Mindy - The Mindfulness Coach

Rohit Joseph Mamutil, Sheila Mallavarapu, Zarin Tasnim Biash
Department of Information Technology
Uppsala University

***Abstract*—This project introduces Mindy, a virtual mindfulness coach designed to support emotional well-being through real-time emotion detection and personalized interactions. Built on the Furhat robot platform, Mindy combines emotion recognition, guided breathing exercises, and natural conversation to create an engaging and supportive user experience. The system includes an emotion detection module that analyzes facial features and mood to tailor its responses, and an interaction module powered by a Gemini-based language model for dynamic conversations. User testing demonstrated positive engagement, with participants appreciating the seamless integration of mindfulness exercises and empathetic interaction. While Mindy shows great promise as an emotion-aware coach, areas such as detecting subtle expressions, offline functionality, and conversational fluidity remain opportunities for future development.**

I. Introduction

This project aims to create a system that analyzes human behaviour and affective states and generates in response a socially adaptive behaviour by the Furhat virtual robot. This system consists mainly three modules:

1) **Emotion Detection Module:** This module explores use of different models and also a hybrid model which was deployed in the project.
2) **Language Processing Module:** This module explores usage of a Large Language Model(LLM) and the challenges involved in it using it for a social robot.
3) **Furhat Control Module:** This module includes gesture control, gaze management and speech generation. This also proposes a formula to calculate parameters to control the robot's gaze.

To act as a midfullness coach, Mindy the furhat was given capabilities to do breathing exercises, and also hold conversations with the user.

II. Team Contributions

Contributions from team members include system architecture design, implementation of emotion detection, and development of conversational flow.

- Zarin focused on the User Perception Sub-System, identifying the optimal approach for emotion detection. She developed a hybrid emotion detection model that integrated three distinct models, working together to provide the most accurate emotion analysis.
- Rohit worked on the Interaction Sub-System, primarily integrating the emotion detection module with Furhat, configuring LLM and system design.
- Sheila concentrated on designing conversational flows, initially using fixed phrases in Python and later employing prompt engineering with an LLM to refine interactions.

All members collaborated on the report, video, presentation, system integration, and conducting a feedback study with voluntary participants. The report discusses key challenges and opportunities in building such an integrated system.

III. Methodology

This project was experimental in nature, focusing on exploring the challenges involved in developing a social robot using Furhat. The goal was to investigate technical and user-interaction challenges based on emotion and identify areas for improvement in the robot's social capabilities.

An iterative development process was employed, enabling continuous refinement based on feedback and observations. The survey included questions about the robot's expressiveness, responsiveness, and ease of interaction. The responses were collected through an online form and analyzed using descriptive statistics to identify trends and areas of improvement.

To gain insights into user perceptions, a survey was conducted with 15 participants aged 22–30.

IV. Overall System Design

The overall system design prioritizes simplicity to facilitate quick development. The Furhat SDK was not utilized; instead, the system relied entirely on remote APIs. The language model (LLM) used was cloud-based, rather than being hosted on a local machine or server. Below is a brief introduction to the classes involved in the program. Threading was used to run the emotion detection in parallel.

A. Classes

1) Furhat Client: Includes function wrappers to call the Furhat RemoteAPI. We have used all the functions offered except visibility API. Additionally, breathing gestures were also added here where each gesture involved a combination of different basic Furhat gesture.

2) Breathing Guide: Has the breathing exercise function, which can be called directly by the LLM with arguments for the time for inhaling, holding the breath, and the number of repetitions of the pattern.

3) Emotion Detection Module: Has functions for detecting the emotion of the user from the webcam when called. The coordinates the Furhat robot needs to gaze at to maintain eye contact are determined in this.

Fig. 1. Model 1 of the User-Perception Subsystem

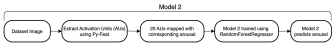

Fig. 2. Model 2 of the User-Perception Subsystem

4) Large Language Model (LLM): The gemini-2.0-flash-exp large language model is configured in this class, along with the predefined responses for better control of the flow of the system.

5) Mood Interpretation: The mood of a user is detected by calculating the exponential moving average (EMA).

V. User Perception Sub-system

A. Design

The User Perception subsystem is designed to detect human emotions from images or live webcam feeds. It leverages facial Action Units (AUs) extracted using the py-feat library to classify emotions into predefined categories such as neutral, happy, sad, surprise, fear, disgust, and anger. The system incorporates human-annotated valence and arousal values as additional features to enhance emotion classification accuracy. The goal is to bridge the gap between dataset-based training and real-world applicability by providing an emotion detection system capable of handling diverse facial inputs, including live feeds. The system performs three key tasks:

- **Feature Extraction:** Extracts AUs from facial images and predicts valence and arousal values.
- **Model Training:** Trains machine learning models using extracted features and annotated labels.
- **Emotion Prediction:** Predicts emotions from live webcam feeds using the trained models.

The process is structured around three core models, as illustrated below:

- **Model 1:** Trained to predict valence values from the extracted AUs using a RandomForestRegressor.
- **Model 2:** Trained to predict arousal values from the extracted AUs using a RandomForestRegressor.
- **Model 3:** Combines the 20 AUs, predicted valence, and predicted arousal values as inputs to a Random Forest Classifier, which outputs the final emotion prediction.

The overall flow integrates these models seamlessly, starting from the extraction of AUs from input images or webcam frames (converted to images), predicting valence and arousal using Models 1 and 2, and then classifying the emotion using Model 3.

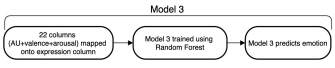

Fig. 3. Model 3 of the User-Perception Subsystem

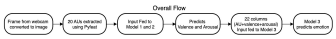

Fig. 4. Overall Flow of the User-Perception Subsystem

B. Implementation

The implementation began with the exploration of the DiffusionFER dataset. This dataset consisted of folders named after emotions, each containing images associated with the folder's label. Additionally, a CSV file provided annotations for each image, including human-labeled emotions and corresponding valence and arousal values. The first step involved extracting AUs from the images using the py-feat library. Initially, all available AUs were used as features to train various machine learning models, and the folder labels were used as ground truth for the emotions. At this stage, we trained several models, including K-Nearest Neighbors (KNN) , Radius Neighbors, Support Vector Classifier (SVC), Decision Tree, Random Forest, XGBoost, Multi-Layer Perceptron (MLP) Classifier, and Naive Bayes [5]. The models achieved moderate accuracy, ranging from 50% to 60%. Next, we realised a notable challenge: for some images, the annotated emotion labels in the CSV differed from the folder names. This discrepancy prompted the decision to prioritize the annotations for model training, as they offered more precise human judgment. We transitioned to using the annotated emotion labels from the CSV file for model training. This shift, combined with hyperparameter tuning using GridSearchCV, significantly improved the system's performance. Weighted f1-score was used as the evaluation metric during tuning to ensure fair consideration of underrepresented classes. With these improvements, the accuracy increased to around 65Realizing the potential of valence and arousal values in better representing emotional dimensions, we expanded the approach to predict these values from the AUs. Two separate RandomForestRegressors (Model 1 and Model 2) were trained for this purpose. Model 1 was designed to predict valence, and Model 2 was designed to predict arousal. The system was then restructured to use the predicted valence and arousal values, along with the original AUs, as features for emotion classification. This multi-stage pipeline, with Model 3 as the final emotion classifier, significantly boosted the accuracy to approximately 83-84%. Among the models tested in this approach, Random Forest and XGBoost emerged as the best-performing models, both in terms of accuracy and weighted f1-score. Ultimately, Random Forest was selected as the final model for its slightly better generalization and interpretability. To integrate the system into a live webcam setup, the trained models were utilized to predict emotions from real-time video feeds. Frames captured from the webcam were processed to extract AUs, which were then used to predict valence and arousal via Models 1 and 2. These values, combined with the AUs, were fed into Model 3 to detect emotions in real time. While the system performed well for neutral, happy, and surprise emotions, it struggled with emotions like sadness, fear, disgust, and anger. This limitation was attributed to the dataset's nature, as it primarily contained

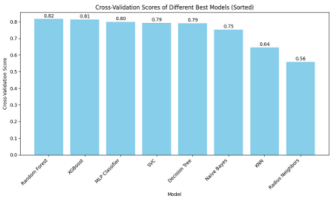

Fig. 5. Cross-Validation Scores of Different Best Models

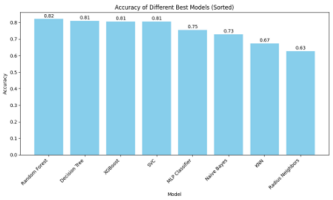

Fig. 6. Accuracy of Different Best Models

cartoon-like faces, making it difficult for the models to generalize to real-world subtle cues. Subsequently, we explored another dataset with greyscale human faces using the FER-2013 dataset [2]. However, this dataset lacked annotations and yielded lower accuracy due to its limited compatibility with the existing pipeline.

Finally, we experimented with another alternate approach where we trained models using only valence and arousal as features, excluding AUs entirely. In simpler words, the AUs extracted from the images were fed as inputs to models 1 and 2 to predict the valence and arousal. Then this valence and arousal (ignoring the 20 AUs this time) were fed as input to model 3 to predict emotion. While this approach slightly improved accuracy to 86%, its real-world performance on the webcam feed was poorer, leading to the decision to retain the original multi-stage pipeline that combined 20 AUs with predicted valence and arousal values to predict the emotion.

C. Results

The results demonstrated significant improvements over time as the system evolved. Initially, using only AUs and folder-based labels, the accuracy was limited to 50-60%. This improved to 65% when annotated labels and hyperparameter tuning were introduced. The integration of valence and arousal as predicted features marked the most substantial advancement, achieving an accuracy of 83-84%. This approach provided a more comprehensive understanding of emotions by incorporating dimensional information. Among the models evaluated, Random Forest and XGBoost consistently delivered the best performance, with Random Forest slightly edging out XGBoost in both accuracy and weighted f1-score. This solidified Random Forest as the preferred model for the final implementation.

However, real-world testing with live webcam feeds highlighted some challenges. The system was highly effective at detecting neutral, happy, and surprise emotions, likely due to the distinct visual cues associated with these expressions. Conversely, it struggled with emotions like sad, fear, disgust, and anger, which often share subtle facial cues that are harder to distinguish. Moreover, the dataset images were animated, which made it harder for the model to work well on human faces.

The experiments with using only valence and arousal as features (excluding 20 AUs) when fed to our model 3 showed promise in terms of accuracy, reaching 86%. However, the live testing revealed its limitations, as the absence of detailed AUs reduced its robustness in handling diverse expressions. This outcome reinforced the importance of combining AUs with valence and arousal to achieve the best balance of accuracy and real-world applicability.

In conclusion, the User Perception subsystem performs well in controlled environments and provides a reliable foundation for emotion detection. While challenges remain in handling subtle or ambiguous emotions, the multi-stage pipeline integrating AUs, valence, and arousal offers a robust and scalable solution for real-time emotion detection.

VI. Interaction Sub-system

A. Design

The Interaction Sub-System facilitates communication between the user and the Furhat robot. The goal of this subsystem is to factor in user's emotion and to create an adaptive interaction environment where the user can communicate with the system while receiving personalized mindfulness coaching.

The interaction sub-system is designed with the following high-level objectives:

- **Emotional Awareness**: Processing the detected user's emotion from the Emotion Detection module and use it in creating respones.
- **Natural Conversation Flow**: Ensuring a conversational experience that feels natural and dynamic. This includes guiding the user through mindfulness exercises, providing feedback, and adjusting the system's responses based on the user's emotional state.
- **User Engagement**: Using the Furhat robot's physical presence (gaze, gestures, and speech) to maintain user engagement and create an empathetic environment for emotional support.

B. Implementation

The interaction subsystem included various components that were essential for facilitating communication between the user and the system.

1) Selection of the Large Language Model: The gemini-2.0-flash-exp large language model has been selected for its focus on speed and efficiency, and advances to function calling. The Gemini based model is also free for non-commercial use upto a certain limit.

2) Prompt Engineering and Rule-Based Execution: The system combines predefined rule-based responses with LLM functionality to provide a seamless user experience. The LLM is instructed to emulate a Furhat robot, engaging users compassionately and maintaining an empathetic tone. It offers

breathing exercises for negative emotions, mindfulness tips or jokes for positive emotions, and predefined responses for common inputs like greetings and exercise requests. When inputs don't match predefined rules, the system appends detected emotions to the input for LLM processing. Gestures triggered by user emotions further enhance communication, ensuring that rule-based actions complement the LLM's context-aware interactions.

3) Emotion Integration with LLM: The EmotionDetectionModule continuously detects emotions via the camera and appends the detected state to each message sent to the LLM. This enables context-aware responses, such as suggesting breathing exercises for negative emotions or cheerful replies for positive emotions, ensuring personalized support.

4) Mood Interpretation: Mood interpretation in our system utilizes an Exponential Moving Average (EMA)-based approach to analyze real-time emotional fluctuations. Detected emotions are assigned a weighted value, updated incrementally using a smoothing factor ($\alpha = 0.2$). The system continuously updates these weights, emphasizing recent emotions while retaining historical context.

The EMA is computed using the following formula:

$$EMA_t = \alpha \cdot E_t + (1 - \alpha) \cdot EMA_{t-1}$$

Where:

EMA_t: Current Exponential Moving Average at time t,

α: Smoothing factor, where $0 < \alpha \leq 1$,

E_t: Current detected emotion's value,

EMA_{t-1}: Previous EMA value.

5) Selected Breathing Patterns: Several evidence-based breathing techniques were implemented in the mindfulness robot to address stress and anxiety:

- **4-7-8 Breathing:** Inhale for 4 seconds, hold for 7 seconds, and exhale for 8 seconds. This pattern reduces stress and induces relaxation [1].
- **Resonance Breathing:** Inhale and exhale for 5 seconds each, promoting calmness and heart rate synchronization [4].
- **Pursed-Lip Breathing:** Exhale through pursed lips for relaxation and improved oxygenation [7].
- **3-6-9 Breathing:** Inhale for 3 seconds, hold for 6 seconds, and exhale for 9 seconds. This technique quickly alleviates stress [3].
- **Humming Bee Breathing:** Exhale while producing a humming sound to enhance relaxation and lower blood pressure [6].

6) Gestures for breathing exercises: The breathing exercises were implemented using keyframe-based animations provided smooth transitions, while parameters like NECK_TILT and BROW_UP created realistic expressions to guide users through inhalation, holding, and exhalation phases.

7) Eye contact: Based on feedback form users, it was noted that user engagement is enhanced when robot could face the user. Our system calculates gaze direction based on the position of detected objects within a given frame. The gaze offset is computed using the following equations:

$$G_x = G_f \cdot \frac{\left(\frac{F_w}{2} - S_x\right)}{F_w}$$

$$G_y = G_f \cdot \frac{\left(\frac{F_h}{2} - S_y\right)}{F_h}$$

Where:

- F_w: Width of the frame or image.
- F_h: Height of the frame or image.
- S_x: X-coordinate of the detected object's bounding box center.
- S_y: Y-coordinate of the detected object's bounding box center.
- G_f: Scaling factor for gaze intensity or sensitivity.
- G_x: Gaze offset in the horizontal direction (left or right).
- G_y: Gaze offset in the vertical direction (up or down).

8) Lighting Cues: Lighting cues were added for the users to know when to speak and listen. Soft amber is used when the robot is talking, and a gentle green for when the user can speak. During the guided breathing exercises, different colors were added during each stage of breathing.

C. Results

1) Evaluation Metrics: The Interaction Sub-System was evaluated based on the following criteria:

a) Engagement Duration: The average engagement duration among the 15 participants was 4.25 minutes. If the prompt was tuned to be more relaxed to include generic topics, the engagement duration could have been longer.

b) Emotion Recognition Accuracy: On average, participants rated the accuracy of the emotion recognition system as 4/5. This rating is relatively higher than the system's actual performance, likely due to users' limited awareness of the changes in their emotional state occurring in the background.

c) Response Time: The response generation process involved multiple sequential steps: detecting user silence after speech, processing the input to text and generate a response via a language model (LLM) on the cloud, and transmitting the output to the Furhat robot for verbal articulation. This multi-step pipeline introduced a delay in response time compared to natural human communication. The response time, defined as the duration from when the user finishes speaking to when the robot begins speaking, is approximately 4.45 seconds.

d) Adaptability to User Emotions: EMA calculation on the detected emotions gave a stable metric. The system was able to execute light conversations when the detected mood was of positive valance and suggested breathing exercises when it detected a mood of negative valance.

2) Strengths:

a) Emotion Awareness: Mindy's ability to recognize and respond to emotions was rated highly with users.

b) Engagement Aids: Visual cues like adaptive lighting and gestures (smiling, nodding, maintaining gaze) contributed to high levels of engagement.

c) Ease of Use: Participants seemed to interact without extensive instructions.

3) Limitations:

a) Internet Dependence: A stable internet connection is required for the system to function, limiting its usability in offline environments.

b) Speech Interruption: When Mindy is speaking, it does not allow interruptions. Users must wait for it to finish before responding. This can lead to reduced conversational fluidity.

4) Suggestions for Improvement:

a) Offline Mode: Integrate offline capabilities to ensure broader applicability in environments with limited connectivity.

b) Interrupt Handling: Implement real-time interruption detection to allow users to interrupt Mindy mid-sentence, making the interactions feel more natural.

5) Overall Evaluation: Mindy demonstrates strong potential as an emotion-aware mindfulness coach. The system effectively detects and responds to user emotions, maintaining engagement through intuitive visual cues and diverse interaction options. However, addressing the limitations of internet dependence and speech interruption handling would further enhance its practicality and usability.

VII. General Discussion

A. Discussion of the Overall Pipeline

The User Perception Sub-System begins by analyzing webcam input to detect faces using OpenCV. Upon detecting a face, the system seamlessly integrates three models to process the data. First, facial Action Units (AUs) are extracted from the input images or webcam frames, which are converted into image format. Models 1 and 2 then predict valence and arousal levels based on the extracted features. Finally, Model 3 classifies the detected emotion. To ensure reliability, a stability check is performed on the detected emotion using an exponential moving average. This approach helps determine the mood by accounting for temporal consistency in emotion detection.

Once the emotion is detected, it is sent to the interaction subsystem, which along with listening to the audio through the microphone and generating responses using a large language model (LLM) responds to the users. This makes it seem like intelligence is being given to the virtual robot.

The user perception subsystem and the interaction subsystem run in separate threads and interact through shared memory between the two subsystems.

B. Challenges Faced

1) Rule-based subsystem vs large language model: Initially, a rule-based approach controlled conversation flow for mindfulness tips, guided breathing, jokes, and stopping sequences. However, minor user response variations (e.g., tense changes or synonyms) disrupted the flow, prompting the shift to a large language model (LLM) with predefined response triggers.

2) Robot Interaction and Conversation Flow: The robot performed actions like synchronized breathing but struggled with conversation dominance, failing to listen while speaking. A listening thread was implemented but led to self-triggered responses. Audio processing with echo cancellation was explored but deemed out of scope for this project.

3) Technical Limitations in Transitioning Between Systems: To gain control back from LLM was a challenge, but adding conditions before passing to LLM and also function calling feature in Gemini helped mitigate it to an extent.

4) Emotion Detection Challenges: Dataset limitations significantly impacted the performance, particularly with the DiffusionFER dataset. Since the dataset primarily consisted of animated images, the models struggled to handle subtle or ambiguous emotions for real human faces, leading to reduced accuracy in real-life emotion detection.

5) Privacy and Security Concerns: User data was not stored at any point, effectively addressing data privacy concerns. However, the use of an online LLM model could still pose potential privacy risks.

C. Use of ChatGPT or Similar Tools

ChatGPT was used to debug and troubleshoot code, providing insights and solutions to errors and inefficiencies and paraphrasing.

VIII. Ethical Issues

Initially designed as a robot receptionist at a mindspace facility, the system was redefined as a mindfulness coach following feedback sessions to address ethical concerns. It prioritizes user privacy by anonymizing questionnaire responses with unique identifiers and processing raw video data locally without storage. Ethical concerns include user overreliance, biases in emotion detection across demographics, and handling critical situations like suicidal tendencies by encouraging professional help. Transparency is ensured by explicitly identifying the system as AI, emphasizing its role as a coach rather than a medical professional. Efforts focus on mitigating misuse risks, such as surveillance, and improving fairness through diversified training datasets.

IX. Conclusion

This project successfully developed Mindy, an emotion-aware mindfulness coach capable of providing personalized support through adaptive interactions. By integrating emotion detection with a language model, the system created a dynamic and empathetic user experience, offering guided mindfulness exercises and engaging conversations. User feedback underscored the strengths of Mindy's design, particularly its ability to adjust to emotional states and maintain user engagement. However, challenges like internet dependence and occasional rigidity in conversation highlight opportunities for improvement. Despite these limitations, Mindy represents an important step toward integrating socially aware robotics with mental health support. Future work will aim to enhance adaptability, ensure broader accessibility, and refine user interaction for a more inclusive and effective experience.

References

[1] G. K. Aktaş and V. E. İlgin. The effect of deep breathing exercise and 4-7-8 breathing techniques applied to patients after bariatric surgery on anxiety and quality of life. *Obesity Surgery*, 33(3):920–929, 2023.

[2] A. Ananthu. FER-2013 Emotion Detection Dataset. Kaggle, 2025. Accessed: 2025-01-16.

[3] National Head Start Association. Exercise 9: Breathing techniques. Accessed: January 17, 2025.

[4] S. Chaitanya, A. Datta, B. Bhandari, and V. K. Sharma. Effect of resonance breathing on heart rate variability and cognitive functions in young adults: A randomised controlled study. *Cureus*, 14(2):e22187, 2022.

[5] Hivi Ismat Dino and Maiwan Bahjat Abdulrazzaq. Facial expression classification based on svm, knn and mlp classifiers. In *2019 International Conference on Advanced Science and Engineering (ICOASE)*, pages 70–75, 2019.

[6] N. Ghati, A. Killa, G. Sharma, B. Karunakaran, A. Agarwal, S. Mohanty, L. Nivethitha, D. Siddharthan, and R. M. Pandey. A randomised trial of the immediate effect of bee-humming breathing exercise on blood pressure and heart rate variability in patients with essential hypertension. *Explore*, 17, 2020.

[7] Berkeley University Health Services. Breathing exercises. Accessed: January 17, 2025.

Appendix

A. Project Specification Document

The detailed project specification document can be found at the following link: https://drive.google.com/file/d/18-QhI7-JqRxNplpDNZi6p8aTk_DSMgka/view?usp=sharing

B. Project Video

You can view the project video here: https://drive.google.com/file/d/17Mh-KXy-46fAe10sU5BtQmKU8k8B6ft-/view?usp=sharing

C. GitHub Repository

The project's source code and related files are available on GitHub: https://github.com/git-sheila/InteractiveIntelligentSystems-Assignments

D. Project Blog

A blog about the challenges in building a social robot can be found at: https://medium.com/@rjmrohit94/challenges-in-building-a-social-robot-ec94ec6fae29

E. User Experience Survey Results

The user experience survey results can be accessed here: https://drive.google.com/file/d/1ldyn5Tx8CYvt2MQBOlGMSgf-RgKF4KIQ/view?usp=sharing

F. Project Resources

Project-related resources, including the specification document, project video, GitHub repository, blog, and survey results, can be accessed in the following Google Drive folder: https://drive.google.com/drive/folders/1FWhxCZZnvO0nB0uxhL_b4q2Q2GaN8nD4?usp=sharing