Affective Computing Use Cases

Dr. Anil Rahate

Artificial Intelligence

AI is every where

AI refers to refers to the machines and the services which are intelligent right and

in making them intelligent what it does it provides them the ability to learn

Tons of use cases in every aspect of human life

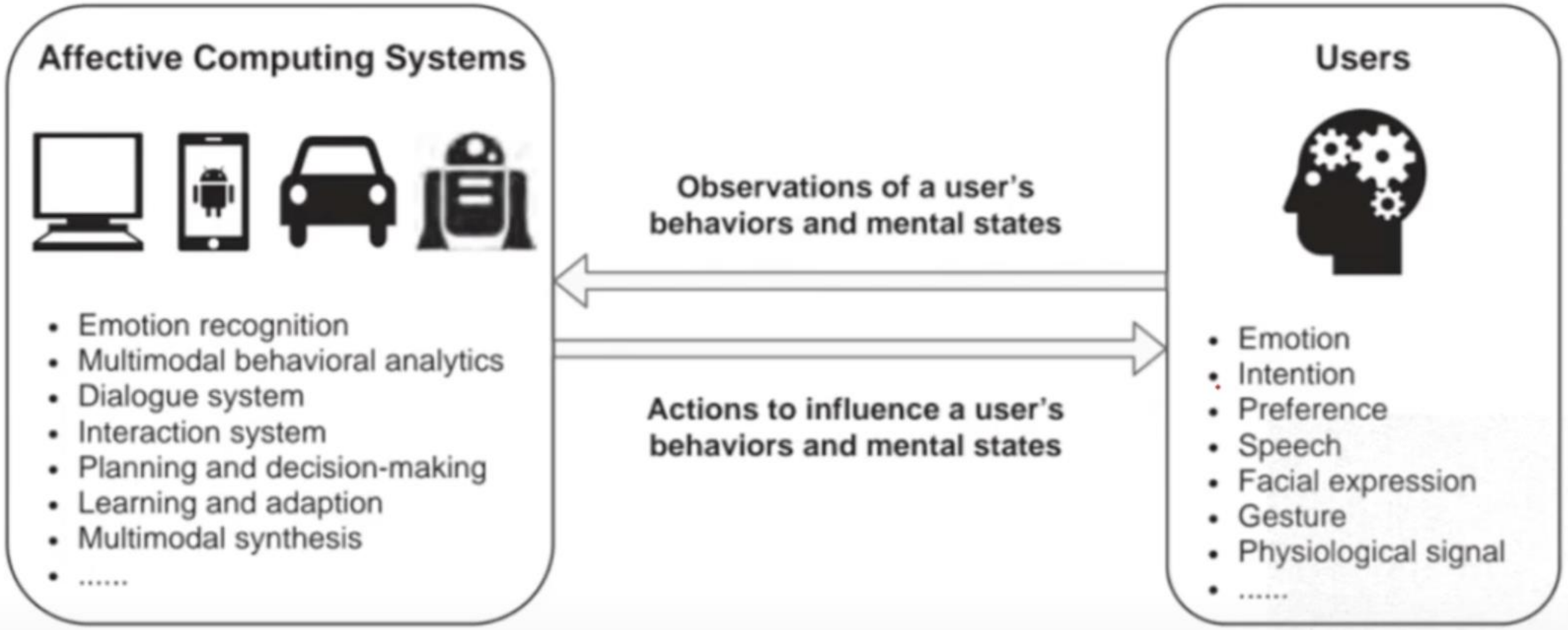
Affective Computing

- **Hypothetical scenario** so let us say there is a baby crying and the mother sees the baby and runs towards, the mother consoles the baby
- In this what all happened - the mother recognized that the baby was panicking was in a certain State and then reacted accordingly
- Now for the machines? are we there yet?

- So, the question is whether intelligent machines can have emotions?
- Can machines be intelligent without emotions
- Affective Computing is is all about making machines and services emotionally intelligent
- Affective Computing lies at the intersection of computer science design and psychology to be able to so that it can provide this emotional intelligence to the machines and the

Affective Computing

The field of affective computing encompasses both the creation of and interaction with machine systems that sense, recognize, respond to, and influence emotions (Picard, 1997; Picard and Klein, 2002).



Affect Sensing

- Affect sensing refers to a system that can recognize emotion by receiving data through signals and patterns (Picard, 1997).
- To accomplish this task, a computer would need to be equipped with hardware and software.
- Affect-sensing systems can be classified by modalities, each of which has a unique signature.

Affective Computing Areas

- Fundamentals of Affective Computing
- Emotion Theory and Emotional Design
- Affect Elicitation
- Emotions in Facial Expressions
- Emotions in Voice
- Emotions in Text
- Emotions in Physiological Signals
- Multimodal Emotion Recognition
- Emotional Empathy in Agents/Machines/Robots
- Online and Adaptive Recognition of Emotions: Challenges and Opportunities
- Use cases/applications
- Ethical Issues: Ethical, legal and Social Implications of Affective Computing

Human Multimodal Communication

Multimodal

Verbal

- **Lexicon**
  - Words
- **Syntax**
  - Part-of-speech
  - Dependencies
- **Pragmatics**
  - Discourse acts

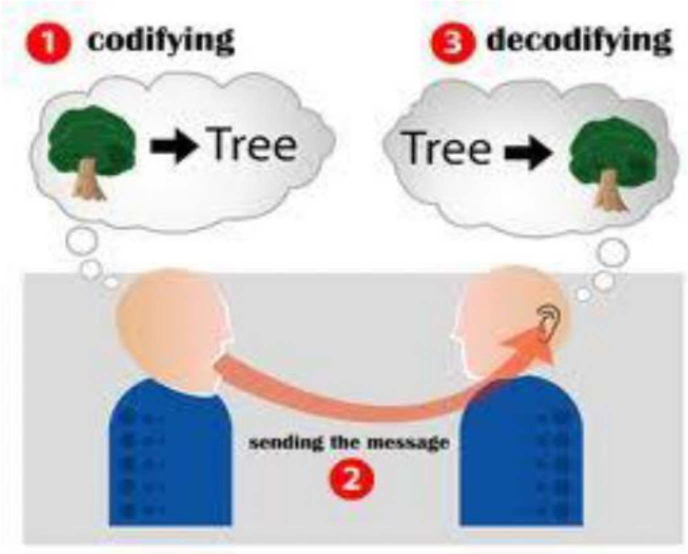
Vocal

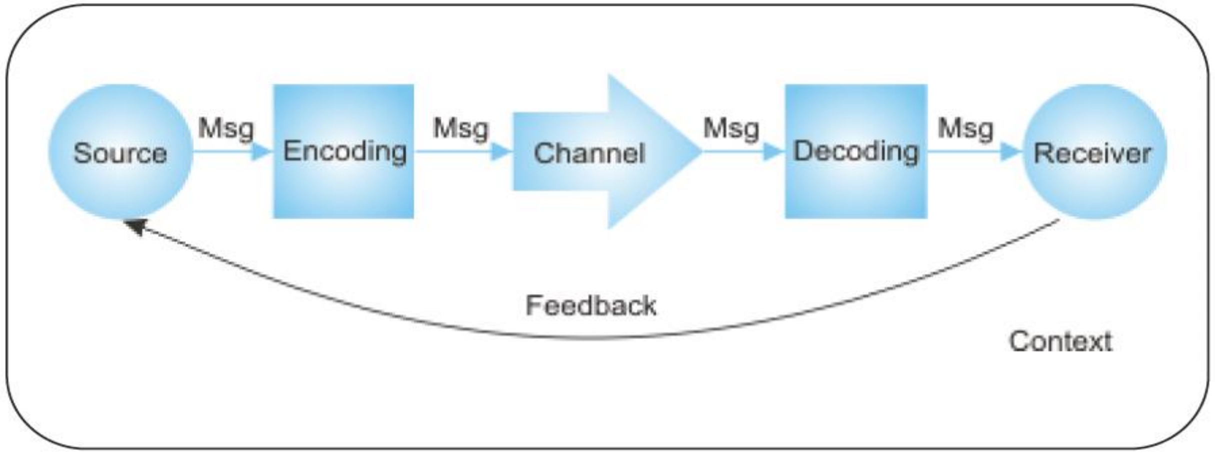
- **Prosody**
  - Intonation
  - Voice quality
- **Vocal expressions**
  - Laughter, moans

Visual

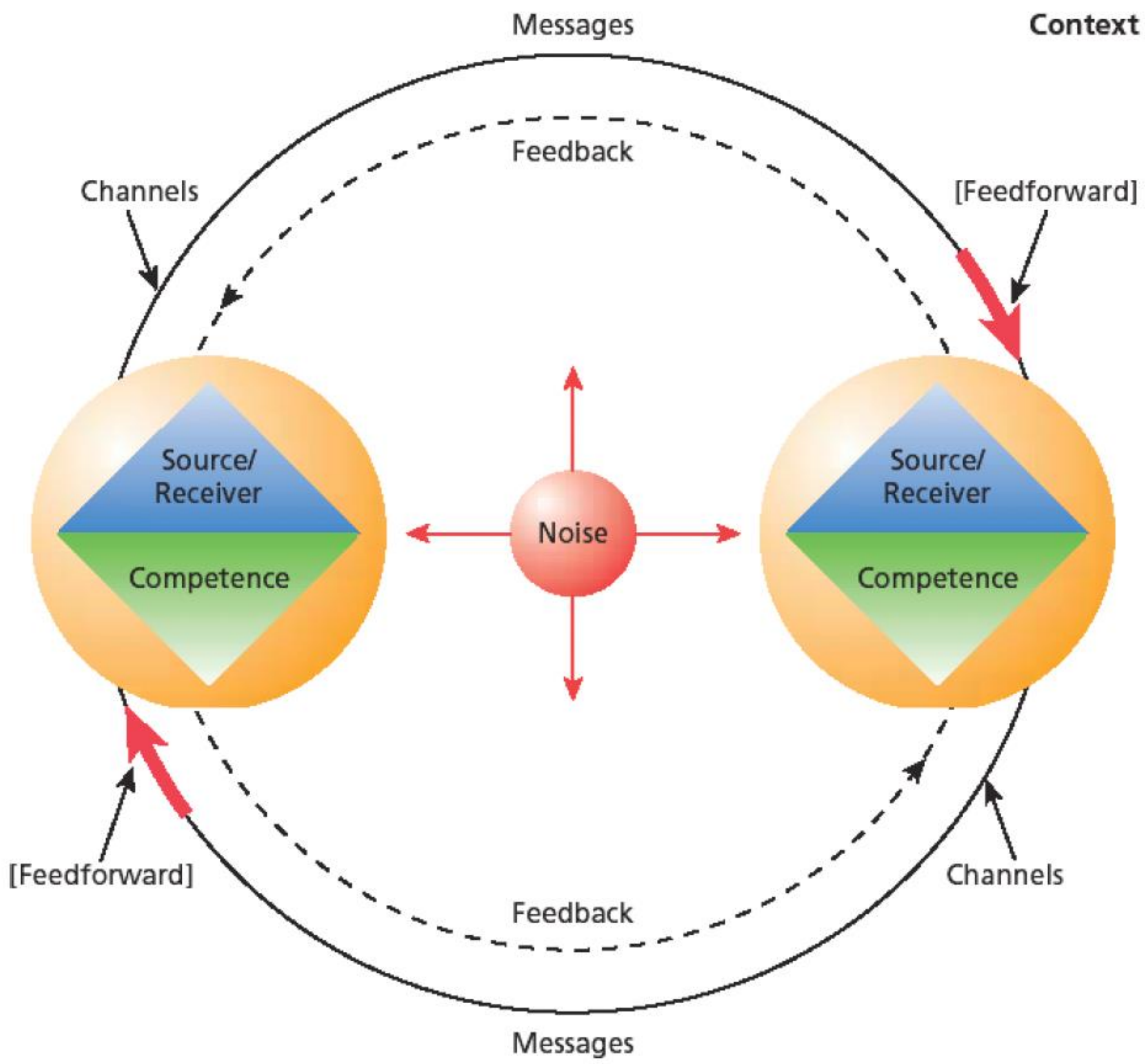
- **Gestures**
  - Head gestures
  - Eye gestures
  - Arm gestures
- **Body language**
  - Body posture
  - Proxemics
- **Eye contact**
  - Head gaze
  - Eye gaze
- **Facial expressions**
  - FACS action units
  - Smile, frowning

Communication Process: Encoder-decoder





Elements of Interpersonal Communication

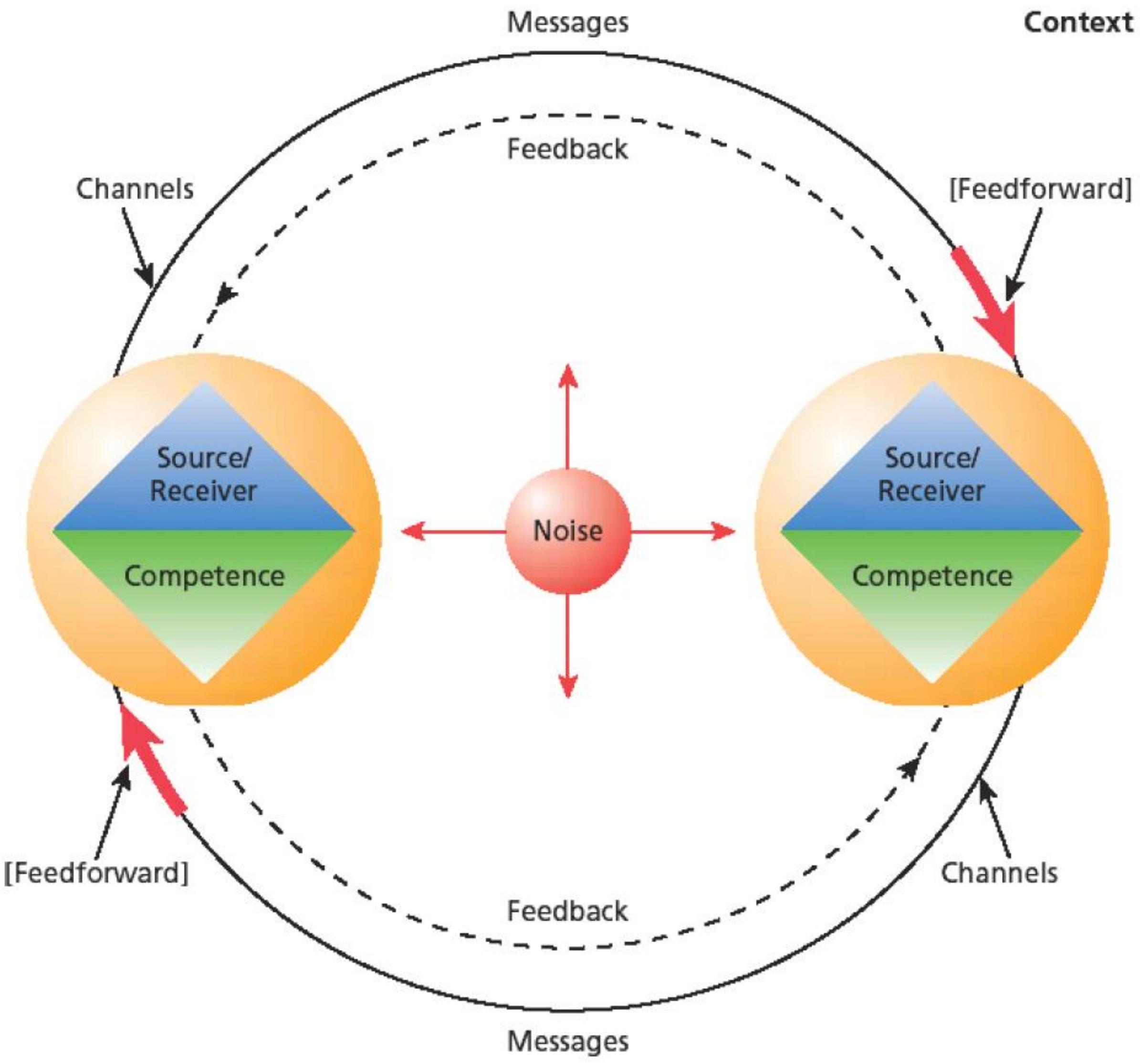


1. Source-Receiver
2. Channels
3. Messages
4. Feedback

Messages

Pu

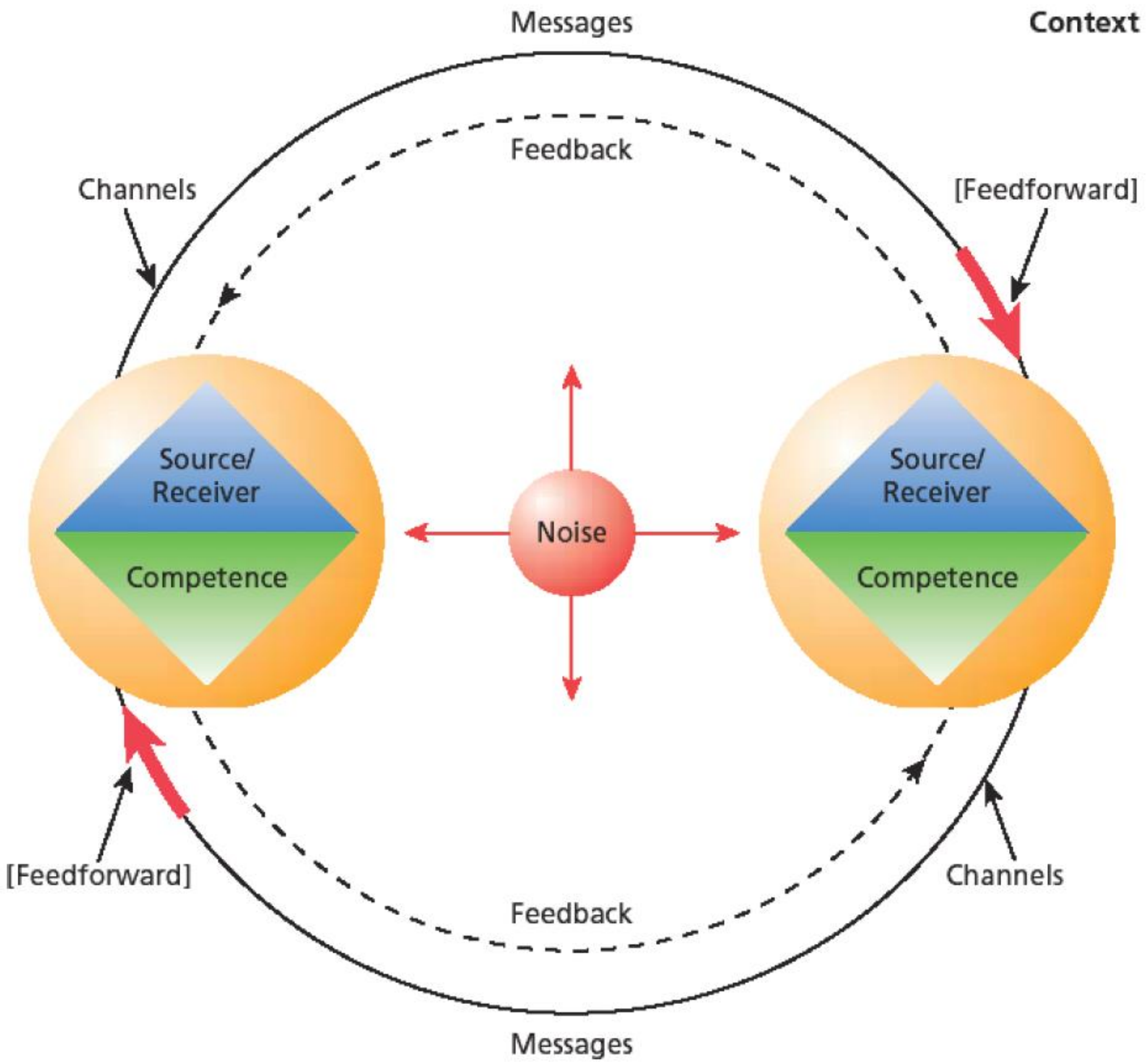
Elements of Interpersonal Communication



5. Types of Noise

- Physical
- Physiological
- Psychological
- Semantic

Elements of Interpersonal Communication



6. Context
   - Physical dimension
   - Temporal dimension
   - Social-psychological dimension
   - Cultural context
7. Competence

Diversity in Dyadic Interactions

Multimodal Affective Computing

Robots

Virtual Humans

Ubiquitous

Mobile

Online

Wide Applicability

Medical


Psychological signals


Suicide prevention

Autistic children

Education


Group learning analytics


Virtual Learning Peer


Public speaking training

Online

Opinion mining

Social influence


Negotiation outcomes

Phenomena

Pathology
- Distress
- Autism

Social
- Empathy
- Dominance

Emotion
- Sentiment
- Frustration

Cognitive
- Attention
- Curiosity

Personality
- Assertive
- Trusting

Multimodal Affective Computing

Behaviors

Verbal
- **Lexicon**
  - Spoken words
- **Pragmatics**
  - Discourse acts

Vocal
- **Prosody**
  - Voice quality
- **Vocal expressions**
  - Laughter, moans

Visual
- **Body language**
  - Head gestures
- **Facial expressions**
  - Smile, frowning

Computation

Statistical analysis
- Variance analysis
- Reliability tests

Prediction models
- Bayesian networks
- Markov fields

Deep learning
- Bayesian networks
- Markov fields

Phenomena

Pathology
- Distress
- Autism

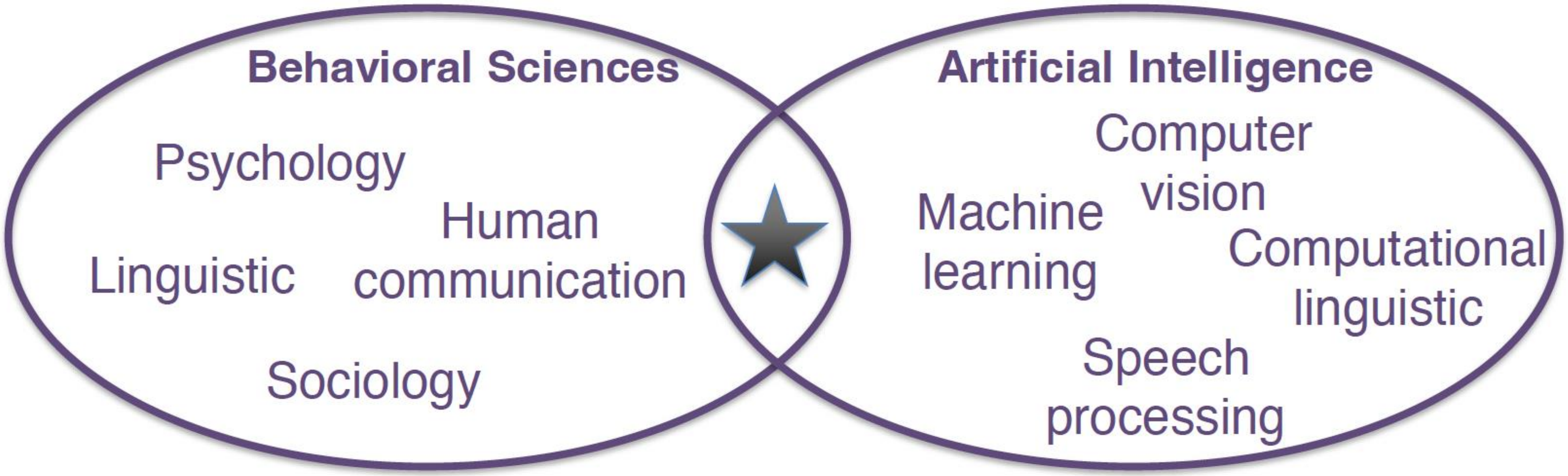
Social
- Empathy
- Dominance

Emotion
- Sentiment
- Frustration

Cognitive
- Attention
- Curiosity

Personality
- Assertive
- Trusting

Multimodal Affective Computing: Behavioral Sciences and Artificial Intelligence



New tools to study human factors with technologies

Brings new or understudied learning challenges

Datasets/ Applications/ Use Cases

- **Affective states**
- **Cognitive states**
- **Personality**
- **Pathology**
- **Social processes**

Sentiment Analysis Task

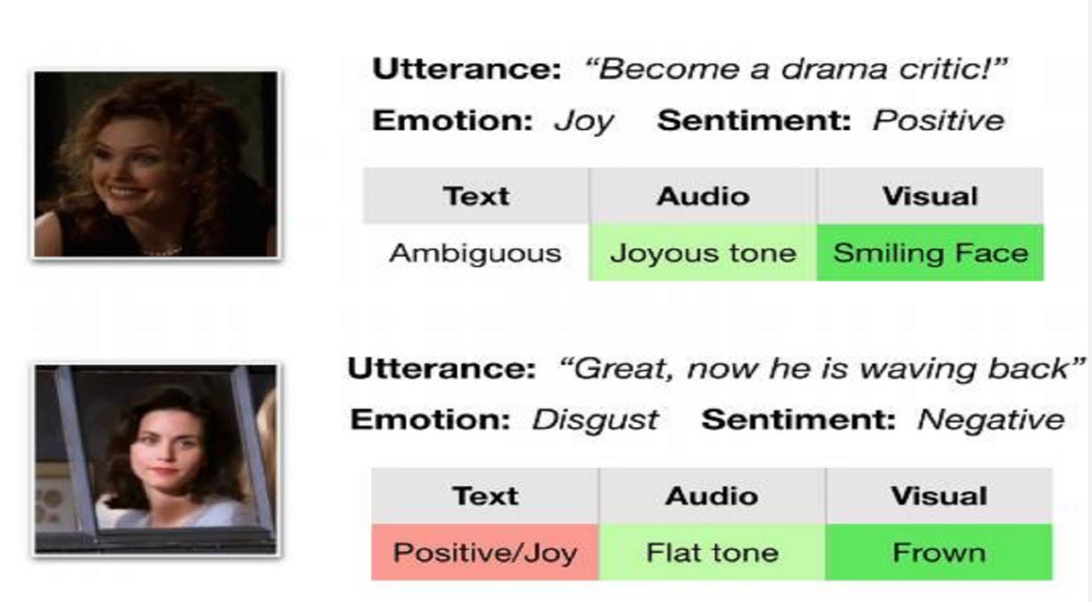


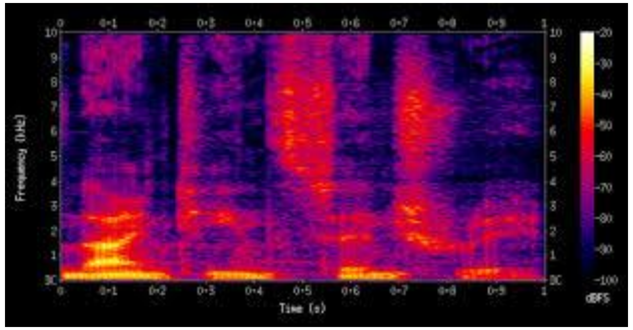
Fig 2: Sentiment Analysis Task

Study of MOSI Multimodal Dataset and Base Papers Implementation

- **MOSI**: Multimodal Opinion-level Sentiment Intensity dataset contains:
  - multimodal observations including transcribed speech (text) visual gestures/features (video) as well as automatic audio (audio),
  - sentiment intensity annotations and
  - alignment between words, visual and acoustic features

Figure 1: Example snapshots of videos from our new MOSI dataset.

- Data set split : Train: 1283, Validation: 229, Test: 686 [Speaker Independent]
- Features & Extraction tools: Visual – 47 (OpenFace) , Audio : 74 (Opensmile), Text: 300 (Glove embeddings)
- Additionally Low dimension data set with Text (300), FACET (20) and Audio (5) is provided as h5py (.h5) files
- 20 sequence length is widely used in research community compared to 50. 20 sequence length also have better accuracies



- **CMU Multimodal SDK:** CMU group provides Multimodal SDK to extract features from .CSD files to required to sequence length and feature dimensions
  - Features are extracted per frame, aligned with word duration and average out at word level to have 1 feature vector per word

Experimentation and Results .. Contd.

Comparison of individual modality performances with fusion

Modality	Accuracy	F1 Score
Language (L)	70.80	71.00
Audio (A)	55.83	42.00
Video (V)	50.87	44.00
Fusion (A, L, V)	73.03	75.00

Co-learning performance for different input modalities

Modalities at Test Time	Accuracy	F1 Score
Language (L)	73.00	76.00
Audio (A)	56.41	51.00
Video (V)	53.50	54.00
Language, Audio (A V)	72.47	76.00
Language, Video (L V)	72.40	76.00
Audio, Video (A V)	54.81	53.00

It is visible that language modality is the dominant modality among all three modalities.

MOSI: Fusion Implementation for Sentiment Regression

- Linear regression models for sentiment score prediction. Sentiment scores provided are between -3 to 3
- Sequence length is kept at 20 words with post padding/truncation
- Mean Absolute Error (MAE) is a metric measured with Mean Square Error as a loss metric.
- Pearson correlation coefficient (Corr) between actual & predicted sentiment score is calculated
- Early stopping is applied for val_loss, Model checkpoint saving on minimum validation loss
- Models are also tried for MAE as loss as some papers used that for MOSI

Data Split	Modality/ Fusion	TFN		Pol & Int		MFN		(MSE Loss)		(MAE Loss)	
		MAE	Corr	MAE	Corr	MAE	Corr	MAE	Corr	MAE	Corr
Test	Text	0.99	0.61	1.196	0.404	1.019	0.607	1.2302/ 1.134	0.4239/ 0.5645	1.2391	0.4030
	Audio	1.23	0.36	1.456	0.125	1.446	0.186	1.4399/ 1.4616	0.0853/ 0.214	1.4397	0.123
	Visual	1.13	0.48	1.442	0.092	1.446	0.155	1.4473/ 1.4515	0.0755/ 0.058	1.4442*	0.089*
	Late			1.179	0.471			1.2463/ 1.1624	0.425/ 0.532	1.2883	0.373
	Early			1.197	0.454			1.2500	0.3760	1.2559	0.3413
	TFN Fusion	0.87	0.70	1.186	0.448			1.1276	0.550		
	MFN Fusion					0.965	0.632				

Tensor Fusion Paper Results:

Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
$\text{TFN}_{language}$	74.8	75.6	38.5	0.99	0.61
TFN_{visual}	66.8	70.4	30.4	1.13	0.48
$\text{TFN}_{acoustic}$	65.1	67.3	27.5	1.23	0.36
$\text{TFN}_{bimodal}$	75.2	76.0	39.6	0.92	0.65
$\text{TFN}_{trimodal}$	74.5	75.0	38.9	0.93	0.65
$\text{TFN}_{notrimodal}$	75.3	76.2	39.7	0.919	0.66
TFN	**77.1**	**77.9**	**42.0**	**0.87**	**0.70**
TFN_{early}	75.2	76.2	39.0	0.96	0.63

Memory Fusion Paper Results:

Task	CMU-MOSI Sentiment				
Metric	BA	F1	MA(7)	MAE	r
SOTA2	$73.9^{\dagger}$	$74.0^{\diamond}$	$32.4^{\S}$	$1.023^{\S}$	$0.601^{\diamond}$
SOTA1	74.6^{*}	74.5^{*}	$33.2^{\diamond}$	$1.019^{\diamond}$	$0.622^{\S}$
MFN l	73.2	73.0	32.9	1.012	0.607
MFN a	53.1	47.5	15.0	1.446	0.186
MFN v	55.4	54.7	15.0	1.446	0.155
MFN (no Δ)	75.5	75.2	34.5	0.980	0.626
MFN (no mem)	76.5	76.5	30.8	0.998	0.582
MFN	**77.4**	**77.3**	**34.1**	**0.965**	**0.632**
Δ_{SOTA}	↑ **2.8**	↑ **2.8**	↑ **0.9**	↓ **0.054**	↑ **0.010**

* batch normalization at the input layer added as suggested in Tensor fusion paper implementation

- TFN: Tensor Fusion Network for Multimodal Sentiment Analysis (MSE Loss used for regression)
- MFN: Memory Fusion Network for Multi-View Sequential Learning
- Polarity and Intensity: the Two Aspects of Sentiment Analysis, Leimin Tian & et. al. In this paper, individual models are run as a part of multi task learning (MAE loss is used for regression)
- MFN and Polarity & Intensity uses the same data split. Data split is not mentioned explicitly for TFN but must be same

MOSEI: Fusion Implementation for Emotion classification

Data set size:

- **Train: 15290**
- **Validation: 2291**
- **Test: 4832**

Data set dimensions

- **Text: 300 (Gloves)**
- **Audio: 74 (Covarep)**
- **Video: 35 (Facet)**

- **6 emotions** are Anger, Disgust, Fear, Happy, Sad and Surprise with Text, Audio, Video modalities
- **Labels**: Some videos have multiple emotions making dataset as - Multi-label, Multiclass.
  - Label values include intensity as well as polarity on Likert scale of 0-3. [0: no evidence of x, 1: weakly x, 2: x, 3: highly x].
  - Average across ratings provided 3 annotators is taken to arrive at final intensity. This results in intermediate values of intensity.
  - Some videos are not having any emotion, those along with corresponding train/test examples are dropped

Emotions + ve example distribution used for binary classification after dropping null value labels

Emotion	Train	% Pos	Valid	% Pos	Test	% Pos
Anger	3443	26%	427	22%	971	24%
Disgust	2720	21%	352	18%	922	22%
Fear	1319	10%	186	10%	332	8%
Happy	4900	38%	633	33%	1576	38%
Sad	3906	30%	576	30%	1334	33%
Surprise	1562	12%	201	10%	479	12%
Total Size	13047		1946		4098	

Imbalance dataset with examples for fear and Surprise are less compared to other emotions

MOSEI: Fusion Implementation for Emotion classification .... Contd.

Implementation using One v/s Rest (Sigmoid, with 0.5 threshold for binary classification]

Modality/ Fusion	Anger				Disgust				Fear			
	WA	WF1	Acc.	F1	WA	WF1	Acc.	F1	WA	WF1	Acc	F1
Text	56.3/ 54.79	69.9/ 70.08	/73.54	/26.0	56.71	72.96	76.40	28.52	50.53/ 51.33	88.19/ 88.12	91.87/ 91.06	2.0/ 6.63
Video	52.66	69.00	76.03	14	50.0	67.67	77.50	0.0	50.88/ 51.36	88.26/ 88.26	91.75/ 91.38	4.0/ 6.0
Audio	52	68			50.0	67.67	77.50	0.0	50/ 49.98	88.01/ 88.0	91.89/ 91.87	0.0/ 0.0
Late Fusion	61	73.3			65.0	75.85	76.00	46.0	50/ 50.0	88.0/ 88.02	91.89/ 91.89	0/ 0
TFN Fusion	50	66										

Modality/Fusion	Happy				Sad				Surprise			
	WA	WF1	Acc	F1	WA	WF1	Acc	F1	WA	WF1	Acc	F1
Text	61.36/ 62.51	63.84/ 63.86	64.17/ 63.56	72.0/ 69.0	55.14	62.47	66.25	31	50.60/ 54.00/ 53.96	82.43/ 61.4/ 61.32	86.01/ 64.9/ 66.12	7.0/ 30.0/ 26.87
Video	67.51	68.32	68.00	73.0	50.49/ 51.98	55.58/ 58.74	66.98/ 66.56	0.0/ 17.0	51.68	58.80	65.27	19.28
Audio	64.96	67.28	67.59	74.0	50.00/ 50.98	54.33/ 56.32	67.44/ 67.56	0.0/ 6.0	50/ 51.76	82.82/ 57.66	88.31/ 76.76	0.0/ 11.00
Late Fusion	67.91	70.13	70.42	77.0	57.68/ 57.13	64.13/ 63.57	65.49/ 64.86	40.0/ 39.0	50/ 57.52	82.82/ 64.31	88.31/ 66.42	0.0/ 38.0
TFN Fusion												

Dataset	MOSEI Emotions											
Task	Anger		Disgust		Fear		Happy		Sad		Surprise	
Metric	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1
LANGUAGE												
SOTA2	$56.0^{\cup}$	$71.0^{\times}$	$59.0^{\S}$	$67.1^{\triangleright}$	$56.2^{\S}$	$79.7^{\S}$	$53.0^{\triangleright}$	$44.1^{\triangleright}$	$53.8^{\wr}$	$49.9^{\wr}$	$53.2^{\times}$	$70.0^{\triangleright}$
SOTA1	$56.6^{\wr}$	$71.8^{\bullet}$	$64.0^{\triangleright}$	$72.6^{\bullet}$	$58.8^{\times}$	$89.8^{\bullet}$	$54.0^{\S}$	$47.0^{\S}$	$54.0^{\S}$	$61.2^{\bullet}$	$\mathbf{54.3}^{\triangleright}$	$85.3^{\bullet}$
VISUAL												
SOTA2	$54.4^{\wr}$	$64.6^{\S}$	$54.4^{\heartsuit}$	$71.5^{\triangleleft}$	$51.3^{\S}$	$78.4^{\S}$	$53.4^{\wr}$	$40.8^{\S}$	$54.3^{\triangleright}$	$60.8^{\bullet}$	$51.3^{\triangleright}$	84.2
SOTA1	$60.0^{\S}$	$71.0^{\bullet}$	$60.3^{\wr}$	$72.4^{\bullet}$	$\mathbf{64.2}^{\heartsuit}$	$89.8^{\bullet}$	$57.4^{\bullet}$	$49.3^{\bullet}$	$57.7^{\S}$	$61.5^{\triangleleft}$	$51.8^{\S}$	$85.4^{\bullet}$
ACOUSTIC												
SOTA2	$55.5^{\triangleleft}$	$51.8^{\triangle}$	$58.9^{\triangleright}$	$72.4^{\bullet}$	$58.5^{\triangleright}$	$89.8^{\bullet}$	$57.2^{\cap}$	$55.5^{\cap}$	$58.9^{\triangleleft}$	$65.9^{\triangleleft}$	$52.2^{\heartsuit}$	$83.6^{\cap}$
SOTA1	$56.4^{\triangle}$	$71.9^{\bullet}$	$60.9^{\S}$	$72.4^{\bullet}$	$62.7^{\S}$	$89.8^{\triangleleft}$	$61.5^{\S}$	$61.4^{\S}$	$\mathbf{62.0}^{\cap}$	$\mathbf{69.2}^{\cap}$	$\mathbf{54.3}^{\triangleleft}$	$85.4^{\bullet}$
MULTIMODAL												
SOTA2	$56.0^{\diamond}$	$71.4^{\flat}$	$65.2^{\#}$	$71.4^{\#}$	$56.7^{\S}$	$\mathbf{89.9}^{\#}$	$57.8^{\S}$	66.6^{*}	58.9^{*}	$60.8^{\#}$	52.2^{*}	$85.4^{\bullet}$
SOTA1	60.5^{*}	$72.0^{\bullet}$	$67.0^{\flat}$	$73.2^{\bullet}$	$60.0^{\heartsuit}$	$\mathbf{89.9}^{\bullet}$	$\mathbf{66.5}^{*}$	$\mathbf{71.0}^{\blacksquare}$	$59.2^{\S}$	$61.8^{\bullet}$	$53.3^{\#}$	$85.4^{\#}$
GMFN	**62.6**	**72.8**	**69.1**	**76.6**	62.0	**89.9**	**66.3**	**66.3**	60.4	66.9	53.7	**85.5**
Δ_{SOTA}	↑ 2.1	↑ 0.8	↑ 2.1	↑ 3.4	↓ 2.2	0.0	↑ 4.8	↑ 4.9	↓ 1.6	↓ 2.3	↓ 0.6	↑ 0.1

Acc & F1 are binary accuracy and F1 scores, WA & WF1 are weighted accuracy and F1 scores

Affective Computing Implementation on MOSI (Multimodal Sentiment) dataset

Multimodal Co-learning: Base paper Implementation – MOSI Aligned Data

- **MOSI dataset – Aligned (Paired/Parallel Data)**

Measure	Text	Audio	Video
Validation MAE	0.9863	1.4017	1.3751
Testing MAE	1.0123	1.4120	1.4486
Co-relation Coeff.	0.6319	0.2342	0.1621
Binary Accuracy	76.67	55.53	49.12
F1 Score	False 0.79/ True – 0.73	False - 0.43/ True – 0.64	False - 0.27/ True – 0.61

Paper: Multimodal Co-learning, Amir Zadeh, Paul L, L. P. Morency, CMU Group, 2020

Category	Same-dataset					Cross-dataset	Cross-domain	
Dataset	CMU-MOSI					SST	IMDB	MDSD
Task	Sentiment					Sentiment	Sentiment	Sentiment
Metric	A^2	F1	A^7	MAE	Corr	A^5	A^2	A^5
(-/S/B/SB)-RNN	70.1	70.1	27.3	1.131	0.541	28.6	60.9	17.5
(-/S/B/SB)-LSTM	77.1	76.9	33.4	0.979	0.636	31.5	75.3	18.9
(-/S/B/SB)-GRU	75.8	75.7	33.4	0.974	0.635	31.1	75.0	19.1
CNN	67.8	67.9	27.0	1.166	0.500	27.9	69.2	13.8
CNN-LSTM	74.2	74.0	29.7	1.092	0.553	28.8	65.5	14.9
MARN-L	75.5	75.6	33.2	1.011	0.626	29.1	75.7	17.5
MFN-L	76.5	76.4	34.5	0.982	0.628	31.3	73.1	19.0
MFN(MCl)	**78.0**	**77.9**	**35.3**	**0.968**	**0.641**	**31.9**	**76.1**	**21.7**

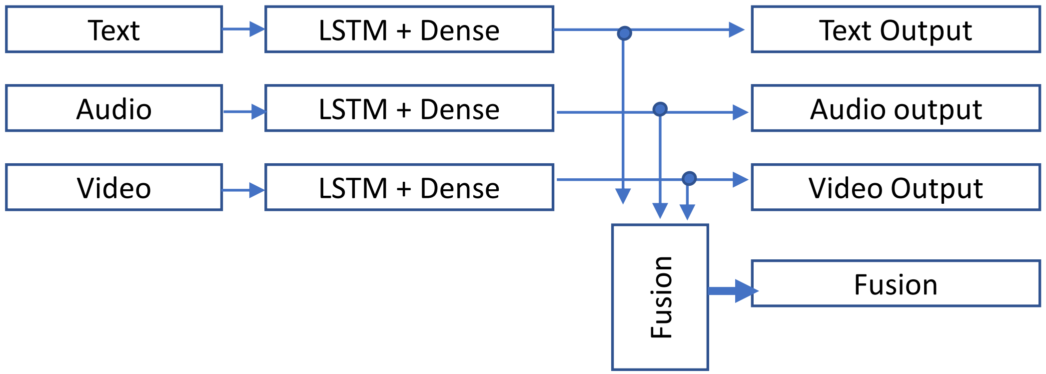
Table 1: Sentiment prediction experiments comparing our proposed MFN model with baseline models. SST, IMDB and MDSD are language-only datasets. All models are trained on CMU-MOSI train set and evaluated on CMU-MOSI, SST, IMDB and MDSD language-only test sets.

Measure	Late Fusion	Text only at Test Time
Validation MAE	1.01653	
Testing MAE	1.0218	1.0186
Co-relation Coeff.	0.6239	0.6177
Binary Accuracy	76.09	75.80
F1 Score	False - 0.78/ True – 0.73	False - 0.78/ True – 0.73

Measure	Intermediate Fusion	Text only at Test Time
Validation MAE	0.9960	
Testing MAE	1.02901	1.0208
Co-relation Coeff.	0.6420	0.6177
Binary Accuracy	74.80	75.80
F1 Score	False - 0.76/True – 0.73	False - 0.78/True – 0.73

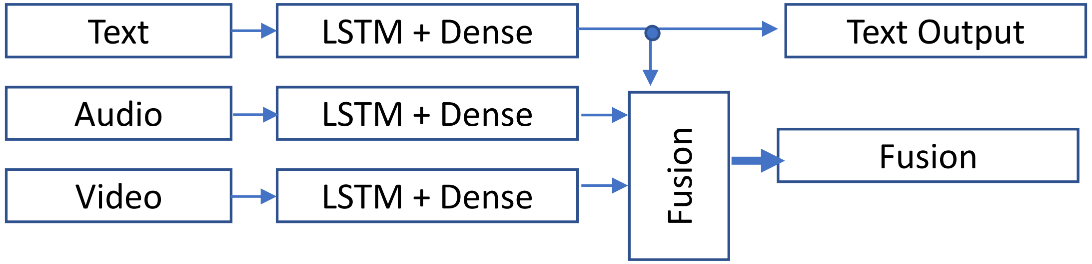
Multimodal Co-learning: Base paper – MOSI Multi-task Aligned data

- **MOSI dataset – Aligned (Paired/Parallel Data)**
- **Multi-Task like model**
- **Multi-task like model with audio and video in un-supervised form**



*a) Multi task like model with Intermediate Fusion **

Measure	Multi_task like model with Intermediate Fusion
Training MAE	Audio:1.2651, Video:1.172, Text: 0.675, Fusion: 0.7190
Validation MAE	Audio:1.3846, Video:1.3826, Text: 1.004, Fusion: 1.018
Testing MAE	Audio:1.4006, Video:1.4799, Text: 1.020, Fusion: 1.038
Co-relation Coeff. - Testing	A: 0.2473, V: 0.0777, T: 0.6345 F:0.6313
Binary Accuracy - Testing	Audio:55.10, Video:49.85, Text: 77.84, Fusion: 76.82
F1 Score - Testing	A: F:0.43/0.63, V: 0.35/0.59, T:0.80/0.74 F: False - 0.79/True – 0.74
Text modality at test time	Audio:1.4558, Video:1.5540, Text: 1.020 , Fusion: 1.0625
Co-relation Coeff. - Testing	T: 0.6345 F:0.6346
Binary Accuracy - Testing	Text: 77.84 Fusion: 76.38
F1 Score - Testing	T:0.80/0.74 F: False - 0.78/True – 0.74



b) Multi task like model with Intermediate Fusion with text & fusion as output i.e. labels of audio and video not used mimic of un-supervised however data is aligned here

Measure	Text	Audio	Video
Validation MAE	0.9863	1.4017	1.3751
Testing MAE	1.0123	1.4120	1.4486
Co-relation Coeff.	0.6319	0.2342	0.1621
Binary Accuracy	76.67	55.53	49.12
F1 Score	False 0.79/ True – 0.73	False - 0.43/ True – 0.64	False - 0.27/ True – 0.61

*__Paper__: Multimodal Multitask Emotion Recognition using Images, Texts and Tags, Mathieu Pagé Fortin, 2019

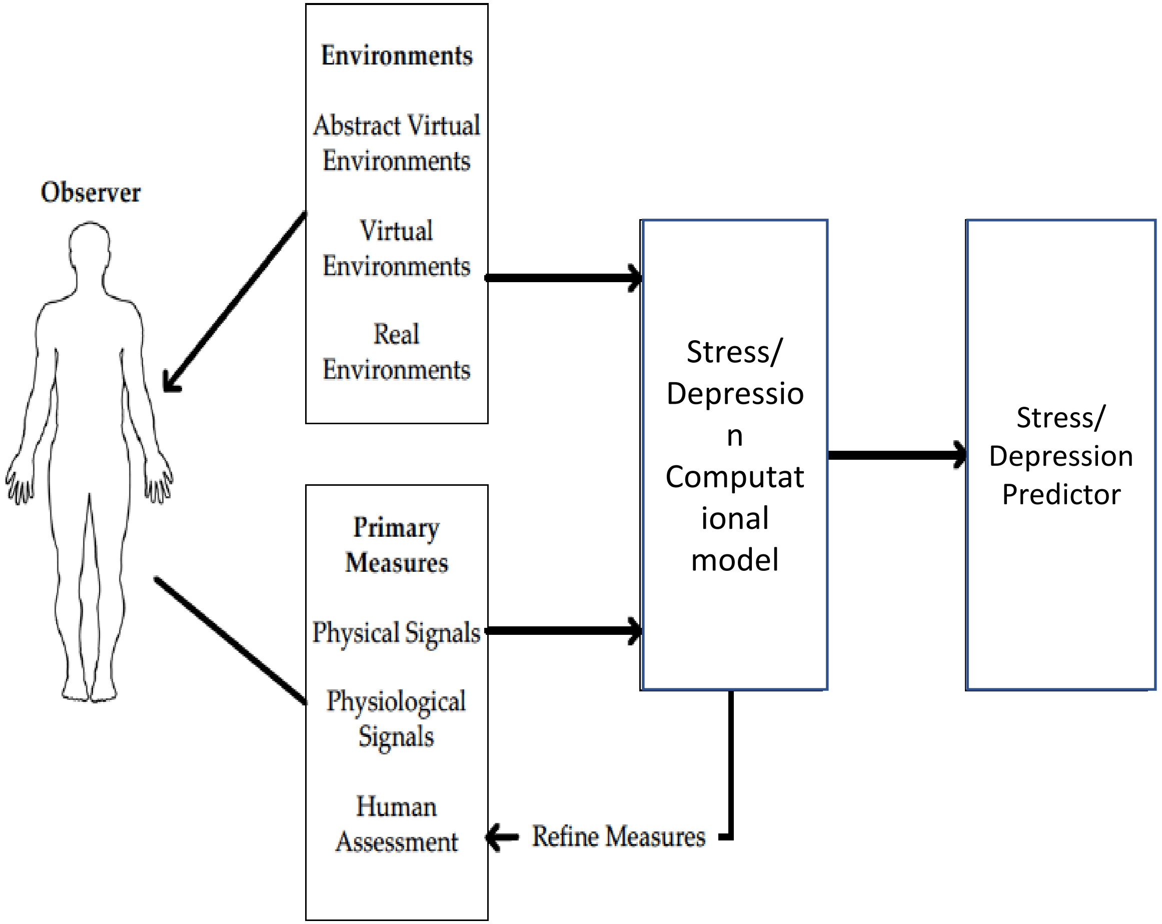
Affective computing:
Stress or Depression Detection

Affective Computing and Multimodal Data Sources

- Affective computing focuses on human sentiments, emotions, stress, depression, engagement, personality assessment, etc.
- Multimodal data such as audio, video, language, heart rates, physiological signals, physical signals, medical signals are available as sources for affective computing and two or more data sources are fused to predict the underlined task.
- Recently, public datasets have become available for sentiment analysis, emotion analysis, stress detection, depression conditions, affective content, personality assessment, EEG, ECG, heart rate, body postures, human-computer interactions.

Sensor Types	• Invasive Sensors • Non-invasive and less or non-obtrusive to natural movement
Physiological Sensors	• Electroencephalogram (EEG) • Electrocardiogram (ECG) • Galvanic skin response (GSR)/ electrodermal activity (EDA) • Blood Volume Pulse (BVP) • Heart rate (HRV) • Breathing patterns • Bio-markers • Medical Tests (MRI, CT etc.)
Physical Sensors	• Eye Gaze, Pupil Dilation • Facial Videos in visible and thermal spectrum • Body postures • Gestures • Voice modulation • Language used
Annotation Process	• Survey Questions (self-assessment) • Interview by expert • Expert assessments • Combinations of above

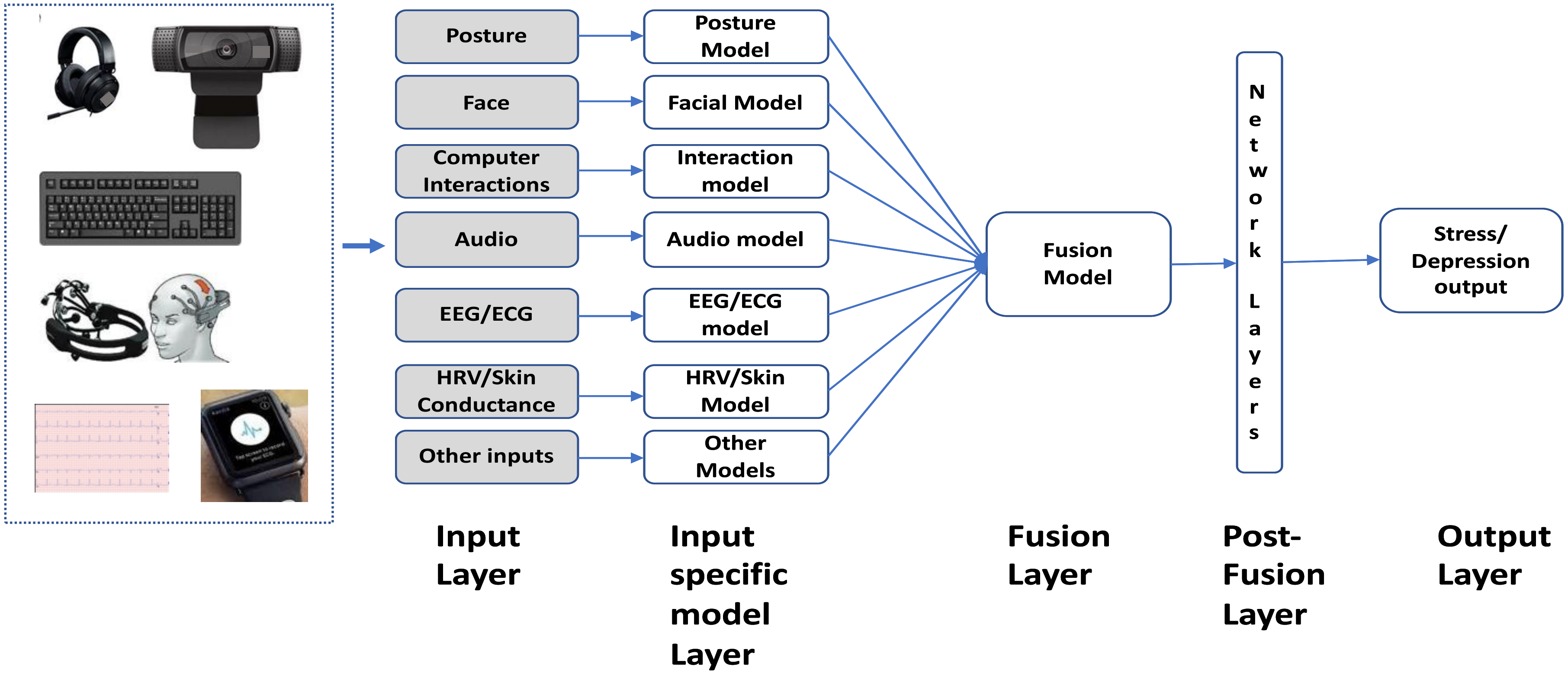
Stress and/or Depression Detection using Multimodal Deep Learning



Simulating Environments	• Abstract virtual like Text
	• Virtual environment like movies, videos, games, emotional content, images, music
	• Conversation with virtual agents
	• Real environments –
	  • interactions with people, leaders
	  • mentally difficult work, mediation
	  • time pressure, interruptions in work,
	  • interview, public speaking,
	• Physical stress – like exercise, long duration work

Reactions	Individual reactions depends on: body conditions, age, gender, experience, mental state, culture
Types of Individuals studied	• Persons studied for stressful conditions • First response – firefighters, disaster management • Arm forces, police, allied personalities • Pilots in flights • Automotive /car Drivers • Medical health professionals and workers • Knowledge workers

Stress or Depression Detection: Representative Model and Datasets



- Continue study of co-learning on SWELL dataset using models and performance baseline from SCAAI group paper. modalities – Key logs, Facial Expressions, Kinect and HRV

- Explore WESAD dataset and study co-learning aspects for stress detection with Physiological (BVP, ECG, EDA, EMG, RESP, & TEMP) & motion (ACC) modalities

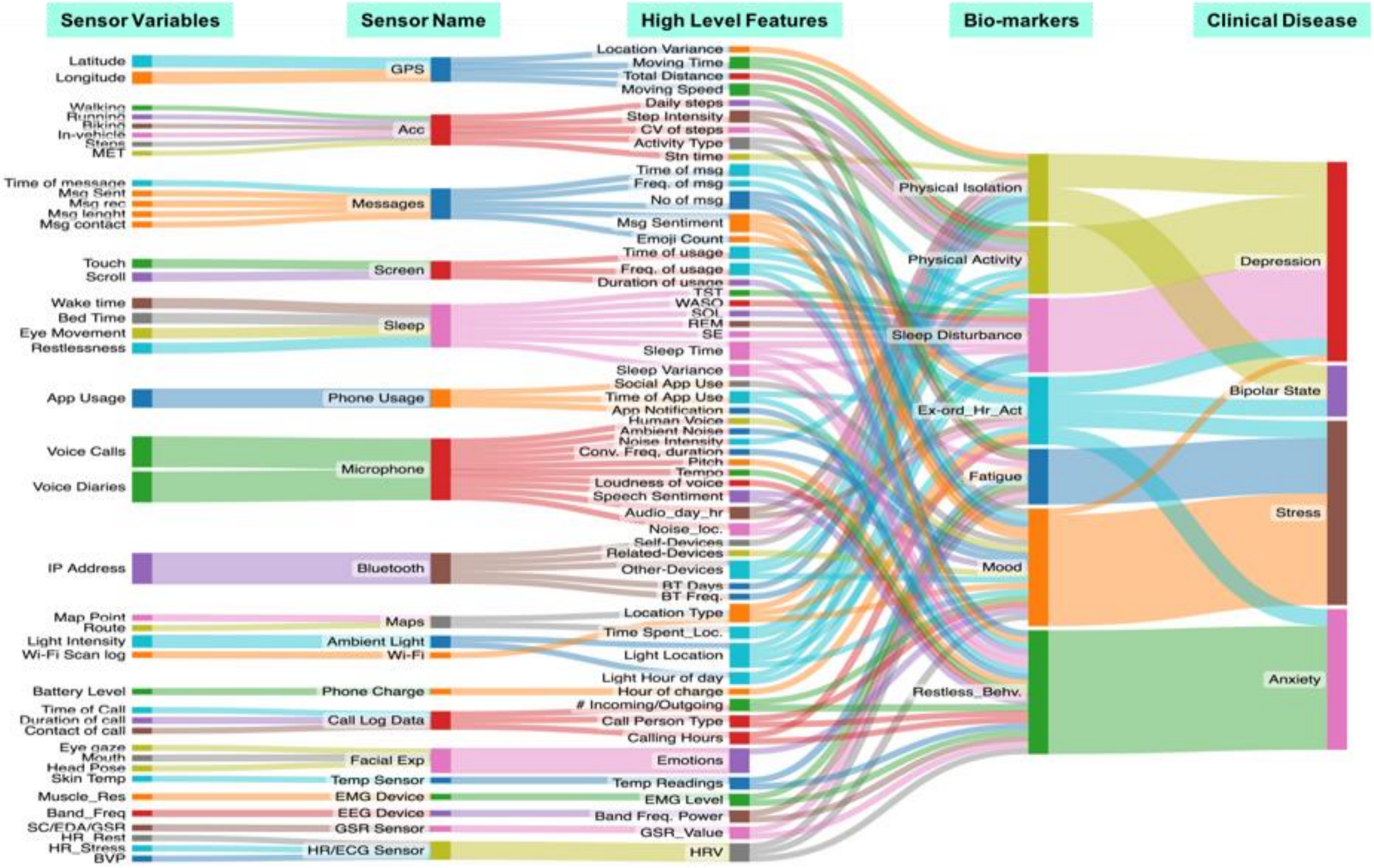
Depression Detection Results on Depressjon & Modma Dataset

The aim is to demonstrate how machine learning can be used for depression detection using the features acquired from digital biomarkers with the help of available secondary datasets.

No	Dataset Name	Model Name	Data pre-processing, feature extraction & model details	Accuracy	Precision, recall, F1 Score
1	Depresjon	Decision Tree	Hourly average of data of motor activity using actigraph	62%	0.44, 0.36, 0.40
2	Depresjon	Random Forest	Data segmentation for 6 hrs. and mean, median and standard deviation for each 6 hrs. segment is taken as features.	80%	0.75, 0.62, 0.66
3	Depresjon	1D CNN	Minute level data	73%	0.81, 0.81, 0.81
4	MODMA	KNN	EEG data. A feature space - mean, median, max, min, amplitude of power signal and alpha, beta, delta, theta waves along with non-linear, linear, and phase lag index	64%	0.43, 1.0, 0.6
6	MODMA	SVM		64%	0.43, 1.0, 0.6
7	Modma	Deep learning	Audio + EEG multimodal deep learning with fusion work is **[in progress]**		

- **Digital Biomarkers Mapping**
- **Multimodal Fusion for Depression Detection**

Walambe R., Rahate A., Ghate R., Tasgaonkar V., Shetty M., Saharan S., Kalnad N., Kotecha K., Depression Detection using Machine Learning based on Digital Biomarkers: Review, Methods, Datasets, Implementations and Future Directions (*under preparation)*



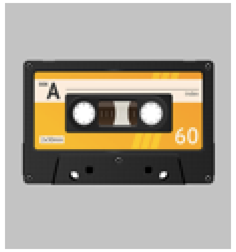
Comprehensive Digital Biomarkers Mapping

Experimentation and Results

MODMA Dataset

(a Multi-modal Open Dataset for Mental-disorder Analysis)

EEG



Audio

Fig. 13: MODMA Dataset

Study items	**Detailed Information**
No. of Participants	24 depressed (condition), 29 healthy (control)
Data Collection Procedure	The 128 electrodes EEG signals of 53 subjects were recorded as both in resting state and under stimulation; the 3-electrode EEG signals of 55 subjects in resting state; the audio data of 52 subjects were recorded during interviewing, reading, and picture description.
Dataset Organization	128 electrode EEG data of 24 depressed and 29 healthy controls, the 3 electrode EEG data of 26 depressed and 29 healthy controls and audio data of 52 participants.
Modalities	EEG data, Audio Data
Rating Scales	DSM-IV (Do, 2011) and PHQ-9 (Kroenke et al., 2001) score > 5 for depressed conditions

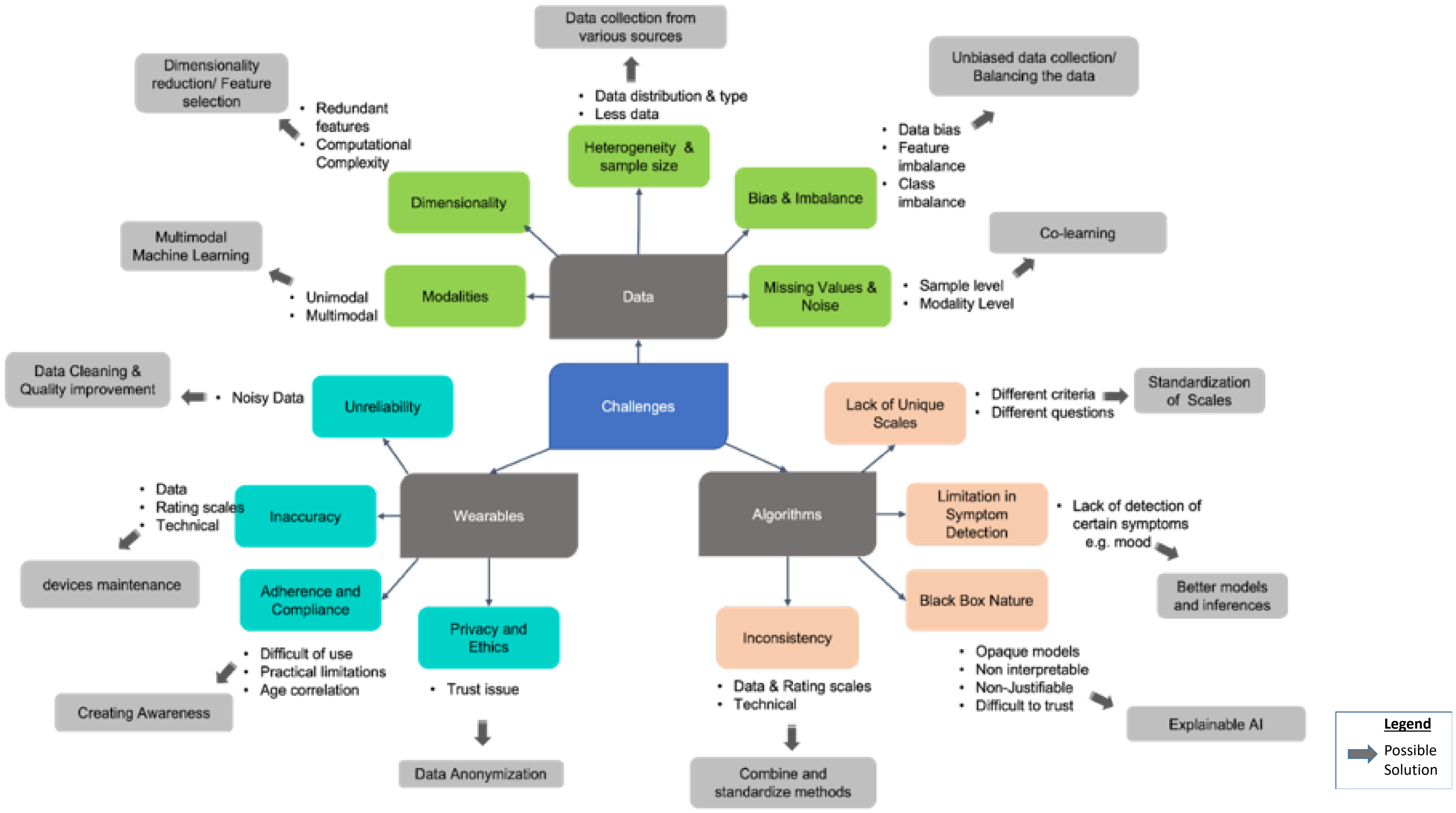
► Cai, H., et. al. (2020). MODMA dataset: a Multi-modal Open Dataset for Mental-disorder Analysis. https://doi.org/10.1038/s41597-022-01211-x

Experimentation and Results .. Contd.

Modality	Model Name	Accuracy	Precision, Recall, F1 Score
EEG	Logistic Regression	73.00%	0.5, 1.0, 0.67
Audio	DNN	74.10%	Class 0: 1.00, 0.58, 0.74 Class 1: 0.58, 1.00, 0.74
Early Fusion (Audio + EEG)	DNN	75.09%	Class 0: 0.71, 0.71, 0.71 Class 1: 0.78, 0.78, 0.78

- 16 EEG channels - mean, median, max, min, and amplitude, etc. as EEG features.
- Mel-frequency cepstral coefficients (MFCCs) as audio features.
- Participant-level features are obtained.
- For multimodal fusion, audio and EEG modality features are concatenated to implement an early fusion model.

Challenges & Recommendations for Depression Detection



Datasets/ Applications/ Use Cases

- **Affective states**
- **Cognitive states**
- **Personality**
- **Pathology**
- **Social processes**

Please refer to PDF shared for multimodal affective datasets

Source: Carnegie Mellon University