Impact of Data Distribution on Action Unit Detection

Saurabh Hinduja and Shaun Canavan

***Abstract*—Detecting action units is an important task in face analysis, especially in facial expression recognition. This is due, in part, to the idea that expressions can be decomposed into multiple action units. In this paper we investigate the impact of data distribution (e.g. number of active action units, and patterns of action units) on the accuracy of detecting action units. To facilitate this investigation, we review state of the art literature, for AU detection, on 2 state-of-the-art face databases that are commonly used for this task, namely DISFA, and BP4D. We also conduct multiple experiments on BP4D to validate our findings, which suggest that there are explicit detection patterns that are directly impacted by distribution of the action units. This pattern exists across a range of classification methods that include convolutional neural networks, long short-term neural networks, and support vector machines. In many works F1-binary scores are used to evaluate the accuracy of action unit detection methods. Our findings also suggest that the patterns strongly impact the F1-binary scores, and that using other metrics such as F1-macro, F1-micro, or Area Under the Curve (AUC) scores along with more balanced data can help with breaking this impact.**

***Index Terms*—action units, detection, distribution**

I. Introduction

Facial expression recognition is a growing field that has attracted the attention of many research groups due in part from early work from Picard [18]. A range of modalities have been found useful for this problem including 2D [3], [27], thermal [17], [24] and 3D/4D data [1], [4]. Promising multimodal approaches to expression recognition have also been proposed [14], [25]. Another interesting approach to facial expression recognition is based on the detection of action units (AU). These works are based on the Facial Action Coding System [8] (FACS), which is seminal work that decomposes facial expressions into a set of action units.

There have been promising approaches to action unit detection that have made use of both hand-crafted features, as well as deep learning. Liu et al. [15] proposed TEMT-NET that utilizes the correlations between AU detection and facial landmarks. Their proposed network performs action unit detection, facial landmark detection, and thermal image reconstruction simultaneously. The thermal reconstruction and facial landmark detection provide regularization on the learned features providing a boost to AU detection performance. Werner et al. [26] investigated action unit intensity estimation using modified random regression forests. They also developed a visualization technique for the relevance of the features. Their results suggest that precomputed features are not enough to detect certain AUs. Li et al. [12] proposed the EAC-Net, which is a deep network that enhances and crops regions of interest for AU detection. They found the proposed approach allows for robustness to face position and pose. They also integrate facial attention maps that correspond to areas of the face with active AUs. Their results suggest using these attention maps can enhance the learning in the network, at specific layers. Romero et al. [19] proposed a Convolutional Neural Network (CNN) to address multi-view AU detection. In their approach they explicitly model temporal information using optical flow, as well as predicting the overall view of a sequence before detecting the AUs. By first predicting the view, their cascaded approach evaluates temporal AU networks trained for the specific view. The Facial Expression Recognition and Analysis challenge (FERA) [21]–[23] also focused on the detection of AUs. This challenge was designed to address the challenge of a common evaluation protocol for AU detection.

Girard et al. [9] conducted an investigation to determine how much training data is needed for AU detection. They investigated 80 subjects, and more than 350,000 frames of data using SIFT features. Their results suggest that variation in the total number of subjects is an important factor in increasing AU detection accuracies, as they achieved their max accuracy from a large number of subjects. Ertugrul et al. [7] investigated the efficacy of cross-domain AU detection. To do this, they reviewed state-of-the-art literature and conducted experiments using both shallow and deep approaches, to address some of the limitations of cross-domain detection. Their results suggest that more varied domains, as well as deep learning can increase generalizability, however, more improvement is needed before applying AU classifiers across different domains. In this paper, we also review state-of-the-art literature in AU detection, however, our focus here is not on cross-domain detection but within-domain. Specifically, we focus on how the number of active AUs and AU patterns (i.e multiple active AUs) impact the detection accuracies. Considering this, the contribution of our work is 3-fold, and can be summarized as follows:

1) We review state-of-the-art literature and give an in-depth analysis on the impact of distribution and patterns on AU detection results. Our results suggest that there is an explicit trend that strongly impacts the F1-binary scores.
2) We analyze the AU data distribution on 2 state-of-the-art datasets, namely DISFA [16] and BP4D [29].
3) We conduct multiple experiments, on BP4D, that suggest using F1 macro/micro, or AUC, as well as a balanced data distribution can result in more accurate analysis of AU detection results.

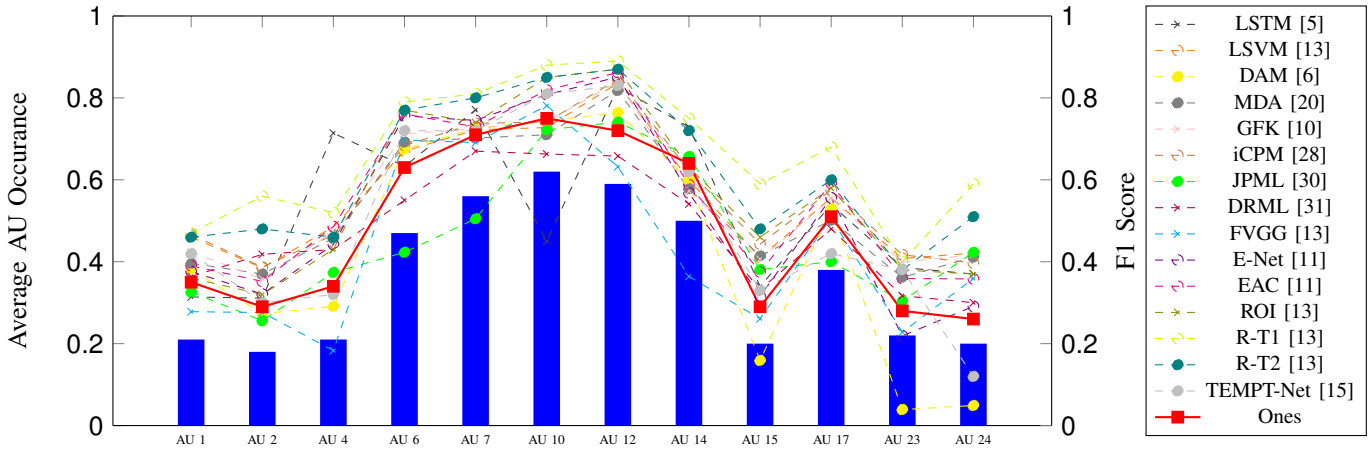


Fig. 1: AU detection accuracy vs occurrence in BP4D. Bars are the average number of AU occurrences, per frame, across all subjects. Line graphs are different F1 scores, of methods in the literature, for each AU. NOTE: "Ones" is what happens when we manually predict all 1's (i.e. AU is active) for each of the AUs. *NOTE: Best viewed in color.*

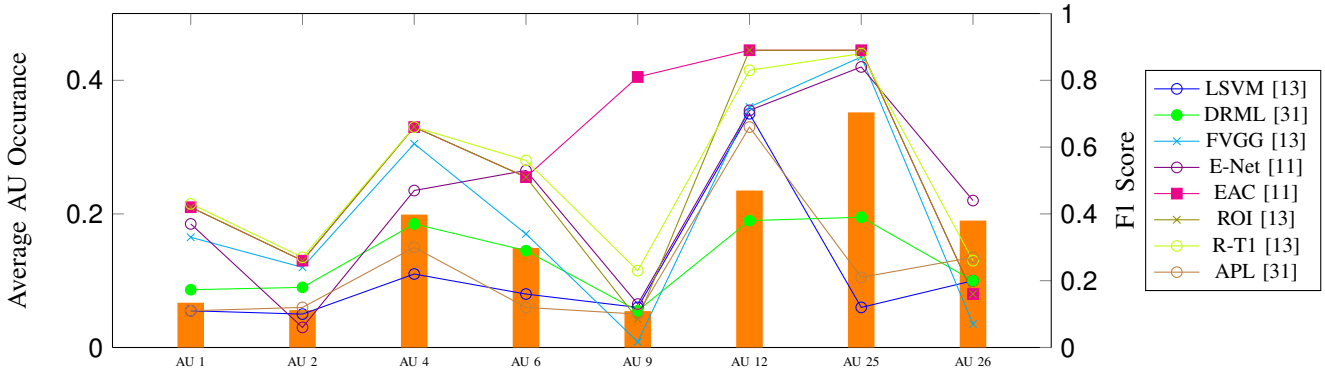


Fig. 2: AU detection accuracy vs occurrence in DISFA. Bars are the average number of AU occurrences, per frame, across all subjects. Line graphs are different F1 scores, of methods in the literature, for each AU. NOTE: The scales on the left and right are different due to the low number of active AUs compared to some of the F1 scores. *NOTE: Best viewed in color.*

II. Data Distribution

To investigate the impact of the distribution and patterns of action units (i.e. multiple active action units), we analyzed two state-of-the-art datasets - DISFA [16], and BP4D [29]. Details on both of these datasets are given in the following subsection.

A. Datasets

DISFA is a spontaneous dataset designed for studying facial action intensity. It contains 27 adults (12 women/15 men) that watched a 4-minute video clip that was meant to elicit spontaneous expressions (i.e. AUs). For our analysis, all frames from this dataset are used, as all are AU annotated (130,815 frames). It is important to note that in this dataset, 66,893 frames have no active AUs ($\approx$ 51% of the data). For this dataset, we investigated the most commonly used AUs from the literature: 1, 2, 4, 6, 9, 12, 25, 26.

BP4D is a multimodal (e.g. 2D, 3D, AUs) facial expression dataset which was used in the Facial Expression Recognition and Analysis challenges in 2015 [21] and 2017 [22] where the focus, in both challenges, was the detection of occurrence and intensity of AUs. There are a total of 41 subjects (23 female/18 male) displaying eight dynamic expressions. For our analysis and experimental design (Section III), we analyze all AU annotated frames (146,847 frames) from this dataset for our investigation. In BP4D, there are 10,630 frames that have no active AUs ($\approx$ 7.2% of the AU annotated frames). For this dataset, we also investigated the most commonly used AUs from the literature: 1, 2, 4, 6, 7, 10, 12, 14, 15, 17, 23, 24.

B. Data Distribution and AU Recognition

We have analyzed state-of-the-art literature regarding their F1-binary scores from AU detection experiments on BP4D and DISFA. As can be seen in Figs. 1 and 2, the occurrence of AUs in both datasets are imbalanced. For example, in BP4D, AU 10 has an average occurrence of 0.62, while AU 2 has an average AU occurrence of 0.18. There is a direct correlation between these occurrences and their F1-binary scores. AU 10 is one of the highest occurring AUs and also one of

TABLE I: Correlation between F1-binary scores and AU distribution. *NOTE: high correlation suggests the F1-binary scores explicitly follow the data distribution trend.*

Method	Correlation	
	BP4D	DISFA
LSTM [5]	0.680	-
LSVM [13]	0.957	0.347
DAM [6]	0.922	-
MDA [20]	0.948	-
GFK [10]	0.951	-
iCPM [28]	0.967	-
JPML [30]	0.869	-
DRML [31]	0.949	0.844
FVGG [13]	0.890	0.785
E-Net [11]	0.944	0.919
EAC [11]	0.953	0.472
ROI [13]	0.966	0.773
R-T1 [13]	0.931	0.816
R-T2 [13]	0.970	-
APL [31]	-	0.5098
Average	**0.921**	**0.683**

TABLE II: Variance and standard deviation of F1-binary scores, for each individual AU, between all investigated methods (Figs. 1, 2 and Table I).

AU	BP4D		DISFA	
	Var	Std	Var	Std
AU 1	0.004	0.060	0.020	0.141
AU 2	0.007	0.083	0.007	0.083
AU 4	0.014	0.118	0.032	0.179
AU 6	0.010	0.099	0.031	0.175
AU 7	0.005	0.073	-	-
AU 9	-	-	0.064	0.253
AU 10	0.011	0.106	-	-
AU 12	0.006	0.080	0.027	0.165
AU 14	0.009	0.096	-	-
AU 15	0.011	0.103	-	-
AU 17	0.004	0.065	-	-
AU 23	0.011	0.104	-	-
AU 24	0.015	0.122	-	-
AU 25	-	-	0.113	0.337
AU 26	-	-	0.012	0.109
Average	**0.0089**	**0.0923**	**0.0383**	**0.180**

the highest F1-binary scores across the literature, with an average F1-binary score of 0.75. Similarly, AU 2 is one of the lowest occurring AU and one of the lowest F1-binary scores across the literature, with an average F1-binary score of 0.36. Considering this, our analysis shows that current state-of-the-art results, in AU detection, follow an explicit trend which is the distribution of the AUs (i.e. the number of active AUs explicitly impacts the F1-binary score of the detection method). To further illustrate this trend, we also calculated the F1-binary score if we were to manually predict all 1's, for all frames (i.e. all AUs are active). As can be seen in Fig. 1, the general trend that the F1-binary scores, for all methods in BP4D, follow is the same as predicting all 1's.

The general trend that F1-binary scores follow can visually be seen in Figs. 1 and 2. To statistically analyze this trend, we calculated the correlation between the data distribution and F1-binary scores of the methods shown in Figs. 1 and 2. We define correlation as $corr = \frac{\sum (x-\overline{x})(y-\overline{y})}{\sqrt{\sum (x-\overline{x})^2 (y-\overline{y})^2}}$, where $\overline{x}$ and $\overline{y}$ are the averages of the distribution and the F1-binary scores, respectively. For BP4D and DISFA, the average correlations are 0.921 and 0.683, respectively. These results suggest that there is high correlation between the data distribution and the reported F1-binary scores for AU detection (Table I). Although there is a general trend of AU occurrence versus F1-binary accuracy, it is important to note there are some anomalies in the F1-binary scores for some AUs and methods. For example, on BP4D, Chu et al [5] use high intensity AUs along with a 3-class problem (i.e. +1/-1, and 0). In DISFA, some of the experiments train on BP4D and test on DISFA, which is a common approach for testing on this dataset, due to the imbalance of active versus inactive AUs. This can explain, in part, some of the lower correlations in BP4D and DISFA (Table I: [5], [11], [12]).

Along with the correlation between the data distribution and F1-binary scores, we also calculated the variance, $var = \frac{\sum (x-\overline{x})^2}{(n-1)}$, and standard deviation, $std = \sqrt{var}$, of the F1-binary scores between each of the methods detailed in Figs. 1 and 2. As shown in Table II, there is a small amount of variance between each of the methods across all studied AUs. In BP4D, the average variance is 0.0089 (std of 0.0923), and in DISFA the average variance is 0.0383 (std of 0.180). This suggests that the investigated F1-binary scores are within a small accuracy range, while following the data distribution trend. While the general variance and standard deviations are low, there are some outliers, especially in DISFA. For example, AU 9 have a variance of 0.064 and standard deviation of 0.253. This can also be visually seen in Fig. 2, with the F1-binary score of Li et al. [11]. As previously detailed, this can partially be explained from training on BP4D and testing on DISFA.

C. AU Patterns

From the AUs we investigated, for each dataset, we are interested in which AU patterns occur most often. We define a pattern as the active and inactive AUs for each frame. To investigate this, we looked at how many individual patterns exist, as well as the total frame count of each pattern (i.e. how many frames of data have that pattern). There are 1692 different patterns in BP4D and 265 in DISFA, that contain the investigated AUs for each dataset respectively. Tables III and IV show the 5 patterns with the highest, and the 2 patterns with the lowest frame counts in BP4D and DISFA, respectively.

As can be seen in Table III, there are some patterns that only appear 1 time across all frames, while there are patterns that occur over 6000 times across all frames. While there are less overall patterns in DISFA, there is still a large imbalance in the active and inactive AUs per frame. Similarly to BP4D, there are patterns that only occur in one frame, while there are many patterns that occur over 5000 times, including one pattern that occurs in 10,762 frames. In both datasets the pattern with the largest frame count is all 0's (i.e. no active AUs for that frame). This difference in the number of patterns contributes, at least

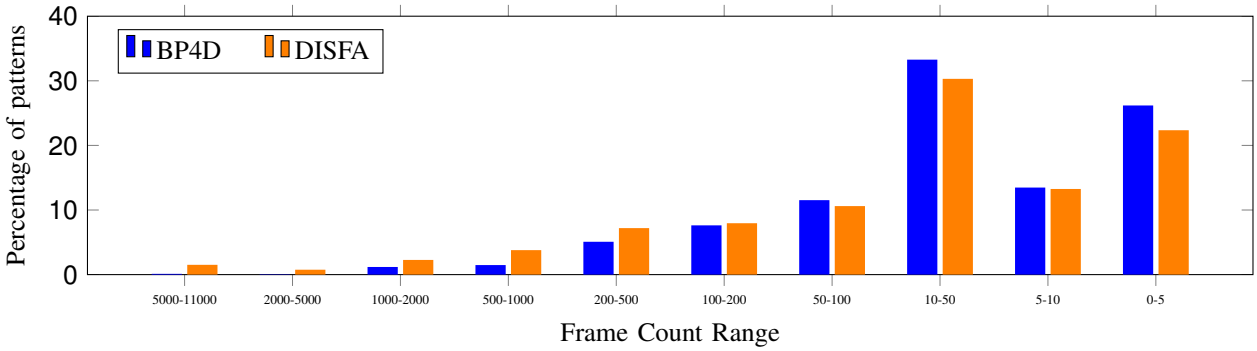


Fig. 3: Comparison of range of frame counts vs. percentage of patterns in that range for BP4D and DISFA.

TABLE III: AU patterns (active and inactive AUs per frame) from BP4D, along with total number of frames with that pattern. *NOTE: top 5 rows are patterns with highest counts; bottom 2 rows are patterns with lowest counts.*

Frame Count	Pattern											
	1	2	4	6	7	10	12	14	15	17	23	24
10630	0	0	0	0	0	0	0	0	0	0	0	0
8402	0	0	0	1	1	1	1	1	0	0	0	0
6883	0	0	0	1	1	1	1	0	0	0	0	0
3571	0	0	1	0	0	0	0	0	0	0	0	0
1814	0	0	0	0	0	1	1	0	0	0	0	0
1	0	0	1	0	1	1	0	0	1	1	1	0
1	1	1	0	1	1	1	1	0	1	1	1	1

TABLE IV: AU patterns (active and inactive AUs per frame) from DISFA, along with total number of frames with that pattern. *NOTE: top 5 rows are patterns with highest counts; bottom 2 rows are patterns with lowest counts.*

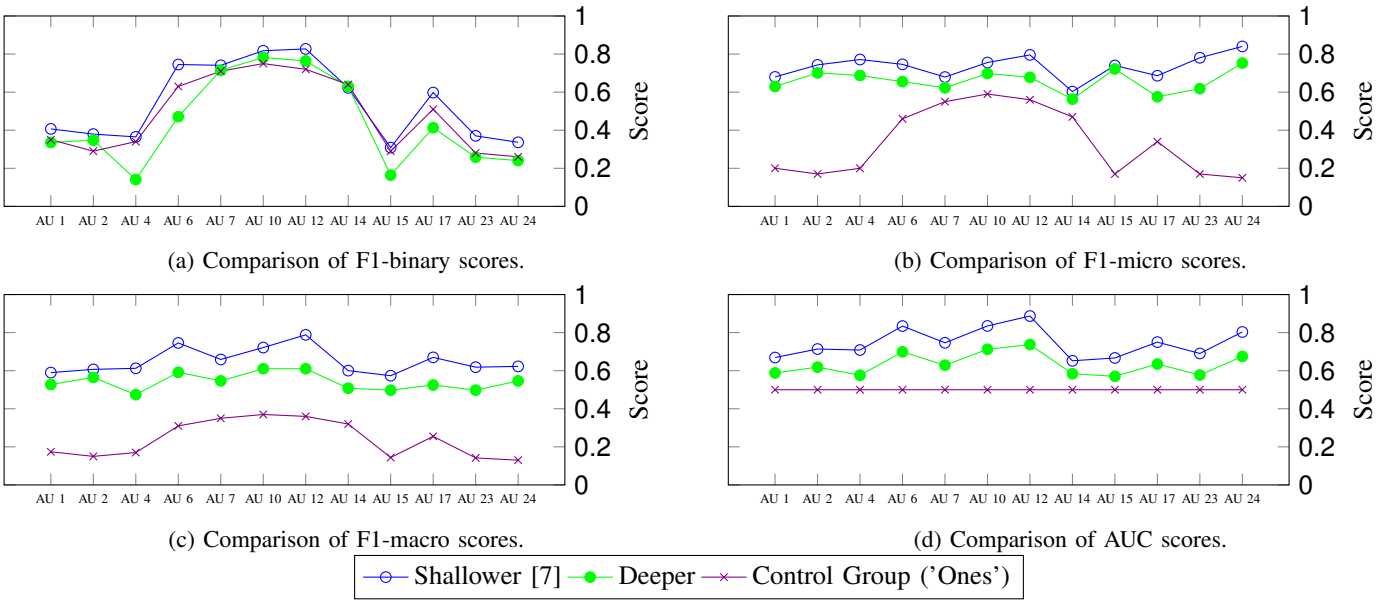
Frame Count	Pattern							
	1	2	4	6	9	12	25	26
66893	0	0	0	0	0	0	0	0
10762	0	0	0	0	0	0	1	0
7112	0	0	1	0	0	0	0	0
5222	0	0	0	0	0	1	1	0
5148	0	0	0	1	0	1	1	0
1	0	0	0	0	0	1	1	0
1	0	0	0	0	0	0	1	0

partially, to the imbalance of AUs in these datasets. It is also important to note that for both datasets, there is a similar trend in the percentage of patterns that exist in a range of frame counts (Fig. 3). For example, $>$72% of the patterns in BP4D and $>$65% of the patterns in DISFA occur $<$50 times in both datasets. On the other hand, $<$0.2% of the patterns in BP4D and $<$2% of the patterns in DISFA occur $>$5000 times.

III. Experiments and Results

To further investigate the impact of data distribution and patterns on AU detection, we conducted in-depth experiments on BP4D. We chose BP4D for our experiments as DISFA contains a large imbalance of active versus inactive AUs [12]. We evaluated the impact of two different convolutional neural networks (CNN) for this investigation (e.g. shallower versus deeper CNNs). First, we implemented the CNN as detailed by Ertugral et al. [7] for our shallower CNN, as this contains three convolutional layers and two fully connected. For our deeper CNN, we used a network which had two CNN layers with filter size of 8 and 16 followed by max pool layers, followed by two more CNN layers, with filter sizes of 16 and 20; another max pool layer and batch normalization. All CNNs used had a kernel of (3,3). There were three dense layers, before the output layer, with 4096, 4096 and 512 neurons respectively, relu activation function was used and dropout of 0.4. In addition to these two networks, we also had a control group called 'Ones', in which we detected all AUs as active in all the frames. This control group has a trend that largely follows the data distribution, and is used as a basis for comparisons for our shallower [7] and deeper networks.

To investigate the impact of the AU patterns on detection accuracy, we calculated the F1-binary, F1-micro, F1-macro, and Area Under the Curve (AUC) scores. F1 score is defined as $F1 = \frac{2 \times TruePositives}{2 \times TruePositives + FalsePositives + FalseNegatives}$ [2] and AUC is defined as the area under the graph between True positive rate V/S False positive rate. F1-binary is F1 score for the positive class and does not consider the negative class where as F1-macro is the simple average of F1 scores of all classes. F1-micro is the weighted average of the F1 score of all the classes, with more weight being given to the class which has a higher occurrence in the data. To facilitate our investigation, we conducted two experiments. First, using the entire dataset, we detected multiple AUs (i.e. the entire sequence/pattern). For this experiment, we used all AU-labeled frames from BP4D, and we refer to this as *experiment 1* in rest of paper. Secondly, we detected individual AUs by balancing the data to have an equal number of frames where the AU was active and inactive. We refer to this as *experiment 2* in the rest of the paper, and was done to test what impact balancing the data and removing the patterns has on AU the detection accuracy. Both experiments were subject-independent (i.e. same subject does not appear in training and testing), and the subjects in the each fold were fixed so that both experiments trained and tested the same subjects and images. We detected 12 AUs using three-fold cross-validation.



(a) Comparison of F1-binary scores.

(b) Comparison of F1-micro scores.

(c) Comparison of F1-macro scores.

(d) Comparison of AUC scores.

Fig. 4: Comparisons of different accuracy metrics on multiple networks (*experiment 1*).

TABLE V: Correlation between investigated metrics and data distribution for *experiment 1*.

Metric	Correlation		
	Shallower	Deeper	Ones
F1-binary	0.9758	0.9489	0.9912
F1-macro	0.7570	0.6349	0.9961
AUC	0.5795	0.6406	N/A
F1-micro	-0.2656	-0.3118	0.9913

A. Experiment 1 (Multi-AU detection)

When using the 'Ones' control group as a baseline, it can be seen that there is a high correlation between the data distribution and F1-binary, macro, and micro scores (Table V). There is an average correlation, with the data distribution, of .9928 across the three metrics. While the accuracies vary between the different metrics, it can be seen that the trend is similar (Fig. 4). For the control group, the AUC correlation is N/A as a score of 0.5 was obtained for each AU (Fig. 4d).

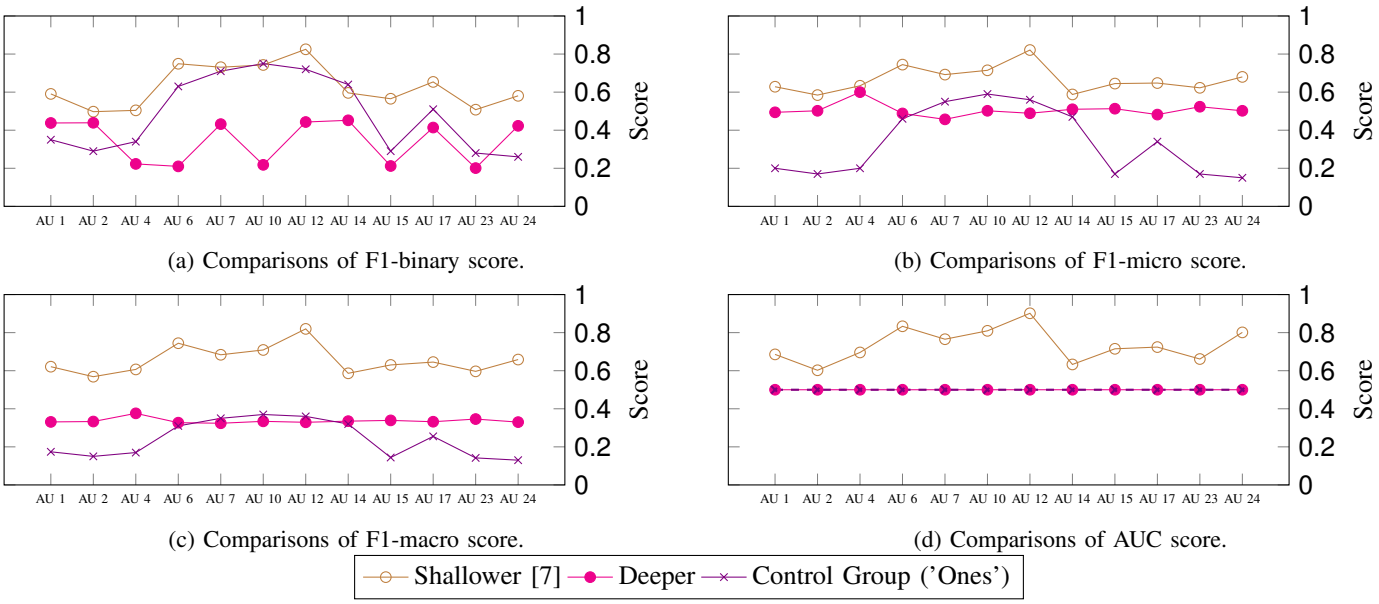
As can be seen in Fig. 4a, when comparing the F1-binary scores of the tested shallower [7] and deeper networks to the control group there is little difference. It can be seen that all three of them follow a similar trend, which is the distribution of the data (distribution can be seen in Fig. 1). This suggests that the F1-binary score may not be an accurate metric to distinguish between correct detection and guessing (i.e. "guessing" all AUs as ones/active). This can be explained, in part, since the F1-binary score only looks at the positive classes [2]. This can also be seen in Table V (first row), as there is a high correlation between the F1-binary score of both networks and the data distribution. We also calculated the correlation across each AU of both networks to the control group (i.e. how correlated are the F1-binary accuracies for each AU). This resulted in correlations of 0.98 and 0.94 for the shallower [7] and deeper networks, showing both give similar results to detecting all AUs as active.

We also looked at using F1-micro and macro as the metrics for AU detection accuracy. F1-micro does not follow the control group trend. It has a correlation of -0.29 and -0.31 for the shallower [7] and deeper networks with the control group (Fig. 4b). It also had a low negative correlation with the data distribution for both networks (Table V). This can be explained by F1-micro more heavily weighting the negative class for low occurring AUs. F1-macro also does not follow the control group trend (Fig. 4c). Although F1-macro is more correlated compared to F1-micro, with correlations of 0.74 and 0.62 for the shallower and deep networks, it is less correlated compared to F1-binary (Table V).

The final metric we looked at, for AU detection accuracy, was AUC. This metric has a lower correlation, with the data distribution, compared to F1-macro. It can also be seen that it does not follow the data distribution trend (Fig. 4d). Again, as AUC for this experiment was 0.5, we were unable to calculate the correlation between the control group and the shallower and deeper networks. It is also important to note that the correlations between the control group and F1-binary, macro, and micro closely resemble the correlation with the data distribution. This is due to the high correlation of the control group with the data (i.e. correlations are close to one).

B. Experiment 2 (Single-AU Detection)

To validate that *data distribution and patterns* have an impact on F1-binary scores we used our balanced data and trained separate networks for each individual AU (i.e. single AU detection). Similar to *experiment 1*, we calculated the correlations between the F1-binary, macro, micro and AUC scores compared to the data distribution for the shallower [7] and deeper networks (Table VI). In *experiment 1*, the



(a) Comparisons of F1-binary score.

(b) Comparisons of F1-micro score.

(c) Comparisons of F1-macro score.

(d) Comparisons of AUC score.

Fig. 5: Comparisons of different accuracy metrics on different networks (*experiment 2*). *NOTE: Best viewed in color - (d) has 2 overlapping lines as the deeper network and the control group has an AUC score of 0.5 for all AUs.*

TABLE VI: Correlation between Metric and data distribution (on balanced dataset and independent AU). *NOTE: Deeper network AUC N/A due to score of 0.5 for all AUs.*

Metric	Correlation	
	Shallower	Deeper
F1-binary	0.8783	0.1121
F1-macro	0.6869	-0.4281
AUC	0.5840	N/A
F1-micro	0.6290	-0.4679

correlations with the data distribution, for each metric, across the two network architectures was similar (e.g. F1-binary correlation of 0.9758 and 0.9489 for the shallower and deeper networks, respectively). Conversely, for *experiment 2*, the correlations, between the two networks, are not similar. For example, the correlation between F1-micro for the shallower network is 0.6290, while the deeper network has a correlation of -0.4679. Although the correlations differ between the two networks, in general F1-binary follows the data distribution, while the others do not (Fig. 5). Similar to *experiment 1*, the AUC for our control group was 0.5 for all AUs, however, the AUC for the deeper network was also 0.5 for all AUs (5d). This can be explained, in part, by the general performance of the deeper network as seen in Fig. 5. Each metric, for the deeper network, generally performed poorly which could indicate that it had difficulty learning individual AUs, resulting in an AUC of 0.5 (i.e. similar to a random guess of one for all frames).

For *experiment 2*, we *do not* claim this as a solution to the problems encountered when using F1-binary score as the accuracy metric for AU detection. This experiment was conducted to validate that AU patterns, as well as class imbalance contributes to the F1-binary scores following the data distribution trend. While the overall AU detection scores are lower for this experiment, the score was not the goal, but showing that the trend can be broken when patterns are removed and the classes are balanced. There are two major concerns with this experimental design being a solution. First, in a real-world setting balancing AUs is a difficult problem as many AUs are active at the same time as others (i.e. patterns). This causes balancing issues as balancing one AU can cause an imbalance in another AU. Second, and complimentary to the first concern, our results suggest that higher AU scores can be achieved with a multi-AU detection approach, especially as seen in our deeper network. These results validate other work that has shown the same thing [13].

IV. Conclusion

We have presented results that suggest data distribution and AU patterns (i.e. multiple active AUs), directly impact F1-binary scores causing a trend across multiple databases. We have reviewed state-of-the-art literature showing this trend exists across multiple works that make use of the F1-binary metric. We have also shown that this trend can be broken by removing the patterns (i.e. single AU detection), as well as balancing the AUs. Although this can help break the trend, we *do not* recommend it as a possible solution due to lower accuracies across multiple metrics, as well as the difficulty of balancing AUs in a real-world setting.

Our results suggest that different metrics besides for F1-binary could be a potential solution to breaking this trend. Considering this, we have detailed results of F1-micro, F1-macro, and AUC metrics. We conclude that the inclusion of all of these metrics for AU detection can give a better analysis and more confidence in the accuracy of results.

REFERENCES

[1] A. T. Chang, F-J .and Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni. Expnet: Landmark-free, deep, 3d facial expressions. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 122–129, 2018.

[2] Z. Chase Lipton, C. Elkan, and B. Narayanaswamy. Thresholding classifiers to maximize f1 score. *arXiv preprint arXiv:1402.1892*, 2014.

[3] J. Chen, Z. Chen, Z. Chi, and H. Fu. Facial expression recognition in video with multiple feature fusion. *IEEE Transactions on Affective Computing*, 9(1):38–50, 2018.

[4] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou. 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In *Computer Vision and Pattern Recognition*, pages 5117–5126, 2018.

[5] W. Chu et al. Learning spatial and temporal cues for multi-label facial action unit detection. In *Face and Gesture Recognition*, 2017.

[6] L. Duan, I. W. Tsang, D. Xu, and T.-S. Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 289–296. ACM, 2009.

[7] I. Ertugrul et al. Cross-domain au detection: domains, learning, approaches, and measures. In *Face and Gesture Recognition*, 2019.

[8] E. Friesen and P. Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3, 1978.

[9] J. M. Girard, J. Cohn, L. A. Jeni, S. Lucey, and F. De la Torre. How much training data for facial action unit detection? In *Face and Gesture Recognition*, volume 1, pages 1–8, 2015.

[10] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073. IEEE, 2012.

[11] W. Li, F. Abtahi, Z. Zhu, and L. Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 103–110. IEEE, 2017.

[12] W. Li, F. Abtahi, Z. Zhu, and L. Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2583–2596, 2018.

[13] W. Li et al. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Computer Vision and Pattern Recognition*, 2017.

[14] Y. Li, J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia. Mec 2017: Multimodal emotion recognition challenge.

[15] P. Liu, Z. Zhang, H. Yang, and L. Yin. Multi-modality empowered network for facial action unit detection. In *Winter Conference on Applications of Computer Vision*, pages 2175–2184, 2019.

[16] M. Mavadati, M. Mahoor, K. Bartlett, P. Trinh, and J. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[17] T. Nguyen, K. Tran, and H. Nguyen. Towards thermal region of interest for human emotion estimation. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 152–157, 2018.

[18] R. Picard. *Affective computing*. MIT press, 2000.

[19] A. Romero, J. León, and P. Arbeláez. Multi-view dynamic facial action unit detection. *Image and Vision Computing*, 2018.

[20] Q. Sun, R. Chattopadhyay, S. Panchanathan, and J. Ye. A two-stage weighting framework for multi-source domain adaptation. In *Advances in neural information processing systems*, pages 505–513, 2011.

[21] M. Valstar et al. Fera-2015 - second facial expression recognition and analysis challenge. In *Face and Gesture*, 2015.

[22] M. Valstar et al. Fera 2017 - addressing head pose in the third facial expression recognition and analysis challenge. In *Face and Gesture*, 2017.

[23] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *Face and Gesture*, pages 921–926, 2011.

[24] S. Wang, B. Pan, H. Chen, and Q. Ji. Thermal augmented expression recognition. *IEEE transactions on cybernetics*, 48(7):2203–2214, 2018.

[25] X. Wei, H. Li, J. Sun, and L. Chen. Unsupervised domain adaptation with regularized optimal transport for multimodal 2d+ 3d facial expression recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 31–37, 2018.

[26] P. Werner, S. Handrich, and A. Al-Hamadi. Facial action unit intensity estimation and feature relevance visualization with random regression forests. In *International Conference on Affective Computing and Intelligent Interaction*, pages 401–406, 2017.

[27] H. Yang, U. Ciftci, and L. Yin. Facial expression recognition by de-expression residue learning. In *Conference on Computer Vision and Pattern Recognition*, pages 2168–2177, 2018.

[28] J. Zeng, W.-S. Chu, F. De la Torre, J. F. Cohn, and Z. Xiong. Confidence preserving machine for facial action unit detection. In *Proceedings of the IEEE international conference on computer vision*, pages 3622–3630, 2015.

[29] X. Zhang, L. Yin, J. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[30] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, 2015.

[31] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.