PREDICTING INVESTMENT DECISIONS WITH A MULTIMODAL ANALYSIS OF ACOUSTIC AND BODY EXPRESSION FEATURES

GUUS VAN DONGEN

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

u189576

COMMITTEE

dr. M. Junge
dr. G. Saygili

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

December 4th, 2023

WORD COUNT

8067

ACKNOWLEDGMENTS

First and foremost, I wish to express my sincere appreciation to my supervisor, Dr. M. Junge, whose unwavering guidance and support proved indispensable throughout the entirety of the project. The weekly sessions played a pivotal role in refining the quality of the research, characterized by her constructive feedback and insightful suggestions. Furthermore, I extend my gratitude to my fellow students, with whom I had the privilege of engaging in meaningful discussions, sharing insights, and collectively evaluating the progress of the project. Their invaluable input has significantly contributed to the research. Lastly, I would like to convey my appreciation to W. Liebregts for providing and maintaining the entrepreneurial pitch dataset.

PREDICTING INVESTMENT DECISIONS WITH A MULTIMODAL ANALYSIS OF ACOUSTIC AND BODY EXPRESSION FEATURES

GUUS VAN DONGEN

Abstract

This research investigates the predictive effectiveness of utilizing multiple nonverbal behavioral cues during entrepreneurial pitches to determine their success in securing funding. Employing a multimodal approach, this study integrates both the acoustic and body expression modalities to predict an investment score, while previous research has explored these modalities in isolation. The models implemented in this study adopt a deep learning framework, employing either a Gated Recurrent Unit or a Long Short-Term Memory layer to capture temporal information within a pitch. The findings present promising results for predicting the likelihood of investment using a multimodal strategy that incorporates both acoustic and body expression modalities. Specifically, a late fusion multimodal model, consisting of the best single feature representation from each modality, has proven to be the most predictive.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

1.1 *Data Source and ethics*

The dataset is gathered by W.J. Liebregts, D. Urbig, and M.M. Jung (Liebregts et al., 2018-2023). The dataset is unpublished, and ongoing efforts are in place to collect additional data. It consists of video data featuring entrepreneurial pitches delivered by university students in the Entrepreneurship course. Only pitches from students who provided explicit consent for their data to be used in research are included in the dataset; pitches from students who did not grant consent are excluded. The dataset's owners have granted consent for its utilization in research, facilitated by the signing of a non-disclosure agreement. The images used

in this research are predominantly created specifically for this study. In instances where external images are incorporated, the source is explicitly cited, and the images are governed by a non-commercial license.

1.2 *Code*

The code is primarily handwritten, sometimes based on example code that is listed within the Google Drive of the dataset, hence it has similarity with studies that use the same dataset (e.g. Goossens et al., 2022; Jung et al., 2023; Van Aken et al., 2023). The code of this project is included in the Google Drive of the dataset. Code snippets from the documentation of the following packages may have been employed:

Library	**Version**
Numpy	1.21.6
Scipy	1.7.3
Matplotlib	3.5.3
XlsxWriter	3.1.9
Openpyxl	3.1.2
Keras	2.10.0
Tensorflow-gpu	2.10.0
Scipy	1.7.3
Pandas	1.1.5
MoviePy	1.0.3
Scikit-learn	1.0.2
Shap	0.42.1
Pydub	0.25.1
openSMILE	3.0
OpenPose	1.7.0

1.3 *Technology*

ChatGPT-3.5 was utilized for generating LaTeX code for tables and figures. The assessment of spelling and grammar correctness was conducted using Grammarly. It is important to note that no tool was employed for the generation of entire textual content; rather, tools like Grammarly or ChatGPT-3.5 were employed at most to revise specific segments of the text.

2 INTRODUCTION

This thesis seeks to contribute to the domain of entrepreneurial decision-making, specifically focusing on business pitches that aim to acquire investments. In Section 2.1, the research motivation is given, including the context of the research, relevant prior studies, and the scientific as well as societal significance of the research. In Section 2.2, the main research question is addressed, accompanied by the introduction of requisite sub-questions crucial for addressing the main research question. In Section 2.3, a brief overview is provided of the key findings from the thesis.

2.1 *Research motivation*

Startup companies frequently deliver their business proposals to potential investors in pursuit of funding to facilitate business expansion. Within a brief time frame, the pitcher communicates the core elements of the proposal, and the rationale behind their funding needs, and explains why an investor should contemplate investing in their startup company. Investment decisions are frequently made based on limited information, including subjective statements and nonverbal cues exhibited by the pitcher (Raab et al., 2020). In recent years, there has been an increasing focus on research aimed at understanding how an investor makes decisions concerning investments (Clarke et al., 2019).

Research in this domain holds the potential to offer invaluable insights for entrepreneurs aspiring to enhance their pitching skills. Additionally, it has the potential to provide insight into the decision-making processes for investors, pitchers, and researchers. Exploring this field of research can provide insight into the relationship between nonverbal cues such as body expressions, and vocal behavior displayed by the pitcher.

The entrepreneurial decision-making process, which involves determining the most suitable course of action based on the available information, is marked by high levels of uncertainty (Shepherd et al., 2015). The entrepreneurial decision-making process consists of decisions made by entrepreneurs themselves as well as those made by external actors with immediate repercussions for the entrepreneur (Shepherd, 2011). At the heart of effective business operations lies the pivotal role played by entrepreneurial decision-making.

A pitcher influences the decision-making process during an entrepreneurial pitch through both verbal and nonverbal cues. According to (Clarke et al., 2019), research often looks at the effects of verbal and nonverbal cues in isolation. Consequently, integrating various behavioral cues within a unified analysis holds the potential to yield valuable insights into

the interplay among these behavioral cues and their collective influence on the entrepreneurial decision-making process. behavioral cues that have a strong influence on the decision-making process are physical appearance, gestures and posture, face and eye behavior, vocal cues, space and environment (Liebregts et al., 2020).

A substantial proportion of communication emanates through body expressions, such as posture and gestures. Theoretical frameworks on gestures suggest that they play a pivotal role in conveying crucial information, facilitating cognitive processes, and enhancing learning and memory (Clough & Duff, 2020). Likewise, studies have reported that facial expressions and eye behavior during social interactions, coupled with gestures, influence the individuals' perceptual assessments (Ciuchta et al., 2017; Nagy et al., 2012). Another behavioral cue to consider is vocal behavior, which transmits information through acoustic features like amplitude, pitch, tempo, and tone. Vocal behavior encapsulates both an individual's personality traits and emotional state. (Warner & Sugarman, 1986) suggests that vocal information is embedded in the speech style and significantly influences the perception of the decision-maker in terms of personality dimensions.

behavioral cues can be categorized into various modalities, and previously investigated modalities for forecasting investment likelihood, in four out of the six sessions of the entrepreneurial dataset (Liebregts et al., 2018-2023), involve verbal, visual, and vocal modalities (Goossens et al., 2022; Jung et al., 2023; Van Aken et al., 2023). Notably, (Van Aken et al., 2023) achieved commendable results in predicting the vocal modality by extracting acoustic features from the entrepreneurial dataset. The research within the visual modality has concentrated on body expressions, facial expressions, head movement, and mimicry (Prabawa et al., 2022; Stoitsas et al., 2022). (Jung et al., 2023) attained commendable results through the incorporation of body expressions coupled with deep learning networks, such as a Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM).

This thesis endeavors to integrate the visual and vocal modalities, both of which have demonstrated predictive value in prior research. It adopts a multimodal approach that transcends the isolation of these modalities and captures the temporal aspect through the utilization of deep learning. The multimodal approach enables the exploration of the interplay between the visual and vocal modalities, shedding light on their collective impact on the likelihood of investment. Furthermore, the thesis investigates whether a combined utilization of these modalities enhances the predictive accuracy of the investment likelihood. The thesis aims to broaden the research scope by including two additional in-person pitch sessions (19 pitches), bringing

the total to six sessions (44 pitches) within the entrepreneurial dataset (Liebregts et al., 2018-2023).

2.2 *Research strategy*

To address the problem statement, the following research strategy has been devised comprising the main research question:

> *Can the likelihood of investment be predicted from acoustic and body expression features from a pitcher during an entrepreneurial pitch by using multimodal analysis?*

To evaluate the main research question, the following sub-questions need to be answered:

RQ1 *How do the acoustic features of openSMILE and VGGish or a combination of both compare in predicting the likelihood of investment?*

RQ2 *Which distinct body expression feature set or combination of features of OpenPose has the best performance in predicting the likelihood of investment?*

RQ3 *How does a Gated Recurrent Unit model perform in comparison to a Long Short-Term Memory model in predicting the likelihood of investment?*

RQ4 *How does early fusion of the multimodal compare to late fusion in predicting the likelihood of investment?*

2.3 *Thesis findings*

The findings indicate that both acoustic and body expression features possess the capacity to discern temporal patterns, and, to some degree, can be used to predict the likelihood of investment. Combining single feature representations within an unimodal enhances predictability, particularly with the adoption of a late fusion approach. Notably, the LSTM model demonstrates superior performance over the GRU model when applied to unimodal data with a singular feature representation. However, in the context of combining feature representation and multimodal approaches, the GRU models surpass the LSTM models. The combination of the modalities slightly improves its accuracy in predicting the likelihood of investment.

3 LITERATURE REVIEW

3.1 *Related work*

In the entrepreneurial field, not all decisions originate solely from entrepreneurs, some decisions are made by external entities, like fundraising decisions. An integral facet of the entrepreneurial process involves securing financial support from investors to facilitate business expansion (Liebregts et al., 2020). The subject of decision-making holds a well-established position of interest within the field of entrepreneurship research (Shepherd, 2011; Shepherd et al., 2015). Entrepreneurs make numerous decisions on a daily basis, often navigating through circumstances characterized by high risk and uncertainty (Baron, 1998). Considering the uncertainty inherent in operating a business and decision-making, entrepreneurs frequently resort to a set of flexible decision-making principles (Dew et al., 2009; Sarasvathy, 2008). In an instance where a decision involves other individuals and the entrepreneur is unable to gather additional information about the counterpart, stereotyping can significantly influence the entrepreneur's assessments and decisions (Bodenhausen, 1993; Greenwald & Banaji, 1995). This holds particularly valid when the decision-making process is impacted by one or more social interactions between the entrepreneur and the involved parties (Huang et al., 2013). In the existing entrepreneurship literature, considerable attention has been directed towards funding decisions, particularly those made by investors (Chen et al., 2009; Huang & Pearce, 2015).

The prevailing evidence unambiguously indicates that behavioral cues exhibited during human-to-human interactions have a significant influence over the decision-making process (Ambady & Rosenthal, 1992; Bonaccio et al., 2016; Mcneill, 2005). This has also been extensively demonstrated concerning entrepreneurial decision-making with particular emphasis has been placed on decisions related to hiring or employment decisions (Koch et al., 2014). Numerous studies have explored entrepreneurial pitches as a specific context wherein investors assess business ideas, observing that their funding decisions are typically influenced by the verbal and nonverbal expressions of entrepreneurs (Ciuchta et al., 2017; Clarke et al., 2019; Pollack et al., 2012). Based on prior research, both verbal and nonverbal elements are recognized as pivotal in decision-making processes within the entrepreneurial context (Chen et al., 2009; Clark, 2008). Consequently, behavioral cues including both verbal and nonverbal cues during social interactions have been demonstrated to have a substantial influence on individuals' decision-making process in an entrepreneurship environment (Liebregts et al., 2020).

The persisting inquiry revolves around the extent to which verbal and nonverbal behavioral cues can be used to predict decision-making processes within the entrepreneurial context and the consequential value of distinct behavioral cues. As stated in Section 2.1, prior research of the entrepreneurial decision-making process is exploring these behavioral cues, some of which have yielded promising outcomes (e.g., Goossens et al., 2022; Jung et al., 2023; Van Aken et al., 2023) on the entrepreneurial dataset (Liebregts et al., 2018-2023). (Goossens et al., 2022) explored the influence of vocal behavior on funding decisions. The research sought to determine whether deep learning methods could be utilized to predict the decisions of investors based on vocal behavior. The application of deep learning methodology employing a Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) model resulted in a predictive accuracy of 77.8% for predicting investors' decisions.

Subsequently, (Van Aken et al., 2023) used a comparable deep learning methodology, extending its application to behavioral cues, such as acoustic and linguistic modalities, through a multimodal approach. The acoustic and linguistic features are derived from recordings of entrepreneurial pitches using a combination of handcrafted and deep features extracted through tools such as openSMILE, VGGish, LIWC, and Longformer. The multimodal approach facilitates the integration of feature representations, even between different modalities, thereby enhancing the predictive accuracy of investment likelihood. The results of a multimodal approach in the entrepreneurial context have demonstrated promise, attaining an average Mean Absolute Error (MAE) of 13.91 across the initial four sessions of the entrepreneurial dataset. This was accomplished through the utilization of both acoustic and linguistic features, complemented by a multimodal approach utilizing early fusion.

(Jung et al., 2023) explored the nonverbal behavioral cues of body expressions. The study included both traditional regression models as well as deep recurrent regression models, capable of capturing the temporal aspect from the recordings. The study extracted body expression features using OpenPose, a tool designed to capture anatomical key points of the human body. The deep recurrent regression models, such as the GRU and LSTM, demonstrated superior performance compared to traditional regression models like decision trees, random forests, and support vector machines. The findings of this study indicated that employing a GRU model for temporal modeling of body expressions resulted in the most favorable performance (average MAE of 16.9).

3.2 *State-of-the-art methods*

In Section 2.1, the problem statement is established, and in Section 2.2, the research strategy is formulated in main and sub-questions based on Section 3. In this section, the state-of-the-art methods are discussed for integrating acoustic and body expression features in a multimodal analysis, based on prior research outlined in Section 3.1.

Acoustic features constitute a fundamental basis of human interaction. These acoustic features are extensively utilized to analyze human behavior, including the decision-making processes. The extraction of features from raw audio signals and the subsequent creation of feature representations constitute an integral component of vocal behavior analysis. Broadly speaking, two distinct approaches exist for feature engineering related to acoustics. Feature representations are either handcrafted using domain knowledge or learned through the utilization of deep learning algorithms (Swain et al., 2018).

The extraction of handcrafted audio features is commonly performed using a tool known as openSMILE (e.g., Goossens et al., 2022; Marchi et al., 2016; Sun et al., 2020; Van Aken et al., 2023). The openSMILE feature extraction toolkit integrates feature extraction algorithms from both the speech processing and the Music Information Retrieval communities. It supports a range of audio low-level descriptors, including CHROMA and CENS features, loudness, Mel-frequency cepstral coefficients, perceptual linear predictive cepstral coefficients, linear predictive coefficients, line spectral frequencies, fundamental frequency, and formant frequencies. Additionally, delta regression and various statistical functionals can be applied to these low-level descriptors (Eyben et al., 2010).

The extraction of audio features through deep learning is commonly performed utilizing a tool called VGGish (e.g., Goossens et al., 2022; Sun et al., 2020; Van Aken et al., 2023). The VGGish architecture is pre-trained on an extensive YouTube dataset (Yu et al., 2020), incorporating 128-dimensional embeddings for each AudioSet segment generated through a VGG-like audio classification model (Hershey et al., 2017).

The extraction of body expression features commonly employs deep learning algorithms due to their adeptness in efficiently recognizing the body posture and gestures of individuals. Two notable instances of such algorithms, capable of extracting body posture and gestures from video data, include OpenPose and DensePose (Cao et al., 2017). (Jung et al., 2023) utilized OpenPose to extract anatomical key points. Subsequently, mathematical calculations were performed to determine angles, distance ratios, and area ratios between various body parts based on the anatomical key points. OpenPose has multiple deep learning models designed for feature

extraction from the body, face, and hands (Cao et al., 2019; Simon et al., 2017; Wei et al., 2016). The features extracted from OpenPose can function as a baseline for comparing the results of the mathematical calculations, or they can be integrated into a combined feature representation.

When behavioral cues exist in isolation, they are commonly referred to as unimodal features, comprising a specific behavioral cue such as vocal behavior, or gestures and posture. In human-to-human communication, reliance is often not placed on a single unimodal feature, but rather on the combination of various unimodal features. The established literature consistently outlines the superiority of multimodal approaches over unimodal ones (D'mello & Kory, 2015; Stoitsas et al., 2022; Van Aken et al., 2023). The integration of unimodal models into a multimodal model is known as fusion, and there are two types: early fusion at the feature level and late fusion at the decision level. Early fusion involves concatenating the feature vectors from different modalities into a single vector. On the other hand, late fusion entails training separate unimodal models for each modality, and the results are then fused to make a final decision (Poria et al., 2017).

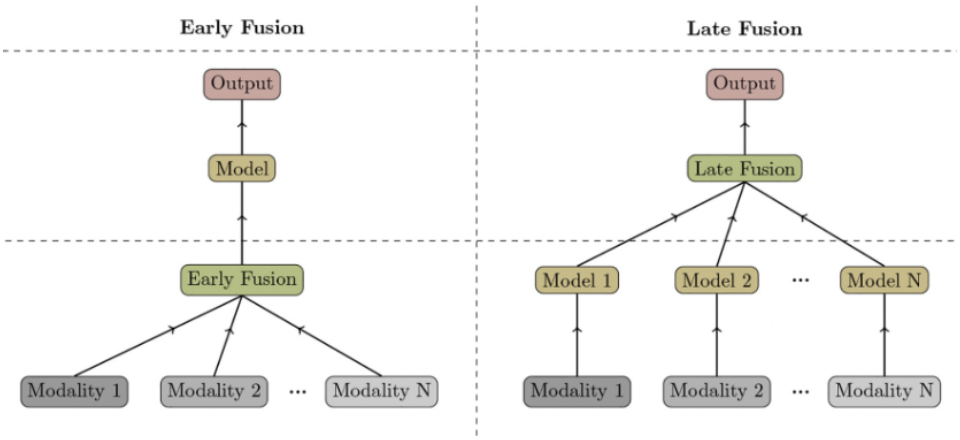

Figure 1: Visualization of early and late fusion, image source: Zhang et al., 2021

4 METHODOLOGY

4.1 *Entrepreneurial dataset*

The dataset utilized in this study comprises video recordings capturing entrepreneurial pitches (Liebregts et al., 2018-2023). The in-person video recordings contain 44 pitches. Among these, 25 pitches were recorded between 2018 and 2020, featuring a resolution of 1080p at 25 frames per second (FPS). Simultaneously, the remaining 19 pitches were recorded in the subsequent years, 2022 and 2023, with a resolution of 720p and a frame rate

of 29.97 FPS. The video recordings contain the pitch of a business proposal by a university student who participates in an entrepreneurship course. The student presents the proposal to a panel of three experienced investors. Following the pitch, the investors are provided with the opportunity to pose inquiries related to the business proposal in a question-and-answer (Q&) session. In this study, the analysis focuses specifically on the pitch segment within the video, the Q&A sessions are being excluded during the pre-processing stage. The 44 pitches were recorded across six distinct sessions, each characterized by a consistent set of investors. However, varying sessions involve different investors. Each session contains 5-10 pitches of approximately 3 minutes. After each session, each investor independently asses the pitches by assigning an investment probability rating (i.e., 0-100), and the investors complete a survey containing nonbehavioral cues, such as demographics, entrepreneurial traits, and entrepreneurial competencies.

Due to the COVID-19 pandemic, the format of pitch sessions underwent a transition from an in-person setting to an online environment. The online setting, however, introduces additional constraints compared to the in-person setting. Notably, online sessions often exhibit notable drawbacks, such as compromised audio and video quality, and a tendency for pitchers to be seated, thereby constraining their body movements (Kuhn & Sarfati, 2021). Furthermore, (Kuhn & Sarfati, 2021) explored the impact of transitioning to online settings on the perception of social signals by investors. The results indicate that acoustic features fulfill a more prominent role in the assessment of pitches within online settings. Therefore, it is probable that the association between acoustic and body expression features differs between in-person recordings and online recordings. Hence, the six in-person sessions, totaling 44 videos, are utilized for training the models and comparing results.

Each pitch is associated with three investment probability scores provided by the investors. However, predictive models are designed to estimate a singular likelihood of investment. Two options for creating a singular likelihood of investment exist: the first involves averaging the scores, and the second selects the highest likelihood of investment score. In this study, the latter option is adopted for several reasons. Firstly, the primary objective of the pitcher is to secure capital, and achieving this goal does not necessarily require convincing every individual investor. Secondly, instances may arise where certain investors are less enthusiastic due to factors such as a mismatch in the background with the business proposal's sector or lack of experience in the pitch's industry. However, as long as at least one investor is receptive, the overarching goal is achieved. Averaging the probabilities in such scenarios might yield a lower score. Lastly, previous research conducted on the dataset has consistently utilized

the highest probability score as a metric, and deviating from this standard makes it difficult to compare results. In figure 2, the distribution of the highest investment scores can be seen for each in-person session.

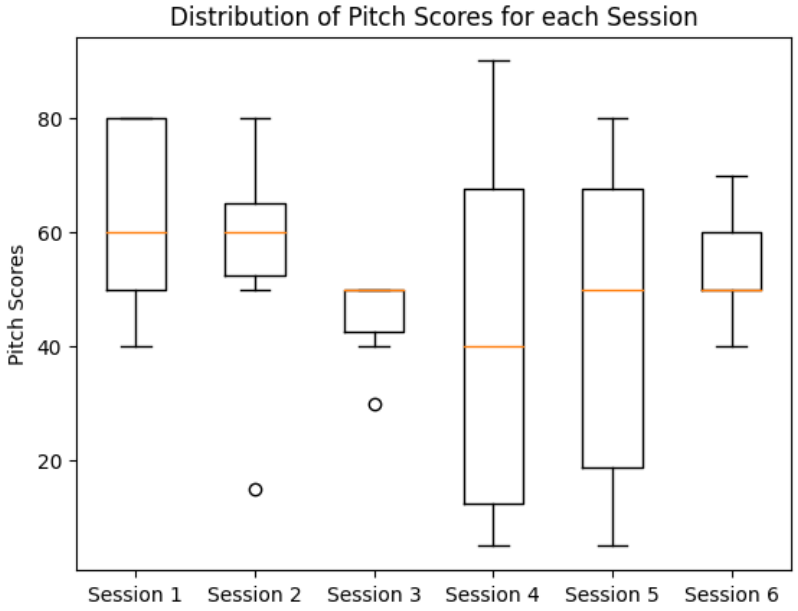

Figure 2: Box plots visualizing the distribution of the highest investment scores by each in-person session.

4.2 *Feature Extraction*

In this section, we will delve into the pre-processing procedures applied to the video data. Subsequently, the pre-processed data will be employed for the extraction of acoustic and body expression features. This extraction process will be facilitated through the utilization of the openSMILE toolkit, VGGish, and OpenPose.

4.2.1 *Pre-processing data*

Before the extraction of features from the data and the training of models on these features, various pre-processing steps must be executed. The raw data consists of video files in the MPEG-4 format and the following steps are taken to pre-process the data:

1 The initial step involves removing videos from the dataset for which no consent has been granted by the pitcher, as well as excluding videos from the online setting. This process results in a reduction of the dataset from 70 videos to 44 distributed across 6 sessions.

2 The subsequent step involves identifying the start and end times of each pitch and subsequently trimming the video to match the actual duration of the pitch utilizing the MoviePy package in Python. The trimmed video incorporates both audio and video data, and it is stored in MPEG-4 format.

3 The third step involves determining the duration of the videos and establishing a uniform length for all videos in the dataset. In this instance, the second-longest pitch, totaling 5 minutes and 17 seconds, is used as a benchmark, since the longest pitch exceeds 9 minutes. Given the utilization of a Recurrent Neural Network (RNN), specifically a GRU or LSTM, it is necessary that every video in the dataset conforms to the same predetermined length. Consequently, videos shorter than the specified duration undergo zero-padding to meet the set length, while videos exceeding the predetermined length are trimmed accordingly.

4 The fourth step involves extracting audio data from videos. This process involves utilizing the MoviePy package in Python to extract audio data from the MPEG-4 files. Subsequently, the extracted audio data is stored in WAV files, chosen for their uncompressed nature, facilitating the preservation of more information for the feature extraction processes.

5 The fifth stage involves segmenting the audio into chunks with the Pydub package in Python. Given that VGGish extracts a feature vector every 0.96 seconds, it necessitates dividing the audio data into chunks of 0.96 seconds. This segmentation ensures that the handcrafted features, extracted using openSMILE, align with the same temporal rate as VGGish. Consequently, openSMILE utilizes corresponding data segments, ensuring alignment with the feature extraction rate employed by VGGish.

6 The sixth stage involves extracting a frame from the video data at a consistent rate of 0.96 seconds, aligning with the temporal rate established in the preceding step. This synchronization is crucial to maintaining a uniform temporal rate for both the body expression features and acoustic features. The frame extraction is accomplished using the OpenCV package in Python. The decision to extract a frame every 0.96 seconds was prioritized over averaging frames within the same interval. This choice was made based on OpenPose's preference for images of higher quality, essential for accurate determination of anatomical key points. Averaging frames sometimes presented challenges for OpenPose in identifying these key points.

7 The final step involves generating a singular prediction score as stated in Section 4.1. The investment probabilities of each investor are stored in a comprehensive spreadsheet. This spreadsheet includes crucial details regarding the pitches, including the pitch ID, session ID, and file path. The highest score for each pitch is determined and stored in a distinct column within the spreadsheet. The column is structured for convenient loading into Python using the Pandas package.

4.2.2 *Acoustic features*

After the pre-processing of the dataset into usable audio data, suitable for the extraction of acoustic features, the subsequent step involves feature extraction within two distinct categories. The first category involves explainable and handcrafted features, while the second category involves features derived from deep learning methodologies. The extraction of handcrafted features involves the utilization of the openSMILE toolkit, as detailed in Section 3.2. Concurrently, the VGGish tool is utilized to extract deep learning features.

Within the openSMILE toolkit, the extraction process involves employing the *extended Geneva Minimalistic Acoustic Parameter Set*, resulting in the extraction of 88 features. The feature set is conceived as a fundamental standard acoustic parameter set, designed with the aim of establishing a shared baseline for research within the acoustic domain (Eyben et al., 2016). The features include Low-Level Descriptors (Table 1) and Functionals (Table 2), extracted at a regular interval of 0.96 seconds, as outlined in step 5 of Section 4.2.1.

The extracted features are structured into a matrix of dimensions T x 88 for each pitch, where T denotes the quantity of 0.96-second segments that align with the duration of the pitch. To capture temporal information in the audio signal, a model utilizes either a Gated Recurrent Unit (GRU) or a Long Short-Term Memory (LSTM) layer, Both are categorized as Recurrent Neural Networks (RNN). The GRU and LSTM layers require uniform input shapes, requiring consistency in the dimension denoted by T for each pitch. Shorter pitches undergo zero-padding to establish feature vectors of equal length, while longer pitches are trimmed, following the procedure outlined in step 3 of Section 4.2.1. This approach is outlined as a standard method for achieving uniform feature vector lengths in the audio modality (Han et al., 2020).

Meanwhile, VGGish undertakes the transformation of audio input into a semantically meaningful 128-dimensional embedding. This embedding is generated at regular intervals of 0.96 seconds for each segment of audio. Consequently, a feature set of dimensions T x 128 is produced for every pitch, with T representing the count of 0.96-second intervals aligning with

Table 1: Low-Level Descriptors (LLD) features extracted with OpenSMILE

Feature Group	Description
Waveform	Zero-Crossings, Extremes, DC
Loudness	Energy, Intensity, Auditory model loudness
FFT spectrum	Phase, Magnitude
ACF, Cepstrum	Autocorrelation, Cepstrum
Mel/Bark spectrogram	Bands o-N
Semitone spectrogram	FFT based and filter based
Cepstral	Cepstral features (e.g., MFCC, PLPCC)
Pitch	Fo via Autocorrelation and sub-harmonic summation, smoothed by Viterbi algorithm
Voice Quality	HNR, Jitter, Shimmer, Voice Probability
LP	LPC Coefficient, reflect coefficient, residual Line Spectral Pairs (LSP)
Auditory	Auditory spectra, psychoacoustic sharpness
Formants	Centre frequencies and bandwidths
Spectral	Energy in N user-defined bands, roll-off points, centroid, entropy, y
Tonal	CHROMA, CENS, CHROMA-based features

Table 2: Functionals extracted with OpenSMILE

Category	Description
Extremes	Extreme values, positions, and ranges
Means	Arithmetic, quadratic, geometric
Moments	Standard deviation, variance, kurtosis, skewness
Percentiles	Percentiles and percentile ranges
Regression	Linear and quad, approximation coefficients, regression error, and centroid
Peaks	Number of peaks, mean and standard deviation peak distance, mean and standard deviation peak amplitude
Segments	Number of segments based on delta thresholding or various fixed thresholds, mean and standard deviation segment length
Sample values	Values of the contour at configurable relative positions
Times/durations	Up- and down-level times, rise and fall times, duration
Onsets	Number of onsets, relative position of first and last on- and offset
DCT	Coefficients of the DCT
LPC	Autoregressive coefficients
Zero-Crossings	Zero-crossing rate, Mean-crossing rate

the duration of the pitch. The identical procedure, involving zero-padding and trimming to a standardized size of 5 minutes and 17 seconds, as outlined in Section 4.2.1, is applied to the pitches. This process yields a feature matrix with dimensions 330 x 128, thereby constructing a comprehensive and temporal representation of the acoustic modality.

4.2.3 *Body expression features*

The body expression modality engages in feature extraction through the utilization of OpenPose, since OpenPose yielded promising results in prior studies conducted on the same dataset (Jung et al., 2023). OpenPose utilizes video data to identify the body parts of the pitcher. To align with the temporal rate established in the acoustic modality, a frame from the video is extracted at intervals of 0.96 seconds, as outlined in step 6 of Section 4.2.1. The quality of the extracted frame is important for OpenPose's accurate recognition of anatomical key points. However, due to computational efficiency, the frames are downscaled to a resolution of 368x368, as recommended by Simon et al., 2017. This uniform resolution across individual frames and pitches enables the use of automatic feature scaling by OpenPose, resulting in the normalization of feature values within the range of 0 to 1.

OpenPose has multiple models designed for the extraction of anatomical key points from the human body. Within the scope of this study, the selected models are BODY_25 and FACE. The BODY_25 model captures 25 anatomical key points of the full body for each frame. However, as the recordings solely focus on the upper half of the pitchers, the key points corresponding to the legs and feet are discarded. The FACE model extracts 70 anatomical key points specifically from the facial region of the pitcher.

Following the extraction, the anatomical key points from both models are stored in separate JSON files for each individual frame. Subsequently, these key points are aggregated for each JSON file, and their alignment is mapped based on Figure 3. The resulting features are then compiled into a CSV file, wherein the confidence statistic for each feature is discarded.

In addition to the features derived from the BODY_25 and FACE models, Jung et al., 2023 computed affective body expression features based on perception, recognition, and generation, using data from the BODY_25 model. These affective body expression features consist of seven angles, five distance ratios, and three area ratios, collectively capturing nuances in body openness and alterations in body posture throughout the temporal domain (Table 3).

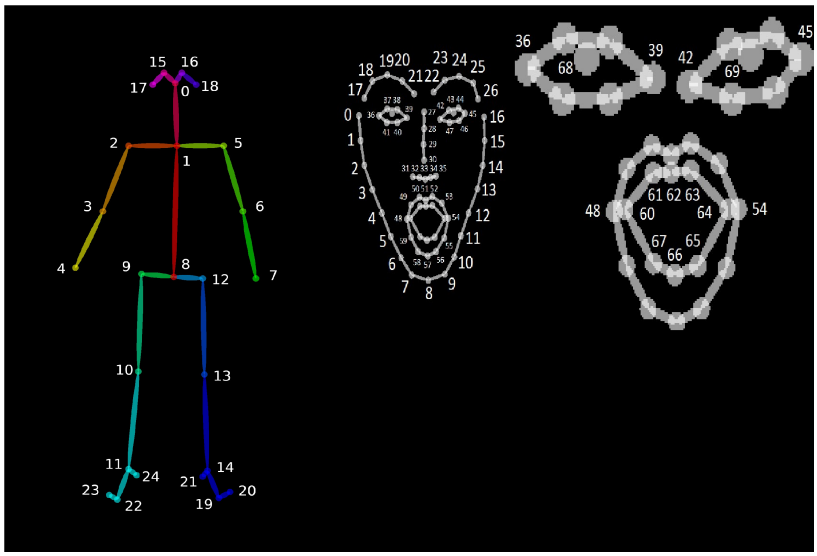

Figure 3: The anatomical key points extracted with the BODY_25 model on the left, and FACE model in OpenPose. *Source: OpenPose Github*

Table 3: Extracted Affective Body Expression Features, based on Jung et al., 2023

Feature Type	Body Parts
Angle	Left upper arm - left lower arm
	Right upper arm - right lower arm
	Left shoulder - neck
	Right shoulder - neck
	Left shoulder - left upper arm
	Right shoulder - right upper arm
	Neck - nose
	Left wrist - nose / left wrist - mid hip
	Left wrist - neck / left wrist - mid hip
Distance Ratio	Right wrist - nose / right wrist - mid hip
	Right wrist - neck / right wrist - mid hip
	Left wrist - right wrist / mid hip - nose
	Left wrist - right wrist - neck / left wrist - right wrist - mid hip
Area Ratio	Right wrist - nose - mid hip / left wrist - nose - mid hip
	Right wrist - neck - mid hip / left wrist - neck - mid hip

4.3 *Experimental Setup*

This section focuses on delineating the experimental setups pertaining to the unimodal feature representations of each modality. Subsequently, the singular feature representations will be amalgamated to form a com-

prehensive combined feature representation for each modality. Finally, a detailed discussion on the multimodal approach involving these modalities will be presented. This discussion will encompass the experimental setup for reproducibility purposes and the evaluation metrics employed in this study.

4.3.1 *Unimodal models: single feature representations*

The primary objective of the initial phase of the experimental setup is to address the sub-questions about the optimal performance of feature sets derived from the acoustic and body expression modalities in predicting the likelihood of investment. Additionally, this phase aims to offer partial insights into the comparative effectiveness of Recurrent Neural Network (RNN) types, specifically the GRU and LSTM.

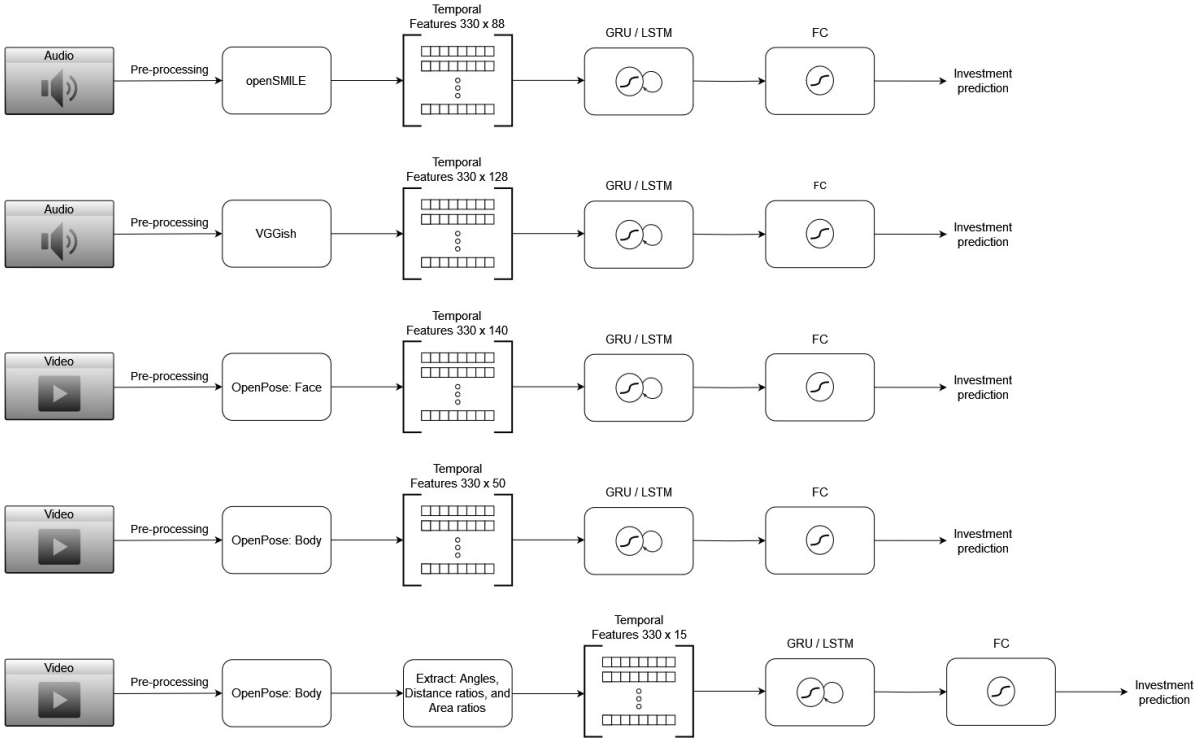

Figure 4: The workflow of the unimodal single feature representations.

The procedural sequence unfolds as follows: initially, the dataset undergoes pre-processing steps, subsequently, features are extracted from the data using diverse tools, as outlined in Section 4.2. Following this, the extracted features are input into a single GRU or LSTM layer, as Recurrent Neural Networks (RNN) can capture the temporal dynamics inherent in human behavior. The output of the RNN layer is directed into a regression

layer, generating a continuous value that represents the probability of investment in an entrepreneurial pitch. Figure 4 illustrates the workflow corresponding to each feature representation. This approach was utilized in the following studies: (Soleymani et al., 2019; Tavabi et al., 2020; Van Aken et al., 2023).

For each model, hyperparameter tuning is conducted through the implementation of a grid search algorithm. This algorithm systematically explores all possible combinations of predefined parameters. The combination of parameters yielding the lowest validation score on the validation set is selected to train the model. Explicit procedures for training and evaluating the models are elaborated upon in Section 4.4.

4.3.2 *Unimodal models: combining feature representations*

The second phase of the experimental setup continues to address the sub-questions concerning the optimal performance of feature sets derived from the acoustic and body expression modalities in predicting the likelihood of investment. However, in this phase, the focus shifts to the integration of single feature representations into a combined feature representation. Additionally, the evaluation compares the performance of the best-performing GRU and LSTM models for each modality. This assessment is crucial for determining their integration within the multimodal approach.

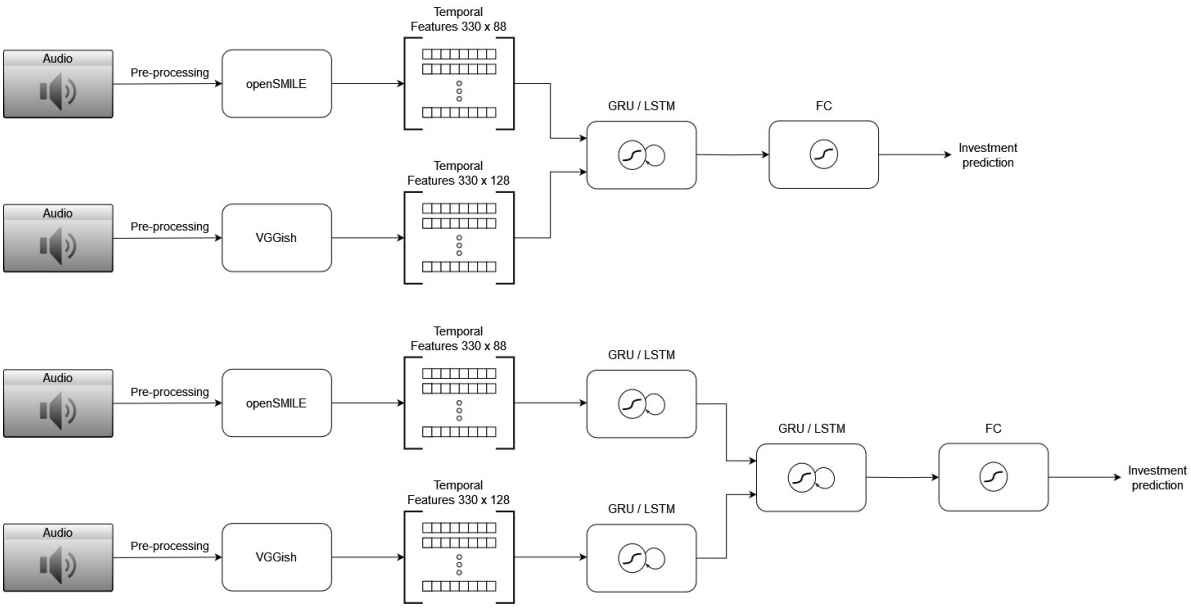

Figure 5: The workflow of the acoustic combined feature representations.

To address the first sub-question, *How do the acoustic features of openSMILE and VGGish or a combination of both compare in predicting the likelihood*

of investment?, the hand-crafted feature representation from openSMILE will be combined with the deep learning features from VGGish. In Figure 5, the workflows illustrate the combination of these feature representations. The first representation utilizes early fusion, achieved by concatenating the feature representation into a single feature vector, which is then input into a GRU or LSTM layer. Subsequently, the regression layer predicts an investment score. The second representation utilizes late fusion, wherein the single feature representations are input into a GRU or LSTM layer. The outcomes of these layers are subsequently combined into another GRU or LSTM layer. Finally, the output of the combined layer is fed into the regression layer, which predicts an investment score.

To address the second sub-question, *Which distinct body expression feature set or combination of features of OpenPose has the best performance in predicting the likelihood of investment?*, The features extracted from OpenPose, including BODY_25, FACE, and affective body expression, will be combined into a unified feature representation. The feature representations will be combined with an early fusion and late fusion approach, as shown in Figure 6. The early fusion approach works the same as the early fusion approach of the acoustic modality, where the single feature representations are concatenated into a single feature vector and are then fed into a single GRU or LSTM layer. Whereas the late fusion approach utilizes a GRU or LSTM layer for every single feature representation. The outcomes of these layers are then aggregated into a singular GRU or LSTM layer. Ultimately, the results from the single GRU or LSTM layer are directed into a regression layer for predicting the investment probability.

4.3.3 *Multimodal models*

After identifying the optimal feature sets for each modality and the most effective RNN, as detailed in the preceding subsections of 4.3, multiple multimodal approaches are developed for answering the fourth sub-question *How does early fusion of the multimodal compare to late fusion in predicting the likelihood of investment?* The multimodal models are designed to predict the probability of an investment score by leveraging two distinct modalities. The multimodal models are constructed using the Keras module within the Tensorflow package. Each model utilizes two types of fusion, early fusion and late fusion. In early fusion, the single feature vectors from different modalities are concatenated into one feature vector. The combined vector is then fed into an RNN, and the resulting output from the RNN is directed to a regression layer for predicting the investment score. In late fusion, an RNN computes the output for each modality, and these individual outputs are then input into a single RNN layer. The resulting output from the

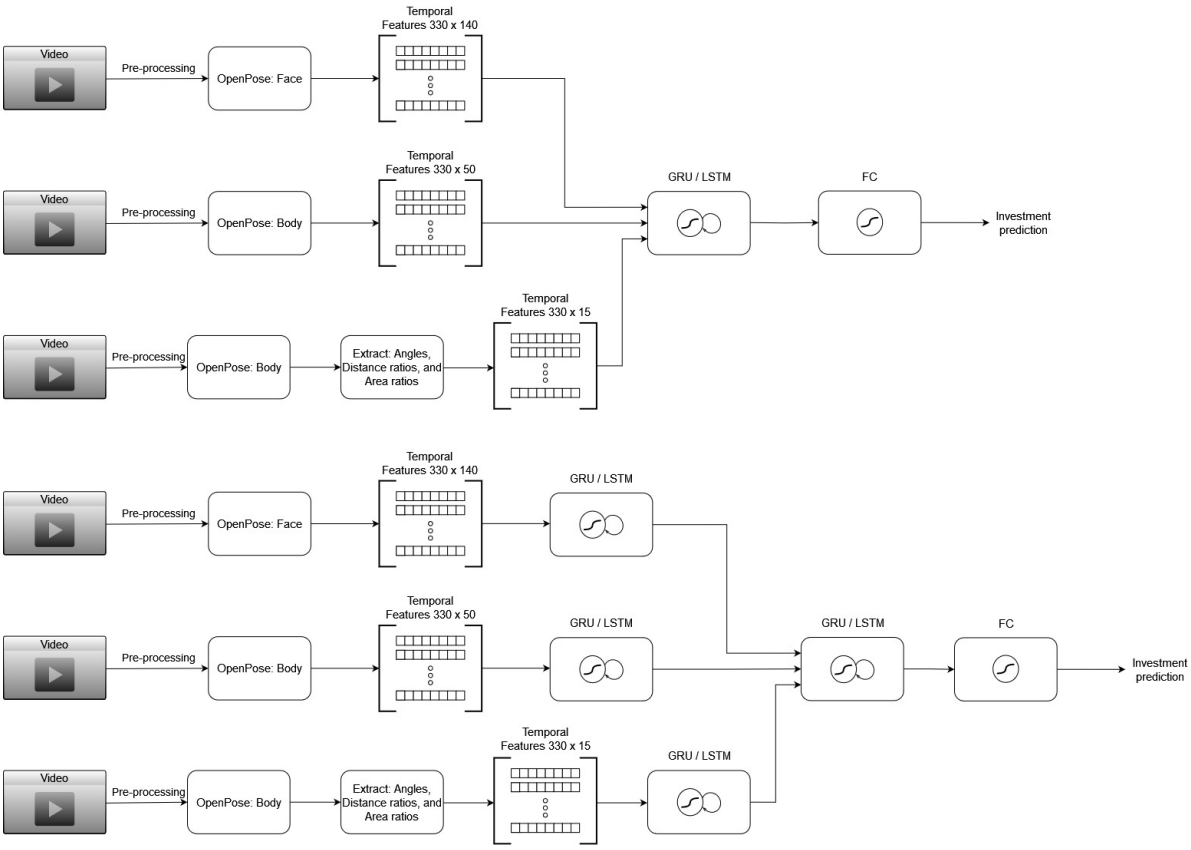

Figure 6: The workflow of combining the feature representations utilizing early and late fusion on the body expression features.

single RNN layer is subsequently fed into a regression layer for predicting the investment score.

The multimodal models will utilize both the best-performing single feature representation of each modality and the most effective combination of feature representations for each modality. The results obtained from these models collectively contribute to addressing the main research question. The optimal GRU or LSTM model, identified as the best performer for each modality, will also be incorporated into the multimodal models. In Figure 7, the workflow of the multimodal model integrating the best-performing single feature representations along with their respective best-performing RNN for each modality is illustrated. The initial workflow demonstrates an early fusion approach, while the subsequent workflow illustrates the late fusion approach.

In addition to the multimodal model incorporating the best-performing single feature representations for each modality, there will be another multimodal model utilizing a combination of the best-performing feature

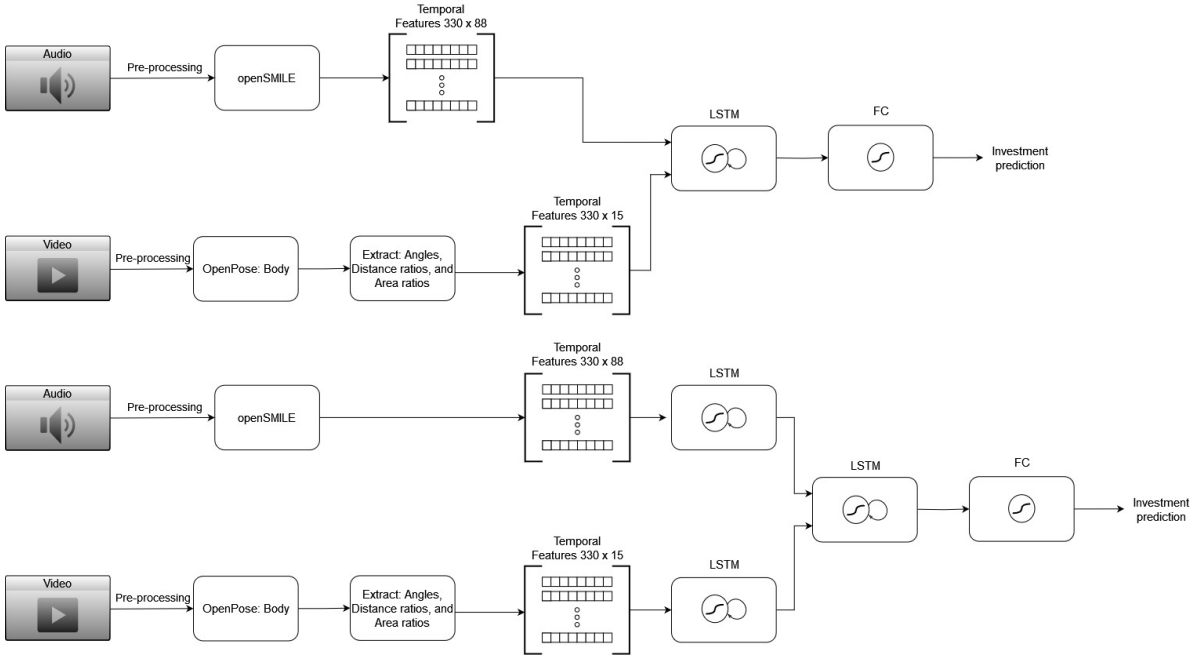

Figure 7: The multimodal model workflows with the best-performing single feature representations of each modality

representations for each modality. Illustrated in Figure 8, the early fusion multimodal model integrates the combined feature representations, feeding them into a bidirectional GRU to capture temporal information inherent in the features. The GRU output is then passed through a fully connected layer, transforming the task into a regression problem and producing an investment score as the output for every single pitch.

In Figure 9, the multimodal model incorporating the combined acoustic and body expression feature representations is visualized. The individual acoustic feature representations are input into an LSTM layer, and the outputs are subsequently directed into another LSTM layer to generate an outcome for the combined feature representation. Simultaneously, the singular body expression feature representations are fed into a GRU layer, and the outcomes of the GRU layer are then input into another GRU layer, producing an outcome for the combined feature representation. Finally, the outcomes from the acoustic and body expression modalities are directed in a final GRU layer, and the resulting output is fed into a regression layer, predicting an investment score for each pitch. The choice of utilizing a single RNN layer between the modalities was made to effectively capture the temporal aspects of both modalities within a unified RNN layer.

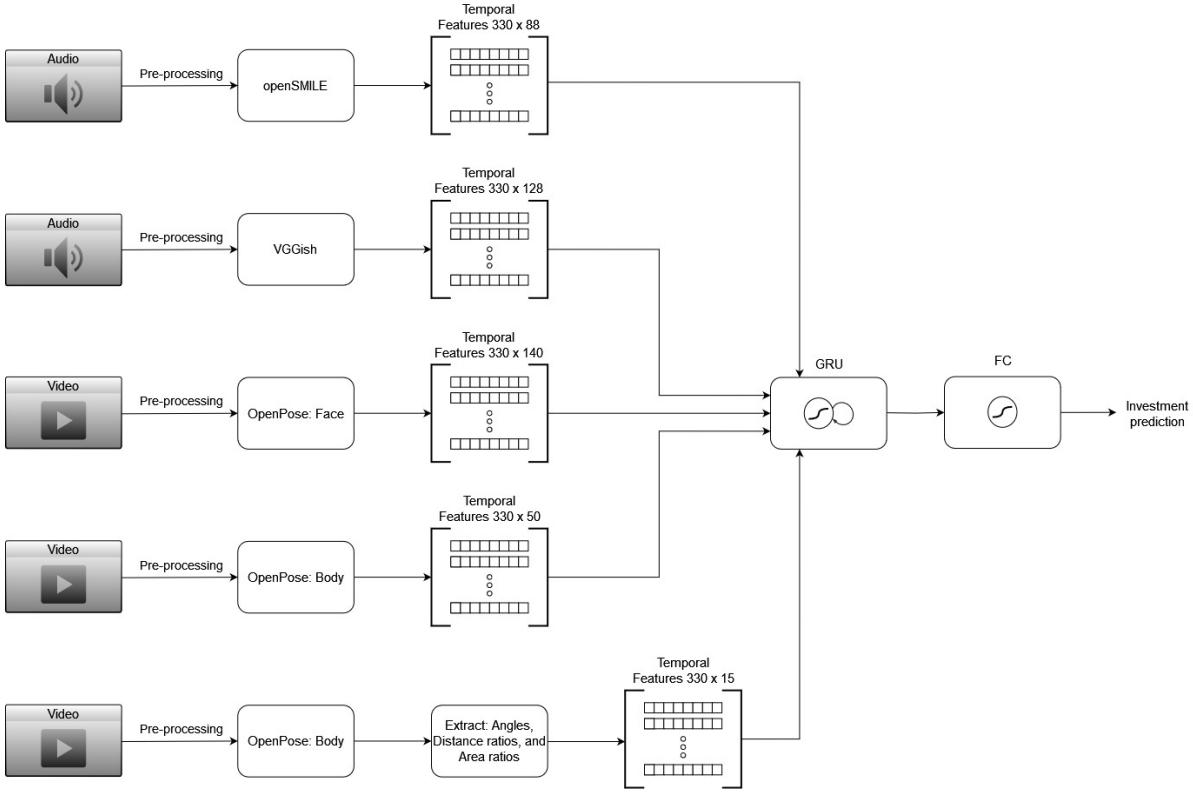

Figure 8: Multimodal model workflow utilizing early fusion on the best-performing combination of feature representations of each modality.

4.4 *Training and Evaluation*

In this section, the methodologies for training and evaluating the models described in Section 4.3, are explained. Before training the models, the dataset is split into a training, validation, and test set. While the conventional approach for dividing the dataset involves random sampling, it was decided to create a fold for each specific session. This decision ensures that each fold, or test set, is representative of an actual session. Consequently, the 44 in-person recordings are distributed over 6 folds, where each fold includes all the pitches from a session. The number of pitches within each fold varies, ranging from 5 to 10 pitches per fold. In Table 11, a comprehensive overview of the pitches and their distribution across the folds is provided.

After partitioning the dataset into training, validation, and test sets, a Grid Search algorithm is employed to optimize the hyperparameters for the models. The hyperparameter tuning process for each model involves running 10 epochs for every possible combination of hyperparameters. The

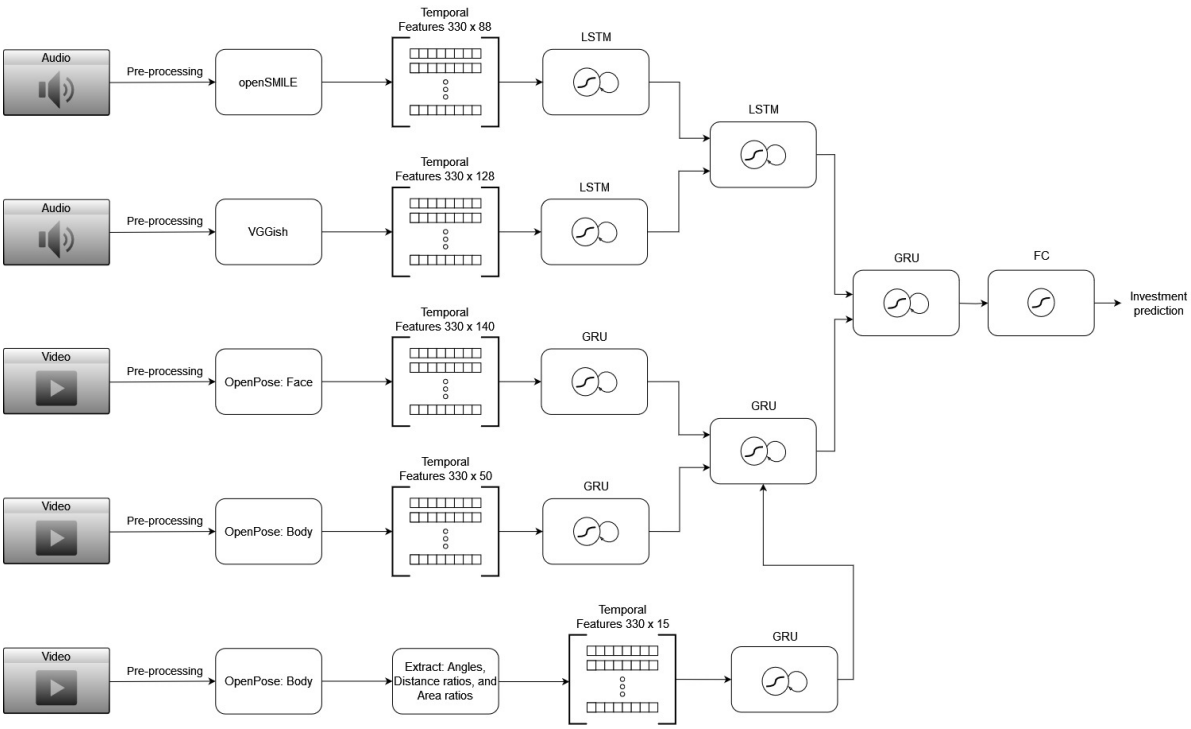

Figure 9: Multimodal model workflow utilizing late fusion on the best-performing combination of feature representations of each modality.

hyperparameter space is detailed in Table 4. The hyperparameters that result in the lowest validation score are selected as the optimal hyperparameters.

Hyperparameters	**Explored Values**
Number of units	64, 128, 256
Learning rate	0.01, 0001
Drop-out rate	0, 0.1
Optimizer	Adam, SGD

Table 4: The possible hyperparameters utilized in the Grid Search algorithm.

After determining the optimal hyperparameters, the models are trained on the training and validation sets with a batch size of 5 for a total of 300 epochs. Subsequently, the performance of the trained models is evaluated using the test set. The performance of all models is evaluated using the Mean Absolute Error (MAE) as the chosen metric. This decision facilitates direct comparisons between different models and aligns with the widespread acceptance of MAE for assessing regression problems.

Previous research on the dataset has also employed MAE as an evaluation metric, enabling comparisons between the current models and prior work, such as Van Aken et al., 2023 and Jung et al., 2023. The MAE is calculated by summing the absolute errors and dividing by the sample size.

5 RESULTS

In this section, the outcomes of the conducted experiments, as outlined in 4.3, are presented. The structure of the chapters mirrors that of the methodology, with a focus on addressing the sub-questions, as outlined in 2.2.

5.1 *Unimodal models: single feature representations*

The findings derived from the single feature representations within the acoustic modality are provided in Table 5. The outcomes are presented through the utilization of the Mean Absolute Error (MAE) evaluation metric for each specific fold, alongside the computation of the average MAE across the entire model. The best-performing model within the modality is emphasized through the application of bold formatting.

The handcrafted features derived from openSMILE exhibit superior performance compared to the deep learning features obtained from VGGish, for both the GRU and LSTM models. This observation is noteworthy, especially in light of the findings in Van Aken et al., 2023, where the identical feature extraction methodology resulted in the deep learning features outperforming the handcrafted features. A plausible explanation for this disparity could be attributed to the inclusion of two additional sessions, wherein the handcrafted features demonstrate better generalization capabilities across the dataset in contrast to the deep learning features. An intriguing observation arises from the acoustic singular feature representation results, particularly when employing the first fold as the test set. The substantial variations in the outcomes strongly suggest that the generalization of the other five folds onto the first fold is notably challenging within the context of the acoustic modality. This observation is particularly noteworthy, especially when taking into account the relatively limited distribution of pitch scores for the first session. A potential explanation for this could be the relatively high mean of the pitch scores, as illustrated in Figure 2.

The outcomes derived from the single feature representations within the body expression modality are detailed in Table 6. An intriguing observation is the consistent trend wherein all feature sets within the body expression modality tend to outperform their counterparts in the acoustic modality,

Feature Set	MAE 1	MAE 2	MAE 3	MAE 4	MAE 5	MAE 6	**avg MAE**
openSMILE (GRU)	22.28	11.57	13.37	27.36	22.56	12.18	18.22
VGGish (GRU)	32.67	21.20	14.98	30.61	24.08	9.53	22.18
openSMILE (LSTM)	14.85	11.43	15.05	30.83	23.17	8.82	**17.36**
VGGish (LSTM)	25.35	17.63	14.45	30.83	23.64	12.21	20.69

Table 5: The single feature representation results of the acoustic modality.

particularly evident in the first fold. Across both modalities, the LSTM models consistently outperform their GRU counterparts. Additionally, a noteworthy pattern emerges in the uniform results observed in the fourth fold, particularly within the LSTM models. The models utilizing singular feature representations encounter challenges in effectively generalizing onto this fold, often resorting to predicting a mean value based on the remaining folds.

Feature Set	MAE 1	MAE 2	MAE 3	MAE 4	MAE 5	MAE 6	**avg MAE**
Aff. Body Expr. (GRU)	19.95	15.15	6.97	31.21	23.32	17.35	18.99
BODY_25 (GRU)	14.54	16.40	12.70	27.21	25.23	10.20	17.71
FACE (GRU)	15.79	15.93	8.51	30.83	23.02	8.28	17.06
Aff. Body Expr. (LSTM)	15.43	15.54	7.36	30.83	23.17	8.44	**16.80**
BODY_25 (LSTM)	16.13	15.71	9.92	30.83	23.12	8.69	17.40
FACE (LSTM)	16.47	15.85	8.20	30.83	24.67	8.35	17.40

Table 6: The single feature representation results of the body expression modality.

The most effective feature set within the body expression modality is the affective body expression feature set, particularly when employed in combination with an LSTM model. Nevertheless, the BODY_25 and FACE models demonstrate substantial competitiveness, surpassing the performance of feature sets within the acoustic modality. Intriguingly, when utilizing a GRU model, these models even outperform the affective body expression feature set. The affective body expression feature set is the best-performing single feature representation while having the least amount of features.

5.2 *Unimodal models: combined feature representations*

In Table 7, the outcomes of the combined feature representation in the acoustic modality are presented. Both early fusion and late fusion approaches are employed for both GRU and LSTM layers. Notably, the LSTM

Feature Set	MAE 1	MAE 2	MAE 3	MAE 4	MAE 5	MAE 6	**avg MAE**
EF GRU openSMILE + VGGish	26.65	11.90	19.37	33.02	23.25	8.51	20.45
LF GRU openSMILE + VGGish	28.44	**9.74**	19.97	26.56	23.69	15.72	20.69
EF LSTM openSMILE + VGGish	23.54	16.39	14.35	25.79	23.22	8.36	18.61
LF LSTM openSMILE + VGGish	15.69	13.81	15.91	**24.87**	23.37	8.62	**17.05**

Table 7: The combined feature representation results of the acoustic modality.

layers consistently demonstrate superior performance compared to the GRU layers within the acoustic modality, with the late fusion approach outperforming the early fusion approach. The best-performing average MAE model is denoted in bold, alongside the best scores of each fold across all models in Tables 7 and 8.

Feature Set	MAE 1	MAE 2	MAE 3	MAE 4	MAE 5	MAE 6	**avg MAE**
EF GRU: Aff. Body Exp., BODY_25, FACE	**12.83**	17.32	**6.74**	29.04	**22.04**	8.43	**16.07**
LF GRU: Aff. Body Exp., BODY_25, FACE	16.05	15.75	16.09	30.45	24.30	**7.09**	18.29
EF LSTM: Aff. Body Exp., BODY_25, FACE	15.75	16.00	7.43	30.83	24.73	8.78	17.25
LF LSTM: ff. Body Exp., BODY_25, FACE	18.55	15.52	11.32	30.70	23.10	9.80	18.17

Table 8: The combined feature representation results of the body expression modality.

In Table 8, the results of the combined feature representation within the body expression modality are presented. In contrast to the acoustic modality, the GRU layers exhibit superior performance compared to the LSTM layers, and early fusion consistently outperforms late fusion. Notably, the body expression modality continues to face challenges in generalizing onto the second and fourth sessions, in contrast to the combined feature representation of the acoustic modality, which demonstrated optimal performance on these folds. Whereas the combined body expression features exhibit remarkable performance on the remaining folds. An additional observation is that the combined feature representations demonstrate the capability to make predictions on the fourth fold, an improvement compared to the performance of single feature representations.

5.3 *Multimodal models*

Derived from the outcomes of the unimodal models, two multimodal models were devised. The first integrated the most effective single feature representations from the acoustic and body expression modalities. Meanwhile, the second multimodal model incorporated the optimal combined feature representations from both modalities.

The initial multimodal model incorporated handcrafted features from openSMILE and affective body expressions derived from the BODY_25 model. Both features demonstrated optimal performance when integrated with an LSTM layer. In Table 9, the results of the multimodal model are presented, revealing that a late fusion multimodal model approach surpasses the outcomes of the unimodal models. Nevertheless, it is crucial to acknowledge that while the overall performance improved, the individual performance on specific folds decreased. This implies that the multimodal model excels at incorporating results from both modalities and tends to average the outcomes of both layers, aligning with its anticipated behavior. Looking ahead, for further enhancement in results, a potential strategy involves selecting the best-performing modality on a specific fold rather than averaging the results of both modalities.

Feature Set	MAE 1	MAE 2	MAE 3	MAE 4	MAE 5	MAE 6	**avg MAE**
EF: openSMILE + Aff. Body. Exp. (LSTM)	18.24	13.31	16.33	26.02	23.77	10.54	18.04
LF: openSMILE (LSTM) + Aff. Body. Exp. (LSTM)	14.28	11.28	12.41	26.21	23.16	9.01	**16.06**

Table 9: The single feature representation multimodal result.

The second multimodal model integrates the combined feature representations from both the acoustic and body expression modalities. The combined feature representation of the acoustic modality performed best utilizing an LSTM layer, meanwhile the combined feature representation of the body expression modality performed best utilizing a GRU layer. In Table 10, the outcomes of the multimodal model utilizing combined feature representations are displayed. The results suggest that the multimodal models exhibit commendable performance for both early and late fusion, although they do not surpass the performance of the multimodal model employing single feature representations. Similar observations are noted for this multimodal model; while achieving high scores on every fold, they do not outperform the unimodal models on specific folds.

Feature Set	MAE 1	MAE 2	MAE 3	MAE 4	MAE 5	MAE 6	**avg MAE**
EF GRU: Acoustic & Body expressions	15.14	17.12	13.15	29.56	22.97	8.49	17.74
LF: Acoustic (LSTM) + Body expressions (GRU)	15.91	15.79	12.07	26.54	23.40	8.64	**17.06**

Table 10: The combined feature representation multimodal result.

6 DISCUSSION

6.1 *Summary and discussion of results*

This study aims to address the main research question, *Can the likelihood of investment be predicted from acoustic and body expression features from a pitcher during an entrepreneurial pitch by using multimodal analysis?* To achieve this goal, the results of the sub-questions will be examined, leading to a comprehensive discussion. Subsequently, the findings will be synthesized to provide a conclusive response to the main research question.

To answer the first sub-question, the findings indicate that a feature set, combining handcrafted features from openSMILE with deep learning features from VGGish, surpasses the performance of individual single feature representations. To address the second sub-question, the results suggest that a feature set that combines affective body expressions with the BODY_25 and FACE features exhibits superior performance compared to the individual single feature representations.

Addressing the third sub-question concerning the optimal choice between a GRU or LSTM model, it is contingent upon the modality and feature representation. Specifically, the LSTM model exhibits superior performance on single feature representations and the combined feature representation within the acoustic modality. Nevertheless, it is evident that within the combined feature representation of the body expression modality, the GRU models outperform their LSTM counterparts.

To address the fourth sub-question, it is observed that, in general, the late fusion approach tends to outperform early fusion. The only exception to this trend was noted in the case of the combined feature representation for the body expressions, where both early fusion models outperformed their late fusion counterparts. However, across all multimodal models and the acoustic modality, the late fusion approach consistently exhibited superior performance compared to early fusion.

To address the main research question, both the acoustic and body expression modalities demonstrated the ability to yield satisfactory results in predicting the likelihood of investment. However, it is noteworthy that

the single feature representations faced challenges, particularly evident on the fourth fold. This limitation was successfully addressed by combining the feature representations for each modality, and the multimodal model approach introduced a notable improvement in consistency across all individual folds.

6.2 *Comparison to literature*

The first study for comparison is Van Aken et al., 2023, which adopted a methodology similar to the one employed in this study. In that research, focusing on the first four sessions of the dataset Liebregts et al., 2018-2023, the best outcome entailed an average Mean Absolute Error (MAE) of 13.91. This was achieved by employing the best feature set of each modality through an early fusion approach. Conversely, combining feature representations for each modality resulted in inferior performance. Similar trends were observed in the current study, where the combination of multiple feature sets within a modality into a multimodal model tended to significantly increase model complexity. Furthermore, the less successful single feature representations had a detrimental impact on the overall model performance.

An interesting deviation lies in the fact that early fusion appeared to perform better in the previous study Van Aken et al., 2023, while late fusion demonstrated superior performance in the current study. This discrepancy highlights the nuanced influence of fusion strategies on model outcomes and underscores the importance of considering contextual factors in multimodal analyses. An additional intriguing deviation lies in the performance of acoustic features. In this study, handcrafted features surpassed the performance of deep learning features, contrary to the findings in Van Aken et al., 2023 where better performance was observed with deep learning features. A plausible explanation for this shift could be attributed to the increased size of the dataset, expanding from 25 pitches in the previous study to 44 pitches in the current study. The larger dataset may have introduced variations and complexities that impacted the relative performance of handcrafted and deep learning features, emphasizing the sensitivity of results to dataset characteristics and scale.

The second study chosen for comparison is that of Jung et al., 2023, where the focus was on utilizing body expressions to predict investment outcomes. This study presented intriguing results concerning affective body expressions, a feature set also employed in the current study. In Jung et al., 2023, which concentrated on the first four sessions of the dataset, a GRU model yielded an MAE of 16.9, while the LSTM model produced a higher MAE of 20. This presents a notable deviation from our findings, as

in our study, the LSTM model outperformed the GRU model for single feature representations.

Additionally, an interesting difference arises in the performance ranking of body expression and acoustic features. In the current study, body expression features outperformed acoustic features, whereas prior results from Jung et al., 2023 and Van Aken et al., 2023 indicated that acoustic features yielded superior results in the first four sessions compared to body expression features. This discrepancy underscores the variability in outcomes across studies and emphasizes the need for careful consideration of contextual factors and dataset characteristics.

6.3 *Scientific and societal impact*

The outcomes of this study underscore the efficacy of combining different modalities to enhance investment predictions compared to unimodal model approaches. Particularly, previously studied unimodal models such as the affective body expression features, which have historically been examined in isolation, exhibit improved prediction capabilities when integrated into a multimodal framework. The insights gained from this study hold potential for future advancements in the realm of entrepreneurial decision-making. Practical applications may involve developing tools wherein entrepreneurs can upload their pitches for analysis and receive feedback in the form of identified points of improvement to enhance their pitching skills. This approach aligns with the broader trend of leveraging multimodal analyses to provide nuanced insights and support decision-making processes.

6.4 *Limitations and future directions*

In terms of limitations, it is important to acknowledge that the models in this study were trained on a relatively small dataset consisting of 44 pitches. Furthermore, it is essential to note that the pitches are not based on real-world scenarios, introducing challenges in interpreting investment probability, as real-world investment decisions are typically binary (yes or no), rather than expressed as a percentage. Additionally, previous studies utilized SHAP (SHapley Additive exPlanations) to establish feature importance. However, due to the lack of transparency in the gradient within the Tensorflow package for GRU and LSTM models, implementing SHAP analysis for these models is currently not feasible. Models that do not incorporate a time series approach may also overlook the temporal aspect, which would not be captured in a SHAP analysis of such models. These limitations should be taken into consideration when interpreting the results and applying the findings to real-world scenarios.

This study focuses on nonverbal behavior cues, but future research could explore the integration of non-behavioral cues, such as the demographics and entrepreneurial traits of both the pitcher and the investors. These non-behavioral cues may play a crucial role in determining investment probability, as factors like investor experience in the pitch's sector can significantly impact investment decisions. For instance, an inexperienced investor in the pitch's sector might have a different influence on investment probability compared to an experienced one. Recognizing and incorporating non-behavioral cues is essential, as investment probability is not solely dependent on the pitch's quality but also on various contextual and individual factors. Exploring these additional dimensions could provide a more comprehensive understanding of the dynamics involved in entrepreneurial investment decision-making.

7 CONCLUSION

In this thesis, the prediction of investment likelihood was explored through the utilization of acoustic and body expression features extracted from recordings of entrepreneurial pitches. To capture the temporal dynamics inherent in the videos, deep learning models with Recurrent Neural Networks were employed. A multimodal approach was adopted, combining the acoustic and body expression modalities through both early and late fusion techniques applied to the feature representations. This comprehensive methodology aimed to leverage the synergies between different modalities and temporal aspects to enhance the accuracy of investment predictions.

The presented findings reveal promising results in predicting the likelihood of investment for entrepreneurial pitches using acoustic and body expression features. Satisfactory performance has been achieved on this dataset through the implementation of a multimodal model, integrating the best-performing features from each modality using late fusion. Notably, in the experiments, body expression features generally outperformed acoustic feature representations. Furthermore, the combination of feature representations in unimodal models demonstrated promising results specific to each modality. Overall, the late fusion approach consistently outperformed early fusion, highlighting its effectiveness in leveraging the strengths of each modality for enhanced prediction accuracy.

REFERENCES

Ambady, N., & Rosenthal, R. W. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*, 256–274. https://api.semanticscholar.org/CorpusID:31293145

Baron, R. A. (1998). Cognitive mechanisms in entrepreneurship: Why and when enterpreneurs think differently than other people. *Journal of Business Venturing*, *13*(4), 275–294. https://doi.org/https://doi.org/10.1016/S0883-9026(97)00031-1

Bodenhausen, G. (1993). Emotion, arousal, and stereotypic judgments: A heuristic model of affect and stereotyping. In D. Mackie & D. Hamilton (Eds.), *Affect, cognition, and stereotyping: Interactive processes in group perception* (pp. 13–37). Academic Press.

Bonaccio, S., OReilly, J., O'Sullivan, S., & Chiocchio, F. (2016). Nonverbal behavior and communication in the workplace: A review and an agenda for research. *Journal of Management*, *42*. https://doi.org/10.1177/0149206315621146

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., & Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1302–1310. https://doi.org/10.1109/CVPR.2017.143

Chen, X.-P., Yao, X., & Kotha, S. (2009). Entrepreneur passion and preparedness in business plan presentations: A persuasion analysis of venture capitalists' funding decisions. *Academy of Management Journal*, *52*, 199–214. https://doi.org/10.5465/AMJ.2009.36462018

Ciuchta, M., Letwin, C., Stevenson, R., & McMahon, S. (2017). Betting on the coachable entrepreneur: Signaling and social exchange in entrepreneurial pitches. *Entrepreneurship Theory and Practice*, *2017*, 15065–15065. https://doi.org/10.1177/1042258717725520_

Clark, C. (2008). The impact of entrepreneurs' oral 'pitch' presentation skills on business angels' initial screening investment decisions. *Venture Capital*, *10*(3), 257–279. https://doi.org/10.1080/13691060802151945

Clarke, J. S., Cornelissen, J. P., & Healey, M. P. (2019). Actions speak louder than words: How figurative language and gesturing in entrepreneurial pitches influences investment judgments. *Academy of Management Journal*, *62*(2), 335–360.

Clough, S., & Duff, M. C. (2020). The role of gesture in communication and cognition: Implications for understanding and treating neurogenic communication disorders. *Frontiers in Human Neuroscience*, *14*. https://doi.org/10.3389/fnhum.2020.00323

Dew, N., Read, S., Sarasvathy, S., & Wiltbank, R. (2009). Effectual versus predictive logics in entrepreneurial decision making: Differences between experts and novices. *Journal of Business Venturing*, *24*, 287–309. https://doi.org/10.1016/j.jbusvent.2008.02.002

D'mello, S., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*, *47*, 1–36. https://doi.org/10.1145/2682899

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., & Truong, K. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing [Open access]. *IEEE transactions on affective computing*, *7*(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on Multimedia*, 1459–1462. https://doi.org/10.1145/1873951.1874246

Goossens, I., Jung, M., Liebregts, W., & Onal Ertugrul, I. To invest or not to invest: Using vocal behavior to predict decisions of investors in an entrepreneurial context [12th international workshop on human behavior understanding, HBU ; Conference date: 21-08-2022 Through 21-08-2022]. English. In: 12th international workshop on human behavior understanding, HBU ; Conference date: 21-08-2022 Through 21-08-2022. 2022, August, 1–14. https://www.cmpe.boun.edu.tr/hbu/2022/index.html

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Han, W., Jiang, T., Li, Y., Schuller, B., & Ruan, H. (2020). Ordinal learning for emotion recognition in customer service calls, 6494–6498. https://doi.org/10.1109/ICASSP40776.2020.9053648

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R., & Wilson, K. (2017). Cnn architectures for large-scale audio classification. In *International conference on acoustics, speech and signal processing (icassp)*. https://arxiv.org/abs/1609.09430

Huang, L., Frideger, M., & Pearce, J. (2013). Political skill: Explaining the effects of nonnative accent on managerial hiring and entrepreneurial

investment decisions. *The Journal of applied psychology*, *98*. https://doi.org/10.1037/a0034125

Huang, L., & Pearce, J. L. (2015). Managing the unknowable: The effectiveness of early-stage investor gut feel in entrepreneurial investment decisions. *Administrative Science Quarterly*, *60*(4), 634–670. https://doi.org/10.1177/0001839215597270

Jung, M. M., Van Vlierden, M., Liebregts, W., & Önal Ertuğrul, I. (2023). Do body expressions leave good impressions? - predicting investment decisions based on pitcher's body expressions. *Companion Publication of the 25th International Conference on Multimodal Interaction*, 36–40. https://doi.org/10.1145/3610661.3617156

Koch, A., D'Mello, S., & Sackett, P. (2014). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *The Journal of applied psychology*, *100*. https://doi.org/10.1037/a0036734

Kuhn, N., & Sarfati, G. (2021). Zoomvesting: Angel investors' perception of subjective cues in online pitching. *Journal of Entrepreneurship in Emerging Economies*, *ahead-of-print*. https://doi.org/10.1108/JEEE-09-2021-0363

Liebregts, W., Darnihamedani, P., Postma, E., & Atzmueller, M. (2020). The promise of social signal processing for research on decision-making in entrepreneurial contexts. *Small Business Economics*, *55*(3), 589–605. https://doi.org/10.1007/s11187-019-00205-1

Liebregts, W., Urbig, D., & Jung, M. (2018-2023). Survey and video data regarding entrepreneurial pitches and investment decisions. *[Unpublished raw data]*.

Marchi, E., Eyben, F., Hagerer, G., & Schuller, B. W. (2016). The promise of social signal processing for research on decision-making in entrepreneurial contexts. *INTERSPEECH*, 1182–1183.

Mcneill, D. (2005, January). *Gesture and thought*. https://doi.org/10.7208/chicago/9780226514642.001.0001

Nagy, B. G., Pollack, J. M., Rutherford, M. W., & Lohrke, F. T. (2012). The influence of entrepreneurs' credentials and impression management behaviors on perceptions of new venture legitimacy. *Entrepreneurship Theory and Practice*, *36*(5), 941–965. https://doi.org/10.1111/j.1540-6520.2012.00539.x

Pollack, J., Rutherford, M., & Nagy, B. (2012). Preparedness and cognitive legitimacy as antecedents of new venture funding in televised business pitches. *Entrepreneurship Theory and Practice*, *36*. https://doi.org/10.1111/j.1540-6520.2012.00531.x

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*. https://doi.org/10.1016/j.inffus.2017.02.003

Prabawa, A., Jung, M., Stoitsas, K., Liebregts, W., & Ertuğrul, I. (2022). Predicting probability of investment based on investor's facial expression in a startup funding pitch. *Proceedings of BNAIC/BeNeLearn 2022*.

Raab, M., Schlauderer, S., Overhage, S., & Friedrich, T. (2020). More than a feeling: Investigating the contagious effect of facial emotional expressions on investment decisions in reward-based crowdfunding. *Decision Support Systems*, *135*, 113326. https://doi.org/https://doi.org/10.1016/j.dss.2020.113326

Sarasvathy, S. (2008). Effectuation: Elements of entrepreneurial expertise. *Effectuation: Elements of Entrepreneurial Expertise*, *243*. https://doi.org/10.4337/9781848440197

Shepherd, D. A. (2011). Multilevel entrepreneurship research: Opportunities for studying entrepreneurial decision making. *Sage Journals*, *37*(2).

Shepherd, D. A., Williams, T. A., & Patzelt, H. (2015). Thinking about entrepreneurial decision making: Review and research agenda. *Journal of Management*, *41*(1), 11–46. https://doi.org/10.1177/0149206314541153

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. *CVPR*.

Soleymani, M., Stefanov, K., Kang, S.-H., Ondras, J., & Gratch, J. (2019). Multimodal analysis and estimation of intimate self-disclosure. *2019 International Conference on Multimodal Interaction*, 59–68. https://doi.org/10.1145/3340555.3353737

Stoitsas, K., Önal Ertuğrul, I., Liebregts, W., & Jung, M. M. (2022). Predicting evaluations of entrepreneurial pitches based on multimodal nonverbal behavioral cues and self-reported characteristics. *Companion Publication of the 2022 International Conference on Multimodal Interaction*, 121–126. https://doi.org/10.1145/3536220.3558041

Sun, L., Lian, Z., Tao, J., Liu, B., & Niu, M. (2020). Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism. *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, 27–34.

Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: A review. *International Journal of Speech Technology*, *21*. https://doi.org/10.1007/s10772-018-9491-z

Tavabi, L., Stefanov, K., Zhang, L., Borsari, B., Woolley, J., Scherer, S., & Soleymani, M. (2020). Multimodal automatic coding of client behavior in motivational interviewing [International Conference on Multimodal Interfaces 2020, ICMI 2020 ; Conference date: 25-10-2020 Through 29-10-2020]. In N. Berthouze, M. Chetouani, & M. Nakano (Eds.), *Proceedings of the 2020 international conference on multimodal interaction* (pp. 406–413). Association for Computing Machinery (ACM). https://doi.org/10.1145/3382507.3418853

Van Aken, P., Jung, M. M., Liebregts, W., & Onal Ertugrul, I. (2023). Deciphering entrepreneurial pitches: A multimodal deep learning approach to predict probability of investment. *Proceedings of the 25th International Conference on Multimodal Interaction*, 144–152. https://doi.org/10.1145/3577190.3614146

Warner, R., & Sugarman, D. (1986). Attributions of personality based on physical appearance, speech, and handwriting. *Journal of Personality and Social Psychology*, *50*, 792–799. https://doi.org/10.1037/0022-3514.50.4.792

Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. *CVPR*.

Yu, H., Chen, C., Du, X., Li, Y., Rashwan, A., Le Hou, P. J., Yang, F., Liu, F., Kim, J., & Li, J. (2020). Tensorflow model garden.

Zhang, Y., Sidibé, D., Morel, O., & Mériaudeau, F. (2021). Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, *105*, 104042. https://doi.org/https://doi.org/10.1016/j.imavis.2020.104042

APPENDIX A

Table 11: The in-person recorded pitches with consent for each fold.

Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Fold 6**
Ziggurat	LittleSister	tALste	Ar-T-ficial	HomePage	Parkscout
PREA	FLIPR	Choos3 Wisely	Recipe-Me	RoundAbout	EvaluateMe
Young Boosters	Bubble Pop	SmArt	Salix	AVG OK	BellyBuddy
Whitebox	RecognEyes	StudentFood	Peech	SmartGrade	Lookasa
Soccer Academy	HOTIDY	wALste	HoodFood	Thousand Aiyes	Data Quench
	FitPoint	Chattern	LockUp	MyVeggie	Spot
	SOLON	FindIt		Kap in Kaart	Daily Dutch
				Sovereignty Consulting	Aimed Decisions
				Lendo	Conquse
				CertifAI	