HHS Public Access

Author manuscript

Pattern Recognit Lett. Author manuscript; available in PMC 2024 July 31.

Published in final edited form as:

Pattern Recognit Lett. 2024 June ; 182: 111–117. doi:10.1016/j.patrec.2024.04.016.

Time to retire F1-binary score for action unit detection

Saurabh Hinduja[a,*], **Tara Nourivandi**[b], **Jeffrey F. Cohn**[a], **Shaun Canavan**[b]

[a]Department of Psychology, University of Pittsburgh, Pittsburgh, USA

[b]Department of Computer Science and Engineering, University of South Florida, USA

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. To evaluate systems that detect action units, F1-binary score is often used as the evaluation metric. In this paper, we argue that F1-binary score does not reliably evaluate these models due largely to class imbalance. Because of this, F1-binary score should be retired and a suitable replacement should be used. We justify this argument through a detailed evaluation of the negative influence of class imbalance on action unit detection. This includes an investigation into the influence of class imbalance in train and test sets and in new data (i.e., generalizability). We empirically show that F1-micro should be used as the replacement for F1-binary.

Keywords

Action units; Data imbalance; Machine learning; F1 score

1. Introduction

In areas of machine learning, having measures that can exhibit truly and realistically how well a model works is the key to progress and success. Multiple measures and metrics have been used to evaluate classifiers in the past. One of these metrics is accuracy, which calculates the ratio of the number of correctly classified samples over the total number of samples. When used with imbalanced datasets, accuracy overestimates the power of the classifier when the majority classes are considered [1]. In other words, accuracy can still result in high performance when only the majority class is selected. Another metric used to evaluate classifiers is the F1-binary score (also known as F1 score). For this measure, two metrics of precision and recall are added to the calculation. If we count the results associated with the considered class as positive results and all others as negative results, we

*Corresponding author. SAH273@pitt.edu (S. Hinduja).

CRediT authorship contribution statement

Saurabh Hinduja: Conceptualization, Investigation, Methodology, Validation, Writing – original draft. **Tara Nourivandi:** Visualization. **Jeffrey F. Cohn:** Supervision, Writing – review & editing. **Shaun Canavan:** Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

can introduce precision and recall as the number of correct positive results over all positive results, and the number of correct positive results over all correct results respectively.

A reliable classifier predicts minority classes as accurately as majority classes in an imbalanced dataset. For real-world datasets, in which the number of samples can vary considerably among classes, the prediction accuracy is often high in majority and low in minority classes. Here, F1-binary acts poorly in providing an informative evaluation of classifiers. For instance, in “Mammography Data Set”, the number of healthy patients is much greater compared to cancerous ones and the F1-binary score measure does not reflect the real effectiveness of a classifier [3].

In many datasets that are commonly used for action unit (AU) detection, the available AUs are often imbalanced [4]. In these datasets, the number of samples in classes corresponding to some movements of lips, mouth, and eyes are noticeably greater than in other classes. The imbalance in these datasets makes F1-binary score an unsuitable measure to evaluate models that detect facial action units [5]. The datasets with imbalanced classes are difficult to learn from, as there are fewer positive samples, which results in skewed performance metrics. Some works have also used AUs for expression recognition. Tian et al. [6] used AUs that occer on the lower face for expression analysis. Liu et al. [7] constructed an AU facial graph called the deep action units graph network, for facial expression recognition. Yao et al. [8] used active learning along with a support vector machine to classify AUs for expression recognition. Shang et al. [9] use the co-occurrence of AU patterns and the attention to the region of occurrence of AUs.

There have also been encouraging works for mitigating class imbalance. Alvarez et al. [10] extend the capabilities of prototype based classifiers using fuzzy similarity relation to make them sensitive to class imbalanced data. GANS have also been used to generate synthetic data to help balance datasets [11]. Focal Loss [12] was developed for dealing with class imbalance. This loss function gives more weight to hard examples, which improves the performance of the minority class. Recently, Dablain et al. [13] introduced DeepSMOTE to generate synthetic images from deep features. This approach leverages SMOTE [14] along with features from an encoder/decoder network to generate the images. While these approaches address class imbalance from a model approach, we are motivated to address it from an evaluation metric approach as some researchers have questioned the use of F1-binary. This is due, in part, to how it treats precision and recall equally, when they are conceptually distinct [15]. To the best of our knowledge this is the first work to investigate the influence of class imbalance on F1-scores, and in train and test sets. We argue that it is time to retire F1-binary score for AU detection. We justify this argument through multiple experiments and contributions.

1. The negative influence of class imbalance on F1-score is detailed across three publicly available datasets that are commonly used for AU detection.
2. The influence of class imbalance in train and test sets, and in new data (i.e., generalizability) is investigated.

3. In order to retire F1-binary score, an appropriate metric is needed for replacement. We discuss multiple metrics and empirically show how F1-micro is what should be used as a replacement for F1-binary.

2. Class imbalance and action unit detection

2.1. Datasets

DISFA [34] is a spontaneous dataset designed for studying facial action intensity. It contains 27 (12 female and 15 male) adult subjects watching 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 frames (130, 815). For this dataset, the most commonly used AUs from the literature are AU1, AU2, AU4, AU6, AU9, AU12, AU25, AU26.

BP4D [4] is a multimodal facial expression dataset with a total of 41 subjects (23 female and 18 male) displaying 8 dynamic expressions (happy, surprise, sad, startled, skeptical, embarrassed, fear, and pain). We analyze (Section 3) all AU annotated frames (146,847) from this dataset. For this dataset, the most commonly used AUs from the literature are AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17, AU23, AU24.

BP4D+ [35] is a multimodal emotion-based dataset, similar to BP4D. It includes 140 subjects (82 female and 58 male) with age range of 18 to 66 years old. Ethnicities include 15 Hispanic, 64 Caucasian, 15 African American, 46 Asian, and 1 identified as other. Similar to BP4D, each subject participated in tasks, however, 10 tasks were used. In total 197 875 frames are used from this dataset, and the same AUs as BP4D(Fig. 3).

2.2. Class imbalance and AU detection

It has been shown that AU class imbalance has a negative impact on evaluation metrics [5]. To further extend this notion, we will show that class imbalance and F1-binary score are directly correlated. To facilitate this, we have analyzed state-of-the-art literature regarding their F1-binary scores from AU detection experiments on BP4D, DISFA, and BP4D+. As can be seen in Figs. 1–3, the occurrence of AUs in each dataset is imbalanced. For example, in BP4D, AU10 has an average occurrence of 0.62, while AU2 has an average AU occurrence of 0.18. There is a direct correlation between these occurrences and their F1-binary scores. AU10 is one of the highest occurring AUs and also one of the highest F1-binary scores across the literature, with an average F1-binary score of 0.75. Similarly, AU2 is one of the lowest occurring AUs 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 correlated to the imbalance of the AUs. 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 follow, for all methods in BP4D, is the same as labeling all 1's.

The general trend that F1-binary scores follow can visually be seen in Figs. 1–3. To statistically analyze this trend, we calculated the correlation between the class imbalance and F1-binary scores of the methods shown in Figs. 1, 2, and 3. We define correlation as

$corr = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2(y-\bar{y})^2}}$ where $\bar{x}$ and $\bar{y}$ are the averages of the classes and the F1-binary score, respectively. For BP4D, BP4D+, and DISFA the average correlations are 0.907, 0.76, and 0.843, respectively. These results suggest that there is high correlation between the class imbalance and the reported F1-binary scores for AU detection (Table 1). This further suggests that AUs with high F1-binary score have a lot of training data, and AUs with low F1-binary score have a small amount. Although there is a general trend of AU occurrence versus F1-binary score, it is important to note there are some anomalies from some AUs and methods. For example, on BP4D, Chu et al. use high intensity AUs, equal to or higher than A-level for positive samples, and the rest for negative. In DISFA, some of the experiments train on BP4D and test on DISFA, which is a common approach, due to the imbalanced occurrence of AUs. This can explain, in part, some of the lower correlations in Table 1 (e.g. [16,23,24]).

Along with the correlation between the class imbalance and F1-binary score, we also calculated the standard deviation of the F1-binary scores between each of the methods detailed in Figs. 1, 2, and 3. There is a small amount of variance between each of the methods across all studied AUs. In BP4D, the average variance (std) is 0.0079 (0.0890), 0.002 (0.0452) in BP4D+, and in DISFA the average variance is 0.0324 (0.18). This suggests that the investigated F1-binary score are all similar within a small range. While the general standard deviations are low, there are some outliers, especially in DISFA. For example, AU9 has a standard deviation of 0.253. This can also be visually seen in Fig. 2, with the F1-binary score of Li et al. [24]. This can again be partially explained from training on BP4D and testing on DISFA. See Table 2 for the standard deviation between all methods. This analysis naturally leads to the question of why is this specific trend occurring in AU detection? To answer this question, we next investigate the influence of action unit base rates, as well as the influence of different networks architectures.

3. Influence of AU base rates and network architectures on detection

Section 2 showed that there is a correlation between reported F1-binary scores and the AU base rates of the evaluated datasets. The base rate of a facial action unit is the proportion of time that an AU is present in a given database. The base rate is calculated by dividing the total duration of time that an AU is present by the total duration of time that the face is visible in the database [36]. For further analysis, to investigate the influence of base rates and different network architectures we conducted in-depth experiments on BP4D. We chose BP4D for our experiments, as DISFA and BP4D+ contain a larger imbalance of active versus inactive AUs [37]. We wanted to investigate the impact on multiple metrics, that are similar to F1-binary score. We also calculated the Negative agreement, F1-macro, and F1-micro. F1-binary score is defined as $F1-binary = \frac{2\times TP}{2\times TP+FP+FN}$ Negative agreement is defined as $NA = \frac{2\times TN}{2\times TN+FP+FN}$ F1-micro is defined as $F1_{micro} = 2 * \frac{precision_{micro} * recall_{micro}}{precision_{micro} + recall_{micro}}$ and F1-macro is defined as $F1_{macro} = \frac{F1-binary+NA}{2}$ where TP= True Positive, FP= False Positive, FN= False Negative, and TN= True Negative. To calculate the per-AU scores, in

our experiments, there are two classes which are positive and negative. More specifically, the presence of the AU is the positive class and absence of the AU is the negative class.

F1-macro is the simple arithmetic mean of F1-binary scores of all classes. **F1-micro** is the harmonic mean of the micro-average precision and recall [38]. F1-micro treats each sample equally, regardless of the class it belongs to. In micro-average, we compute the scores by aggregating the contributions of both classes. F1 micro is calculated using total number of True Positives (TP), False Positives (FP) and False Negatives (FN) across both the classes. To facilitate our investigation, we conducted two experiments. First, we investigate the impact of varying train and test set AU base rates on different metrics. For this experiment we chose the four AUs with the highest base rates (AU6, AU10, AU12, and AU17). Second, we investigate the impact of two different network architectures on these metrics. We also have a control group called ‘Ones’, in which we labeled all AUs as active in all the frames. Both experiments were subject-independent.

3.1. Impact of base rates

In this section, we investigate the impact of varying AU train and test base rates for F1-binary, F1-micro, F1-macro, and negative agreement. More specifically, this investigation will give us insight into which evaluation metric is more stable when imbalanced data is being evaluated (i.e., the varying base rates simulate more or less imbalance in the data). Figs. 4–7 show the corresponding evaluation metric when the training and testing base rates are varied. For example, in Fig. 4(a) for AU 6 and F1-binary, the first row shows the F1-binary score when we have 20% base rate for testing, and each column shows the varying training base rates (from 20% to 100%).

To facilitate our investigation into the impact of base rates, the data was split into an 80/20 subject-independent train/test split and the base rates, of the action units, were varied from 20% to 100% of the original base rate. To better understand the impact of base rates it is important that there is a significant change in the base rates. Considering this, the following AU base rates were used: [20%, 40%, 60%, 80%, 100%]. All combinations of training and testing base rates were considered. For example, 20% AU base rate training data was used and 20%, 40%, 60%, 80%, and 100% AU base rates were used for testing data. Same was done for all training and testing base rates. To facilitate this, the CNN architecture from Ertugral et al. [39] was implemented. This network has three convolutional layers followed by two fully connected layers. The convolutional layers have 64, 128, and 128 filters respectively each with a stride of 2. ReLU and batch normalization is applied to the output of each convolutional layer. Max pooling is then applied after batch normalization, which is then connected to a fully connected layer of 400 neurons followed by the output layer with 12 neurons. We modified the output layer to predict only the target AU. The target AUs, for this experiment, were selected for their high initial base rates: AU6, AU10, AU12 and AU17.

3.1.1. F1-Binary—When the base rates for both training and testing data are varied, F1-binary tends to have high variability that is negatively impacted by the change in base rate. For example, AU6 is susceptible to changes in base rate across the testing set. As can

be seen in Fig. 4, given a training base rate of 20%, as the AU test base rate increases from 20% up to 100%, the F1-binary generally increases (e.g., 0.36 F1-binary score at 20% up to 0.64 F1-binary score at 100% base rate). The F1-binary scores for AU10 vary across both training and testing. Here, training base rates appear to have a larger impact on F1-binary with an F1-binary score of 0.38 at 20% base rate, up to 0.7 at 100% base rate. An interesting base rate to note, for AU10, is the training AU base rate of 40%, where the F1-binary scores are all ≤0.093. This could be explained, in part, by a poor split in the data causing the network to not learn, as far as this metric is concerned. As can be seen in Fig. 4, the F1-binary scores for AUs 12 and 17, also vary across different AU base rates for both training and testing. This type of behavior can help give further justification for the results seen in Figs. 1–3, where the F1-binary scores are highly correlated with the AU base rates.

3.1.2. Negative agreement—Compared to F1-binary score, negative agreement (NA) has less variation across the different base rates for training and testing, as can be see in Fig. 5. There is some influence, however, the effect of test base rate is opposite to that on F1- binary - as the test set base rate increases the NA scores decreases. Although there is some decrease here, the amount is much less. For example, with AU6, given a training base rate of 20%, the NA is 0.82 when the test base rate is also 20%. When the test base rate increases to 100%, the NA lowers to 0.74. A change of 0.08, compared to F1-binary where the change is 0.28, from 20%–100% testing base rates, with 20% training base rate. Similar, however, to F1-binary, there are some interesting base rates where the NA drops significantly. For example, given AU10 and a 60% training base rate, the NA for all testing base rates is ≤0.096. This can again be explained, in part, by a bad split of the data resulting in the network not being able to learn, as far as this metric is concerned.

3.1.3. F1-macro—F1-macro is another metric that does not have such a significant variation across training and testing base rates, which can be seen in Fig. 6. For example, when varying the testing base rate of AU17 with 60% training base rate, the different between min and max F1-macro is 0.02. 100% testing base rate has an F1-macro of 0.59, 40% testing base rate has an F1-macro of 0.61. Another example of this can be seen in AU12, where 40% training base rate varies between 0.57 with 20% testing base rate to 0.63 at 60% and 80% testing base rate. Another interesting note about F1-macro is that this is the first metric, we have discussed in this paper, that do not have a base rate where the metric where all values are extremely low (e.g., ≤1). This gives some initial justification for F1-macro being a better evaluation metric compared to F1-binary and NA.

3.1.4. F1-micro—Similarly to F1-macro, F1-micro does not have a lot of variation when the base rates are changed (Fig. 7). However, compared to F1-macro, we can see where there are instance of varying the base rates, where there is almost no change in the F1-micro score. For example, given AU6 with 20% and 40% training base rates, the difference between the min and max scores, across all testing base rates, is 0.7 and 0.72, respectively. This can also be seen in AU17 with 60% and 80% testing base rates. The min and max F1-micros scores, across all training base rates, is 0.7 and 0.78, respectively. Sections 3.1.1–3.1.3, showed F1-macro had less variation, across base rates, compared to F1-binary, and NA.

Here, however, it can be seen (Fig. 7) that F1-micro further improves on F1-macro when varying base rates across both training and testing sets.

3.2. Impact of network architectures

In this section, we investigate the impact of network architectures (Fig. 8) on each of these metrics. To do this, we evaluated the impact of two different convolutional neural networks (CNN). First, we implemented the CNN as detailed by Ertugral et al. [39], which we refer to as *Network 1*. Second, we used a network with two convolutional layers with filter size of 8 and 16 followed by max pool layers, followed by two more convolutional 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. We refer to this as *Network 2*.

When using the 'Ones' control group as a baseline, it can be seen that there is a high correlation between the class imbalance and F1-binary, negative agreement, F1-macro, and micro scores (Table 3). There is an average correlation, with the class imbalance, of .9915 across the four metrics. While the accuracies vary between the different metrics, it can be seen that the trend is similar (Fig. 8). For *Network 1* and *Network 2*, it can be seen (Table 3) that F1-binary has a high correlation with the class imbalance. F1-macro is less correlated compared to F1-binary, however, it is still a relatively high correlation with an average of 0.7 across the two networks. Conversely, F1-micro has a negative correlation across both of the networks showing, again, that it is not as susceptible to the class imbalance that is found in AUs. This can be explained by F1-micro more heavily weighting the negative class for low occurring AUs. Similar to F1-micro, Negative agreement also has a negative correlation across both networks. In fact, it can be seen (Table 3) that it is more negatively correlated, to the class imbalance, compared to F1-micro. While these results could suggest Negative agreement is a better metric, taking into account the results shown in Section 3.1, F1-micro is better suited. More specifically, it is negatively correlated to the class imbalance *and* it has less variance across changes in training and test set base rates.

Further justification, for retiring F1-binary for AU detection, can be seen in Fig. 8(a). When comparing the F1-binary of the tested networks to the control group there is little difference. It can be seen that all of them follow a similar trend, which is the class imbalance. This suggests that F1-binary 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 F1-binary only looks at the positive classes. This can also be seen in Table 3 (first row), as there is a high correlation between the F1-binary of both networks and the class imbalance. We also calculated the correlation across each AU of both networks to the control group. This resulted in correlations of 0.98 and 0.94 for *Network 1* and *Network 2*, respectively, showing both give similar results to labeling all AUs as active. As can be seen in Fig. 8(d), unlike F1-binary, F1-micro does not follow the class imbalance. The F1-micro scores across each of the AUs is relatively stable (i.e., class imbalance does not impact it). These results further motivate and justify our argument that F1-micro is a suitable replacement for F1-binary. For these experiments, it is also important to note that

the correlations between the control group and all metrics closely resemble the correlation with the class imbalance. This is due to the high correlation of the control group with the data (i.e. correlations are close to one).

4. Limitations and future work

There are some limitations to our work. First, when analyzing base rates, we have only selected four (AU6, AU10, AU12, and AU17) with initially high base rates. A larger selection of AUs should be investigated with varying base rates (e.g., low and high initial base rates). Second, when analyzing networks, only two CNN-based architectures were evaluated. More varied network architectures should be evaluated. These include, but are not limited to, LSTM-based architectures [23], Transformer-based models [40], and other large-scale AU detection networks [41]. Lastly, three datasets where evaluated in this work, however, it can be argued that BP4D+ is an extension of BP4D (i.e., they are similar/the same). Considering this, larger and more varied AU detection datasets, as well as in-the-wild datasets, should be evaluated.

5. Conclusion

We have presented results showing that F1-binary is negatively influenced by class imbalance, and that varying the base rates of AU training and testing data also negative influences those scores. To facilitate these experiments, we reviewed state-of-the-art literature that evaluated DISFA, BP4D, and BP4D+ for AU detection. The reported F1-binary scores show that they are similar across a small range, and that each of the proposed methods have a high correlation between the class imbalance and the reported scores. When AU training and testing base rates are changed, F1-binary is highly susceptible to these changes. This influence can have significant impact on F1 scores. For example, high F1 scores can be reported for folds (e.g., cross-validation) where large amount of training and/or testing data is available for specific AUs. Conversely, those results could change dramatically given different folds where the AU base rates are changed (e.g., lower). An interesting finding from our work is that network architectures have comparatively little impact on AU detection scores. Results across literature, and our experiments (Section 3.2) show small changes in scores. This is due to the AU base rates having such a significant impact on the scores (Section 3.1).

Our results show that *F1-binary is not a suitable metric* to evaluate AU detection results. Considering this, we argue that F1-binary should be retired for AU detection and a suitable replacement should be used. We argue that F1-micro should be used as this replacement as it is least susceptible to class imbalance and varying base rates. Using the more stable metric, F1-micro, would allow for more fair comparisons in the literature. This is due to F1-micro being more robust to varying base rates has less impact compared to F1-binary.

Data availability

The authors do not have permission to share data.

References

[1]. Sokolova M, et al. , Beyond acc, F-score and ROC: A family of discrim measures for performance eval, in: Advances in AI, 2006.

[2]. Liu P, et al. , Multi-modality empowered network for facial action unit detection, in: WACV, IEEE, 2019.

[3]. He H, et al. , Learning from imbalanced data, IEEE Trans. Knowl. Data Eng (2009).

[4]. Zhang X, et al. , Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database, IVC (2014).

[5]. Jeni LA, et al. , Facing imbalanced data–recommendations for the use of performance metrics, in: ACII, IEEE, 2013.

[6]. Tian Y.-l., et al. , Recognizing lower face action units for facial expression analysis, in: FG, IEEE, 2000, pp. 484–490.

[7]. Liu Y, et al. , Facial expression recognition via deep action units graph network based on psychological mechanism, IEEE Trans. Cogn. Dev. Syst 12 (2) (2019) 311–322.

[8]. Yao L, Wan Y, Ni H, Xu B, Action unit classification for facial expression recognition using active learning and SVM, Multimedia Tools Appl. 80 (2021) 24287–24301.

[9]. Shang Z, Du C, Li B, Yan Z, Yu L, MMA-Net: Multi-view mixed attention mechanism for facial action unit detection, Pattern Recognit. Lett 172 (2023) 165–171.

[10]. Rodríguez Alvarez Y, García Lorenzo MM, Caballero Mota Y, Filiberto Cabrera Y, García Hilarión IM, Machado D de Oca Montes, Bello Pérez R, Fuzzy prototype selection-based classifiers for imbalanced data. Case study, Pattern Recognit. Lett 163 (2022) 183–190.

[11]. Sampath V, Maurtua I, Aguilar Martin JJ, Gutierrez A, A survey on generative adversarial networks for imbalance problems in computer vision tasks, J. Big Data 8 (2021) 1–59. [PubMed: 33425651]

[12]. Mukhoti J, et al. , Calibrating deep neural networks using focal loss, Adv. Neural Inf. Process. Syst 33 (2020) 15288–15299.

[13]. Dablain D, et al. , DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data, IEEE Trans. Neural Netw. Learn. Syst (2022).

[14]. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res 16 (2002) 321–357.

[15]. Hand DJ, et al. , F*: an interpretable transformation of the F-measure, Mach. Learn (2021).

[16]. Chu W-S, et al. , Learning spatial and temporal cues for multi-label facial action unit detection, in: IEEE FG, 2017.

[17]. Duan L, et al. , Domain adaptation from multiple sources via auxiliary classifiers, in: ICML, 2009.

[18]. Sun Q, et al. , A two-stage weighting framework for multi-source domain adaptation, Adv. NeurIPS (2011).

[19]. Gong B, et al. , Geodesic flow kernel for unsupervised domain adaptation, in: CVPR, IEEE, 2012.

[20]. Zeng J, et al. , Confidence preserving machine for facial action unit detection, in: ICCV, 2015.

[21]. Zhao K, et al. , Joint patch and multi-label learning for facial action unit detection, in: CVPR, IEEE, 2015.

[22]. Zhao K, et al. , Deep region and multi-label learning for facial action unit detection, in: CVPR, IEEE, 2016.

[23]. Li W, et al. , AU detection with region adaptation, multi-labeling learning and optimal temporal fusing, in: CVPR, IEEE, 2017.

[24]. Li W, et al. , Eac-net: A region-based deep enhancing and cropping approach for facial AU detection, in: FG, IEEE, 2017.

[25]. Ertugrul I, et al. , D-pattnet: Dynamic patch-attentive deep network for AU detection, Front. Comput. Sci (2019).

[26]. Corneanu C, et al. , Deep structure inference network for facial action unit recognition, in: ECCV, 2018.

[27]. Shao Z, et al. , Deep adaptive attention for joint facial action unit detection and face alignment, in: ECCV, 2018.

[28]. Shao Z, et al. , JAA-Net: Joint facial action unit detection and face alignment via adaptive attention, IJCV (2021).

[29]. Shao Z, et al. , Facial action unit detection using attention and relation learning, TAC (2019).

[30]. Yang J, et al. , FAN-Trans: Online knowledge distillation for facial action unit detection, in: CVPR, IEEE, 2023.

[31]. Tallec G, et al. , Multi-order networks for action unit detection, IEEE TAC (2022).

[32]. Miriam Jacob G, et al. , Facial action unit detection with transformers, in: CVPR, IEEE, 2021.

[33]. Song T, et al. , Uncertain graph neural networks for facial action unit detection, in: AAAI, 2021.

[34]. Mavadati SM, et al. , Disfa: A spontaneous facial action intensity database, IEEE TAC (2013).

[35]. Zhang Z, et al. , Multimodal spontaneous emotion corpus for human behavior analysis, in: CVPR, IEEE, 2016.

[36]. Ambadar Z, et al. , Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions, Psychol. Sci 16 (2005) 403–410. [PubMed: 15869701]

[37]. Li W, et al. , Eac-net: Deep nets with enhancing and cropping for facial action unit detection, PAMI (2018).

[38]. Takahashi K, et al. , Confidence interval for micro-averaged F 1 and macro-averaged F 1 scores, Appl. Intell (2022).

[39]. Ertugrul IO, et al. , Cross-domain au detection: Domains, learning approaches, and measures, in: FG, IEEE, 2019.

[40]. Jacob GM, et al. , Facial action unit detection with transformers, in: CVPR, IEEE, 2021.

[41]. Jyoti S, et al. , Expression empowered residen network for facial action unit detection, in: FG, IEEE, 2019.

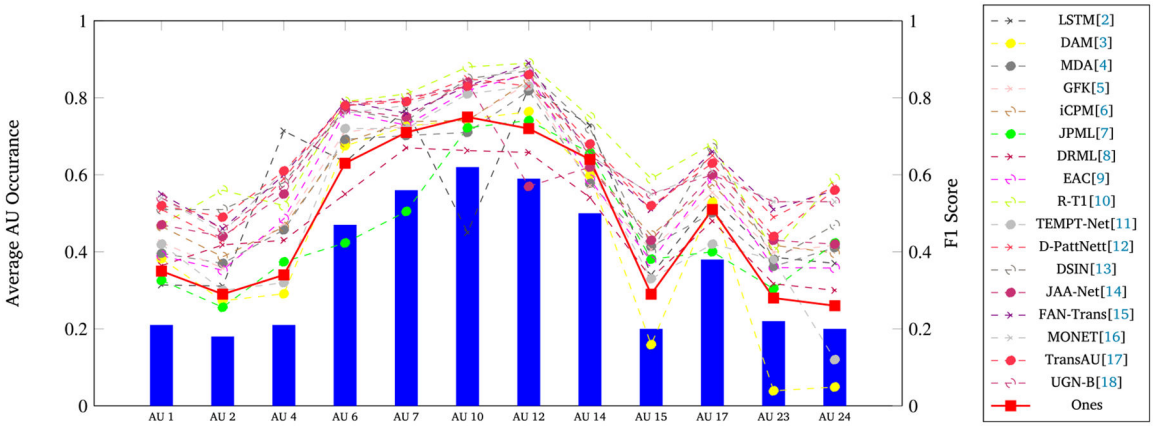

Fig. 1.

AU detection F1-binary score vs. occurrence in BP4D. Bars are the average number of AU occurrences, per frame, across all subjects. Line graphs are different F1-binary scores, of methods in the literature, for each AU. “Ones” is manually labeling all 1’s (i.e. AU is active) for each of the AUs. Here, we can see that each of the compared works have similar trends. In other words, the AUs with a higher average occurrence result in a higher F1-binary score, while the AUs with the lower average occurrence result in a lower F1-binary score (see [2]).

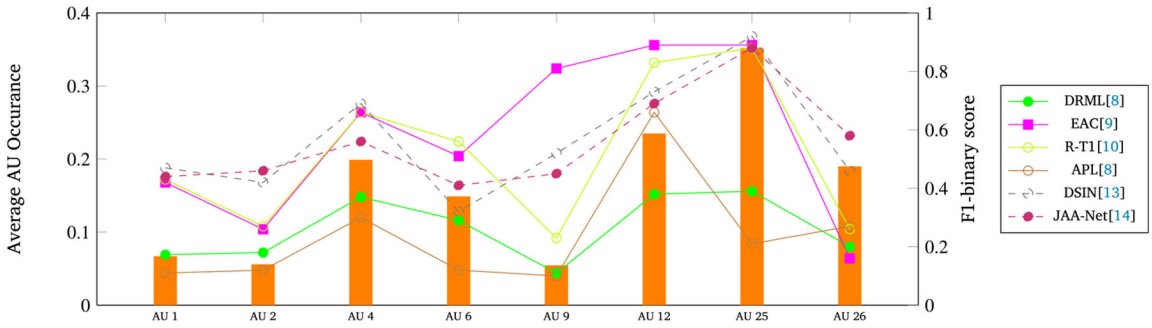

Fig. 2.

AU detection F1-binary score vs. occurrence in DISFA. Bars are the average number of AU occurrences, per frame, across all subjects. Line graphs are different F1-binary scores, of methods in the literature, for each AU. The scales on the left and right are different due to the low number of active AUs compared to some of the F1-binary scores. Here, we can generally see that each of the compared works have similar trends. In other words, the AUs with a higher average occurrence result in a higher F1-binary score, while the AUs with the lower average occurrence result in a lower F1-binary score.

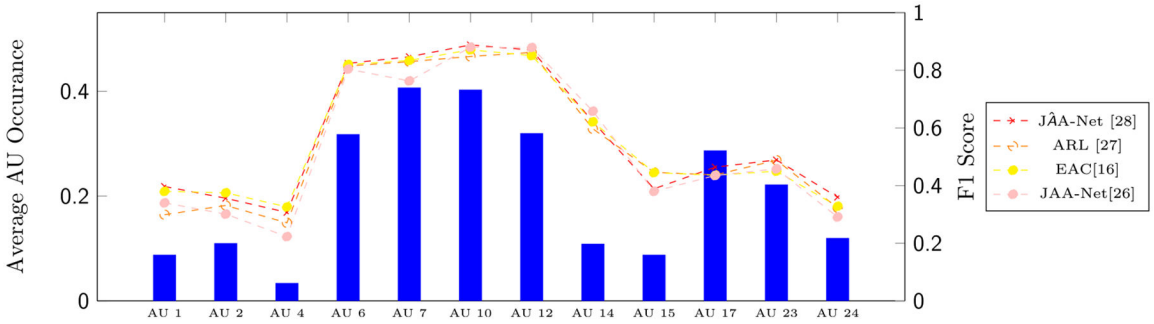

Fig. 3.

AU detection F1-binary score vs. occurrence in BP4D+. Bars are the average number of AU occurrences, per frame, across all subjects. Line graphs are different F1-binary scores, of various methods, for each AU. The scales on the left and right are different due to the low number of active AUs compared to some of the F1-binary scores. The AUs with a higher average occurrence result in a higher F1-binary score, while the AUs with the lower average occurrence result in a lower F1-binary score.

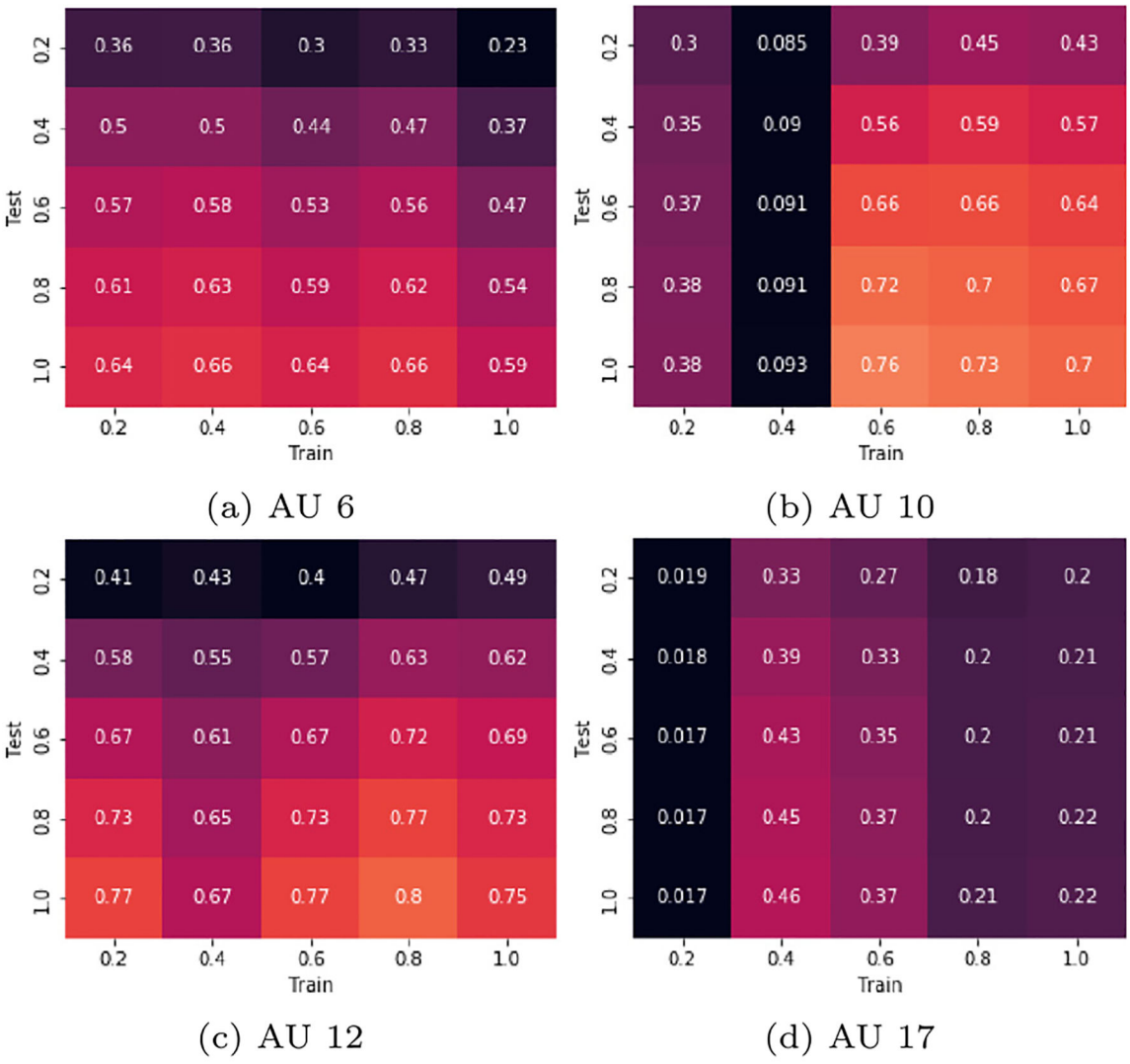

Fig. 4.

F1-binary metrics. Columns represent varying train base rates from 20% to 100%. Rows represent varying test base rates from 20% to 100%. These matrices show the F1-binary score for each combination of train and test base rates. For example, the top left corner is the F1-binary score with 20% base rate for training and testing data.

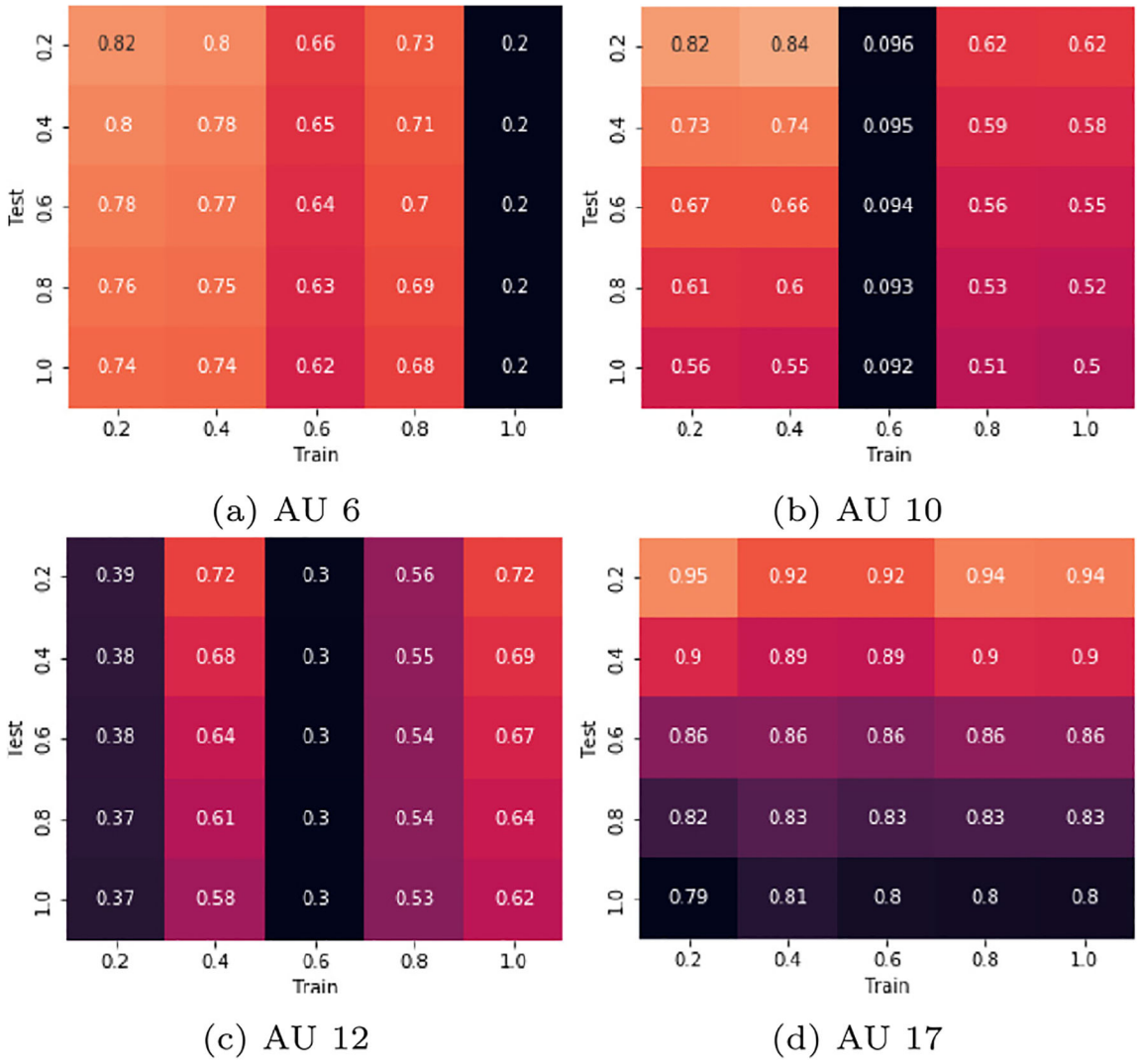

Fig. 5.

Negative agreement metrics. Columns represent varying train base rates from 20% to 100%. Rows represent varying test base rates from 20% to 100%. These matrices show the negative agreement for each combination of train and test base rates. For example, the top left corner is the negative agreement score with 20% base rate for training and testing data.

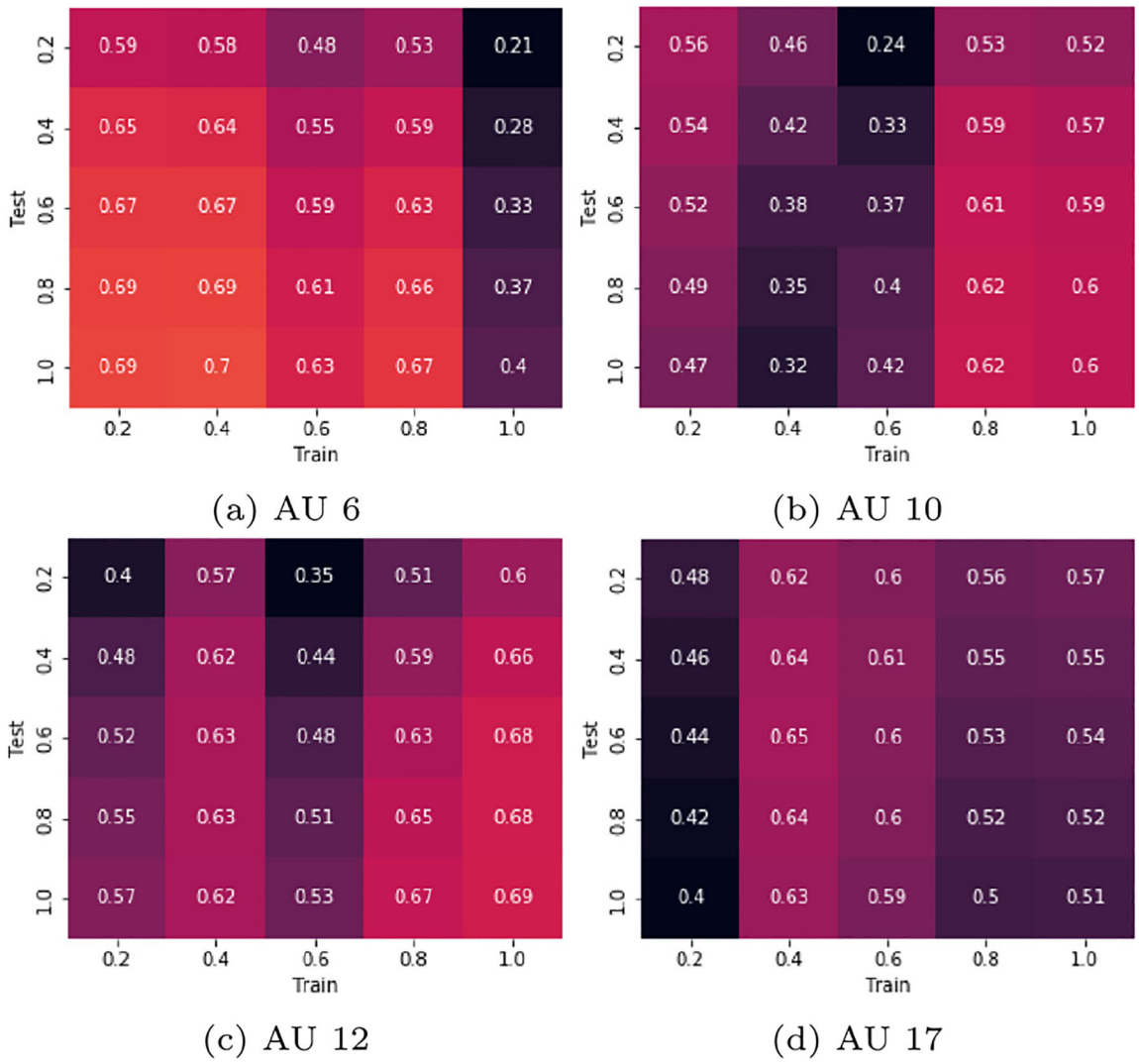

Fig. 6.

F1 macro metrics. Columns represent varying train base rates from 20% to 100%. Rows represent varying test base rates from 20% to 100%. These matrices show the F1-macro score for each combination of train and test base rates. For example, the top left corner is the F1-macro score with 20% base rate for training and testing data.

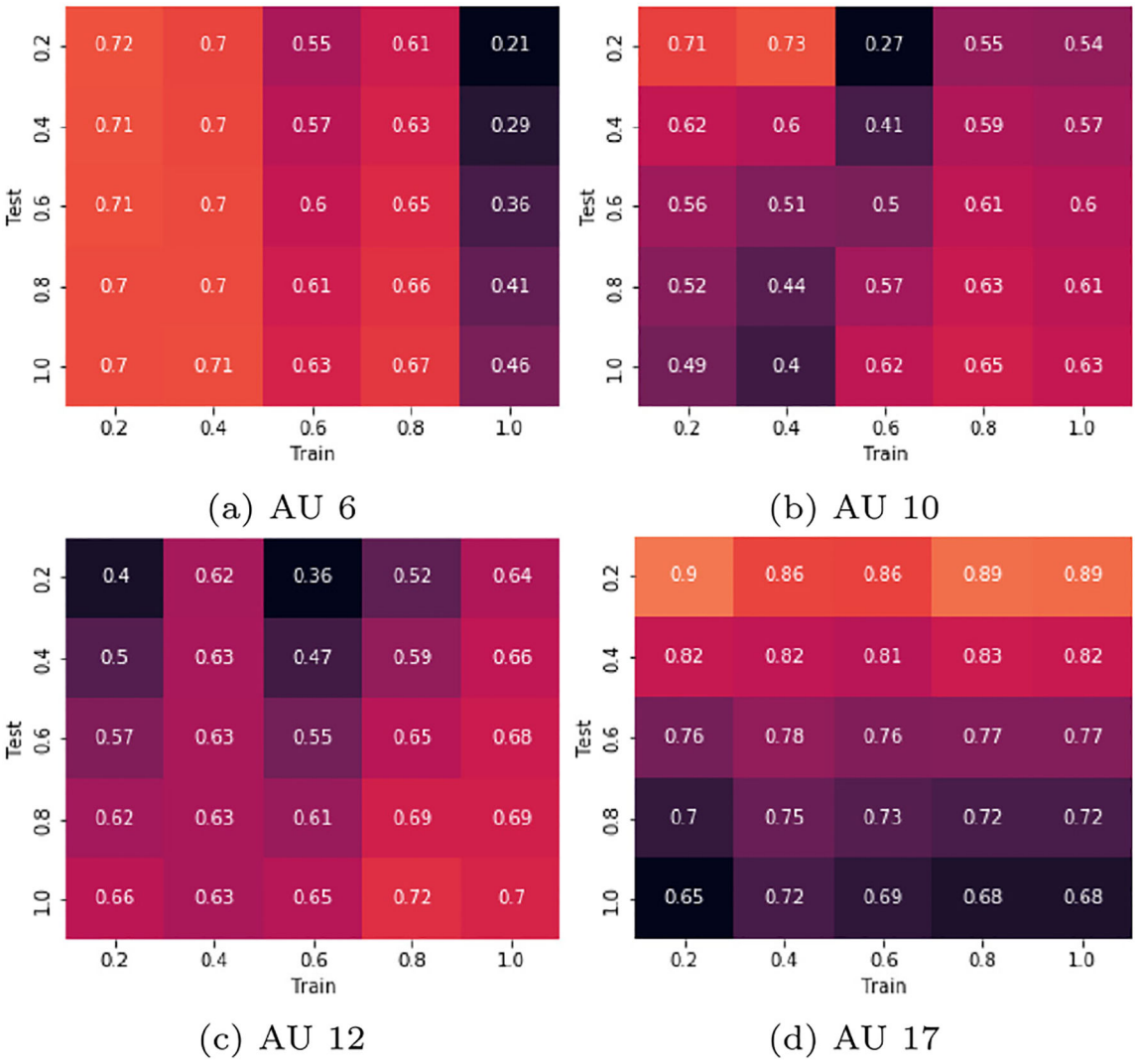

Fig. 7.

F1 micro metrics. Columns represent varying train base rates from 20% to 100%. Rows represent varying test base rates from 20% to 100%. These matrices show the F1-micro score for each combination of train and test base rates. For example, the top left corner is the F1-micro score with 20% base rate for training and testing data.

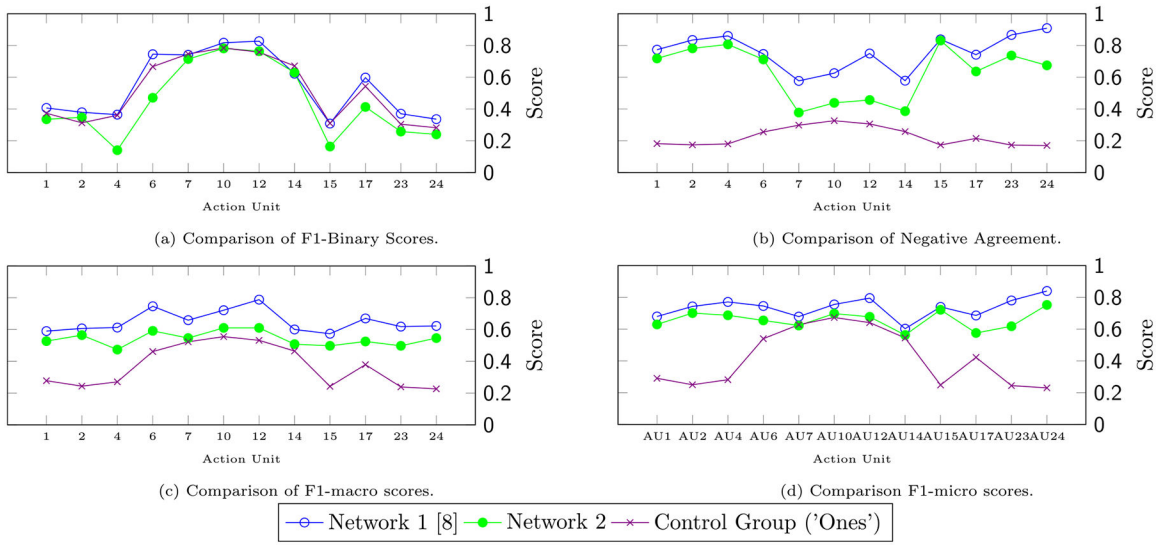

Fig. 8.

Impact of network architecture on different evaluations metrics. It can be seen that F1-binary has a similar trend as seen in Figs. 1–3, for networks 1 and 2, and the control group. Conversely, F1-micro, does not have the same trend. It can be seen that networks 1 and 2 have higher, and more stable, scores across all AUs.

Table 1

Correlation between F1-binary score and AU class imbalance, for state-of-the-art methods on BP4D, BP4D+, and DISFA dataset. '–' means dataset not evaluated by cited work.

Method	Correlation		
	BP4D	DISFA	BP4D+
LSTM [16]	0.680	–	–
DAM [17]	0.922	–	–
MDA [18]	0.948	–	–
GFK [19]	0.951	–	–
iCPM [20]	0.967	–	–
JPML [21]	0.869	–	–
DRML [22]	0.949	0.844	–
FVGG [23]	0.890	0.785	–
EAC [24]	0.953	0.472	0.84
R-T1 [23]	0.931	0.816	–
APL [22]	–	0.5098	–
D-PattNett [25]	0.946	–	–
DSIN [26]	0.931	0.792	–
JAA-Net [27]	0.847	0.918	0.833
JÂA-Net [28]	0.863	0.774	0.87
ARL [29]	0.808	0.954	0.827
FAN-trans [30]	0.932	–	–
MONET [31]	0.93	–	–
TransAU [32]	0.94	–	–
UGN-B [33]	0.95	–	–
Average	**0.907**	**0.76**	**0.843**

Table 2

Standard deviation of F1-binary, for each individual AU, between all evaluated methods across the BP4D, BP4D+, and DiSFA. Evaluated methods are the same as seen in Figs. 1–3. As can be seen in this table, the standard deviation is low between all evaluated approaches, across all datasets. This suggests that each of the approaches have a similar F1-binary score for each AU on BP4D, BP4D+, and DISFA. This can be seen across all AUs as the average standard deviation is low for each dataset.

AU	Standard deviation		
	BP4D	DISFA	BP4D+
AU 1	0.0671	0.1418	0.0759
AU 2	0.0853	0.1301	0.0613
AU 4	0.1197	0.1702	0.0825
AU 6	0.0935	0.1559	0.0203
AU 7	0.0660	–	0.0418
AU 9	–	0.2541	–
AU 10	0.1006	–	0.0214
AU 12	0.0917	0.1455	0.0216
AU 14	0.0885	–	0.0979
AU 15	0.0965	–	0.0524
AU 17	0.0847	–	0.0191
AU 23	0.0663	–	0.0202
AU 24	0.1076	–	0.0282
AU 25	–	0.3172	–
AU 26	–	0.1573	–
Average	**0.0890**	**0.1840**	**0.0452**

Table 3

Correlation between F1-binary, Negative Agreement, F1-macro, and F1-micro with class imbalance. This is done for two different networks and the "Ones' control group. As can be seen in this table, different network architectures have similar correlations. For example, F1-binary has a high correlation across both networks, and F1-micro has a negative correlation across both networks. This suggests that the specific network does not have a high impact on the metrics obtained and the class imbalance plays a larger role.

Metric	Correlation		
	Network 1	Network 2	Ones
F1-binary	0.9758	0.9489	0.9912
NA	−0.8359	−0.9031	0.9874
F1-macro	0.7570	0.6349	0.9961
F1-micro	−0.2656	−0.3118	0.9913