ELEC-E5531 - Speech and Language Processing Seminar V D

Prediction of Depression Severity Based on the Prosodic and Semantic Features with Bidirectional LSTM and Time Distributed CNN

Juho Ylä-Outinen & Wong Wei Kang

Aalto University
School of Electrical
Engineering

Table of contents

Aalto University
School of Electrical
Engineering

Introduction: Depression

- WHO: over 264 million people suffering in 2017
- Diagnosis - semi-structured interviews (DSM) and questionnaires (PHQ, BDI)
- PHQ-8 measures fatigue and anxiety
- Observation: depression affects prosody and linguistic content
  - Speech: depressive qualities
  - Content: less variability

Task: Depression Severity

- Multimodal approach: audio and text features
- Features extracted frame-by-frame → majority voting
- 5 classes: healthy, mild, moderate, moderately severe, severe
- Metric: F1 score on sequence level > patient level

The Dataset (DAIC-WOZ)

- 189 recorded interviews with sound and transcripts
- Virtual interviewer controlled by human
- Duration 5 to 20 min
- Difference in control and experiment groups
- Training set class distribution: 16, 37, 30, 15, 9

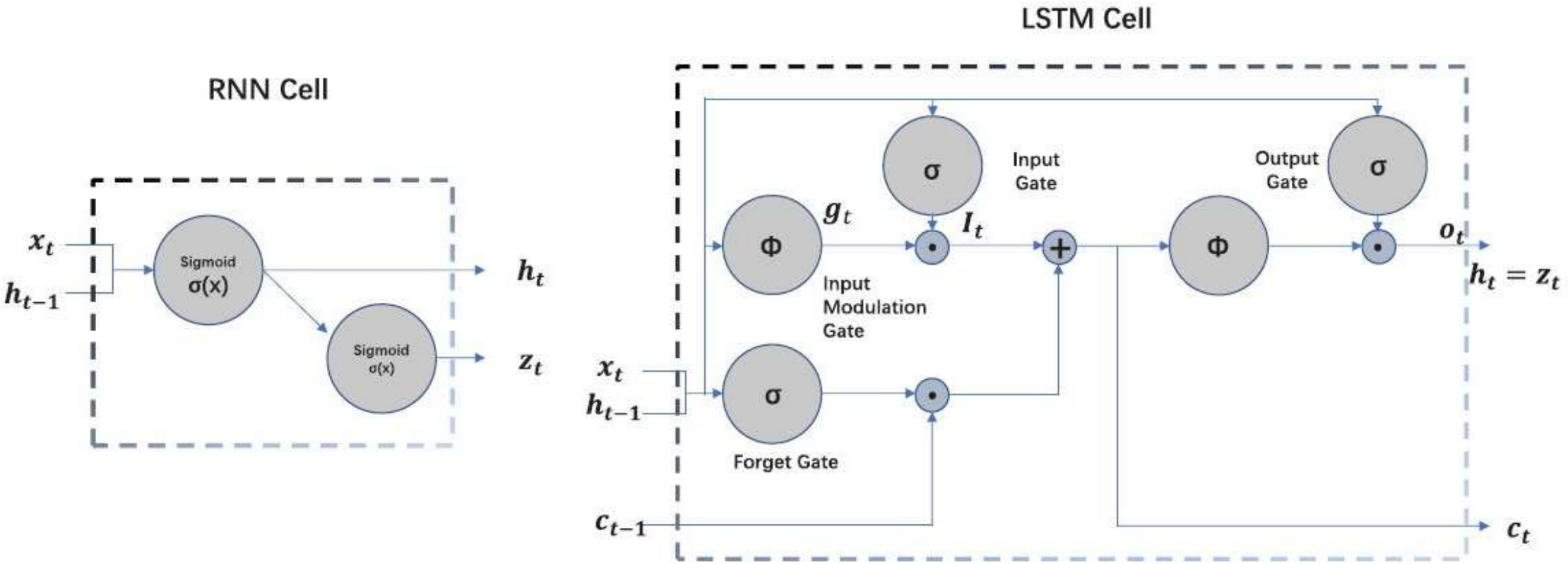
The Dataset (DAIC-WOZ)

	Training set		Validation set		Test set	
ID	303, 304, 305, 310, 312, 313, 315, 316, 317, 318, 319, 320, 321, 322, 324, 325, 326, 327, 328, 330, 333, 336, 338, 339, 340, 341, 343, 344, 345, 347, 348, 350, 351, 352, 353, 355, 356, 357, 358, 360, 362, 363, 364, 366, 368, 369, 370, 371, 372, 374, 375, 376, 379, 380, 383, 385, 386, 391, 392, 393, 397, 400, 401, 402, 409, 412, 414, 415, 416, 419, 423, 425, 426, 427, 428, 429, 430, 433, 434, 437, 441, 443, 444, 445, 446, 447, 448, 449, 454, 455, 456, 457, 459, 463, 464, 468, 471, 473, 474, 475, 478, 479, 485, 486, 487, 488, 491		302, 307, 331, 335, 346, 367, 377, 381, 382, 388, 389, 390, 395, 403, 404, 406, 413, 417, 418, 420, 422, 436, 439, 440, 451, 458, 472, 476, 477, 482, 483, 484, 489, 490, 492		300, 301, 306, 308, 309, 311, 314, 323, 329, 332, 334, 337, 349, 354, 359, 361, 365, 373, 378, 384, 387, 396, 399, 405, 407, 408, 410, 411, 421, 424, 431, 432, 435, 438, 442, 450, 452, 453, 461, 462, 465, 466, 467, 469, 470, 480, 481	
Dataset profile for depression level classifcation						
	Female	Male	Female	Male	Female	Male
#Healthy	7	9	2	3	3	2
#Mild	12	25	7	6	9	11
#Moderate	10	20	5	2	7	3
#Moderately severe	10	5	2	2	1	4
#Severe	5	4	3	3	4	3
Dataset profile for depression detection						
#Subjects w/o significant symptom (PHQ-8≤10)	27	49	12	11	17	16
#Subjects w/ significant sympotom (PHQ-8>10)	17	14	7	5	7	7

Table: “Gender distribution over all groups and dataset partitions”

	healthy	mild	moderate	mod.sev.	severe
train	16	37	30	15	9
validation	5	13	7	4	6
test	5	20	10	5	7

Methods - LSTM-RNN



Picture: A schematic of an RNN cell and an LSTM cell used in the paper.

Methods - Attention mechanism

Attention mechanism

- Encoder-decoder compresses the data into a vector.
- Attention mechanism alleviates the burden of compression.
- Does this by selecting a subset of encoded vectors.

Audio Features

- **Audio Features extracted by “A Cooperative Voice Analysis Repository for Speech Technologies (COVAREP)”**
- **Categories: Glottal flow features, voice quality features and spectral features**
  - *Glottal flow features: NAQ, QOQ, H1-H2, PSP, MDQ, Peak slope, Rd*
  - *Voice quality features: F0, VUV*
  - *Spectral features : MCEP, HMPDM, HMPDD*
- **Final feature dimension of 74 for each frame**

Audio Features

Feature Category	Feature Name	Description
Glottal Flow Features	Normalized Amplitude Quotient (NAQ)	Measures voice sharpness, indicating vocal fold tension.
	Quasi Open Quotient (QOQ)	Reflects time vocal folds are open, indicating speech breathiness.
	Rd (Open quotient derivative)	Derivative measure related to vocal folds' opening speed, showing speech dynamism.
	H1-H2 (Harmonic differences)	Amplitude difference between first two harmonics, revealing voice quality.
	Maxima Dispersion Quotient (MDQ)	Indicates breathy voice quality by measuring dispersion of glottal pulse maxima.
	Parabolic Spectral Parameter (PSP)	Quantifies spectral decay, offering clues about vocal tract shape.
	Peak Slope	Evaluates the spectral peak slope, related to vocal effort and stress.

Audio Features

Feature Category	Feature Name	Description
Voice Quality Features	Fundamental Frequency (F0)	Basic pitch of voice, varies with emotional states and stress.
	Voice Unvoiced (VUV) Decision	Determines voiced vs. unvoiced segments, affecting emotional tone.
Spectral Features	Mel-Frequency Cepstral Coefficients (MCEP)	Reflects short-term power spectrum, related to emotional timbre.
	Harmonic Model and Phase Distortion Mean (HMPDM)	Analyzes voice harmonic structure and phase distortion for quality changes.
	Harmonic Model and Phase Distortion Deviation (HMPDD)	Measures deviations in harmonic model and phase distortion, indicating voice quality variations.

Audio Model

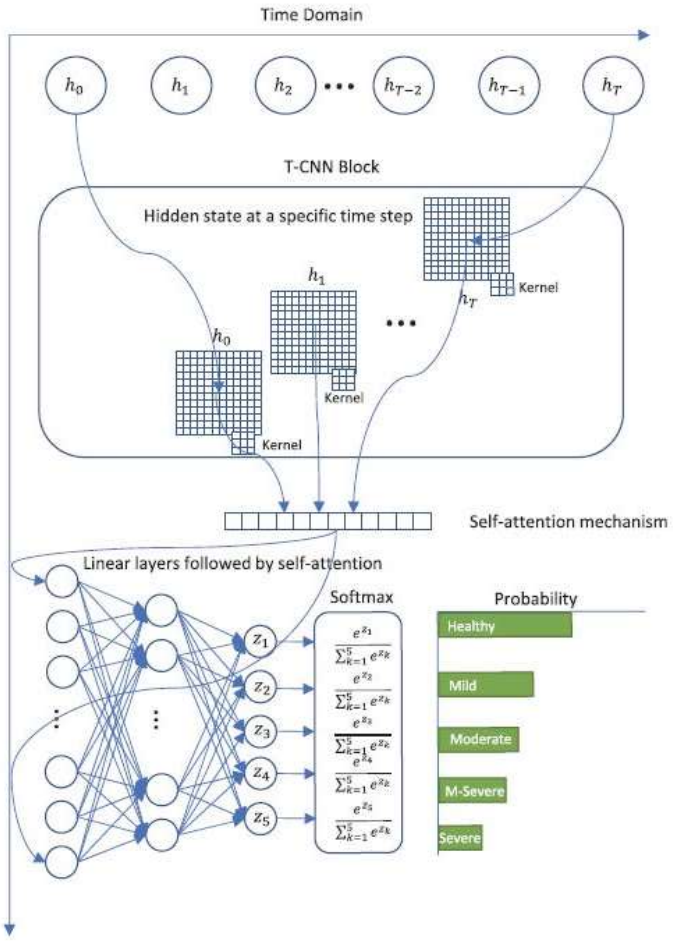


Fig. 5: The structure of the T-CNN model and the following linear neural network.

- **Hybrid of Bi-LSTM and T-CNN Model**
  - Bi-LSTM: Enhances the model's understanding of the temporal sequence.
  - T-CNN: Applies convolution across time, capturing spatial and temporal relationships.
  - Bi-LSTM encoder followed by 5-block T-CNN with 3 layers
  - Global-Average Pooling layer and 2 linear layers before the final output layer

Text Features and Model

- **Feature Preprocessing:**
  - Text normalization, stop word removal, and lemmatization using NLTK
  - Use of sliding window technique for segmenting text into manageable pieces
  - GloVe (Global Vectors for Word Representation) for dense word representations, followed by concatenation into sentence embeddings
- **Model:**
  - Utilization of Bi-LSTM to capture contextual dependencies within text
  - Attention mechanism applied adaptively to select depression-sensitive hidden states
  - 2 linear layers and a final output layer

Fused Text-Audio Joint Model

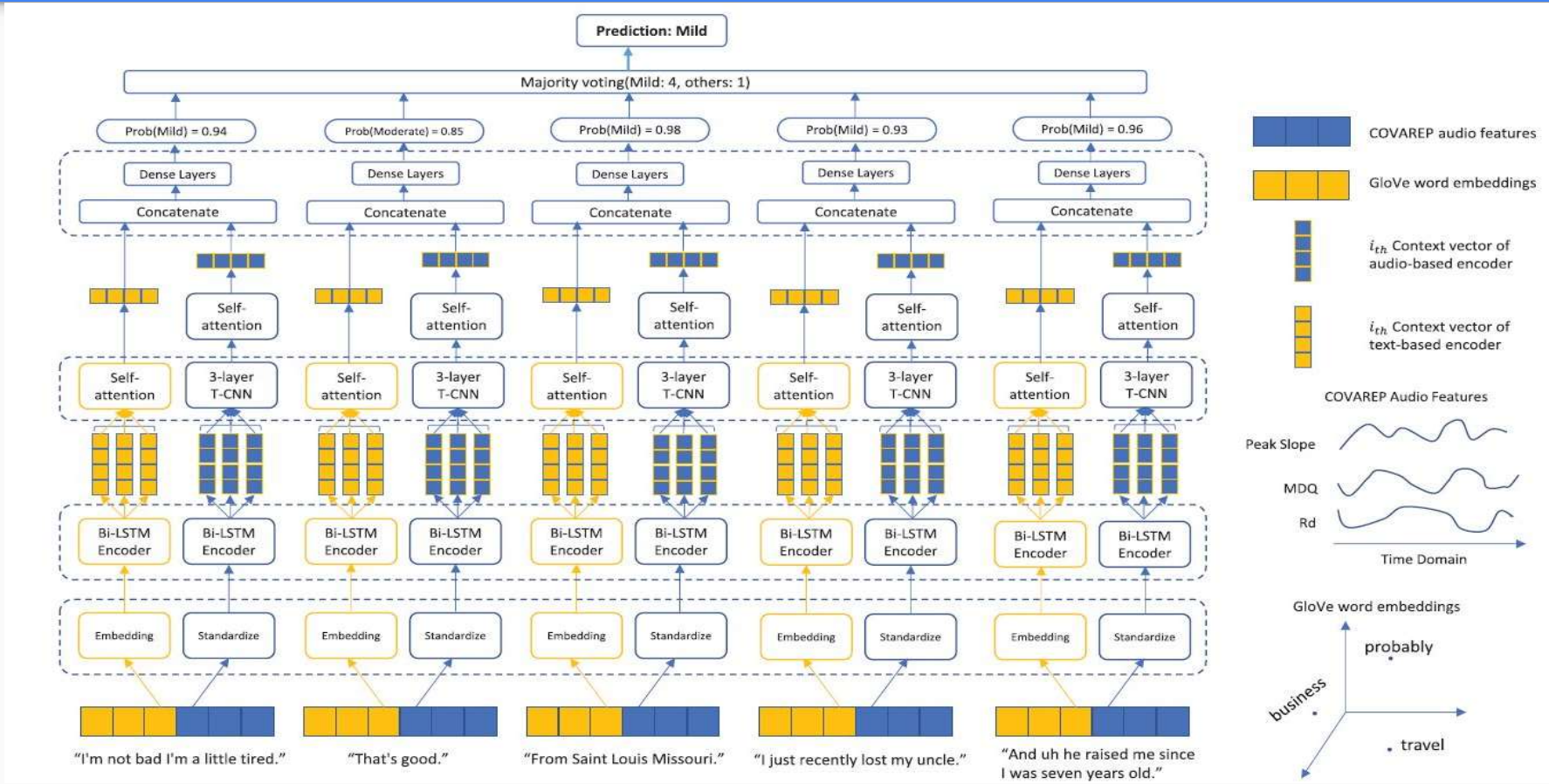
- **Fused multimodality model comprised of text and audio model, followed by a shared late fusion neural network**
- **For each segment, output size of the two subnetworks is changed from 5 to 32, turning them into feature extractors**
- **Late fusion neural network concatenates outputs of text and audio to form feature set that includes both semantic and prosodic information**
- **Fusion strategies:**
    1. Initial Fusion Model which fuses text models with varying window sizes with an audio model of constant configuration. Global Max Pooling layer is then implemented to align audio and text before final output layer
    2. Same as #1 but with an attention mechanism during feature alignment phase*
    3. Same as #1 but with two attention mechanism during feature alignment phase and fusion phase**
- **Majority Voting of segment-wise predictions to obtain the patient-level prediction**

* Feature alignment phase refers to the phase after each subnetwork generates high-level feature representations

** Fusion phase refers to the phase when these high-level feature representations are concatenated or fused

Aalto University
School of Electrical
Engineering

Fused Text-Audio Joint Model



Block diagram of the proposed multimodality depression level prediction algorithm given a specific example. Audio features are fed into the network through the input layer. After batch normalization, the input data is fed into the Bi-LSTM and time-distributed CNN block. In this proposed design, we have five time-distributed CNN blocks followed by a single-layer Bi-LSTM. The detailed architecture of each block is illustrated and explained in the remainder of this paper

Results of the Audio Modality

TABLE 4: Results of the Baseline Audio Models

Models	Random Forest		Test Baseline Madhavi et al. [66].	Yang et al. [67]
	Mean	St. dev	Mean	Mean
Accuracy	0.3192	0.0085	0.7500	0.8273
Precision	0.3206	0.0064	0.7200	0.7930
Recall	0.3184	0.0040	0.7500	1.0000
F1 Score	0.3168	0.0076	0.7300	0.8850

Baseline Models and Innovations

- Random Forest: An ensemble of 100 decision trees analyzing non-stationary audio feature series pre-processed for stationarity.
- CNN Approaches: Highlighting methodologies from Madhavi et al. (2 convolutional layers) and Yang et al. (3 convolutional layers), focusing on high-level feature extraction from frequency spectrograms.

Results of the Audio Modality

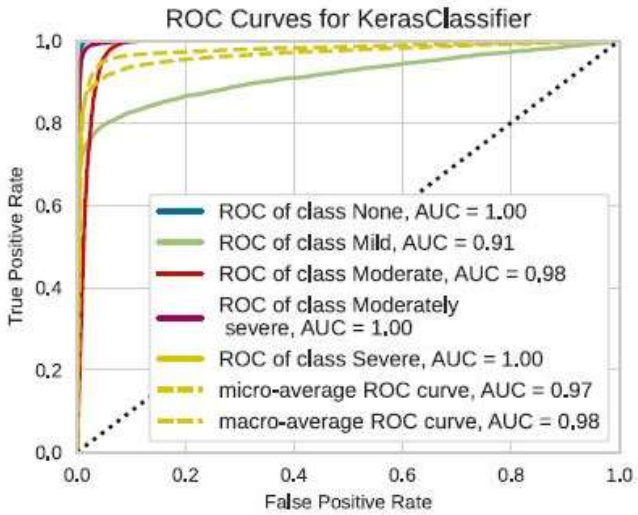
TABLE 5: A Comparative Study of Different Proposed Audio Models

Models	Experimental Settings	Accuracy	F1	p-value
LSTM + FC	TS=16, HU=73, LHU=(128,64,5), Adam	0.5674 ± 0.0034	0.5650 ± 0.0042	$<.01$
Bi-LSTM + FC	TS=16, HU=73, LHU=(128,64,5), Adam	0.8717 ± 0.0013	0.8818 ± 0.0013	$<.01$
LSTM + T-CNN	TS=16, HU=73, #TCNNB=5, #KRNL=(64,64,64,128,256), KS=(3,3,3,3,9), Adam	0.8698 ± 0.0897	0.8609 ± 0.0988	$<.01$
Bi-LSTM + T-CNN	**TS=16, HU=73, #TCNNB=5, #KRNL=(64,64,64,128,256), KS=(3,3,3,3,9), Adam**	$\mathbf{0.9871 \pm 0.0009}$	$\mathbf{0.9870 \pm 0.0009}$	$\mathbf{<.01}$

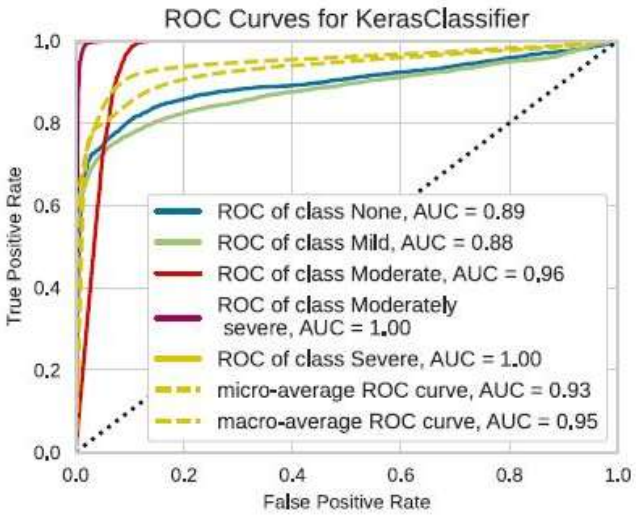
#TCNNB: Number of T-CNN blocks #KRNL: Number of conv kernels in each T-CNN block #KS: Kernel size

- Bi-LSTM + T-CNN significantly outperformed other configurations with an accuracy of 98.71% and F1-Score of 0.987
- The addition of bidirectional processing and convolutional layers for feature extraction enhances the model's accuracy and F1 scores by capturing temporal and spatial relationships

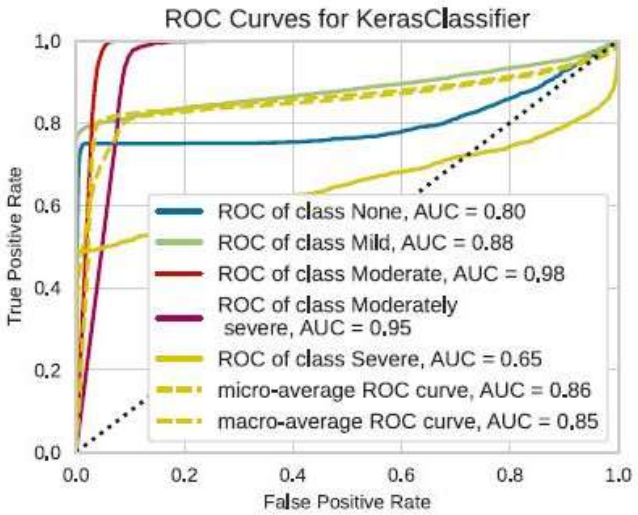
Results of the Audio Modality



(a) The ROC of 16-timestep model on DAIC-WOZ



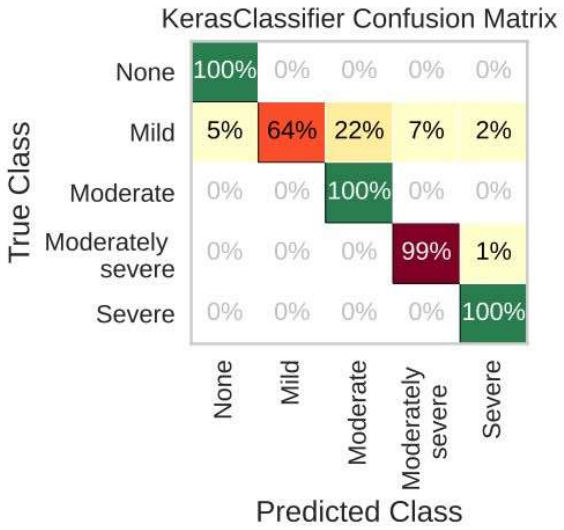
(b) The ROC of 32-timestep model on DAIC-WOZ



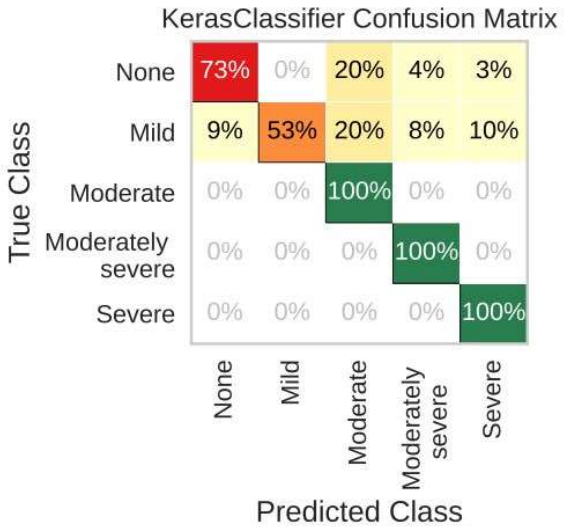
(c) The ROC of 64-timestep model on DAIC-WOZ

- Micro-Average and Macro-average AUC both decrease with a longer sequence
- AUC of “Mild” is the lowest for 16-timestep model
- AUC of “Severe” decreases significantly when we move from 32-timestep model to 64-timestep model
- Longer sequence makes it more challenging to predict severity due to introduction of noise which can mislead the model
- LSTM’s capabilities suffer as sequence length increases

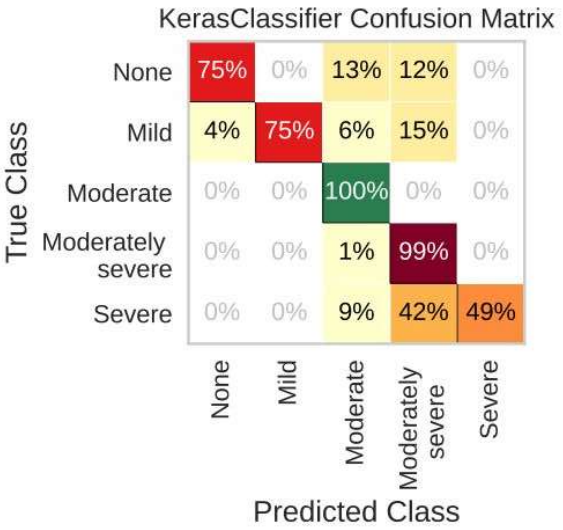
Results of the Audio Modality



(a) Confusion matrix of 16-timestep model on DAIC-WOZ



(b) Confusion matrix of 32-timestep model on DAIC-WOZ



(c) Confusion matrix of 64-timestep model on DAIC-WOZ

Aalto University
School of Electrical
Engineering

- Micro-Average and Macro-average AUC both decrease with a longer sequence
- AUC of “Mild” is the lowest for 16-timestep model
- AUC of “Severe” decreases significantly when we move from 32-timestep model to 64-timestep model
- Longer sequence makes it more challenging to predict severity due to introduction of noise which can mislead the model
- LSTM’s capabilities suffer as sequence length increases

Results of Text Modality

TABLE 6: A Comparative Study of the Proposed Text Models

Models	Experimental Settings	Accuracy	F1	Micro-average AUC
LSTM + FC	TS=64, HU=100, LHU=(256,128,5), Adam, Stopwords	0.9091	0.9094	0.9738
LSTM + FC	TS=64, HU=100, LHU=(256,128,5), Adam, No stopwords	0.9792	0.9754	0.9897
Bi-LSTM + FC	TS=64, HU=100, LHU=(256,128,5), Adam, Stopwords	0.9617	0.9610	0.9908
Bi-LSTM + FC	**TS=64, HU=100, LHU=(256,128,5), Adam, No stopwords**	**0.9685**	**0.9709**	**0.9925**

TS: Timestep; HU: #Hidden units in LSTM; LHU: #Hidden units in linear layers

TABLE 7: A Comparative Study of the Text Model with Different Window Size

Window Size	Accuracy	Precision	Recall	F1 Score
16	0.8254	0.8318	0.8340	0.8141
32	0.8256	0.8371	0.8465	0.8260
64	**0.8778**	**0.8779**	**0.8782**	**0.8705**
128	0.8409	0.8599	0.8430	0.8304

- Bi-LSTM performs better due to learning of more contextual information
- Removal of stopwords worsened model performance
- Window size of 64 yields best results
- Any larger window size will have diminishing returns

Results of fused model

TABLE 8: A Comparative Study of Our Proposed Patient-Level Methods and the State of The Art

Model	Experimental Settings	Accuracy	F1	Sensitivity	Specificity
UniLSTM as encoder	WIN=16, Stride=64	0.8604	0.8579	0.9844	0.8182
	WIN=32, Stride=64	0.9209	0.9188	0.9647	0.9777
	WIN=64, Stride=64	0.8674	0.8682	0.9705	0.9888
BiLSTM as encoder	WIN=16, Stride=64	0.9488	0.9500	0.9735	0.9444
	WIN=32, Stride=64	0.9186	0.9191	0.9852	0.9700
	WIN=64, Stride=64	0.8535	0.8546	0.9647	0.8778
BiLSTM as encoder	WIN=16, Stride=64, attention	0.8419	0.8427	0.9735	0.8222
	WIN=32, Stride=64, attention	**0.9581**	**0.9580**	**0.9824**	**1.0000**
	WIN=64, Stride=64, attention	0.9093	0.9086	0.9706	0.9889
UniLSTM as encoder	WIN=16, Stride=64, attention(aligning&fusion)	0.8977	0.8973	0.9559	0.9889
	WIN=32, Stride=64, attention(aligning&fusion)	0.9326	0.9315	0.9735	0.9889
	WIN=64, Stride=64, attention(aligning&fusion)	0.8581	0.8615	0.9412	0.8889
BiLSTM as encoder	WIN=16, Stride=64, attention(aligning&fusion)	0.8491	0.8439	0.9353	0.9000
	WIN=32, Stride=64, attention(aligning&fusion)	0.9047	0.9103	0.9941	0.9000
	WIN=64, Stride=64, attention(aligning&fusion)	0.6279	0.6560	0.7500	1.0000
Srimadhur et al. [68]	End to end convolutional neural network	0.7464	0.7750	0.74	0.8
Alhanai et al. [62]	Combination of LSTM and CNN	*	0.77	0.83	*
Niu et al. [69]	Hierarchical context-aware graph attention model	*	0.92	0.92	*

Conclusion

- Proposed multimodal configuration achieved the highest F1-score of 0.9580 on the patient-level depression detection task, even outperforming previous state-of-the-art models such as graph attention models
- Balance between increasing segment length and retaining information must be maintained
- Use of attention mechanism during the feature alignment phase improves performance, but addition of anymore will cause overfitting
- Future studies on how to represent audio and text features during the whole interview should be carried out
- Semantic and contextual integration, such as NLU and Psycholinguistics, could be further carried out

Questions

Question 1: What does PHQ-8 measure and how?

Question 2: How is a T-CNN used in the paper different from a standard CNN?

References

K. Mao, W. Zhang, D. B. Wang, et al., “Prediction of depression severity based on the prosodic and semantic features with bidirectional LSTM and time distributed CNN,” IEEE Transactions on Affective Computing, vol. 14, no. 3, pp. 2251–2265, 2023, Cited by: 13; All Open Access, Green Open Access. doi: 10.1109/TAFFC.2022.3154332. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125318336&doi=10.1109%2fTAFFC.2022.3154332&partnerID=40&md5=4c0893558c65b06265a31f6b4f7c7584.

THANK YOU!

THANK YOU! Any questions?

Aalto University
School of Electrical
Engineering