Utrecht University
Faculty of Science
Department of Information and Computing Sciences
MSc Artificial Intelligence

A multimodal approach to working alliance detection in therapist-patient psychotherapy using deep learning models

Master Thesis

Rivka Vollebregt

8842507

Project supervisor Prof. dr. Albert Ali Salah
Second examiner dr. Itır Önal Ertuğrul

Abstract

The working alliance between therapist and patient is a critical component of a therapeutic process and is an important part of preventing ruptures and drop-outs. The purpose of this research was to gain a better understanding of the working alliance so therapists can get more insight into what factors patients take into account in their perception of it, and so working alliance might be predicted automatically to detect a low working alliance early in the therapeutic process. In this study, we analyzed the relationship between various indicators across multiple modalities and the perception of working alliance according to the perception of the patient, the therapist, and independent observers. All three perceptions showed a difference in their perception of the working alliance in terms of which features were most important. The features that were used stemmed from the conversational modality, affect analysis from audio and text, facial emotion analysis, and head movements. Several features have been identified as strong predictors of the working alliance. It showed that therapists base their perception of the working alliance often on contrasting features to the patient, underscoring the importance of this research in informing the therapists of this difference and preventing low working alliance from the patient resulting in low therapeutic outcome. We have designed multiple regression models which could not show strong prediction performance due to the small dataset. Furthermore, a support vector machine (SVM) model was designed that based on the automatically extracted features, predicts with 80–90 per cent accuracy whether a therapy session had a high or low WAI score. This research was a starting point to explore the complex process that underlies a patient's perception of the working alliance and provided more understanding so that future research can analyze each aspect of predicting working alliance with automatic tools in depth.

Table of Contents

List of the most important Abbreviations

Abbreviation	Full Form
AI	Artificial Intelligence
WAI	Working Alliance Inventory
WAI-S	Working Alliance Inventory Short form
WAI-SRT	Working Alliance Inventory Short Revised Therapist
NLP	Natural Language Processing
WER	Word Error Rate
MER	Match Error Rate
WIL	Word Insertion Likelihood
DER	Diarization Error Rate
JER	Jaccard Error Rate
AUs	Action Units
FACS	Facial action coding system
LIWC	Linguistic Inquiry and Word Count
RF	Random Forest
LSTM	Long-term Short-term Memory neural network
BERT	Bidirectional Encoder Representations from Transformers
VAD	Voice Activity Detection
SVM	Support Vector Machine
kNN	k-Nearest Neighbours
SVR	Support Vector Regression
XGB	Extreme Gradient Boosting
MRMR	Maximum Relevance Minimum Redundancy
CV	Coefficient of Variation

1. Introduction

This thesis proposal dives into the therapeutic alliance, or working alliance, which is the relationship between a patient and his or her therapist. Establishing the quality of the working alliance is beneficial as it gives the therapist a method of improving the therapeutic quality. Manually evaluating the working alliance can be time-consuming. The development of an automatic indicator identification system to detect the working alliance is, therefore, a good improvement over the method of manual detection. The purpose of this thesis proposal is to devise a system that can detect visual and textual indicators that reflect the state of the working alliance. This thesis proposal describes the research that will be conducted to find features in video recordings of psychotherapy sessions that are indicative of working alliance, which will be used with a machine learning model to predict the working alliance between therapist and patient in a psychotherapy session.

1.1 Relevance

Therapy is a large part of the mental health system and has become increasingly important in recent years, as there has been a 13% rise in the prevalence of mental health issues and substance use disorders between 2007 and 2017[1]. This rise is felt by psychotherapists who reported an increase in demand for anxiety and depression treatments in 2021 compared to 2020, resulting in less available treatment (14). Because of this increasing pressure on psychotherapy to combat mental health issues in perhaps fewer sessions due to the rising demand, it is important to have an effective treatment during psychotherapy sessions.

1.2 Working alliance

A widely-used factor that measures the quality of the relationship between therapist and patient in psychotherapy is the "working alliance" (47; 37). Working alliance can be defined as the "collaborative and affective bond between the therapist and patient" (89). This alliance is considered to be one of the most important factors in the success of therapy, as it encompasses the emotional bond, trust, and understanding that develops between the therapist and the patient. Research on the working alliance has found that a strong alliance is associated with better outcomes in therapy. A meta-analysis of studies by Horvath and Symonds (1991)

[1] World Health Organization, `https://www.who.int/health-topics/mental-health#tab=tab_2`

found that the alliance was the most consistent predictor of therapy outcome across a wide range of therapies and patient populations (62). A more recent meta-analysis by Martin, Garske, and Davis (2000) confirmed that the alliance is positively correlated with symptom reduction and therapeutic progress (89).

The working alliance is typically assessed using self-report measures, such as the Working Alliance Inventory (WAI), which assesses the patient's perceptions of the alliance with the therapist (61). Other measures, such as the Therapist Alliance Scale (TAS), assess the therapist's perceptions of the alliance (82). Several factors contribute to the formation of a strong working alliance. These include: Empathy: The therapist's ability to understand and validate the patient's feelings and experiences (120). Authenticity: The therapist must be his or her authentic self in the therapeutic relationship (120). Goals: The therapist's and patient's shared understanding of the goals and objectives of therapy (62). Tasks: The therapist's and patient's shared understanding of the tasks and activities that will be undertaken in therapy (62). Bond: The emotional connection between the therapist and patient (61).

It is important to note that the working alliance may fluctuate throughout the course of therapy, and can be strengthened or weakened by various factors such as the therapist's interventions, the patient's attitude, and therapist-patient communication. Therefore, the therapist needs to keep track of the alliance and make any necessary adjustments in his therapeutic methods (59). Accurately predicting the working alliance can help the therapist to gain a better understanding of the working alliance itself and methods to increase it to also increase the quality of the sessions.

While much research has already gone into the effectiveness of psychotherapy, the mechanism underlying working alliance is still largely unknown. Currently, the analysis of psychotherapy sessions is done manually, scoring the interaction between the patient and therapist by a qualified psychotherapist. This work is very labor-intensive and is sensitive to bias due to the subjectivity of the observer's judgment. Therefore, this project aims to use deep learning to automatize the analysis of the interaction between the patient and the therapist during psychotherapy sessions. This will not only save valuable time by removing the labor-intensive scoring of the sessions but will also provide more insight into the efficiency of specific components of the psychotherapy session. The benefit of this will be twofold. First, it can help improve the effectiveness of the psychotherapy sessions by helping the therapist improve their methods based on the patient's perception of the working alliance. Second, an accurate prediction of working alliance can make psychotherapy sessions more effective, relieving some of the pressure on the mental health system by achieving the same outcome in fewer sessions.

1.3 Working Alliance Inventory

The Working Alliance Inventory (WAI) is a widely used measure of the working alliance in psychotherapy. Developed by Bordin (1979), the WAI is a self-report questionnaire that assesses the patient's perceptions of the therapeutic alliance in three areas: goals, tasks, and bond (18). The questionnaire consists of 36 items that assess the perception of the agreement and collaboration of the patient and the therapist regarding therapy goals, tasks, and the therapeutic relationship. The WAI can be filled in by the therapist, the client, or an observer, so the perception of the working alliance can be measured from the patient's, the therapist's, and an outside perspective. There are different versions of the WAI for the therapist and the patient to fill in which are called the Working Alliance Inventory-patient version (WAI-C), and the Working Alliance Inventory-Therapist version (WAI-T). For the outcome of psychotherapy, the patient's perception of the working alliance is most relevant. It's worth mentioning that there exists also a short version of the WAI which consists of 12, instead of 36 items. This short version is called the Working Alliance Inventory-Short form (WAI-S). There is also a short version of the WAI for the therapist which consists of 10 questions and is called the Working Alliance Inventory-Short Revised Therapist (WAI-SRT). In this research, we use the WAI-S for capturing the patient's and observer′s perception of the working alliance and the WAI-SRT for the therapist's perception of the working alliance. Both questionnaires are added to Appendix 9.

The WAI has been found to have high reliability and validity in multiple studies. For example, a study by Horvath and Symonds (1991) found that the WAI had a high level of internal consistency (Cronbach's alpha = 0.86) and good test-retest reliability (r = 0.77) across different types of psychotherapy (62).

In addition, research has shown that WAI is related to therapeutic outcomes. For example, a meta-analysis by Martin et al. (2000) found that the working alliance, as measured by the WAI, was significantly related to therapy outcome (89), which was supported by a meta-analysis from Horvath and Bedi (2002) (104).

1.3.1 Working Alliance Inventory in different languages

The WAI has been used in different cultures and languages, and it has been shown to have good reliability and validity in different cultures such as in Chinese and Spanish (64; 28). However, this study uses a database that was recorded with Dutch-speaking therapists and patients. Up till now, there has been no extensive validation of the WAI in the Dutch language, which requires some apprehension when analyzing the WAI results (129). Cultural bias will also likely play a role in using a Dutch dataset, as results from a Dutch language inventory will differ from, for instance, a British inventory, due to cultural bias. An indication of this is that Dutch people are known around the world as being very direct in their speech, which may cause distinct differences in the use of language as British people. This does not mean that the WAI or language processing in Dutch datasets can not be used, but it does suggest some caution in comparing this study's result with English-speaking WAI studies.

1.4 Machine learning to predict working alliance

Machine learning is a powerful tool that can be used to predict the working alliance in psychotherapy. The ability to predict the alliance early in therapy can help therapists to identify potential issues and take steps to strengthen the alliance, which can ultimately lead to better therapy outcomes.

One approach to predicting the working alliance with machine learning is to use natural language processing (NLP) techniques to analyze the transcriptions of therapy sessions. For example, in a study by Wampold et al. (2015), researchers used NLP to analyze the transcripts of therapy sessions and found that certain linguistic markers, such as the use of first-person pronouns and positive emotion words, were positively associated with the working alliance (143).

The use of machine learning techniques requires some caution as it does bring out ethical concerns. It should, therefore, be guided by ethical considerations, such as ensuring that the working alliance predictions are used as an additional statistic to help the therapist gain insight into the process to improve therapy outcomes, and are not used to replace the therapist's judgment or fully relied upon.

1.4.1 Bias in the machine learning model

Predicting a working alliance using a machine learning model instead of human observers might reduce the subjectivity that goes into the prediction. Humans are always biased in their observations because of (sub)conscious stereotypes, their current state, and recent experiences. For example, a human observer might pay more attention to a specific indicator such as facial expressions if he or she has recently read a study on the importance of facial expressions. An often-used argument for implementing automated systems over humans is that they are objective and not prone to bias like humans are. While automated systems have the advantage of being consistent in their predictions, they are not necessarily unbiased. Since we train and test the model using human predictions as a reference, the model will likely also incorporate some bias that the human observers might have had. However, the model is trained on human predictions provided by multiple people, which will result in an average bias of these people. Thus, when predicting a working alliance in a newly recorded video, the model will likely be less subjective than a human observer would have been as it is trained to be the average of all human observers.

1.5 Purpose of this research

The purpose of this research is to design a machine learning system that can receive a video recording as input and based on the indicators it can extract from that video, produce the predicted working alliance as output.

To be able to design such a system, we need to find out which indicators (such as gestures) are useful to extract from the video and how to extract them. We will find out which features are good indicators of working alliance by comparing the predicted working alliance score of the machine learning system based on the extracted features, with the working alliance score as provided by the Working Alliance Inventory. If a feature or a combination of features correlates very highly with the WAI score, then we assume that it is highly indicative of a working alliance.

The initial features that we are going to extract are based on the findings of previous research on the working alliance. By going through the many studies that have researched working alliance and their mechanisms, we end up with a list of about 10 features (this number can change during the study) that are likely to be good indicators of the working alliance. Then, we extract these features from the entire dataset and use them as input for our machine learning model which will have a deep learning architecture. The model will perform supervised training on part of the dataset where it will learn to adjust its weights to arrive at the right prediction (which is given by the WAI: the working alliance score).

After training, we will use a test set to determine the performance of the model and measure the accuracy of its prediction. Furthermore, we would specifically like to know which features are good predictors, therefore, we will calculate the predictive value of each feature which will give us a list of how much each feature is indicative of the working alliance. With this result, we will be able to build a second model incorporating only the features that turned out to be good predictors for working alliance and test if that model can accurately predict the WAI score.

We assume that the WAI is highly correlated with the actual working alliance and thus that the model will be correlated with the actual working alliance, but since we do not specifically test this, we are unable to conclude this with certainty. However, we can use the outcome of the model to gain insight into the working alliance. First, we will know what the indicators are for the working alliance. Second, we will have a model that will hopefully be at least as accurate as humans, to replace the labor-intensive work of therapists in analyzing the working alliance.

1.6 Research questions

As mentioned in the introduction, predicting the working alliance using automatic methods is an important improvement in the field of psychotherapy. Therefore, we have attempted to create a machine learning model that can predict the working alliance or the WAI score that is used as a measurement for the working alliance. Such a model will have automated a process that takes up valuable time of licensed therapists and contributes to the understanding of the patient's perception of the relationship with the therapist.

Research question: To what extent can a machine learning model predict the quality of the working alliance between therapists and patients using visual and textual features extracted from psychotherapy sessions?

The above-mentioned research question is a very broadly formulated question that we will specify further in this research. The machine learning model encompasses a broad scope of possibilities. Therefore, our first task in this research is to select a few possible machine learning algorithms that have proven successful for this prediction task in related research and test them using our dataset. This part of the thesis forms its sub-experiment with its sub-research question formulated below.

Sub-research question 1: Which machine learning model demonstrates the highest predictive performance for predicting the working alliance in psychotherapy sessions based on multimodal features?

This sub-research question specifically states which model is best suited instead of has the highest accuracy as we will be looking at more aspects than just accuracy. Aspects such as speed, generalizability, and explainability are also important to take into account in the context of working alliance prediction.

It is also interesting to examine whether the indicators of the therapist or those of the patient are more predictive of working alliance. There is some debate among scientists about whether incorporating the indicators of the therapist is useful for predicting the patient's perception of the working alliance, as we will discuss later on in this thesis. Therefore, we also wish to examine the difference in predictability between the therapist's and patient's indicators which is formulated in sub-research question 2.

Sub-research question 2: In a psychotherapy session, what is the difference in predictability between indicators displayed by the therapist and indicators displayed by the patient for predicting working alliance?

The concept of multimodality and the predictability of the individual indicators as opposed to the combination of indicators is also interesting to study. We assume that using multimodality will increase the prediction accuracy of the model compared to using a single modality. However, this has not yet been tested with as many modalities as used in this research, thus we want to examine Also, we would like to know the predictability of each indicator so we can create a second model with only the most predictive indicators. Using only a subset of the possible indicators to predict the working alliance, without compromising accuracy, will be beneficial as it could cut down on the time that the model needs to predict and it will result in a less complex model. To specify the aim of this sub-examination, sub-research question 3 is formulated to provide direction.

Sub-research question 3: What specific features among the visual and textual indicators extracted from psychotherapy sessions exhibit the strongest predictive power for determining the perception of the working alliance quality?

By answering these questions, we will provide a valuable contribution to the field of psychotherapy research and hopefully also provide a tool for therapists to improve their insight into and the quality of psychotherapy.

1.7 Main takeaways of this thesis

This thesis aims to contribute to the existing knowledge on working alliance prediction and multimodal predictability using automated tools. The most straightforward contribution is a better understanding of the working alliance perception by the patient and the therapist which can help therapists to make adjustments in their therapeutic style to improve the working alliance perception of the patient and improve therapeutic outcomes, as we discussed in Section 1.5. This thesis also contributes to the field of Artificial Intelligence by applying different models to an authentic dataset and evaluating how well each model performs for their applied purpose. We use multiple pre-trained models for extracting indicators from the data and we also use models for predicting the working alliance based on the combined extracted features.

There are multiple main takeaways from this thesis, and we will give a short overview so the reader is informed before reading the specifics underlying these findings. We tested the correlation between each extracted indicator and the WAI score and found several features that showed a significant correlation. There were many significant correlations which would be too unclear to state in this Section, but we will discuss all correlations in Section 4.1 and they are displayed in Figure 8.3.1. Important is that most modalities had some significant correlations with the WAI, highlighting the multimodal characteristic of the working alliance. There were different indicators that showed significant correlations for the patient, the therapist, and the observer, indicating that their perception of the working alliance is based on different indicators.

The performance of the multimodal machine learning models highlights the importance of a large training dataset for predicting the working alliance as it is a complicated concept with various elements that can only be modeled well with a large and comprehensive dataset. We compared different models and found that the Random Forest (RF), Support Vector Regressor (SVR), and Extreme Gradient Boost (XGB) regression models showed promising results that should be investigated further. Further, the most significant finding is the Support Vector Machine (SVM) that was able to classify the therapeutic sessions with 80-90 per cent accuracy into having a low or high working alliance based on the multimodal extracted indicators.

1.8 Structure of the thesis

Chapter 2 provides an overview of the field and the related work in this area. Chapter 3 outlines our main methodology. This will be followed by Chapter 4 providing experimental results. Then, their discussion and our conclusion in Chapter 5 and finally the main conclusion and some final remarks in Chapter 6. The Appendices (8, 9) show additional data and figures for a more detailed understanding of the results.

2. Related work

2.1 Working alliance and related concepts

2.1.1 Definition and components of working alliance

In this research, we look at working alliance as reflective of psychotherapy effectiveness, as it has been shown that working alliance is a reflection of the patient's preference in therapy condition (90). Patients who receive therapy that matches their preferred therapy conditions show greater improvement in the treatment outcome (132). Not only is the working alliance the most widely used measure of process-outcome association in psychotherapy research (122), but it is also shown to have a direct association with the psychotherapy outcome (47; 37). Therefore, it is important to have an indication of the state of the working alliance so intervention is possible in cases where it is low to improve treatment outcomes.

The working alliance consists of three components: an agreement of goals, collaboration on tasks, and the strength of the bond between patient and therapist (60). The bond component refers to the emotional connection and trust that develops between the therapist and patient, while the tasks component refers to the specific actions that the therapist and patient undertake to achieve the therapeutic goals. The goals component refers to the shared understanding of what the patient hopes to achieve through therapy.

As mentioned in the introduction, the working alliance can be measured using the working alliance inventory (WAI). This self-report questionnaire was developed by Bordin in 1979 and has been used in many studies on working alliance since then (18). The full WAI consists of 36 items that assess the perception of the three components of the working alliance: goal, task, and bond. Therefore, using the WAI, the strength of the working alliance can be measured for each component individually which creates an insight into the state of the therapeutic relationship and can assist the therapist in improving specific aspects of the working alliance. It includes questions such as "I believe my therapist likes me" and "My therapist and I are working towards the same goal".

There also exist several shorter versions of the WAI which still capture the three components of the working alliance well (56). In this research, the WAI was filled in by the patients, by the therapists, and also by independent observers. A short

WAI (WAI-S) that consists of only 12 items was used in this research and was filled in by the patients and by the observers. The therapists filled in a different WAI questionnaire, the Working Alliance Inventory - Short Revised -Therapist (WAI-SRT), which consisted of only 10 questions. Each of the questions in the WAI-S and the WAI-SRT corresponds to a specific component of the working alliance. Table 2 shows to which component of the working alliance the specific questions of the WAI-S and the WAI-SRT correspond.

Table 2. Questions of the WAI-SRT and WAI-S that correspond to the working alliance components: bond, goal, and task

Component	Questions of the WAI-Scale	
	WAI-SRT	WAI-S
Bond	2, 5, 7, 9	3, 5, 7, 9
Goal	3, 4, 8	1, 4, 8, 11
Task	1, 2, 10	2, 6, 10, 12

2.1.2 Related concepts: trust, symmetry, rapport

The real relationship

The relationship between therapist and patient plays a significant role in the success of psychotherapy, regardless of the specific type of treatment used (106). The working alliance is an important part of the therapeutic relationship, as it describes the collaboration between therapist and patient. However, it does not describe the totality of the relationship, as that also contains a personal bond, or as it is often termed *the real relationship*. The relationship between therapist and patient is often divided into the working alliance and the real relationship. However, there exist different opinions on whether the real relationship is part of the working relationship ((103)), or whether a distinction must be made between the bond component of the working alliance and the bond component of the real relationship (51). According to the latter, a greater emotional bond and a greater agreement on goals and tasks will entail a stronger working alliance as the concepts of the real relationship and the working alliance is intertwined. They provide a very convincing argument for why the real relationship and working alliance should be seen as intertwined, but still as distinct concepts. The argumentation was a result of research into the effect of rupture in the therapeutic process by (51).

A rupture is defined as a breach in the therapeutic alliance resulting from a disagreement between the therapist and patient on treatment goals, a lack of collaboration, or a breach in the emotional bond. A rupture can occur in the therapeutic alliance and the personal bond between therapist and patient, but both have different impacts on the therapeutic process, as shown by (51). This study showed that a rupture in the working alliance can be repaired, but a rupture in the personal bond can not. This finding argues for a distinction between the concepts of working alliance and the real relationship. However, a rupture in both has an impact on the therapeutic process and can strain a positive therapeutic outcome.

Rapport

Working alliance is closely related to several other concepts that are important in the field of psychotherapy. One of these related concepts is therapeutic rapport, which refers to the process of building a trusting relationship between the therapist and the patient. According to Newhill et al. (2003), therapeutic rapport is essential for effective therapy because it allows the patient to feel safe and understood, which in turn allows them to explore their thoughts and feelings more deeply (103).

While rapport and working alliance are often used interchangeably, there is a slight difference between the concepts. Rapport refers to the positive relationship and understanding established between a therapist and patient, characterized by mutual trust and respect, open communication, and a sense of comfort and safety. The working alliance, on the other hand, refers to the collaborative partnership formed between the therapist and patient to work towards achieving therapeutic goals. While the two concepts are related, they are distinct in that rapport focuses only on the relationship itself, while the therapeutic alliance focuses on the specific goals, tasks, and the bond (relationship) being worked on in therapy. Rogers (2015) tested the relationship between rapport and alliance in the context of student learning and found evidence that suggests that working alliance is the broader concept, incorporating rapport (121).

Patient engagement

Patient engagement is also closely related to the concept of working alliance. As noted by Duncan, Miller, Wampold, and Hubble (2010), patient engagement refers to the active participation of the patient in therapy, and it is considered a key predictor of therapeutic outcome (41; 105). When patients are actively engaged in therapy, they are more likely to achieve their therapeutic goals and have better outcomes. Patient engagement can be seen as overlapping with the working alliance, as high engagement goes hand in hand with a high working alliance, but they are distinct in the sense that engagement is an expression of the working alliance. In practice, this means that if a patient's view of the working alliance is very high, then the engagement of the patient is higher. This has been shown by Sturgiss et al. (2016) where in a study on working alliance in obesity treatment, a strong working alliance was found to be linked to a high level of patient participation in the weight management program, which showed itself by a high number of appointments attended (130). Thus, while engagement is often not distinctly measured alongside working alliance, it can not be separated from the topic of the working alliance altogether as it is implicitly incorporated into the working alliance.

Trust

Similar to the concepts of rapport and engagement, trust is also a concept that is related to working alliance. Trust is an essential component of a bond and it is crucial in the therapeutic alliance. Studies have shown that a strong therapeutic alliance is positively correlated with trust (50). While engagement can be seen as a *consequence* of a strong working alliance, trust is a *component* of the working alliance. More specifically, trust is an important part of the bond between patient and therapist and, therefore, a good measure of the bond component of working alliance (50). Trust refers to the patient's perception of the therapist's ability to understand and help them. It also refers to the therapist's trust in the engagement and dedication of the patient in the therapeutic process. It is important to have trust in a therapeutic relationship because it helps create a sense of safety and security, which lowers the threshold for the patient to share personal and sensitive information which is important to work towards change.

Trust can be built in various ways, such as by the therapist showing active and non-judgmental listening and being open and honest with the patient. Trust can also be strengthened by the therapist's ability to establish clear goals and tasks, and by maintaining consistency and reliability in the therapeutic relationship. Trust can be measured by various tools such as the Trust in Physician Scale (TIPS) (6).

While these concepts are all related to the working alliance, we will not use them explicitly in this thesis. However, as they are often used in related work, it is important to know how they relate to the working alliance to be able to place this thesis in the context of previous and future studies.

2.1.3 Limitations of measuring working alliance

While the working alliance has been widely recognized as a valuable measure in the field of psychotherapy, there are some disadvantages to its use as well. Scientific studies have highlighted several limitations to using working alliance as the only measure of therapeutic effectiveness.

Subjectivity

One of the limitations of using the working alliance as a measure of therapeutic quality is its subjectivity. Working alliance is a subjective measure that is influenced by the perceptions and expectations of both the therapist and client (Bordin, 1979) (18). This can lead to different interpretations of the alliance and can impact its reliability as a measure of therapeutic effectiveness.

Cultural sensitivity

Lack of cultural sensitivity: Some studies have found that working alliance may not be an appropriate measure for all cultural groups, as cultural differences can influence perceptions of the therapeutic relationship (140). As an example, Hynes (2019) found that the engagement of East Asian Americans in psychotherapy is much lower than that of white and minority groups (69). This will cause the working alliance measured in psychotherapy of East Asian people to be lower than that of white and minority groups, but the working alliance measure will not provide a correct explanation for this. Namely, the lower working alliance is likely due to the family-oriented mindset that is present in East Asian culture. Since cultural sensitivity is not incorporated in the WAI, the reason behind the low working alliance will not show from using the WAI. Hynes demonstrates in this paper how cultural differences might influence the working alliance between therapist and patient.

2.1.4 Benefits of measuring working alliance

Despite the limitations and potential drawbacks of using working alliance as a measure in psychotherapy, it remains a widely used and valuable tool in the field. There are several reasons why the working alliance is an important measure to consider, even in light of its limitations. Most of the advantages have already been described, but to bring the most important one to the attention again: working alliance has been shown to have a consistently positive relationship with the therapy outcome and is, therefore, a reliable tool to measure the state of the professional relationship between therapist and patient (62; 106).

In conclusion, while the working alliance can be a valuable measure in psychotherapy, it is important to be aware of its limitations. It must be realized that the working alliance should be seen as a tool in psychotherapy sessions, but not as the definitive truth about the relationship between therapist and patient. Using the working alliance as part of a comprehensive evaluation of the therapeutic relationship, rather than as a sole measure, can give the therapist a more complete and accurate picture of the state of the therapeutic relationship.

2.2 Measurement of working alliance

2.2.1 How to measure

In the previous section, we established that measuring working alliance is an invaluable part of the research into the therapeutic relationship between therapist and patient. To measure the working alliance one can look at different elements that are observable in a therapeutic session that are indicative of the working alliance. Such indicators can be extracted from different domains, such as facial expressions from the visual domain, or the type of words used in a conversation from the linguistic domain. Some of these indicators can be measured manually, which is often done by having master psychology students observe a recording of a therapeutic session and annotate the time and frequency of each indicator observed.

While manual annotation works reasonably well, it is very inconvenient as it is time-consuming and limited by the indicators that can be observed. For example, a human observer can extract head gestures such as nodding but will be unable to annotate all exact facial expressions that are visible on someone's face during the session as these expressions are often too quick and inconspicuous for humans to pick up on.

The working alliance can also be determined without using specific indicators, but by filling in the WAI as described in the previous section. This questionnaire indicates the perception of the working alliance of the therapist and the patient. However, filling out a questionnaire is also time-consuming for the patient and the therapist and is, therefore, not done in every session. Thus, the WAI does not provide immediate feedback on the perception of the working alliance after every therapeutic session.

A better alternative to using manual annotation and using the WAI is to automate the calculation of the working alliance using computational methods. Computers can be trained to extract many types of indicators from a recording of a therapy session and process them immediately to predict a working alliance score. Language processing in particular is a good example of the superiority of automation working alliance prediction because humans will have to manually count every type of word that is said during the conversation to determine the usage of specific word categories that can be predictive of the patient's emotional state while computers can do this almost instantaneously. Therefore, it will be very beneficial and time-saving to use automated methods to predict working alliance to give feedback to the therapist.

The use of automated systems for detecting working alliances has its benefits, but it also requires a large amount of training. The automated system must learn to distinguish which indicators are representative of a high working alliance and which indicators are representative of a low working alliance. For this, the system needs a ground truth measure which is a label or score of every psychotherapy session that states what the working alliance score is. Using this score, the system can check whether its prediction of working alliance for a specific session is correct or whether it needs to adjust itself to improve the prediction accuracy. The ground truth must be accurate, otherwise, the system would try to improve itself incorrectly.

The ground truths in this thesis will be WAI scores that were measured in each psychotherapy session and that will serve as the actual working alliance score of the session. We will use the WAI scores as rated by the patients, the therapists, and the observers to have an extensive view of the perception of the working alliance according to multiple parties.

The scores provided by the patients are likely most useful for the goal of improving therapeutic outcomes, as the perception of the patients is most important for this aspect. There are differences in the WAI ratings between these three parties, as we have mentioned before, therapists are often wrong about the working alliance perception of the patient. It will be interesting to see whether the indicators that predict the WAI scores will also be different between the three parties. See Section 5.1 for a discussion of this point in light of our analysis.

2.2.2 Performance measures

The indicators that the automated system will extract have to be analyzed to find out whether they are good indicators of the working alliance or not. The performance measures of the indicators will differ per indicator as to what is custom in related studies. For example, the transcription of the speech from the psychotherapy session will be evaluated using the Word Error Rate (WER), the Match Error Rate (MER), and Word Insertion Likelihood (WIL), as these three measures have been shown to give a good representation of transcription accuracies (97; 119).

The accuracy of the speech diarization will be evaluated using the Diarization Error Rate (DER) and the Jaccard Error Rate (JER) (107; 123). Further explanation on these measures can be found in Section 3.4.2. We will extract multiple indicators from the data, for instance, facial emotional expressions. We will test the correlation between each individual feature and the WAI scores using a Spearman correlation test. Further, we will train machine learning models on the combined extracted feature set. We will test how well the models can predict the WAI scores by evaluating their fit to the data using the R-squared, the Root Mean Squared Error (RMSE) values, and the Coefficient of Variation (CV). A more extensive explanation and argumentation behind the choice of these methods is given in Section 3.5.3.

2.3 Automatic detection of relevant indicators

This section will explain the indicators of working alliance that will be used in this thesis and will give an overview of the related studies using each indicator.

2.3.1 Facial Indicators

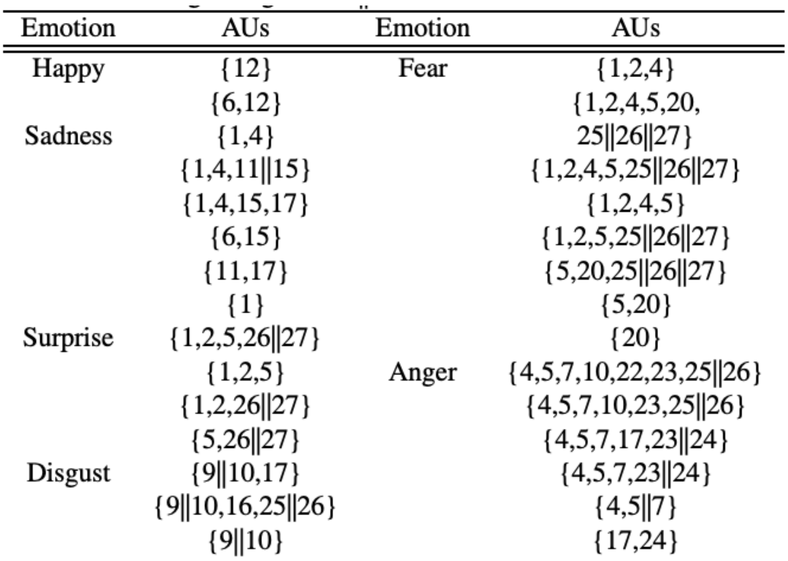
Measuring facial movement using the Facial Action Coding System

One of the most significant parts of bodily indicators is facial analysis. Facial action coding system (FACS) is a widely used method for measuring and describing facial expressions (44). It was developed by Paul Ekman and Wallace Friesen in the 1970s and is based on the idea that facial expressions are composed of basic, universal actions of the muscles in the face. An advantage of using FACS is that facial movement is an objective measure without requiring the interpretation of humans. Therefore, it is a method that can be used to objectively taxonomize facial movement. It does this by classifying facial movement into different components of single muscle activity called Action Units (AUs). These, in turn, can be used to classify facial expressions or head movements.

Facial Expressions

Being able to recognize facial expressions or emotions using computational approaches is beneficial as studies have shown that the level of empathy shown by the therapist explains nine per cent of the variance in therapeutic outcome (1). It is, therefore, crucial for the therapist to be able to recognize the emotion and the emotional intensity of their patients. However, therapists are not very good at perceiving the intensity of the emotions of their patients. To demonstrate, Machado et al. (1999) found that trained therapists do not perform better than psychology students in recognizing emotion intensity (88). Using computational approaches to measure emotion and emotion intensity can support the therapist in interpreting the patients, thus being able to show empathy more appropriately.

FACS has been used in research into expressions by using certain combinations of AUs as underlying specific emotions (78; 128; 146; 91; 10). FACS was originally meant as an anatomical facial muscle detection, and not as an expression classification system. However, psychological research has developed a method of exploring the AUs to detect facial emotional expressions.



Emotion	AUs	Emotion	AUs
Happy	{12}	Fear	{1,2,4}
	{6,12}		{1,2,4,5,20,
Sadness	{1,4}		25||26||27}
	{1,4,11||15}		{1,2,4,5,25||26||27}
	{1,4,15,17}		{1,2,4,5}
	{6,15}		{1,2,5,25||26||27}
	{11,17}		{5,20,25||26||27}
	{1}		{5,20}
Surprise	{1,2,5,26||27}		{20}
	{1,2,5}	Anger	{4,5,7,10,22,23,25||26}
	{1,2,26||27}		{4,5,7,10,23,25||26}
	{5,26||27}		{4,5,7,17,23||24}
Disgust	{9||10,17}		{4,5,7,23||24}
	{9||10,16,25||26}		{4,5||7}
	{9||10}		{17,24}

Figure 1. Rules for mapping Action Units to emotions according to a rule-based FACS method. This table is from (138).

Earlier approaches to detecting emotional expressions from FACS were rule-based and relied on a set of predefined rules for which AUs must be activated for a specific emotion. (138) uses a rule-based FACS method to measure emotions from AUs and a figure from this paper is shown in Figure 1.

Valstar and Pantic (2006) researched facial emotion encoding in videos with biologically-inspired artificial neural networks (ANN) and the logical rule-based method described above (138). They found that the ANN outperformed the classic rule-based methods, suggesting that using machine learning techniques such as ANN should become the standard for classifying emotion based on AUs. Since this study was in 2006, deep learning methods have become the most common method in detection, prediction or classification tasks, due to their very high accuracy (83). Recent research reinforced this finding, for instance, the study by Siam et al. (2022) proved that using the FACS system as an input for a machine learning classifier, can recognize facial expressions with a high accuracy of 97% (127).

An advantage of using a rule-based method over deep-learning models is that they are easy to implement and do not require extensive training or much computational power. However, rule-based methods are limited in their capabilities because they are very inflexible and are thus unable to adjust predictions to a specific person, and although they are robust, they are not as accurate as some deep-learning methods (92). Another advantage of using deep-learning models is that they can include contextual information such as previously detected emotions. Also, the working alliance is very dynamic as it can have different meanings at different times during the therapeutic process. Thus, having a flexible deep-learning model is likely to better represent the fluctuating working alliance than a rule-based model.

A limitation of using facial expressions as an indicator for working alliance prediction is that not every facial movement or expression is representative of an emotion. Also, contrary to Ekman's theory it has been shown that emotions are not as universal as previously thought. Some studies found more than the six basic emotions proposed by Ekman (29). Further, there are also differences in the display of emotion between cultures. Boiger et al. (2018) studied facial expressions in Japanese, US, and Belgian cultures and they found that the distribution of emotions differed between these cultures (15). Moreover, the emotion behind facial expressions can only be accurately detected if the context is known. To demonstrate this, the photos in Figure 2 show two similar facial expressions, but in photo A the underlying emotion seems to be anger, while in photo B the context is included which also shows a fist pumped into the air which is a sign of extreme happiness. The context in this photo was Serena William winning the US Open tennis game in 2008. It is likely that without the context of the full photo or even the context of the won game, the correct emotion of ecstatic happiness would not be detected.

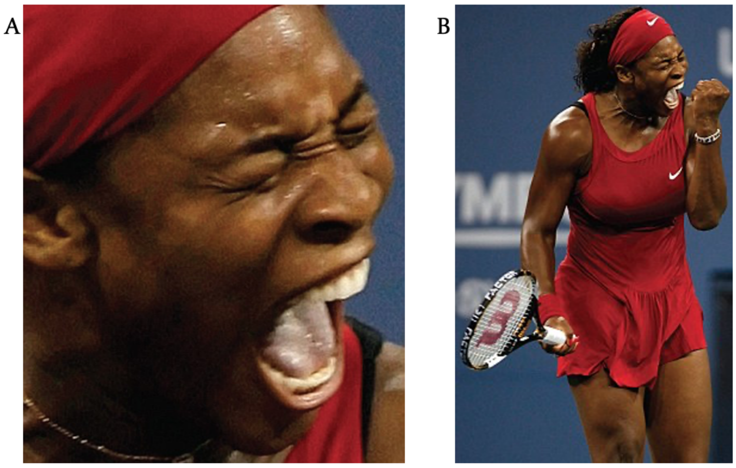


Figure 2. Photo of Serena Williams winning the 2008 US Open. A) cropped photo to show the facial expression. B) Full photo
Source: (4)

Fortunately, this challenge that comes with detecting emotion from facial expressions is less of an issue in this thesis, because the facial expression is one of many indicators that will be used in a multimodal prediction model. The context belonging to a facial expression is, therefore, included with the facial expression meaning that the accuracy of detecting the correct emotion will likely be increased compared to predicting emotion using only the facial expression.

Limitations of the FACS

There are some limitations of FACS, as it can only detect clear changes in muscle movement and is unable to capture subtle muscle movement (43). Moreover, FACS was not originally meant as an emotion classification system and does, therefore, not detect other physiological changes corresponding to an emotion such as tears, changes in skin color, or breathing. This limits its accurate representation of an emotion.

Measuring trust using Action Units

As mentioned in Section 2, trust is a component of the working alliance. Therefore, trust is likely a good indicator of the state of the working alliance. AUs are not only useful in analyzing facial expressions but they can also be used to determine trust between two people by analyzing the synchrony of the displayed AUs during social interaction (93). A requirement for measuring trust based on the synchrony of displayed AUs is, however, that both people will need to be visible in the video. Unfortunately, this is often not true in the dataset used in this research (see Section 3.2).

Microexpressions

In recent years, emotion recognition based on AUs has made significant progress (84; 110; 147; 148). Whereas most research focuses on macro expressions, microexpression studies have also made progress in their classification (84). Microexpressions are brief, spontaneous facial expressions that occur in response to emotions. They are thought to reflect unconscious emotional responses and have been used as an indicator of working alliance in psychotherapy. Microexpressions are potentially better than facial expressions as they reflect a person's inner motives and emotions as they are the result of unconscious processes.

Research has shown that therapists who can accurately identify their clients' microexpressions are more likely to form strong working alliances with their clients. For example, a study by Datz et al. (2019) found that therapists who had psychoanalytic backgrounds received higher WAI scores compared to therapists without psychoanalytic backgrounds, suggesting that having been trained to recognize microexpressions strengthens the working alliance (30).

A limitation of microexpressions is that they require a high frame rate and resolution of the video to be captured. Unfortunately, as with the trust indicator based on synchrony, the dataset is unsuited for the detection of this feature. The dataset that was used does not have a high enough frame rate to capture microexpressions. However, we did want to mention them here because they are likely a promising feature indicative of the working alliance and thus could be looked at in future research.

Eye-contact

The gaze direction of someone's eyes is an important measure in the psychotherapeutic context as it can be used to detect eye contact and aversion to eye contact between therapist and patient. Research has shown that eye contact can indicate a strong bond (74), whereas aversion to eye contact is indicative of someone trying to attenuate the interpersonal relationship and thus is a sign of decreased working alliance (53; 5). A therapist making eye contact is also perceived by the patient as engaging and friendly, stimulating a strong and positive working alliance (49). Fortunately, eye contact can be measured using the OpenFace toolbox (9), as it automatically detects someone's gaze and can thus estimate whether the gaze of two people is towards each other.

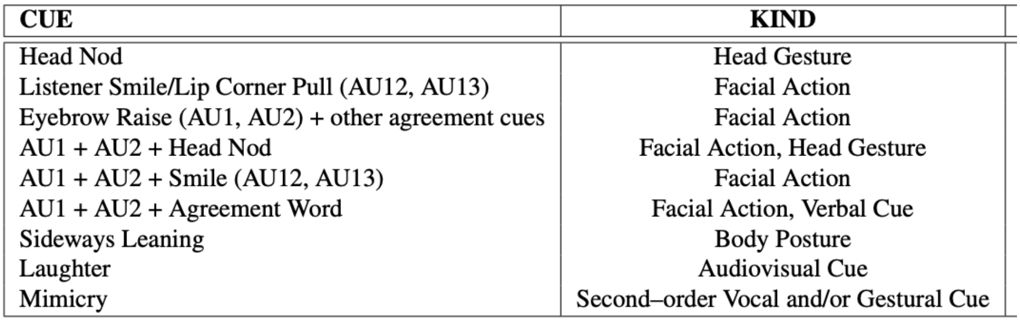
The difference in cultural perception of eye contact should be taken into account. In more Eastern cultures, as opposed to Western cultures, eye contact is seen as a sign of respect and authority. Therefore, a person making eye contact is a sign of establishing authority over someone else. In the case of psychotherapy, this will work detrimentally as it can negatively impact the working alliance perception if there is authority behavior from the therapist (68). In this thesis, we work with a dataset consisting of Dutch citizens, which means that we don't have to take the cultural difference of eye-contact perception into account, but further research containing people with an Eastern culture should be aware of this.

2.3.2 Body analysis

Head Gestures

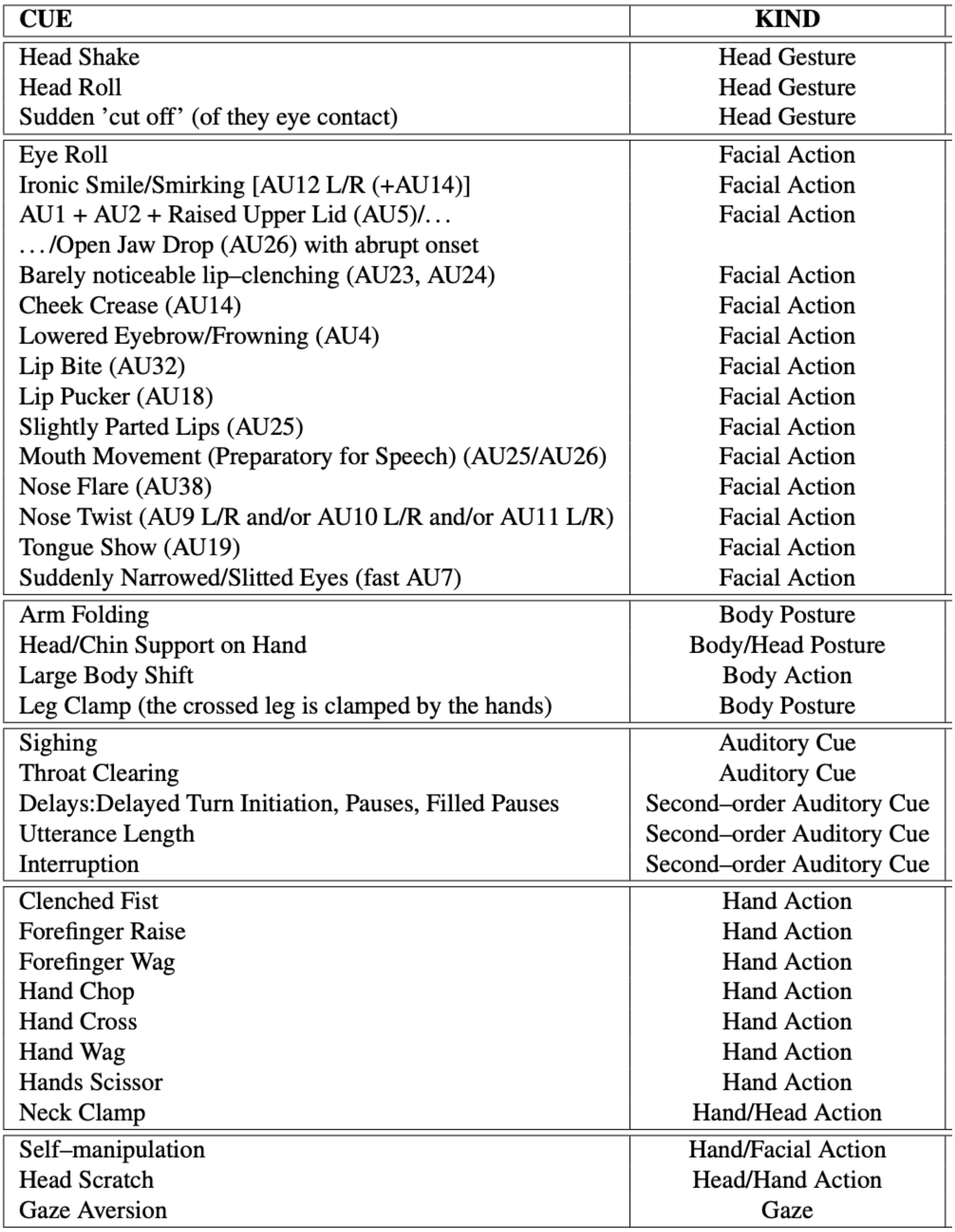
Head gestures, such as nodding and shaking, are nonverbal behaviors that involve movement of the head and neck. These gestures can convey various emotions, attitudes, and meanings and can play an important role in the formation of a strong working alliance in psychotherapy.

The working alliance is a representation of how much the patient and therapist are aligned. Therefore, an important indicator of the working alliance is the amount of agreement or disagreement between people. If the patient displays a lot of disagreement, then the therapist and patient are likely not aligned and thus don't have a good working alliance. Whereas a lot of agreement from the patient signals a strong working alliance with the therapist as that is indicative of the therapist and the patient having the same views. It is, therefore, important to pay attention to extracting agreement and disagreement from the interaction between patient and therapist. Research into the signaling of agreement and disagreement has found which gestures are indicative of agreement and disagreement, so we can use that to extract them as indicators for working alliance. A meta-analysis by Bousmalis et al. (2009) on the use of gestures to signal agreement or disagreement found that head nods are typical signs of agreement and head shakes or tilts are typical signs of disagreement (19). This paper gives an overview of all bodily, head, and facial movements that can be linked to either agreement or disagreement. This overview is very relevant for this thesis as the mentioned gestures and the research supporting them are given, and can thus be used to extract from our dataset. An overview of these gestures is from (19) and displayed in Table 3 and Table 4.



CUE	KIND
Head Nod	Head Gesture
Listener Smile/Lip Corner Pull (AU12, AU13)	Facial Action
Eyebrow Raise (AU1, AU2) + other agreement cues	Facial Action
AU1 + AU2 + Head Nod	Facial Action, Head Gesture
AU1 + AU2 + Smile (AU12, AU13)	Facial Action
AU1 + AU2 + Agreement Word	Facial Action, Verbal Cue
Sideways Leaning	Body Posture
Laughter	Audiovisual Cue
Mimicry	Second–order Vocal and/or Gestural Cue

Table 3. Visual indicators that are characteristic of agreement. This Table is from (19).



CUE	KIND
Head Shake	Head Gesture
Head Roll	Head Gesture
Sudden ’cut off’ (of they eye contact)	Head Gesture
Eye Roll	Facial Action
Ironic Smile/Smirking [AU12 L/R (+AU14)]	Facial Action
AU1 + AU2 + Raised Upper Lid (AU5)/...	Facial Action
.../Open Jaw Drop (AU26) with abrupt onset	
Barely noticeable lip–clenching (AU23, AU24)	Facial Action
Cheek Crease (AU14)	Facial Action
Lowered Eyebrow/Frowning (AU4)	Facial Action
Lip Bite (AU32)	Facial Action
Lip Pucker (AU18)	Facial Action
Slightly Parted Lips (AU25)	Facial Action
Mouth Movement (Preparatory for Speech) (AU25/AU26)	Facial Action
Nose Flare (AU38)	Facial Action
Nose Twist (AU9 L/R and/or AU10 L/R and/or AU11 L/R)	Facial Action
Tongue Show (AU19)	Facial Action
Suddenly Narrowed/Slitted Eyes (fast AU7)	Facial Action
Arm Folding	Body Posture
Head/Chin Support on Hand	Body/Head Posture
Large Body Shift	Body Action
Leg Clamp (the crossed leg is clamped by the hands)	Body Posture
Sighing	Auditory Cue
Throat Clearing	Auditory Cue
Delays:Delayed Turn Initiation, Pauses, Filled Pauses	Second–order Auditory Cue
Utterance Length	Second–order Auditory Cue
Interruption	Second–order Auditory Cue
Clenched Fist	Hand Action
Forefinger Raise	Hand Action
Forefinger Wag	Hand Action
Hand Chop	Hand Action
Hand Cross	Hand Action
Hand Wag	Hand Action
Hands Scissor	Hand Action
Neck Clamp	Hand/Head Action
Self–manipulation	Hand/Facial Action
Head Scratch	Head/Hand Action
Gaze Aversion	Gaze

Table 4. Visual indicators that are characteristic of disagreement. This Table is from (19).

This paper also mentions body gestures which are also indicative of agreement or disagreement and therefore also useful for indicating alliance. However, a study by Muller et al. (2022) found that body posture is not as strong a predictive feature for agreement detection as head gestures are (99). In this study, they researched backchanneling activities such as gestures: nodding, shaking, gaze, body posture, and vocal features such as humming. It was found that head gestures were most predictive of agreement while body posture and vocal features were least predictive.

A study by Vail et al. (2021) looked at social and bodily features concerning working alliance (137). They studied six features: head nods, head shakes, speaking turn length, the waiting time (the duration of the pause between the end of one speaker's turn and the beginning of the other speaker's turn), listening nods, and listening shakes (137). They found that the behavior of a person corresponded more with their rating of working alliance than the rating of their partner. This tells us that it is more useful to look at the behavior of the patient than that of the therapist if we want to know the patient's estimation of the working alliance. In this thesis, we want to specifically know the patient's perception of the working alliance, therefore, we should focus mainly on the patient's behavior and less on the therapist's behavior. Also, head gestures were found to be specifically predictive of the patient's rating of the working alliance.

A limitation of many studies on gestures is that they do not use naturalistic data but simulated datasets. While the assumption is that many detection systems that are trained on simulated datasets can be applied in naturalistic settings, this is not yet tested extensively.

Body Posture

Interestingly, while (99) found that body posture was not very indicative of the working alliance, other studies have found the contrary: body posture can be predictive of the bond between two people. An open body posture is characterized by a slightly forward lean and arms in an open (not crossed) position and is perceived by patients as more empathetic (22). For instance, research by de Roten et al (1999) has shown that physical proximity, along with other nonverbal behaviors, can be a predictor of therapeutic alliance (33). An open body posture and small physical distance can signal intimacy, trust, and mutual understanding, which are all important components of a strong therapeutic alliance (39). Thus, a small distance between two people or people leaning toward each other is associated with a higher working alliance.

A disadvantage of body posture is that it is difficult to analyze in a therapeutic recorded setting, as the visual does not always include the full body of the therapist and the patient. However, if possible, body posture can be a good indicator of the working alliance.

Backchanneling

Another good predictor of engagement in a conversation, and thus of working alliance is backchanneling, which entails the reaction of the listener during the speaker's turn (149). Backchanneling refers to the cues that listeners use to signal to speakers that they are paying attention and engaged in conversation. It can be both vocal using utterances like "yes", "hmm", or nonvocal where the listener shows head gestures like nodding or shaking.

In psychotherapy, backchanneling is an important aspect of the therapeutic relationship and can play a crucial role in forming a strong working alliance. It can be used as a measure of engagement but also as a measure of agreement (99), both of which can be potentially good predictors for the perception of the working alliance. A study by Bavelas et al. (2000) into the role of backchanneling in storytelling, showed that narrators are better at telling a story when listeners display backchanneling behavior, compared to when the listener displays no backchanneling behavior (11). In this study, a narrator was reading a story, and the listeners were tasked to show either passive behavior (no backchanneling), non-specific backchanneling (humming, nodding), or specific backchanneling behavior such as exclaiming at specific moments in the story. The more specific the backchanneling behavior was, the better the narrator was able to read the story. This shows that the listener's behavior during a conversation is important.

Backchanneling has also been researched for use in human-robot interaction where it has been shown that backchanneling stimulates engagement of the participant with the robot (73). In a study where people interacted with the smart speaker Alexa, it was found that having Alexa use backchanneling would result in people speaking to it for longer periods and would cause more sustained user engagement (98).

Interestingly, (99) found that nonvocal behavior is the best indicator of engagement and agreement in a conversation. In this study, they measured the predictive value of vocal, head, and bodily features for engagement and agreement in a conversation. They found that nonvocal behavior specifically head pose and the combination of head and body pose were most predictive. The vocal features were found to be not predictive of engagement and agreement in a conversation. Since engagement and agreement both correspond to the working alliance, we can assume that extracting backchanneling behavior as a predictive feature for the working alliance is likely very beneficial. We can also assume that nonvocal behavior such as head movement is probably a good indicator of the working alliance.

Synchrony

Synchrony is not just important to determine trust between therapist and patient, it plays a larger role in reading working alliance from a psychotherapy session. A study by Ramseyer and Tschacher (2011) showed that high nonverbal synchrony, so the synchrony between the movement of therapist and patient, was indicative of a strong working alliance perception by the patient (115). Building on this finding, these authors released a paper in 2014 in which they explored the relationship between specific head and body gestures and therapeutic outcome (116). The results of this study showed that synchrony of head gestures predicted the overall outcome of the psychotherapy sessions, while synchrony of bodily gestures did not. However, synchrony of bodily gestures was predictive of session outcome, while synchrony of head gestures was not. This suggests that there are distinct systems that underlie head and body gestures. Synchrony in head gestures is more indicative of the working alliance over time, while synchrony of bodily gestures is more indicative of the working alliance at a certain moment.

Another interesting characteristic of synchrony is that is known for being a good indicator of trust between people (118; 93). Moreover, movement synchrony has been shown to correlate with rapport. In a study by Bernieri (1988), participants had to observe videos of an interaction between a teacher and student and rate the strength of the relationship (13). The relationship was rated as being stronger in the clips where there was a lot of movement synchrony, compared to the clips with little movement synchrony. A more recent study found that movement synchrony increases affiliation (63). Since affiliation is part of the bond component of the working alliance, this indicates that synchrony will affect the working alliance in a therapeutic setting as well. The relation between synchrony and affiliation also works the other way around as Lakin and Chartrand (2003) have shown (81). In this study, it was found that a feeling of high affiliation and rapport between people will cause increased use of mimicry.

To summarize, interpersonal synchrony is an important aspect of social interaction and can have a significant impact on social perception. Interpersonal synchrony between two individuals can lead to the formation of a strong therapeutic alliance and increase the social perception of likability, trustworthiness, and competence (94; 81). Synchrony is likely a very interesting feature to include as a possible predictor for working alliance. However, synchrony again requires video of both the therapist and the patient which is only possible in a small subset of our dataset.

As we have seen, research has shown the influence of movement synchrony on aspects related to working alliance. There is, however, another method of measuring synchrony that applies to a dataset that lacks visuals of both therapist and patient. Namely, synchrony in language, or language entrainment. This will be discussed in the next section on using voice and speech features as possible indicators for the working alliance.

2.3.3 Voice and speech

Language Entrainment

Language entrainment is a phenomenon where speakers start to adapt their language to each other's language over time. This means that the language styles become more similar as time progresses.

In the field of human-robot interaction has been shown that children prefer a robot that uses language entrainment as opposed to a robot that does not (80). The robot using speech entrainment resulted in an increase in children's engagement and an improvement in the child's perception of the relationship.

Research has also shown that language entrainment can play an important part in the forming of the working alliance in psychotherapy. This was shown in a study by Vail et al. (2022), who researched language entrainment in therapist-patient interactions (136). The authors used a similar dataset to the current study and found that the language entrainment of the therapist significantly impacted the patient's perception of the working alliance. Additionally, they found that the language entrainment of the patient was a strong predictor of their perception of the working alliance.

This study used a method known as reciprocal linguistic style matching (rLSM) metric to predict language entrainment. rLSM is a method that can determine how much the language styles of two people change to be more similar over time. It simply tests whether the people adapt their language style to match that of the other person. This matching can occur in different aspects of language style, such as word choice or grammar. In this context, rLSM is a predictor of the working alliance between the therapist and the patient. If the language style of the therapist and patient become more similar during a psychotherapy session, it can indicate a stronger working alliance between them. Conversely, if the language styles of the therapist and the patient become more dissimilar during the session, it can indicate a weaker working alliance. Vail et al. found that more language entrainment by the therapist was associated with a higher perceived working alliance by the client (136). This finding demonstrates that the use of language in psychotherapy can be an important indicator of the working alliance and, therefore, a feature worth exploring.

A limitation of this study is that the authors did not include other social behaviors such as gestures, facial expressions, and posture in their analysis. If language entrainment is used along with other possible indicators such as gestures, it may explain a smaller percentage of the variance of working alliance than in the study of Vail et al., because another indicator could be confounding factors that cause both language entrainment and working alliance.

Another study that proved the importance of language entrainment in working alliance prediction linked specific word categories to the WAI questions (12). The words were manually sorted into different dialogue acts: inform, agreement, offer, feedback functions, and request. It was found that the inform, agreement and offer dialogue acts correlate with the task component (questions 1, 2, 12), the feedback dialogue act correlates with the goal component (question 11), and the request act with the bond component of the working alliance (question 5). Overall, higher use of feedback, the inform, and the request dialogue acts correlate significantly with a higher WAI rating.

Word choice

Language is an important part of psychotherapy sessions as it conveys the information and the topics that are discussed. Both the content of the spoken language and the style with which the content is conveyed can tell a lot about the state of the speaker. In a therapeutic context, we mainly look at the language of the patient and can infer a lot about that person, from current mood to characteristics of personality (58; 111).

If we start to look at language a little closer, we can see that it can be divided into two main categories: function words and content words. Content words are the type of words that we attribute meaning to, they convey information in a message. Function words, or style words, refer on the other hand to more grammatical constructs in the sentence. These are words such as “the", “it", and “was". Although content words convey meaning, they make up only 0.05% of the English language, while style words make up about 55% of language (134).

Linguistic Inquiry and Word Count (LIWC) is a computational tool used to analyze text for various linguistic and psychological dimensions (113). LIWC uses a dictionary of words and categorizes them based on various dimensions, such as emotions, cognitive processes, and social processes (112). It has been widely used in various research fields, including psychology, sociology, and linguistics, to analyze written and spoken text. LIWC analyses the words used in a text and sorts them into categories and many subcategories. It counts the total number of words in every category and provides percentages of how much every category was present in the text. To give an example, the word 'cries' is part of several categories and subcategories, namely: sadness, negative emotion, overall affect, and a present tense verb. These categories range from semantic to grammatical characteristics, which demonstrates how LIWC can analyze a text comprehensively.

In the context of psychotherapy, LIWC has been very useful in studying the language patterns of therapists and patients and how they relate to the working alliance (Ryu et al.). For example, a study by Negri et al. (2019) found that patients who use higher levels of emotional language and positive emotion words in the first psychotherapy session are associated with having a stronger working alliance at the end of the first session (101).

LIWC can even be used to detect a rupture in psychotherapy sessions. This came forth from a study by Jacques and Dykeman (2022) in which they explored the linguistic features of three rupture types: confrontation, withdrawal, and mixed rupture to detect when these ruptures occur, how often, and discover the implications for the working alliance between therapist and patient (70). This study showed that using the LIWC analysis, ruptures could be detected which we know from previous research can be very damaging to the working alliance (51). Therefore, it is beneficial to try to detect these ruptures early in therapy before they deteriorate the working alliance so much that it can not be repaired and the progress in therapy stalls. A limitation of LIWC is that it does not detect sarcasm and has difficulty with idioms. This is, however, only a very small part of language and is, therefore, not a large factor in the language analysis results.

LIWC has also been adapted for the Dutch language by Peter Boot (17). Therefore, using a dataset with Dutch spoken language does not hinder the analysis of word choice using LIWC.

To summarize, we learn from these studies that studying word choice in the speech of patients and therapists will likely provide good indicators of working alliance.

Turn-taking

A key component of psychotherapy is turn-taking since it involves a rhythmic speech exchange between the therapist and the client. An alternating-speaker dialogue is important in building a therapeutic relationship between the therapist and the patient. Turn-taking in a psychotherapeutic setting is different from typical dialogues as the length of the speaking turns of the patient is typically longer than the length of the speaking turns of the therapist.

Turn-taking can be a good indicator of the quality of a conversation, as was shown in a study by Cassel (2004), in which they examined the impact of social language in an intelligent system in a robot that interacted with children by listening to and telling stories (24). It was found that children had a better relationship with the robot if there was room for long pauses between or within turns that would give the listener room to take over the turn.

In Section 2.3.2 we described a paper by Vail et al. (2021) in which the predictive value of head gestures and turn-taking for working alliance was researched (137). From this paper, we learned that the patient's perception of the working alliance is best predicted by the patient's behavior, and not necessarily the therapist's behavior. However, this paper showed some other findings which will be discussed in this section on turn-taking behavior. It was found that both head gestures and turn-taking behavior were predictive of working alliance, but that they were both indicators for a specific component of the working alliance. Head gestures were found to be more reflective of the task-oriented components and turn-taking behavior was found to be more reflective of the bond-oriented component.

Another study by Bayerl et al. (2022) found a correlation between some specific turn-level and turn-taking features and the working alliance (12). Specifically, it was shown that an equal engagement between patient and therapist is correlated with a higher working alliance score for the task, goal, and bond components (WAI-S questions 1, 2, 5, 11, 12). Furthermore, the unpredictability of the therapist's turn-taking strategy (turn-level freedom) was positively correlated with the task component of the working alliance, meaning that an unpredictable strategy of the therapist (e.g. different lengths of feedback signals, giving patients different amounts of speaking time) has a positive effect on the working alliance. Also, a lot of short moments where both speakers speak at the same time are positively correlated with task (WAI-S question 2,12) and bond (WAI-S question 5). These are likely short feedback utterances that are being given by the listener that stimulate the working alliance. Moreover, this study also found a positive correlation between the minimum observed speech rate of the therapist and the task (WAI-S question 8) and goal (WAI-S question 9) components of the working alliance. It is hypothesized by the authors that the reason behind this correlation is that the therapist is comfortable talking faster if there is more mutual trust between the patient and the therapist.

While turn-taking has not been extensively researched in a therapeutic context, the studies mentioned above provide strong evidence for its important role in working alliance. Thus, turn-taking will be an interesting feature to study in the context of this thesis.

Affect analysis

Extracting affectual features such as sentiment (positive/negative), arousal and valence, or specific emotions from a conversation can be a useful indicator of people's emotional state. Along with gestures and facial movement, features such as the tone of voice, intonation, volume, and pace are indicators of someone's emotional state. Extracting the indicators from speech that represent emotion can be done in different ways, as speech consists of many different elements that each require different analysis methods. For example, speech has acoustic features such as frequency and lexical features such as word choice. These features are so distinct that they require different methods of analysis which we will explain here.

One of the different aspects of speech is acoustic features, which can be extracted using a technique called acoustic analysis. In this method, the acoustic features of the speech are analyzed. Such features include pitch, energy, frequencies, and pause time in between speeches. Research has found that certain acoustic features are indicative of emotion. For example, if speech has a high pitch and a loud volume, this can indicate excitement or happiness (57), while the pace of speech is correlated with levels of depression: a slow pace can indicate sadness or depression (3).

A second method of speech analysis that is often used in speech research is prosodic analysis, in which the rhythm, stress, and intonation patterns in speech are extracted (79). This method can reveal the speaker's emotional state by detecting changes in speech features. Sudden changes in the tone of voice, pitch, or loudness can indicate changes in an emotional state, which makes it a useful method for emotion detection in speech.

In most trained models that extract emotion from speech, both acoustic features and prosodic features are used to make a prediction. In the case of machine learning models, these features are often not explicitly extracted from the data to base an emotion prediction on, but during training these models automatically learn distinctive features.

A third method to extract emotion from a conversation is a semantic approach called a lexical analysis, which involves analyzing the word usage of speech to determine the speaker's emotional state. In this method, the speech is transcribed into text which is analyzed to extract the word choice, the frequency of words from specific categories, and the context in which they are used to determine the emotional content of the text. For example, the use of words such as "sad" can indicate a negative emotional state, while the use of words such as "excited" can indicate a positive emotional state. Lexical analysis can be done using LIWC as described in Section 2.3.3.

The features used in acoustic and prosodic analyses can also be used for separating speech from different speakers, also called speaker diarization. Every person has a different combination of features that are produced by the differences in the vocal tract and oral anatomy. The energy in speech between people is different, which can be extracted using the Mel-frequency Cepstral Coefficients (MFCC). MFCCs are a collection of features that are extracted from a speech signal that together represent the anatomy of one's vocal tract. They are extracted by splitting the frequency range of a speech signal into various frequency bands using the Mel frequency scale. An advantage of using MFCCs is that they accurately reflect the voice of a person in a concise and low-dimensional way which makes them quick to process. They offer a representation of speech signals that is resilient to variations in speaker, accent, and emotional state which is why they have gained popularity in speech analysis. This technique has proven very efficient in speech recognition and speaker diarization (48).

To conclude, these three approaches can be used to extract emotion from speech (audio) and in the case of lexical analysis, also from the text. Extracting emotion is beneficial as we have also seen in Section 2.3.1 because the level of empathy shown by the therapist explains nine per cent of the variance in therapeutic outcome (1). Therefore, being able to extract empathy from the therapist's speech and gestures is useful, but it also requires checking if the therapist's empathy is displayed at the correct moment when the patient is emotional. This last element requires extracting emotion from the patient's gestures and speech, using the methods described here. Moreover, a meta-analysis by Peluso and Freund (2018) has shown a direct link between the emotional expression of the patient and therapist and the therapeutic outcome (109). Since the working alliance is also predictive of the therapeutic outcome, it can be useful to analyze the emotional state as there might be a direct relation between working alliance and emotional expression. For example, the emotional expression could be an indicator of the working alliance and since the working alliance is an indicator of the therapeutic outcome, the relationship between emotional expression and the therapeutic outcome could be indirect.

2.3.4 Multimodality

Many of the studies mentioned in previous sections make use of unimodal models, which means that their model input is one type of measure, for instance, only gesture input or only vocal input. While this works well in finding a correlation between one feature and the output, for instance, finding whether frequencies in voice data can predict working alliance, there is a method that has more potential prediction capability. In recent years, multimodal models have proven very effective in predicting complex concepts as complex phenomena are often composed of various modalities, as was demonstrated in previous studies with working alliance (133). Multimodality measurement refers to using multiple methods to study a certain phenomenon. This often results in more accurate predictions as phenomenon such as angry behavior is expressed in different modalities. For example, anger is expressed by raising the volume of one's voice, balling fists, and displaying a specific facial expression. These different modalities together form the expression of a certain feeling or emotion. Therefore, the use of multimodality enhances the validity of the findings and enables a more comprehensive and nuanced outcome of the topic being studied. Multimodality can be used in machine learning models where multiple modalities can be combined as input for a prediction task, such as image classification, speech recognition, or natural language processing.

A study by Schirmer et al. (2017) has shown that using multimodality to detect emotion from the face, voice, and touch results in faster and more accurate emotion judgments compared to unimodality (125). Another advantage of using multimodality in machine learning models is that it can help to overcome the limitations of single-modality models, such as the presence of noisy or incomplete data.

To the best of our knowledge, there has up till now not been a multimodal machine learning model that uses the multimodality of facial, bodily, textual, and vocal features to predict working alliance, which is why we aimed to fill that gap with this thesis. However, emotion detection research has already studied the properties of multimodal models. As we have seen, emotion detection can be predictive of the working alliance, which makes it interesting to look at as it is a part of the working alliance.

A large meta-research has found that multimodality improves emotion classification using computational analyses by 9.38 per cent on average (the more representative median corresponded to a 6.60 per cent improvement) over unimodality (38). In this analysis, over 90 studies between 2003 and 2013 were compared in which both unimodal and multimodal emotion classification systems were used. The classification accuracy with which emotion could be detected using unimodal and multimodal systems was compared, as well as the types of multimodal features measured. This study showed that multimodality has a higher classification accuracy for detecting emotion compared to unimodality. Since emotion is also a predictor of working alliance (see Section 2.3.1), we can view this meta-study as evidence that exploring multimodality in detecting working alliance is likely to provide higher accuracy than using unimodality.

While multimodality has its advantages in many fields, its use in machine learning models can also have its downsides. One major challenge in using multimodality in machine learning models is the integration of multiple data sources, which can lead to issues with data compatibility, data quality, and data interpretation. While multimodality can improve recognition performance, it also increases the complexity of the model and can lead to overfitting and decreased generalizability of the model. If multiple modalities are included in the model, a bad predictive modality may corrupt a good predictive modality thereby decreasing the accuracy of the model. Therefore, it is necessary to first research the predictive value of each modality before combining them in a multimodel model. This is also the reason why in many studies, and also this thesis, we first do extensive research to find out whether a feature is indicative of the working alliance or not. This feature selection allows us to only include features that are likely to be good predictors of the working alliance. A detailed description of some feature selection methods is described in Section 3.5.3.

Another challenge in using multimodality in machine learning models is the potential for increased computational demands, leading to longer training and prediction times. Multimodality models can require a significant amount of computational labor, particularly when working with large amounts of data. Using multiple modalities in a machine learning model can also lead to more complexity in the model which will make optimization more difficult.

Despite these disadvantages, which require some caution when implementing a multimodal model, studying the use of multimodality in working alliance prediction is likely very beneficial. In this thesis, we build on the findings of two very relevant papers from Vail et al. (2021) en Bayerl et al. (2022), (137; 12), by combining conversational, speech, textual, and visual features to try to predict working alliance accurately in a multimodal model.

2.4 Prediction of working alliance

Working alliance can be measured by the WAI as we have seen in Section 1.3, but predicting working alliance without direct information from the patient and therapist themselves is less clear-cut. There have been several studies that have attempted to predict working alliances based on the indicators that we have mentioned in the previous section, all using different statistical tests and model architectures. Here, we will describe the models that each research used and evaluate their performance.

2.4.1 Pearson Correlation

While a Pearson correlation analysis is not a machine learning model, it can still be a valuable tool for evaluating correlations between features and the working alliance, as was shown by (12). We shortly describe this study and specifically the features used in this study in Section 2.3.3. The analysis in this study was done by testing the correlations between the extracted features and the WAI scores. A Pearson correlation was done with each feature and each question of the WAI which resulted in a large number of statistical tests, but multiple strong correlations of features with the WAI. As sub-research within this thesis, we aim to try to replicate the results of this study by Bayerl et al. (2022). However, we use a Spearman correlation and explain this decision in section 3.3.4.

2.4.2 SVR, Elastic Net, and RF

The study by Vail et al. (2021) tested the predictability of head gestures and turn-taking on working alliance (137) (see Section 2.3.2). They tested three different algorithms, Support Vector Regression (SVR), Elastic Net, and Random Forests (RF), to compare their performance in predicting the working alliance. These algorithms were chosen in this study because they performed well on small datasets. In studies on psychotherapy, datasets are often small because of the difficulty of collecting data. Not everyone wants their data recorded because there is very sensitive and private information being shared. Therefore, algorithms that perform well on smaller datasets are very useful in studies like this one. (137) tested each model against a subset of the WAI, the bond, task, and goal components, and found that overall the SVR and Elastic Net performed best in predicting working alliance.

2.4.3 (RI)-CLPM, Multilayer-perceptron and MLM

In 2022, Vail et al studied how language entrainment was predictive of working alliance (136). In this study, they tested the performance of a random intercept cross-lagged panel model (RI-CLPM) and two types of Multilayer-perceptrons (MLPs) against two baseline models, a cross-lagged panel model (CLPM) and a multilevel linear model (MLM). They tested the performance of the RI-CLPM and the MLPs not using the prediction accuracy but with a method called structural equation modeling (SEM). SEM is a statistical analysis method used to evaluate multivariate causal relationships. It can find structural and causal relationships in data. In this study it means that the models are evaluated using model fit, meaning how well the model fits the data. An advantage of using SEM over standard machine learning models is that they have added interpretability and a causal analysis can be done. The RI-CLPM and the MLPs are evaluated to see if they perform better than the baseline models. The MLPs are two models, a model with one hidden layer (MLP-1) and a model with two hidden layers (MLP-2). The result of this study was that the RI-CLPM had the best fit for the data, followed by CLPM. The MLPs had a very bad fit compared to the rest of the models.

From this study, we can conclude that a RI-CLPM is an interesting model option to look at for multimodal prediction, especially considering the advantages of establishing causal relationships and interpretability.

2.4.4 fCNN and RF

The use of deep learning in predicting working alliance has been researched before by Zhou et al. (2022) who tested a fully connected neural network (fCNN) and random forest algorithm (RF) in predicting working-class alliance in first sessions psychotherapy (150). This research used self-reported indicators, which were multiple questionnaires filled in by the patient about, for instance, socio-demographic status and clinical preferences, and multiple questionnaires filled in by the therapist about, for instance, psychotherapy style and intervention orientations. Several features were extracted from these questionnaires that served as input for the predictive models. They found that the fCNN outperformed the RF and was thus a better choice for predicting the working alliance with these features. What is interesting about this study is that it used a large amount of data (325 patients and 32 psychotherapists) and was tested on data recorded in a different setting. The training data was recorded in a University counseling center and the testing data was recorded in general hospital counseling sessions. The large dataset and different settings underscore the validity of this study as well as its generalizability.

An important takeaway from this study is that using deep learning models will likely perform better for predicting working alliances than using Random Forest algorithms. Moreover, it was also found that a model with both the therapist and the patient features performed better than a model with only the patient features. This is an interesting finding that is supported by earlier research by Doyran et al. (2019) (40) in a study on child play therapy. This study also found that including information about the therapist's face and speech improved the prediction accuracy of emotion in children compared to only using information about the children's face and speech.

An improvement in accuracy upon including both therapist and patient has not always been found in research on the working alliance. As we mentioned in Section 2.3.2, Vail et al. (2022) found that the (head)gestures of the therapist are not predictive of the perception of the working alliance of the patient (136). They found that the behavior that someone displayed was only indicative of the perception of the working alliance of him or herself and not of the other person.

An explanation for this contradictory finding could be the difference in the indicators used in the study. Vail et al. (2022) used (head)gestures and speech indicators (136) whereas Doyran et al. (2019) used facial and word choice indicators (40) and Zhou et al. (2022) used information from questionnaires filled in by therapist and patient as indicators (150). Since these indicators are all different modalities it could explain the difference in findings as perhaps some therapist-produced indicators are predictive of the working alliance perception of the patient and some are not. Using multimodality in this thesis provides an opportunity to study which of these indicators are good predictors of the working alliance perception of the patient, which could give some more insight into these contradictory findings.

2.4.5 LSTM, RNN, transformer model based on language

Where the previous study tested the accuracy of a deep learning model against a less complex Random Forest algorithm, Lin et al. (2022) compare multiple deep learning models for the prediction of working alliance (85). The deep learning algorithms tested in this study predicted working alliance based on dialogue classification were a Recurrent Neural Network (RNN), a Long-term Short-Term Memory neural network (LSTM), and a transformer model. The input used in this study to base the prediction on is threefold: the first feature is the working alliance embedding, which is the concatenation of the sentence embedding vector and the psychological state vector. The second feature is the working alliance score, which uses the state vector. The third feature is the embedding which is the baseline that uses the sentence embedding vector. The results show that the LSTM model with the patient's dialogue features as input yields the best results in predicting a working alliance. Therefore, using an LSTM will be worth exploring in the experimental phase of this thesis.

3. Methodology

3.1 Outline of the methodology

The aim of this thesis was to predict the working alliance using a machine-learning algorithm. The source of the input was a large dataset of recorded psychotherapy sessions consisting of mono-sound Dutch audio, and visuals of either the therapist or patient and in some cases both. From this dataset, the audio, textual, and visual features that were discussed in Chapter 2 were extracted. The visual features were extracted with the public toolboxes OpenFace (9) and focused on facial emotion recognition and head gestures of the patient during speaking and listening. The audio features were affect features which were detected with a pre-trained Wav2Vec2 model that was fine-tuned for arousal, valence, and dominance detection. Third, the audio was transcribed into text using WhisperX (8) and diarized by PyAnnote(117) and conversational features were extracted. Also, sentiment analysis was applied to the text, and the features of positive/negative sentiment, specific emotions, arousal, and valence were extracted using multiple Bidirectional Encoder Representations from Transformers-based models (BERT) (36). Finally, multiple regression models and a classification model were trained on the multimodal features for predicting the working alliance WAI scores.

In this Chapter, we will go into each component outlined above and describe the methods used to arrive at the resulting WAI prediction models and feature analyses. There are multiple models used in this thesis, most of which are off-the-shelf models and pipelines that we applied to our data.

3.2 Dataset

The data used in this research stemmed from a large dataset consisting of recorded CBT and IPT psychotherapy sessions for patients with a diagnosis of major depressive disorder according to the DSM V criteria (n = 200). This data was recorded in a study by Bruijniks et al. (2020) who researched the effect of once- versus twice-a-week therapy on the outcome in depressed patients (23). The mean age of the participants was 37.85 years (+/- sd 12.26) and 61.5% of the participants were female. There were 76 therapists with an age range of 25–61 years, of which 81.6% was female. There are 12-20 sessions taped per patient. Many sessions include WAI-S and WAI-SRT scores as rated by the patient and the therapist respectively, and also include observational codings on the quality and psychological processes (WAI-S scores), as rated by experts. The recording of this dataset was agreed upon by the Medical Ethical Committee of VU Medical Centre Amsterdam (registration number 2014.337) and with full knowledge and consent from the participating patients and therapists.

The content of the recordings is very diverse as they differ in the visibility of the patient and therapist, angle, lighting conditions, and audio quality. This made the feature extraction with the deep learning models more difficult but had the advantage of being more representative of various therapy setting conditions and can thus be more easily generalized to different clinical settings. The language spoken in this dataset is Dutch.

Since there are multiple sessions recorded per patient, some of the data was dependent. Wherever possible, this was taken into account in the analysis by providing a model with the information that sessions with the same ID number belonged to one group.

3.2.1 Descriptive analysis

The dataset used in this research was a subset of the data from (23). There were a total of 438 sessions with 89 patients and there were on average 5-10 sessions per patient available. There were WAI scores filled in by the patients, by the therapists, and by observers. There was however, not for every session a WAI score meaning that the total dataset was smaller than 438 sessions (see Section 3.3.5 for more information on the final dataset).

The WAI scores are generated by a 7-point Likert scale. The WAI questionnaire consisted of 12 questions for the patient and observer versions and 10 questions for the therapist version. The questions of the WAI questionnaires can be divided into three categories: bond, goal, and task. We conducted multiple analyses and used the individual WAI score per question (Spearman correlations). We also tested the features against the task, bond, and goal component scores, which were calculated by summing up the WAI values from the questions corresponding to that component, and the total WAI score which was calculated by summing up the value from all questions in the WAI.

The division of which question corresponds to which component is visualized in Table 2. The WAI-S questionnaires filled in by the patient and the observers had two questions, number 4 and 10, which were formulated negatively. To exemplify, question 4 was formulated:

"The client and therapist have different ideas about what the client's real problems are."

A high score for this question would mean a low working alliance, whereas question 3 was formulated positively:

"There is mutual liking between the client and therapist.".

A high score for this question would mean a high working alliance. Therefore, we pre-processed the WAI-S scores by calculating the maximum score of a question (7) minus the actual score to convert them into a positively formulated question.

3.2.2 Transformer models

Since many of the models used in this research make use of Transformers, we wanted to dedicate a section to explaining the specifics of the Transformer architecture. Transformer models were designed for Natural Language Processing tasks but can be applied to any task involving sequential text, image, or video data. They emerged as a replacement for other neural networks like Recurrent Neural Networks (RNN) to solve the problem of sequence transduction (a task that entails converting an input sequence to an output sequence). To perform sequence transduction, a model must have a way of storing information, a memory. While an RNN model is able to remember some information from short sentences, it is unable to remember the information from longer input. In an RNN architecture, new information slowly replaces the old information as the amount of input becomes larger. Therefore, old information is not remembered well. This is a problem when processing large amounts of text as it can't use information from a few sentences back to understand the context of a new sentence. A Long-Short Term Memory model (LSTM) was designed to solve this problem, as one of the main properties is the ability to store and 'remember' large amounts of information. While it does this well, it is still unable to cope with long input sentences. This is because an LSTM calculates the probability of a word that is far removed from the one currently being processed as being related, to be diminishing exponentially with distance. This means that when information needs to be used from sentences earlier in the text to understand a new sentence, an LSTM will not be able to do this well. The solution to the problems posed by the RNN and LSTM is an attention mechanism as is found in Transformers.

The transformer architecture primarily consists of two blocks: an encoder block (with six identical encoder layers) and a decoder block (with six identical decoder layers). The encoder processes the input sequence and creates a representation of it and the decoder uses the encoder's representation to generate an output sequence. Each encoder layer consists of two sub-layers: a multi-head self-attention sublayer and a fully-connected feed-forward network sublayer. The multi-head attention sublayer processes the input by applying an attention mechanism multiple times in parallel, each time paying attention to a different part of the input sequence to capture all diverse relationships of the input. The second sublayer, the feed-forward neural network, has as its purpose to introduce non-linearity and enable the model to capture more complex relationships between words. It is composed of two linear functions with a ReLu activation function in between.

After each encoder block, layer normalization is applied to normalize the input batch of each layer, and residual connections are applied. Residual connections help combat the vanishing gradient problem by allowing some information from previous layers to reach the next layer by bypassing the current layer. A second advantage of residual connections is that they ensure that the representation of the input accurately represents the meaning of the input. As input passes through the layers, their representation can change a lot, possibly leaving out some important characteristics. By allowing some information to continue to the next layer without being processed, it saves the information of the input's true meaning. It can thus help to retain important information from earlier layers.

The decoder block is formed in much the same way as the encoder block, except that the decoder layers have two multi-head self-attention sublayers. The additional sublayer performs attention over the output of the encoder and its purpose is to help the decoder to focus on the most important parts of the input when generating the output. Additionally, the self-attention mechanism in the decoder layer is slightly different as masking is applied. The process of masked self-attention prevents the decoder model from looking at the words that come after the current word by applying a mask to block the attention to future positions. This prevents the decoder from knowing any information about the next positions via the attention mechanism. It thus ensures that the decoder only considers previously generated output.

The decoder output is passed through a linear transformation and a softmax activation function to produce the final output probabilities.

Multi-head attention is the application of the self-attention process in a transformer

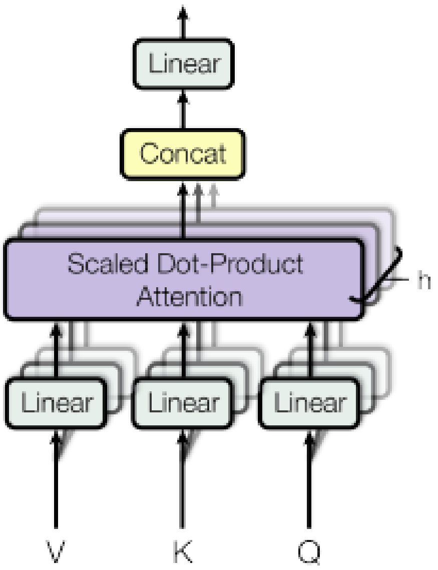


Figure 3. Multi-head attention process as used in the Transformer architecture. This figure is from (141)

but performs multiple sets of attention computations in parallel. Each set is referred to as a "head", explaining the name multi-head attention. The transformer model can jointly attend to data from various representation subspaces at various positions thanks to multi-head attention (see Figure 3).

The attention computations in the multi-head attention process is a method called Scaled Dot-Product Attention. The input of the Scaled Dot-Product Attention consists of queries (Q), keys of dimension d_k (K), and values of dimension d_v (V). Q, K, and V are three matrices that were trained during the training process. The first step of the attention calculation is to take the dot product between Q and K, which is then divided by the square root of d_k to prevent the values from becoming too large. Then, the softmax function is applied to this outcome to obtain attention weights on the values. Finally, the outcome of the softmax function is multiplied by the values matrix V. The formula of the Scaled Dot-Product Attention calculation is given in Equation 3.1 and the process is displayed in Figure 4.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.1}$$

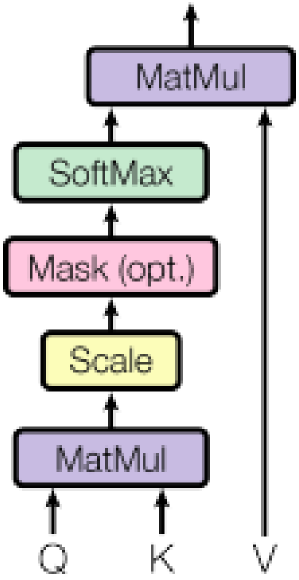


Figure 4. Scaled Dot-Product Attention as used in the Transformer architecture. This figure is from (141)

3.3 Indicators

An advantage of the dataset was that it included both sound and video, allowing the extraction of auditive, textual, and visual indicators for predicting the working alliance. The auditive and textual indicators that were used were mainly based on a paper by Vail et al. (2022) who showed the different types of indicators that corresponded with the three elements of the working alliance according to their definition (136). The visual indicators that were extracted depended on the content of the video and were therefore different for each session. If possible, gaze, (head) gestures, facial expressions, and the physical distance between therapist and patient were extracted to create a multimodal set of indicators for the working alliance prediction. In this Section, we will describe all the indicators that were extracted from the sessions. An overview of the indicators per modality is given in Table 5.

Table 5. Overview of the feature categories from different modalities used to predict the working alliance.

Features	
Audio affect	*Arousal* *Valence*
Text affect	*Text emotions* *Nr positive utterances text* *Nr negative utterances text* *Arousal* *Valence*
Conversation dynamics	
Facial emotion recognition	
Head movements	

3.3.1 Textual feature extraction

Transcription

In order to be able to analyze the linguistics of the session, the speech had to be transcribed into text so methods such as emotion extraction could be applied. The speech-to-text transcription was done using the WhisperX library which is a language transcription tool with multilingual options, among which is a Dutch language transcription option to transcribe speech to text (8). The WhisperX library is a Python wrapper that uses the Whisper library for transcription. WhisperX uses batched interference which allows for faster transcription of audio data. In this thesis, a batch size of 5 was chosen as it would enable a moderately fast transcription without taking up too much GPU RAM. There are different pre-trained models available, all of which are trained on a different size dataset. The large-v2 model, which was the most accurate and largest model available when doing this analysis, consists of 1550 million parameters. Due to the limited GPU RAM availability, this largest model could unfortunately not be used for analysis, but the large-v1 model was used which was slightly smaller but still performed well (for WER, see results section). WhisperX also applied Voice Activity Detection (VAD) before processing the audio data to isolate the sections of audio containing human speech and leave out sections containing background noises. This process reduces hallucinations (incorrectly transcribing non-speech sounds) without degrading the word error rate (WER) of the transcriptions.

Description of the Whisper Pipeline

The WhisperX pipeline (see Figure 5) consists of multiple steps of pre-processing data before it uses the Whisper model to perform the transcription. In this section, we will explain the steps of the WhisperX pipeline and go into more detail on the Whisper and PyAnnote diarization model that it uses to get a complete description of the speech-to-text process that we used in this research.

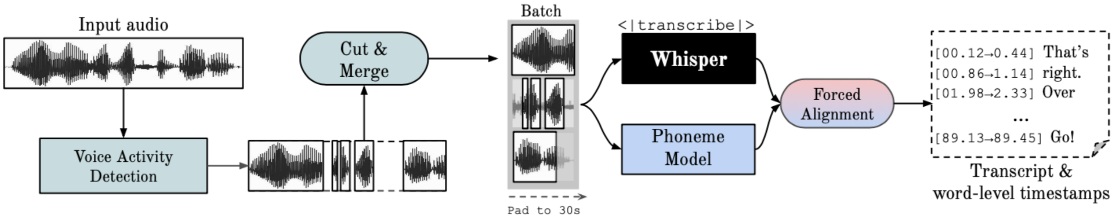


Figure 5. WhisperX transcription Pipeline. This figure is from (8)

Voice Activity Detection

The first step in the WhisperX model is applying Voice Activity Detection (VAD) to the raw audio. The VAD used by WhisperX was the PyAnnote VAD model (21; 20). By applying VAD, the audio data will be split into segments where there is only speech present, so the segments containing only noise or non-speech sounds will be left out. This increases accuracy as it reduces the number of errors made in clustering the segments. A second advantage is that it will reduce timestamp inaccuracies as each speech segment automatically contains a start and end timestamp boundary. This helps reduce the problem of Whisper's inaccuracy with produced timestamps.

Segment creation

The second step in the WhisperX pipeline is the cut & merge. In this process, the segments produced by the VAD are re-evaluated based on length. Very long segments will be split into smaller segments, with a maximum length of thirty seconds, at the times when the voice activity is lowest (this is most likely to be the end of a sentence). This step is important to reduce high memory consumption when feeding the segment to the Whisper model. Additionally, very short segments contain too little information for Whisper to reliably make a transcription as it requires some context to do so. Specifically, if there are two possibilities for a transcription of a word in a sentence that is equally likely based on the audio, the context of other sentences will likely clarify which word the transcription should

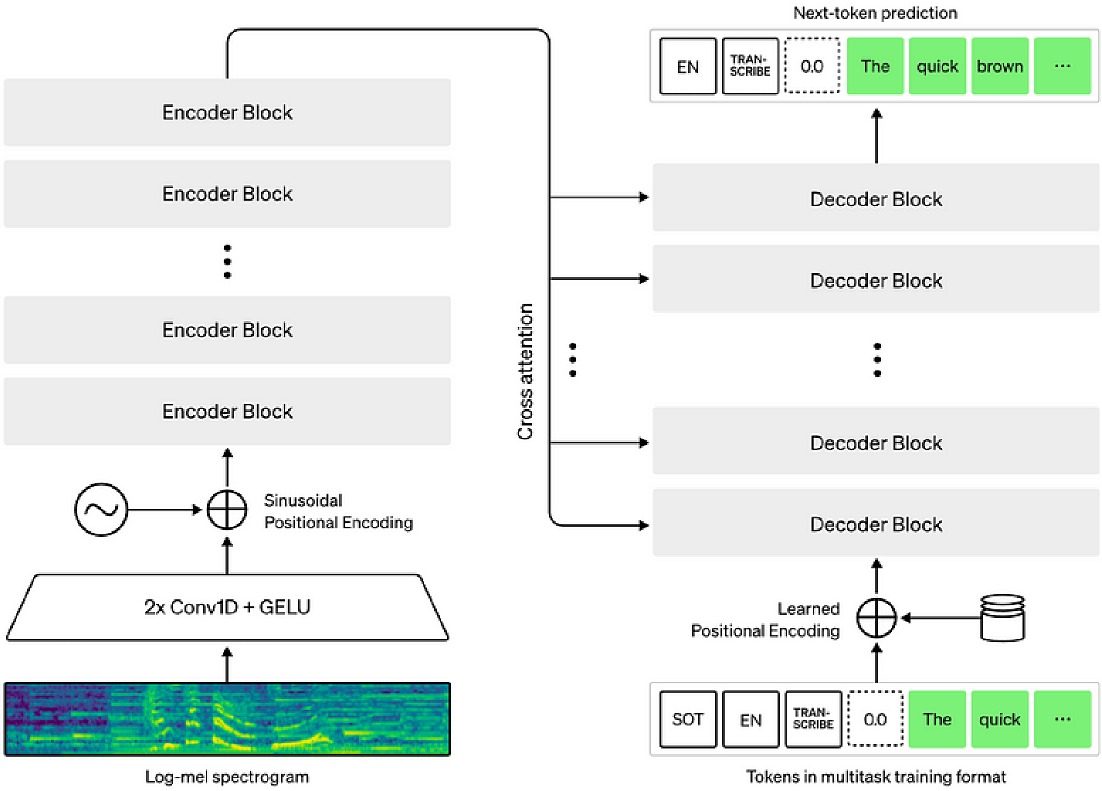


Figure 6. Whisper Architecture. This figure is from (114)

be. Therefore, very short adjacent sentences will be merged together until they have a length closest to the thirty seconds that Whisper is trained on to provide as much contextual information as possible.

Transcribing the segments

The third step is the Whisper transcription. Each segment is fed to the Whisper model independently without a learning effect on the previous sentences as that will reduce the risk of hallucinations that have been shown to occur in Whisper when using longer input.

Whisper Architecture

Whisper is a transformer-based model that works based on an encoder-decoder architecture with an attention mechanism (114). In our case, Whisper is fed 30 seconds of raw audio by WhisperX. The input audio segment is first re-sampled to 16 kHz after which it is converted into a log-mel spectrogram format. Then, it is passed through two convolutional layers with a width of three and two and a GELU activation function to extract the most relevant features from the data. As can be seen in Figure 6, sinusoidal position embeddings are applied to the output of the convolutional layers before passing to the encoder blocks. The sinusoidal position embeddings create a representation of the position of each word spoken in the audio so the relative order of the words in the sentence is remembered. The next step in the Whisper architecture is a standard transformer encoder-decoder structure. The encoder blocks are multiple LSTM layers that encode the segment after which the decoder aims to predict the correct words spoken. After the encoding blocks, a final layer normalization is applied to the encoder output. The positional encoding that was captured by applying the sinusoidal position embeddings to the data, is applied during the decoder process to give information about the correct word order in the sentence.

While Whisper is similar to other language models like Wav2Vec2 in its function (representing language in an encoded way), it also differs in a number of ways. First, Whisper was trained on a very large amount of data. Namely, 125,000 hours of English translation data and 680,000 hours of noisy speech training data in 96 different languages. Second, Whisper was trained in a supervised way, while a model such as Wav2Vec2 was trained in a self-supervising way. Third, Whisper is a very generalizable model that can be applied to many different functions, for instance, language detection, speech-to-text, or VAD. This makes Whisper very robust in different language settings and accurate to use. In fact, Whisper has been shown to be as accurate as humans in transcribing speech to text, as can be seen in Figure 7 where the average Word Error Rate of Whisper is 8.81 compared to a Word Error Rate of human transcription of 7.61-10.5.

Forced alignment for accurate timestamps

To continue in the pipeline of WhisperX, after transcribing the segments with Whisper, forced alignment is applied to the segments that were generated as output by Whisper. The process of forced alignment enables the production of word-level timestamps. Forced alignment is a process in which the transcriptions are matched with the audio segments. A phoneme model represents the phonemes (smallest unit of speech) and can match the specific words in the transcription with a specific part of the audio segment, thus allowing word-level timestamps to be created. The phoneme model used by WhisperX in this research was a Wav2Vec2 model (for a detailed explanation of this model, see Section 3.3.2), that was trained on 960 hours of data and used a greedy decoding strategy (select the word with the highest probability).

The output of the WhisperX pipeline is a list with the transcriptions and word-level timestamps. This can be combined with a diarization of the speakers to label each utterance as corresponding to either the therapist or the patient. The diarization

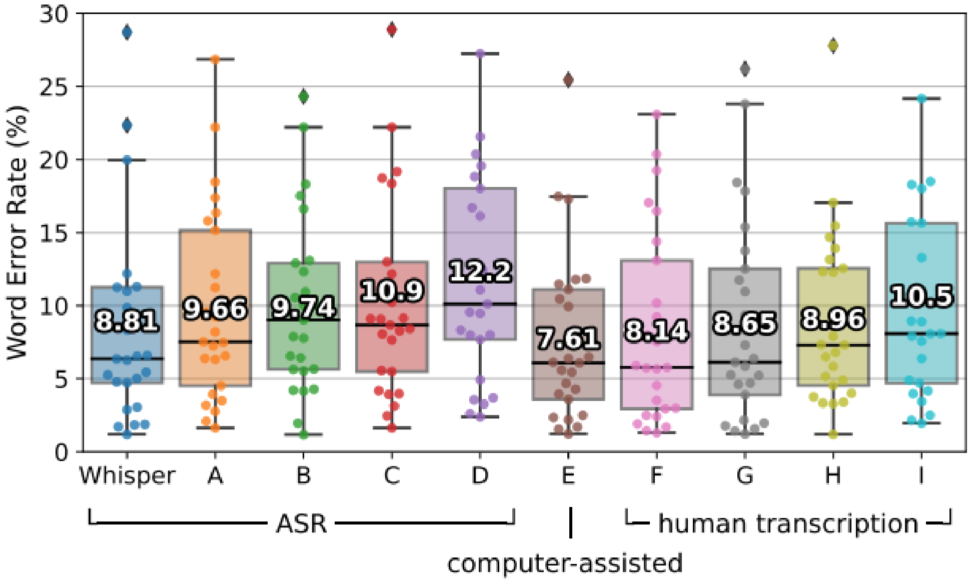


Figure 7. Performance of Whisper against other Automatic Speech Recognition (ASR) models and human transcription. This figure is from (114). The average of Word Error Rate (WER) for Whisper (dark blue) is similar to the WER of the human-made transcriptions showing that Whisper is an accurate transcription model.

was done using PyAnnote which will be explained in the next section.

Diarization

After the transcription, the audio was diarized using the speaker_diarization@2.0 function from PyAnnote library (20; 21).

PyAnnote pipeline

The pipeline for the PyAnnote diarization model is visualized in Figure 8. The architecture that PyAnnote uses is a convolutional neural network as designed by (117) and is named SincNet. The diarization process consists of multiple steps of which the first is Voice Activity Detection. This is the same VAD system as was used in the WhisperX pipeline, which was described in Section 3.3.1.

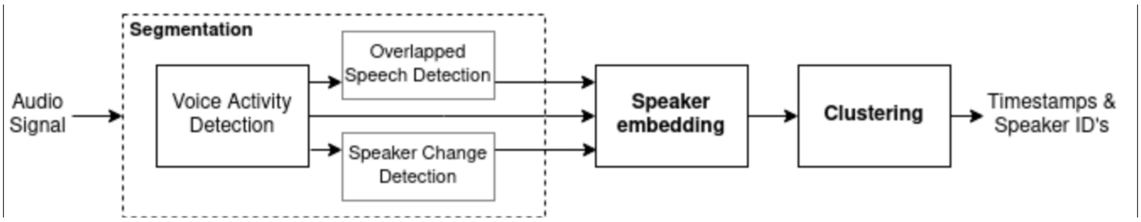


Figure 8. PyAnnote Diarization Pipeline. This figure is from (75)

After the VAD, PyAnnote simultaneously performs Speaker Change Detection

(SCD), in which the moments in the audio are detected where there is a change between speakers, and Overlapped Speech Detection (OSD), in which audio segments are detected in which two or more speakers speak simultaneously.

The next step is speaker embedding where the voice characteristics of each speaker are captured in a vector. After the embedding, the audio segments are clustered by classifying each audio segment as the speaker that closest matches the embedding. The clustering system uses a combination of cosine distance metrics, centroid linkage, and hierarchical agglomerative clustering. Every element of the PyAnnote model (VAD, SCD, OSD) was trained independently and combined into a large pipeline.

The output of the diarization pipeline is a list of the timestamps and the corresponding speaker IDs, so each moment in the audio file that contains speech is labeled as a specific speaker, in our research either the patient or the therapist.

The PyAnnote pipeline was trained and fine-tuned on a variety of datasets, ranging from television shows, news broadcasts, and podcasts. The diverse training data results in a generalizable and robust audio diarization model.

Specific changes made for this research

The hyperparameters given to the PyAnnote diarization model in our research were a max_speaker number of two (as we knew there would always be two speakers in the audio). Further, it used an automatic segmentation and clustering threshold to determine the difference between noise, speech, and different speakers. We chose an automatic threshold as the audio quality and presence of noise were very different between patients and between sessions depending on the placement of the camera (which contained a microphone) relative to the speakers.

Combining transcription and diarization

The first attempt at transcription and diarization used the libraries PyAnnote and Whisper separately which resulted in a mismatch of the timestamps between the speakers and the utterances. It wasn't until halfway through the analysis that the WhisperX library became available which improved the timestamp results due to forced alignment (8). The improvement of WhisperX over the Whisper library is that it is capable of performing re-alignment of the produced timestamps, to combine the transcription and diarization into one output of a list of utterances, the corresponding speaker of each utterance and the timestamps corresponding to the start and end of each utterance within the audio file. The issue with combining a transcription and diarization tool is often that the timestamps that mark the beginning and end of an utterance are not identical, due to small differences in trimming of silence before or after an utterance, which makes it difficult to align the transcription and diarization. WhisperX therefore improves this issue by performing re-alignment of the timestamps. Having accurate timestamps is an important contribution to this type of research, not only to have a correct matching of the diarization and transcriptions, but it also allows for an accurate linking of the visual behaviors with the vocal utterances. Thus, this allows for the estimation of synchrony between facial and bodily gestures and the speech of a person.

Conversational features

From each session, we extracted conversational (turn-level and turn-taking) features such as described by (12). In this paper, these features showed high correlations with the WAI-S ratings, and could specifically be linked to predict one of the three subsets of WAI-S questions corresponding to the bond, task, and goal components. Therefore, we extracted these features according to the methods described in (12) to see if we could replicate the results. As shown in Table 5, the turn-level features provide information about the individual turns and speakers in the session, including the number of turns, duration, word count, and speech rate. The conversational features, such as participation equality and turn-level freedom, represent the dynamics of the interaction between the therapist and the patient. All the extracted conversational features and their calculation are given in Table 6 for the turn-taking features and in Table 7 for the turn-level features.

The participation equality and turn-level-freedom are a bit more intricate than the number of turns feature. Thus, for clarity, we will give the equations and descriptions here as well.

The calculation of the participation equality was done according to Equation 3.2. The PEQ stands for participation equality, the SSD stands for the Sum of Squares of Differences. This is the sum of the squares of the difference between each speaker's speech duration and the average speech duration. It represents the deviation of each speaker from the average turn duration. The Maximum Possible Duration is the sum of all the duration of each speaking turn and represents the total possible speech duration.

$$\mathrm{PEQ} = 1 - \frac{\mathrm{SSD}}{\text{Maximum Possible Duration}} \tag{3.2}$$

The turn-level freedom is calculated according to Equation 3.3. The FCOND stands for turn-level-freedom which is a measure of how freely speakers alternate during the conversation. Cond_Ent_Sum is the sum of the conditional entropy values which measures the uncertainty of the prediction of the next speaker in the text given the current speaker. It represents the predictability of the turn-taking pattern. The Max_Cond_Ent_Sum is the sum of the maximum possible conditional entropy values, which is the maximal level of uncertainty of the prediction of the next speaker given the current speaker.

$$\mathrm{FCOND} = 1 - \frac{\mathrm{Cond\_Ent\_Sum}}{\mathrm{Max\_Cond\_Ent\_Sum}} \tag{3.3}$$

Table 6. Overview of all turn-taking conversational features and their description.

Feature	**Description**
Participation equality	A measure of how evenly speech time is distributed among speakers.
Overlapping turns	The percentage of overlapping turns in the conversation.
Turn-level freedom	A measure of how freely speakers alternate during the conversation. It considers the order of speaker turns and calculates a score that indicates the level of turn-taking flexibility.

Table 7. Overview of all turn-level conversational features and their description.

Feature	Description
Number of turns	Total number of turns. Calculated by WhisperX.
Turns therapist/patient	The number of turns for each speaker. Calculated by WhisperX.
Average turn length therapist/patient	The average speaking time per turn in seconds. Calculated by dividing the total speech duration by the number of turns.
Wordcount therapist/patient	The total number of words spoken by each speaker. Calculated by splitting the text into words and counting them for each segment.
Speech rate therapist/patient	The rate at which words are spoken by each speaker, usually measured in words per second (wps). Calculated as the ratio of the word count to the total speech duration.
Speech duration therapist/patient	The total time spent speaking by each speaker. Calculated by summing the durations of all speaking segments for each speaker.
Duration percentage therapist/patient	The percentage of time spent speaking by each speaker in relation to the total speaking time. Calculated by dividing the individual speaker's speech duration by the sum of both speakers' speech durations.

Affect analysis from text

We performed an affect analysis of the text that was transcribed by WhisperX. This affect analysis consisted of multiple features. Most features were extracted with a model that was a fine-tuned version of the Bidirectional Encoder Representations from Transformers (BERT) language model, which is why we will go into detail on the design of the BERT language model and the specific enhancements that were made for each fine-tuned affect analysis model (36).

Arousal and valence from text

For the first textual affect feature, we used robbert-v2-dutch-sentiment, a fine-tuned model of the Dutch language model RobBERT with which we extracted the sentiment (positive and negative) (34). The model was fine-tuned on the dbrd dataset consisting of 110k book reviews which are annotated with sentiment polarity labels (139).

Second, we used the VADER sentiment analyzer to extract the polarity (positive/negative) and the intensity of the emotion present in the text (a summation of the intensity of each word in the text) (67). While the polarity is not very useful for our research (it was designed for analyzing book review sentiment), the strength of the emotion present could potentially be very useful. VADER was originally designed for classifying reviews as positive or negative, and to our knowledge, it has not been tested as a classifier of sentiment in therapeutic sessions. However, it has been shown to have the highest accuracy (77 per cent accuracy) in predicting sentiment compared to other language models such as Text Blob (74 per cent accuracy), thus we wanted to test its performance on a new context (16).

VADER is a very basic rule-based model based on a Lexicon containing words and their respective polarity and intensity scores. These scores are obtained through human annotation. The specific formula for the inner workings of the VADER analyzer is not publicly available, yet we do know that it consists of a couple of components. First, there is a sentiment intensity calculation on a sentence which increases and decreases the relative importance of each word in the sentence based on a set of grammatical rules. With this calculation, not every word gets the same weight when calculating the total sentiment score to better capture the content of a sentence. An example of a grammatical rule is the occurrence of the word 'very' in a sentence. The presence of this word indicates that the word following it should have more weight than without the presence of the word 'very'. Second, VADER performs valence shifting and amplification, which is a process to capture the nuance in a text that is caused by its relation to previous sentences. For instance, the occurrence of the word 'nevertheless' in a sentence indicates a relation with previous sentences which can influence the strength of the sentiment in a sentence. Lastly, VADER takes the overall sentiment in a text and the frequency of the most common words in the text into consideration to adjust the sentiment scores per sentence if necessary.

Since the intensity score of the emotion that VADER outputs is similar to the valence which we also extracted from the audio using a fine-tuned version of the Robust Wav2Vec2 model (65), and the XLM-RoBERTa-large model (27) to extract valence from the text, it would be interesting to compare the correlation between valence and the strength of the emotion extracted with different models. An important note in this sentiment analysis is that the VADER sentiment analyzer was trained on an English lexicon and can thus only be applied to English text. Therefore, we translated the Dutch text into English.

Specific emotions from text

The third feature that was extracted from the text was the presence of a specific emotion. This was extracted using the EmoRoBERTa model (71), which is able to classify 28 different types of emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise and neutral (77). The emotion model was a fine-tuned version of the Robustly Optimized BERT Pretraining Approach (RoBERTa) which is a variant of the BERT model and was trained on the GoEmotions dataset which contains 58000 labeled Reddit comments with the 28 emotions mentioned. We chose an emotion classifier that was trained on various emotions, rather than the six basic emotions as labeled by Eckman (42), to better represent the complexity of the human psychological process. Since emotional display in a therapeutic setting is common, it won't be sufficient to only extract negative or positive sentiments as they might be reactions to a personal psychological dilemma or thought rather than a response to the therapist's or patient's behavior. Thus, we decided to extract as many details about the displayed emotions as possible.

Language choices

The sentiment and emotions were extracted per utterance but combined to form session-level features. This resulted in a count for each feature: how many utterances displayed positive and negative sentiment and how many times every possible emotion occurred in a session. The EmoRoBERTa model and the NLTK's Vader sentiment model are trained in English rather than the Dutch language. Therefore, we translated the Dutch transcriptions into English using the Argos Translate package (Finlay and Argos Translate). This specific package was chosen as it included the Dutch language and was able to translate offline which was a necessity due to the sensitivity of the dataset.

The choice to translate the text into English instead of analyzing the Dutch text was motivated by the scarcity of models pre-trained on Dutch datasets and the scarcity of Dutch annotated emotion datasets. Of the latter, there was only one to our knowledge, the EmotioNL dataset which consists of a thousand utterances from TV shows and is annotated with arousal, valence, dominance and specific emotion (32). We trained the RobBERT model (34) on this EmotioNL dataset but found that the performance was very low (12% testing accuracy). Also, the language used in the EmotioNL dataset does not contain proper spelling and grammatical usage which would make it likely very difficult to apply it to the transcriptions of this research. Unfortunately, there is no large Dutch dataset with emotion annotations yet and thus no pre-trained model is able to process the Dutch language to perform affect analysis.

We also extracted a fourth set of affect features from the text: arousal and valence. As previously mentioned, the model that was used in this was the XLM-RoBERTa-large (27). This pre-trained model is fine-tuned on 34 publicly available datasets containing arousal and valence ratings for a large corpus of words. Among the datasets that the model was trained on, was also a dataset in the Dutch language by Moors et al. (2012) consisting of 4300 Dutch words labeled for valence, arousal, and dominance (96).

To clarify which affect features were extracted from the text, we created an overview of the features and their corresponding models in Table 8.

Table 8. Sentiment Analysis Models used for extracting affect features from text.

Features	**Model**
Sentiment	robbert-v2-dutch-sentiment (34)
Arousal & Valence	XLM-RoBERTa-large (27)
Polarity & Emotion Strength	NLTK's VADER sentiment analyzer (67)
Specific Emotion	EmoRoBERTa (71)

Architecture of the BERT model

BERT is a language model that has proven quite revolutionary in the field of natural language processing(36). It uses a Transformer model with bidirectional encoding in contrast to the more common directional model, which means that it can read an entire sentence of words at once instead of reading it sequentially from left to right or right to left. The bi-directionality allows BERT to learn the context of a word based on the other words in the sentence, in other words, it learns the relationship between the words in a sentence. Directional models like Wav2Vec or GloVe use word embeddings to represent language. Every word is represented by a vector, and that vector will remain the same regardless of the context of the sentence. These models have a large disadvantage because they are unable to distinguish between meanings in the case of ambiguous words. As an example, the word bridge can have two meanings, either a game or a construction to cross some otherwise unpassable grounds. BERT, however, can do this very well as it takes the context of the sentence into account when representing an ambiguous word, thereby knowing the true meaning of the word. In the sentence 'He likes playing bridge with a good friend', the directional models will not be able to distinguish the meaning of bridge from its other meaning of a structure, whereas BERT will take the other words in the sentence into account and see that the verb 'playing' heightens the possibility of bridge indicating a game in this sentence.

Another advantage of BERT compared to directional models is that it can be trained without datasets of labeled data, but rather can train on a corpus of unsupervised data. BERT was trained on an enormous dataset consisting of 3.3 billion words (2.5 billion words from Wikipedia and 800 million words from Google's BooksCorpus). This results in a large language model of 350 million parameters for the large BERT model and 110 million parameters for the base BERT model.

To achieve bidirectionality, BERT is trained on two tasks, Masked Language Modelling (MLM), and Next Sentence Prediction (NSP). MLM is a method where about 15 per cent of the words (in the BERT training) are randomly selected to be masked, meaning that they are replaced with a 'blank' spot. The model has to learn to predict which words were replaced by the blank sport which it can do based on the context (the other words in the sentence). This learns BERT to not predict the next word in a sentence but rather to evaluate the words left and right of the missing word to grasp the context and meaning of a sentence. This makes BERT less of a language prediction model and more of a language representation model. In NSP the model receives two sentences and must state whether the sentences follow each other or are unrelated. The purpose of NSP is for BERT to learn to understand the coherence in a text, not just within, but also between sentences. This will help BERT to grasp the broader context of a document and enable it to perform well on tasks such as question answering and sentiment analysis.

As mentioned, BERT is a multi-layer bidirectional Transformer encoder with a self-attention mechanism. We will not further explain the architecture here as we refer the reader to Section 3.2.2 where the transformer model is explained in detail. Specific to BERT, however, is the design of 24 hidden layers of size 1024 and 16 self-attention heads (12 hidden layers with size 768 as 12 self-attention heads for the BASE model).

Architecture of the RoBERTa model

RoBERTa is an improved version of the BERT model (86). The authors of the paper wanted to replicate the BERT model but found that it had more potential if there were slight adjustments made. Thus, while the architecture of RoBERTa is similar to BERT, its training is different causing it to be a more robust and improved version of BERT. Therefore, the name R(obustly) o(ptimized) BERT a(pproach). The changes that were made in RoBERTa compared to BERT are the use of dynamic masking, full sentences without NSP loss, large mini-batches, a larger byte-level BPE, and longer training periods on more data. We begin with the use of dynamic masking. This is similar to the masking during the MLM that BERT is trained on, but instead of masking one word in a sentence, multiple consecutive words are masked. This stimulates the model to rely more on contextual information in a larger range. Second, it was found that training the model in the NSP task on full sentences that are sampled contiguously from multiple documents decreases the ability to learn long-range dependencies. Therefore, RoBERTa is not pre-trained on an NSP task. Third, using a large batch size of 8 times the size of BERT's increases the accuracy of RoBERTa on the MLM task and specific tasks it is applied to (examples of end-tasks include sentiment analysis and text classification). Also, RoBERTa uses a larger vocabulary size of Byte-level encoding compared to BERT (50K compared to 30K). This allows for a better understanding of the text by reducing ambiguity as a larger vocabulary means a smaller chance of multiple words sharing the same prefix or suffix which causes ambiguity, and expands the number of rare words seen. Lastly, RoBERTa is trained on three additional datasets, CC-NEWS (54), OPENWEBTEXT (52), and STORIES (135), which means it is trained on about 10 times more data. Moreover, it was trained for a much longer time than BERT with 500K pre-training steps compared to the 100K pre-training steps for BERT. The authors note that even after this increase in training length, the model does not appear to overfit and would, therefore, perform even better with additional training.

The technical specifics of the RoBERTa training were an Adam optimizer with a polynomial decay of the learning rate lr = 10e-6. It used a ramp-up period of 1000 iterations and during the training, the learning rate gradually increased. To prevent overfitting, a weight decay of 0.1 is used which acts as a penalty term to the loss function stimulating the weights to stay small, and a dropout of 0.1 is used to randomly set 10 per cent of the model's units to zero to prevent heavy reliance on few features.

Architecture of the RobBERT model

The RobBERT model is a Dutch version of the RoBERTa model, that uses the same architecture and training specifics, but with a Dutch language corpus (34). There are two different versions of RobBERT, the v1 and v2 but the one used in this research is the v2. In the v1 version only the pre-training corpus was a Dutch corpus, whereas in the v2 version, both the corpus and the tokenizer were Dutch. The corpus that RobBERT was trained on was the multilingual Open Super-large Crawled Aggregated coRpus (OSCAR) (131). The Dutch subset of this corpus has a size of 39GB with 6.6 billion words. The tokenizer in v2 was the same as for RoBERTa but with a Dutch vocabulary originating from OSCAR.

In the affect analysis we did in this research, we used two fine-tuned models of RobBERT and RobBERTa, so we tested sentiment analysis in Dutch and English. The first Dutch-based sentiment analysis was a version of RobBERT that was fine-tuned on the Dutch Book Reviews dataset (DBRD) which contains 118,516 book reviews from hebban.nl. Of these reviews, about 22,000 were labeled as positive or negative as these were used to train RobBERT with a 90 per cent training set and 10 per cent test set. RobBERT was trained for 2000 iterations on this portion of the dataset with a batch size of 128 and a learning rate was 5e-5. Interestingly, RobBERT outperforms other BERT models that were trained for sentiment analysis (34).

Architecture of the EmoRoberta model

The second sentiment analysis model based on BERT was a fine-tuned model of RobBERTa called EmoRoberta (72). It was trained on the GoEmotions dataset consisting of 58000 Reddit comments which are labeled for 28 emotions (35): admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise and neutral. EmoRoberta was trained on this dataset for 10 iterations with a batch size of 16 and a learning rate of 5e-5.

3.3.2 Affect analysis from Audio

We extracted arousal, valence, and dominance scores from the audio of the sessions. The motivation behind this was threefold. First, it has been shown that arousal is much more apparent in audio than in text, therefore it will likely be a useful addition to extract this from audio. The Wav2Vec2 model for emotion is trained to extract the features of arousal, valence, and dominance, therefore we extracted all three. Second, it was interesting to compare the arousal and valence scores that were extracted from text and from audio. We would expect a high correlation between these as they are the same features on the same dataset which will testify to the validity of the arousal and valence scores (not for the dominance scores as these were not extracted from the text). On the other hand, a low correlation would be interesting as it could indicate that text and audio are so inherently different from each other that the resulting features show this as well. The third reason for extracting these scores was the availability of manually annotated arousal and valence scores for this dataset. We wanted to perform an automatic extraction of these features to compare their performance with the manually annotated scores. This was mostly due to the contribution it would have to other research on this dataset, but it would also be interesting to evaluate the correlations between manual and automatic annotation for this research as well.

The manual annotations were one rating for every five-minute segment of each video. Therefore, we also split the video sessions into five-minute segments and applied the valence, arousal, and dominance extraction for each segment. After the manual check, we averaged the segment ratings into one score per video so it could be added to the working alliance prediction model as an input feature.

The extraction of the arousal, valence, and dominance features was done using the audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim model from HuggingFace that was applied to the audio from each video. This model was created by fine-tuning Wav2Vec2-Large-Robust on the MSP-Podcast corpus(v1.7), containing 100 hours of annotated speech data originating from podcast recordings.

Architecture of the Robust Wav2Vec2 emotion model

The Wav2Vec2-Large-Robust model is a transformer-based model that is designed to represent raw audio data in a self-supervised learning approach. It is pre-trained on large datasets to learn how to represent raw audio as a vector space encoding. Then, the model is fine-tuned on labeled Speech Emotion Recognition datasets to be able to use for emotion analysis. It is similar to Bert's masked language model that we applied for affect extraction from text, but it is specifically re-trained and applied for speech.

The Wav2Vec2 model consists of three main components:

- Feature Encoder
- Transformer
- Quantization Module

Feature encoder

The Feature encoder has a task to reduce the dimensionality of the data. It is indicated in Figure 9 as *Z*. It takes raw audio (*X* in Figure 9) with a sample rate of 16 kHz as input and outputs feature vectors representing the audio. The process between the input and output is as follows: First, the audio is normalized to a zero mean and unit variance. Then, there are 7 convolutional layers with 512 channels per layer, with decreasing kernel width and stride as we progress in the network, that the audio is passed to. This is followed by a Gaussian Error Linear Unit (GELU) activation function. The resulting output is a series of latent representations that represent the essential characteristics of the audio.

Transformer blocks

The output of the feature encoder layer is fed to the 24 transformer blocks (12 blocks for the base model, but 24 for the large model), where it first goes through a feature projection layer to increase the dimension to 1024 to match the size of the convolutional layers (512 dimensions for the base model). Each transformer block performs two main operations: self-attention and feed-forward neural networks. The self-attention operation allows the model to weigh the importance of every latent representation relative to the others by calculating an attention score to represent the relevance of each token compared with the other in the input sequence. The feed-forward neural networks capture the local relationships for each latent representation and the attention mechanism captures the global relationships between the latent representations. Where the feature encoder encodes the raw audio into latent representations, the transformer component of the Wav2Vec2 model builds a conceptualized representation, which is indicated in Figure 9 by *C*.

In contrast to the original transformer model, the Wav2Vec2 model has an alteration. Traditional transformers use fixed positional embeddings to encode the absolute position of each element in a sequence. However, in this case, instead of using fixed positional embeddings, it uses a convolutional layer to create relative positional embedding.

Quantization module

To be able to perform self-supervised training the data needs to be represented as discrete units. Therefore, the third component of the Wav2Vec2 model, the quantization module, takes the continuous data of the latent representations and converts it into quantized discrete speech units. The units consist of codewords which are retrieved from codebooks. Codebooks consist of a predefined set of speech sounds that can be combined to create a speech unit. Wav2vec uses 2 codebooks with 320 possible words in each group which can be combined to form 320x320=102400 possible speech units. The quantized latent representations are indicated in Figure 9 as Q.

Training datasets

The Wav2Vec2 model was pre-trained with several large speech audio datasets. Two of the datasets, the Libri-Light and the CommonVoice datasets, consisted of clean read-out audio and text data. Two datasets, Switchboard and Fisher, consisted of noisy telephone data. During pre-training, the goal was to minimize the total loss (indicated in Figure 9 as L) which consists of the contrastive loss and diversity loss. The contrastive loss is given in Equation 3.5 and measures how well the model performs in the self-supervised task of distinguishing correct quantized representations. The diversity loss is given in Equation 3.4 and acts as a normalization to make sure that the model does not favor a few of the codewords in the codebooks, but rather uses all of them. By minimizing the diversity loss, the model will be more comprehensive and better able to represent speech in multiple different contexts.

$$\mathcal{L}_{\text{diversity}} = -\frac{1}{GV}\sum_{g,v}\log(p_{g,v}) \tag{3.4}$$

$$\mathcal{L}_{\text{contrastive}} = -\log\left(\frac{\exp(\text{sim}(C_t, Q_t)/\kappa)}{\sum_{\tilde{Q}}\exp(\text{sim}(C_t, \tilde{Q})/\kappa)}\right) \tag{3.5}$$

The Wav2Vec2 model was fine-tuned on the MSP-Podcast corpus consisting of English podcast recordings, where the speech segments are manually annotated with emotion labels using attribute-based descriptors (activation, dominance, and valence) and categorical labels (anger, happiness, sadness, disgust, surprised, fear, contempt, neutral and other). The annotations are created using crowdsourcing (87). The fine-tuned model creates a measure between 0 and 1 for arousal, valence, and dominance. Further details on the finetuned Wav2Vec2 emotion model are present in a paper by Wagner et al. (2023) (142) and for the pre-trained model robust Wav2Vec2, the reader is referred to (65).

While the robust Wav2Vec2 model is trained on 53 languages, including Dutch, the fine-tuning on an emotion dataset was done using an English dataset, due to the lack of a Dutch emotion dataset, and thus not specifically designed for Dutch valence, arousal, and dominance detection.

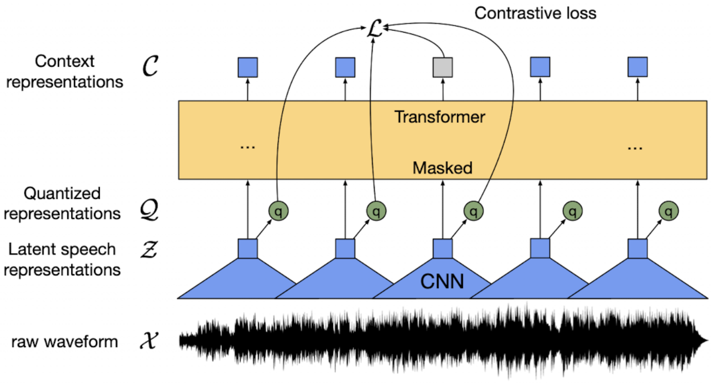


Figure 9. Architecture of the Wav2Vec2 model. This figure is from (7).

3.3.3 Facial and Gesture feature extraction

Facial Emotion Recognition

In Section 2.3.1, we have seen that an automatic neural network or SVM outperforms a rule-based emotion classification based on artificially extracted AUs. However, the dataset that we used is very limited in its size which means it is not enough to train an SVM or ANN. This leaves us with two options, to either classify the AUs in a rule-based manner or to use the AUs as input in the working alliance prediction model.

One of the most prominent indicators is the facial expressions of the patient and the therapist. From these, synchrony between both can be calculated which is known to be a good indicator of trust between people (118; 93). As mentioned previously, the facial analysis will be done using OpenFace (9) which provides the AUs of the face in a temporal sequence of video, and a confidence measure for each prediction. While this method has been widely used for facial analysis (95; 100; 55), it does come with some challenges that require pre-processing of the data. For instance, when participants move a hand for their face or move their head, it can cause inaccuracies in the AU prediction. Prevention of such inaccuracies has been determined in previous research using OpenFace to predict facial expression (93) and will be used in this study for similar causes.

Limitation of the dataset

Due to the variety of the visual information in the dataset, there will not be a frontal view of the face available in every session which could impair emotion recognition performance. However, research has shown that the eyes and mouth are most important for recognizing emotion (2; 76) and that even with only a few visible AUs, emotion can be predicted relatively well (145). We will try to build on the findings of (145) by improving the computational vision of emotion recognition by focusing specifically on the mouth and eye region and giving more importance to the corresponding AUs.

To summarize, the facial features we will extract are facial expressions (based on AUs), gaze, synchrony in facial expression, and open-mouth detection (for speech detection).

Gestures

Another possible important indicator to predict a working alliance is gestures. As described in the related research, gestures during the listening times, also named non-vocal backchanneling, are especially important as they reflect the engagement of the listener in the conversation.

We used OpenFace to extract the shakes and nods of the head which indicate disagreement and agreement respectively. The head movement was calculated by tracking the distance that specific AUs moved over time.

Thresholding OpenFace detections

First, we applied a threshold on the OpenFace data that allowed only faces with more than 75 per cent to be kept as we could be fairly certain that faces with less than 75 per cent certainty were not faces. The 75 per cent threshold was picked based on inspection of the data, as a threshold lower than 75 per cent resulted in more than two faces that were recognized (which is not possible in our dataset), and a threshold higher than 75 per cent missed some faces. In an ideal dataset, the required threshold would be higher but due to the visual imperfections of the dataset (especially profiles instead of frontal faces) the certainty of a detected face was lower causing us to have to lower our threshold.

After filtering out the incorrectly detected faces, we calculated the distance traveled along the pitch (vertical movement for detecting head nods) and yaw dimensions (horizontal movement for detecting head shakes) that were extracted per video frame by OpenFace over a rolling window of one second. According to the method described by (137), a threshold over the pitch and yaw movement was applied to select only the top quartile of distance traveled within the one-second window. By selecting only the largest distance traveled, the most prominent movement will be extracted. Then, the head gestures are segmented meaning that frames that show consecutive movement are identified and scanned for gaps between the movement so as to split two separate movements if present.

Finally, the timestamps of the head gestures of each face are compared to the transcription file to split the head gestures that occurred during the speaking time of speaker 1 or during the speaking time of speaker 2 (or in some cases not while someone was speaking). With this comparison was determined how many head gestures were displayed during listening or speaking time for each face ID.

3.3.4 Evaluation method for individual features

The presence of a correlation between the individual features and the working alliance scores was tested between a single feature and the working alliance score. We tested against every WAI question (10 or 12 items) to be to compare with the (12) paper.

The WAI scores were normally distributed and were tested with a Shapiro-Wilks test (p-value of 0.9) meaning that a Pearson correlation test could be used. However, since the WAI scores are ordinal data (scores from the Likert 7-point scale) a Spearman test is more suited for this data. A Bonferroni correction was applied as we did multiple statistical tests on the same dataset. The features that showed a significant correlation from the Spearman test were evaluated further by plotting the feature values against the component of the WAI score it correlated with to see if there was really a correlation. In some cases, the Spearman test showed a significant correlation but when the feature was plotted against the WAI score, there was no clear correlation visible. The data was too spread out so a correlation did not show in the plot. An example of such a plot is given in Figure 10. In these cases, the dataset is likely too small to say with certainty that there is a correlation, but we also can't state with certainty that there is no correlation based on the plot. Therefore, all correlations that were significant according to the Spearman test were taken into account in this thesis. If a correlation was present in the Spearman test as well as in the plot, this feature was labeled as a significant predictor. These features are described in Section 4.1.

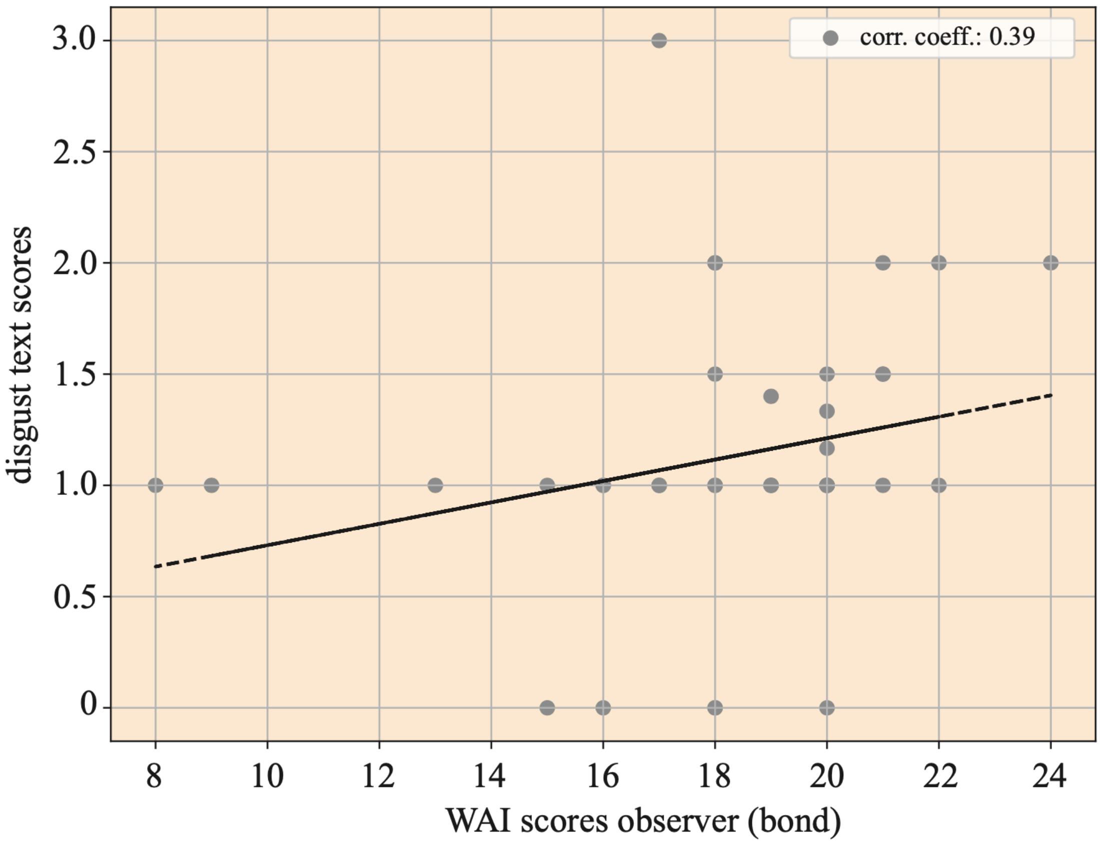


Figure 10. Example of a significant correlation but an unclear correlation when plotted.

3.3.5 Construction of the final feature dataset

After the initial feature extraction, the data from all modalities were combined and linked to the correct WAI score to be used as input for the model. This process was relatively straightforward as we merged the feature from each modality per session together, but we had to make a few important decisions regarding the WAI score. There are on average 2 or 3 sessions for which the WAI questionnaire was filled out, therefore there were on average 2-4 sessions in between two ratings. We chose to not add the sessions together but rather to only use the session that directly preceded the WAI rating as that would likely be most meaningful. Due to the long time periods in between sessions, the working alliance of previous sessions will fluctuate much and is therefore likely to produce noise rather than meaningful information. Alternative choices would have been to average all sessions preceding a rating moment or give different weights to the importance of each session depending on how close to the rating it occurred.

Availability of the features

The features could not be extracted for each session. Most prominently, the facial features could only be extracted for a subset of sessions because there were many videos where the faces were unrecognizable. Also, in some sessions, the diarization performed very poorly (the cause for this was manually checked and was due to poor audio quality or very similar voices), which caused the conversational features of that session to be unusable. Therefore, we decided to construct the final dataset where each session needed to have a WAI score or it was discarded. For most features, we kept the rows where all features were present thereby discarding some sessions for which not all features were extracted well. This choice was made because keeping these incorrect features would likely have a large impact on a dataset this small. The facial features did not have to be present in every session of the dataset, because these were so few that keeping only the rows with all features including facial features would result in a dataset with ca 10 sessions. Nevertheless, the facial features were added and we filled the empty spaces with an average value of the feature. This makes it more difficult to detect any significant influence of the feature but allows the presence of every feature in the dataset to train the final model. For the individual feature analysis, there were many more sessions with ratings present than in the final combined dataset, as well as for the facial analysis. We did not remove any outliers as doing so on a dataset this small would not be justified.

The selection of the final features resulted in three datasets, one for the therapist, one patient, and the observer, of which the sizes were: 37, 43, and 34 respectively. The reason for the small datasets is the limited number of WAI scores present. Without the WAI scores, the sizes would be 166. The therapist dataset consisted of 21.7k turns, the patient dataset consisted of 25k turns and the observer dataset consisted of 20k turns. All datasets had an average turn length of ca 9.6 tokens.

The individual feature analysis was done on multiple datasets with a larger size which were merged for the combined feature analysis using the machine learning models. Table 9 shows the number of data points (sessions) of each feature's dataset. This table shows that the conversational level features had the largest dataset (around 95 sessions), and as explained, the visual datasets were the smallest (around 30 sessions).

Table 9. Number of sessions in the individual datasets for each type of feature.

Feature	Patient scores	Therapist scores	Observer scores
Audio affect	43	40	35
Text emotions	52	44	39
Text affect	75	79	71
Conversational level	91	100	88
Facial emotions	28	38	34
Head Movements	24	33	29

3.4 Pre-analysis

As part of a pre-analysis, we tested the quality of transcriptions and diarization produced by multiple systems before deciding which transcription and diarization system to use. Our first requirement for picking out a system was that it would protect the privacy of the data, either by being available offline or through a secure system.

The choice of which systems to test in the pre-analysis was based on related research and the qualification for the privacy requirement.

For the transcriptions, this resulted in the choice of the systems Whisper (halfway through the final analysis WhisperX became available which uses Whisper but with small improvements), Google speech-to-text, and Word Online. The methods that we compared are Google speech-to-text, Word Online, and Whisper. The latter two would have had the possibility of using through a secure connection (Word online over the university server or buying a private workspace from Google), while Whisper was available to use as an offline system.

The diarization systems that were compared were PyAnnote (21; 20), Word Online, Agglomerative clustering (GMM) and Google's speech-to-text. At the time of this thesis, Google did not have a diarization method for Dutch audio. However, since diarization mostly depends on sub-linguistic features such as voice pitch and MFCC features, we still wanted to evaluate the performance of Google's diarization and therefore chose to test the Diarization systems of English and

German audio. English was chosen because it is the most common language in audio training sets and is therefore likely most established and fine-tuned in its diarization performance. German was chosen due to its closely relatedness to the Dutch language.

We performed the pre-analysis comparison tests with a test video that closely represented the real dataset but was publicly available and could therefore be used without having to purchase a private workspace from Google. The test video showed a psychologist and patient having a therapeutic session in Dutch.

3.4.1 Comparison of transcription systems

We created a transcription using each system and manually annotated a ground truth transcription to compare the performance. The systems were evaluated on the Word Error Rate (WER), Match Error Rate (MER), and Word Insertion Likelihood (WIL), as these have been shown to give a good representation of transcription accuracies (97). As explained by (97), the WER, MER, and WIL can give different outcomes as they each measure the transcription performance in a slightly different way. In our usage case, it is most important to have as little incorrect information in the transcription as possible. This is more important than having information missing as one incorrect word in the transcription can influence the further language analysis, for instance, an incorrectly transcribed word or an additional inserted word such as 'crying' will influence emotion recognition from the text to a negative emotion such as sadness while this may not be accurate. On the other hand, if a spoken word is left out, this might cause an emotion recognition to be missed or classified as less strongly present which is less harmful than an incorrect insertion or transcription could potentially be. From the explanations of the concepts of WER, MER and WIL below it will become clear that the MER therefore the performance measure that is most applicable to our study as it represents the probability of a given match being incorrect. However, we still include WER and WIL to get a more complete view of the transcription accuracy.

Transcription evaluation methods

The word error rate is a measure that quantifies the accuracy of the transcription based on three elements. It takes substitutions into account (incorrectly transcribed words), insertions (additional words in the transcription compared to the ground truth) and deletions (missing words in the transcription compared to the ground truth). It is calculated by the Equation 3.6. Where S represents the number of substitutions, I represents the number of insertions, D represents the number of deletions and N represents the total number of words (N = Substitutions + Deletions + Correct Words). So, the WER is determined by dividing the total number of errors by the total number of words in the ground truth text.

As mentioned, the MER is the most useful measure for this study. It represents the probability of a match between the transcribed and the ground truth texts being incorrect. It is calculated by dividing the total number of errors (S+D+I) by the total number of words (N) and the insertions (I), as displayed in Equation 3.7.

Lastly, we have the WIL as a performance measure that represents the probability of insertions in the transcription. It is calculated by dividing the number of insertions (I) by the total number of words in the reference (N), see Equation 3.8.

$$WER = \frac{S + D + I}{N} \tag{3.6}$$

$$MER = \frac{S + D + I}{N + I} \tag{3.7}$$

$$WIL = \frac{I}{N} \tag{3.8}$$

Important to note is that the transcription evaluations only look at the transcribed text and don't take timestamps or diarization into account. While having accurate timestamps is important to be able to link the transcriptions and variations, it is not possible to compare the timestamps of the transcriptions unless they are word-level timestamps. This is because each transcription model will split the utterances in its own way as there is often no clear-cut beginning and ending of a sentence. Therefore, when comparing the timestamps of the two models, the timestamps will likely differ without being necessarily incorrect. However, since the timestamps are important in this research, we did try to evaluate them manually, by reading the transcription alongside the video to see if they match. In all cases, the transcription timestamps matched the video and the whisper models even provided accurate word-level timestamps.

Comparison results

As we can see in Figure 11, the whisper large-v2 model has the lowest scores for WER, MER, and WIL, followed by the smaller whisper models large-v1 and medium. The Google speech-to-text has the highest error rates (WER = 0.39) and thus the lowest scoring transcription accuracy. Based on this analysis, the decision was made to use the Whisper transcription tool. Ideally, the large-v2 model would be used but since the resources of computational power are limited, we had to use the smaller, but still very good-performing whisper large-v1 model.

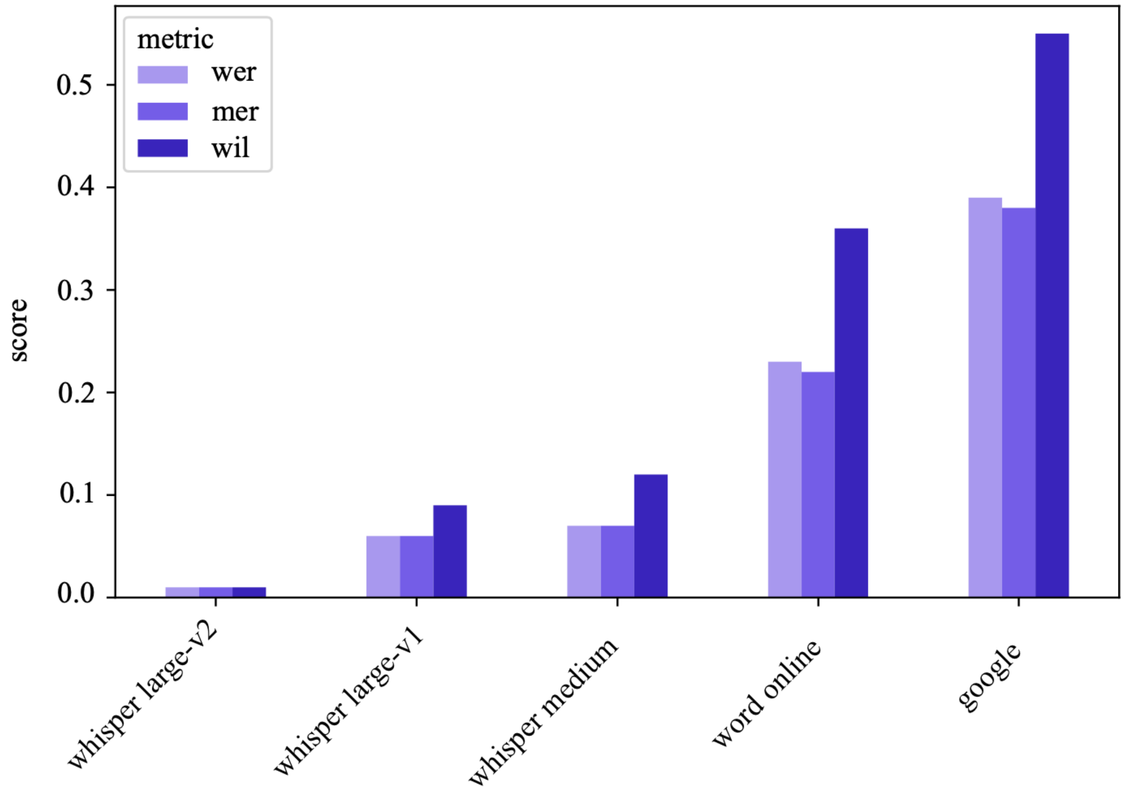


Figure 11. Performance Metrics for Different Transcription Models on a Test Video.

Performance on real dataset

The Whisper transcription performed really well on this test set with good audio quality. But to get an idea of how it performs on our dataset with a worse audio quality we manually transcribed two sessions and compared the results with Whisper's transcription. As is visible in Table 10, the transcription accuracy on the real dataset is much less accurate than on the test video. Interesting is that the performance on video 6013 session 2 is good, but the performance on video 1025 session 6 is much less accurate. On manual inspection is this due to the audio quality and due to an unbalanced volume between the two speakers. The therapist was sitting very close to the microphone than the patient causing the imbalance in volume which could disturb the transcription performance. The Whisper system automatically balances the volume in a pre-processing step. Unfortunately, the audio of the patient was very hard to understand in this specific session as the microphone did not pick up the speech of the patient well, even if the volume was increased.

Table 10. Evaluation using the Word Error Rate (WER), Match Error Rate (MER), and Word Insertion Likelihood (WIL) of the Whisper-produced transcriptions on a test video and the therapeutic dataset.

Dataset	**WER**	**MER**	**WIL**
Test Video	0.0085	0.0085	0.0141
1025 session 6	0.1791	0.1782	0.2339
6013 session 2	0.0993	0.0986	0.1479

3.4.2 Comparison of diarization systems

The most widely used and reliable evaluation metric for diarization is the Diarization Error Rate (DER) (107). The DER is calculated by dividing the total percentage of errors in the diarization including False alarm (FA), missed detection of speech (Missed), and an incorrect speaker label (Incorrect), by the total amount of time (T), see Equation 3.9.

A second more recent measure is also an established way of evaluating the diarization, the Jaccard Error Rate (JER). This Method has been introduced in the Third DIHARD Challenge and separates itself from the DER by evaluating each speaker with equal weight (123). DER measures the error rate for the entire text while with JER, the error rates are first computed per speaker and then averaged. The formula for the JER is given in Equation 3.10, where N represents the total number of speakers, FA_i represents the total time that the wrong speaker was assigned as a speaker i, $Missed_i$ represents the total amount of time that should have been attributed to speaker i, but was instead mislabelled and $TOTAL_i$ represents the duration of all speaker segments together.

The JER score has the advantage of being more representative of the true error rate if the division of speech between the speakers is very unequal. While we don't expect this to occur in this dataset, we still include the JER to give a complete picture of the diarization performance and to make it easier to compare the findings in this thesis with other research that uses the JER rating as an evaluation measure.

$$DER = \frac{FA + Missed + Incorrect}{T} \tag{3.9}$$

$$JER = \frac{1}{N}\sum_{i=1}^{N}\frac{FA_i + Missed_i}{TOTAL_i} \tag{3.10}$$

Comparison Results

From the pre-analysis on diarization appears that the PyAnnote system has the best performance on diarization as it has a DER of 1.7% compared to the closest of 14.1% from the GMM model, and it has a JER of 0.14 compared to the closest value of 0.23 of the GMM.

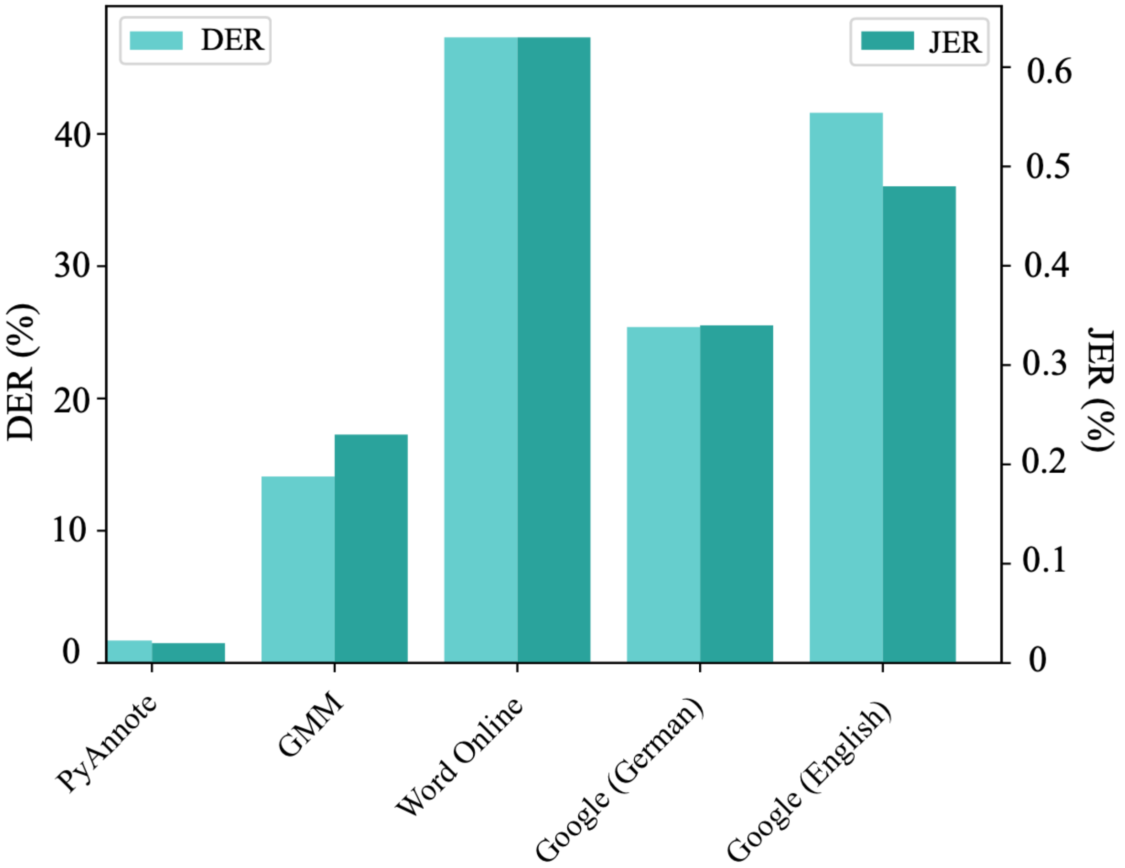


Figure 12. Performance Metrics for Different Diarization Models on a Test Video.

Performance on real dataset

Because the PyAnnote diarization performed very well on this test video, we wanted to see how it would perform on the dataset where the audio is less clear and often contains some background noise. As is visible in Figure 12, the diarization of PyAnnote in the real dataset is less accurate than in the test video, but still a lot better than the other systems we tested. Therefore, the decision to choose PyAnnote as a diarization system was clear. The videos for this analysis were randomly selected from the dataset. The first 10 minutes of each video were analyzed.

Table 11. Evaluation using the DER and JER of the PyAnnote-produced diarizations on a test video and the therapeutic dataset.

Dataset	Diarization Error Rate (%)	Jaccard Error Rate (%)
1025 session 6	9.1	0.13
6013 session 2	6.3	0.12
Test Video	1.7	0.02

The diarization performance on the dataset shows that the DER and JER are higher for the videos from the dataset compared to the test video, as expected, see Table 11. However, the DER and JER are still much lower than the scores for test videos in the other systems. Based on this pre-analysis outcome the decision was made to use PyAnnote as a diarization method for this thesis.

Moreover, using PyAnnote had another advantage as well as there was already an option within the WhisperX library to automatically perform diarization with PyAnnote. This, however, was not yet available until halfway through the analysis, so when this tool became available the diarization and transcriptions were extracted again using WhisperX.

3.5 Model choices and architectures

This section will describe the models used to predict working alliance based on the extracted features. For the initial analysis, we used four relatively simple models, the (categorical classifier) support vector machine (SVM), the k-Nearest Neighbours (kNN), the Support Vector Regressor (SVR), and the Elastic Net. These models were chosen as they handle a small dataset well without much chance of overfitting, and they provide clear and explainable results without requiring a lot of optimization and training time. The SVR and Elastic Net were chosen based on their success in a similar research (137). However, the complexity of the working alliance perception is likely much better modeled by more complex architectures that can model the longitudinally of the dataset such as described in Section 1.4.

We also applied more complex models to test if they would be able to represent the data complexity well. For this, we used a Random Forest, Multilinear Regression model, and XGB Regressor. The Random Forest was chosen as it is able to capture complex non-linear relationships in a dataset, making it very suited to model something as complex as the working alliance. The multilinear Regression was chosen because it is a very interpretable model. It does not handle non-linear relationships but is able to perform well without a lot of hyperparameter tuning, which is beneficial to prevent overfitting. The XGB Regressor was chosen as it is a very effective model for smaller datasets; it is able to combine multiple weakly predicting features to enhance their predictive capabilities. Moreover, like the Random Forest model, regularization and specific overfitting parameters can be applied (maximum depth of the model), to prevent overfitting.

All models were optimized using a GRID search to find the best hyperparameter settings. The data was divided into a train and test set (80/20 split). The models were optimized on the training set and the performance was tested on the test set. The models were evaluated using the MSE, RMSE, and R-squared values, and cross-validation was applied with five folds. Additionally, for the SVM model, it was possible to extract a confusion matrix which allowed us to get a better understanding of the incorrect predictions. The features were normalized before being fed to the models to a mean of zero and a standard deviation of one. The train/test ratio for the models where applicable was 80/20.

3.5.1 Simple models

Support Vector Machine

The SVM is a categorical test that predicts the category that a sample belongs to. We used this method to also try a categorical approach which might be more robust in a small dataset like ours. The samples with corresponding WAI scores were divided into high and low based on the quantile approach. The data was divided into four quantiles based on the median the lowest two quantiles were marked as low WAI and the highest two quantiles were marked as high WAI. We chose this approach instead of an equal sample split based on the mean because it is possible that there is an unequal number of high WAI and low WAI scores (low WAI occurs less often than a high WAI) and we wanted to model this as accurately as possible. Therefore, we chose the quantile median approach which did not result in two equal sets of samples, but we believed modeled the WAI score better.

The hyperparameters were equal for the patient and observer models. They contained a linear kernel, a regularisation parameter (C) of 1, and a scale kernel coefficient (gamma). The hyperparameters for the therapist model were somewhat different and contained a radial basis function (RBF) kernel, a regularisation parameter (C) of 10, and a scale kernel coefficient (gamma).

k-Nearest Neighbours

We trained multiple kNN models for the three datasets. The optimized hyperparameters of the models for the patient and observer were again identical with an optimal number of neighbors (n=9) and uniformly distributed weights.

The therapist model showed an ideal number of seven neighbors, but also uniformly distributed weights. A Euclidean distance measure was used for all models.

SVR

The hyperparameters of the SVR were different between the patient, therapist, and observer. The patient model showed the best performance with a regularisation parameter (C) of 10, a degree of 2, an auto kernel coefficient (gamma), and an RBF kernel.

For the observer, the hyperparameters were a regularisation parameter (C) of 1, a degree of 2, a scale kernel coefficient (gamma), and a linear kernel. For the therapist, the hyperparameters were a regularisation parameter (C) of 0.1, a degree of 2, an auto kernel coefficient (gamma), and a polynomial kernel.

Elastic Net

The Elastic Net model for the patient had an alpha of 1 (a value used to weight the contribution of the L1 penalty for the loss function) and an l1_ratio (regularization parameter) of 0.9. The observer model's design had an alpha of 10 and an l1_ratio of 0.5. The therapist model's design had an alpha of 10 and an l1_ratio of 0.2.

3.5.2 Moderate complexity

Random Forest

The random forest has a risk of overfitting on small datasets which is why we implemented some prevention methods. The model had to have a minimum sample per leaf of 3 (1 or 2 resulted in a better R-squared but a stronger overfit of the model), a maximum of 100 estimators, and a maximum depth of 15.

The hyperparameters for the optimized patient model were a maximum depth of 15, a log2 method for selecting the features (max_features), a minimum sample per leaf of 3, and a minimum sample split of 10 and 100 estimators.

The observer model was designed with a maximum depth of 15, an sqrt method for selecting the features (max_features), a minimum sample per leaf of 4, and a minimum samples split of 2 and 50 estimators.

The therapist model had similar hyperparameters to the observer model with a maximum depth of 15, an sqrt method for selecting the features (max_features), a minimum samples per leaf of 3, and a minimum samples split of 5 and 50 estimators.

Multilinear Regression model

Unlike the other models, the Multilinear Regression model did not have hyperparameters to optimize. The implementation was, therefore, relatively straightforward and without GRID-search hyperoptimization tuning resulting in different models for the patient, therapist, and observer.

XGB Regression model

The XGBoost regression model for the patient contained the hyperparameters of a maximum depth of 3, a minimum child weight of 3, 100 estimators, and a learning rate of 0.2. The observer's model was optimized to show a maximum depth of 5, a minimum child weight of 1, 300 estimators, and a learning rate of 0.01. The therapist's model was used with a maximum depth of 4, a minimum child weight of 1, 300 estimators, and a learning rate of 0.1. These hyperparameters were again obtained by performing a GRID-search hyperoptimization tuning.

3.5.3 Packages used

All analyses were done in Python and the packages used to create the models were Sklearn (108), XGBoost (25) and Statsmodels (126). Matplotlib (66) and Seaborn (144) were used for plotting the data and Adobe Illustrator for finalizing the figures for use in this thesis.

Model input

As input for the models as part of the combined feature analysis, all features were initially used as they could potentially be a good predictor if combined with other features. The models were, however, also trained with the selection of features that were better predictors of the WAI score. This is a process called feature selection and is often performed on datasets with multiple features to remove noise and optimize the predictive value of the feature dataset. The feature selection was done using two methods: selecting the features that showed a significant predictive value from the Spearman test, and a method called Maximum Relevance Minimum Redundancy (MRMR).

Maximum Relevance Minimum Redundancy feature selection

This method calculates the relevance and redundancy value for each feature relative to the target variable according to the scoring formula described in Equation 3.11. The scoring formula is the key component of the feature selection technique as it calculates the relevance and the redundancy of each feature to determine the importance of predicting the outcome variable. The MRMR provides a value for the trade-off between redundancy and having valuable information to predict the outcome variable. The relevancy is determined by the $MI(X_i, Y)$ component which is a quantification for the mutual information between the feature X_i and the target variable Y. A higher mutual information value means that the feature has a higher relevance. The redundancy is determined by this component of the formula: $\frac{1}{k}\sum_{j=1}^{k} MI(X_i, X_{S_j})$. It accounts for the mutual information between feature X_i and all the other features (X_{S_j}), thereby quantifying how much of the relevancy of a feature is already explained by other features. The total MRMR score is calculated by subtracting the redundancy score from the relevance score $Score_i = MI(X_i, Y) - \frac{1}{k}\sum_{j=1}^{k} MI(X_i, X_{S_j})$. The component k is the total number of features, so this score is iteratively calculated for each feature.

$$Score_i = MI(X_i, Y) - \frac{1}{k}\sum_{j=1}^{k} MI(X_i, X_{S_j}) \tag{3.11}$$

where X_i is the feature being considered, Y is the target variable, X_{S_j} are the features already selected, and k is the number of selected features.

The MRMR was calculated for each modality individually to retain as many sessions as possible without having to lose sessions due to merging, except for the combined audio and text modality as these likely contained some overlapping features (both arousal and valence scores), so the MRMR scores for these modalities were calculated together. The resulting positive MRMR scores were used as a feature selection dataset as model input.

Model evaluation

We used different model evaluation methods. We chose the R-squared, Root Mean Squared Error (RMSE), and the Variance of Coefficient (CV). Most studies using regression models use either RMSE, Mean Squared Error (MSE), or Mean Absolute Error (MAE). We chose to use the RMSE as it is a common metric used in related studies, specifically Vail et al. (2021) so it allowed for the comparison of results (137). The RMSE measures the average size of the error between the predicted values and the true values of the outcome measure. The R-squared formula is given by:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}. \tag{3.12}$$

It represents the model's prediction accuracy. A lower RMSE value means better performance. In this research, we calculated the baseline average of the error (the average error if the mean value of the WAI scores is predicted every time), and compared the RMSE of the model with the baseline. If the RMSE of the model is lower than the baseline, its performance is better than average.

We also used the R-squared value as that has been found to be a robust evaluation method for smaller datasets (26). The R-squared represents the fit of the model to the data; how much variance of the outcome variable is explained by the model. It has a range of 0 to 1, with 0 explaining none of the variance and 1 perfectly explaining the variance of the outcome variable. While it seems unintuitive, a negative value is possible as well if the model has a very bad fit to the data. From the formula can be seen that this is possible if the SS_{res}, the sum of squared residuals is larger than the SS_{tot}, the total sum of squares that represents the total variance in the data. This occurs when the model's fit is worse than a horizontal line, meaning that the model fits the data very poorly.

We calculated the CV as well as a measurement that represents the relative variability of a dataset compared to its mean. It gives an indication of the variability of the dispersion of the data points in a model. The CV is calculated by dividing the MSE value by the baseline MSE value, see Equation 3.13. A value < 1 is generally considered as a low variability whereas a value > 1 is considered as a high variability. It indicates the size of a standard deviation relative to the dataset mean. It represents a similar aspect of the model evaluation as the RMSE but is normalized with the baseline value making it easy to compare between models using different datasets (such as the patient, therapist, and observer WAI scores).

$$CV = \frac{\text{RMSE}}{\text{Mean}} \tag{3.13}$$

4. Results

In this Chapter, we will describe the results from the individual feature analysis and the results from the combined feature analysis using prediction models. First, the correlation analysis between the individual features and the WAI will be described per modality. Thus, we will go through each modality and describe the important findings for the patient, therapist, and observer. For a few significant correlations, visualization will be presented in a scatterplot to provide a better understanding of the structure of the data and the correlations. Second, the combined feature analysis will be described. This section is divided into three subsections, one for the patient results, one for the therapist results, and one for the observer results. In the following section, the results of the MRMR feature analysis will be mentioned, but to avoid distractions from the main results, the visualization of the MRMR feature analysis is presented in the Appendix rather than in this Chapter.

4.1 Correlation analysis of individual features

This Section will describe per modality, the features that showed significant Spearman correlations with the WAI score of either patient, therapist, or observer. An overview of all significant features is displayed in the Appendix 8.3.1.

4.1.1 Conversational features

Correlation results with the patient scores (WAI-S)

Patients experience a positive correlation, calculated with the Spearman test, between the total duration of the patient's speech, as well as the average duration of the patient's speech (turn length), and the bond component of the working alliance (0.39). The participation equality shows a positive correlation with the goal and task components of the working alliance as well (0.30 and 0.31), showing that equality in the participation of both the therapist and the patient, is indicative of a stronger working alliance between them.

A negative correlation is seen between the average turn length of the therapist's speech and the bond component of the working alliance. (-0.31) The total duration of the therapist's speech, as well as the average turn length of the therapist, are negatively correlated with the task component of the working alliance (-0.31). Notable is the negative (-0.3) correlation that patients experience between the speech rate of the patient and the bond component of the working alliance, where this was positively correlated in the eyes of the therapist. There is also a negative correlation between the turn-level-freedom and the task component of the working alliance (-0.30).

Correlation results with the therapist scores (WAI-SRT)

Therapists experience a positive correlation, calculated with the Spearman test, between the total number of turns during the session, and the task component of the working alliance (0.34). A positive correlation with the goal component of the working alliance is found for both the number of overlapping segments (0.35), i.e. the times that both the therapist and the patient are speaking at the same time. The speech rate of the patient was found to be positively correlated with the goal and task components (0.32 and 0.29). A positive correlation was found between the turn-level-freedom and the task component of the working alliance (0.30).

Furthermore, the average turn length of the therapist and the patient has a negative correlation with the task component (-0.30 and -0.25). However, the duration percentage of the therapist has a positive correlation with the task component (0.24), see Figure 13 for a visualization of this correlation.

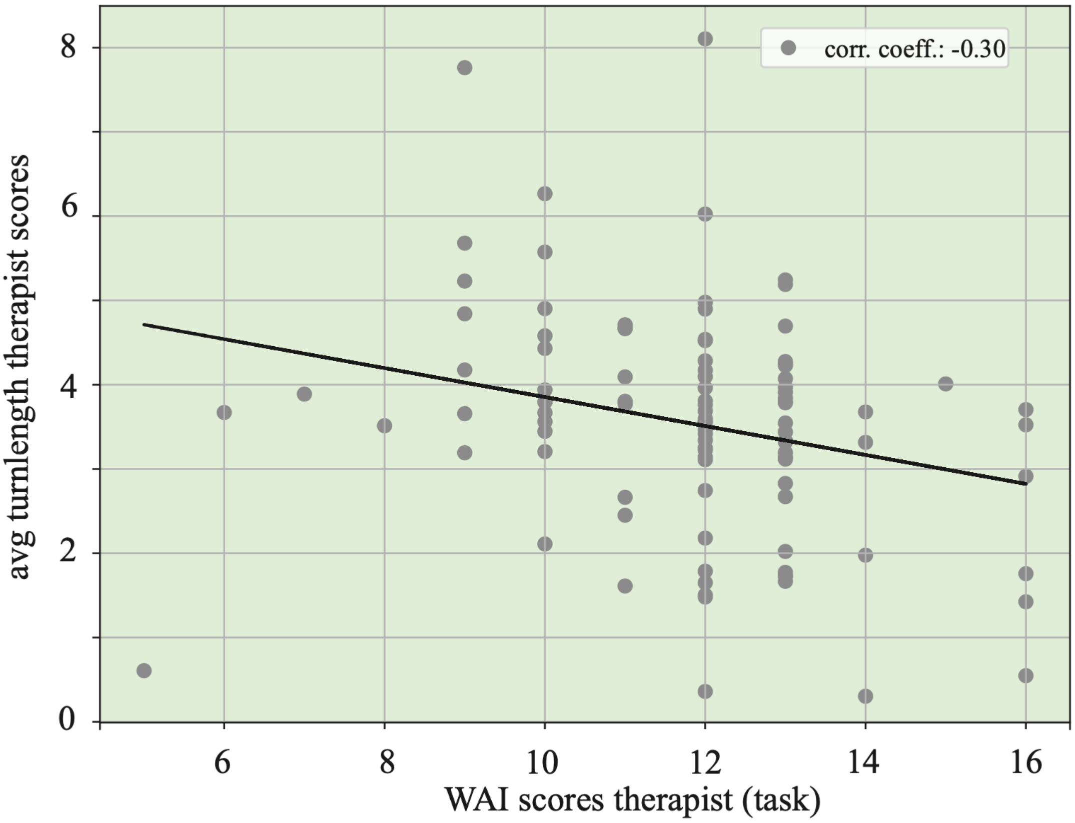


Figure 13. Example of a significant correlation between the average turn length of the therapist and the task component of the working alliance.

Correlation results with the observer scores (WAI-S)

Observation of the videos of the therapy sessions by a neutral observer shows a positive correlation between the total number of turns in the conversation and the task component of the working alliance (0.33). The turn-level-freedom is positively correlated with the bond component of the working alliance (0.28).

The average turn length of the patient's speech shows a negative correlation with questions that are related to the task component of the working alliance (-0.27). The average turn length of the therapist is negatively correlated with the goal component of the working alliance (-0.32). The number of overlapping segments, i.e. the times that both the therapist and the patient are speaking at the same time, is positively correlated with questions related to the task component of the working alliance (0.25).

4.1.2 Text Affect and emotion features

The text effect features are the minimum and maximum values of arousal and valence. We chose the minimum and maximum values as well as the mean since that better represented the data as it showed the range of the arousal and valence instead of only the average.

Correlation results with the patient scores(WAI-S)

The Spearman correlation test showed that there was a significant negative correlation between the minimum arousal that the patient shows and the bond and task components (-0.30 and -0.29).

Amusement showed a positive correlation with the bond component of the WAI (0.40). Also, fear had a positive correlation with the bond component (0.39).

Correlation results with the therapist scores (WAI-SRT)

The therapist scores show no significant correlation with both patient and therapist arousal features.

As for the emotion features, disgust was positively correlated with the bond component (0.38) and the number of sentences containing positive sentiment in the text was negatively correlated with the task component (-0.38).

Correlation results with the Observer scores (WAI-S)

There is a positive correlation between the approval in the text and the task component (0.44), see Figure 14 for a visualization of this correlation.

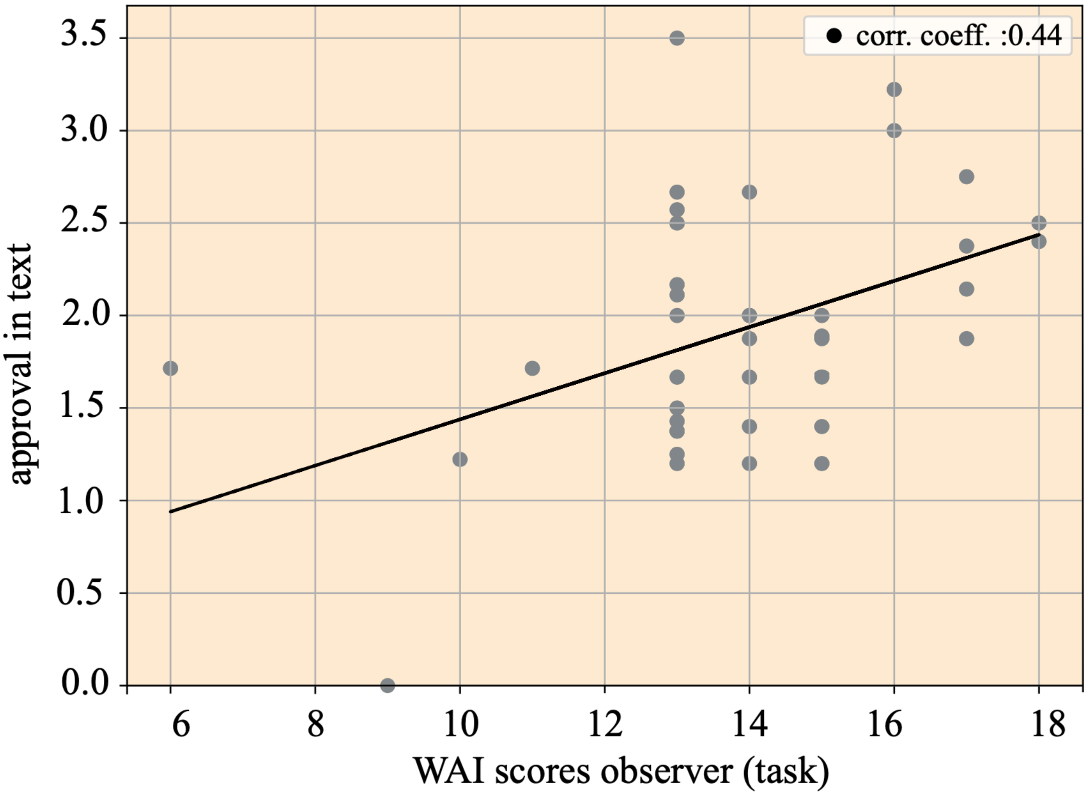


Figure 14. Example of a significant correlation between the presence of approval in the text and the task component of the working alliance.

Likewise, there is a negative correlation between disapproval and the bond component (-0.44). Also, remorse is positively correlated with the task, bond, and goal components (ranging from 0.42 to 0.46). Confusion and excitement are negatively correlated with the bond component (-0.47 and -0.40) and disgust is positively correlated with the bond component (0.39).

4.1.3 Speech Affect features

Correlation results with the patient scores (WAI-S)

There are no significant correlations between the patient scores and the valence and arousal features from the audio.

Correlation results with the therapist scores (WAI-SRT)

The maximal arousal is positively correlated with the bond and task components (0.47 and 0.42), see Figure 15 for a visualization of this correlation.

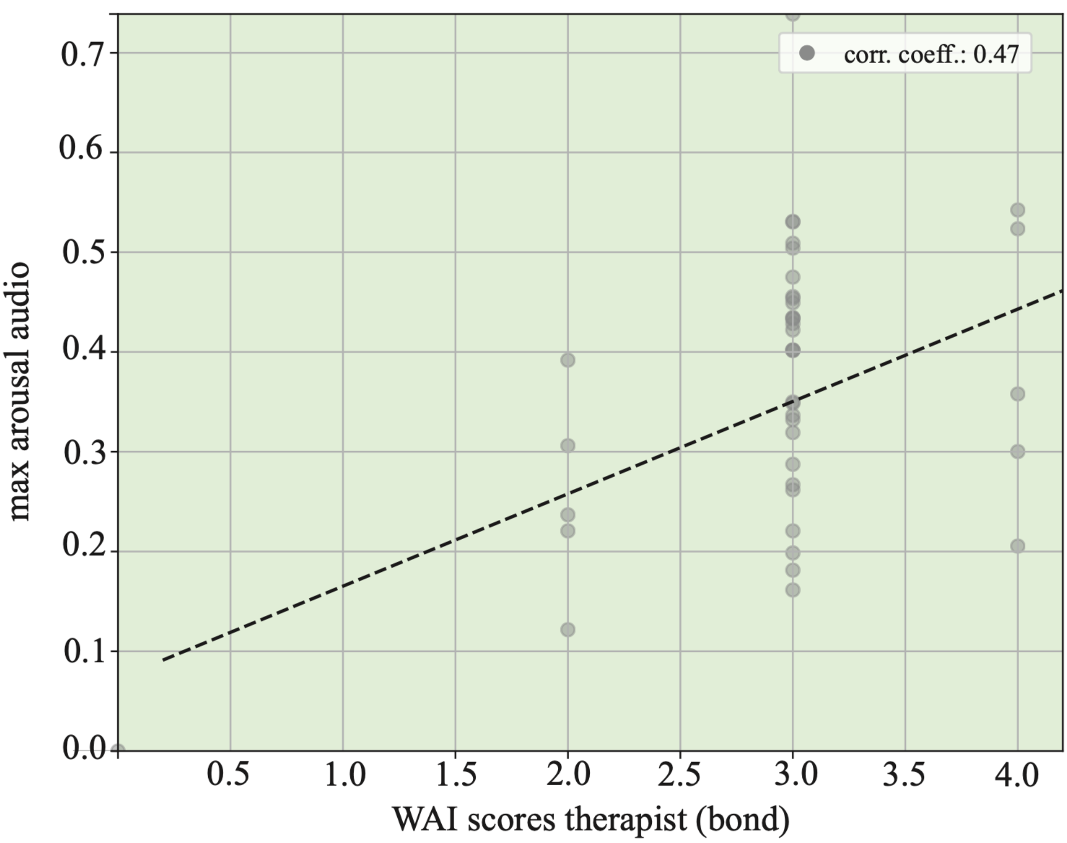


Figure 15. Example of a significant correlation between the maximum arousal in the audio and the task component of the working alliance.

The maximal valence is positively correlated with both bond, task, and goal (0.63, 0.54, and 0.47). The minimal valence of the audio is positively correlated with the bond and goal components (0.57 and 0.47).

Correlation results with the Observer scores (WAI-S)

The maximal arousal is positively correlated with the bond component (0.41) and the maximal valence is positively correlated with both bond, task, and goal (0.56, 0.53, and 0.50). Further, the minimum valence is also positively correlated with the bond, task, and goal components of the working alliance (0.53, 0.45, and 0.47).

4.1.4 Facial features

Correlation results with the patient scores (WAI-S)

The patient WAI scores show a significant negative correlation between the anger emotion and the goal component of the working alliance (-0.63), but also a positive correlation between happiness and the task component (0.62), see Figure 16 for a visualization of this correlation.

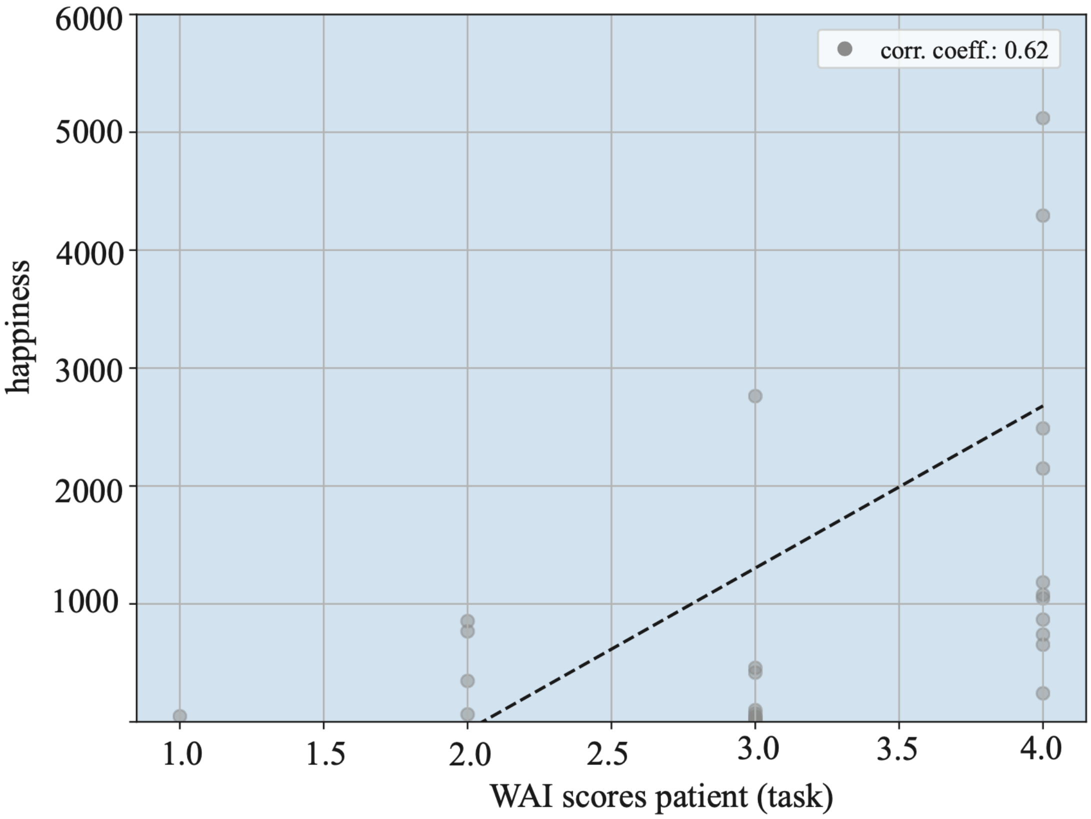


Figure 16. Example of a significant correlation between happiness and the task component of the working alliance.

Correlation results with the therapist scores (WAI-SRT)

The therapist's WAI scores show a negative correlation between sadness and the goal component of the working alliance (-0.41). Moreover, a neutral facial expression is positively correlated with the bond and goal components of the working alliance (0.48 to 0.47).

Correlation results with the observer scores (WAI-S)

The observer WAI scores show only one significant correlation for the facial features, namely a negative correlation between anger as a facial expression and the task component of the working alliance (-0.50).

4.1.5 Head movement

Correlation results with the patient scores (WAI-S)

The patient scores show a positive correlation with the patient's head movement behavior and a negative correlation with the therapist's head movement behavior. The nods displayed by the patient during the listening periods are positively correlated with the task and goal components (ranging from 0.50 to 0.54). The nods and shakes displayed by the therapist during listening behavior are negatively correlated with the bond component (-0.77), see Figure 17 for a visualization of the correlation.

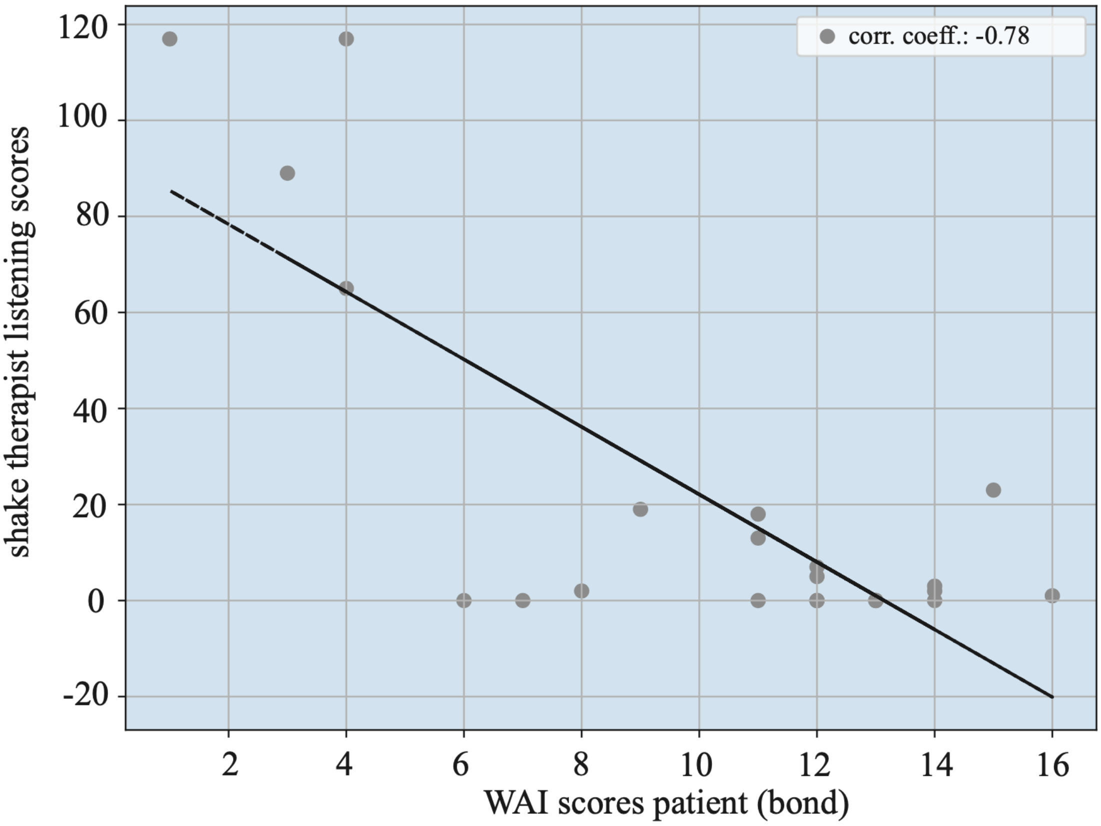


Figure 17. Example of a significant correlation between the head shakes of the therapist during listening and the bond component.

Correlation results with the therapist scores (WAI-SRT)

The therapist scores of the WAI show no correlations with the listening or speaking behavior of either the patient or the therapist.

Correlation results with the observer scores (WAI-S)

The observer WAI scores also show only negative correlations between the listening head movements and the working alliance. Interestingly, the observer scores are correlated with the behavior of only the patient. The head nods and shakes are correlated with the bond and goal components of the working alliance (-0.58 and -0.56).

4.2 Combined feature analysis

4.2.1 Feature selection comparison

We used two feature selection methods, as mentioned in Section 3.5.3, the MRMR feature selection and the manual selection of features that showed a significant correlation with the WAI score according to the Spearman correlation test. Figure

18 and Figure 19 show a comparison of the RMSE and the R-squared (fit to the data) for different feature selection methods and the baseline of no feature selection. The RMSE and the R-squared values both show that there is no specific feature selection method that works better than the others in every case. This plot is meant to give an idea of the comparative values between the feature selection methods. The actual values are less important, but we use them to compare which feature selection method overall seems to perform best, thus showing the lowest RMSE and highest R-squared.

For the Therapist, the no-feature selection method shows the overall lowest RMSE and highest R-squared, followed by the MRMR feature selection. We will, however, continue to use a feature selection method because fewer features as model input reduce the chances of overfitting and is, therefore, good practice to implement.

For the patient, the Spearman feature selection has lower RMSE and higher R-squared values compared to the rest.

Finally, for the observer, there is no specific improvement that either one of the feature selection methods shows, as the no feature selection shows a high RMSE for some features and the MRMR selection for other models. The Spearman feature selection seems to perform worse for the observer scores.

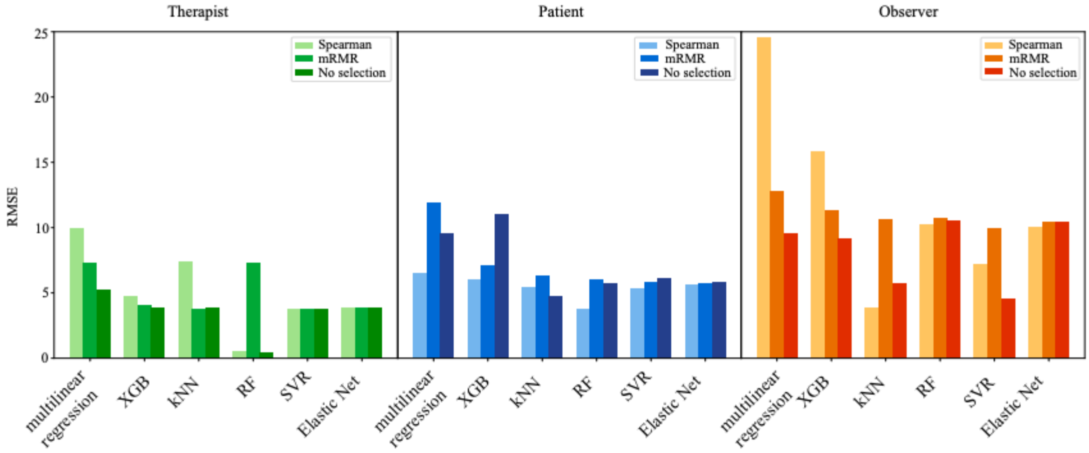


Figure 18. Comparison of Feature Selection Methods for the patient, therapist, and observer datasets (RMSE). For each rater, the RMSE values per model per feature selection method are plotted. There is no feature selection method that shows the lowest RMSE scores consistently for each prediction model.

The difference between the feature selection methods of the Spearman correlations and the MRMR selection is that they select features in a fundamentally different manner. The Spearman selects features that have a significant correlation with

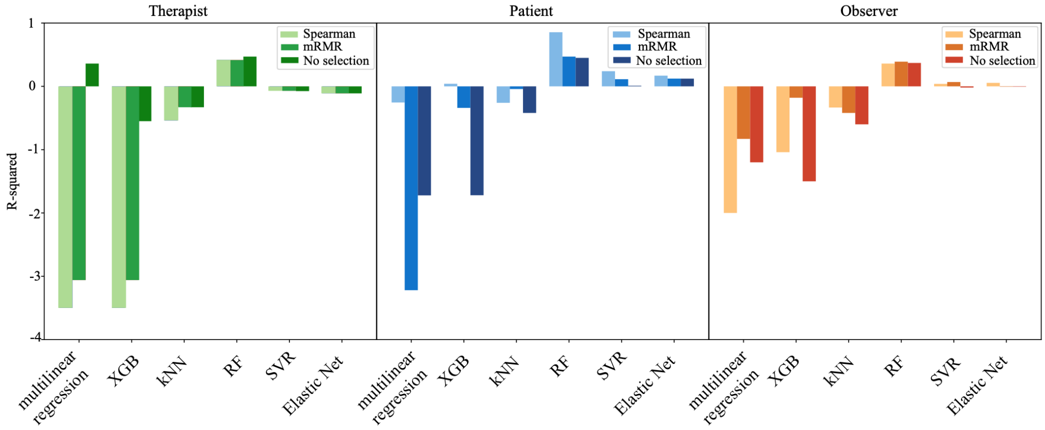


Figure 19. Comparison of Feature Selection Methods for the patient, therapist, and observer datasets (R-squared). For each rater, the R-squared values per model per feature selection method are plotted. There is no feature selection method that shows the highest R-squared scores consistently for each prediction model.

the WAI and does not take non-significant relevance into account. The MRMR selection starts with all features and removes those that are either irrelevant or highly redundant (as they are too highly correlated with other features). Therefore, the MRMR feature selection method includes more features and better captures the relationship between a multimodal feature set and an outcome measure such as the WAI. While the features with a positive MRMR score might not be predictive on their own, they might be predictive as part of a set of features.

Based on theoretical knowledge, the MRMR model should perform best, since Spearman feature selection ensures only the use of features that are significant predictors, the MRMR takes all features and their interrelatedness into account. The Spearman test can show a result of two features with high predictive correlations with the WAI but that also have a high correlation with each other, meaning that this feature will be "over-represented" in the final model. The MRMR feature selection, on the other hand, also takes redundancy into account to only select features that contribute in a unique way to the overall prediction to keep the necessary features to a minimum without reducing the predictive ability of the model. Therefore, we will implement an MRMR feature selection and only keep the features with positive importance values before training the prediction models.

4.2.2 Patient results

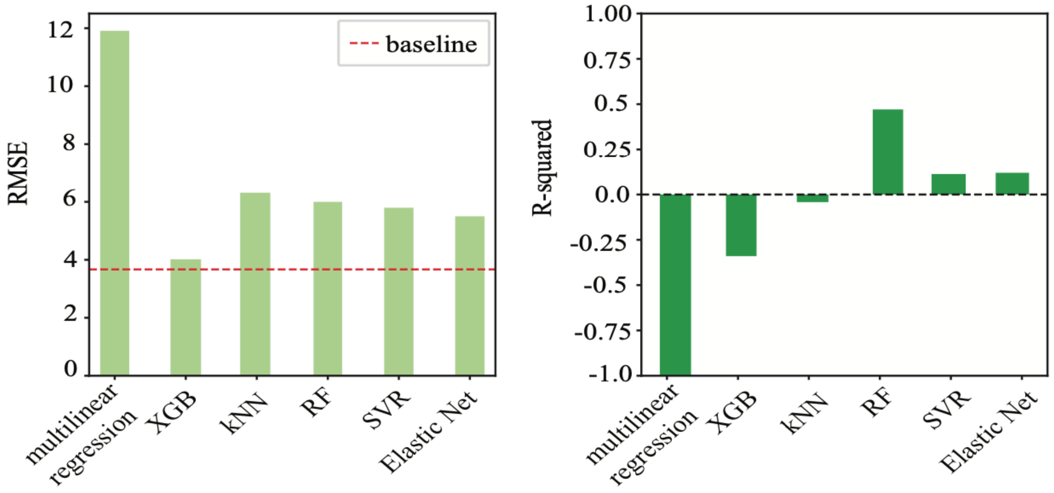


Figure 20. Comparison of predictive performance for different regression models on the patient WAI scores based on the RMSE and R-squared. No model shows an RMSE score lower than the baseline value, and only the RF shows a moderately high R-squared value.

The RF model was the only model with a moderately high positive R-squared value. The multilinear regression, XGB, and kNN models all had negative R-squared values, whereas the SVR and Elastic Net had a very low R-squared of around 0.1, see Figure 20.

The RMSE baseline for the patient WAI scores was 5.80 and no model showed an RMSE lower than the baseline (4.82), suggesting that was not able to predict the WAI score for the patient data on a higher than the mean level. The RF, SVR, and Elastic Net models all showed an RMSE score around the baseline (varying from 5.78 to 5.83), meaning that the model did not perform better than the baseline. The multilinear regression XGB and kNN models had a higher RMSE than the baseline and combined with a negative R-squared value showed to not model the WAI scores well.

We also tested the data categorically using an SVM classifier. The SVM had a cross-validation accuracy of 0.81 with a standard deviation of 0.17. The precision, recall, and f1-score was 0.75 for the high category and 0.80 for the low category.

Feature importance according to the MRMR selection The MRMR feature selection analysis showed that all modalities contributed to the prediction, but on average the conversational features show a higher importance. Interestingly, both the average change of some features within a session as well as the average value of that feature for a session contributed to the prediction without being redundant. Thus, we know that the change within a session is also an important aspect to take into consideration when predicting the working alliance. The features that showed the highest importance were in the textual domain the emotions of amusement, curiosity (average change), approval, and fear, and the minimum arousal of the therapist, and from the conversational modality the speech rate of the therapist, the participation equality, and the turn length of the patient and therapist. Most of these features also showed a high correlation with the WAI score in the Spearman correlation test. For more details, the reader is referred to Appendix 8.5.1.

4.2.3 Therapist results

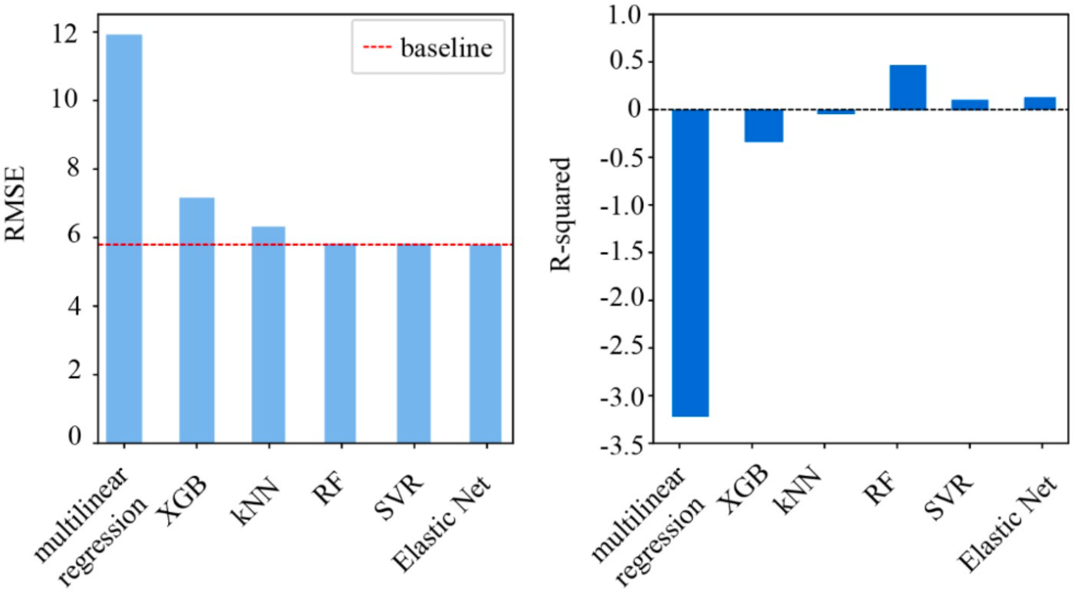


Figure 21. Comparison of predictive performance for different regression models on the therapist WAI scores based on the RMSE and R-squared. No model shows an RMSE score significantly lower than the baseline value, and only the RF shows a moderately high R-squared value.

In the therapist scores, we see a very similar trend to the patient scores with only the RF having a moderately large positive R-squared value, see Figure 21. The RMSE scores for the RF, SVR, and Elastic Net were higher than the baseline value, which was 3.67 for the therapist results. The RMSE of the Multilinear regression is again higher than the baseline, similar to the patient results. The XGB regressor has an RMSE that is closest to the baseline of all models, however, it is still above the baseline. The kNN RMSE is also larger than the baseline, in contrast to the patient scores. The R-squared value was again moderately high for the RF model (0.47), and the SVR and Elastic Net showed a small positive value (ca 0.12). Overall, there were no models that showed very good performance, and no RMSE lower than the baseline.

The SVM had a cross-validation accuracy of 0.93 with a standard deviation of 0.08. The precision was 0.71 for the high category and 1.00 for the low category. The recall was 1.00 for the high category and 0.33 for the low category. The f1-score was 0.83 for the high category and 0.50 for the low category.

Feature importance according to the MRMR selection The MRMR feature selection results of the therapist show that the conversational features are less important than for the patient. Also, text emotion features have a higher importance than for the patient, which is manifested in more emotions in the text with a moderately high importance level. The most predictive features according to the MRMR selection are the minimal valence of the patient, the presence of curiosity in the text, the speech rate of the patient and therapist, and the turn length of the therapist. For more details, the reader is referred to Appendix 8.5.2.

4.2.4 Observer results

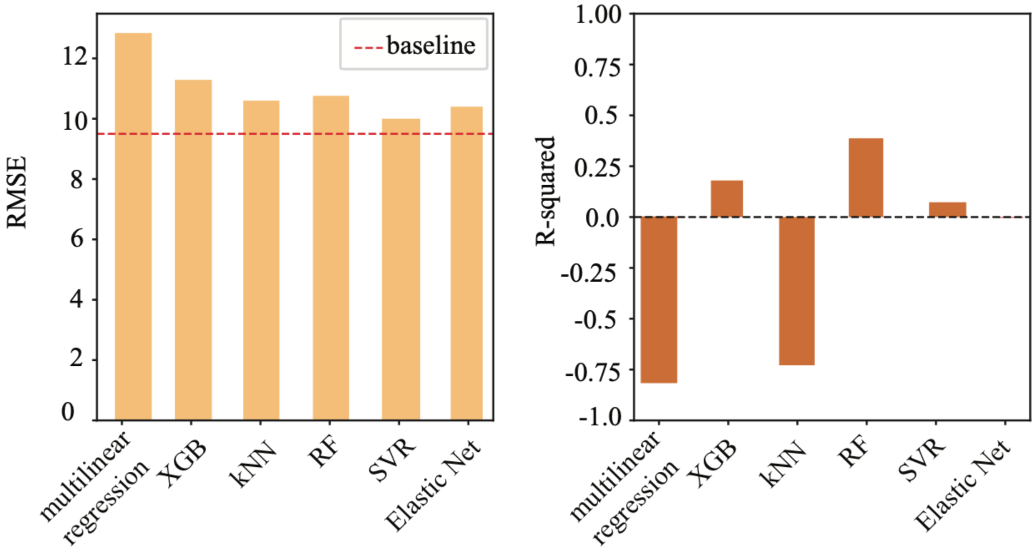


Figure 22. Comparison of predictive performance for different regression models on the observer WAI scores based on the RMSE and R-squared. No model shows an RMSE score lower than the baseline value, only the RF shows a moderately high R-squared value and the XGB shows a moderate R-squared value.

The observer R-squared scores showed a positive R-squared value for the XGB, RF, and SVR models with the largest positive value again for the RF model with 0.39, see Figure 22. The RMSE scores, however, did not show any models that scored lower than the baseline. As for the therapist scores, the kNN model and the

RF, SVR, and Elastic Net all showed similar RMSE scores just above the baseline. Figure 22 does not show it very clearly, but the Elastic Net had a negligibly small negative R-squared of -0.005)

The SVM had a cross-validation accuracy of 0.93 with a standard deviation of 0.09. The precision was 0.80 for the high category and 1.00 for the low category. The recall was 1.00 for the high category and 0.67 for the low category. The f1-score was 0.89 for the high category and 0.80 for the low category.

Feature importance according to the MRMR selection The MRMR feature selection for the observer scores reveals that there are two features that specifically show a very high importance, the turn-level-freedom and the average change of the arousal scores of the therapist. Further, the distribution of the feature importance over the modalities is similar to the patient feature importance. For more details, the reader is referred to Appendix 8.5.3.

4.2.5 Coefficient of variation scores

The coefficient of variance (CV) values are highest for the multilinear model (2.05), around 1.3 for the XGB regressor, and around 1 for the other models. A score of around 1 suggests that the RMSE is about the same as the natural variability found in the data. This pattern was consistent throughout the data for the patient, therapist, and observer. The only mentionable difference is that the CV value for the multilinear model is lower for the observer scores, 1.35 instead of around 2. For the complete tables on the patient, therapist, and observer dataset, see Appendix 8.4

5. Discussion

The goal of this research was to gain a better understanding of the working alliance so therapists can get more insight into their patient's perception of it. In the related works section, it is discussed how emotions, affect and turn-taking behavior can be indicative of the working alliance. An automatic prediction of the working alliance is time-saving and can be used to detect a low working alliance early in a therapeutic process which can prevent ruptures and drop-outs. It was therefore interesting to see what features are most indicative of the patient's perception of the working alliance, but also interesting to see how that compares to the observer's and therapist's perception of the working alliance. Furthermore, a model that can automate the Working Alliance Inventory (WAI) process for use as a detection tool of the working alliance in the therapeutic process is even more valuable.

The results show that the average change of certain features within a session and the average value of those features contributed to the prediction without being redundant. The most important features were emotions such as amusement, curiosity, approval, fear, and minimum arousal of the therapist, as well as the speech rate of the therapist, participation equality, and turn length of the patient and therapist. These features also showed a high correlation with the Working Alliance Inventory (WAI) score in the Spearman correlation test. Most regression models showed a low prediction performance due to a shortage of data points. The RF model showed a moderately large positive R-squared value, indicating a moderately good prediction performance. However, due to the nature of the RF model, it was likely overfitting on the small dataset. The SVM classifier showed an 80-90 per cent prediction accuracy on classifying whether a set of features from a session corresponded to a low or a high WAI score.

5.1 Explaining most interesting findings

An issue with the concept of the working alliance is that most interpretations of the data are subjective and not factual. If a correlation is present between a feature and the WAI score, multiple explanations could be attached to that correlation, each one plausible. For instance, a positive correlation between the facial expression happiness and the WAI score could be because the patient is happy if he/she notices a good connection with the therapist, but also because they are discussing some happy topic that leaves the patient feeling more optimistic. Both can cause the WAI to be higher but for different reasons. The former will likely be correlated with a good therapeutic outcome as well as the WAI, whereas the latter might not be beneficial as the problems of the patients are not solved by only discussing optimistic topics, and thus the therapeutic outcome might not necessarily be positive. We have explained each significant feature correlation and have tried to link the correlation to previous research as much as possible. However, a level of speculation in the explanations can not be prevented, and the reader should be aware of this. In some cases, the correlation was unexpected and no likely explanation was known, so we have only stated the unexpectedness without trying to fully explain it to keep as close to the factual truth as possible.

5.1.1 Patient WAI ratings

Conversational Modality

As we have seen in Section 4.1, the patient shows significant correlations between some features and the WAI score. For the conversational modality, the patient relies on multiple features for the working alliance perception. The patient views an equal participation rate as a better working alliance, specifically on the goal and task components (participation equality). This means that it is important for the patient that the therapist and patient both have an equal turn length as that means that both contribute to the goal and task. While participation equality is a measure of the difference in turn duration between patient and therapist, we also took the turn duration into account as a separate feature (speech duration for the total and average turn length for the average). Interestingly, a longer turn duration of the patient was positively correlated with a high bond perception, and a longer turn duration of the therapist was negatively correlated with a high bond perception. This means that longer turns of the therapist are rated as negative and longer turns of the patient as positive. In conjunction with the finding on participation equality, we can state that the therapist had on average longer turn durations than the patient, which caused an imbalance in the turn lengths. When looking at the data, we see that the average turn length of the therapist is higher than the patient's, as the average turn length of the therapist was 3.5 seconds compared to the average turn length of the patient of 2.7 seconds. Therefore, in the sessions with a shorter turn length for the therapist, the participation equality was higher as both speakers had similar turn lengths. A high participation equality results in a subsequent high working alliance.

The turn-level-freedom was negatively correlated with the working alliance (task), meaning that the high predictability of conversation was viewed as negative by the patient. The therapist shows a reverse correlation as high predictability is viewed as having a positive effect on the working alliance. A possible explanation for this could be that patients perceive high predictability as restrictive and would prefer a more open and free conversation and thus view high turn-level-freedom as a low working alliance, while therapists might view high predictability as providing structure and working effectively on the tasks, leading to a more positive view of the working alliance.

Patients view a high speech rate of themselves as belonging to a low working alliance, specifically the bond. A possible reason for this could be that when patients feel rushed, they talk faster but perceive a rushed feeling as negative for the bond with the therapist.

Textual modality

Patients also use the emotions of amusement and fear for their perception of the working alliance, specifically the bond component. A possible reason for this might be that showing these emotions could signify that patients trust their therapist enough so that they are comfortable sharing their emotions, which enhances the bond between patient and therapist.

There is a negative correlation between the minimal arousal in the patient's word choice and the bond and task component of the working alliance. A possible reason behind this is that low minimal arousal in the patient's language might indicate that they are not actively engaging in the conversation. This could lead to a weaker perception of the working alliance.

Facial modality

Happiness as a facial expression is correlated with a high working alliance, specifically the task component. A high display of anger on the patient's face, on the other hand, is correlated with a low working alliance perception of the goal component. A likely reason is that patients who perceive the working alliance as high might experience more positive emotions like happiness, which shows in their facial expressions. A high working alliance could cause a feeling of optimism and satisfaction which can lead to the display of happiness. On the other hand, if patients perceive the working alliance as low, they might feel disappointed or frustrated with the therapeutic process which might show on the patient's face as angry emotions.

Movement modality

The head movement of the patient, both the nods and the shakes, that occur during listening behavior are positively correlated with the task and goal components of the working alliance. This is likely a result of the process called backchanneling which we discussed in Chapter 2.1. This indicates that the patient is engaged in the conversation and wants to contribute, which is characteristic of a high perception of the working alliance.

Unexpectedly, the therapist's head movement, both the nods and the shakes, that occur during listening times of the therapist, are negatively correlated with the bond component of the working alliance. This means that if the therapist shows a lot of head movement, the patient perceives this as a sign of a low bond with the therapist. Instead of viewing this head movement of the therapist as the backchanneling that shows engagement, patients could view the excessive head movement of the therapist as distracting or indicating that the therapist is judging the patient. This might make the patient feel unsafe and cause a low perception of the working alliance bond.

5.1.2 Therapist WAI ratings

Conversational Modality

As appears from the results, the therapist has multiple different features that are correlated with their perception of the working alliance. The total number of turns in a conversation is positively correlated with the task component of the working alliance. More turns in a session indicate a more conversational structure with much back and forth between therapist and patient. This correlation indicates that the therapist views a more conversational structure as positive for the task component. The number of overlapping segments and the speech rate of the patient are also viewed as positive for the working alliance (goal and task). This could be caused by the therapist viewing a high overlap and high speech rate of the patient being very engaged in the conversation and contributing to the task and goal of the therapeutic process.

Contrary to the patient's perception of the working alliance, a high turn-level-freedom is linked to a high working alliance task component in the therapist ratings. In addition, if the therapist has a high percentage of speaking time, the working alliance task score is also high. Unsurprisingly, the average turn length of the patient is negatively correlated with the therapist's perception of the task because this means less speaking time for the therapist. A possible reason for this correlation is that the therapist feels free to become engaged and more dominant in the conversation if he/she feels that the working alliance task is high. The therapist wants to contribute and does so more when the working alliance perception is high. However, as we have seen, the patient does not interpret this high speech duration of the therapist in the same way. This means that there is a mismatch between how the patient and therapist observe their behaviors and its impact on their perception of the working alliance. For therapists, it is useful to know about this mismatch and adjust their method of evaluating the patient's perception of his/her behavior.

Textual modality

There is one emotion in the text that has a significant correlation with the bond component of the working alliance, which is disgust. A possible reason for this is that the therapist feels it is a sign of a strong bond when a patient is willing to talk about emotional topics that raise feelings of disgust. An important note to consider is that disgust is part of the topics of the conversations as it is a textual feature and not necessarily an emotion that is felt in the moment by the patient or therapist.

Further, a high number of sentences that contain positive sentiment is correlated to a lower task perception by the therapist. While the reason for this is unclear, a possibility is that the therapist perceives a high number of positive sentiment sentences as superficial, and which show an unwillingness of the patient to discuss deeper emotional issues or feelings. The therapist might feel that this hinders the advancement of working on the therapeutic tasks.

Speech modality

For the speech modality, it was found that the maximum arousal (bond and task), the maximum valence (bond, task, and goal), and the minimum valence (bond and goal) were all positively correlated with the perception of the working alliance by the therapist.

When a high maximum level of arousal and high maximum and minimum valence is displayed during the session, this might be perceived by the therapist as a high emotional engagement and openness between therapist and patient. The patient needs to feel comfortable to express strong emotions. Thus, if strong arousal and valence are displayed the therapist can interpret that as a better connection, more engagement, and a higher working alliance. Since the patient scores did not correlate with the speech modality, this suggests that therapists use the arousal and valence levels in a conversation more strongly than the patients for their estimation of the working alliance.

Facial modality

A high presence of sadness in the facial expression of the patient has a negative correlation with the goal component of the therapist's ratings. The therapist might perceive much sadness and emotional content as a divergence from the goal-related conversations to which the therapist might want to draw the focus. This mismatch could result in the therapist seeing this emotional content as a step away from working on the goals and thus in a lower rating for the goal component of the working alliance.

A high neutral count of facial expressions is correlated with a high bond and goal component of the therapist's WAI scores. While there is no obvious link between the neutral facial expression and the working alliance, it might be because patients who show much neutrality are actively listening and processing the therapist's explanations and advice. This might be perceived by the therapist as an engaged and eager patient who wants to work on his/her problems. This could affect the perception of the working alliance goal positively.

Movement modality

There are no specific head movements that were found to correlate with the therapist's working alliance. This finding was supported by Vail et al. (2021) and already discussed in Section 5.5, and was likely because therapists focus more on conversational behavior than head movement behavior for their perception of the working alliance (137). Moreover, the therapist was often not visible in the videos meaning that there is less data on the therapist's head movements. Since the head movements are more strongly linked to the person's behavior than their counterpart (as discovered by (137)), this lack of much head movement data could explain the lack of correlation.

5.1.3 Observer WAI ratings

Conversational Modality

Like in the therapist ratings, there is a positive correlation between the number of turns in a conversation (task), the turn-level-freedom, and the working alliance (bond).

Moreover, the negative correlation between the average turn length of the patient (task) and therapist (goal) and the working alliance is also present in the observer ratings. Also in agreement with the therapist ratings is that a high overlap between the patient's and therapist's speech has a positive impact on the perception of the working alliance. Observers perceive this as a sign of a strong task component, while therapists see this as a sign of a strong goal component.

Textual modality

The observers perceive the presence of approval (task), remorse (bond, task, and goal), and disgust (bond) in the text as having a positive effect on the working alliance. The presence of remorse and disgust can mean that the patient is willing to be open and honest about his/her feelings indicating that they trust the therapist and feel comfortable sharing their inner feelings. It can also indicate that the patient is self-reflective and able to acknowledge these feelings and show personal growth, which the observer would rate as a high working alliance. Moreover, approval might specifically indicate that the patient and therapist are in line about the methods and tasks for the therapeutic process which is of course characteristic of a positive task component of the working alliance.

Disapproval, confusion, and excitement were all perceived as being characteristic of a low bond component of the working alliance. The presence of disapproval, being the opposite of approval, might indicate that the patient and therapist are not in line with each other. Interestingly, this is correlated with the bond component instead of the task component. This could be because an uttering of disagreement can be a personal attack and harm the bond between therapist and patient more than it does the task component as negative emotions tend to feel more personal and harsh than positive emotions. The presence of confusion might indicate a mismatch between therapist and patient as they do not fully understand each other. The cause of the negative correlation between excitement in the text and the bond is not as clear, but perhaps excitement can be viewed as restlessness by the observer in which case it can be perceived by the observer as a disruption and a difficulty in communication between therapist and patient harming the bond.

Speech modality

The observer shows the same correlations within the speech modality as the therapist did. The maximum arousal, maximum, and minimum valence are positively correlated with the bond (also task and goal for the valence features). This suggests that the observers, like the therapists see a high display of emotion as positive for the working alliance, perhaps because patients also feel like this indicates that the patient is comfortable and trusts the therapist to talk about his/her deeper feelings.

Facial modality

The observer views a high display of anger on the face of the patient as being a sign of a low working alliance, specifically the task component. The patient showed the same correlation but with the goal component of the working alliance. We explained that a facial emotion of anger could be a sign of dissatisfaction or even frustration which the observer classifies, like the patient itself, as a sign of a low working alliance. The observer might feel like it hinders working on the task component.

Movement modality

The head movement of the patient is a significant predictor for the observer's perception of the working alliance between therapist and patient. Specifically, if there are a lot of head movements by the patients during periods when he/she is listening to the therapist this is perceived by the observer as negative. Perhaps the observer perceives this behavior as the patient being distracted or not paying attention to the therapist which causes the observer to take it into account as a sign of low goal and bond between patient and therapist.

Overall, we found that every three raters had a different way of looking at the sessions and the interaction between therapist and patient and thus perceived the working alliance somewhat differently. The observer ratings are in most cases more similar to the therapist ratings but since each rater's dataset had different sessions for which WAI ratings were available, we can't compare the feature importance between the patient, therapist, and observer in detail.

5.2 Evaluation of the combined feature WAI prediction models

We have seen the results of the models fitted to the data in Section 4.2.2 to Section 4.2.4. There exist similar trends between patient, therapist, and observer ratings, but also a few differences. In this section, we will evaluate, compare, and explain these model results and place them in the context of predicting the working alliance. First, we will look at the patient, therapist, and observer results and compare them. After this, we will compare the categorical SVM results with the regression models. Finally, we will make a general conclusion about the model results.

5.2.1 Overview of the model results

Patient model results

As we can see in Section 4.2.2, the models showed very different performances, but none showed evaluation scores indicating that they were able to model the WAI scores well. This means that none of the models was able to predict the working alliance well. The RF model showed the best performance with a moderately high positive R-squared value and RMSE score around the baseline. This suggests that the RF was able to capture a part of the variety in the data well. Due to the nature of the RF design, dividing the data into branches of a tree to capture the differences between the data, it is likely that an RF would overfit with a dataset smaller than 30 items. While measures were applied to prevent overfitting, such as a minimum leaf size of 3 and a maximum depth of the trees, the positive RF results have to be interpreted with some care. However, since the performance of the RF on an unseen test set was slightly worse but not considerably (R-squared of ca 0.30), the RF is still the best-performing model on this data of the chosen models.

The SVR and Elastic Net models showed a performance of around the baseline, indicating that achieved a similar level of predictive performance as simply using the mean value. Since they showed a slight positive R-squared value of around 0.12 they will likely perform better with a larger dataset. The Multilinear Regression, XGBoost (XGB), and k-Nearest Neighbors (kNN) showed considerably worse performance with negative R-squared values and RMSE scores above the baseline. This suggests that these models struggled to capture the underlying patterns in the data causing a bad fit and inaccurate prediction of the working alliance.

Therapist model results

The patterns seen in the therapist outcomes are similar to those in the patient outcomes. The RF model shows the best performance while the other models have difficulty capturing the underlying pattern of the dataset. The RMSE values show that none of the models were better than a baseline prediction.

Observer model results

The observer results show again a similar trend of the models not being able to score a better (lower) RMSE than the baseline prediction. However, the R-squared values were positive for the XGB model as well, in contrast to the patient and therapist scores. This suggests that the XGB model can explain a part of the variance of the observer WAI scores. The scores of the other models were similar to the previous scores and indicate that the RF and SVR models can capture a small portion of the WAI scores, but the Multilinear regression and kNN models are not. The Elastic Net model is negligibly small, suggesting that it does not perform well in capturing the complexities of the working alliance.

Another performance measure was the CV whose results are given in Section 4.2.5 which represents the relative variability of a dataset compared to its mean. We explained that CV values >1 suggest that the model's errors are relatively larger compared to the natural variability in the data, but that a CV of <1 suggests that the model's predictions are within a reasonable range of expected fluctuations. We mainly focus on the RMSE and R-squared scores to evaluate the models as the CV and the RMSE partly represent similar aspects of the model's performances (the relative variance). However, we use the CV as a check to compare the performance across the datasets, as the baseline RMSE of each dataset is different, and calculating the CV makes it easy to compare patient, therapist, and observer scores. The most important takeaway from the CV results is that their consistency across the patient, therapist, and observer scores indicates that the models face similar challenges in predicting the WAI on this dataset. The only notable change in CV scores was between the observer scores and the patient and therapist scores in the Multilinear regression and XGB model results. The observer results show a lower CV score for the Multilinear regression compared to the patient and therapist, but a higher CV score for the XGB regression model. This means that the XGB model probably struggles more with predicting WAI scores in the observer dataset, whereas the Multilinear regression model might struggle less in this dataset. However, since there is no consistent good prediction performance across the models, it is more likely that the CV difference is due to a meaningless fluctuation in the stability of WAI predictions, especially since the other models show no difference in CV values for the observer dataset. Nevertheless, this might be something to look into in similar research on this dataset to exclude a structural difference between these datasets.

5.2.2 Explaining negative R-squared

The calculation and definition of the R-squared value were explained in Section 3.5.3, where was stated that it was a value between 0 and 1 with a higher value indicating a better explainability of the variance in the data. In that section, we also explained that a negative R-squared was possible if the sum of the squared residuals was larger than the total sum of squares. This happens when the model's fit to the data is worse than a horizontal line so it fits very poorly to the data. The models that show a negative R-squared in our results are therefore very bad at predicting the WAI scores.

5.2.3 Categorical v numeric

Since this research uses a small dataset, a categorical SVM model was fitted on the data as well. By dividing the dataset into low or high WAI scores, some of the natural variability, that hinders the ability of regression models to find the underlying pattern in the data, would be eliminated.

The SVM classifier shows good prediction results. Since we applied cross-validation on unseen data, we know that it is not overfitting. The accuracy of the SVM is relatively high across all three datasets, ranging from 0.81 to 0.93. This suggests that it is successful in categorizing the features from a therapy session into high or low WAI scores.

Table 12. Performance results of the SVM Classifier on patient, therapist, and observer WAI score prediction

Dataset	Accuracy	Precision	Recall	F1-Scor
		(High-Low)	(High-Low)	(High-Low)
Patient	0.81	0.75-**0.80**	0.75-**0.80**	0.75-**0.80**
Therapist	0.93	0.71-**1.00**	**1.00**-0.33	**0.83**-0.50
Observer	0.93	0.80-**1.00**	**1.00**-0.67	**0.89**-0.80

From the recall, we can see how the SVM performs for each category. The precision, which measures the proportion of correctly predicted instances in a category, is higher for the low category. The recall, which measures the proportions of instances that were predicted correctly, was higher overall for the high category (except in the patient data). The F1-score gives a balanced value for the model's performance and is reasonable for most categories and models, except for the therapist's low category which scores at 0.50 which indicates that the SVM struggles with predicting the low WAI sessions in the therapist dataset.

To summarize, the SVM models for the therapist and observer dataset show a high accuracy but struggle with predicting the low WAI class. The SVM trained on the patient dataset also performs reasonably well but struggles more with the high WAI category and seems to perform well in the low WAI category.

If we compare the SVM classifier results with the regression model results, the classifier is much better at predicting the working alliance than the regression models. While a categorical prediction offers less insight into the working alliance than a regression model, it is still useful for the therapist to know whether a patient's perception of the working alliance is high or low.

5.2.4 General conclusion on model performance

In conclusion, the SVM classifier outperformed the regression models in predicting the working alliance scores. Since the regression models performed very poorly, we are unable to state with certainty whether a classification model is better suited for predicting working alliance scores than a regression model. The RF was the best-performing regression model and was able to capture a moderate portion of the variance in the data. However, due to the limited dataset size, these findings should be interpreted carefully as the performance of the models is expected to improve with a larger dataset.

Overall, this research highlights the complex nature of the working alliance and the importance of a detailed and large dataset for modeling the working alliance. The difficulty of modeling the data better than a baseline RMSE indicates the complexity of capturing psychological processes using automatic predictive models. However, the importance stimulates further research into refining approaches to provide a deeper understanding of the factors underlying the working alliance.

5.3 Answering the research questions

Research question: To what extent can a machine learning model predict the quality of the working alliance between therapists and patients using visual and textual features extracted from psychotherapy sessions?
We found evidence that a machine learning model can likely predict the working alliance using visual and textual features, but we can't conclude with certainty as the regression models were unable to perform well due to the small dataset limitation. The SVM showed good accuracy in predicting whether a session had a low or high WAI score. Thus, using the SVM we were able to predict high and low working alliances with 80 to 90 per cent accuracy. A regression model, however, was not predictive.

Sub-research question 1: Which machine learning model demonstrates the highest predictive performance for predicting the working alliance in psychotherapy sessions based on multimodal features?
The model evaluation scores (RMSE and R-squared) indicate that the SVM classifier and the RF were the best-performing models (see Section 4.2). However, the answer to this question remains ambiguous as the models in general were unable to fully observe the trend in the features due to the limited dataset size. With sufficient data, the models with the greatest likelihood of a good performance are the XGB regressor, the RF, SVR, and Elastic Net (as these are powerful regression models that show overall low RMSE values). An interesting possibility is that neural networks could also perform well, as they have done in previous research (see Section 2.4), but this will have to be tested in future research.

Sub-research question 2: In a psychotherapy session, what is the difference in predictability between indicators displayed by the therapist and indicators dis-

played by the patient for predicting working alliance?
The indicators showed a similar level of predictability between the patient, therapist, and observer WAI scores, but differences in which specific features contributed to its WAI score prediction. There was a difference in the importance of which modality was predictive for which perception, as conversational features were more correlated to the patient WAI scores, and the audio affect features were more correlated to the therapist WAI scores. Furthermore, the WAI of patients and therapists was more informed by the rater's behavior than that of the other speaker. The observer results do not necessarily resemble the patient or therapist results, indicating that the three different raters each have different underlying features for perceiving the working alliance. In general, the difference in predictability between indicators displayed by the therapist, and indicators displayed by the patient is that the important specific features differ, but the overall predictability based on the features does not.

Sub-research question 3: What specific features among the visual and textual indicators extracted from psychotherapy sessions exhibit the strongest predictive power for determining the perception of the working alliance quality?
Due to the size limitation of the dataset, feature selection did not reveal a significant improvement in the predictive performance of the prediction models. We can conclude that most modalities contain a few features that contribute to the working alliance prediction (see Section 5.1 for an overview of which features per modality), based on the MRMR feature selection and significant correlations between the features and the WAI score from the Spearman test. Additionally, the conversational features show the highest predictability.

5.4 Comparison with Bayerl et al. (2022) paper

Earlier work from Bayerl et al. (2022), shows that the conversational features have a strong correlation with the components of the working alliance, indicating that the entirety of these features can contribute to a better understanding of the working alliance between therapist and patient. Our research shows similar findings, although the features have correlations with other questions of the WAI than what was described by Bayerl et al. (12). However, when looking at the correlation between the feature and the component of the working alliance, similar patterns arise. We will compare the correlations found by Bayerl and the correlations found in this research, and explain the differences where desired and possible.

The correlations found by Bayerl were between conversational features and the WAI ratings according to observer-generated ratings. The features and ratings are the same as in this research, although we also include patient and therapist WAI ratings. The dataset that Bayerl used consisted of 5.9K turns with an average turn length of 12 tokens. In comparison, our observer dataset consisted of 20k turns with an average turn length of 9.6 tokens. This means that Bayerl has around 70.1k words for their 23 recorded conversations compared to 200k words in our dataset for 34 sessions. An important difference, of which it is uncertain whether it can influence the results, is that the language of the corpus used by Bayerl is Italian and the language of our dataset is Dutch.

Participation equality

Bayerl found that participation equality was correlated with the task bond and goal components with a correlation ranging from 0.42 to 0.53. We did not find a correlation between participation equality and the WAI within the observer dataset. However, there was a positive correlation within the Patient dataset between the participation equality with the goal and task components (0.30 and 0.31), which were also the strongest correlations in the Bayerl research. Thus, these results are in agreement with the correlation between participation equality, goal, and task found by Bayerl.

Turn-level-freedom

The turn-level freedom that measures the predictability of the turn-taking structure is found by Bayerl to be positively correlated with all questions of the task component (Spearman correlation ranging from 0.43 and 0.54). Our research shows a positive correlation between the turn-level-freedom and the bond component (0.28). For the therapist dataset, we found a positive correlation with the task component (0.30). These results support the findings of Bayerl.

Therapist turn duration

As a small footnote in their paper, Bayerl also mentions that a negative correlation is seen with the therapist's total percentage of turn duration. They don't give any further information or correlation coefficients, but these results correspond to our finding that the average turn length of the therapist is negatively correlated with the task component in the patient and therapist ratings (-0.30 and -0.25 for the therapist scores and -0.31 correlation for the patient scores). The average turn length of the therapist in the observer dataset shows a negative correlation with the goal component (-0.32). Unexpectedly, however, this finding is accompanied by a positive correlation between the duration percentage of the therapist and the task component. This suggests that the therapist views longer total speaking times for him/herself as productive for the task-oriented discussions while having relatively few turns with a long turn length is viewed as inhibiting the goal component. Thus, according to the therapist, it is more constructive if his/her role is to engage actively in conversation with many turns, but not speak for too long at a time. This is in line with the finding of Bayerl that the therapist talking for a large total percentage of the session negatively impacts the working alliance.

Therapist speechrate

Bayerl found that the minimal speech rate of the therapist is positively correlated with the bond and task components (0.69 and 0.60), while no correlation with the average speech rate of the therapist has been found in our data. The lack of correlation in our data is likely because we did not take the minimal speech rate into account, but only considered the average speech rate. The MRMR selected features found that the speech rate of the therapist does contribute to the total WAI score prediction, suggesting that speech rate is a predictive feature for the WAI, but not strongly enough to result in a significant correlation on its own.

Number of overlapping turns

Further, the number of overlapping turns has been found by Bayerl to be positively correlated with the task (0.42 and 0.47) and bond (0.42) components. In our research, we found a positive correlation between the overlapping turns and the task component of the working alliance as well (0.25). In the therapist scores a positive correlation with the goal component was found (0.35).

Total number of turns

Furthermore, our observer ratings show that there is a positive correlation between the number of turns and the task component (0.33). Bayerl, however, did not take this feature into account in its research.

Correlations with session segments

In addition to the features described above, Bayerl also divided the sessions into three segments, the first and last 15 per cent of the total duration were classified as the start and end of the session and the remaining 70 per cent as the middle segment. They found some features to only have a significant correlation in specific segments. We did not split up the sessions into these components as the recordings did not necessarily start at the beginning and end of the session. Hence, it would be inaccurate to divide up the recordings into segments using percentages. However, we will evaluate their segment results and compare them to our own.

Standard deviation of therapist turn duration
Bayerl found a negative correlation between the standard deviation of the therapists' turn length and the goal component. Bayerl hypothesizes that long stretches of the therapist talking to the patient indicate schooling the patient instead of talking. If that is followed by segments of conversation, that creates a high standard deviation and a negative impact on the goal component. Our observer results support their hypothesis as they also show a negative correlation between the average turn length of the therapist and the goal component of the working alliance (-0.32). This, in combination with a positive correlation between the turn-level-freedom and the bond component (0.28) indicates that the conversation is not predictable, which happens when the therapist's turn lengths are very diverse. The therapist has long turns during schooling moments, but also shorter turns during a quick back-and-forth conversation. This diverse turn-length results in a low turn-level-freedom, which in turn means a lower WAI score, which is what our research and Bayerl found.

Total turn duration patient
Bayerl also found that the total duration of the patient's speech during the middle segment of the session is positively correlated with the bond. In the patient WAI results, it has been found that patients experience a positive correlation between the duration of the patient's speech and the bond component of the working alliance (0.39). These results support each other and the finding that the more the patient talks (engages in the conversation) during therapy, the more confidence he/she has that the therapist can help him or her. This hypothesis is based on the correlation with question 5 of the WAI which states that: *"The client feels confident in the therapist's ability to help the client."*.

Median turn duration patient
Bayerl further showed that the median duration of the patient speaking in all three segments of the session is positively correlated with the task. Bayerl hypothesizes that the more regularly the patient engages in the conversation with the therapist, the more he/she is motivated to work on problems correctly, due to the correlation with the specific WAI question 12: *"The client believes that the way they are working with his/her problem is correct"*. This finding is closely related to the previous one, as the total and median duration of the patient's speech are likely highly correlated. However, Bayerl does not mention the correlations between features so it is difficult to state with certainty.

Explaining differences between the results from Bayerl et al. (2022) and our findings

The correlations from the Bayerl results were not always found in the observer's data, but sometimes in the therapist's or patient's data. This can be explained by the highly fluctuating consistency of the inter-annotator agreement (IAA) and the correlations to dialogue features as there were multiple questionnaires with very low internal consistency (3 of the 23 with lower than acceptable performance). This could mean that every observer looks at the sessions differently and perhaps some observers evaluate the session more from the viewpoint of the therapist and some from the viewpoint of the patient. It is also possible that the IAA between our observer ratings and the Bayerl ratings is not very high explaining the change in feature significance.

Most differences are likely due to a difference in the transcription and diarization. Although the methods of segmentation and extracting overlapping segments are similar between Bayerl and our research, we have used automatic transcription and diarization tools whereas they have manually transcribed and diarized the data. Since we found that the diarization does not perform as well as manual diarization would on some noisy audio, this could explain why there is more noise in our dataset due to some errors in the diarization.

This difference in diarization assignment is likely also the cause of the difference in correlation strength. The correlations that Bayerl found are generally higher than our correlations, which can be due to the higher level of misclassifications in the diarization. An automatic diarized dataset contains more noise and is more spread out compared to Bayerl with a (nearly) perfect speaker diarization.

Despite the high correlations that Bayerl found, it is good to mention their limited dataset size. Moreover, they only include 3 participants with multiple sessions. Their paper does not mention how they handled the dependency in the data, but finding a correlation in only three patients has the risk of limited generalizability. Especially in the significant correlation of the number of overlapping turns the slope corresponding to the correlation is 0.02 and 0.03 which means that with a change in the WAI, there is only a very small change in the number of overlapping turns. It is, therefore, possible that this correlation would disappear if more data is added. It would be interesting to obtain the data from this paper and perform a more extensive comparative study where the dependency of the Bayerl data is taken into account and a Spearman correlation can be applied like in this research.

Overall, most findings of Bayerl et al. (2022) are aligned with our conversational results which underline the importance of including conversational features in the prediction of the working alliance (12). Since the subject and methodology of the Bayerl research are so closely related to ours, it would be interesting to cooperate and compare the two datasets more in-depth.

5.5 Comparison with Vail et al. (2021) paper

5.5.1 Summary of the main similarities and differences

The research by Vail et al. (2021) is also closely related to ours (137). Their analysis focuses primarily on head gestures and turn-taking behaviors as features predictive of the WAI. In this Section, we will compare Vail's findings and our own, particularly in terms of dataset, feature importance, and model performance. The main takeaway of this comparison is that the features that proved important for the working alliance were similar between Vail and our research. Specifically, the head gestures during listening periods were significant predictors for patient working alliance ratings in both studies. Also, our study confirmed Vail's finding that therapists relied more on conversational features like turn length and wait time. Also, both studies found that therapists relied more on conversational behavior than head movements and both studies showed that head gestures were more reflective of the task component of the working alliance, while conversational behaviors tended to be more reflective of the bond component.

The model performance in our research was lower than Vail's research. Vail found that the SVR and Elastic Net models performed best, whereas in our research the VR, RF, and Elastic Net had similar RMSE values, but were still around or above the baseline RMSE, which indicates poorer model performance.

5.5.2 Comparison of the study designs

An important difference is that their dataset was with English-speaking patients and therapists and was larger than ours. They have 266 audiovisual recordings from 39 unique patients and 11 unique therapists. The sessions lasted 50.3 minutes on average. Their speaking turns were defined slightly differently, with a separation between turns of minimally 1 s whereas, for us, this difference had to be at least 0.5 s. The diarization was more accurate in Vail's study as two microphones were used. The extraction of the head nods and shakes was performed using the same methodology as this research.

Vail uses six features in their study: head nods, head shakes, speaking turn length, wait time (pause length between the end of the partner's turn and the start of the speaker's), listening nods, and listening shakes. We have extracted these features as well in our research, so we will compare the results. Since Vail takes the comparison of importance between the different features into account, we will use the feature importance scores from the MRMR feature selection as the comparative evaluation measure between our research and Vail.

Figure 23 shows the MRMR feature importance results for the patient dataset containing only the head movement and the turn length feature that Vail also uses. This Figure shows the relative importance of each feature to the prediction of the three WAI components. We used the three components instead of the total WAI to be able to compare our findings with Vail.

Figure 24 is a similar visualization but of the relative importance of each feature for the therapist's WAI score. The difference between the feature importance for the patient and therapist thus becomes clear by comparing these two Figures.

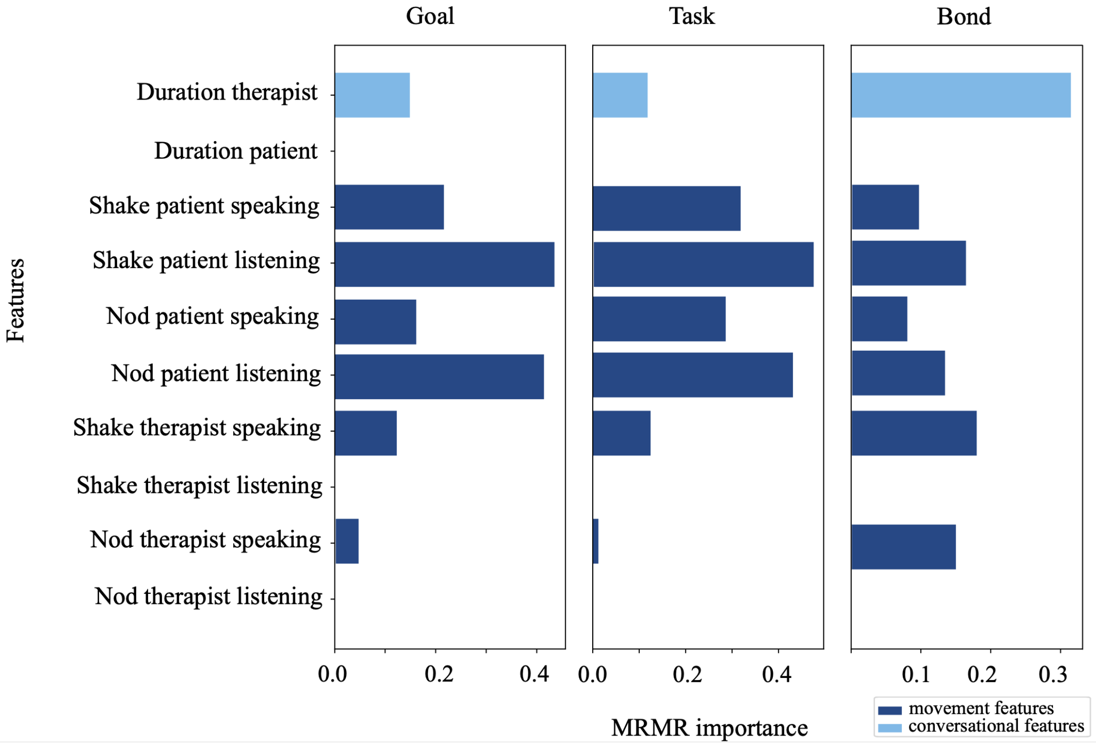


Figure 23. Positive feature importance of the goal, task, and bond components of the patient WAI

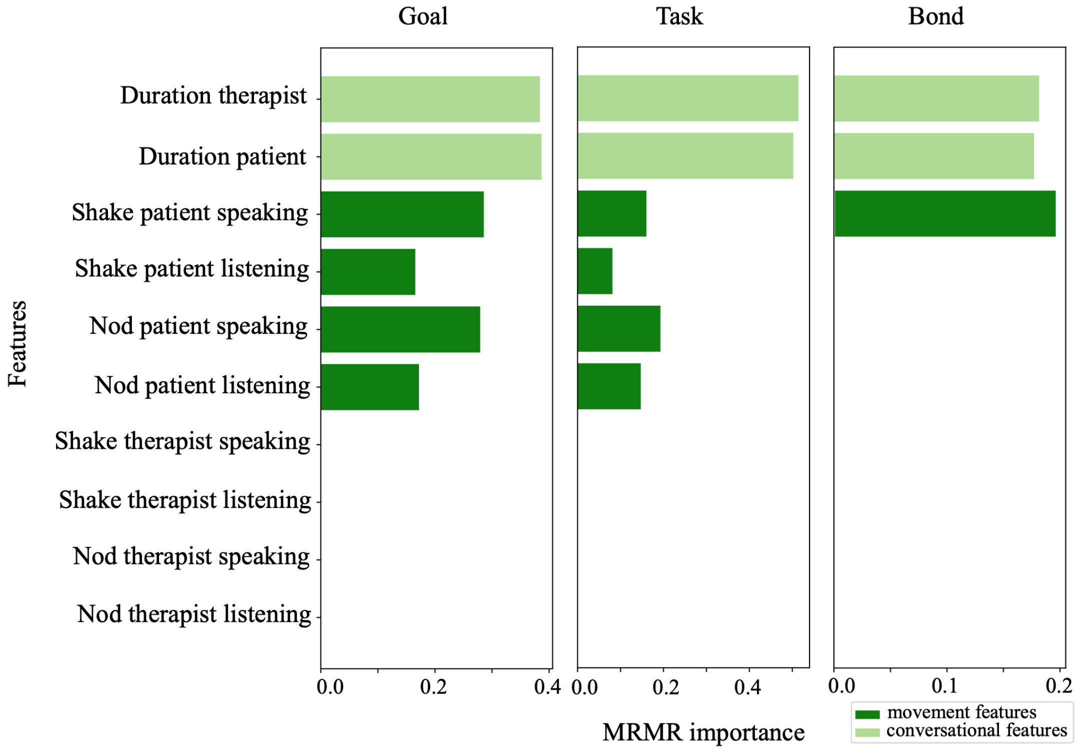


Figure 24. Positive feature importance of the goal, task, and bond components of the therapist WAI

Feature importance results comparison between Vail et al. (2021) and our research

Vail found that head gestures displayed by the patient during listening periods are significant predictors for the patient's working alliance ratings. Therapists did not have as strong an association between their head movements and the WAI ratings. Also, conversational features such as turn length and wait time appeared more important for the therapist's perception of the working alliance.

It was also found in this research that both speakers relied more on their own behavior than on the other speaker's behavior. Moreover, in some cases, the WAI ratings are misinformed by the other speaker's behavior. This was the case in head movement: when patients nodded more often, the patient had higher WAI ratings, whereas this behavior made the therapist rate the WAI as lower. This also worked the other way around.

Our results reveal partially the same relations between the features and the WAI. We analyzed a subset of our data, namely the features that correspond to the features used by Vail which were the head nods and shakes displayed by therapist and patient during speaking and listening times, and the average (speaking) turn length of patient and therapist. We calculated the MRMR importance values for this subset of features against the total WAI score for the patient and therapist.

As can be seen in Figure 23, the head gestures displayed by the patient during listening periods show high importance for predicting the patient's WAI scores. The head gestures of the therapist, on the other hand, show very little importance and even negative importance during listening times (this is not visible in Figure 23 as only the positive importance values are displayed for clarity).

The second finding of Vail is also supported by our findings, that the therapists relied more on conversational behavior than on head movements. This can be seen by comparing the relative importance of the conversational features and the movement features in Figure 24. The conversational features show a consistently high importance in task, goal, and bond components whereas the head movements show much lower importance values. Interesting to note is that therapists look at the average turn lengths of themselves as well as the patient.

A difference between our results and Vail's is that we found that therapists do not rely on their own head movements for their perception of the working alliance, but on the patient's head movements during speaking periods (see Figure 24).

Vail also found a trend between the types of features and the components of the working alliance. Head gestures appear to be more reflective of the task component of the working alliance, while conversational behaviors tend to be more reflective of the bond component.

This can be analyzed in our data by comparing the MRMR importance scores for the bond, goal, and task components of the WAI scores of the therapist and the patient datasets. As can be seen in Figure 23, where the feature importance for predicting the patient WAI scores is displayed, the head movement features are more important for the task component of the patient WAI, whereas the conversational behavior is more important for the bond component of the patient WAI.

This trend is not seen in the therapist scores, as the conversational behavior has a high importance score across all components of the WAI, see Figure 24. Moreover, the importance of the head movement scores even decreases in the bond component plot if we look at the absolute values (ca 0.20 in the bond component compared to ca 0.5 in the task component). Notable is a decrease in the number of important head movement features between the task and the bond components, suggesting that these might be more important for the task component than for the bond.

Model performance results comparison between Vail et al. (2021) and our research

We have explained the different models (SVR, Elastic Net, and RF) that this study uses in Section 2.4.2 so we will not explain these in detail again, but rather focus on the comparison of the RF, Elastic Net and SVR model performances of Vail and ours.

Vail evaluated their models by looking at the RMSE and whether it was lower than the baseline. Their argumentation was that RMSE has as much benefit over other metrics as the R-squared error and its stability in smaller datasets. Using this measure, they found that the SVR and Elastic Net models performed best.

When we look at our findings, we see that SVR, RF, and Elastic Net are generally the best-performing from the regression models. All three have very similar RMSE values but unfortunately, they are still around or above the baseline RMSE, meaning that the model is unable to predict the WAI scores well for either patient or therapist. We also took the R-squared measure into account and noticed that the RF model had significantly higher R-squared values than the SVR or Elastic Net models. However, as we discussed previously, the RF model is likely overfitting. This means that the model performance in our research is not as good as the model performances from Vail. Nevertheless, we know the limitations of our research and dataset (we will discuss this further in Section 5.7), which is why our findings do not erase the credibility of the Vail findings, but rather stimulate us to investigate the best model for working alliance prediction further.

Overall, when comparing our findings with Vail, we were able to replicate some of the findings on the feature importance of the patient and therapist WAI ratings. An added advantage of our research is that we supplement the head movement and conversational features with more features from various modalities. Perhaps in the future, the authors of the Vail et al. (2021) paper will focus on researching more modalities as well in order to facilitate another comparative research.

5.6 Most important takeaway for therapists

Part of the purpose of this research was to provide therapists with more insight into what factors patients take into account in their perception of the working alliance and the largest mismatches between the patient and therapist in their perceptions. Therefore, we will here give a short overview of the most important take-away messages for therapists to use in their future practice.

The most important for therapists is the confirmation of the idea that therapists predict the working alliance according to their patients often incorrectly. This was most prominent in conversational features and head movement. While the head movement of therapists during listening turns is correlated with a higher working alliance by the therapist, it is correlated with a lower working alliance as rated by the patient. Thus we can say that it is important for a high working alliance perception of the patients that the therapists do not talk for long periods at a time and show enough space for the patients to talk. Also, therapists should watch their behavior when patients are speaking and be aware that much head movement and a show of high arousal are perceived by the patient as a sign of low working alliance. Perhaps because this shows as being distracted or judgemental.

Also, where therapists view high predictability in the session's conversational structure as positive, patients view this as belonging to a low working alliance. Furthermore, the results on emotion display suggest that patients appreciate an open conversation with space to show their emotions, especially negative emotions.

For reading the patient's response during a session, the therapist can pay attention to facial expressions. Anger is a sign of a low working alliance, and a high head movement during listening times is a sign of a high working alliance.

While we have not been able to produce a robust and highly predictive model for capturing the working alliance perception automatically, we hope that these findings prove useful for therapists in gaining a bit more understanding of the patient's working alliance perception during a psychotherapy session.

5.7 Limitations

There are a couple of limiting factors in this research, the first of which is the WAI scores. As mentioned in Section 1.3, the WAI scores provide a rating of the perception of the working alliance according to the person who fills it in. Although the reliability of the questionnaires has been proven, the WAI is still only a momentary reflection of the working alliance between therapist and patient. We use the WAI scores as the golden standard in our research, but in truth, it is influenced by multiple factors such as mood, topic of the therapy session, age, or culture. Thus, it is important to keep in mind that it is an indirect measure of the working alliance and not in every case the perfect score to represent the working alliance. Nevertheless, it is the most reliable measure in the form of a questionnaire we can use in research.

The most prominent limitation is the dataset, both its size and quality. To clarify, the number of videos is quite extensive (ca 400 videos), but the number of WAI scores that are available is much lower. For the patient, therapist, and observer, between 50-100 videos with ratings were available.

The quality of the dataset also inhibits the availability of a good-sized train and test set for the models. Very few (ca 100) videos had video quality that was good enough to extract Facial Action Units from, often this could not be done due to the lack of faces present in the video or because of lighting conditions (the person was sitting in front of a window with strong sunlight).

These imperfections of the dataset caused the final usable number of sessions to be very low. The models trained on this dataset were either overfitting or did not show any good fit to the data due to the regularization penalties we had to apply to prevent overfitting.

Another limitation of this research is that many of the features depend on each other, for instance, the textual affect analysis depends on the transcription. The transcription and diarization are the most fundamental parts of this thesis as they provide most of the features (the turn-taking and turn-level features), and the text affects analysis and emotion extraction are also dependent on the accuracy of the transcription. Moreover, the diarization is important for differentiating between the speakers, which is used to distinguish the textual affect displayed by the therapist or patient. Furthermore, the head movement is calculated based on the Facial Action Units and how they move in space over time. Therefore, if there is an error in extracting the Facial Action Units, both the facial expression recognition and the head movement features will have the same systematic error. The large number of features and modalities requires much scrutiny and close attention to prevent any such systematic errors from occurring.

The availability of Dutch-trained models is another limitation that might have had an impact on the results. The audio features were extracted with a Wav2Vec2 model that was fine-tuned on an English corpus consisting of speech segments with emotion labels. This model was used to extract the arousal and valence from the audio of the sessions, but since the dataset was Dutch there was a language mismatch. We tried to find an audio dataset of Dutch recordings but there was none available with affect labels. Therefore, the model was unfortunately not optimized for our dataset. However, multiple studies have shown that cross-language acoustic emotion recognition is possible and if the languages are related can still perform well (45; 102).

The same problem existed for the textual affect model. This was a RoBERTa-based model that was fine-tuned for extracting emotions from the text. While the RoBERTa model is specifically fine-tuned on Dutch text, this is only a general language model and not fine-tuned for emotion extraction. We used a RoBERTa-based model that was fine-tuned on multiple emotion-annotated textual datasets in the English language and applied that to our transcriptions. Since the emotion labels from the model are linked to English words, we translated the transcriptions into English before running the emotion extraction model on it. We manually inspected the translations and these captured the same emotions as the Dutch transcriptions, so we don't expect a large negative impact of the cross-lingual system. However, it is a limitation important to mention as ideally the RoBERTa model should be trained on a Dutch emotion annotated dataset. We found one dataset with such annotations, but it was not publicly available (31). Perhaps future research could create a publicly available similar dataset, but for this research, the suitability of the textual affect analysis model was limited.

The vocal analysis of this dataset has some challenges. The audio was recorded with a single microphone meaning that the two speakers (the patient and the therapist) are not separated. As studies have shown that the patient and therapist's vocal indicators have distinct results, it is important to separate the spoken audio of both people (136).

Another important limitation is that all feature extraction methods were done with off-the-shelf tools and were not extensively optimized by fine-tuning on this dataset. This project has many different elements that all require some literary research to understand its role in working alliance, a different model to extract the features, and an analysis of the specific feature. Ideally, much more time would have had to go to each element of the research and especially explore the possibilities of the dataset and feature recombination. To give an example, after all, features were extracted, we looked at the change of each feature within a session and were able to extract secondary features, the slope of the change of each feature within the time span of a session. This proved to be quite meaningful as it better represented such elements as the changing dynamics of arousal or the length of each speaker's turn throughout a session. There are likely many more secondary features that can prove meaningful in representing the data, but this requires much time to study the data manually. Furthermore, the original plan was to also perform a LIWC analysis to study the content of the language used in this dataset. Unfortunately, we did not have time to conduct this analysis and include it in this thesis.

Also, computational resources were limited during the time of this research despite that they were required to run the largest and most accurate models. To give an example, the WhisperX model had different pre-trained versions of which we used large-v1. The large-v2 model was more accurate but required a GPU RAM of at least 12GB which was not available at the time. Therefore, we had to select less accurate models which probably influenced the accuracy of the results. These two limitations naturally bring us to potential future work where we can sketch a design of this research without limitations.

5.8 Future work

5.8.1 Specific recommendation for further investigation into this research

The findings of this study, while not showing overwhelming and clear outcomes of the regression model performances, are very promising. The individual feature analyses reveal several key features that are correlated with the WAI scores, underscoring their capabilities of predicting the working alliance perception. Furthermore, the SVM model shows a good prediction capability indicated by its high accuracy. These positive findings assure us that we are on the right track and leave us to believe that with a few alterations to the study design and a larger dataset, a good working regression model can be created.

In the previous section, we mentioned the importance of extracting each modality with as much accuracy as possible, especially the transcription and diarization, as many features are dependent on each other. We also mentioned how this was difficult with the limited resources of time and computational power. We propose that this dataset be studied again more extensively where more time can be taken to optimize the models used to extract each modality's features. Also, with a strong enough computational set-up, no compromises on the sizes of the pre-trained models have to be made which, for instance, allows the use of the large-v2

WhisperX model with more accurate transcriptions and diarization. Moreover, as we noticed even during the relatively short period of this thesis project, the fields of speech-to-text systems and affect analysis are progressing incredibly fast. We had to re-run the Whisper model multiple times as there became updates available that improved the accuracy which we wanted to use to stay as much up-to-date on the latest developments as possible. Therefore, we expect that during the coming years, more and better pre-trained models in the field of affect analysis, diarization, and facial recognition will become available. With better feature extraction models that can be fine-tuned on similar datasets, a replication of this research will likely be very fruitful.

To specify the above-mentioned model improvement further, the audio Wav2Vec model was trained on an English affect and emotion corpus. As mentioned in the limitations, there was no dataset containing annotated audio samples available in the Dutch language to fine-tune or re-train this model on. If such a dataset can be created in future research, the audio affect model could be fine-tuned to predict Dutch audio. The textual affect model encountered the same problem and as we described in the limitations, creating a dataset with annotations and using that to fine-tune the Dutch-language-based RoBERTa model would improve the design of the study and likely the results as well.

Another adaptation to the study's design is the availability of more WAI scores. While it is impossible to add more patient or therapist-generated WAI scores after the therapeutic process has been finished, it is possible to let more observers rate the sessions and extend the observer-based WAI score dataset. However, because we have seen that there is a difference in the feature selection between patient and observer, this difference would have to be taken into account. If such differences are mapped out more clearly, the observer ratings could be used to predict the patient's perception of the working alliance as well. With more WAI ratings, more sessions can be used resulting in a larger dataset with (potentially) more predictive power.

An effective adaptation in the research design on the level of dataset acquisition would be the recording of the sessions. Needless to say, videos with a better frontal view of the faces without direct outside daylight in the frame would improve the quality and especially quantity of the facial features enormously. Not only the video quality is important, but the audio quality is important as well. For any research that requires speech-to-text, it would be very efficient to record the audio using one microphone per speaker.

In this research, a lot of time and energy went into finding an accurate diarization system that could be merged with an accurate transcription system. Although the PyAnnote system performs relatively well (in cases with two clear voices), the diarization error can be avoided altogether by using two microphones (one per speaker) to record audio. However, it remains uncertain whether this will impact the natural setting of the therapy as it may influence the patient's sense of comfort by having to notice being recorded so explicitly. Perhaps two microphones close to the speakers, but not necessarily pinned onto the speaker would be the best solution. This might pick up audio from the other speaker as well but would still be easy to separate due to the difference in volume. If multiple microphones are not possible, a more accurate diarization system can also lower diarization error.

As mentioned before, during this research the development of systems that could diarize and transcribe audio simultaneously was still very much in development and we therefore expect that within the next year, the toolbox WhisperX will have developed its tools much more accurately. It is therefore recommended to rerun the data when such a tool becomes available to see how much the transcription and diarization accuracy can be improved and what effect that will have on the WAI prediction results.

As mentioned in Section 5.7, we did not have time to perform a LIWC analysis and study the content of the language used in the sessions. In Section 2.3.3 we have seen that many studies have found that the content of language and especially language entrainment are related to the working alliance and can thus be good predictors. Therefore, a future study on this dataset should include a LIWC analysis and subsequent language entrainment extraction.

The final recommendation for conducting similar research as this one is to focus more on longitudinal analysis. The original aim was to incorporate the longitudinal aspect as well since the dataset contains multiple sessions per patient, but this requires more WAI scores to be able to use multiple sessions. A longitudinal analysis allows for the investigation of how the perception of the working alliance changes over the course of a psychotherapeutic treatment process. In these results, we have seen that the change of specific features within a session is informative such that it allows us to assume that investigating the change of features over multiple sessions might be informative of the change in working alliance perception as well. We have tried to study this in our dataset by looking at the change in features for the patients where the WAI changed a lot over the sessions but found that the features that showed a lot of change differed per patient. There could be an interaction effect of specific features meaning that the trend in high WAI change over the treatment course can not be detected by manually inspecting the dataset. We only found four patients (in the patient WAI score dataset) that showed a high change in WAI score (at least 5 points change) so the effect of a longitudinal analysis of the features could not be found in so few patients which is why it was not included in the results. However, with a larger dataset, a longitudinal analysis can reveal new patterns in the features and the working alliance perception that can improve the WAI prediction and its subsequential understanding of the working alliance perception.

5.8.2 General application

The concept of working alliance and its predictors remains an interesting topic with high importance and should be studied much more in the future. The progress of automatic video processing tools allows for a deeper understanding of the factors that contribute to the perception of the working alliance and its differences between therapist and patient.

Much research into the automatic analysis of working alliances will be beneficial in gaining a better understanding, but here we highlight several specific approaches that we believe will be especially useful.

First, in this research, we used multiple pre-trained models in a transfer learning capacity. Specific research into the feasibility of using pre-trained models for this purpose can help detect discrepancies in their application, so the difference between their desired performance and their actual performance can be analyzed and their inaccuracies better understood. An example is the BERT-based affect analysis models that were applied to the transcriptions to extract affect, sentiment, and emotions. The BERT models were trained on multilingual datasets but not fine-tuned on a Dutch affectual dataset. Therefore, we had to translate the text into English to extract emotional features, while it would have been much better to use a Dutch dataset to fine-tune the model for the extraction of emotional features from Dutch text. Unfortunately, we could not find a suitable Dutch emotion and affect dataset to perform this extra fine-tuning, but future research should attempt to create such a dataset, perform the fine-tuning on the Dutch dataset, and re-extract the text-based features. A deeper exploration of the models in not just the textual, but any domain has the potential to save computational resources and improve their results.

Second, the trained models should be validated on different therapy datasets or across other medical conversational settings between therapist and patient. While we performed cross-validation on unseen sessions from this dataset to evaluate the models' performances, it would be interesting to research the generalizability of the model in different settings as well.

Third, where AI models come into play, so come interpretability and ethical concerns, especially in psychotherapeutic settings. Ethical concerns had to be taken into consideration for this research as well, as privacy preservation allowed us to only use offline models and secure servers to run them on. Interpretability is very important as well, as the main goal of this research is to better understand what features underlie the perception of the working alliance. Therefore, for each model that can predict the WAI based on the extracted features, it must be interpretable how the predictions are made. This information includes which features are used for the prediction, their importance (weight), and what interactions between features occur. We did not specifically look into this in our research as our main concern was creating a model that could predict the WAI scores well. It is, however, worth exploring the interpretability of different models as well as finding methods to explain the predictions of more intricate models such as neural networks.

Further, future research might include a comparative study between the AI-based working alliance predictions, the self-report WAI measurements, and other measures such as the therapeutic outcome that represents the working alliance. Since the WAI is used as a golden standard to train the prediction models against, it is an indirect measure of the working alliance. Due to the lack of a perfect working alliance measure, the WAI is the best choice for this. However, it is possible that the features that the AI models extracted can predict the therapeutic outcomes better than the WAI can. In this case, the features are even more valuable as a model can be trained to predict therapeutic outcomes based on the automatic features without needing the intermediate WAI scores. Such a trained model might be able to provide real-time feedback to the therapist based on the detected features during the session. This could help therapists adapt their approach to improve the therapeutic outcome.

Finally, a study looking at multilingual differences will be interesting as well. Detecting the potential differences in important features for working alliance perception between different languages and cultures can help create generalizable models capable of modeling language-specific patterns and predicting working alliances across different languages.

6. Conclusion and final remarks

To conclude, in this thesis project the relationship between various indicators across multiple modalities and the perception of the working alliance has been studied. Several features have been identified as strong predictors of the WAI and we designed an SVM model able to predict whether a therapeutic session had a low or high WAI score based on the automatically extracted features. This research was a start to explore the intricate process that underlies a patient's perception of the working alliance and provided a better understanding so future research can analyze each aspect of predicting working alliance with automatic tools in depth.

Unfortunately, we were unable to answer the research question of *"To what extent can a machine learning model predict the quality of the working alliance between therapists and patients using visual and textual features extracted from psychotherapy sessions?"* with much certainty due to the limited data availability. However, we have found various patterns between the features and the working alliance score which suggests that an accurate working alliance prediction model is possible. Therefore, our findings enable a lot of future research to take on the subject of predicting working alliances with precision and good indications.

With the predictive features found in this research, we hope to give therapists some more clarity in how their patients judge the working alliance and what indicators they can pay specific attention to. This will hopefully reduce drop-outs, improve therapeutic outcomes, and relieve the pressure on the delicate but important mental health sector.

7. Acknowledgements

I am very grateful for the people I have met during this project and all the interesting knowledge they have taught me.

I would especially like to thank the following people:

- Prof. Dr. Albert Ali Salah, my supervisor who, despite having many different projects going on simultaneously, made time to think along, and provide new ideas and constructive feedback that allowed me to improve during this thesis. I'm also grateful for his enthusiasm for the project and for stimulating me to achieve the best results.
- I would also like to thank my second supervisor, Dr. Itır Önal Ertugrul, for her very useful feedback on my thesis proposal and her willingness to dive deeper into the field of working alliance and read and evaluate my thesis.
- Dr. Sanne Bruijniks, who guided me in this project on psychology-oriented topics. I'm very grateful for the time she took to think along and offer her expertise on the processing of working alliance scores.
- Dr. Heysem Kaya, for the interesting meetings and conversations about AI and its applications and for his useful advice on handling speech and emotional feature processing. Unfortunately, his designed model on facial emotion expression prediction did not work well on our dataset, but his willingness to help was certainly appreciated!
- Annika Vollebregt, who offered her experience in research and taught me the best practices in data visualization in a clear and comprehensive manner.
- Olav Vollebregt, who supported and helped me through this project, and for his help in providing feedback on my thesis.

Bibliography

Abargil, M. and Tishby, O. (2022). How therapists' emotion recognition relates to therapy process and outcome. *Clinical Psychology & Psychotherapy*, 29(3):1001–1019.

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., and Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature*, 433(7021):68–72.

Albuquerque, L., Valente, A. R. S., Teixeira, A., Figueiredo, D., Sa-Couto, P., and Oliveira, C. (2021). Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan. *PloS one*, 16(4):e0248842.

Allen, G. (2011). Angry or ecstatic? context is everything in reading facial emotion, say psychologists. `https://www.dailymail.co.uk/sciencetech/article-2045209/Angry-ecstatic-Context-reading-facial-emotion-say-psychologists.html`. Daily Mail Online.

Ambady, N., Koo, J., Rosenthal, R., and Winograd, C. H. (2002). Physical therapists' nonverbal communication predicts geriatric patients' health outcomes. *Psychology and aging*, 17(3):443.

Anderson, L. A. and Dedrick, R. F. (1990). Development of the trust in physician scale: a measure to assess interpersonal trust in patient-physician relationships. *Psychological reports*, 67(3_suppl):1091–1100.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Bain, M., Huh, J., Han, T., and Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH*.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE.

Bartlett, M. S., Hager, J. C., Ekman, P., and Sejnowski, T. J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, 36(2):253–263.

Bavelas, J. B., Coates, L., and Johnson, T. (2000). Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941.

Bayerl, S. P., Roccabruna, G., Chowdhury, S. A., Ciulli, T., Danieli, M., Riedhammer, K., and Riccardi, G. (2022). What can speech and language tell us about the working alliance in psychotherapy. *arXiv preprint arXiv:2206.08835*.

Bernieri, F. J. (1988). Coordinated movement and rapport in teacher-student interactions. *Journal of Nonverbal behavior*, 12(2):120–138.

Bethune, S. (2021). Demand for mental health treatment continues to increase, say psychologists. `https://www.apa.org/news/press/releases/2021/10/mental-health-treatment-demand`.

Boiger, M., Ceulemans, E., De Leersnyder, J., Uchida, Y., Norasakkunkit, V., and Mesquita, B. (2018). Beyond essentialism: Cultural differences in emotions revisited. *Emotion*, 18(8):1142.

Bonta, V., Kumaresh, N., and Janardhan, N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, 8(S2):1–6.

Boot, P., Zijlstra, H., and Geenen, R. (2017). The dutch translation of the linguistic inquiry and word count (liwc) 2007 dictionary. *Dutch Journal of Applied Linguistics*, 6(1):65–76.

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice*, 16(3):252.

Bousmalis, K., Mehu, M., and Pantic, M. (2009). Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In *2009 3rd international conference on affective computing and intelligent interaction and workshops*, pages 1–9. IEEE.

Bredin, H. and Laurent, A. (2021). End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., and Gill, M.-P. (2020). pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Brugel, S., Postma-Nilsenová, M., and Tates, K. (2015). The link between perception of clinical empathy and nonverbal behavior: The effect of a doctor's gaze and body orientation. *Patient education and counseling*, 98(10):1260–1265.

Bruijniks, S. J., Lemmens, L. H., Hollon, S. D., Peeters, F. P., Cuijpers, P., Arntz, A., Dingemanse, P., Willems, L., Van Oppen, P., Twisk, J. W., et al. (2020). The effects of once-versus twice-weekly sessions on psychotherapy outcomes in depressed patients. *The British Journal of Psychiatry*, 216(4):222–230.

Cassell, J. (2004). Towards a model of technology and literacy development: Story listening systems. *Journal of Applied Developmental Psychology*, 25(1):75–105.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Chicco, D., Warrens, M. J., and Jurman, G. (2021). The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7:e623.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Corbella, S. and Botella, L. (2004). Psychometric properties of the spanish version of the working alliance theory of change inventory (watoci). *Psicothema*, pages 702–705.

Cowen, A. S., Keltner, D., Schroff, F., Jou, B., Adam, H., and Prasad, G. (2021). Sixteen facial expressions occur in similar contexts worldwide. *Nature*, 589(7841):251–257.

Datz, F., Wong, G., and Löffler-Stastka, H. (2019). Interpretation and working through contemptuous facial micro-expressions benefits the patient-therapist relationship. *International Journal of Environmental Research and Public Health*, 16(24):4901.

De Bruyne, L., De Clercq, O., and Hoste, V. (2021a). Emotional robbert and insensitive bertje: combining transformers and affect lexica for dutch emotion detection. In *Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), held in conjunction with EACL 2021*, pages 257–263. Association for Computational Linguistics.

De Bruyne, L., De Clercq, O., and Hoste, V. (2021b). Prospects for dutch emotion detection: Insights from the new emotionl dataset. *Computational Linguistics in the Netherlands Journal*, 11:231–255.

de Roten, Y., Darwish, J., Stern, D. J., Fivaz-Depeursinge, E., and Corboz-Warnery, A. (1999). Nonverbal communication and alliance in therapy: The body formation coding system. *Journal of clinical psychology*, 55(4):425–438.

Delobelle, P., Winters, T., and Berendt, B. (2020). Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diener, M. J., Hilsenroth, M. J., and Weinberger, J. (2007). Therapist affect focus and patient outcomes in psychodynamic psychotherapy: A meta-analysis. *American Journal of Psychiatry*, 164(6):936–941.

D'mello, S. K. and Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36.

Dowell, N. M. and Berman, J. S. (2013). Therapist nonverbal behavior and perceptions of empathy, alliance, and treatment credibility. *Journal of Psychotherapy Integration*, 23(2):158.

Doyran, M., Türkmen, B., Oktay, E. A., Halfon, S., and Salah, A. A. (2019). Video and text-based affect analysis of children in play therapy. In *2019 International Conference on Multimodal Interaction*, pages 26–34.

Duncan, B. L., Miller, S. D., Wampold, B. E., and Hubble, M. A. (2010). *The heart and soul of change: Delivering what works in therapy*. American Psychological Association.

Ekman, P. (1992). Are there basic emotions? *Psychological Review*, 99(3).

Ekman, P. and Friesen, W. V. (1976). Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1:56–75.

Ekman, P. and Friesen, W. V. (1978). Facial action coding system: a technique for the measurement of facial movement.

Feraru, S. M., Schuller, D., et al. (2015). Cross-language acoustic emotion recognition: An overview and some tendencies. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 125–131. IEEE.

Finlay, P. and Argos Translate, C. Argos Translate.

Flückiger, C., Del Re, A. C., Wampold, B. E., and Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, 55(4):316–340.

Friedland, G., Vinyals, O., Huang, Y., and Muller, C. (2009). Prosodic and other long-term features for speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5):985–993.

Fuertes, J. N. (2019). *Working alliance skills for mental health professionals*. Oxford University Press, USA.

Fuertes, J. N., Toporovsky, A., Reyes, M., and Osborne, J. B. (2017). The physician-patient working alliance: Theory, research, and future possibilities. *Patient Education and Counseling*, 100(4):610–615.

Gelso, C. J. and Kline, K. V. (2019). The sister concepts of the working alliance and the real relationship: on their development, rupture, and repair. *Research in Psychotherapy: Psychopathology, Process, and Outcome*, 22(2).

Gokaslan, A., Cohen, V., Pavlick, E., and Tellex, S. (2019). Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

Grumet, G. W. (1983). Eye contact: The core of interpersonal relatedness. *Psychiatry*, 46(2):172–180.

Hamborg, F., Meuschke, N., Breitinger, C., and Gipp, B. (2017). news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Haque, A., Guo, M., Miner, A. S., and Fei-Fei, L. (2018). Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv preprint arXiv:1811.08592*.

Hatcher, R. L. and Gillaspy, J. A. (2006). Development and validation of a revised short version of the working alliance inventory. *Psychotherapy research*, 16(1):12–25.

He, L., Lech, M., Maddage, N. C., and Allen, N. B. (2011). Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech. *Biomedical Signal Processing and Control*, 6(2):139–146.

Hirsh, J. B. and Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of research in personality*, 43(3):524–527.

Horvath, A. and Bedi, R. (2002). The alliance,[w:] psychotherapy relationships that work: Therapist contributions and responsiveness to patients,(red.) norcross jc.

Horvath, A. O. (1981). *An exploratory study of the working alliance: Its measurement and relationship to therapy outcome*. PhD thesis, University of British Columbia.

Horvath, A. O. and Greenberg, L. S. (1989). Development and validation of the working alliance inventory. *Journal of counseling psychology*, 36(2):223.

Horvath, A. O. and Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of counseling psychology*, 38(2):139.

Hove, M. J. and Risen, J. L. (2009). It's all in the timing: Interpersonal synchrony increases affiliation. *Social cognition*, 27(6):949–960.

Hsu, S. and Yu, C. K.-C. (2017). A hong kong study of working alliance inventory short form–therapist. *Asia Pacific Journal of Counselling and Psychotherapy*, 8(2):87–100.

Hsu, W.-N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., Kahn, J., Lee, A., Collobert, R., Synnaeve, G., et al. (2021). Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *arXiv preprint arXiv:2104.01027*.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Hwang, W.-C. and Wood, J. J. (2007). Being culturally sensitive is not the same as being culturally competent. *Pragmatic Case Studies in Psychotherapy*, 3(3).

Hynes, K. C. (2019). Cultural values matter: The therapeutic alliance with east asian americans. *Contemporary Family Therapy*, 41(4):392–400.

Jacques, J. and Dykeman, C. (2022). Psycholinguistic markers of therapeutic rupture types.

Kamath, R., Ghoshal, A., Eswaran, S., and Honnavalli, P. (2022a). An enhanced context-based emotion detection model using roberta. In *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–6.

Kamath, R., Ghoshal, A., Eswaran, S., and Honnavalli, P. (2022b). An enhanced context-based emotion detection model using roberta. In *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–6. IEEE.

Kawahara, T., Uesato, M., Yoshino, K., and Takanashi, K. (2015). Toward adaptive generation of backchannels for attentive listening agents. In *International workshop serien on spoken dialogue systems technology*, pages 1–10.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63.

Khoma, V., Khoma, Y., Brydinskyi, V., and Konovalov, A. (2023). Development of supervised speaker diarization system based on the pyannote audio processing library. *Sensors*, 23(4):2082.

Kim, M., Cho, Y., and Kim, S.-Y. (2022). Effects of diagnostic regions on facial emotion recognition: The moving window technique. *Frontiers in Psychology*, 13.

Kim, T. and Vossen, P. (2021). Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

Kohler, C. G., Turner, T., Stolar, N. M., Bilker, W. B., Brensinger, C. M., Gur, R. E., and Gur, R. C. (2004). Differences in facial expressions of four universal emotions. *Psychiatry research*, 128(3):235–244.

Koolagudi, S. G., Kumar, N., and Rao, K. S. (2011). Speech emotion recognition using segmental level prosodic analysis. In *2011 international conference on devices and communications (ICDeCom)*, pages 1–5. IEEE.

Kory-Westlund, J. M. and Breazeal, C. (2019). Exploring the effects of a social robot's speech entrainment and backstory on young children's emotion, rapport, relationship, and learning. *Frontiers in Robotics and AI*, 6:54.

Lakin, J. L. and Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological science*, 14(4):334–339.

Lambert, M. J. and Barley, D. E. (2001). Research summary on the therapeutic relationship and psychotherapy outcome. *Psychotherapy: Theory, research, practice, training*, 38(4):357.

Li, S. and Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215.

Li, Y., Huang, X., and Zhao, G. (2021). Micro-expression action unit detection with spatial and channel attention. *Neurocomputing*, 436:221–231.

Lin, B., Cecchi, G., and Bouneffouf, D. (2022). Working alliance transformer for psychotherapy dialogue classification. *arXiv preprint arXiv:2210.15603*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lotfian, R. and Busso, C. (2019). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Machado, P. P., Beutler, L. E., and Greenberg, L. S. (1999). Emotion recognition in psychotherapy: Impact of therapist level of experience and emotional awareness. *Journal of Clinical Psychology*, 55(1):39–57.

Martin, D. J., Garske, J. P., and Davis, M. K. (2000). Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. *Journal of consulting and clinical psychology*, 68(3):438.

Marziliano, A., Applebaum, A., Moyer, A., Pessin, H., Rosenfeld, B., and Breitbart, W. (2021). The impact of matching to psychotherapy preference on engagement in a randomized controlled trial for patients with advanced cancer. *Frontiers in Psychology*, 12:637519.

McIntyre, G., Göcke, R., Hyett, M., Green, M., and Breakspear, M. (2009). An approach for automatically measuring facial activity in depressed subjects. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–8. IEEE.

Mellouk, W. and Handouzi, W. (2020). Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science*, 175:689–694.

Meynard, A., Seneviratna, G., Doyle, E., Becker, J., Wu, H.-T., and Borg, J. S. (2021). Predicting trust using automated assessment of multivariate interactional synchrony. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE.

Miles, L. K., Nind, L. K., and Macrae, C. N. (2009). The rhythm of rapport: Interpersonal synchrony and social perception. *Journal of experimental social psychology*, 45(3):585–589.

Monaro, M., Maldera, S., Scarpazza, C., Sartori, G., and Navarin, N. (2022). Detecting deception through facial expressions in a dataset of videotaped interviews: A comparison between human judges and machine learning models. *Computers in Human Behavior*, 127:107063.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A.-L., De Schryver, M., De Winne, J., and Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 dutch words. *Behavior research methods*, 45:169–177.

Morris, A. C., Maier, V., and Green, P. (2004). From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

Motalebi, N., Cho, E., Sundar, S. S., and Abdullah, S. (2019). Can alexa be your therapist? how back-channeling transforms smart-speakers to be active listeners. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pages 309–313.

Müller, P., Dietz, M., Schiller, D., Thomas, D., Lindsay, H., Gebhard, P., André, E., and Bulling, A. (2022). Multimediate'22: Backchannel detection and agreement estimation in group interactions. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7109–7114.

Namba, S., Sato, W., Osumi, M., and Shimokawa, K. (2021). Assessing automated facial action unit detection systems for analyzing cross-domain facial expression databases. *Sensors*, 21(12):4222.

Negri, A., Christian, C., Mariani, R., Belotti, L., Andreoli, G., and Danskin, K. (2019). Linguistic features of the therapeutic alliance in the first session: a psychotherapy process study. *Research in Psychotherapy: Psychopathology, Process, and Outcome*, 22(1).

Neumann, M. et al. (2018). Cross-lingual and multilingual speech emotion recognition on english and french. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5769–5773. IEEE.

Newhill, C. E., Safran, J. D., and Muran, J. C. (2003). *Negotiating the therapeutic alliance: A relational treatment guide.* Guilford Press.

Norcross, J. C. (2002). *Psychotherapy relationships that work: Therapist contributions and responsiveness to patients.* Oxford University Press.

Norcross, J. C. and Goldfried, M. R. (2005). *Handbook of psychotherapy integration*. Oxford University Press.

Norcross, J. C. and Wampold, B. E. (2011). Evidence-based therapy relationships: research conclusions and clinical practices. *Psychotherapy*, 48(1):98.

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., and Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Peluso, P. R. and Freund, R. R. (2018). Therapist and client emotional expression and psychotherapy outcomes: A meta-analysis. *Psychotherapy*, 55(4):461.

Peng, X., Gu, Y., and Zhang, P. (2022). Au-guided unsupervised domain-adaptive facial expression recognition. *Applied Sciences*, 12(9):4366.

Pennebaker, J. W. (2011). Your use of pronouns reveals your personality. *Harvard Business Review*, 89(12):32–33.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count [computer software]. *Mahway: Lawrence Erlbaum Associates*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Ramseyer, F. and Tschacher, W. (2011). Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *Journal of consulting and clinical psychology*, 79(3):284.

Ramseyer, F. and Tschacher, W. (2014). Nonverbal synchrony of head-and body-movement in psychotherapy: different signals have different associations with outcome. *Frontiers in psychology*, 5:979.

Ravanelli, M. and Bengio, Y. (2018). Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop (SLT)*, pages 1021–1028. IEEE.

Rennung, M. and Göritz, A. S. (2016). Prosocial consequences of interpersonal synchrony: a meta-analysis. *Zeitschrift für Psychologie*, 224(3):168.

Rodríguez, L., García-Varea, I., and Vidal, E. (2010). Multi-modal computer assisted speech transcription. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 1–7.

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of consulting psychology*, 21(2):95.

Rogers, D. T. (2015). Further validation of the learning alliance inventory: The roles of working alliance, rapport, and immediacy in student learning. *Teaching of Psychology*, 42(1):19–25.

Rubel, J. A., Zilcha-Mano, S., Giesemann, J., Prinz, J., and Lutz, W. (2020). Predicting personalized process-outcome associations in psychotherapy using machine learning approaches—a demonstration. *Psychotherapy Research*, 30(3):300–309.

Ryant, N., Church, K., Cieri, C., Du, J., Ganapathy, S., and Liberman, M. (2020). Third dihard challenge evaluation plan. *arXiv preprint arXiv:2006.05815*.

Ryu, J., Heisig, S., McLaughlin, C., Katz, M., Mayberg, H., and Gu, X. A natural language processing approach reveals interpretable linguistic features of therapeutic alliance in psychotherapy. *Available at SSRN 4276250*.

Schirmer, A. and Adolphs, R. (2017). Emotion perception from face, voice, and touch: comparisons and convergence. *Trends in cognitive sciences*, 21(3):216–228.

Seabold, S. and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Siam, A. I., Soliman, N. F., Algarni, A. D., El-Samie, A., Fathi, E., and Sedik, A. (2022). Deploying machine learning techniques for human emotion detection. *Computational Intelligence and Neuroscience*, 2022.

Skiendziel, T., Rösch, A. G., and Schultheiss, O. C. (2019). Assessing the convergent validity between the automated emotion recognition software noldus facereader 7 and facial action coding system scoring. *PloS one*, 14(10).

Stefens, M., Rondeel, E., Templin, J., Brode, D., de Waart, E., de Jong, R., ten Hoeve-Rozema, J., Waringa, A., Reijnders, J., Jacobs, N., et al. (2022). Longitudinal measurement invariance of the working alliance inventory-short form across coaching sessions. *BMC psychology*, 10(1):277.

Sturgiss, E. A., Sargent, G., Haesler, E., Rieger, E., and Douglas, K. (2016). Therapeutic alliance and obesity management in primary care–a cross-sectional pilot using the working alliance inventory. *Clinical obesity*, 6(6):376–379.

Suárez, P. J. O., Sagot, B., and Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Swift, J. K., Callahan, J. L., and Vollmer, B. M. (2011). Preferences. *Journal of clinical psychology*, 67(2):155–165.

Tal, S., Bar-Kalifa, E., Kleinbub, J. R., Leibovich, L., Deres-Cohen, K., and Zilcha-Mano, S. (2022). A multimodal case study utilizing physiological synchrony as indicator of context in which motion synchrony is associated with the working alliance. *Psychotherapy*.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Trinh, T. H. and Le, Q. V. (2018). A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Vail, A., Girard, J., Bylsma, L., Cohn, J., Fournier, J., Swartz, H., and Morency, L.-P. (2022). Toward causal understanding of therapist-client relationships: A study of language modality and social entrainment. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, pages 487–494.

Vail, A. K., Girard, J., Bylsma, L., Cohn, J., Fournier, J., Swartz, H., and Morency, L.-P. (2021). Goals, tasks, and bonds: Toward the computational assessment of therapist versus client perception of working alliance. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8.

Valstar, M. F. and Pantic, M. (2006). Biologically vs. logic inspired encoding of facial actions and emotions in video. In *2006 IEEE International Conference on Multimedia and Expo*, pages 325–328. IEEE.

van der Burgh, B. and Verberne, S. (2019). The merits of universal language model fine-tuning for small datasets - a case with dutch book reviews. *CoRR*, abs/1910.00896.

Vasquez, M. J. (2007). Cultural difference and the therapeutic alliance: an evidence-based analysis. *American Psychologist*, 62(8):878.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., and Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., and Ahn, H.-n. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empiricially," all must have prizes.". *Psychological bulletin*, 122(3):203.

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, A., Ram, Y., Yarkoni, T., Williams, M. L., Evans, C., Fitzgerald, C., Brian, Fonnesbeck, C., Lee, A., and Qalieh, A. (2017). mwaskom/seaborn: v0.8.1 (september 2017).

Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., and Kissler, J. (2017). Mapping the emotional face. how individual face parts contribute to successful emotion recognition. *PloS one*, 12(5).

Wolf, K. (2022). Measuring facial expression of emotion. *Dialogues in clinical neuroscience*.

Yao, A., Shao, J., Ma, N., and Chen, Y. (2015). Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *Proceedings of the 2015 acm on international conference on multimodal interaction*, pages 451–458.

Yao, L., Wan, Y., Ni, H., and Xu, B. (2021). Action unit classification for facial expression recognition using active learning and svm. *Multimedia Tools and Applications*, 80(16):24287–24301.

Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting, 1970*, pages 567–578.

Zhou, Y., Chen, X.-y., Liu, D., Pan, Y.-l., Hou, Y.-f., Gao, T.-t., Peng, F., Wang, X.-c., and Zhang, X.-y. (2022). Predicting first session working alliances using deep learning algorithms: A proof-of-concept study for personalized psychotherapy. *Psychotherapy Research*, pages 1–10.

8. Appendix A: Additional Data

8.1 Transcription comparison results

Table 13 shows the transcription scores for all tested systems. The evaluation measures were the Word Error Rate (WER), Match Error Rate (MER), and Word Insertion Likelihood (WIL). The best-performing model is highlighted.

Table 13. WER, MER, and WIL scores for different systems on the test video

Model	WER	MER	WIL
whisper large-v2	0.01	0.01	0.01
whisper large-v1	0.06	0.06	0.09
whisper medium	0.07	0.07	0.12
word online	0.23	0.22	0.36
google	0.39	0.38	0.55

8.2 Diarization comparison results

Table 14 shows the diarization scores for all tested systems. The evaluation measures were the Diarization Error Rate (DER) and the Jaccard Error Rate (JER). The best-performing model is highlighted.

Table 14. Diarization performance comparison for different systems

Model	DER (%)	JER (%)
PyAnnote	1.7	0.02
GMM	14.1	0.23
Word Online	47.3	0.63
Google Speech-to-Text (German)	25.4	0.34
Google Speech-to-Text (English)	41.6	0.48

8.3 Total list of features

Figure 25 shows all features per modality as they were extracted from the dataset. As an additional secondary feature, the average change during a session was extracted for the features from all modalities.

(a) The conversational, speech, and textual affect features sorted per modality that were extracted from the dataset and used for WAI prediction.

Conversational Features	**Speech**	**Textual Affect**
Number of turns	Max/min/avg arousal	Max/min/avg arousal patient
Turns therapist/patient	Max/min/avg valence	Max/min/avg valence patient
Avg turn length therapist/patient		Max/min/avg arousal therapist
Wordcount therapist/patient		Max/min/avg valence therapist
Speech rate therapist/patient		
Speech duration therapist/patient		
Duration percentage therapist/patient		
Participation equality		
Overlapping turns		
Turn-level freedom		

(b) The textual emotions, facial emotions, and head movement features sorted per modality that were extracted from the dataset and used for WAI prediction.

Textual Emotions		**Facial Emotions**	**Head Movement**
Admiration	Amusement	Neutral	Shake/nod patient listening
Anger	Disgust	Anger	Shake/nod patient speaking
Annoyance	Approval	Disgust	Shake/nod therapist listening
Caring	Confusion	Fear	Shake/nod therapist speaking
Curiosity	Desire	Happiness	
Disappointment	Excitement	Sadness	
Embarrassment	Fear	Surprise	
Gratitude	Grief	Contempt	
Joy	Love	Pain	
Nervousness	Optimism	Unknown	
Pride	Realization		
Relief	Remorse		
Sadness	Surprise		
Neutral			

Figure 25. The features sorted per modality extracted from the dataset and used for WAI prediction.

8.3.1 Significant features from the Spearman correlation

Table 15 shows an overview of the features per modality for each rater (patient, therapist, and observer) with a significant Spearman correlation result ($p<0.05$). The features displayed in black color have a positive correlation with the WAI score (meaning that the higher the value of the feature is, the higher the perception of the working alliance is), whereas the features in red color have a negative correlation with the WAI score (meaning that the higher the value of the feature is, the lower the perception of the working alliance is).

Table 15. Overview of all significant features according to the Spearman correlation test sorted per modality and WAI category. The features in red display negative correlations with the WAI, whereas the features written in black display positive correlations with the WAI.

	Conversational	**Speech**	**Text**	**Facial**	**Movement**
Patient	Speech duration patient Turn length patient Participation equality Turn length therapist Speech duration therapist Speech rate patient Turn-level-freedom	Max arousal Max valence Min valence	Amusement Fear Min arousal patient	Anger Happiness	Listening nods patient Listening nods therapist Listening shakes therapist
Therapist	Number of turns Overlapping segments Speech rate patient Turn-level freedom Turn length therapist Turn length patient Duration % therapist	Max arousal Max valence Min valence	Disgust Positive sentiment	Anger Sadness Neutral	
Observer	Number of turns Turn-level freedom Turn length patient Turn length therapist Overlapping segments	Max arousal Max valence Min valence	Approval Disapproval Remorse Confusion Excitement Disgust	Anger	Listening nods patient Listening shakes patient

8.4 Model Performance Metrics

The three Tables in this Section show a more detailed evaluation of the regression machine learning models used for the combined feature analysis. The models Multilinear regression, XGB, kNN, SVR, Elastic Net, and RF were trained on the extracted features from different modalities, and their performance on how well they could predict the WAI scores was tested. The evaluation metrics are CV, RMSE, and R-squared scores, and a baseline MSE score is given as a reference for the baseline prediction performance. For an interpretation of these results, the reader is referred to Section 5.2, where the models are evaluated. The performance of the models on the patient WAI scores can be found in Section 8.4.1, the performance of the models on the therapist WAI scores can be found in Section 8.4.2 and the performance of the models on the observer WAI scores can be found in Section 8.4.3.

8.4.1 Patient scores

Table 16 shows the evaluation scores of the tested machine learning regression models on the patient WAI scores.

Table 16. Model Performance Metrics Patient

Model	CV	RMSE	R-squared	Baseline MSE
Multilinear	2.05	11.91	-3.21	34.88
XGB	1.23	7.16	-0.34	34.88
kNN	1.09	6.32	-0.04	34.88
SVR	1.00	5.82	0.11	34.88
Elastic Net	1.00	5.78	0.12	34.88
RF	1.03	6.00	0.47	34.88

8.4.2 Therapist scores

Table 17 shows the evaluation scores of the tested machine learning regression models on the therapist WAI scores.

Table 17. Model Performance Metrics Therapist

Model	CV	RMSE	R-squared	Baseline MSE
Multilinear	2.01	7.34	-3.06	32.43
XGB	1.12	4.08	-3.06	32.43
kNN	1.16	4.21	-0.32	32.43
SVR	1.04	3.81	-0.07	32.43
Elastic Net	1.06	3.87	-0.11	32.43
RF	1.09	3.97	0.42	32.43

8.4.3 Observer scores

Table 18 shows the evaluation scores of the tested machine learning regression models on the observer WAI scores.

Table 18. Model Performance Metrics Observer

Model	CV	RMSE	R-squared	Baseline MSE
Multilinear	1.35	12.84	-0.82	56.09
XGB	1.35	11.29	-0.18	56.09
kNN	1.11	10.6	-0.63	56.09
SVR	1.05	10.0	0.07	56.09
Elastic Net	1.09	10.40	-0.01	56.09
RF	1.13	10.76	0.39	56.09

8.5 MRMR feature selection

8.5.1 Patient MRMR results

The MRMR feature selection showed the importance of each feature and how much it contributed to the final WAI score prediction. Figure 26 shows the different modalities and the features that contributed to the prediction positively. As can be seen in this Figure, all modalities contributed to the prediction, but on average the conversational features show a higher importance.

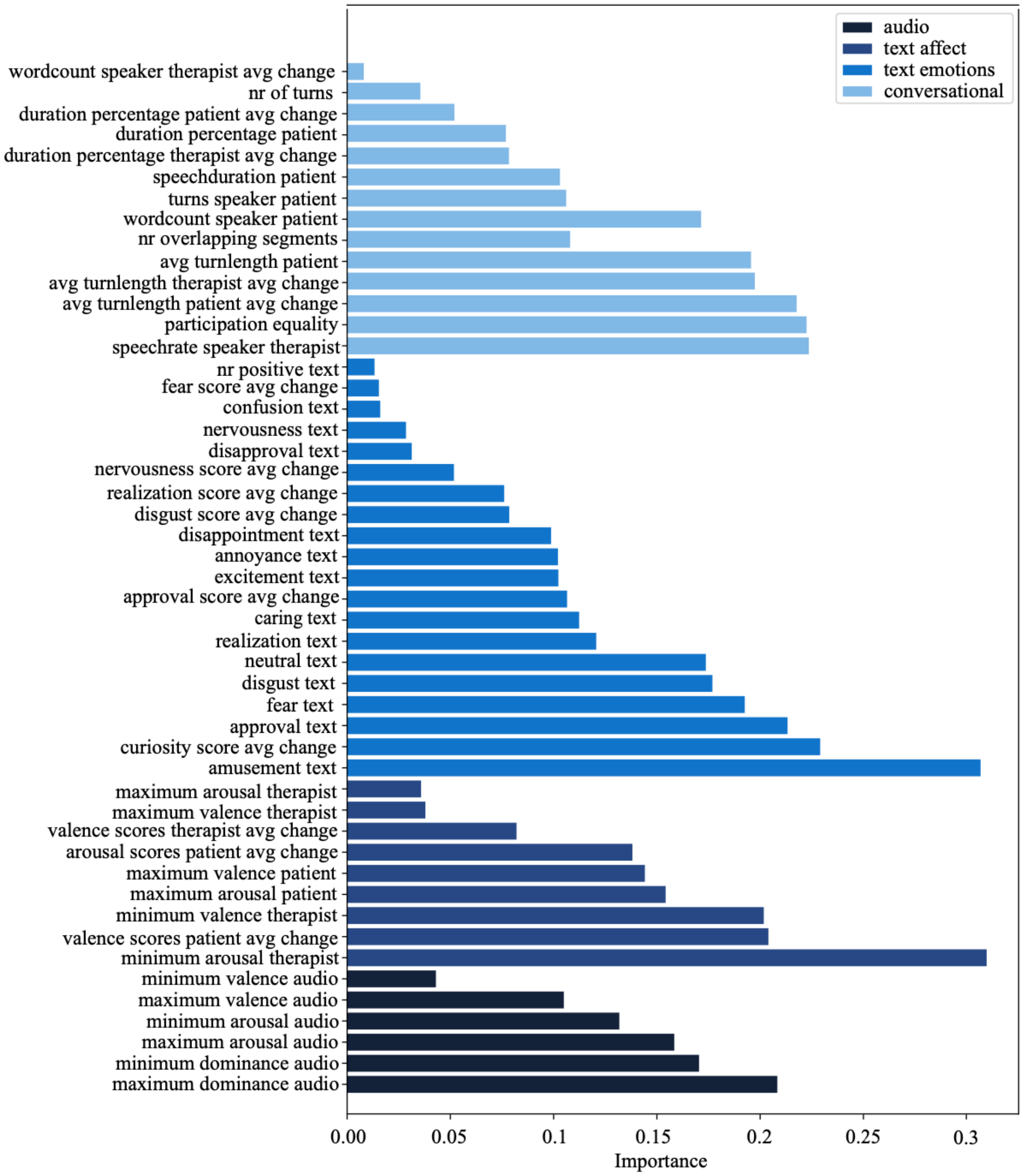


Figure 26. Comparison of models (patient)

8.5.2 Therapist MRMR results

For the therapist, the conversational features are less important than for the patient. This can be observed in Figure 27, where there are fewer conversational features with a high importance. Also, text emotion features have a higher importance than for the patient, which shows in more emotions in the text with a moderately high importance level.

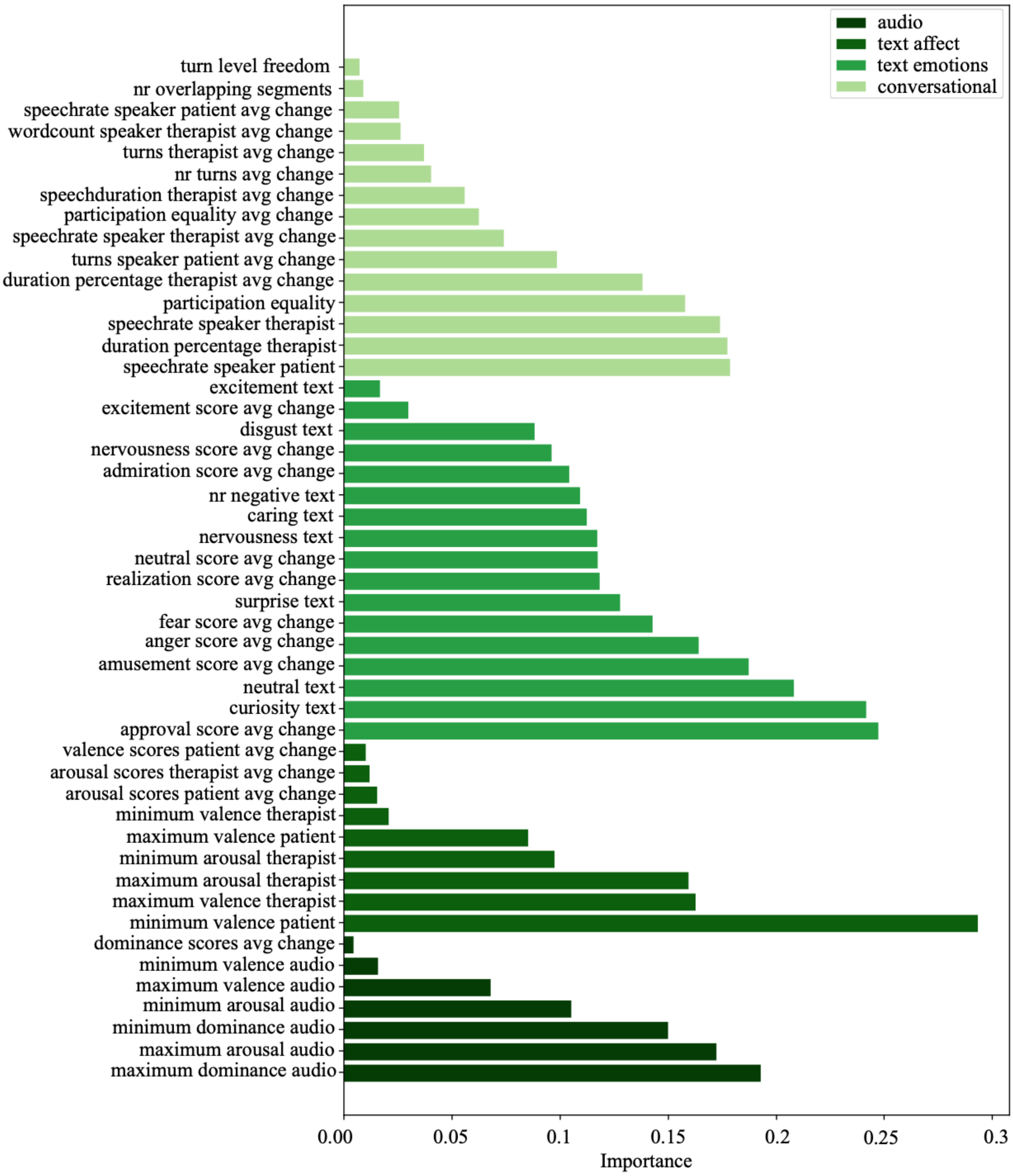


Figure 27. Comparison of models (patient)

8.5.3 Observer MRMR results

The importance of the features for predicting the observer WAI score according to the MRMR selection is given in Figure 28. As can be seen, there are two features that show a very high importance, the turn-level-freedom and the average change of the arousal scores of the therapist.

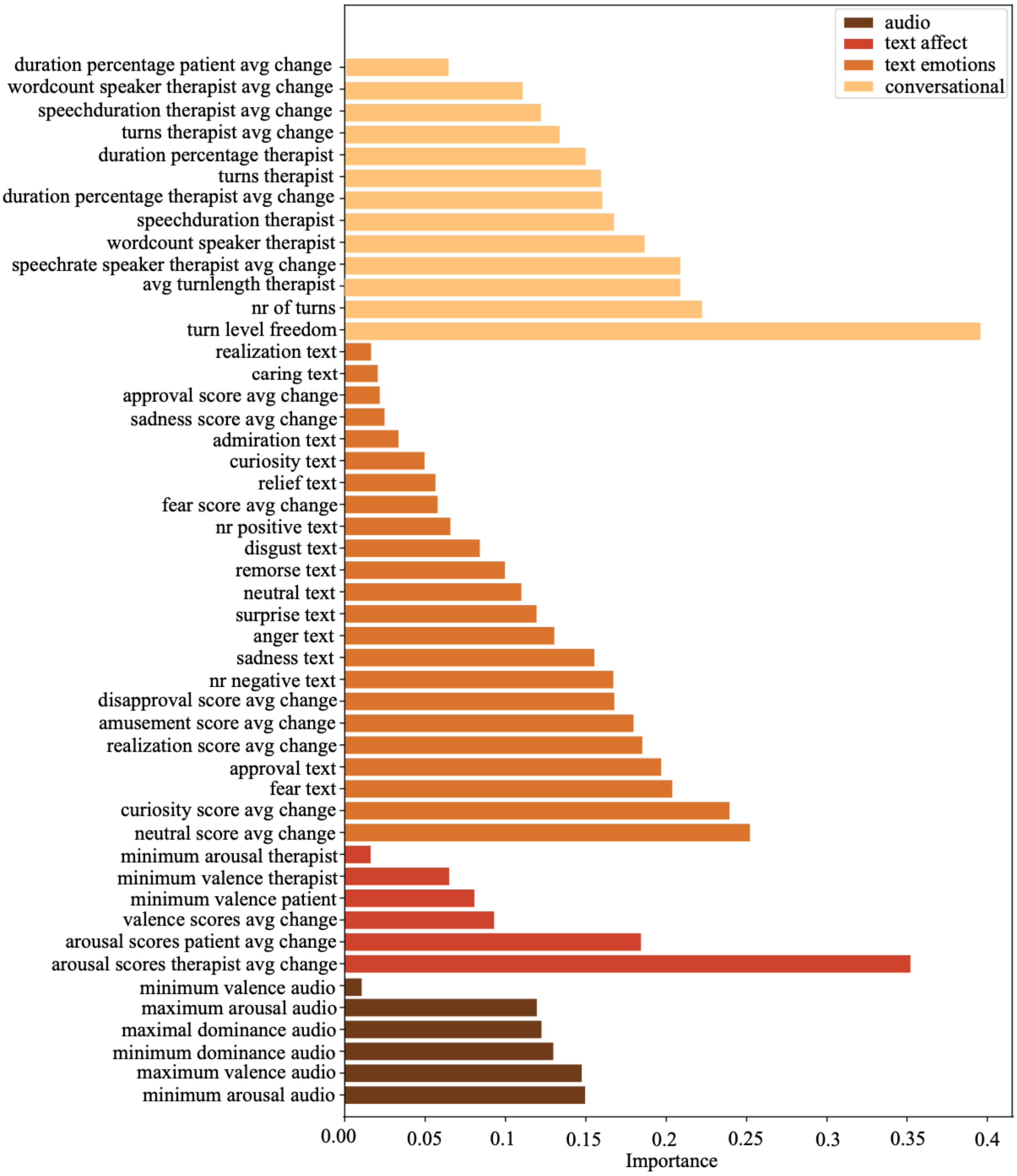


Figure 28. Comparison of models (patient)

9. Appendix B: WAI-S and WAI-SRT forms

In this Chapter are the two WAI forms added as they were used for gathering the Working Alliance Inventory score in the dataset that was used in this research. The WAI-S is the patient's version, but the observer's WAI-S form consisted of the same questions only phrased differently (a third-person perspective instead of a first-person perspective). The WAI-SRT is the questionnaire that the therapists filled out in sessions 4, 8, 12, 16, and 20.

9.1 WAI-S

1. Een resultaat van deze sessies is dat het voor mij duidelijker is hoe ik zou kunnen veranderen.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

2. Wat ik doe in therapie, geeft mij een nieuwe kijk op mijn probleem.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

3. Ik geloof dat mijn therapeut(e) mij aardig vindt.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

4. Mijn therapeut(e) en ikzelf werken samen bij het bepalen van de doelstellingen voor mijn therapie.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

5. Mijn therapeut(e) en ik respecteren elkaar.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

6. Mijn therapeut(e) en ik werken naar de doelstellingen toe die we beiden goedkeurden.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

7. Ik voel dat mijn therapeut(e) mij apprecieert.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

8. Wij zijn het eens over wat voor mij belangrijk is om aan te werken

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

9. Ik voel dat mijn therapeut(e) om mij geeft, zelfs wanneer ik dingen doe die hij/zij niet goedkeurt.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

10. Ik voel dat de dingen die ik in therapie doe, mij zullen helpen om de veranderingen die ik wil, te bereiken.

    ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

11. We hebben ons een goed begrip gevormd van het soort veranderingen die goed zouden zijn voor mij.

    ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

12. Ik geloof dat de manier waarop we aan mijn probleem werken, de juiste is.

    ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

9.2 WAI-SRT

Working Alliance Inventory – Short Revised - Therapist (WAI-SRT)

Instructions: Below is a list of statements about experiences people might have with their client. Some items refer directly to your client with an underlined space – as you read the sentences, mentally insert the name of your client in place of ___ in the text.

IMPORTANT!!! Please take your time to consider each question carefully.

1. ___ and I agree about the steps to be taken to improve his/her situation.

①	②	③	④	⑤
Seldom	Sometimes	Fairly Often	Very Often	Always

2. I am genuinely concerned for ___'s welfare.

⑤	④	③	②	①
Always	Very Often	Fairly Often	Sometimes	Seldom

3. We are working towards mutually agreed upon goals.

①	②	③	④	⑤
Seldom	Sometimes	Fairly Often	Very Often	Always

4. ___ and I both feel confident about the usefulness of our current activity in therapy.

①	②	③	④	⑤
Seldom	Sometimes	Fairly Often	Very Often	Always

5. I appreciate ___ as a person.

⑤	④	③	②	①
Always	Very Often	Fairly Often	Sometimes	Seldom

6. We have established a good understanding of the kind of changes that would be good for ___.

⑤	④	③	②	①
Always	Very Often	Fairly Often	Sometimes	Seldom

7. ___ and I respect each other.

①	②	③	④	⑤
Seldom	Sometimes	Fairly Often	Very Often	Always

8. ___ and I have a common perception of his/her goals.

⑤	④	③	②	①
Always	Very Often	Fairly Often	Sometimes	Seldom

9. I respect ___ even when he/she does things that I do not approve of.

①	②	③	④	⑤
Seldom	Sometimes	Fairly Often	Very Often	Always

10. We agree on what is important for ___ to work on.

⑤	④	③	②	①
Always	Very Often	Fairly Often	Sometimes	Seldom