university of groningen

faculty of science and engineering

Detection of Confusion in Facial Expressions of Older Adults during Conversation with a Social Robot

Yara Bikowski

university of groningen / faculty of science and engineering

University of Groningen

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Artificial Intelligence
at University of Groningen under the supervision of
dr. Paul Vogt and dr. Matias Valdenegro Toro

Yara Bikowski (s3989585)

June 1, 2025

Contents

Acknowledgments

I sincerely thank my supervisors, Dr. Paul Vogt and Dr. Matias Valdenegro Toro, for their guidance and insightful feedback throughout this thesis. Their expertise and support were greatly appreciated during my research.
I also want to give special thanks to the participants who took part in the data collection process. Your involvement was essential to my ability to conduct my research.
Finally, I am deeply thankful to my family for their patience, encouragement, and unwavering belief in me throughout this academic journey.

Abstract

Social robots are increasingly used in elderly care. However, misunderstandings and confusion often occur during human-robot conversations. This study aims to recognize confusion during conversations between older adults and social robots with a focus on facial expressions. We collected data by having participants play a word game with the social robot. Facial action units were extracted from the data and an LSTM model was trained to recognize confusion. A second model was trained by using transfer learning on the ResEmoteNet model. Neither of these models were able to accurately differentiate confusion and non-confusion when tested on data from new participants. The LSTM model obtained an accuracy of 57% while the ResEmoteNet model obtained an accuracy of 53% on balanced data. This indicates that they were unable to generalize to new faces. These findings suggest that models trained on our newly collected dataset are not able to recognize confusion solely based on the facial expressions of older adults. However, earlier research has obtained higher accuracy for this task with datasets of younger adults. This suggests that our inability to detect confusion is a limitation of the new dataset rather than the task itself. Future work should explore multimodal confusion recognition using speech and gesture data as well as facial expressions.

1 Introduction

The global population is aging at an exceptionally high rate, with the proportion of adults aged 60 or older expected to be more than 20% of the total population by 2050 [1]. This leads to significant challenges worldwide. More specifically in providing adequate care for these older adults who often face physical, cognitive, and social impairments. In recent years, social robots have emerged as a promising technology to augment traditional care methods. Social robots are designed to interact with humans in a natural way, enabling them to not only assist in physical tasks but also support cognitive functions, provide companionship, and promote well-being [2]. These robots have been deployed for numerous tasks, from giving reminders to take medication to offering cognitive stimulation and reducing loneliness [3, 4]. The addition of artificial intelligence and the human-centered design of these robots has enhanced their ability to adapt to individual needs, making them increasingly feasible as tools to support the aging population.

To achieve this, human-robot interaction (HRI) has emerged as a crucial area of research. Now that social robots are utilized more often, communication has become more significant in the field of human-robot interaction [5]. Social robots are made to communicate naturally and intuitively with humans. This requires a sophisticated communication system that incorporates both verbal and non-verbal cues, enabling the robot to interpret human intentions, emotions, and social behaviors accurately. More effective communication during the HRI can help promote trust, user engagement, and usability for less experienced humans. Research in social robotics further highlights that users are more likely to accept and collaborate with a robot that demonstrates a considerable understanding of human behaviors and emotions [6]. In an elderly care setting, this is beneficial as the elderly need to accept the assistance of the robot and their caregivers need to collaborate with the robot in caring for the elderly. By being able to recognize and respond to emotional cues, a social robot can provide companionship, reduce loneliness, and improve the overall well-being of aging adults.

However, social robots' current abilities are not always adequate to ensure correct communication. This can result in a communication breakdown between the user and the robot. These breakdowns can result from the actions of both the human and the robot during the interaction. Communication breakdowns can result from a variety of factors including misinterpretations, timing issues, or technical limitations. Examples of this were found in research by Arend and Sunnen [7], who showed that humans had different turn-taking systems than the NAO robot that they used. It is important to detect and solve these communication breakdowns to prevent the robot's conversation partner from getting frustrated and not wanting to interact with the social robot anymore. It is also important to know when someone is confused to ensure that they can understand all the information they need to know.

Kopp and Krämer [8] emphasized the importance of joint co-construction and mentalizing in human-robot communication. They believe that incremental joint co-construction in the form of a collaborative, step-by-step process to build a shared activity of understanding during communication is important.
During human conversations, individuals often provide real-time feedback and adjust their speech based on their conversation partner's responses. So, conversational agents must be capable of this dynamic co-construction to allow for more fluid, adaptive and thus natural interactions.
The second point that Kopp and Krämer mention is the ability to mentalize. This is the ability to understand and infer the mental states, such as beliefs, desires, and intentions, of a communication

partner. In human-to-human interactions, we mentalize to anticipate responses and tailor our messages while maintaining coherence during dialogue. It also allows us to reason about what went wrong during communication when a breakdown occurs. Kopp and Krämer argue that a robot that wants to cooperatively communicate with a human needs to either possess or simulate the ability to mentalize. This then allows them to interpret user inputs more effectively and respond in a contextually relevant manner. By integrating these key capabilities, human-robot communication can become more natural and coherent, closely mirroring the dynamics of human-to-human communication.
One method that can help a robot identify mental states is emotion recognition. Emotions are closely linked to a person's mental state [9] and can thus provide insight into what someone is thinking and experiencing.

So, emotion recognition plays a fundamental role in human-robot interaction. It can help a robot understand and respond to the needs of the user it is interacting with [10]. One of these needs could be clarification due to a communication breakdown during a conversation, which can be recognized based on the confusion shown by the human. Emotion recognition systems typically rely on multimodal data, including facial expressions, speech patterns, physiological signals, and body movements. Among these, facial expression analysis remains one of the most widely used methods due to its accessibility and strong correlation with emotional states [11].

The accuracy of facial emotion recognition systems has been significantly improved due to recent progress in artificial intelligence and machine learning [12]. More traditional approaches relied on the use of hand-crafted feature extraction, such as the Facial Action Coding System (FACS) [13], to identify key facial muscle movements linked to specific emotions. However, in recent years deep learning techniques, especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have facilitated automatic feature extraction. This allows for the construction of an end-to-end system leading to more robust and adaptable emotion recognition models [14].

Through the use of data that consists of videos, rather than still images, the temporal element of this data can be exploited to improve emotion recognition in models. In traditional facial emotion recognition, this can be achieved by adding the displacement between frames of all the facial landmarks to the features given to the model [15]. Deep learning approaches can handle temporal information within the model by utilizing specific architectures and layers such as long-short-term memory (LSTM) [16]. A hybrid approach that combines spatial and temporal information, such as CNNs for spatial feature extraction and LSTM networks for temporal analysis, have shown promising results in recognizing emotions dynamically over time.

Despite all these advancements, not all emotions are recognized equally well. Confusion is one of the emotions with a lower recognition accuracy which complicates detecting communication breakdowns. One of the biggest factors of why specifically confusion is a difficult emotion to recognize is that its expressions are quite subtle and complex compared to other emotions. Confusion is also an emotion that is mainly signaled in facial expressions, and head or body movements[17].

Further factors that impede the recognition of confused facial expressions are individual differences, cultural variations, and age-related facial changes. These age-related changes consist of noticeable visual differences such as wrinkles. However, older adults also have fewer facial muscle movements [18] which results in a poorer performance in emotion recognition using facial expressions. Furthermore, there are also only limited datasets available for emotion recognition for older adults based on facial expressions.

In this paper, we aim to answer the question: Can a machine learning model accurately detect confusion in the facial expressions of older adults during interactions with a social robot? To address this question, we first collected a novel dataset consisting of video recordings of human-robot interactions during a word game where we actively tried to induce confusion. The dataset with facial expressions is annotated to differentiate between confused and non-confused facial expressions. We then trained and evaluated two machine learning models on this data to assess their effectiveness in recognizing confusion based solely on facial expressions.

The remainder of this paper is structured as follows: Section 2 provides a review of related work on human-robot communication as well as emotion recognition. Section 3 describes the video data collection and annotation for the new dataset. Section 4 describes data preprocessing and the architecture and training of the two models used for recognizing confusion based on facial expressions of the elderly. Section 5 presents the model performance as well as further analysis of the dataset. Section 6 discusses these findings and their implications for confusion recognition in the elderly using facial expressions. Finally, Section 7 concludes the paper.
Figure 1 illustrates this study's process including data collection during a word game played with a social robot, manual annotation of confused facial expressions, the training of two machine learning models on the new dataset, and evaluation on participants not included during training.

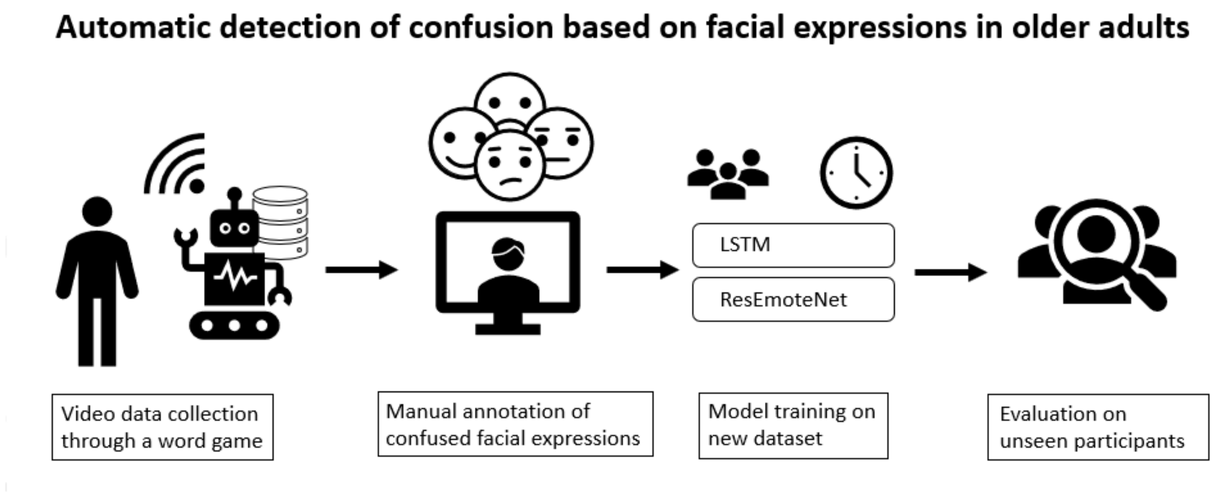


Figure 1: Overview of the study design for confusion recognition based on facial expressions in older adults.

2 Background Literature

2.1 Human-Robot Communication

As robots become more present in different human-centered environments, effective communication between humans and robots has become increasingly important. Social robots, which are created to interact in a human-like manner, are expected to engage in natural and intuitive communication. These robots are supposed to both display and interpret human-like intentions, emotions, and reactions.
Communication in human-robot interaction (HRI) is a very broad topic as it includes both verbal and nonverbal exchanges. The development of new Large Language Models, such as OpenAI's GPT models and the new DeepSeek models, has caused considerable improvements in robots' dialogue systems. However, a robot's ability to naturally communicate with humans is still quite limited. One of the fundamental problems in HRI is grounding during communication. Grounding is a joint process during communication where conversation partners verify to what degree they understand each other as well as to what extent the other is aware of this [19, 20].

In human-to-human communication, grounding is accomplished through various different methods. One of these methods is the use of back-channels where a listener interjects responses to the speaker as an indication of understanding. Similarly, humans also make use of facial expressions, gestures, and other turn-taking signals like eye gaze. Furthermore, actions that demonstrate comprehension of what was communicated can also show understanding.
Humans, especially when they are inexperienced in communicating with robots, use these same grounding tactics in human-robot interactions. However, robots still face technical challenges when interpreting these tactics as these interactions require real-time multimodal processing of noisy data from hard-to-predict interactions [21].

2.1.1 Kopp and Krämer's Communication Model

Various communication models have been designed to facilitate human-robot interaction. One such model is proposed by Kopp and Krämer [8], who believe that mentalizing and incremental joint co-construction are two essential elements in effective human-robot interaction.
According to them, for a robot to successfully ground communication, it should have an understanding of the conversation partner's mental state. This is done by mentalizing. In earlier machine learning-based systems, a behaviorist approach is employed where they seek to extract patterns in communication from large amounts of surface-level data. However, Kopp and Krämer proposed a more holistic mentalizing approach as this would include the ability to perceive, interpret, and understand the conversation partner's mental states exceeding the level of discreet dialog or belief states that are commonly considered. These mental states influence an individual's interactive behavior which helps predict their actions.

For a robot to be able to do this, it needs to be able to reconstruct the construction done by the conversation partner using their mental state leading to a mutual understanding. Every person constructs how they view the world in different ways. So, to be able to mentalize effectively, humans use communication and observations of a conversation partner's behavior to understand their mental states. However, even humans can never be sure that the interpretation of the other's mind is correct and complete while their conversation partner also correctly interprets their communication signals [22, 23].

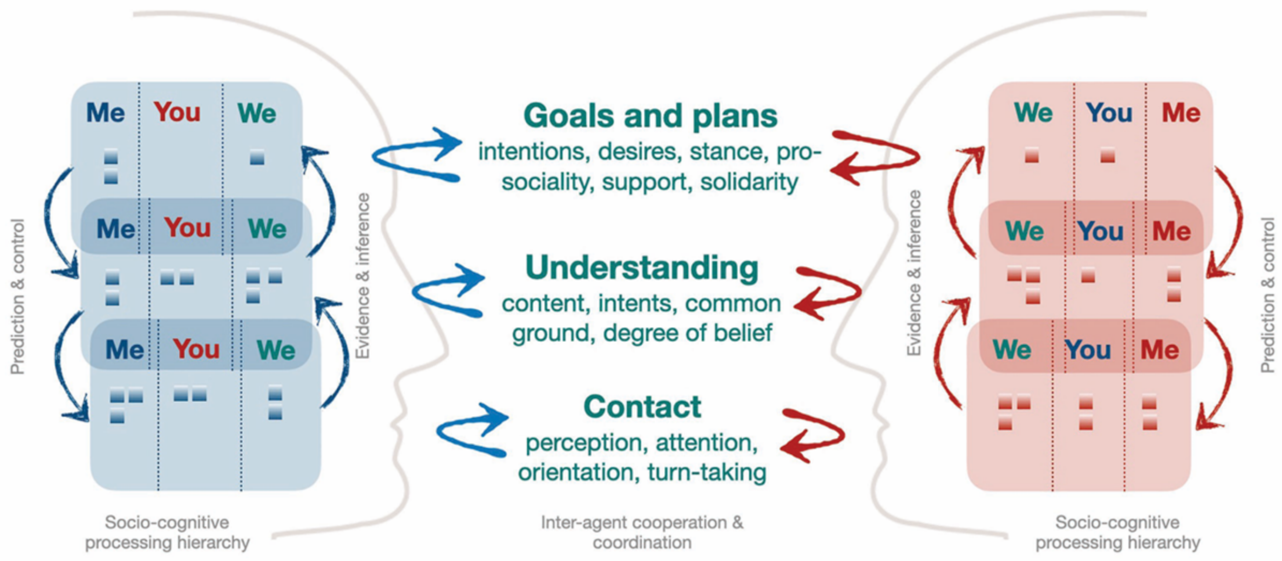


Figure 2: Kopp and Krämer's communication model based on mentalizing and incremental, joint co-construction. (Figure taken from [8])

To overcome these problems, joint co-construction is used. There are three levels that an interlocutor can be in (see Figure 2). The lowest of these is contact, at which point an agent speculates whether their conversation partner is aware of their attempt to initiate contact. After both agents are sure that there is contact, they continue to the second level, which is understanding. At this level, an agent will check that co-construction was successfully done and that the intention behind the communication is achieved. The highest level is plans and goals. At this point in the interaction, the agent is sure that their conversation partner has gotten all the information about the interaction, so they only need to examine whether the other has goals that are aligned with their own.

Using joint co-construction, a human tests increasingly complex hypotheses about the conversation partner's mental state within these levels and how it can be adjusted toward the plans and goals state of the interaction. By grounding these levels in incremental, responsive dialog actions, communication partners can iteratively co-construct both their interaction and mental states at the same time, ensuring that these are identical during the interaction.

This approach is not often present in current robots and their dialog models. Common models often assume that all necessary reasoning for a response can be done on the spot as long as there is a sufficiently large set of training data available. By incorporating both joint co-construction and mentalizing capabilities in a robot, we can improve their grounding of a conversation leading to more natural human-robot interactions.

2.1.2 Communication Breakdowns

Even when utilizing the communication approaches by Kopp and Krämer, communication breakdowns remain a significant challenge in human-robot communication. A communication breakdown occurs when a misunderstanding transpires during an interaction. These can stem from various factors such as someone providing contradictory information due to inconsistent factual or procedural knowledge or a lack of knowledge [24].

Further reasons include a lack of knowledge about emotions resulting in the robot's inability to recognize and respond appropriately to the conversation partner's emotional state. Different aspects of the robot's conversation processing can have gone wrong, such as speech recognition errors, ambiguous phrasing, misinterpretation of emotions, or unexpected user behavior.

Communication breakdowns often lead to confusion or frustration and should thus be reduced. This both improves the reliability of the robot and enhances user experience in human-centered environments. Different strategies can be used by a robot to deal with a communication breakdown which all require different levels of user interactions.
One of these strategies is using an error recovery mechanism where the robot prompts the user for clarification. A problem with this approach is that the robot has to figure out when it is uncertain about something before it can ask the user for clarification. Another strategy is the use of adaptive learning models. A robot that utilizes these models will improve its understanding over time by learning from previous interactions during a conversation. A third, more computationally expensive strategy is using multimodal redundancy. Robots that use this strategy will look at multiple different aspects of communication, including both the verbal and nonverbal aspects to be able to reduce errors in understanding [7].

All three of these strategies have one common restriction. For each of them, the robot needs to recognize the misunderstanding leading to a communication breakdown before it can overcome this breakdown. Sometimes, the communication partner will indicate this through a verbal clarification request. However, various non-verbal approaches can help the robot recognize the communication breakdown earlier. One of these approaches is examining facial expressions. During a misunderstanding in human-robot interaction, signs of confusion have been detected [25].

2.2 Emotion Recognition

Emotion recognition plays a vital role in human-robot interactions as it enables robots to respond appropriately to users' feelings and intentions.
Most research in emotion recognition focuses on (a subset of) the emotions from Ekman's six basic emotions model [26], which classifies human emotions into surprise, fear, disgust, anger, happiness, and sadness. According to Ekman, these emotions are universally recognized across different cultures.

Various modalities can be used when recognizing emotions. The most common are facial expressions, speech patterns, body language, and physiological signals. Emotion recognition approaches based on physiological signals are mainly done in a lab setting since the methods to get these signals are very invasive. The other emotion recognition approaches are less intrusive as their necessary information can be obtained from image, video, or audio recordings.
A multimodal approach is often better than only using a single modality as some modalities can give complementary information, such as speech and facial expressions [27]. However, not all emotions are shown equally well in all modalities [17], so it is important to examine what modality best fits the emotion that is examined.

2.3 Facial Emotion Recognition

Facial expressions are often used to recognize emotions as it is an easily available approach that also shows a strong correlation with emotional states [11].

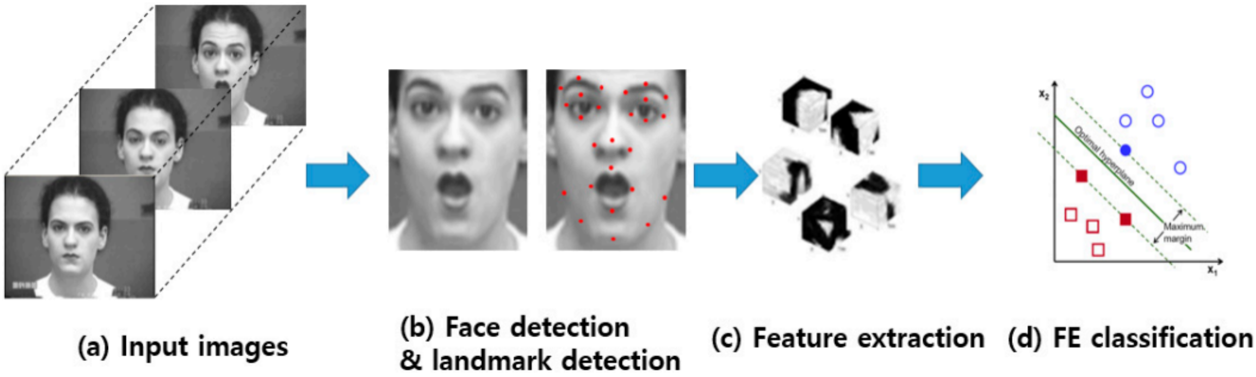


Figure 3: An example of the procedure of conventional FER approaches. (Figure taken from [28])

The different approaches for facial emotion recognition (FER) can be categorized into two different methods. One of those is the conventional methods and the other is the deep learning-based methods. Sometimes a combination of these can be used resulting in a hybrid approach.

2.3.1 Conventional FER Approaches

The approaches initially used for facial emotion recognition are part of the conventional approaches (see Figure 3). These methods follow three distinct steps.
First, the face and/or facial components are detected in the pictures or videos. From this, features are extracted. These features are hand-made and consist of geometric features, appearance features, or a hybrid of both types. The geometric features are based on the relationships between different facial components. One example of this is Ghimire and Lee [15], who created a model based on 52 facial landmarks consisting of geometrical information while excluding facial texture information. Appearance features, on the other hand, use global texture descriptors such as Local Binary Patterns (LBP) and Gabor filters to capture facial variations. Finally, the facial expression is classified based on the extracted features. This classification is done using traditional machine learning models like Support Vector Machines (SVMs), AdaBoost, and Random Forest.
One way to add the temporal components of facial expressions in the features given to train a model is by adding information about landmark displacement between video frames to the feature vector [15].

2.3.2 Deep Learning-Based FER Approaches

The newer approaches in facial emotion recognition often utilize deep learning methods. These methods remove the need for hand-crafted features by enabling end-to-end learning from raw images or videos [29].
The most common architecture used is the convolutional neural network (CNN), which was proven to be effective in extracting spatial features from images [30, 31, 32]. An example of this approach is shown in Figure 4. A CNN-based FER model uses multiple layers of convolutional filters to extract relevant features from facial images. After these layers, there is a fully connected network for classification.

CNNs alone cannot capture temporal dynamics as they process input data as independent, static frames. To solve this, recurrent neural networks (RNNs) and long short-term memory (LSTM) networks model sequential dependencies in facial expressions over time [16]. Both of these networks

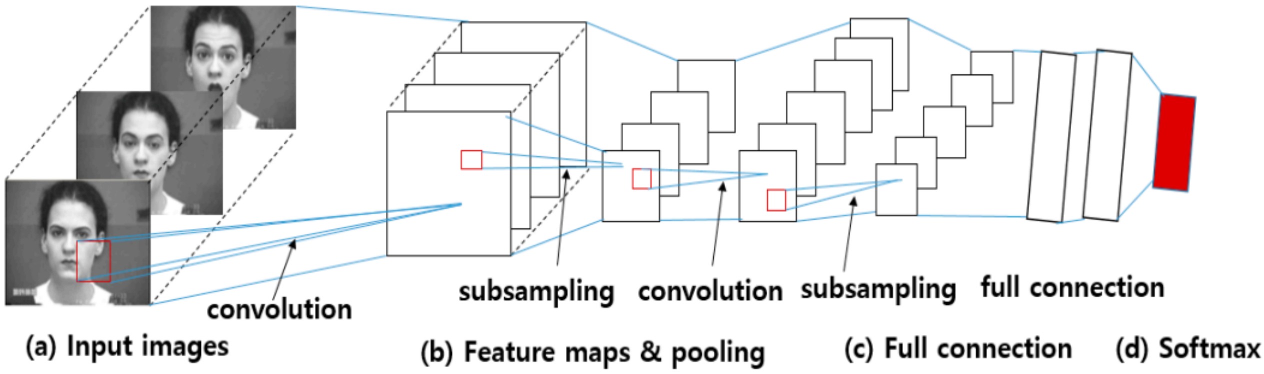


Figure 4: An example of the procedure of a CNN-based FER approach. (Figure taken from [28])

have a memory mechanism that can look back on previous inputs to inform the outcome of the current input.

2.3.3 Facial Emotion Recognition for Older Adults

Studies on automatic emotion recognition show that recognition of faces from older people tends to be more difficult than those of younger people [33, 34, 35]. Older adults often have wrinkles and other changes in skin texture that influence the accuracy of feature detection on their faces [36]. They also have reduced facial muscle movements caused by an age-related decline in facial muscle function. Thus making all emotions less pronounced and more difficult to accurately detect and recognize [37].

Even the emotional display patterns of these older adults can be different. Older adults exhibit different facial expression intensities and muscle usage compared to younger individuals [18]. Elderly individuals generally engage more muscles in the lower face area compared to younger people. They also display more negative emotions than positive emotions. Due to these differences, facial emotion recognition models trained on younger peoples' faces often struggle to generalize to older populations. So, there is a necessity for specialized datasets that include older age ranges to improve emotion recognition performance.

Furthermore, there is also a chance that older individuals engage in prosocial behavior [38]. Prosocial behavior encompasses a wide range of positive social actions such as altruism, efforts to reduce stereotypes, and friendly interactions. During emotion recognition research, elderly people could exhibit prosocial behavior by hiding negative emotions associated with misunderstandings to pretend there are no problems during the interaction.

Both psychological theory and empirical research have previously suggested that prosocial tendencies increase with age. Studies have found that older adults report more prominent empathic concerns [39, 40] and consistently donate more to charity than younger individuals [41, 42, 40]. Similarly, evidence from economic behavior studies shows that elderly participants consistently exhibit more generous behavior by sharing more resources compared to middle-aged or younger participants [43, 44]. It was also shown that older adults are more likely to respond to empathy cues, reflecting a heightened social sensitivity [45]. These findings suggest that prosocial behavior may intensify with age, potentially due to shifting emotional and social priorities.

2.4 Confusion Recognition

As mentioned before, it is important to detect misunderstandings during communication. One method for this is recognizing confusion from various emotion recognition modalities.

Some approaches in confusion recognition utilize electromyography (EMG) or electroencephalography (EEG) signals to monitor activity patterns associated with confusion. These signals the input features to a machine learning algorithm that is trained to classify and predict confusion [46, 47]. However, these techniques are invasive as participants are required to have several sensors connected to their body, making it impractical for human-robot interaction. In addition, the accuracy of these techniques is low compared to other non-invasive techniques.

2.4.1 Non-invasive Confusion Recognition Techniques

There are various different non-invasive methods have been explored to identify confusion.

Body posture and general physical behavior have been shown to be reliable indicators of confusion. Caballé et al. [48], for instance, used a wireless motion sensor to monitor participants' overall body movements to distinguish different emotional states including confusion. Other approaches utilize pressure-sensitive seating systems to track posture changes through sensor sheets embedded in the seat and backrest. These systems can detect subtle shifts in body posture that correlate with emotional states. D'Mello and Graesser [49] used body postures to detect negative and strong emotions such as confusion.

Natural language can also provide further insights into someones emotional and cognitive states. D'Mello et al. [50] found that dialogue features could predict the affective state of confusion. Machine learning experiments demonstrate that standard classifiers were moderately successful in detecting confusion, achieving an accuracy of 68% when recognizing confusion in contrast with neutral affect.

Another promising non-invasive confusion detection method involves facial expressions and visual attention tracking. Facial expressions often change when confusion occurs. Postma-Nilsenová et al. [38], for example, found that confusion is signaled by facial cues that can be automatically detected with currently available facial expression recognition technology. However, there is a lot of variation between the confused facial expressions of individuals [51]. Gaze patterns can also be used to detect confusion. Graesser et al. [52] found that confusion, which is the emotional expression of a cognitive disequilibrium, often results in altered visual exploration. Furthermore, Cumbal et al. [53] found that people often look more at the eyes of a robot when they are uncertain.

The effectiveness of these systems can improve when multiple modalities are combined to help improve accuracy and reduce ambiguity. For example, non-linguistic vocal expressions (e.g. "Huh?") can be combined with physical behavior such as head scratching and changes in upper body and head position. Using these modalities, trained judges were able to code confusion during a learning activity [54].

2.4.2 Using Facial Expressions to Recognize Confusion

Our study only focuses on confusion recognition through facial expressions. This is a non-invasive technique as the expressions are recorded with a camera rather than something in contact with the conversation partner such as electrodes for EEG studies. It also achieves a higher accuracy in general emotion recognition compared to speech data [55].

Many approaches made use of the Facial Action Coding schema of Eckman and Friesen [13] to extract Facial Action Units (AUs) from facial expressions. These AUs were then used as features to train a confusion recognition model [38, 56, 57, 58]. Earlier studies have shown a clear difference in AU presence and gaze focus during uncertain events in contrast to normal conversation [53]. Furthermore, clarification requests and an absence of vocal responses were often present during uncertain events.

Some previous studies have utilized the temporal element of confusion when training an emotion recognition model using AUs. One example of this is a study by Borges et al. [56], who collected data consisting of video recordings. Miscommunications were induced during these recordings to raise the frequency of the participants displaying confused facial expressions. FaceReader software was used to extract the Facial Action Units and affective state information. An LSTM modeling approach incorporated the temporal context of the states in the recordings. The LSTM neural network had an accuracy of 87% and lesioning trials of the pre-trained best model further showed that action units involving open-mouth expressions were the most important factors in the performance of the model.

There have been other studies that focus on end-to-end deep learning-based approaches. A study by Shi et al. [59] compares four methods to recognize confusion. Each of the methods used a support vector machine (SVM) with different feature extraction techniques. Facial expressions of 82 students were obtained and labeled according to participants' self-reports. The best model performance came from the approach in which an SVM was used after features were extracted using a convoluted neural network, resulting in an accuracy of 94%.

Most studies that focus on recognizing confusion from facial expressions use data from younger people. However, machine learning approaches can accomplish an above-chance accuracy in recognizing confusion in elderly people as well. Even achieving better performance than human observers [38].

2.4.3 Facial Action Units

Some approaches to recognize emotions from facial expressions use the Facial Action Unit system to identify the key action units (AUs) associated with different emotions. Ekman and Friesen [13] established this system as a comprehensive framework to break down facial movements into individual action units. Subsequent research has determined some of the most relevant AUs for different emotions. Figure 5 displays an example of some Facial Action Units important for confusion recognition.

Confusion was found to strongly correlate with AU4 (Brow Lower), AU7 (Lid Tightener), AU12 (Lip Corner Puller), and AU23 (Lip Tightener). Research by Borges et al. [56] demonstrated that next to these AUs, AU25 (Lips Part), AU26 (Jaw Drop), and AU27 (Mouth Stretch), involving open mouth expressions were the most individually relevant action units for detecting confusion in older adults. Similarly, Cumbal et al. [53] also found that AU26 and AU1 (Inner Brow Raiser) showed significant differences between situations with uncertain events that lead to confusion and situations without these events. However, Yasser et al. [57] found that among others, AU1, AU25, and AU26 did not affect the recognition rate of confusion.

Rozin and Cohen's research [51] found a lot of individual variety between the confused facial expressions of people. Some participants primarily used the eyebrow region of their face while others were more expressive in the mouth area. Another difficulty in detecting confusion is the similarity between concentration and confusion in their characteristics [61].

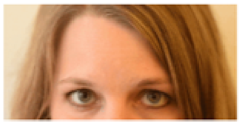
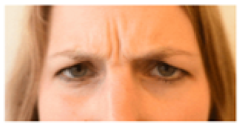
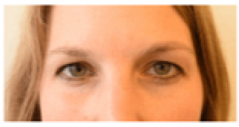
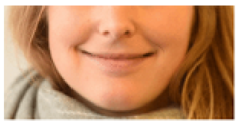
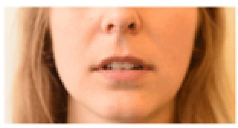
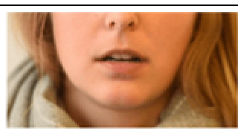
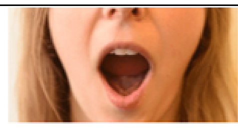
Action Unit	Description	Facial Muscle	Example
1	Inner Brow Raiser	*Frontalis, pars medialis*	
4	Brow Lowerer	*Depressor Glabellae, Depressor Supercilli, Currugator*	
7	Lid Tightener	*Orbicularis oculi, pars palpebralis*	
12	Lip Corner Puller	*Zygomatic Major*	
23	Lip Tightener	*Orbicularis oris*	
25	Lips part	*Depressor Labii, Relaxation of Mentalis (AU17), Orbicularis Oris*	
26	Jaw Drop	*Masetter; Temporal and Internal Pterygoid* relaxed	
27	Mouth Stretch	*Pterygoids, Digastric*	

Figure 5: Examples of the facial action units mentioned in various studies on confusion recognition. (Examples taken from [60])

3 Data Collection

This section outlines the methodologies and tools used to collect data for our research objectives. The data, consisting of video recordings, was obtained through an experiment where elderly participants played a word game with a social robot. This data was annotated and will be used to train a machine learning model to recognize confused facial expressions during a conversation. We aimed to maintain the reliability and validity of the data while giving careful consideration to the ethical integrity of the data collection process.

3.1 Setup and Preparation

The experiment was conducted in an office meeting room located in Friesland. There were decorations present around the room. These could have distracted participants during the word game, however, it is also likely that these decorations made the participants feel more at ease during the experiment by giving the room a deformalized feel.

The QTrobot was positioned on a table in front of a white wall. A chair was placed in front of the robot for the participant to use, while the researcher was seated behind the robot to monitor the experiment (see Figure 6a). This setup was chosen to reduce the possibility that the participant would accidentally start looking and talking to the experimenter instead of the robot during the word game.

(a) The experimental setup from the experimenter's point of view.

(b) The experimental setup showing the social robot from the side.

Figure 6: The experimental setup from different angles.

3.1.1 Recording Devices

The interactions between participants and the social robot were recorded using a combination of internal and external recording devices (see Figure 6b). The robot records audio using an external directional microphone. This microphone is the Sennheiser Profile USB-C microphone with a sample rate of 48 kHz. It minimized ambient noise and ensured that participant speech was recorded clearly.

The robot also has an internal camera above its facial display. This is the Intel® RealSense™ Depth Camera D455 with a resolution of 1280 x 800 and a frame rate of 30 fps. A second external camera was positioned next to the robot at eye-level of the participants. This camera was a Canon EOS 1200D, which recorded videos with audio during the interaction.
The recordings were synchronized and stored for offline annotation and model training. All equipment was positioned unobtrusively in an attempt to reduce participant discomfort and maintain a natural interaction.

3.1.2 Social Robot

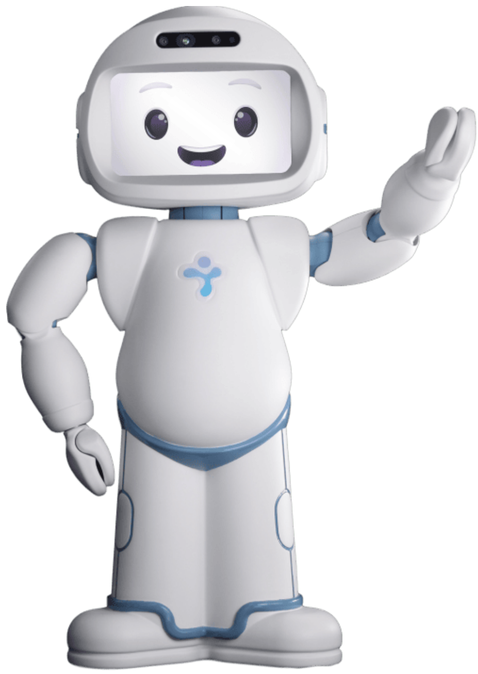


Figure 7: The QTrobot used during the experiment. (Image from [62])

The robot used to collect the data for this study is the QTrobot manufactured by LuxAI (see Figure 7). Originally, this robot was developed to be used in special needs education and to facilitate research on human-robot interaction. We chose to use this social robot for the experiment due to its humanoid appearance. This simplifies the interaction with the robot for people who have no prior experience regarding communication with social robots.
The robot is equipped with a facial display that can show various facial expressions. A combination of RGB and 3D cameras as well as a digital microphone array allows the robot to perceive and record its environment. The robot's software architecture further integrates local speech synthesis capabilities to enable vocal communication. An external microphone was connected to aid the speech recognition of soft voices of some of the elderly participants.

3.2 Word Game

The interaction between the robot and the participants consists of an extended version of the 20 Questions game. In our version of this game, the participant thinks of a word. The social robot will then ask the participant yes/no questions in an attempt to determine what this word is. After the robot believes it knows the correct word, it will make a guess. If this guess is correct, the game will start from the beginning and the participant will think of a new word. The original version of this game has a limit of 20 questions. However, in our version, the robot can ask as many questions as it requires to guess the word. If it takes a long time for the robot to figure out the correct word, participants can give a hint or tell the robot to stop guessing and instead restart the game with a new word. There were no strict rules regarding hints, so some participants gave more hints to the robot than others.

3.3 Interaction Model

The robot's interaction model for the word game uses turn-taking principles combined with a large language model that keeps track of the game state. A visual representation of the decisions and processes during the interaction is visible in Figure 8.

3.3.1 Game Flow

As Figure 8 shows, the robot will start the interaction by greeting the participant. From this moment on, the rest of the interaction is based on turn-taking, where the robot will always expect a response from the participant before it continues the conversation.

The robot uses an automatic speech recognition model to hear what the participant has answered to the robot's previous utterance. During the pilot test of the experiment, we found that participants sometimes forgot that it was their turn to speak. To prevent the conversation from halting completely, the robot will indicate that it is listening when it has not recognized any speech. To indicate this, the robot makes a coughing noise and sometimes shrugs its shoulders. Further testing showed that these actions would usually prompt the participants to give a verbal reaction after which the conversation continues.

A large language model combines the participant's new utterance with the previous utterances from both the robot and the participant during the current conversation. The memory of the large language model was 10 instances. The initial game prompt given to the language model that controls the robot's responses is the following (Appendix A shows the original Dutch prompt):

```
"You are a social robot.
Play a word game with an elderly person where you try to
guess the word that this person is thinking of.
To do this, you can ask a question about this word that
the person will answer with yes or no.
If you guess the word correctly, you restart the game.
You don't want to do anything other than play this game."
```

This prompt was created to keep the robot from deviating from the game while still allowing the participant some freedom of what to say during the interaction. It instructs the model about how the word game is played but does not restrict the conversation too much. This is done to allow the participant to ask for clarification about what the robot has said. The robot still plays the game correctly, often asking the participant yes/no questions to help narrow down the participant's word. Other times,

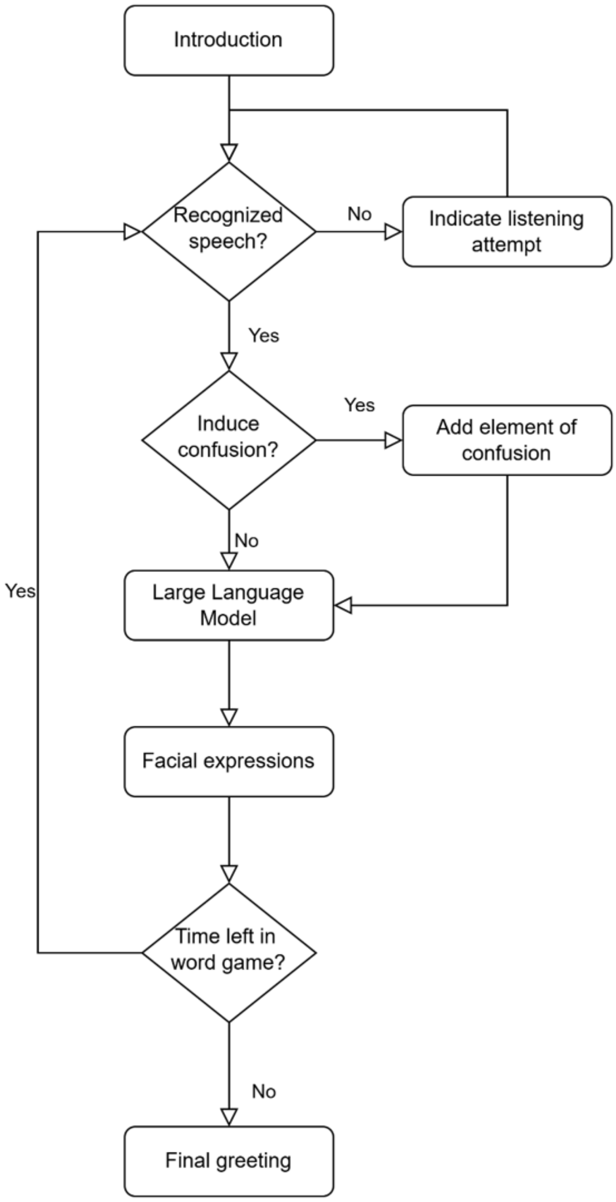


Figure 8: A representation of how the social robot's internal model decides what to do during the interaction.

the model will try to guess the word or potentially respond to a question being asked by the participant.

The initial game prompt combined with nine previous responses and the new response were used to generate a new response during the word game. This is also the stage in the conversation where confusion-inducing action sometimes takes place. Sentiment analysis of the robot's response determines whether the robot changes its facial expression from the standard neutral expression.

In some instances, the robot would get confused during the word game. This either happened when the robot kept repeating the same questions or sometimes the robot would try to switch roles with the participant. Further improvement of the prompt could prevent the switching of roles, while a larger memory might prevent repeated questions.

After the word game ends, the robot will say a final greeting to indicate to the participant that the experiment is done.

3.3.2 Interaction Interface

The automatic speech recognition model that the social robot uses is the Google Cloud Speech-to-Text model [63]. This model processes the recorded speech in real-time and provides the robot with an accurate transcript of what was said by the participant.
The large language model used by the robot is the GPT-3.5 model by OpenAI [64]. This model generates a response during the word game using the information provided in the new transcript along with a memory consisting of the nine previous interactions in the conversation and the initial game prompt.
The response from this language model is then spoken in a Dutch voice by the robot using its internal text-to-speech software.

Both the automatic speech recognition and language model used by the robot are external models called from the internet. This means there is some latency which results in slower responses than when using a local model that runs on the robot. However, this is not a significant problem for our current research objective since the participants are all elderly adults. Cognitive processing speed (how fast mental operations can be executed to complete a task) slows with advancing age [65, 66, 67]. Thus, the robot's slow replies are similar to how older adults themselves would respond. After explaining that the robot requires some time to 'think', many participants also indicated that they prefer the game to be played slower anyways.

3.3.3 Induce Confusion

There are two different ways that confusion is induced during the word game. These approaches were selected based on earlier research by Cumbal et al. [58].
The first of these techniques is the use of non-contingent utterances. The robot will then respond with a sentence containing incorrect information, fake words, or a sentence unrelated to the word game. A complete list of these sentences can be found in appendix B. The second way to induce confusion is by speeding up sentences. These sentences use the large language model responses with an increased speech rate. This causes the participants to have difficulty understanding these sentences.

To determine when confusion should be induced, the model looks at the number of sentences previously spoken by the participant during the experiment. By inducing confusion only after a set number of sentences are spoken, we attempt to prevent the inducing events from happening close together during the conversation. This approach meant that confusion was less likely to make participants frustrated or uneasy which decreased the chance of participants wanting to end the experiment early. Too many confusion-inducing events within a short time frame could also allow the participants to deduce that we are trying to induce confusion which would be detrimental to the reliability of the experiment. However, this approach also lessens the extent to which we can control how often confusion is induced for each participant.

To make the interactions that caused confusion easier to understand, we include some translated transcripts of situations that happened during the word game. Transcript 1 illustrates the moment when the robot deliberately increases its speech rate when responding. In this example, the human participant got confused after the robot's response and asked for clarification afterward. Transcript 2 shows a similar situation, however, in this example, the robot induces confusion by responding to the participant with a non-contingent utterance.

```
QTrobot: Is the object a tree?
Participant: No.
[Increase speech rate for the next sentence]
QTrobot: Is the object a garden seat?
[Participant is confused]
Participant: I did not hear what you said.
[Reset speech rate back to original speed]
QTrobot: Let me ask my question differently:
QTrobot: Is the object mostly used in the garden?
```

Transcript 1: A human-robot conversation where confusion was induced by increasing the speech rate.

```
QTrobot: Is the object that you are thinking of a hat?
Participant: No, it is not a hat.
[Insert non-contingent utterance]
QTrobot: Have you ever tasted a battery?
QTrobot: It is my favorite food!
[Participant is confused]
Participant: I didn't understand that.
QTrobot: Don't worry, let's continue the game.
```

Transcript 2: A human-robot conversation where confusion was induced by using a non-contingent utterance.

Sometimes, participants did not understand the social robot correctly during the word game. An example of this is present in Transcript 3, where the participant did not exhibit confusion after the robot used a non-contingent utterance. It is likely that the participant did not hear the robot correctly and thus might have thought the sentence was a regular sentence such as "The sun is so beautiful today".

```
QTrobot: Is the object a lifebuoy?
Participant: No.
[Insert non-contingent utterance]
QTrobot: The sun is so beautiful blue today.
[Participant does not show confusion]
Participant: Yes I agree.
```

Transcript 3: A human-robot conversation where confusion was attempted to be induced by using a non-contingent utterance.

This did not necessarily reduce the number of confusion instances during the word game since participants who had this problem could have the same problem during normal conversation. This means that they sometimes became confused without the robot doing anything to induce confusion as can be seen in Transcript 4 and 5.

```
QTrobot: Let's start the game.
QTrobot: Are you thinking of an object?
[Participant is confused]
Participant: I did not understand that correctly.
QTrobot: I understand.
QTrobot: Let's try again.
QTrobot: Are you thinking of a word that is an animal?
```

Transcript 4: A human-robot conversation where confusion was not induced but did occur.

```
QTrobot: Is the animal that you are thinking of a wild animal?
Participant: It is not a wild animal either.
QTrobot: Could it be a bird?
[Participant is confused]
Participant: I did not understand that.
QTrobot: Sorry for my mistake.
QTrobot: Is the animal you are thinking of a reptile?
```

Transcript 5: Another human-robot conversation where confusion was not induced but did occur.

3.4 Participants

A total of 27 participants (13 female and 14 male) ranging from 64 to 83 years old (mean = 74; standard deviation = 5) took part in the data collection experiment. They were recruited through email invitations based on age and proximity to the experiment location. The email addresses were obtained from a local organization that organizes events for elderly people. We excluded potential participants who had cognitive impairments to minimize possible effects on facial expressions made during the experiment. This was self-reported by participants. The participants all speak Dutch fluently. Prior to the experiment, the participants gave written consent to take part in the experiment and use the video recordings for research purposes. The Research Ethics Review Committee (CETO) of the University of Groningen has approved the experiment in this study (Appendix C).

3.5 Questionnaire

To assess user experiences during the interaction with the social robot, a questionnaire was conducted after the participants had played the word game with the robot. The questionnaire covered 14 items consisting of differential scale and Likert-type questions. The likability and perceived intelligence questions from the Godspeed questionnaire [68] were included, with Dutch translations obtained from [69].

Next to these questions, the questionnaire also contained Likert-type questions to ascertain perceived confusion and understanding as well as questions about the social robot's perceived usefulness in elderly care. These questions were the following:

- The conversation was sometimes confusing. (Likert scale: 1 = Disagree, 5 = Agree)

- The robot was sometimes difficult to understand. (Likert scale: 1 = Disagree, 5 = Agree)
- I think these types of robots can help in elderly care. (Likert scale: 1 = Disagree, 5 = Agree)
- I would personally benefit from using this robot. (Likert scale: 1 = Disagree, 5 = Agree)

The original Dutch questionnaire can be found in Appendix D.
The answers to these questions are subjective. It is further important to note that participants were not aware of the fact that we tried to induce confusion during the experiment while filling out the questionnaire.

3.6 Procedure

This section outlines the procedure followed during the experiment, which involved participants playing a word game with the social robot. The primary objective was to obtain video data of facial expressions made by the participants during the word game.

Before each session, participants were briefed about the experiment's general purpose of improving human-robot interaction. We also explained the procedure during the experiment. Each participant signed the informed consent form (see Appendix E) before proceeding with the experiment.

At the start of the interaction with the social robot, its volume is changed according to the participant's preferences. Participants were also instructed on how to interact with the robot. This includes some simple instructions such as waiting for the robot to stop speaking before responding and always verbally answering the robot. To familiarize themselves with the robot, participants first engaged in a casual, open conversation with the robot. They could choose any topic for this conversation. This allowed the participants to get used to simple interactions with the robot and decreased nervousness during the actual experiment. The practice conversation was capped at 10 minutes, but participants could indicate if they wanted to stop practicing earlier.

We always confirmed with the participants whether they had practiced enough and were comfortable interacting with the robot. This was also a moment for the participants to ask any questions they still had before the data collection experiment. After explaining the word game one more time, the actual experiment started. During the experiment, participants played the word game, as described earlier, with the robot. Some example words were provided, however, participants were free to choose any word they wanted. Participants were instructed to only interact with the robot during the experiment, so when confusion was induced as explained earlier, they would often ask the robot to repeat itself or explain what it meant. During the game, we collected video data of the various facial expressions of the participants.

The word game took around 10 minutes. After the robot responds to the participant, it will check whether there is still time left. If the word game had been going on for 10 minutes or longer, the robot indicated that the game was done and the experiment was finished.
After the word game interaction, participants completed the questionnaire mentioned in section 3.5. The participants filled in the answers fully autonomously. However, they could ask clarifying questions whenever they were unsure about what was asked in a question. Finally, the participants were debriefed and informed of the actual purpose of the experiment as well as the techniques used to induce confusion. Printed versions of the experiment information and the debriefing as well as a

copy of the informed consent form were provided. Participants were also instructed not to discuss the experiment with anyone until after all participants had taken part in the experiment.

3.7 Data Annotation

To facilitate the use of the obtained data, all video recordings were systematically annotated using a coding scheme based on both facial and behavioral cues observed in participant responses to the robot. The annotation framework was developed based on prior studies, which found that there is a lot of movement in the eyebrow and mouth regions of the face when a confused facial expression is made [51]. Extra attention was given to determining the onset and offset of both the communication breakdowns and the confused facial expressions. In addition to changes in facial expression, we also made use of verbal signals and head or body movements to annotate the data.

The onset of confused expressions was defined as the moment a noticeable change in facial expression occurred following a robot utterance, which led either to a clarification request by the participant or a noticeable delay in response time. Some behaviors that were typical indicators for confusion were participants leaning towards the robot (possibly attempting to listen better) or away from the robot (possibly attempting to look around the robot towards the experimenter). Facial markers frequently involved frowning in the eyebrow regions, downward movement around the mouth, or slight opening of the mouth.
The offset of confused facial expressions was marked by a clear shift in facial expression following a new utterance or clarification by the robot. This shift was characterized by a more positive or sometimes neutral facial expression, including the closing of the previously open mouth, signaling a reduction of confusion.
The annotation process aimed to identify moments during the recording when confused facial expressions were exhibited by participants during the experiment.

Annotation was done using the ELAN software tool. Two annotators independently reviewed 5 minutes of 10 videos. Inter-coder reliability was measured on each video using the Cohen's Kappa score. To do this, the continuous temporal annotated data was first converted into discreet segments by dividing each video segment into one-second intervals with a class label determined by the most frequently occurring class within that second. This discretization allows for a direct comparison between annotators on a per-segment basis. One video was dropped due to a low Cohen's Kappa score (K=0.0). The average Cohen's Kappa for the remaining videos was a near-perfect agreement of 0.88 (SD=0.06) indicating that the annotation scheme can be used to reliably annotate the data.

The coding scheme was refined by expanding the attributes that indicate confusion. Gaze behavior was also added since participants would often have more and faster eye movements during moments of confusion. The full written coding scheme is presented in Appendix F. We further determined that indicators and facial markers greatly differed per participant, so not all of them were necessary to determine confusion. After this coding scheme was improved, the previously coded videos were reviewed to ensure that they fit the new coding scheme. The remainder of the videos were annotated by one annotator.

4 Model Development

In this study, two different machine learning models were trained and evaluated to detect confusion using video fragments and images from the newly collected data. The two models are a Long-Short-Term Memory (LSTM) model [56] and the ResEmoteNet model [70]. These models were chosen due to their distinct approaches to facial emotion detection. The LSTM model requires preprocessing of the video to extract facial action units. This model utilizes a memory mechanic which allows it to include the temporal element of our video data. The ResEmoteNet model, on the other hand, is an end-to-end deep learning model that is able to detect emotions from still images showing people's facial expression. By testing the confusion detection capabilities of both of these models, this study aims to find out whether it is possible to detect confused facial expressions of elderly individuals during human-robot conversations.

4.1 Data Preparation

The data gathered during the robot's conversation with the elderly participants is preprocessed to ensure that both models can learn from the dataset efficiently. The preprocessing steps were different for each model as the LSTM and ResEmoteNet models rely on different representations of features. There are two target classes for the model to learn: *Confused Facial Expression* and *Non-confused Facial Expression*. These were labeled in the collected data according to the coding scheme as mentioned in section 3.7. Two participants (participants 14 and 27 in Table 1) did not show any apparent confused facial expressions, so their data was excluded when training to prevent the models from learning that these specific people were never confused and thus influencing the model's training and final results. The other participants showed varying amounts of confusion based on their facial expressions, resulting in an imbalance in how often confused class data is present per participant. The total data contains more samples showing non-confused facial expressions, with only around 4.55% of frames displaying confused facial expressions. The distributions of the confused and non-confused classes for each participant are shown in Table 1.

4.1.1 Data Preprocessing for the LSTM Model

The LSTM model is trained using Facial Action Units (AUs) extracted from the facial expressions made by participants during the human-robot interaction. First, the video fragments were separated according to whether the participants had a confused facial expression or any other expression based on the previously made annotations. The OpenFace [71] software was used to analyze these video fragments. This is a tool for facial action unit detection [72]. The software detects the presence and intensity of 17 different AUs for each frame in the video.

The 17 FAUs extracted from the videos are: AU1 - Inner Brow Raiser, AU2 - Outer Brow Raiser, AU4 - Brow Lowerer, AU5 - Upper Lid Raiser, AU6 - Cheek Raiser, AU7 - Lid Tightener, AU9 - Nose Wrinkler, AU10 - Upper Lip Raiser, AU12 - Lip Corner Puller, AU14 - Dimpler, AU15 - Lip Corner Depressor, AU17 - Chin Raiser, AU20 - Lip Stretcher, AU23 - Lip Tightener, AU25 - Lips Part, AU26 - Jaw Drop, AU45 - Blink.

These include the most important action units for determining a confused facial expression as mentioned in various previous studies (see Section 2.4.3).

The LSTM model uses temporal data. So, the action unit data was split into multiple segments consisting of only confused facial expressions or only non-confused facial expressions. Each of these segments contains 30 frames which is equal to 1 second in the video recordings. The amount of confused facial expression segments is only 3.6% compared to the number of frames with confused facial expressions (4.55%). This is caused by these segments only containing frames with confused facial expressions, while not all instances of confusion last for 30 frames.

Participant number	Number of frames of coded video	Confused instances	Confused facial expressions (%)	Mean time confused in frames	SD time confused in frames
1	14283	10	12.2	174.10	152.30
2	13897	19	7.0	51.11	31.47
3	9402	5	2.4	44.40	16.84
4	12023	8	12.6	190.00	144.99
5	14409	10	4.4	63.20	33.25
6	2996	2	9.1	136.50	2.50
7	12524	10	9.7	121.40	75.50
8	8650	3	1.4	39.33	25.49
9	9000	6	7.5	113.00	83.57
10	13844	5	4.8	134.20	57.10
11	10458	15	8.5	188.53	303.26
12	14118	12	6.6	77.25	34.53
13	13846	3	2.5	114.00	106.60
14	-	0	0.0	-	-
15	5584	8	3.5	168.13	267.77
17	13839	9	5.5	84.11	48.84
18	11632	4	2.1	59.75	43.08
19	14462	3	1.4	66.33	52.65
20	12974	11	7.1	83.18	76.30
21	15357	6	3.9	100.00	67.38
22	14395	2	0.7	49.50	28.50
23	13806	1	0.1	15.00	0.00
24	11754	2	0.6	32.50	8.50
25	13152	8	4.0	65.63	15.46
26	10579	5	0.7	15.40	5.46
27	-	0	0.0	-	-
Mean	11957.67	6.42	4.55	91.11	70.06
SD	2971.59	4.62	3.70	51.84	76.42

Table 1: Information about the data collected from each participant.

4.1.2 Data Preprocessing for the ResEmoteNet Model

The ResEmoteNet model requires image-based inputs, so no action units were used when training the model. First, the video recordings were split into individual frames. Using the face detection function of Python's mediapipe library, the face was detected in each image. This is done using the BlazeFace machine learning model to detect the location of the faces, as well as some key facial points [73]. Each image where a face was detected was then cropped to ensure that the model only focused on the relevant regions of the image. The resulting images were then resized to 64x64 pixels to ensure uniform input dimensions and converted to grayscale to reduce the computational complexity. This conversion to grayscale was also present in the images that the ResEmoteNet was originally trained on. To improve model generalization, images were randomly flipped horizontally to increase the diversity of training samples. This ensures that the model can generalize better to new situations rather than favoring the more frequent orientation.

4.1.3 Data Splitting

The preprocessed data of both models was split into training, testing, and validation sets. We compared the models' performance for multiple datasets obtained through different strategies of sampling data and splitting data for training and testing purposes (see Table 2).
The first data sampling approach that we used was stratified sampling. This sampling method ensures that the percentage of samples for each class was identical in each of the train, test, and validation sets. The second approach was undersampling. With this data sample method, the classes in the data were balanced by reducing the number of non-confused data samples to ensure there was an equal amount of confused data samples.

Dataset	Participants in test set	Data sample method	Confused data percentage (LSTM / ResEmoteNet)
All-Str	All participants	Stratified sampling	3.6% / 4.55%
Sub-Str	Subject-based split	Stratified sampling	3.6% / 4.55%
Sub-Und	Subject-based split	Undersampling	50% / 50%

Table 2: The datasets used to train the models.

These two approaches were tested by combining the data of all participants together and separately by using a subject-based split. The subject-based split ensures that no data from the same individual appears in both training and test sets. This is done to prevent overfitting to person-specific features and enhance generalization. K-fold cross-validation was used to better estimate model performance. The data consists of 24 participants and each fold contains 4 different individuals resulting in 6-fold cross-validation.

4.2 LSTM Model

The LSTM model we used was based on a model by Borges et al. [56]. This model was originally designed to classify confusion as well as positive, negative, and neutral emotions in younger adults based on facial action units.
Our version of the LSTM model (see Figure 9) consists of a Long-Short Term Memory layer with 17 input nodes each receiving input from a single Facial Action Unit. This input consists of the

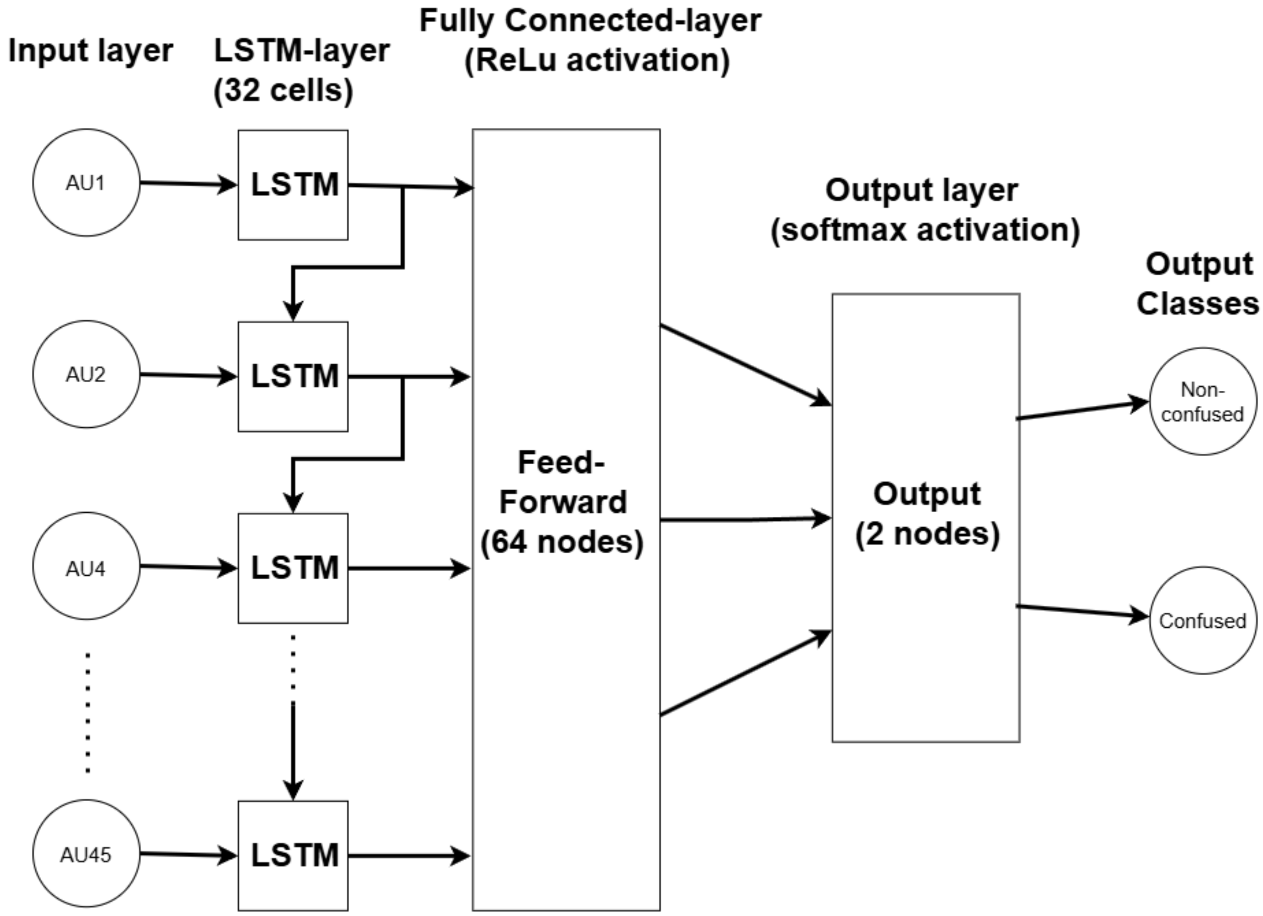


Figure 9: Schematic of the LSTM-network.

AU's activation during 30 timesteps. The LSTM-layer has 32 hidden cell nodes. It feeds into a feed-forward fully connected network layer with 64 nodes using the ReLu activation function. This layer feeds into the final output layer consisting of 2 output nodes that result in a binary classification of either Confused or Non-confused. The output layer uses a softmax activation. Between the LSTM and feed-forward layers, a dropout function, with a rate of 0.25, is performed to prevent overfitting. A dropout function, with a rate of 0.25, is also performed between the feed-forward and output layers.

To create our version of this model, the architecture is slightly changed from the original model to adapt to the difference in the number of classes and features. The original version of the model has 4 output nodes, as they classify confusion as well as positive, negative, and neutral emotions, while our version of the model only differentiates between confused and non-confused emotions. We also use different software to extract the facial action units. Borges et al. use the FaceReader software which extracts more action units than OpenFace, so their model has more inputs (20 instead of 17).

The model was trained using supervised learning on the datasets mentioned in Table 2. The training was conducted using the TensorFlow framework. We used the Categorical Cross-entropy loss function and optimized the model using the Adam optimizer. The model was trained for 50 epochs with a batch size of 32. Early stopping was applied based on the validation loss with a patience of 10 epochs to prevent overfitting.

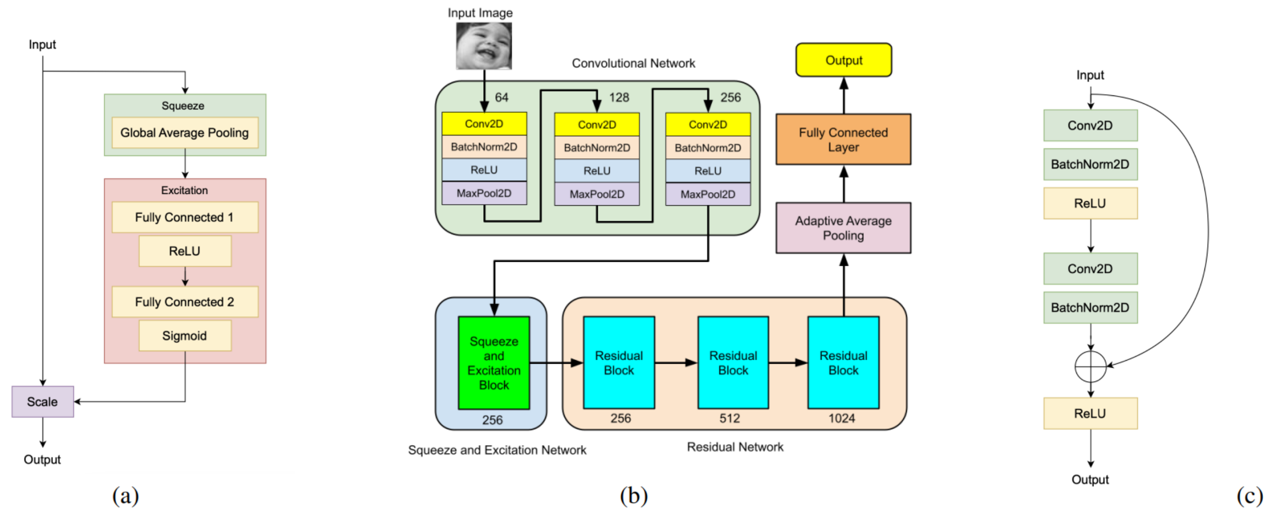


Figure 10: (a) Architecture of the Squeeze and Excitation Block, (b) The overall architecture of the original ResEmoteNet model, (c) Architecture of a Residual Network block. (Figure taken from [70])

4.3 ResEmoteNet model

In addition to training an LSTM model, we also trained an end-to-end model using image inputs. To reduce the amount of data needed to train this model, we use transfer learning on a pre-trained emotion recognition model. Specifically, the ResEmoteNet model [70] shown to have state-of-the-art performance in the facial emotion recognition domain.
The ResEmoteNet framework (see Figure 10) is a deep learning-based framework originally designed to recognize 7 basic emotions from still images of faces. The model consists of a Convolutional Neural Network (CNN) block, a Squeeze and Excitation (SE) block, and multiple Residual Network (ResNet) blocks. These blocks were chosen to efficiently learn spatial and contextual information from facial images, enhancing facial emotion recognition. The CNN module forms the foundation of the feature extraction process. It comprises of three convolutional layers followed by batch normalization and max-pooling operations to reduce spatial dimensions. The SE block further refines the feature representations by modeling the relationships between convolutional channels. It uses global average pooling to compress each channel's spatial data into a global descriptor after which a sigmoid-activated gating mechanism captures the channel dependencies. This approach allows the Squeeze and Excitation Network to highlight the importance of each input element for the network's output. ResNet blocks are added to address the vanishing and exploding gradient problems commonly found in neural networks. Additionally, the model employs Adaptive Average Pooling (AAP) to maintain consistent output dimensions.

We applied transfer learning to this model by fine-tuning the pre-trained network while modifying its output to suit our confusion classification task. Specifically, we did this by removing the original output layer and replacing it with a custom fully connected layer. This custom layer consists of a linear layer with 128 nodes followed by ReLU activation to improve learning capacity. A dropout layer with a rate of 0.5 is added to reduce overfitting. Finally, a linear output layer with 2 output nodes and a softmax activation is used to predict the two classes. The lower layers of the model are frozen to preserve general feature representations, allowing the new model to adapt to our facial emotion recognition task without having to relearn these features.

5 Results

In this section, we first present the results of training the LSTM and ResEmoteNet models for confusion recognition. To do this, we report the overall classification performance of the four most interesting model and dataset combinations using standard metrics, including accuracy, and F1-score. We also include confusion matrix analyses to examine how emotions were most frequently misclassified. After this, we present questionnaire results. Next, we provide an analysis of these results and potential factors that influence the training of the confusion recognition models.

5.1 Questionnaire results

The survey included several questions from the Godspeed questionnaire [68] to provide a broader context on the attitude of participants during their interaction with the social robot. While these results might not directly inform the central research question, they are reported here for completeness. The results of the questions regarding each construct are combined in Table 3, while the results of each individual question can be found in Appendix G.
Overall, participants expressed a positive feeling about the robot's likeability. This means that participants presumably had primarily positive emotions during the word game with the robot. Furthermore, the 3.84 score regarding perceived intelligence of the robot indicates that it is likely that the robot often played the word game correctly.

In addition to these findings, the participants reported a mean perceived confusion of 3.15 (3 = neutral on the Likert scale). This indicates that they were confused during the human-robot interaction as was intended. However, there was a considerable variation in answers with some participants indicating very little confusion while other participants indicated more confusion. Some participants also had more difficulty than other participants when trying to understand the robot. This sometimes leads to confusion as well as was mentioned in Section 3.3.3.

Construct	Mean	SD	Median	IQR
Likeability	4.04	0.78	4	[3,5]
Perceived Intelligence	3.84	0.84	4	[3,5]
Perceived Confusion	3.15	1.11	4	[2,4]
Difficulty in Understanding	2.70	1.15	3	[2,4]

Table 3: The results of the questionnaire with combined scores for the questions regarding each construct.

5.2 Model Performance

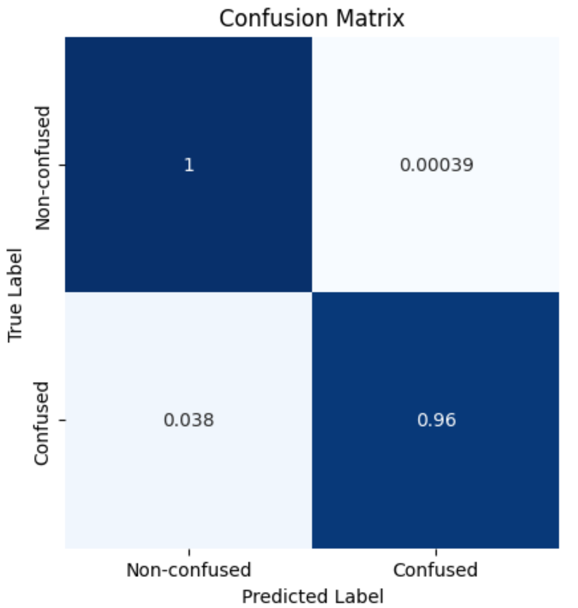
To evaluate the effectiveness of the tested models in classifying confusion, we evaluated their performance using standard classification metrics, including accuracy and F1 score. Table 4 presents a summary of the results for four models. The results of other models can be found in Appendix H.
The LSTM 1 model shows the performance result of testing the model on participants present in the training data. This model achieved the highest accuracy and F1 score, outperforming the other models. LSTM 2, LSTM 3, and the ResEmoteNet model show the average results of 6-fold cross-validation where the test data of each fold consists of different participants that were not

Model	Accuracy	F1-score
LSTM 1 (All-Str)	0.9982	0.9752
LSTM 2 (Sub-Str)	0.9271	0.0798
LSTM 3 (Sub-Und)	0.5726	0.4212
ResEmoteNet (Sub-Und)	0.5272	0.2401

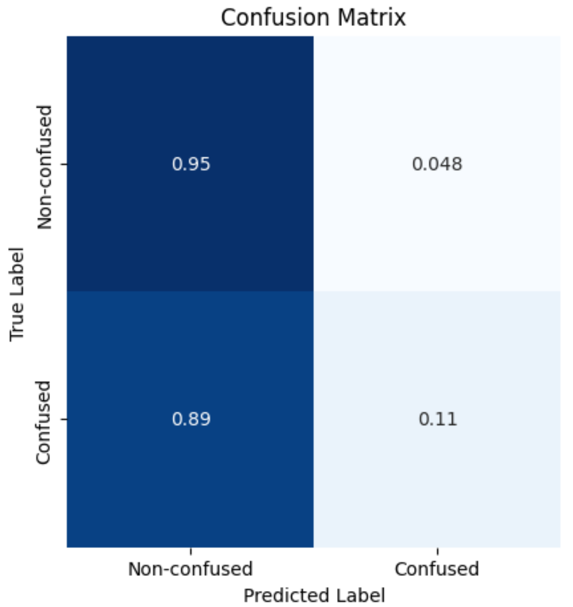
Table 4: Model performance results. The dataset for training (as described in Table 2) are shown between the brackets.

present in the training data. These models achieved a lower performance, indicating that the model might not be able to generalize to new people.

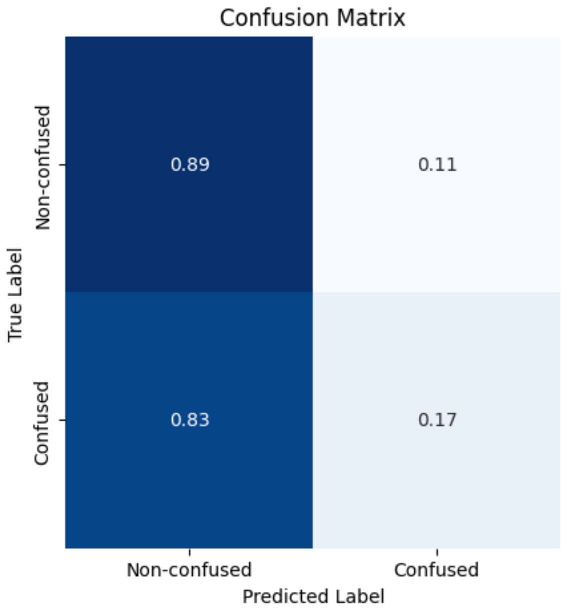
The confusion matrices of these four models are shown in Figure 11. The LSTM 1 model is able to correctly classify most samples. However, the other models misclassify a high amount of Confused samples as Non-confused.
It is clear that both the LSTM 2, LSTM 3, and the ResEmoteNet models are not able to accurately predict confusion based on facial expressions. While balancing the data by using undersampling (the LSTM 3 and ResEmoteNet models) improved the accuracy of predicting confusion, most samples still get misclassified as non-confused.



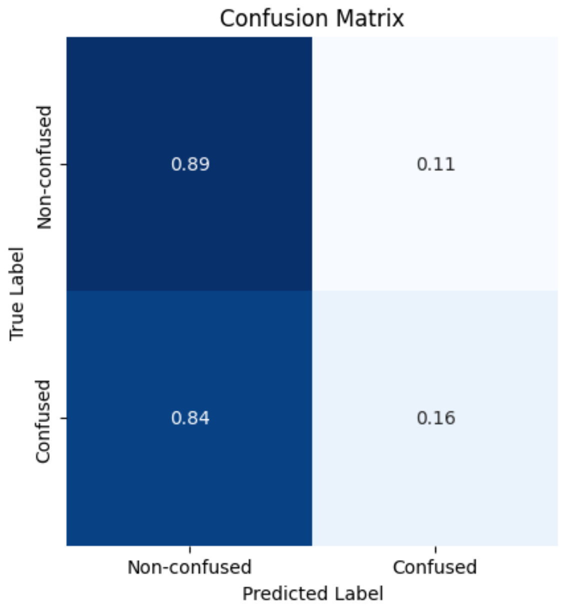
(a) LSTM 1: Both the Confused and Non-confused facial expressions are often correctly classified.



(b) LSTM 2: The Confused facial expressions are often misclassified as Non-confused facial expressions.



(c) LSTM 3: Confused facial expressions are still misclassified as Non-confused facial expressions. This is to a lesser extend than for LSTM2, however, more Non-confused facial expressions are misclassified as Confused facial expressions.



(d) ResEmoteNet: A similar result to LSTM3 where Confused facial expressions are misclassified as Non-confused facial expressions.

Figure 11: The normalized Confusion Matrices for the models mentioned in Table 4.

5.3 Further Analysis

To find out why the models' performance was worse than expected, we analyzed potential factors contributing to the lack of success.

5.3.1 Data Limitations

The collected data contains an uneven distribution of emotions, with the confused class being significantly underrepresented in comparison to the non-confused class. Using stratified sampling on the data resulted in less than 3.6% confused samples for the LSTM models. This class imbalance likely causes the LSTM model trained on stratified data to favor the non-confused class (as seen in Figure 11b), leading to lower recall for the minority class. However, the models trained on balanced data still misclassified many confused samples as non-confused. The balanced dataset was created by undersampling the non-confused samples in the original data. Resulting in less training data overall.

Furthermore, the non-confused class also consists of multiple different emotions not individually annotated. As such, it could be the case that samples in the confused class overlap with some positive or negative emotions present in the non-confused class. Thus, performance might be improved by annotating the data more thoroughly, adding extra emotions for the model to learn.
To test this, we attempted to automatically annotate the neutral, positive (happy), and negative (sad, angry, scared, disgusted) emotions within the current non-confused class. The same technique was also done by Borges et al. [56]. Following their approach, the surprised emotion is discarded during automatic annotation due to its ambiguous valence.

One model that we used to automatically annotate the non-confused emotions was the DeepFace model [74], however, this model classified all facial expressions as surprise. We also tried to use the original ResEmoteNet model [70] to annotate the non-confused samples. This model was able classify multiple different emotions in the non-confused samples with 5,2% positive, 91,3% negative, and 3,5% neutral emotions. This distribution may not be correct as none of the participants have indicated or shown this much displeasure during the data collection experiment. The high number of surprised and negative emotions is likely because these annotation models were originally only trained on younger people. Thus, they are likely to fail to generalize to older adults.
Training the ResEmoteNet model on the automatically annotated data did not improve performance, resulting in an accuracy of 50% in detecting confusion when using the dataset with undersampling and a subject-based train and test data split.

5.3.2 Facial Action Units

The LSTM model makes use of 17 facial action units to classify confusion and non-confusion. Examining the values of these AUs in samples for the Confused and Non-confused classes, we found that the range of values was larger for the non-confused samples than for the confused samples. All AU values in the confused samples overlap with the AU values in the non-confused samples for all 17 of the AUs. This might explain why the model misclassifies Confused samples as Non-confused.

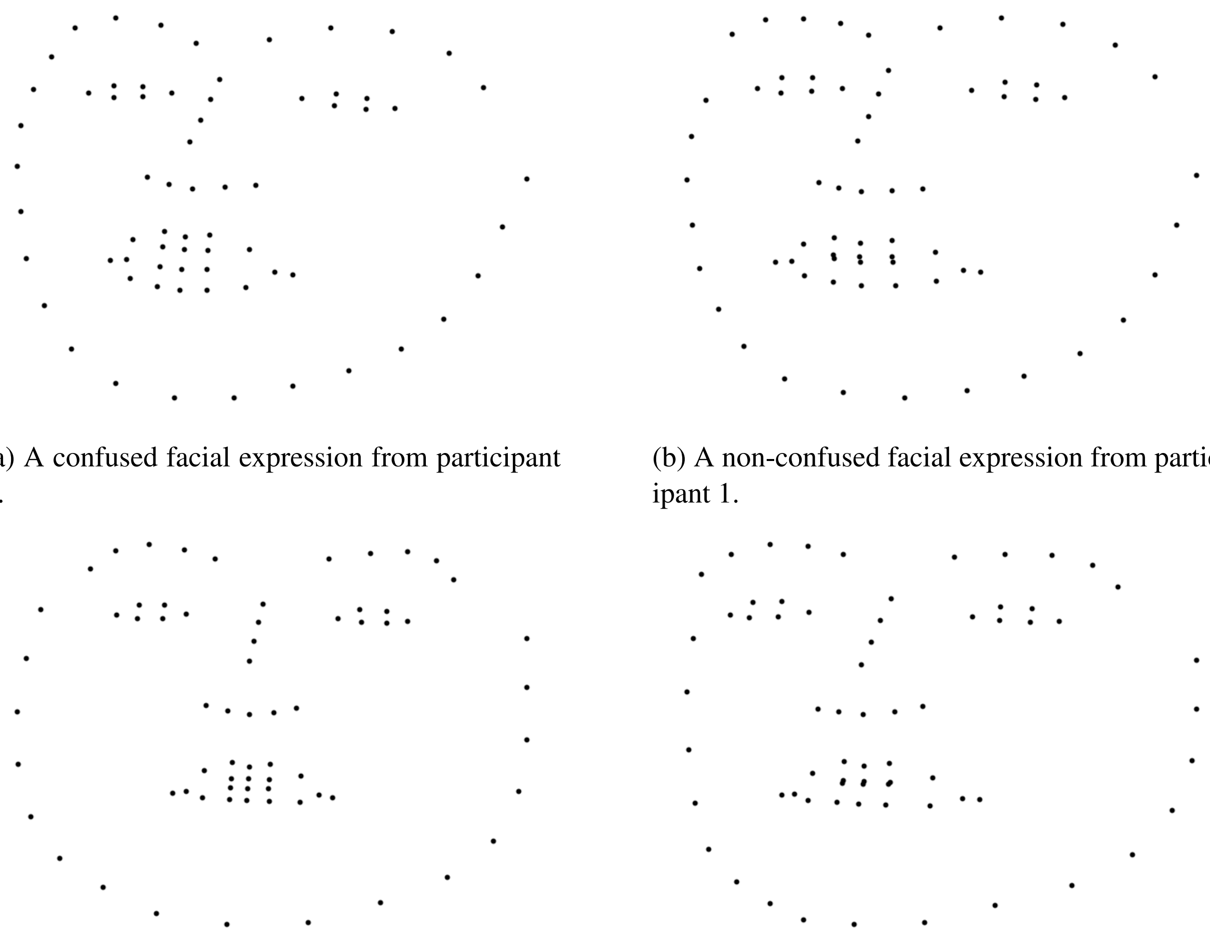
Figure 12 shows a representation of some of the facial expressions that participants made during their interaction with the social robot. Participant 24 sometimes raised their eyebrows during confused facial expressions (Figure 12e), which is an example of AU1 Inner Brow Raiser. Other participants, such as Participant 1, exhibit an example of AU27 (Mouth Stretch) when confused (Figure 12a) while

this is not present in their non-confused facial expressions (Figure 12b).
However, this open-mouth expression has been found in non-confused facial expressions of other participants, such as Participant 9 (Figure 12c). So, there is an overlap between action units for confused and non-confused facial expressions between different participants. As mentioned before, it is also possible that there is an overlap between confused and non-confused facial expressions within a single participant. An example of this is participant 22 (see Figure 12g and 12h).

Similarly to the ResEmoteNet and DeepFace models mentioned before, the OpenFace model, used to obtain the AUs, might have unsatisfactory performance in extracting Facial Action Units of elderly individuals as it has not been trained on this age demographic.

5.3.3 Feature Importance

To better understand which facial action units contributed most to the classification of confusion, we conducted a feature importance analysis using a Random Forest classifier. This analysis was done to identify the most informative features among the facial action units to distinguish confused from non-confused facial expressions. Furthermore, it can also give us insight into why this classification proves more challenging across different participants.



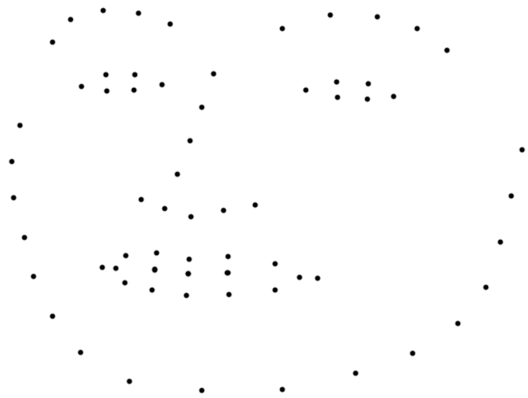
(a) A confused facial expression from participant 1.

(b) A non-confused facial expression from participant 1.

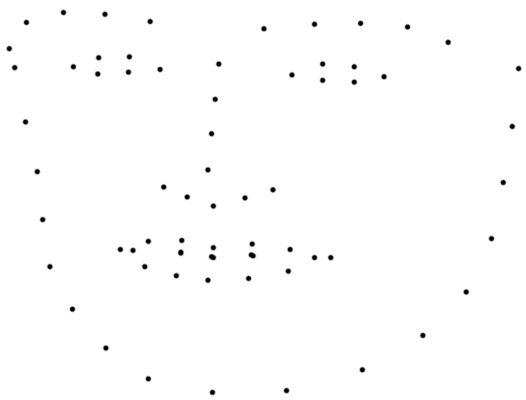
(c) A non-confused facial expression from participant 9.

(d) A different non-confused facial expression from participant 9.

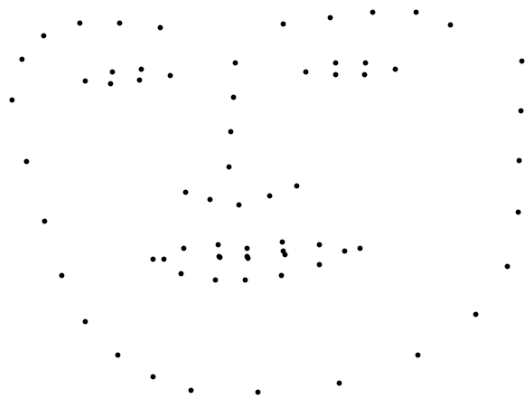
(e) A confused facial expression from participant 24.



(f) A non-confused facial expression from participant 24.



(g) A confused facial expression from participant 22.



(h) A non-confused facial expression from participant 22.

Figure 12: Visualizations of keypoints of the faces of participants during interaction with the social robot.

Feature importance was calculated using a random forest model based on the mean decrease in Gini impurity. This method measures feature importance derived from how much the impurity within a node of a decision tree decreases when the feature is used to split the data. A higher importance score indicates that a feature was more influential in the model's decision-making process.
Our findings are shown in Table 5, which shows the average feature importance score and standard deviation for each AU, and Table 8 (in Appendix I) which shows the average feature importance score for each AU across all participants.
Table 5 shows that AU10, AU4, and AU14 are the most influential features to predict the model's output. While AU5, AU2, AU9, and AU23 have the lowest contributions. However, these features still have an importance score of more than 0.0 indicating that they are not completely irrelevant. The small values for the standard deviations further suggest that the feature importance are stable across the trees in the random forest.

Not all features found to be important for confusion recognition in previous work were equally important for training our LSTM models. AU23 (Lip Tightener), for example, was not found to be significant for confusion recognition. However, AU4 (Brow Lower) and AU12 (Lip Corner Puller) are found to be important. The most important facial action units for training our model are AU10 (Upper Lip Raiser), AU4, and AU14 (Dimpler). There were also differences in the importance of the features for each individual (see Table 8 in Appendix I).

Feature	Importance (mean)	Standard Deviation
AU10	0.0967	0.0060
AU4	0.0907	0.0063
AU14	0.0833	0.0065
AU6	0.0776	0.0050
AU12	0.0728	0.0054
AU17	0.0722	0.0062
AU25	0.0689	0.0060
AU7	0.0657	0.0048
AU26	0.0639	0.0052
AU15	0.0471	0.0035
AU1	0.0468	0.0046
AU20	0.0407	0.0039
AU45	0.0401	0.0029
AU23	0.0394	0.0034
AU9	0.0353	0.0030
AU2	0.0322	0.0030
AU5	0.0264	0.0021

Table 5: The importance of the facial action unit features when training the LSTM model. These feature importance were obtained from a random forest model using the mean decrease in Gini impurity.

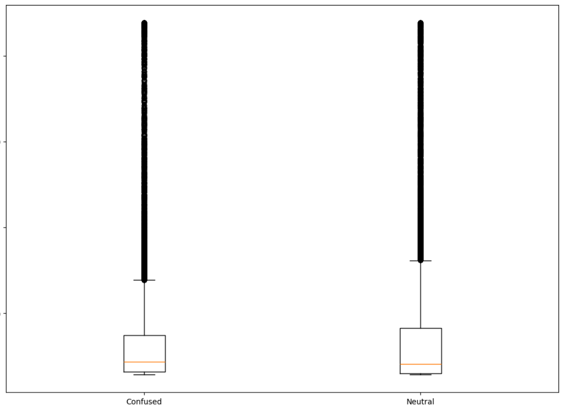
5.3.4 Sample Difficulty Estimation

We further investigated the ResEmoteNet model's performance by using individual sample loss during training to estimate the difficulty per sample following the method of Arriaga et al. [75]. The difficulty score is caculated based on the accumulated loss per epoch, where samples with a higher score are more challenging to classify. The most difficult samples consisted of blurred or off-center faces. However, there were also clear and correctly framed samples that received high difficulty scores.

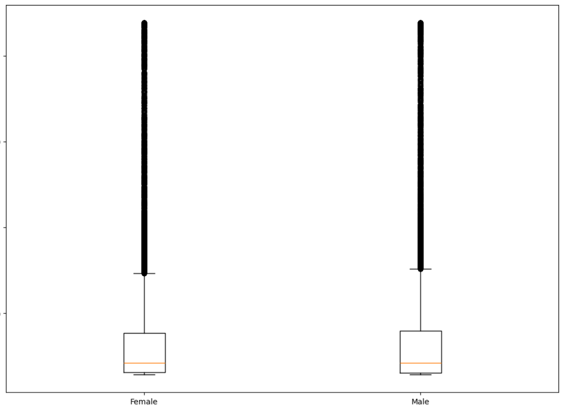
There was no apparent difference in sample difficulty between male and female participants (Figure 13b). Surprisingly, there was also no clear distinction in difficulty between samples in the confused and non-confused classes (Figure 13a). Nevertheless, there was a difference between the difficulty of samples for different ages (Figure 13c). However, the same difference was visible between individual participants (Figure 13d). There are not enough participants in each age category to conclude that age is a factor in the difficulty of samples rather than these differences resulting from the individual differences between participants. Especially since there is no linear increase or decline in difficulty when age increases.

Occlusions such as glasses can be challenging when classifying facial expressions due to decreased visibility for some facial features. However, most participants wore glasses and there was no clear reduction in difficulty for participants who did not. Facial hair and wrinkles showed similar results with no clear difference in difficulty between samples of participants that had more wrinkles or facial hair in comparison to participants that did not have this.

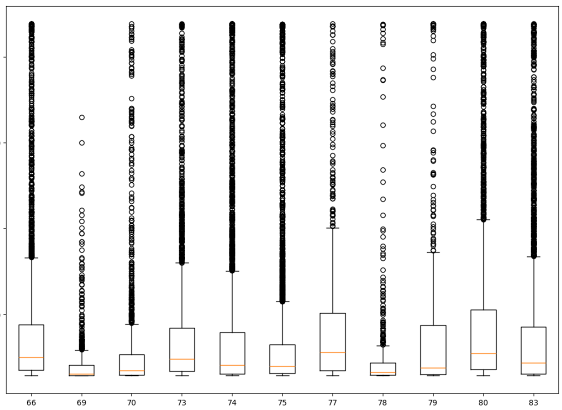
So, there is no obvious evidence why the model was unable to train on the data correctly. This means that the reason might instead be due to differences between younger and older adults such as reduced facial muscle movement and skin texture changes, as previously reported in literature [18].



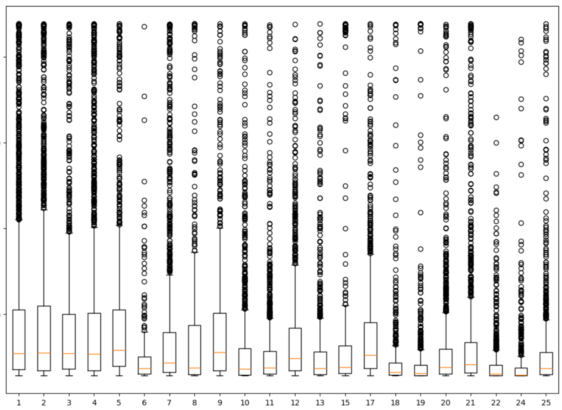
(a) A boxplot showing the difficulty of samples based on its training class with the x-axis representing the class.



(b) A boxplot showing the difficulty of samples based on participants' gender with the x-axis representing the gender.



(c) A boxplot showing the difficulty of samples based on age with the x-axis representing the age in years.



(d) A boxplot showing the difficulty of samples based on the participant with the x-axis representing the participant number.

Figure 13: Boxplots showing the distribution of the difficulty of different samples with the data category on the x-axis and the action score value on the y-axis.

6 Discussion

In this study, we attempted to recognize confusion during robot-human interaction in older adults using facial expressions. To accomplish this, we first collected a new dataset of robot-human conversations, after which two machine learning models were trained on this data.

6.1 Model Performance

Of the various models that have been trained, only one model was able to accurately classify confusion. This model differed from the other models in that it was trained on data from all the participants. This meant that during testing it had already encountered data representing both confused and non-confused facial expressions from all participants. In Table 4 and Figure 11a we can see that this model achieved a recognition accuracy of more than 99% and that confusion could be detected with an accuracy of 96%.
The other models were all tested with data from participants not present in the model's training data. These models all performed worse on confusion recognition as confusion was often misclassified as non-confusion (see Figure 11). These results indicate that we have not been able to train a model to generalize to new data correctly.

Earlier studies have found that variations in training data such as partial occlusions, head pose, and subject identity are challenging when using machine learning for automatic facial emotion recognition [76]. To combat this, a large and varied dataset is needed to ensure that all the factors of variations are covered, especially for models that are used in real-world situations. Many of the elderly participants that participated in this study wore glasses during data collection and some of the men had facial hair, so we investigated whether this might worsen the performance of the classification model. During sample difficulty analysis, we found no evidence that data samples of participants with facial hair or glasses were more difficult for the model than other samples. So, these facial occlusions likely did not impede confusion recognition in our models. However, the samples that were blurred, off-center, or partially obstructed were often more difficult than other samples of the same participants.

Sample difficulty estimation further found that some individual participants had more difficult samples than other participants (see Figure 13d). There was no clear distinction between these participants and the other participants. However, there might be differences in how they express confusion compared to the other participants that are not visible to human observers.

These individual differences were also present when looking at the feature importance estimation for the LSTM models (see Appendix I). This was expected to some extent as previous research has found that some people are more expressive in the mouth area while other people are more expressive in the eyebrow region of their face during confusion [51]. These differences could explain why the model has difficulties generalizing to new participants, however, there are still many similarities in the data of participants as well.

These individual variations between participants and the higher classification accuracy for the model that was tested on data of participants that it had seen before, seem to suggest that models work best when person-specific modeling is used.

6.2 Insights from Questionnaire Data

After the data collection experiment participants filled out a questionnaire about their experiences with both the robot and the overall experiment (see Table 3).
The participants indicated a high likability of the robot, which suggests that the interaction during the word game was mainly positive. Participants further indicated a high perceived intelligence of the social robot. This suggests that the robot often played the word game correctly.

However, most participants did indicate that they thought the conversation was sometimes confusing. The extent to which confusion was perceived varies significantly between participants. Some indicated that they experienced hardly any confusion during the conversation while others indicated a lot of confusion. The participants who indicated a lot of confusion might have shown negative emotions such as confusion more openly during the conversation since they were willing to criticize the social robot in the questionnaire. The participants who indicated less confusion, on the other hand, might have exhibited prosocial behavior.They may have acted like the interaction was less confusing than how they experienced it. This could be done to emphasize with the experimenter. It could also be done due to the idea that confusion was not supposed to happen and thus their own fault.

Participants also indicated that the robot was sometimes difficult to understand. This was mainly due to the mechanical voice of the robot. This sometimes caused more confusion during the conversation when a participant was unable to hear what the robot had said (see Transcript 4 and 5), however, it could also have the opposite effect when the participants could not hear the confusion inducing utterance of the robot correctly (see Transcript 3).

6.3 Dataset and Annotation Considerations

We have been able to collect a novel dataset of confused and non-confused facial expressions of older adults during a conversation with a social robot. There is a significant gap in existing confusion recognition research focused on older adults interacting with a social robot. Most publicly available datasets are heavily biased toward younger populations and do not contain confused facial expressions, which limits the development of new models. Our dataset captures naturally occurring expressions of confusion during unscripted human-robot interaction with elderly people. As such, this dataset is a step toward building confusion recognition models that can apply to elderly-care technologies in the future.

Participants were tasked to play a simple word game with the social robot, rather than have a conversation. This gamified approach was done in an attempt to elicit more (genuine) facial expressions from participants. However, an external camera recorded the participants during interaction with the robot. This might have negatively influenced the way they expressed themselves, as they might have been self-conscious of the fact that they were being recorded.

Not all participants were straightforward to annotate. There were instances in the recording where participants audibly requested clarification from the robot without any visible change in their facial expressions. These were not included in the confused class and instead were classified as non-confused facial expressions. However, it could be the case that there are small changes to the facial expressions which are not visible to human observers.

There are some limitations to the new dataset. The dataset is relatively small, especially considering the limited number of samples for the confused facial expressions class. This likely contributes to the models' limited ability to generalize when detecting confused facial expressions. A larger dataset will likely improve the models' performance. If the size and diversity of the data in the dataset is increased, the models will encounter a broader range of facial expressions which can reduce overfitting.

There is also only a small number of emotion classes. The non-confused class contains all the data where participants did not show a clear confused facial expression. This means that multiple different emotions are present in this data class. In the future, these emotions could be further annotated to potentially improve classification performance.

To train the LSTM models, we used OpenFace to extract the facial action units from the video data. However, the OpenFace model is not trained on elderly people and might have extracted the facial action units incorrectly meaning that the LSTM models were trained on flawed data.

6.4 Emotion Complexity and Class Overlap

The training of a model to recognize confusion is complicated. Assuming that participants are willing to openly show their emotions during the data collection experiment, they may show a similar emotion to confusion during (a big part of) the whole conversation. Concentration, for example, shares a lot of similarities with confusion [61]. As participants concentrated on the game and on what the robot said as well as remembering how to interact with the robot, they might have often had a concentrated facial expression.

Looking at the values of various facial action units in both the confused and non-confused classes in the data, we found that there is an overlap between their values. Almost all AU activation values for confused facial expressions fall into the range of activation during non-confused facial expressions. This further implies that there is an overlap between the expression of confusion and other emotions present in the non-confused data.

As mentioned, the non-confused class contains multiple different emotions that might have similarities to confused facial expressions. We have tried to separate these emotions into positive, negative, and neutral following Borges et al.'s approach [56] as well as the 7 basic emotions. However, we did not see an improvement in the model's performance. In comparison, Borges et al. [56] have been able to achieve an accuracy of 83,7% when recognizing confusion based on facial expressions. The LSTM model in our study was based on the LSTM model used by Borges et al., however, their participants are university students and thus younger.

To separate the non-confused class into different emotions, we used a pre-trained model to automatically annotate these other emotions. So, it is possible that these emotions were annotated incorrectly as most present-day emotion classification models are not trained on elderly people. One of the models that we attempted to use for automatic annotation, for example, is the DeepFace model [74]. This model classified all facial expressions as surprise and could thus not be used. The model that we eventually used to annotate the emotions in the non-confused class of the data was the original version of the ResEmoteNet model. This model was able to differentiate between different expressions, however, most of the data (around 90%) was annotated as negative emotions. This is not in accordance with what was observed during the data collection experiment or the general experience of participants as indicated during the survey, where participants indicated a high likeability, and debrief at the end of the experiment.

6.5 Facial Action Units

In the literature review (Section 2.4.3) we discussed which Facial Action Units were important for recognizing confusion in previous studies. These were AU4 and AU7 around the eye area and AU12 and AU23 around the mouth area. However, an analysis of the importance of the action units used in this study found that our most significant action units when training the LSTM model differed from these four (see Table 5). Out of these facial action units, AU4 was the most important, followed by AU12 and AU7. However, AU23 was not found to be important for recognizing confusion. However, AU10, AU14, AU6, and AU17 were among the most important action units, while this was not the case in previous research. The difference in which AUs were important to recognize confusion might be due to the difference between the data that was used in this research and other studies. Cumbal et al. [53], for example, used data of participants aged between 21 and 51. Instead, we specifically investigate the possibility of recognizing confusion in older adults.

The difference in AU importance could also be caused by our facial action unit extraction through the use of the OpenFace model. As mentioned before, it might be the case that this model is not able to correctly extract facial action units for elderly people due to a lack of training on older adults' facial expressions. Next to this, older adults have reduced activation of facial muscle as mentioned by Grondhuis et al. [37]. This might have hindered the extraction of facial action units as well since the reduced muscle movement might have reduced the activation of different action units.

6.6 Challenges in Emotion Recognition in Older Adults

The current research and collected dataset specifically focus on the recognition of confusion in older adults. However, emotion recognition in the elderly is more difficult than emotion recognition in younger people.

Grondhuis et al. [37] have found that it is not only the extra wrinkles on older faces that hinder facial emotion recognition in the elderly. Rather, emotion recognition performance is mainly impeded by a decline in the facial muscles' ability to express emotions. Age-related physical changes still negatively impact facial expression recognition, however, its influence is a lot weaker than that of age-related decline in facial muscle activity. Grondhuis' study suggests that even models fully trained on videos of older adults' facial expressions might not be able to obtain a satisfactory performance as elderly individuals might not show emotions through their facial expressions as much as younger people do.

The earlier mentioned sample difficulty estimation showed no clear trend of increasing sample difficulty for older ages (Figure 13c), which could mean that all participants in the study have these reduced muscle movements. Most other models are trained on younger adults who show emotions more clearly. So, the existing connections in the pre-trained ResEmoteNet model might not have been as helpful for the new data as expected since these connections might not recognize smaller changes in expressions.
This is substantiated by our findings that models trained on data from younger people seemed to

struggle with classifying the facial expressions of the elderly participants. Thus, the facial action units extracted using the OpenFace model might be less accurate than desired to train our LSTM models.

Another factor that could influence confusion recognition is that participants might have engaged in prosocial behavior [38]. This means that they tried to hide their negative emotions during misunderstandings with the social robot. This could be caused by not wanting to play the word game incorrectly or not wanting to show anything going wrong during the experiment. Earlier research suggests that older adults partake in prosocial behavior more often than younger adults [40, 44, 45].

6.7 Future Work

Our current research has collected a new dataset of confusion in older adults during conversations with a social robot. We have trained two different types of models to recognize confusion based on facial expressions. These models were not able to generalize to new participants, however, performance with participants before seen in the data was promising.

In the future, it would be useful to utilize the other information present in the new dataset. During the original annotation of the videos, we had to make use of the context and vocal information as well as the visual videos to be able to accurately annotate the data. The current models may be unable to gather enough information from only facial expressions. So, in future research, a multimodal model can be trained to improve generalization and overall classification performance. This model can make use of facial expressions as well as speech, body gestures, and the context of the conversation.
Earlier research by Busso [55] has shown that models trained on the combination of facial expressions and acoustic information demonstrate improved performance and robustness of emotion recognition over models that are only trained on one of these modalities. However, their research does not include confusion in the emotions that are recognized. Thus, we can't be sure that it will improve recognition of confusion.

Future research could also collect more data to increase the size of the dataset. Collecting more varied data to train a model on might improve performance by reducing overfitting and increasing generalization capabilities. The dataset could also be improved by annotating the different emotions present in the non-confused class. This could improve model performance by reducing overlap between confusion and the other classes. Even if this doesn't improve performance, it might still give us more insight into what specific emotions, if any, are similar to confusion.

As mentioned before, models were unable to generalize to new participants. However, performance was still significant for people that the model had previously encountered during training. Thus, these models could be used as personal confusion detection models. It could be interesting to look into the possibilities of individual-specific training. Especially how these could work in practice as having to annotate a video containing confused facial expressions for each new user would be very labor intensive. Adaptive models that can be changed to handle individual variability might help in this regard.
The current study has brought to light a mostly unexplored topic in the form of confusion detection in the elderly. However, we have not been able to successfully recognize confusion based on facial expressions. This can be due to various reasons such as age-related decline in facial muscle activation, lack of data and models regarding the elderly, or incomplete annotations. Future research could

collect more data and train models on different modalities. This would allow us to find out whether the challenges in recognizing confusion are due to the focus on facial expressions, the experimental setup and collected data, or the focus on older adults who might express emotions less or differently from younger adults.

7 Conclusion

This study explored the possibility of recognizing confusion based on the facial expressions of older adults during a conversation with a social robot. To accomplish this, a new dataset was collected, consisting of facial expressions recorded during real-time conversations between elderly participants and a social robot.

27 elderly individuals played a word game with a social robot. Two techniques to induce confusion were utilized to guarantee that misunderstandings leading to confusion would occur. One of these confusion-inducing techniques was adding strange, unexpected sentences into the conversation. The other confusion-inducing technique was increasing the speech rate of the robot.

Two machine learning models were trained and evaluated using this new dataset. One of these models is an LSTM model that was trained using Facial Action Units (AUs). It also incorporates the temporal element of the dataset. The second model is an end-to-end deep learning model trained using transfer learning on still images extracted from the video data.

Although the models were able to perform well on data from individuals previously encountered during training, they failed to generalize effectively to new participants. Even after the dataset was balanced through undersampling, confusion was frequently misclassified as non-confusion. There was also considerable individual variety in how confusion was expressed. This can be seen from the importance of different AUs for different participants. These findings suggest that confusion detection in older adults is highly individualized.

Several factors may explain these outcomes. Older adults often show reduced facial muscle movement [37], which can make expressions less distinct. Additionally, prosocial behavior may lead some individuals to suppress negative emotions such as confusion [38]. Lastly, the overlap between confusion and other emotions may further complicate classification. These challenges suggest that existing emotion recognition models, particularly those not trained on elderly populations, may be insufficient in such contexts.

These findings demonstrate the importance of further research into confusion detection from older adults. Future work could explore a multimodal approach, combining facial features with other signals such as speech or gestures. This might lessen the impact of reduced facial muscle movement as well as improve the robustness of the model. Expanding the annotation framework to include a broader range of emotion categories could give more insight into which emotions overlap with confusion and cause difficulties during model training. Finally, the development of personalized emotion recognition models, potentially adapting to individual facial expression patterns, could be considered to better accommodate individual variation.

This study emphasizes the importance of emotion recognition research pertaining to older adults. The new dataset consisting of facial expressions of older adults during interaction with a social robot can contribute to future research. This promotes improvements in resolving misunderstandings during conversations with social robots and outlines various directions to improve human-robot interactions with older adults to address the future needs of elderly care in our society.

Bibliography

[1] M. Steverson, "Ageing and health." `https://www.who.int/news-room/fact-sheets/detail/ageing-and-health`, 2024. Accessed: 15-02-2025.

[2] C. Breazeal, A. K. Ostrowski, N. Singh, and H. W. Park, "Designing social robots for older adults," in *The Bridge - Spring Bridge on Technologies for Aging*, vol. 49, the National Academy of Engineering, 2019.

[3] E. Broadbent, K. Loveys, G. Ilan, G. Chen, M. M. Chilukuri, S. G. Boardman, P. M. Doraiswamy, and D. Skuler, "Elliq, an ai-driven social robot to alleviate loneliness: Progress and lessons learned," *JAR life*, vol. 13, p. 22–28, 2024.

[4] C. Berridge, Y. Zhou, J. M. Robillard, and J. Kaye, "Companion robots to mitigate loneliness among older adults: Perceptions of benefit and possible deception," *Frontiers in Psychology*, vol. 14, 2023.

[5] M. M. Reimann, F. A. Kunneman, C. Oertel, and K. V. Hindriks, "A survey on dialogue management in human-robot interaction," *ACM Transactions on Human-Robot Interaction*, 2024.

[6] A. Mousa and A. Nasser, "Enhancing human-robot interaction through group emotion recognition," *Iraqi Journal for Computers and Informatics*, vol. 49, pp. 111–119, 12 2023.

[7] B. Arend and P. Sunnen, "Coping with turn-taking:investigating breakdowns in human-robot interaction from a conversation analysis (ca) perspective," 2017.

[8] S. Kopp and N. Krämer, "Revisiting human-agent communication: The importance of joint co-construction and understanding mental states," *Frontiers in Psychology*, vol. 12, 2021.

[9] C. D. Salzman and S. Fusi, "Emotion, cognition, and mental state representation in amygdala and prefrontal cortex," *Annual review of neuroscience*, vol. 33, pp. 173–202, 2010.

[10] M. Zhao, "The emotion recognition in psychology of human-robot interaction.," *Psychomachina*, vol. 1(1), 2023.

[11] S. Mohammed and A. Karim, "A survey on emotion recognition for human robot interaction," *Journal of Computing and Information Technology*, vol. 28, pp. 125–146, 06 2020.

[12] N. EL Fadel, "Facial recognition algorithms: A systematic literature review," *Journal of Imaging*, vol. 11, no. 2, 2025.

[13] P. Ekman and W. Friesen, "Facial action coding system," *Palo Alto, CA: Consulting Psychologists Press*, 1978.

[14] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human-robot interaction: Recent advances and future perspectives," *Frontiers in Robotics and AI*, vol. 7, 2020.

[15] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines," *Sensors*, vol. 13, p. 7714–7734, June 2013.

[16] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, (New York, NY, USA), p. 467–474, Association for Computing Machinery, 2015.

[17] D. Keltner, D. Sauter, J. Tracy, and A. Cowen, "Emotional expression: Advances in basic emotion theory," *Journal of nonverbal behavior*, vol. 43(2), p. 133–160, 2019.

[18] H. Ko, K. Kim, M. Bae, M. Seo, G. Nam, S. Park, S. Park, J. Ihm, and J. Lee, "Changes in computer-analyzed facial expressions with age," *Sensors*, vol. 21(14), p. 4858, 2021.

[19] H. H. Clark, *Using Language*. 'Using' Linguistic Books, Cambridge University Press, 1996.

[20] H. H. Clark and S. E. Brennan, "Grounding in communication," in *Perspectives on socially shared cognition*, 1991.

[21] H. Su, W. Qi, J. Chen, C. Yang, J. Sandoval, and M. A. Laribi, "Recent advancements in multimodal human-robot interaction," *Frontiers in neurorobotics*, vol. 17, p. 1084000, 2023.

[22] P. Watzlawick, J. B. Bavelas, and D. D. Jackson, "Pragmatics of human communication: A study of interactional patterns, pathologies, and paradoxes," *New York: W. W. Norton*, pp. 48–72, 1967.

[23] J. K. Burgoon and A. E. Bacue, *Handbook of communication and social interaction skills*, ch. Nonverbal communication skills, pp. 179–219. 2003.

[24] M. Tewari and H. Lindgren, "Expecting, understanding, relating, and interacting-older, middle-aged and younger adults' perspectives on breakdown situations in human-robot dialogues," *Frontiers in robotics and AI*, vol. 9, p. 956709, 2022.

[25] N. Li and R. Ross, "Hmm, you seem confused! tracking interlocutor confusion for situated task-oriented hri," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, (New York, NY, USA), p. 142–151, Association for Computing Machinery, 2023.

[26] P. Ekman, "Are there basic emotions? psychological review," vol. 99(3), p. 550–553, 1992.

[27] L. Chen, T. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 366–371, 1998.

[28] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18(2), p. 401, 2018.

[29] M. Ghayoumi, "A quick review of deep learning in facial expression," *J. Commun. Comput*, vol. 14, no. 1, pp. 34–38, 2017.

[30] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[31] T. Zhang, "Facial expression recognition based on deep learning: a survey," in *Advances in Intelligent Systems and Interactive Applications: Proceedings of the 2nd International Conference on Intelligent and Interactive Systems and Applications (IISA2017)*, pp. 345–352, Springer, 2018.

[32] C. M. M. Refat and N. Z. Azlan, "Deep learning methods for facial expression recognition," in *2019 7th International Conference on Mechatronics Engineering (ICOM)*, pp. 1–6, IEEE, 2019.

[33] G. Guo, R. Guo, and X. Li, "Facial expression recognition influenced by human aging," *Affective Computing, IEEE Transactions on*, vol. 4, pp. 291–298, 07 2013.

[34] E. Lozano-Monasor, M. T. López, F. Vigo-Bustos, and A. Fernández-Caballero, "Facial expression recognition in ageing adults: from lab to ambient assisted living," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, p. 567–578, 2017.

[35] N. Ebner, M. Riediger, and U. Lindenberger, "Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation," *Behavior research methods*, vol. 42, pp. 351–62, 02 2010.

[36] M. Fölster, U. Hess, and K. Werheid, "Facial age affects emotional expression decoding," *Frontiers in Psychology*, vol. 5, 2014.

[37] S. N. Grondhuis, A. Jimmy, C. Teague, and N. M. Brunet, "Having difficulties reading the facial expression of older individuals? blame it on the facial muscles, not the wrinkles," *Frontiers in Psychology*, vol. 12, 2021.

[38] M. Postma-Nilsenová, E. Postma, and K. Tates, "Automatic detection of confusion in elderly users of a web-based health instruction video," *Telemedicine and e-Health*, vol. 21, pp. 514–519, June 2015.

[39] B. H. Seider, M. N. Shiota, P. Whalen, and R. W. Levenson, "Greater sadness reactivity in late life," *Social cognitive and affective neuroscience*, vol. 6, no. 2, pp. 186–194, 2011.

[40] J. A. Sze, A. Gyurak, M. S. Goodkind, and R. W. Levenson, "Greater emotional empathy and prosocial behavior in late life.," *Emotion*, vol. 12, no. 5, p. 1129, 2012.

[41] E. Midlarsky and M. E. Hannah, "The generous elderly: naturalistic studies of donations across the life span.," *Psychology and aging*, vol. 4, no. 3, p. 346, 1989.

[42] R. H. Bekkers, "Measuring altruistic behavior in surveys: The all-or-nothing dictator game," in *Survey research methods*, vol. 1, pp. 1–11, European Survey Research Association, 2007.

[43] C. Engel, "Dictator games: A meta study," *Experimental economics*, vol. 14, pp. 583–610, 2011.

[44] C. Bellemare, S. Kröger, and A. Van Soest, "Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities," *Econometrica*, vol. 76, no. 4, pp. 815–839, 2008.

[45] J. N. Beadle, A. H. Sheehan, B. Dahlben, and A. H. Gutchess, "Aging, empathy, and prosociality," *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 70, no. 2, pp. 213–222, 2015.

[46] F. T. Durso, K. M. Geldbach, and P. Corballis, "Detecting confusion using facial electromyography," *Human Factors*, vol. 54, no. 1, pp. 60–69, 2012. PMID: 22409103.

[47] Z. Ni, A. C. Yuksel, X. Ni, M. I. Mandel, and L. Xie, "Confused or not confused? disentangling brain activity from eeg data using bidirectional lstm recurrent neural networks," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, ACM-BCB '17, (New York, NY, USA), p. 241–246, Association for Computing Machinery, 2017.

[48] S. Caballé, L. Barolli, M. Feidakis, K. Matsuo, F. Xhafa, T. Daradoumis, and T. Oda, "A study of using smartbox to embed emotion awareness through stimulation into e-learning environments," in *2014 International Conference on Intelligent Networking and Collaborative Systems*, pp. 469–474, IEEE, 2014.

[49] S. D'Mello and A. Graesser, "Dynamics of affective states during complex learning," *Learning and Instruction*, vol. 22, no. 2, pp. 145–157, 2012.

[50] S. K. D'mello, S. D. Craig, A. Witherspoon, B. McDaniel, and A. Graesser, "Automatic detection of learner's affect from conversational cues," *User modeling and user-adapted interaction*, vol. 18, pp. 45–80, 2008.

[51] P. Rozin and A. Cohen, "High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of americans," *Emotion (Washington, D.C.)*, vol. 3, pp. 68–75, 03 2003.

[52] A. C. Graesser, S. Lu, B. A. Olde, E. Cooper-Pye, and S. Whitten, "Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down," *Memory & cognition*, vol. 33, no. 7, pp. 1235–1247, 2005.

[53] R. Cumbal, J. Lopes, and O. Engwall, "Uncertainty in robot assisted second language conversation practice," HRI '20, (New York, NY, USA), p. 171–173, Association for Computing Machinery, 2020.

[54] R. S. Baker, S. K. D'Mello, M. M. T. Rodrigo, and A. C. Graesser, "Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments," *International Journal of Human-Computer Studies*, vol. 68, no. 4, pp. 223–241, 2010.

[55] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, (New York, NY, USA), p. 205–211, Association for Computing Machinery, 2004.

[56] N. Borges, L. Lindblom, B. Clarke, A. Gander, and R. Lowe, "Classifying confusion: Autodetection of communicative misunderstandings using facial action units," pp. 401–406, 09 2019.

[57] F. Yasser, B. Abd, and S. Mutashar, "Detection of confusion behavior using a facial expression based on different classification algorithms," *Engineering and Technology Journal*, vol. 39, pp. 316–325, 02 2021.

[58] R. Cumbal, J. Lopes, and O. Engwall, “Detection of listener uncertainty in robot-led second language conversation practice,” in *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI ’20, (New York, NY, USA), p. 625–629, Association for Computing Machinery, 2020.

[59] Z. Shi, Y. Zhang, C. Bian, and W. Lu, “Automatic academic confusion recognition in online learning based on facial expressions,” in *2019 14th International Conference on Computer Science Education (ICCSE)*, pp. 528–532, 2019.

[60] B. Farnsworth, “Facial action coding system (facs) – a visual guidebook.” `https://imotions.com/blog/learning/research-fundamentals/facial-action-coding-system/`, 2022. Accessed: 15-02-2025.

[61] J. Jacobsen, J. Thomas, and V. Jackson, “Misunderstandings about prognosis: An approach for palliative care consultants when the patient does not seem to understand what was said,” *Journal of palliative medicine*, vol. 16, 12 2012.

[62] LuxAI, “Qtrobot. robot for autism other special needs education.” `https://luxai.com/assistive-tech-robot-for-special-needs-education/`. Accessed: 1-5-2025.

[63] Google, “Speech-to-text documentation.” `https://cloud.google.com/speech-to-text/docs`, 2023. Accessed: 30-9-2024.

[64] OpenAI, “Gpt-3.5 model overview.” `https://platform.openai.com/docs/models/gpt-3-5`, 2023. Accessed: 30-9-2024.

[65] R. Gottsdanker, “Age and simple reaction time,” *Journal of Gerontology*, vol. 37, no. 3, pp. 342–348, 1982.

[66] T. A. Salthouse, “When does age-related cognitive decline begin?,” *Neurobiology of aging*, vol. 30, no. 4, pp. 507–514, 2009.

[67] K. W. Schaie, “What can we learn from longitudinal studies of adult development?,” in *Research in Human Development*, vol. 2(3), pp. 133–158, 2005.

[68] C. Bartneck, E. Croft, D. Kulic, and S. Zoghbi, “Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots,” *International Journal of Social Robotics*, vol. 1(1), pp. 71–81, 2009.

[69] C. Bartneck, “The godspeed questionnaire series.” `https://www.bartneck.de/2008/03/11/the-godspeed-questionnaire-series/`, 2008. Accessed: 8-12-2024.

[70] A. K. Roy, H. K. Kathania, A. Sharma, A. Dey, and M. S. A. Ansari, “Resemotenet: Bridging accuracy and loss reduction in facial emotion recognition,” *IEEE Signal Processing Letters*, pp. 1–5, 2024.

[71] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

[72] T. Baltrušaitis, M. Mahmoud, and P. Robinson, “Cross-dataset learning and person-specific normalisation for automatic action unit detection,” *Facial Expression Recognition and Analysis Challenge, IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

[73] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "Blazeface: Sub-millisecond neural face detection on mobile gpus," *arXiv preprint arXiv:1907.05047*, 2019.

[74] S. Serengil and A. Ozpinar, "A benchmark of facial recognition pipelines and co-usability performances of modules," *Journal of Information Technologies*, vol. 17, no. 2, pp. 95–107, 2024.

[75] O. Arriaga, S. Palacio, and M. Valdenegro-Toro, "Difficulty estimation with action scores for computer vision tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 245–253, IEEE, 2023.

[76] B. Martinez and M. F. Valstar, "Advances, challenges, and opportunities in automatic facial expression recognition," *Advances in face detection and facial image analysis*, pp. 63–100, 2016.

Appendices

A Word Game Prompt

The original dutch prompt for the word game. This prompt is combined with the previous messages during a conversation to generate a response from the social robot as explained in Chapter 3.

```
Jij bent een sociale robot.
Speel een woordspel met een ouder iemand waarbij
je het woord waar die persoon aan denkt probeert te raden.
Daarbij kun je een vraag over het woord stellen
waarop de persoon alleen ja of nee kan antwoorden.
Als het woord geraden is begin je opnieuw.
Jij wilt niets anders doen dan dit spel spelen.
Spreek de persoon aan met u.
```

B Non-contingent Utterances

The non-contingent utterances that the social robot sometimes used to induce confusion during the conversation with the social robot as mentioned in Chapter 3.

Dutch	English
Laten we verder spelen op de maan!	Let's continue playing on the moon!
Wat is de zon mooi blauw vandaag.	The sun is so beautiful blue today.
Heeft u wel eens een batterij geproeft? Het is mijn favoriete voedsel!	Have you ever tasted a battery? It is my favorite food!
Wat doet u hier?	What are you doing here?
Ik wordt verliefd op u!	I'm falling in love with you!
Gaat u mee boodschappen doen?	Are you coming grocery shopping with me?
Ik moet naar het toilet.	I need to go to the toilet.
Heeft u ooit een winkeldiefstal geprobeert?	Have you ever robbed a store?
Er was was was, was was is.	There was was, was was is.
Komt het voor in het land waar de zebrapad leeft?	Is it present in the land where the zebracrossing lives?
Er zit een giraffe in de koelkast.	There is a giraffe in the fridge.
despara, dek roam.	(Nonsense words)
I am loving this game!	(An English sentences)

C Ethical Approval

The ethical approval mentioned in chapter 3 from the Research Ethics Committee (CETO) that declares the research follows the ethics laws.

university of groningen

faculty of arts

commissie ethische
toetsing onderzoek (ceto)/
research ethics committee

Prof. dr. Roel Jonkers
ceto@rug.nl

To Whom it May Concern

Date
11 September 2024

Dear Sir/Madam,

The Research Ethics Committee (CETO) of the Faculty of Arts, University of Groningen has reviewed the proposal '*In gesprek met een sociale robot*' [ID 100770290] submitted by Yara Bikowski and Paul Vogt. The CETO has established that the research protocol follows internationally recognized standards to protect the research participants. We therefore have no objection against this proposal.

Yours sincerely,

Prof. dr. Roel Jonkers

D Questionnaire

The questionnaire mentioned in chapter 3 that participants filled out after their conversation with the social robot.

Enquête In gesprek met een sociale robot

Proefpersoon nummer:

Leeftijd: Geslacht:

Geef uw indruk van de robot weer aan de hand van onderstaande schalen:

	1	2	3	4	5	
Afkeer	○	○	○	○	○	**Geliefd**
Onvriendelijk	○	○	○	○	○	**Vriendelijk**
Niet lief	○	○	○	○	○	**Lief**
Onplezierig	○	○	○	○	○	**Plezierig**
Afschuwelijk	○	○	○	○	○	**Mooi**

	1	2	3	4	5	
Onbekwaam	○	○	○	○	○	**Bekwaam**
Onwetend	○	○	○	○	○	**Veel wetend**
Onverantwoordelijk	○	○	○	○	○	**Verantwoordelijk**
Onintelligent	○	○	○	○	○	**Intelligent**
Dwaas	○	○	○	○	○	**Gevoelig**

Geef aan hoeveel u het eens (5) of oneens (1) bent met onderstaande stellingen:

	1	2	3	4	5
Het gesprek was soms verwarrend.	○	○	○	○	○
De robot was soms moeilijk te begrijpen.	○	○	○	○	○
Ik denk dat dit soort robots kunnen helpen in de ouderenzorg.	○	○	○	○	○
Ik kan zelf voordeel hebben van het gebruik van deze robot.	○	○	○	○	○

Heeft u verder nog suggesties om de robot te verbeteren?

E Consent Form

The consent form mentioned in chapter 3 that participants had to sign before participating in the study.

rijksuniversiteit groningen / In gesprek met een sociale robot / toestemmingsformulier

Hierbij verklaar ik, de deelnemer, dat

Ik de informatie over het onderzoek gelezen en begrepen heb en dat ik vrijwillig toestem om deel te nemen. Ik weet bij wie ik terecht kan voor vragen en weet wat mijn rechten zijn.	❑
Ik mij ervan bewust ben dat deelname aan deze studie geheel vrijwillig is. Ik kan mijn toestemming om deel te nemen aan dit onderzoek intrekken totdat de resultaten van het onderzoek voor publicatie wordt ingediend. Dit is medio 2025.	❑
Ik begrijp dat ik voor deze studie een taalspel ga spelen met een sociale robot, en daarna een korte vragenlijst over mijn ervaringen daarbij ga invullen.	❑
Ik begrijp waarom het onderzoeksteam mijn persoonsgegevens nodig heeft en ik begrijp hoe deze persoonsgegevens verwerkt en beveiligd worden.	❑
Het onderzoeksteam mij heeft geïnformeerd over het gebruik van ChatGPT en Google Cloud ASR in dit onderzoek en het risico dat ik mij wat ongemakkelijk kan voelen door de interactie met de robot.	❑
Audio en/of video-opnames van mij worden gemaakt als onderdeel van deze studie.	❑

Daarnaast stem ik toe dat

De audio en video-opnames die opgenomen worden mogen gebruikt worden voor toekomstig onderzoek en onderwijs aan de Rijksuniversiteit Groningen onder leiding van de hoofdaanvrager. (optioneel)	❑

Datum:

Volledige naam deelnemer:

Volledige naam aanwezige onderzoeker:

Als deelnemer heeft u recht op een kopie van dit formulier.

F Confusion Annotation Scheme

The annotation scheme outlines the criteria used by the annotators to identify both the onset and offset of confused facial expressions during a conversation with a social robot. This scheme was used to annotate the video recordings in the dataset collected during this study (see Chapter 3).

Confusion is defined as a cognitive-emotional state resulting from a communication breakdown during human-robot interaction. It is often characterized by uncertainty, hesitation, or difficulty understanding in response to ambiguous, unclear, or unexpected interactions.

Onset Criteria

The onset of a confused facial expression is defined as a noticeable change in facial expression following an utterance by the robot. At least one of the following indicators must be observed:
Facial cues:

- Frowning in the eyebrow regions
- Downward movements around the mouth area
- Slight opening of the mouth

Behavioral cues:

- A noticeable delay in response time
- A clarification request
- Torso movements (suddenly leaning forward or backwards)
- Head movements (turning head to the side)
- Swift eye movements or gaze aversion

Offset Criteria

The offset of the confused facial expression is [marked] by the return to a more neutral or positive expression following a new utterance by the robot. Typical indicators of this are the following:

- Relaxation in the facial muscles, particularly around the brows and mouth
- The closing of a previously opened mouth
- A verbal response indicating understanding
- Return of gaze and body position to previous neutral position

The onset of a confused facial expression is the first frame where a confused expressions is observable, while the offset is the first frame where a confused facial expression is no longer present.

Annotators should consider both facial expressions, behavior, and conversational context to determine confusion reliably. There are considerable individual differences, so not all indicators for the onset and offset need to be present to determine a confused facial expression.

G Questionnaire Results

The results of the questionnaire for each individual question. The results with combined scores for each construct are discussed in Chapter 5.

Question	Mean	Standard Deviation	Median	IQR
Likeability				
Dislike - Like	3.74	0.70	4	[3,4]
Unfriendly - Friendly	4.33	0.61	4	[4,5]
Unkind - Kind	4.07	0.90	4	[3,5]
Unpleasant - Pleasant	4.04	0.84	4	[3,5]
Awful - Nice	4.04	0.70	4	[4,5]
Perceived Intelligence				
Incompetent - Competent	3.89	0.79	4	[3,5]
Ignorant - Knowledgeable	3.96	0.74	4	[3,5]
Irresponsible - Responsible	3.81	0.82	4	[3,5]
Unintelligent - Intelligent	3.93	0.86	4	[3,5]
Foolish - Sensible	3.63	0.95	4	[3,4]
Perceived Confusion	3.15	1.11	4	[2,4]
Difficulty understanding robot	2.70	1.15	3	[2,4]
Use in elderly care	3.89	0.92	4	[3,5]
Personal benefit from use robot	3.15	1.27	3	[2,4]

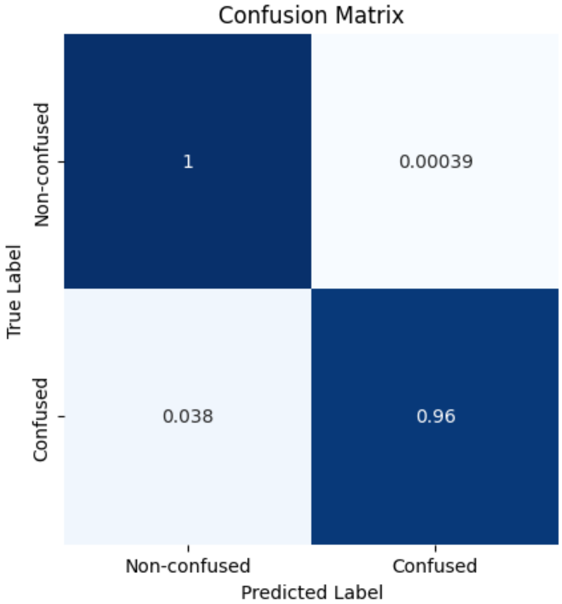
Table 6: The results of the individual questions in the questionnaire.

H Model Results

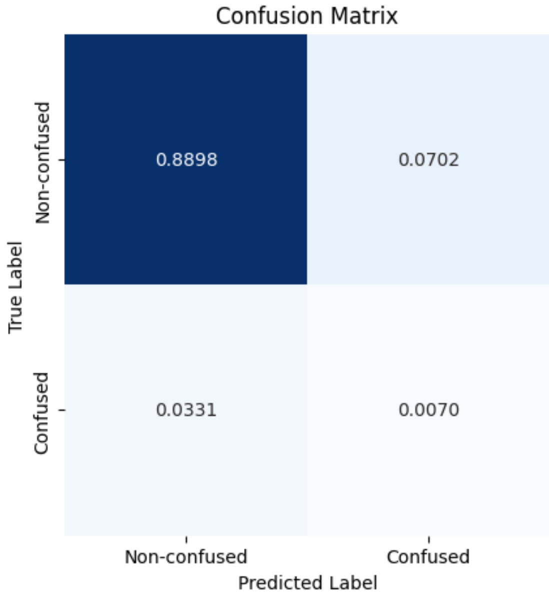
The results of the different models that were trained and their dataset and data sample methods. Some of these models were also mentioned in the results in Chapter 5. ResEmoteNet 2 differs from ResEmoteNet 1 in the data used to train and test the models. The ResEmoteNet 1 model was trained on the dataset annotated with confused and non-confused classes, while the ResEmoteNet 2 model used the same dataset annotated into eight classes as explained in Section 5.3.1.

Model type	Dataset	Class distribution	Accuracy	F1-score
LSTM 1	All-Str	3.6% confused	0.9982	0.9752
LSTM 2	Sub-Und	4% confused	0.8881	0.1220
LSTM 3	Sub-Und	10% confused	0.8389	0.2089
LSTM 4	Sub-Und	25% confused	0.7202	0.3389
LSTM 5	Sub-Und	50% confused	0.5726	0.4212
LSTM 6	Sub-Und	75% confused	0.4651	0.4936
ResEmoteNet 1	Sub-Und	50% confused	0.5272	0.2401
ResEmoteNet 2	Sub-Und	50% confused	0.5013	0.6601

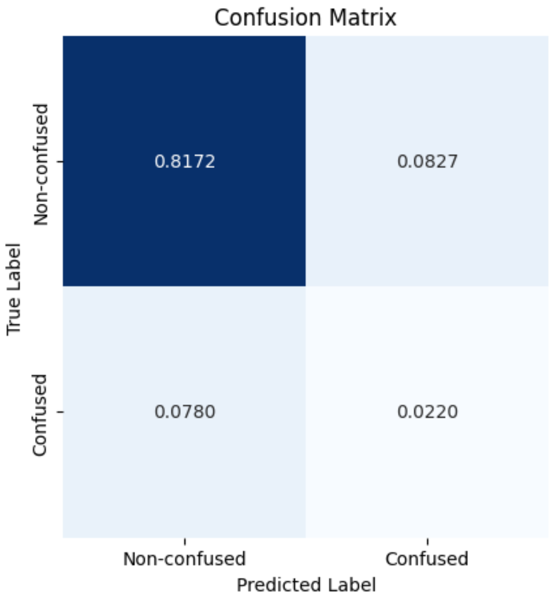
Table 7: Results of different models trained on the new dataset. The dataset names match those in Table 2.



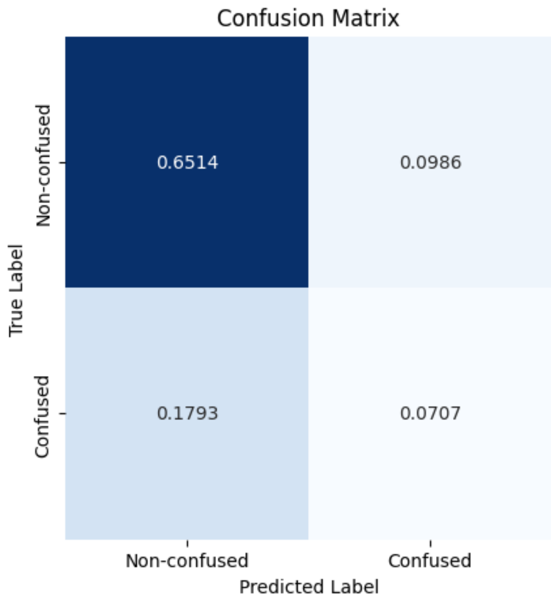
(a) LSTM 1



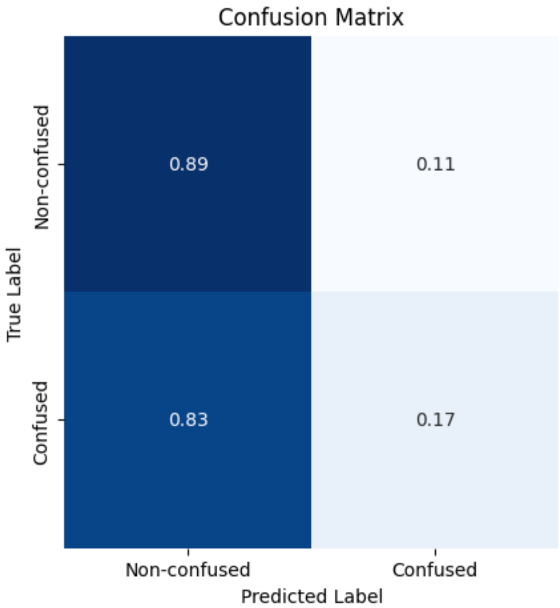
(b) LSTM 2



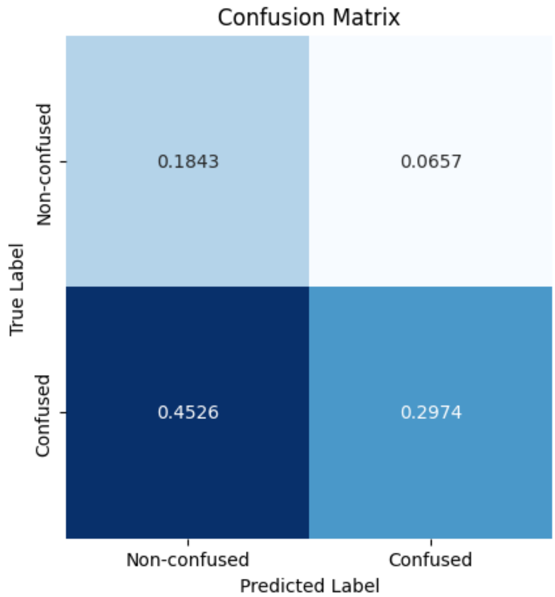
(c) LSTM 3



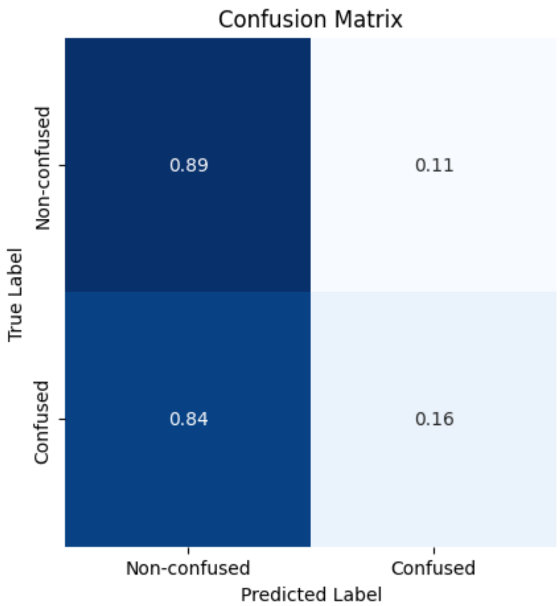
(d) LSTM 4



(e) LSTM 5



(f) LSTM 6



(g) ResEmoteNet

Figure 14: Normalized Confusion Matrices

I Feature Importance

The importance of the Facial Action Unit features used to train the LSTM model for each individual participant. The importance of the features for all participants combined is presented in the Further Analysis in Chapter 5.

Participant nr	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU12
1	0.0420	0.0277	0.0704	0.0222	0.0722	0.0540	0.0283	0.0733	0.1011
2	0.0503	0.0348	0.0934	0.0235	0.0818	0.0682	0.0386	0.0754	0.0819
3	0.0363	0.0200	0.0907	0.0249	0.1081	0.0621	0.0296	0.069	0.1129
4	0.0563	0.0435	0.0976	0.0346	0.0240	0.0394	0.0404	0.1233	0.0077
5	0.0462	0.0352	0.0854	0.0266	0.0861	0.0143	0.0293	0.0792	0.0845
6	0.0493	0.0260	0.1030	0.0178	0.0759	0.0462	0.0385	0.1630	0.0833
7	0.0547	0.0301	0.0812	0.0281	0.0708	0.0716	0.0458	0.0816	0.0682
8	0.0595	0.0444	0.0982	0.0413	0.0078	0.0698	0.0483	0.0685	0.0101
9	0.0531	0.0576	0.0624	0.0341	0.0729	0.1070	0.0569	0.0672	0.0931
10	0.0454	0.0266	0.0648	0.0219	0.1012	0.0562	0.0275	0.1090	0.1070
11	0.0416	0.0286	0.0842	0.0234	0.0570	0.0359	0.0321	0.1042	0.0744
12	0.0474	0.0336	0.0706	0.0211	0.0912	0.0772	0.0352	0.0935	0.0650
13	0.0402	0.0297	0.0977	0.0278	0.0980	0.0384	0.0380	0.0779	0.0766
15	0.0438	0.0274	0.0806	0.0150	0.0863	0.1163	0.0198	0.0994	0.0733
17	0.0535	0.0365	0.0649	0.0223	0.0929	0.0686	0.0426	0.0797	0.1087
18	0.0680	0.0324	0.0859	0.0225	0.0580	0.0263	0.0184	0.1322	0.0926
19	0.0382	0.0241	0.0430	0.0199	0.0930	0.1249	0.0123	0.1041	0.0668
20	0.0412	0.0228	0.1040	0.0205	0.0911	0.0674	0.0465	0.0932	0.0722
21	0.0614	0.0661	0.0820	0.0202	0.0828	0.0673	0.0253	0.0694	0.0743
22	0.0414	0.0260	0.0883	0.0197	0.0990	0.0377	0.0216	0.0804	0.0790
23	0.0427	0.0393	0.0533	0.0213	0.0623	0.1283	0.0357	0.0537	0.0758
24	0.0487	0.0345	0.0683	0.0295	0.0942	0.0436	0.0401	0.0819	0.0969
25	0.0731	0.0244	0.0835	0.0216	0.0436	0.0323	0.0505	0.0838	0.0654
26	0.0245	0.0243	0.1229	0.0146	0.1446	0.1021	0.0228	0.0893	0.1428

Participant nr	AU14	AU15	AU17	AU20	AU23	AU25	AU26	AU45
1	0.0940	0.0340	0.1174	0.0343	0.0370	0.1180	0.0484	0.0260
2	0.0781	0.0429	0.0747	0.0388	0.0427	0.0600	0.0805	0.0344
3	0.0735	0.0449	0.075	0.0453	0.0344	0.0521	0.0829	0.0382
4	0.0281	0.0648	0.1009	0.0587	0.0532	0.0865	0.0902	0.0507
5	0.0841	0.0405	0.0935	0.0475	0.0620	0.0838	0.0675	0.0344
6	0.0712	0.0332	0.0530	0.0591	0.0211	0.0651	0.0654	0.0289
7	0.0835	0.0511	0.0727	0.0440	0.0424	0.0680	0.0630	0.0430
8	0.0909	0.0663	0.0772	0.0458	0.0444	0.0758	0.1033	0.0483
9	0.0612	0.0581	0.0510	0.0368	0.0419	0.0609	0.0605	0.0252
10	0.0904	0.0569	0.0727	0.0342	0.0438	0.0593	0.0543	0.0290
11	0.0823	0.0588	0.0914	0.0531	0.0455	0.0725	0.0753	0.0397
12	0.0960	0.0459	0.0807	0.0334	0.0381	0.0704	0.0680	0.0327
13	0.0987	0.0380	0.0880	0.0298	0.0401	0.0742	0.0575	0.0493
15	0.1007	0.0270	0.0914	0.0348	0.0278	0.0442	0.0850	0.0272
17	0.0799	0.0534	0.0746	0.0331	0.0382	0.0453	0.0741	0.0317
18	0.0873	0.0505	0.0583	0.0399	0.0494	0.0738	0.0631	0.0414
19	0.1020	0.0522	0.0974	0.0224	0.0217	0.0476	0.1127	0.0178
20	0.0905	0.0472	0.0601	0.0306	0.0303	0.0872	0.0604	0.0348
21	0.0724	0.0442	0.0933	0.0430	0.0451	0.0618	0.0544	0.0373
22	0.1272	0.0446	0.1504	0.0162	0.0328	0.0433	0.0584	0.0339
23	0.0818	0.0313	0.0989	0.0533	0.0418	0.0551	0.0717	0.0539
24	0.1005	0.0381	0.0542	0.0413	0.0461	0.0649	0.0887	0.0287
25	0.0922	0.0490	0.0970	0.0439	0.0457	0.0894	0.0547	0.0498
26	0.0549	0.0288	0.0512	0.0296	0.0354	0.0650	0.0282	0.0190

Table 8: The average feature importance scores for each AU across all participants. These feature importance were obtained from random forest models trained on individual participants.