanced data because it focuses on precision and recall for the positive class.

3 Results and Discussion

After training, model performance was evaluated on the held-out chunk set from unseen test subjects using accuracy, macro precision, macro recall, and macro F1-score. Ensemble and neural approaches generally outperformed simpler baselines, highlighting the benefit of non-linear modeling for high-dimensional multimodal feature embeddings, as shown in Table 3.

Table 3: Performance comparison of multimodal classification models

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
Logistic Regression	0.74	0.77	0.77	0.74
Gradient Boosting	0.78	0.77	0.78	0.77
XGBoost	0.77	0.76	0.76	0.76
Naive Bayes	0.63	0.61	0.62	0.61
DNN	0.82	0.82	0.84	0.82

The DNN achieved the highest overall chunk-level accuracy (0.82) and macro-averaged F1-score (0.82) on the held-out subjects. Gradient Boosting and XGBoost models also showed competitive results, confirming that ensemble tree methods effectively capture non-linear dependencies within high-dimensional fused embeddings. Logistic Regression, while simpler, exhibited moderate accuracy (0.74), indicating that linearly separable patterns

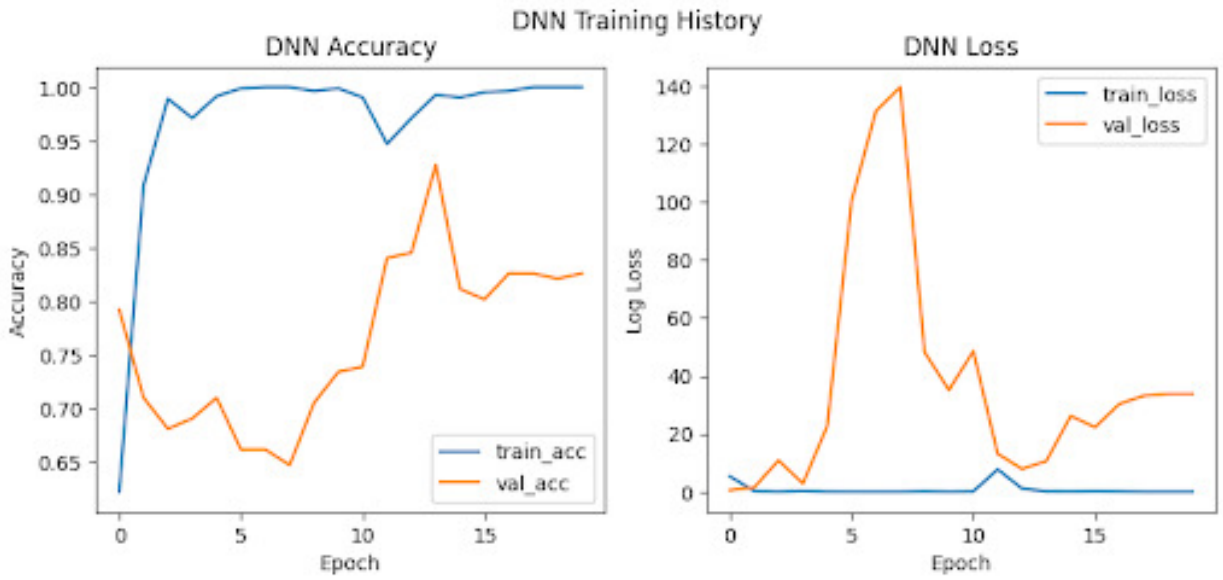

Fig. 1: Accuracy and loss function curves

were partially present. The Naïve Bayes classifier, as expected, underperformed due to its strong independence assumptions and inability to model complex correlations.

A subject-wise five-fold cross-validation analysis was conducted to estimate the stability of the multimodal representation under repeated participant-level validation. This analysis used the same MFCC and VGG16 feature sources but removed chunks with zero-length visual windows, yielding 811 aligned chunks from 68 subjects. Each fold held out complete subjects, and hyperparameters were selected using an inner subject-wise grid search. The resulting cross-validation performance is summarized in Table 4.

Table 4: Subject-wise five-fold cross-validation results

Model	Accuracy	Macro F1	ROC-AUC	PR-AUC
Audio-only LR	0.47 ± 0.12	0.45 ± 0.14	0.41 ± 0.17	0.36 ± 0.10
Video-only LR + PCA	0.66 ± 0.18	0.64 ± 0.19	0.73 ± 0.22	0.71 ± 0.20
Early fusion LR + PCA	0.65 ± 0.17	0.64 ± 0.19	0.73 ± 0.23	0.71 ± 0.20
Early fusion MLP + PCA	0.58 ± 0.12	0.57 ± 0.12	0.62 ± 0.15	0.61 ± 0.16
Late fusion probability average	0.66 ± 0.20	0.65 ± 0.21	0.65 ± 0.30	0.62 ± 0.23

The strongest cross-validation configuration was late fusion by averaging audio and video probabilities, with macro F1 $= 0.65 \pm 0.21$. The corresponding out-of-fold confusion-matrix summary was accuracy $= 0.663$, macro precision $= 0.657$, macro recall $= 0.662$, and macro F1 $= 0.658$ over 811 chunks. Bootstrap confidence intervals further reflected uncertainty due to the modest number of subjects; for the late-fusion model, the 95% interval was 0.63–0.70 for accuracy and 0.63–0.69 for macro F1. A paired comparison between early fusion and the strongest unimodal baseline did not show a statistically significant difference in fold-level macro F1 (Wilcoxon signed-rank test, $p = 0.25$), and the paired McNemar test on out-of-fold correctness was also not significant ($p = 0.388$).

The high dimensionality of the fused vector (8,205 features) relative to the number of unique subjects is an important limitation. The subject-wise split reduces identity leakage, but the reported test metrics remain chunk-level estimates from 14 held-out subjects. Therefore, these results should be interpreted as evidence that multimodal facial and acoustic features contain discriminative signal for suicide-risk classification, rather than as a definitive clinical screening performance estimate. Future validation should include larger cohorts, repeated subject-wise cross-validation, confidence intervals, and external validation on independent clinical recordings. Overall, these findings establish baseline results for multimodal suicide-risk classification and provide a foundation for subsequent experiments with more advanced fusion mechanisms.

4 Conclusion

This study explored the effectiveness of combining low-level acoustic features and deep visual representations for automated suicide-risk classification. Using MFCC descriptors and VGG16-based facial embeddings extracted from structured interview recordings, several unimodal and multimodal models were evaluated. The results indicate that visual information derived from facial expressions carries strong

discriminative value, with the best-performing video-based neural network achieving an accuracy and macro F1-score of 0.82. Audio-only models demonstrated lower but complementary predictive capacity, suggesting that speech characteristics alone may be insufficient for robust suicide detection but remain valuable for future fusion architectures. The findings highlight three key contributions. First, the study provides a reproducible preprocessing and feature-alignment pipeline suitable for small and heterogeneous mental-health datasets. Second, it establishes strong unimodal baselines that can guide the design of more advanced fusion strategies. Third, it demonstrates the potential of deep visual embeddings as a reliable signal for early suicide-risk screening. Future work will focus on expanding the dataset, incorporating additional audio descriptors, and exploring early-, late-, and hierarchical-fusion mechanisms to leverage complementary modalities more effectively. Overall, this study reinforces the promise of multimodal machine learning for improving objectivity and scalability in clinical suicide-risk assessment.

Acknowledgement

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Grant No. AP22786670.

References

[1] World Health Organization, Suicide, 2025, available at: https://www.who.int/news-room/fact-sheets/detail/suicide.

[2] V. Venek, S. Scherer, L.P. Morency and J. Pestian, Adolescent suicidal risk assessment in clinician-patient interaction, IEEE Transactions on Affective Computing **8**, 204–215 (2017).

[3] M. Lotito and E. Cook, A review of suicide risk assessment instruments and approaches, Mental Health Clinician **5**, 216–223 (2015).

[4] S. Ji, S. Pan, X. Li, E. Cambria, G. Long and Z. Huang, Suicidal ideation detection: A review of machine learning methods and applications, IEEE Transactions on Computational Social Systems **8**, 214–226 (2020).

[5] H.C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daume III and P. Resnik, Expert, crowdsourced, and machine assessment of suicide risk via online postings, in Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology, 25–36 (2018).

[6] S. Ji, C.P. Yu, S.F. Fung, S. Pan and G. Long, Supervised learning for suicidal ideation detection in online user content, Complexity **2018**, 6157249 (2018).

[7] S. Min, D. Shin, S.J. Rhee, C.H.K. Park, J.H. Yang, Y. Song and H. Lee, Acoustic analysis of speech for screening for suicide risk: machine learning classifiers for between- and within-person evaluation of suicidality, Journal of Medical Internet Research **25**, e45456 (2023).

[8] H. Wang, J. Li, Q. Zhao, Z. Chen, C. Song, J. Tang and G. Fu, Deep learning-based feature fusion for emotion analysis and suicide risk differentiation in Chinese psychological support hotlines, arXiv preprint arXiv:2501.08696 (2025).

[9] X. Kong, Y. Yao, C. Wang, Y. Wang, J. Teng and X. Qi, Automatic identification of depression using facial images with deep convolutional neural network, Medical Science Monitor **28**, e936409-1 (2022).

[10] Z. Zhang, S. Zhang, D. Ni, Z. Wei, K. Yang, S. Jin and J. Wang, Multimodal sensing for depression risk detection: Integrating audio, video, and text data, Sensors **24**, 3714 (2024).

[11] M. Nykoniuk, O. Basystiuk, N. Shakhovska and N. Melnykova, Multimodal data fusion for depression detection approach, Computation **13**, 9 (2025).

[12] A. Pigoni, I. Tesic, C. Pini, P. Enrico, L. Di Consoli, F. Siri and P. Brambilla, Multimodal machine learning prediction of 12-month suicide attempts in bipolar disorder, Bipolar Disorders (2025).

[13] S. Alghowinem, X. Zhang, C. Breazeal and H.W. Park, Multimodal region-based behavioral modeling for suicide risk screening, Frontiers in Computer Science **5**, 990426 (2023).

[14] A. Shaikh, P. Jangid, M. Rana, S. Sall, R. Rana, D.P. Mhapasekar and A. Jangid, Enhancing suicide ideation detection with multimodal deep learning: ethical, explainable, and privacy-preserving framework, International Journal of Applied Mathematics **38**, 4s (2025).

[15] Y.H. Hu, R.Y. Wu, M.Y. Su, I.L. Lin and C.C. Shen, Multimodal multitask learning for predicting depression severity and suicide risk using pretrained audio and text embeddings: Methodology development and application, JMIR Medical Informatics **13**, e66907 (2025).

[16] G. Jackson-Koku, Beck depression inventory, Occupational Medicine **66**, 174–175 (2016).

[17] L.N. Sobchik, Standardized multifactorial personality inventory SMIL (MMPI): Practical manual, Rech, Saint Petersburg, 219 (2003).

[18] G. Bradski and A. Kaehler, OpenCV, Dr. Dobb's Journal of Software Tools **3**, 2 (2000).

[19] J. Deng, J. Guo, E. Ververas, I. Kotsia and S. Zafeiriou, RetinaFace: Single-shot multi-level face localisation in the wild, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5203–5212 (2020).

[20] S. Tammina, Transfer learning using VGG-16 with deep convolutional neural network for classifying images, International Journal of Scientific and Research Publications **9**, 143–150 (2019).

[21] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[22] M.D. Zeiler and R. Fergus, Visualizing and understanding convolutional networks, in European Conference on Computer Vision, 818–833 (2014).

[23] S. Mascarenhas and M. Agarwal, A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for image classification, in 2021 International Conference on Disruptive Technologies for Multi-disciplinary Research and Applications, Vol. 1, 96–99 (2021).

[24] F. Zheng, G. Zhang and Z. Song, Comparison of different implementations of MFCC, Journal of Computer Science and Technology **16**, 582–589 (2001).
[25] R. Iyer and D. Meyer, Detection of suicide risk using vocal characteristics: Systematic review, JMIR Biomedical Engineering **7**, e42386 (2022).
[26] P. Pujol, D. Macho and C. Nadeu, On real-time mean-and-variance normalization of speech recognition features, in 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, I–I (2006).
[27] N. Fei, Y. Gao, Z. Lu and T. Xiang, Z-score normalization, hubness, and few-shot learning, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 142–151 (2021).

A. Mukhametzhanov received his Bachelor's degree in Information Systems and is currently pursuing a Master's degree in Data Engineering at SDU University, Kaskelen, Kazakhstan. His technical expertise and research interests encompass natural language processing, data analysis, web scraping, backend development, and frontend web application design. As a Junior Research Fellow, he is responsible for data collection, execution of ongoing project tasks, and maintenance of project documentation.

Aisultan Shoiynbek holds a PhD, a Master's, and a Bachelor's degree in Computational Engineering and Software. As Project Leader and Leading Research Fellow, he coordinates all management and organizational activities for the project. His research spans AI-based deception detection, speech emotion recognition for Kazakh and Russian, automatic speech recognition systems for low-resource languages using connectionist temporal classifiers, robust spectral audio feature extraction, and optimization algorithms for neural network models.

Nurzhas Abdikenov received his Bachelor's degree in Big Data Analysis from Astana IT University and is currently pursuing a Master's degree in Computer Science and Engineering at Astana IT University, Astana, Kazakhstan. As a Junior Research Fellow of the project, he is responsible for collecting and structuring multimodal data, conducting preliminary data preprocessing, and transferring processed materials to the analytical group for model development and training. His research interests include data analysis, dataset preparation, multimodal data organization, and experimental reporting.

Bakhtiyor Meraliyev is a PhD candidate in Computer Science at SDU University (2024–2027) and holds a Master's degree in Computer Science. His research interests include ML, data science, NLP, predictive analytics, and intelligent data analysis. As a Senior Research Fellow, he is responsible for developing algorithms and executing the project's current tasks. He has authored publications in leading international conferences—covering topics such as HR analytics, computer vision in educational contexts, and NLP-based social media analysis.

Darkhan Kuanyshbay received his PhD degree in Computer Science and also holds a Master's degree in Information Systems as well as a Bachelor's degree in Computer Systems Processing and Management. As Co-Project Leader and Chief Research Fellow, he is responsible for defining and executing the conceptual framework of the project. His research focuses on speech data collection systems, automatic speech recognition for the Kazakh language—using connectionist temporal classifiers and transfer learning—voice identification techniques, emotion recognition in speech, and optimization algorithms for deep neural network models in speech processing.

Sergey Sklyar received the Candidate of Medical Sciences degree in Psychiatry from the Republican Scientific and Practical Center for Mental Health (2007–2010). His research interests encompass suicide-prevention methodologies, psychotherapeutic and psycho-pedagogical approaches across the lifespan, data analysis and ethical oversight in mental-health interventions, digital cognitive training for children with ADHD, and the assessment of quality of life in older adults. He has published numerous research articles in reputed international journals and conference proceedings in psychiatry and psychology, and serves as a reviewer and editor for scientific journals.

Paulo Menezes received his PhD in Electrical and Computer Engineering (Informatics) from the University of Coimbra, following an MEng in Systems and Automation and a BEng in Electrical Engineering (Informatics) at the same institution. Since July 2024, he has been an Associate Professor in the Department of Electrical and Computer Engineering at the University of Coimbra's Faculty of Sciences and Technology, teaching Computer Architecture, Computer Graphics and Augmented Reality, and Interactive Systems and Robotics. As a senior researcher at the Institute of Systems and Robotics and head of its Immersive Systems and Sensory Stimulation Laboratory, he leads work on human–robot interaction, affective computing, AR/VR applications for therapy and rehabilitation, and serious games for emotional learning; he also serves as WP lead for Robotic Social Interaction in the H2020 RISE Lifebots project, advises multiple PhD candidates, has published extensively in top journals and conferences, contributed to major European research initiatives, and acts as a peer reviewer and editorial board member for leading robotics and AI journals.