Coursework 3: Non-Verbal Affective Computing for Emotion Recognition

Jacob Dean[1] (ID: 20046542)

[1] School of Computer Science, University of Nottingham; psxjd13@nottingham.ac.uk

1. Introduction

Emotion recognition technologies fall into two main categories: verbal systems using text and speech, and non-verbal systems using facial expressions and physiological signals. Non-verbal systems have the distinct advantage of operating subliminally, which opens the door to a much wider range of applications such as detecting stress and monitoring student engagement (EDPS, 2022). This report looks at five research papers which all construct non-verbal emotion recognition models but using diverse methods. The remaining sections are structured as follows: section 2 assesses the methods of each paper, section 3 compares these methods, and section 4 concludes.

2. Research Papers

A. Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features (Lee et al., 2020)

Overview

This paper presents an emotion recognition model using PPG signals and a convolutional neural network (CNN). The key findings are that high accuracy can still be achieved with short duration signals, and that manually selected features can complement those automatically selected by a neural network.

Research Methods

The authors choose to build their model around PPG signals (Fig. 1) because these can be collected by ubiquitous devices such as smartphones and smartwatches, making their model a particularly scalable solution for emotion detection (Chandel et al., 2020).

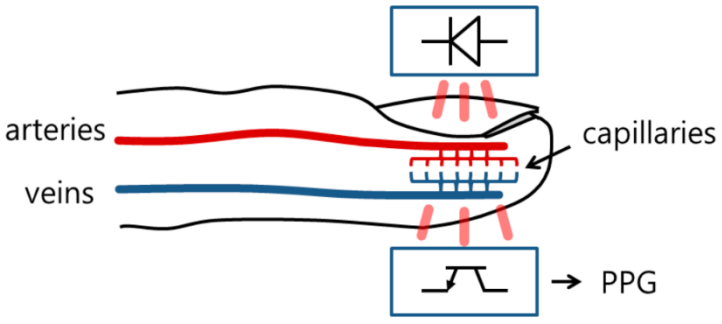

Figure 1. *PPG sensors measure the amount of light passing through one's finger (taken from Park et al., 2017)*

Two graphs are assembled in preparation for manual feature extraction: NN interval (Fig. 2) and power spectral density (PSD) (Fig. 3). This is because each graph contains different statistical information, for instance NN interval reflects heart rate variability whilst PSD indicates signal intensity. The Pearson correlation coefficients of all extracted features are then analysed and only the ten highest are selected. The authors justify their decision to remove surplus features by showing that model accuracy is approximately 1% lower when all features are included.

Each graph is also fed into convolutional layers for automated feature extraction which yields 230 additional features, dwarfing the ten that were manually selected. Despite this imbalance, model accuracy falls by 4.6-6.3% when the manual features are excluded, demonstrating the importance of including both feature sets.

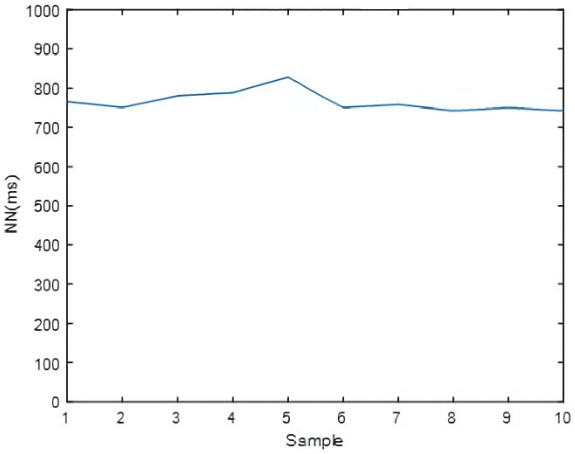

Figure 2. *NN Interval (taken from Lee et al., 2020)*

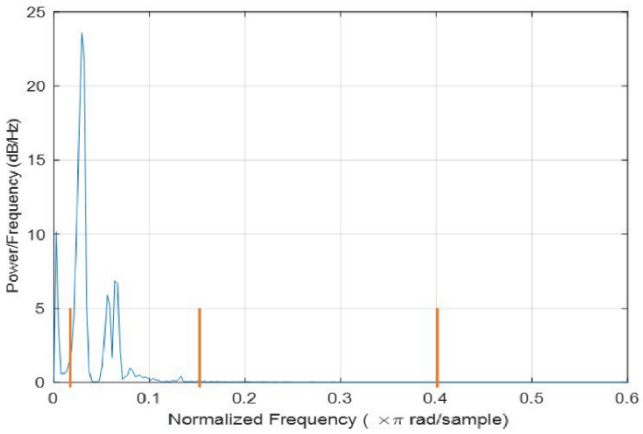

Figure 3. *Power Spectral Density (taken from Lee et al., 2020)*

For classification, the authors apply Russell's Circumplex, a model that describes emotions along two continuous dimensions: valence and arousal (Fig. 4) (Russell, 1980). However, they decide to discretise these dimensions into binary variables which restricts the number of overall classifications to four. The advantage of restricting classifications is that it enables higher accuracies to be attained, but this comes at the cost of making models far less informative since emotions cannot be properly described by only four states.

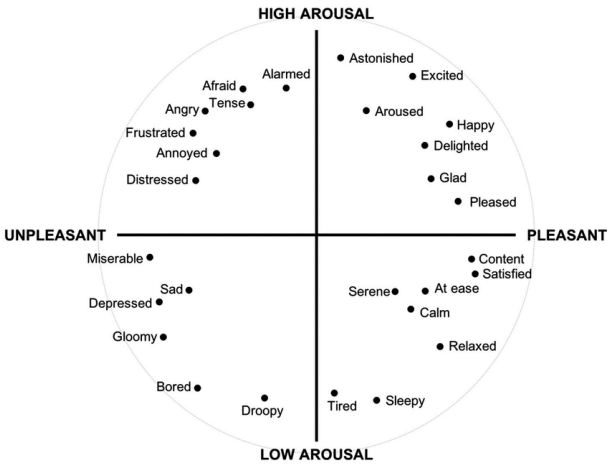

Figure 4. *Russell's Circumplex Model of Affect (taken from Winiger, 2020)*

Finally, the authors compared their results to other studies using the same public dataset (DEAP) and found that they achieved some of the highest accuracies whilst also using the shortest signal duration (10s). However, it would have been better if other studies using PPG had been included in the comparison, since this would have increased the reliability of the authors' conclusions.

B. Wearable emotion recognition system based on GSR and PPG signals (Udovičić et al., 2017)

Overview

This paper presents a model using a fusion of galvanic skin response (GSR) and PPG signals, applied to SVM and KNN models. The key findings are that personalised models perform a lot better than generalised models and that SVM and KNN models perform equally well.

Research Methods

The authors use GSR (Fig. 5) in addition to PPG because both types of signals are often measurable with the same device, allowing more data to be collected at no additional cost.

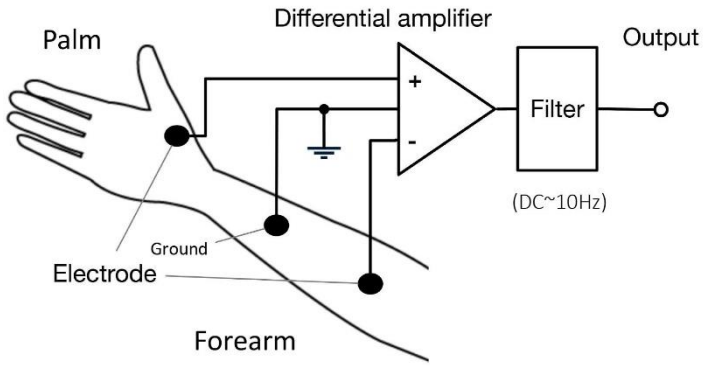

Figure 5. *GSR sensors measure the electrical conductance of skin (taken from Momose et al., 2020)*

Features are manually extracted from both time and frequency domains like we saw in the previous paper, however the authors surprisingly do not create a unified model with both feature sets. A unified model would have likely performed better than the authors' existing models since the feature sets do not overlap.

For classification, the authors implement two relatively simple machine learning methods, SVM and KNN, whereas all the other papers apply deep learning. However, the authors still achieve respectable results compared to these papers as we later see in Figure 15. Avoiding deep learning also made it feasible for the authors to collect their own data, since fewer samples were required. To accomplish this, test subjects were shown random images to elicit emotional responses and were then asked to rate their emotions using self-assessment manikins (Fig. 6) (Bradley & Lang, 1994).

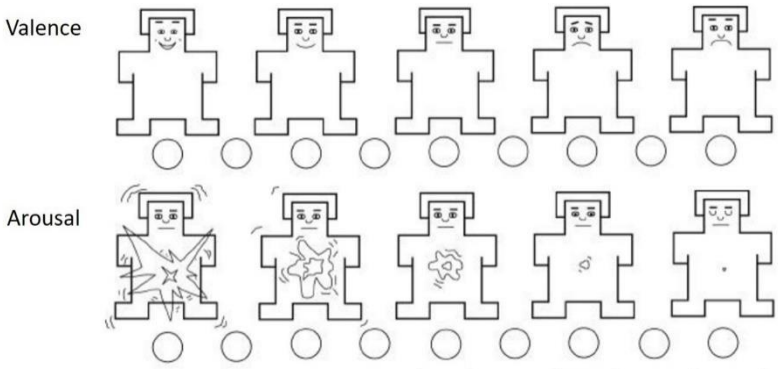

Figure 6. *Self-assessment manikins are used for measuring emotional responses (taken from Reinares-Lara et al., 2019)*

Collecting primary data provided the authors with enough flexibility to compare the performance of personalised and generalised emotion recognition models. They found that personalised models achieved 15-20% higher accuracy than generalised models. This was to be expected since personalised models can better handle individual differences in physiological signals. However, the authors partitioned their data between these two models, whereas more robust results could have been achieved if all data had been utilised at each stage.

C. Facial emotion recognition using convolutional neural networks (FERC) (Mehendale, 2020)

Overview

This paper presents a model using facial images alongside sequential CNNs. The key findings are that background removal is an important preprocessing step and high accuracies can be attained at lower computational complexities.

Research Methods

The author decides to use facial expressions because they often mirror our true emotions, however this practice has been criticised in recent years since facial expressions can also act as a mask to conceal how we truly feel (Lee et al., 2020).

Before classification, facial images were passed through a bespoke CNN for background removal using edge detection filters (Fig. 7) and a skin tone detection algorithm. This was an essential step to prevent predictions from being influenced by people's surroundings. However, there were several shortcomings of the background remover such as limited performance in dealing with rotated faces and greyscale images. Therefore, it may have been preferable to use an existing background removal tool since these are widely available (e.g. Adobe).

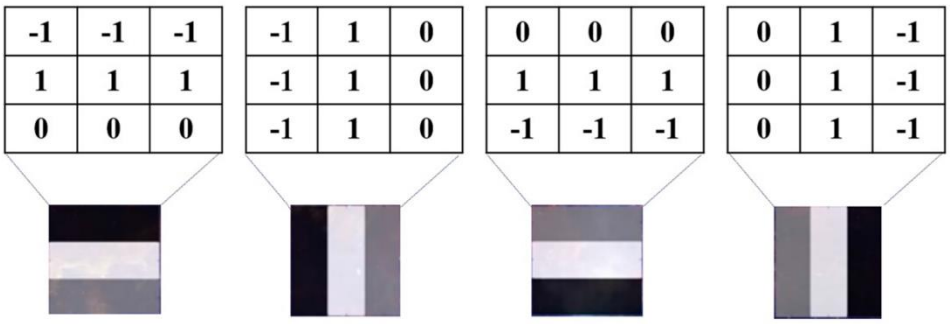

Figure 7. *The four edge detection filters used in the background removal CNN (taken from Mehendale, 2020)*

To reduce computational complexity, the author conducted emotional classification using a separate CNN to the one used for background removal. This appears to have been effective since his approach is shown to have lower complexity (O^4) compared to popular alternatives like GoogleNet (O^5).

The author applies a psychological model based on Ekman's Basic Emotions to categorise facial expressions (Fig. 8) (Ekman, 1992). However, he purposefully excludes two emotions (disgust and contempt) because a high misclassification rate was observed for these. This limits the real-world applicability of the model, since these two emotions would never be properly classified.

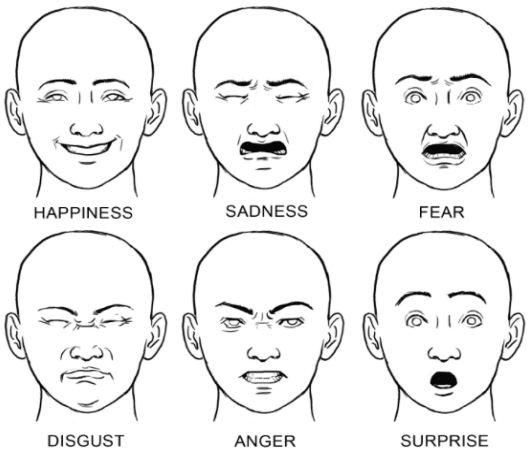

Figure 8. *Ekman's Basic Emotions (taken from Harrison, 2021)*

To evaluate the model, the author compares its performance with other studies that also use facial expressions. Performance appears to be slightly below average, with the model achieving 78-96% accuracy compared to 91.44-99.30% achieved by the other studies.

Additionally, all the other studies consider a wider range of emotions than the author, making their models more appropriate in real-world use.

D. EEG-based emotion recognition using simple recurrent units network and ensemble learning (Wei et al., 2020)

Overview

This paper presents a model using EEG signals and simple recurrent units (SRU). The key findings are that SRU-based models are computationally efficient as well as accurate, and that higher frequency brain waves predict emotions more strongly than lower frequency brain waves.

Research Methods

The authors use EEG signals because they look directly at the brain where emotions originate, and the introduction of dry electrodes in recent years has also made EEG headsets more wearable (Fig. 9).

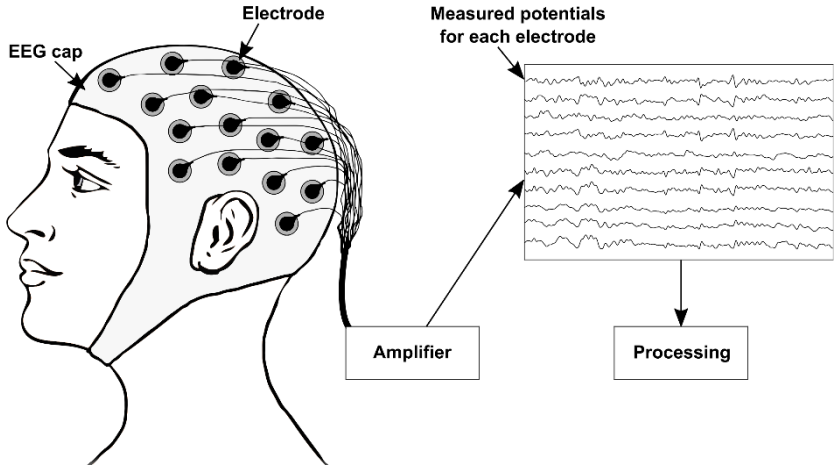

Figure 9. *An EEG headset measures the electrical activity of the brain (adapted from Nagel, 2019)*

Four feature extraction methods are considered at each of the five frequencies of brain wave, yielding 20 total permutations. Interestingly, the authors decide not to fuse these permutations into a single model but instead create 20 base models and combine their classifications using either simple or weighted voting. The best combined model achieves 83.13% accuracy whereas the best base model achieves a slightly lower 80.02% accuracy, demonstrating that voting is a viable method for combining features.

SRU is chosen as the architecture of each base model because it has the same benefit of long-term sequential memory as an LSTM model, but also has significantly lower training time. For example, the authors observed the training time for an LSTM to be over twice as high (136.5s) as an SRU (60.1s). The reason sequential memory is important when processing EEG signals is that they form a time series, unlike static inputs such as facial images.

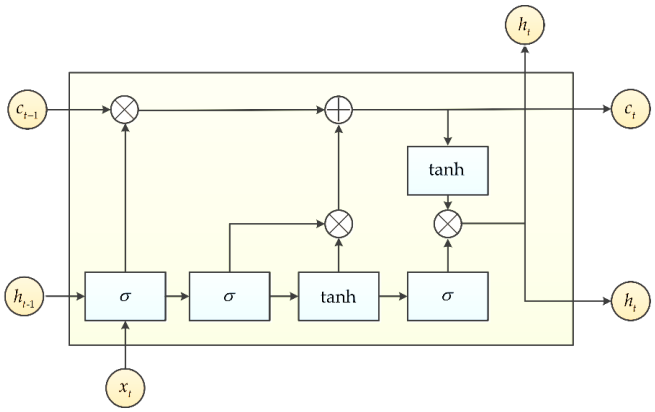

Figure 10. *Structure of an LSTM unit (adapted from Han et al., 2020)*

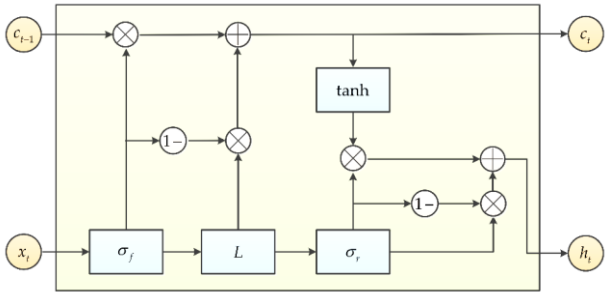

Figure 11. *Structure of an SRU unit (adapted from Han et al., 2020)*

The authors evaluate their results using both ROC curves and confusion matrices. Both these methods provide the same insight that the highest frequency bands (gamma and beta) yield the strongest base models. Simpler models (KNN, Naïve Bayes, SVM) were also trained as benchmarks. This allowed the authors to demonstrate the benefits of using deep learning, since the SRU base models were the best performing in all 20 instances. However, this performance differential would have likely been lower if the benchmark models had been tuned more carefully, for instance by testing different kernels for SVM.

E. Emotion recognition from multiband EEG signals using CapsNet (Chao et al., 2019)

Overview

This paper presents a model using EEG signals alongside a capsule network (CapsNet). The key findings are that spatial information about electrode position can improve model accuracy and that several feature classes can be encoded in a single matrix.

Research Methods

The authors propose a multiband feature matrix (MFM) as a way of capturing spatial information about the origin of each EEG signal because a matrix can be used as a map to represent the positions of each electrode on the head (Fig. 12).

To form the MFM, four copies of these maps are fused together where each copy corresponds to a different frequency band. It is unclear why the fifth frequency band (delta) is excluded. This could be a limitation of designing the MFM as a square matrix since this can only fit four maps. The other limitation is that the MFM fails to capture information about how the EEG signal varies over time, which we saw was an important feature in the previous paper.

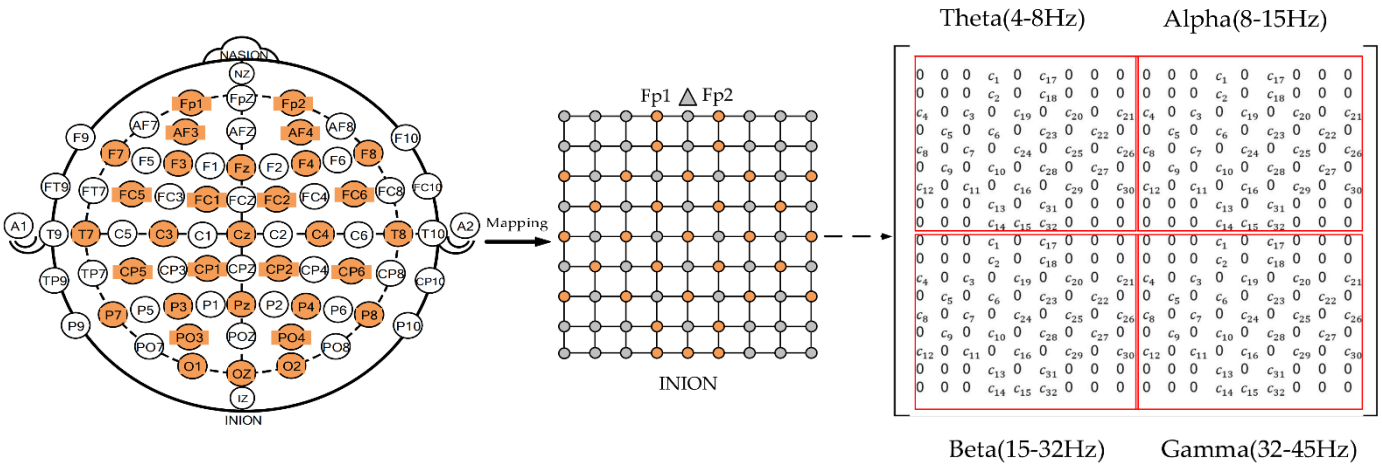

Figure 12. *Mapping from the grid of electrodes to the MFM (adapted from Chao et al., 2019)*

The authors use a CapsNet for classification (Fig. 13). This is a neural network consisting of basic units called capsules that compare patterns in EEG data coming from specific electrode locations. The main benefit of using a CapsNet is that it preserves the spatial information captured in the MFM, whereas a CNN for instance would lose this information at the pooling stage.

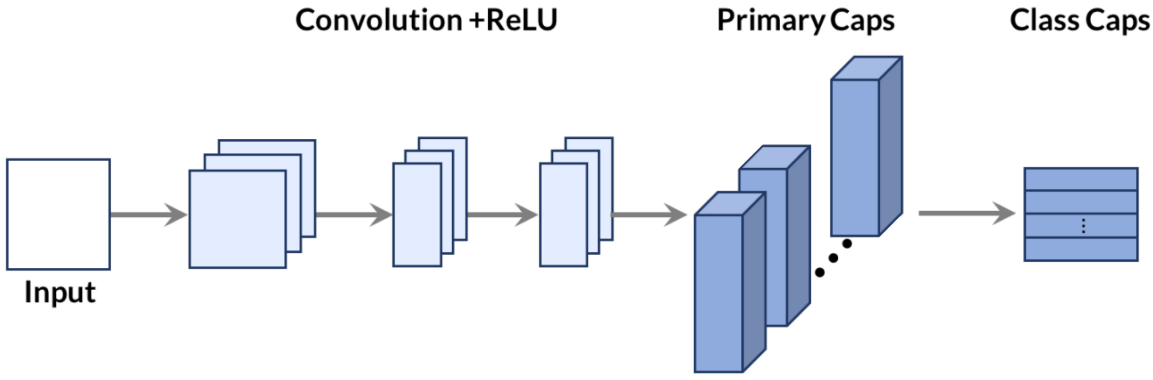

Figure 13. *Basic structure of a capsule network (adapted from Fentaw & Kim, 2019)*

We have seen other papers classify emotions based on valence and arousal, however this paper also includes the 'dominance' dimension, corresponding to the more advanced three-dimensional PAD model (Fig. 14) (Mehrabian, 1996). This provides a more complete representation of emotion by increasing the number of distinct classifications from four to eight.

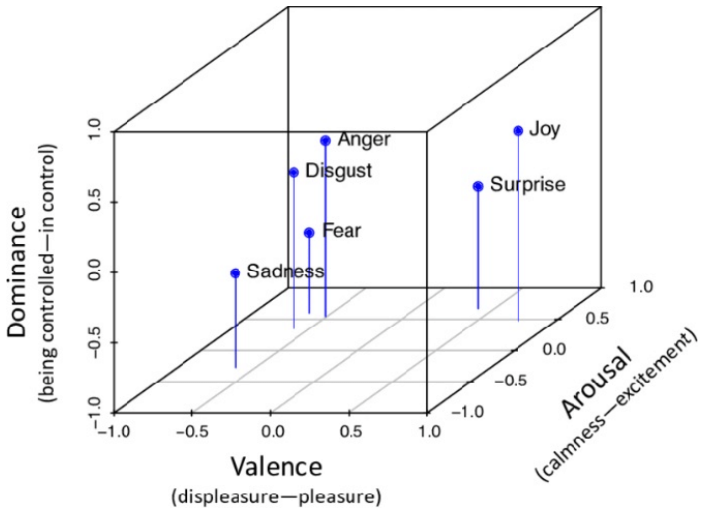

Figure 14. *PAD model (taken from Feidakis et al., 2019)*

To evaluate their model, the authors use benchmarking like the previous paper, however they also include CNN models as a benchmark. Additionally, they apply a Wilcoxon signed-rank test to demonstrate that CapsNet performs better than CNN in all three dimensions at a 5% significance level. This supports the authors' main hypothesis that properly representing the spatial characteristics of EEG signals has a positive effect on model accuracy.

3. Comparison

All five papers use the same overarching structure to build models. First, a dataset is selected, then features are extracted, and finally these features are fed into a machine learning algorithm. Therefore, the differences between the papers lie within each of these stages and this gives rise to a diverse set of results as displayed in Figure 15.

	Emotional Marker	Machine Learning	Feature Domains	Model Style	Emotional States	Overall Accuracy
Lee et al. [A]	PPG	CNN	Temporal & Frequency	Generalised	4	<80.9%
Udovičić et al. [B]	EDA & PPG	SVM	Temporal	Personalised	4	<80.6%
		KNN	Frequency	Generalised		<66.7%
Mehendale [C]	Face	CNN	Spatial	Generalised	5	78%-96%
Wei et al. [D]	EEG	SRU	Temporal & Frequency	Personalised	3	83.13%
Chao et al. [E]	EEG	CapsNet	Spatial & Frequency	Generalised	8	<66.73%

Figure 15. *Comparison of selected characteristics*

Firstly, we see that different papers extract different feature sets depending on the chosen emotional markers. For example, spatial features are critically important for two-dimensional data such as facial images but are irrelevant for PPG signals that are collected from a single location. Papers A and E both find that a diverse fusion of features improves model accuracy, revealing certain areas for improvement. For example, paper C represents facial videos by a single frame, however temporal information could be integrated by comparing multiple frames over time.

Secondly, we see that some papers build personalised models whilst others build generalised models. Paper B demonstrates the superior performance of personalised models, however building these requires copious amounts of personal data. A solution could be to first employ generalised models that have the option of becoming more personalised over time as new data is collected.

Thirdly, we observe a wide range of machine learning algorithms of varying complexity. Paper B uses traditional methods rather than deep learning (SVM, KNN), papers A and C use an established deep learning method (CNN), whilst papers D and E use state-of-the-art deep learning methods (SRU, CapsNet). However, we see that greater complexity doesn't necessarily correlate with better performance, since paper E shows relatively low accuracy values compared to the other papers.

All five papers compare their results with pre-existing literature. This is an effective method for tracking overall progress but is not suitable for pinpointing which innovations led to these gains. Most of the papers addressed this issue by benchmarking their complete models against simpler versions.

None of the papers based on the Circumplex model reported an overall cross-dimension accuracy value. This is why much of the last column in Figure 15 is uncertain. The absence of overall accuracies makes it difficult to compare performance with paper C for instance, which is based on Ekman's model. Therefore, it would be good if reporting overall accuracy became common practice in future studies.

4. Conclusion

To conclude, there is still no consensus on which emotional marker yields the best performing models. However, in recent affective computing literature we have seen that deep learning has become very popular and that models considering a more diverse range of features tend to show higher accuracy. The large datasets required for deep learning also explain why most studies opt to use pre-existing datasets rather than collecting primary data. Only a limited range of emotional states are currently considered, however we can expect more sophisticated classification schemes to become feasible as model performance continues to improve.

5. References

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Chandel, V., Saha, J., Bhattacharyya, C., & Ghose, A. (2020). Real-time robust estimation of breathing rate from PPG using commercial-grade smart devices. *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. https://doi.org/10.1145/3384419.3430458

Chao, H., Dong, L., Liu, Y., & Lu, B. (2019). Emotion Recognition from Multiband EEG Signals Using CapsNet. *Sensors*, *19*(9), 2212. https://doi.org/10.3390/s19092212

EDPS. (2022, May 20). *TechDispatch #1/2021 - Facial Emotion Recognition*. European Data Protection Supervisor. Retrieved 23 May 2022, from https://edps.europa.eu/data-protection/our-work/publications/techdispatch/techdispatch-12021-facial-emotion-recognition_en#:%7E:text=Facial%20Emotion%20Recognition%20(FER)%20is,information%20on%20one's%20emotional%20state.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Feidakis, M., Rangoussi, M., Kasnesis, P., Patrikakis, C., Kogias, D., & Charitopoulos, A. (2019). Affective Assessment in Distance Learning: A Semi-explicit Approach. *The International Journal of Technologies in Learning*, *26*(1), 19–34. https://doi.org/10.18848/2327-0144/cgp/v26i01/19-34

Fentaw, H. W., & Kim, T. H. (2019). Design and Investigation of Capsule Networks for Sentence Classification. *Applied Sciences*, *9*(11), 2200. https://doi.org/10.3390/app9112200

Han, S., Meng, Z., Omisore, O., Akinyemi, T., & Yan, Y. (2020). Random Error Reduction Algorithms for MEMS Inertial Sensor Accuracy Improvement—A Review. *Micromachines*, *11*(11), 1021. https://doi.org/10.3390/mi11111021

Harrison, T. (2021, November 5). *Ekman's 6 Basic Emotions and How They Affect Our Behavior*. The Minds Journal. Retrieved 23 May 2022, from https://themindsjournal.com/basic-emotions-and-how-they-affect-us/

Lee, M., Lee, Y. K., Lim, M. T., & Kang, T. K. (2020). Emotion Recognition Using Convolutional Neural Network with Selected Statistical Photoplethysmogram Features. *Applied Sciences*, *10*(10), 3501. https://doi.org/10.3390/app10103501

Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (FERC). *SN Applied Sciences*, *2*(3). https://doi.org/10.1007/s42452-020-2234-1

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. *Current Psychology*, *14*(4), 261–292. https://doi.org/10.1007/bf02686918

Momose, H., Morimitsu, N., Ikeda, E., Kanai, S., Sakaguchi, M., & Ohhashi, T. (2020). Eyes Closing and Drowsiness in Human Subjects Decrease Baseline Galvanic Skin Response and Active Palmar Sweating: Relationship Between Galvanic Skin and Palmar Perspiration Responses. *Frontiers in Physiology*, *11*. https://doi.org/10.3389/fphys.2020.558047

Nagel, S. (2019). *Towards a home-use BCI: fast asynchronous control and robust non-control state detection*. https://d-nb.info/1201644526/34

Park, C., Shin, H., & Lee, B. (2017). Blockwise PPG Enhancement Based on Time-Variant Zero-Phase Harmonic Notch Filtering. *Sensors*, *17*(4), 860. https://doi.org/10.3390/s17040860

Reinares-Lara, P., Rodríguez-Fuertes, A., & Garcia-Henche, B. (2019). The Cognitive Dimension and the Affective Dimension in the Patient's Experience. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02177

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Udovičić, G., Đerek, J., Russo, M., & Sikora, M. (2017). Wearable Emotion Recognition System based on GSR and PPG Signals. *Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care*. https://doi.org/10.1145/3132635.3132641

Wei, C., Chen, L. L., Song, Z. Z., Lou, X. G., & Li, D. D. (2020). EEG-based emotion recognition using simple recurrent units network and ensemble learning. *Biomedical Signal Processing and Control*, *58*, 101756. https://doi.org/10.1016/j.bspc.2019.101756

Winiger, S. (2020, December 11). *Russells Circumplex model of emotions*. Samim. Retrieved 23 May 2022, from https://samim.io/p/2020-12-11-russells-circumplex-model-of-emotions/