Handling Data Imbalance in Automatic Facial Action Intensity Estimation

Philipp Werner
Philipp.Werner@ovgu.de
Frerk Saxen
Frerk.Saxen@ovgu.de
Ayoub Al-Hamadi
Ayoub.Al-Hamadi@ovgu.de

Otto von Guericke University
Magdeburg, Germany
www.iikt.ovgu.de/nit

Abstract

Automatic Action Unit (AU) intensity estimation is a key problem in facial expression analysis. But limited research attention has been paid to the inherent class imbalance, which usually leads to suboptimal performance. To handle the imbalance, we propose (1) a novel multiclass under-sampling method and (2) its use in an ensemble. We compare our approach with state of the art sampling methods used for AU intensity estimation. Multiple datasets and widely varying performance measures are used in the literature, making direct comparison difficult. To address these shortcomings, we compare different performance measures for AU intensity estimation and evaluate our proposed approach on three publicly available datasets, with a comparison to state of the art methods along with a cross dataset evaluation.

1 Introduction

Facial expression is a central part of non-verbal communication. It reveals information on the affective state of an observed person, which can be used in e.g. pain assessment [28], drowsy driver detection, marketing or human-robot interfaces [25].

The Facial Action Coding System (FACS) [6] is a widely used method for describing and analyzing facial expressions. Based on muscles, it specifies a set of facial movement building blocks that are called action units (AUs). A trained FACS coder decomposes a potentially complex facial expression into the occurring AUs (e.g. see AU 12 and 25 in Fig. 1). Most AUs cannot only be coded regarding their occurrence or absence, but also regarding their intensity. The current version of FACS defines five ordinal intensities, which are usually denoted by the letters A (trace) to E (maximum). In the following we denote these intensities with the numbers 1 to 5, and further add 0 to denote the absence of the respective action unit. Facial expression intensity is linked to the intensity of emotional experiences and essential to analyze the facial dynamics, which e.g. is relevant to assess the authenticity of facial expressions (see [7] for more examples).

Databases Currently there are a few databases that include AU intensity labels according to the FACS 5-point intensity scale. The most relevant regarding the number of coded frames

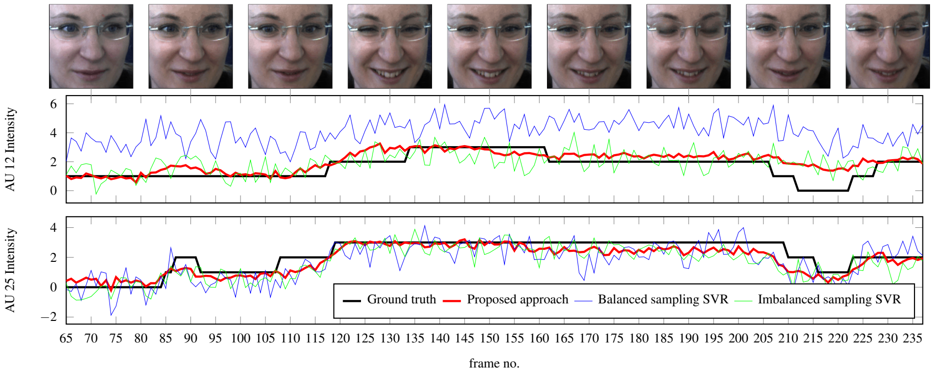


Figure 1: Intensity estimation on test set for AU 12 (lip corner pulling) and 25 (lip parting) shown for a sequence of frames from subject 10 (DISFA dataset). Our proposed approach, the MIDRUS SVR ensemble (red), is compared with balanced sampling SVR used by [7] (green), imbalanced sampling SVR used by [9, 22] (blue), and ground truth (black).

Dataset	Type	Subjects	AUs	Frames	- (0)	A (1)	B (2)	C (3)	D (4)	E (5)
Bosphorus [21]	images / posed	105	33	2,902	90.5%	1.6%	3.3%	2.4%	1.5%	0.6%
DISFA [16]	video / spontaneous	27	12	130,814	87.4%	4.5%	3.5%	3.2%	1.0%	0.3%
UNBC-McMaster [15]	video / spontaneous	25	10	48,398	95.7%	1.4%	1.4%	1.0%	0.5%	0.1%

Table 1: Databases. Number of subjects, AUs, and frames with intensity coding. Distribution of intensities (mean across AUs). Absence is dominant.

and AUs are summarized in Table 1: the Bosphorus database [21], the DISFA database [16], and the UNBC-McMaster database [15]. They differ in the number of subjects, frames, coded AUs, the availability of temporal context, the expression elicitation method, the variability of head poses, and other aspects. Another resource is the recently published BP4D-Spontaneous database [31], but it only comes with intensities for few AUs (5 AUs since FERA 2015 [26]).

An inherent challenge with all AU intensity databases is the imbalance between classes. The lack of a certain facial action is more frequent than its occurrence. If we consider the mean intensity distribution across AUs (see Table 1), about 90% of the samples account for the absence class (intensity zero). The remaining 10% of occurrences split up into five intensities; several classes account for less than 1% of the samples. Most of the AUs occur even less frequently than in this average distribution.

Related Work Standard machine learning methods are often biased towards the majority class, which leads to high misclassification for the minority class [14]. Common solutions can be categorized into three major groups: (1) sampling, (2) cost-sensitive learning, and (3) ensemble techniques. Sampling methods modify the training data to balance the classes; they allow the use of arbitrary classifiers. A common approach is random under-sampling, a method that randomly selects a subset of samples from the majority class to balance the training data. Cost-sensitive learning adjusts the penalties of false positives and false negatives in the learning algorithm. Ensemble techniques train and combine multiple classifiers using (a) sampling strategies or (b) cost-sensitive learning methods. Studies have shown that sampling methods and ensemble strategies outperform cost-sensitive learning because defining an optimal cost-matrix is often the bottleneck [14]. SMOTE [4] is a widely used state of the art minority over-sampling technique, but it is impracticable for large datasets because training time significantly increases. EasyEnsemble [13] is a under-sampling based ensemble method and has been shown to compete with SMOTE in several evaluations [13, 14, 24]. Most of the methods for handling imbalance are designed for binary classification; a multi-class problem is usually reduced to multiple binary problems.

Jeni *et al*. [10] studied the influences of highly imbalanced data on performance measures for action unit recognition (binary classification). Action unit intensity estimation however uses different performance measures not covered in their paper. To the best of our knowledge, no previous work in facial action intensity estimation addresses the imbalance problem adequately, as most works focus on features and machine learning techniques. Girard *et al*. [7] use random under-sampling to roughly balance the training set without further analysis. Sandbach *et al*. [20] reduce imbalance by under-sampling the absence class; they chose to take five times more AU absence samples than the sum of all other classes, but also without further analysis. Rudovic *et al*. [19] and Yang *et al*. [30] pre-segment the database and group several intensities to reduce imbalance, which we discuss in Sec. 3.2. Other authors [1, 22] exclude the absence-class (0) during the intensity estimation, which balances the data but assumes a perfect AU detection before intensity estimation. Many works seem to ignore the imbalance problem and either train with all available data [9, 11, 17] or use sampling methods that keep the imbalance [16, 18].

For classification, Support Vector Machines (SVM) [7, 16, 18] and probabilistic graphical models [17, 19, 20, 30] are often used. Regression models provide a continuous estimation, whereas Support Vector Regression (SVR) [7, 9, 20, 22] has shown good performance. Other authors employ Relevance Vector Regression [11] and Logistic Regression [1]. Often used features include landmarks and geometric features, [1, 11, 18, 19, 30], Gabor-filters [7, 16, 17, 22], and local binary pattern histograms (LBP) [11, 16, 20].

All mentioned works approach the challenging task of action unit intensity estimation. For a more general literature review on facial expression recognition, see [25, 27].

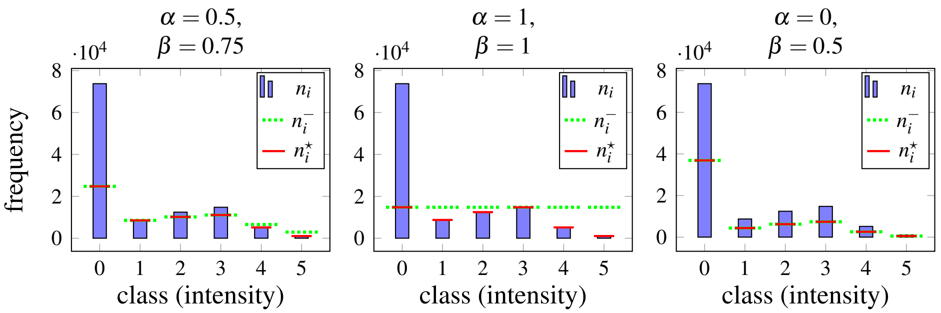
Contributions We propose a novel under-sampling method (MIDRUS, see Sec. 2.1), apply it in an ensemble (Sec. 2.2), and discuss the use of a proper performance measure for AU intensity estimation (Sec. 3.1). The parameters of our proposed method are analyzed in Sec. 4.1. We report the performance of our fully automatic, person-independent approach along with a state of the art comparison (Sec. 4.2) and a cross dataset evaluation (Sec. 4.3).

2 Handling Imbalance During Training

In this Section we propose two new methods to handle imbalance problems: (1) a novel under-sampling method that reduces imbalance and (2) an ensemble method that uses the proposed sampling. Both methods can be applied directly for multiclass problems. Further, they can be combined with various classification and regression techniques.

2.1 Multiclass Imbalance Damping Random Under-Sampling

On the one hand strong imbalance decreases performance on the minority class(es), and on the other hand under-sampling may drop relevant information about the majority class(es). We propose to choose a compromise. Instead of removing the imbalance or ignoring it,



(a) damping **(b) ‘balanced’** **(c) imbalanced**

Figure 2: Multiclass Imbalance Damping Random Under-Sampling (MIDRUS) examples. (a) With $0 < \alpha < 1$, MIDRUS damps imbalance, which improves performance (see Fig. 3). (b) With $\alpha = 1$ (maximum damping) it is equivalent to random under-sampling that balances the two most frequent classes. (c) With $\alpha = 0$ (no damping) it is equivalent to stratified sampling and keeps the imbalance. The number of samples to select is adjusted by β.

we reduce it with a method that we call *Multiclass Imbalance Damping Random Under-Sampling (MIDRUS)*. It is an algorithm with two steps: (1) calculating the number of samples to select from each class, and (2) randomly under-sample the classes without repetition according to the counts calculated in step (1).

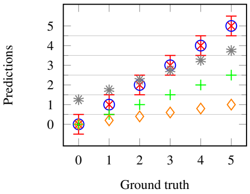
Given that we have M classes $i = 1, \ldots, M$ and n_i is the absolute frequency of class i in the dataset, then the number of samples $n_i^\star$ to select from class i is calculated as follows.

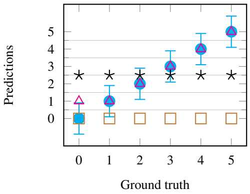
$$n_i^- = \lceil s \cdot (n_i)^{1-\alpha} \rceil, \text{with } s = \beta \frac{n_{f(k)}}{(n_{f(k)})^{1-\alpha}}, \tag{1}$$

$$n_i^\star = \min\{n_i, n_i^-\}. \tag{2}$$

In (1), $\alpha \in [0,1]$ is the imbalance damping parameter. It controls to which extend the imbalance is reduced, i.e. $\alpha = 1$ aims at total balancing of classes, $\alpha = 0$ keeps the imbalance, and an α in between reduces it to a certain degree. With $\alpha > 0$, the term $(n_i)^{1-\alpha}$ calculates new and more balanced class ratios. Next, these are scaled by a common factor s, which controls the total number of samples to be selected. It firstly depends on two parameters: $k \in \{2, \ldots, M\}$ and $\beta \in (0,1]$. Further, the definition of s uses a sorting function $f(k)$ returning the k'th most frequent class. Then β is the sampling fraction of the k'th most frequent class. Usually, k and β are easy to select. We recommend to set k to the number of majority classes plus one, i.e. in the following we set $k = 2$, as in the problem domain of facial action unit intensity estimation there is only one majority class (absence a of facial action). The parameter β should usually be set to one to avoid that minority class samples are discarded. But you may choose $\beta < 1$ due to different reasons, e.g. to reduce the training time or to increase the variance of models in an ensemble (see Sec. 2.1).

Fig. 2 illustrates MIDRUS with three examples. Sub-figure (a) shows a typical use-case with $\alpha = 0.5$; the imbalance is damped by taking the square root of sample counts and scaling the results in a way that 75% of the second most frequent class's samples are selected ($\beta = 0.75$). Two special cases are illustrated in (b) and (c), respectively. With $\alpha = 1$ and $\beta = 1$, MIDRUS balances with the second most frequent class, which is a quite typical approach in the state of the art. With $\alpha = 0$, it is equivalent to stratified random sampling. In





Prediction	MSE$_c^\mu$	MSE$_d^\mu$	MAE$_d^M$	PCC$_c$	PCC$_d$	ICC(3,1)$_d$	ICC(1,1)$_c$	ICC(1,1)$_d$	F1^M
	[9, 11, 20]	[20]	[19, 30]	[9, 11, 20, 22]	[20]	[16, 17, 18, 19, 30]	[7]	[7]	[19, 30]
○ perfect	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
× perfect + noise 0.5	0.251	0.182	0.266	0.846	0.882	0.880	0.835	0.867	0.645
● perfect + noise 0.9	0.812	0.492	0.546	0.664	0.743	0.733	0.611	0.683	0.406
+ 0.5 · perfect	0.176	0.119	1.000	1.000	0.976	0.875	0.780	0.865	0.287
✳ 0.5 · perfect + 1.25	1.406	0.933	0.667	1.000	0.976	0.875	0.008	0.277	0.245
◇ 0.2 · perfect	0.452	0.448	2.000	1.000	0.841	0.411	0.339	0.362	0.159
△ perfect, 0 → 1	0.874	0.874	0.167	0.955	0.955	0.880	0.325	0.325	0.682
□ always 0	0.706	0.706	2.500	undef	undef	0.000	-0.053	-0.053	0.155
⋆ always 2.5	5.623	8.106	1.500	undef	undef	0.000	-0.797	-0.855	0.010
random	8.548	8.548	1.956	-0.003	-0.003	-0.002	-0.412	-0.412	0.081

Table 2: Comparison of performance measures with artificial predictions.

this case, (1) simplifies to $n_i^- = \lceil s \cdot n_i \rceil$, which keeps the class ratios and the imbalance. This is very similar to random sampling, which is often used to reduce the training time.

2.2 MIDRUS Ensemble

As mentioned earlier, it is state of the art to use ensemble methods with sampling strategies to handle imbalance. We propose to combine the MIDRUS method with an ensemble to further improve predictive performance. We use *bagging (bootstrap aggregation)* [2] and apply MIDRUS T times to independently select T subsamples of the training set. We then train T prediction models, each with one of the T selected training subsets.

For aggregation of the model outputs we train a fusion model. But instead of the features vector, this fusion model gets its T-dimensional input vector from the T outputs of the previously trained ensemble models. To train the fusion model, we subsample the training set with MIDRUS again (with the same parameters), and feed the T models with the samples.

Due to the benefit of continuous output we use Support Vector Regression (SVR) models, but the MIDRUS ensemble can also be trained with other models, including classification models. In general, bagging benefits from a large variance in the trained ensemble models. Selecting $\beta < 1$ can be reasonable, as it increases the variance between models, also in the less frequent classes.

3 Handling Imbalance During Testing

3.1 Performance Measures

Performance measures differ in their suitability for imbalanced learning problems. E.g. the widely used accuracy measure can be very misleading for strongly imbalanced problems [5]. In the context of facial action unit intensity recognition, several other measures have been used so far: the Intraclass Correlation Coefficient (ICC) [7, 16, 17, 18, 19, 30], the Pearson Correlation Coefficient (PCC) [9, 11, 20, 22], the Mean Squared Error (MSE) [9, 11, 20], the macro-averaged Mean Absolute Error (MAE^M) [19, 30], and the macro-averaged F-1 measure (F1^M) [19, 30]. Further, several variants of the measures are used. First, the question is whether the measure is applied to a continuous output score, which we denote with a subscript c, or whether it is applied to a discrete or discretized output, which we denote with subscript d. Several measures also differ in averaging across the multiple classes. We denote macro-averaging (which weights the classes equally) with superscript M, and micro-averaging measures (which are dominated by the more frequent classes) with superscript μ. Further, two variants of ICC are used: ICC(1,1) and ICC(3,1) (see [23] for an explanation).

To compare the measures we conduct an experiment with artificial data, which is summarized in Table 2. The ground truth labels of 130k samples are selected according to the mean distribution across AUs in the DISFA dataset (see Table 1). We assume several artificially generated sets of predicitions that are listed in Table 2 and illustrated in the plots above. The mentioned noise is normally distributed with $\mu = 0$, $\sigma = 0.5$, and $\sigma = 0.9$, respectively. The random prediction is uniformly distributed between 0 and 5.

First, we want to emphasize that the variants often differ significantly, depending on whether they are calculated from continuous or discrete scores. E.g. consider differences between MSE$_c^\mu$ and MSE$_d^\mu$, between PCC$_c$ and PCC$_d$, and between ICC(1,1)$_c$ and ICC(1,1)$_d$. Some authors even mix these measures when comparing methods. After our experiments we advice against this practice, and recommend to always use the discrete variant for several reasons: (1) the ground truth is also discrete, so there is no benefit from using the continuous variants, (2) it is possible to discretize continuous model outputs, but not in reverse, (3) using always the same variant will improve comparability.

All considered measures have good characteristics regarding noise and worsen the performance with increasing noise (see ○, ×, ●). However, the measures differ a lot for under-estimation (see +, ✳, ◇), which is a very common phenomenon in AU intensity estimation, especially with regression models. The most misleading measure is PCC, as it is invariant to these linear transformations of the prediction. Although there is a huge qualitative difference between perfect prediction ○ and linear transformations (+, ✳, and ◇), PCC yields the same performance. We conclude that PCC is not suitable to evaluate AU intensity estimation performance. MSE$^\mu$ is also biased towards under-estimation models (+ and ◇), but to a lesser extend. ICC(3,1) is invariant to constant offsets in prediction (see + and ✳), which is less relevant in practice. MAE^M is low for predictions at ($\star$) or near (✳) the mean. The suitability of measures also depends on the importance of the classes. In general, we think that all classes are similarly important, but in practice the majority class usually plays a more important role for a system than the rarely occurring class. Consider a system that can correctly predict all intensities, but completely fails to predict the AU absence (class 0) all the time (△). MAE^M, F1^M, $ICC(3,1)$, and PCC still provide high performances.

Further, the performance level of a trivial classifier (□, $\star$, random) is not self-evident for MSE, MAE and F1. However, authors can avoid this problem by reporting the best trivial performance level along with their results.

Finally, we recommend to use ICC(1,1)$_d$, ICC(3,1)$_d$, or F1^M. However, our experiments show that quantitative results can be misleading. We suggest that authors should report (at least some) results with confusion matrices that summarize the qualitative prediction performance across the whole dataset. See supplemental material for more details.

3.2 Other Pitfalls

There are other pitfalls in action unit intensity estimation that are not exclusively related to imbalance, but cause misleading results. These even occur in papers presented in high impact journals and good conferences. E.g. Rudovic *et al.* [19] and Yang *et al.* [30] pre-segment the database and group several intensities to reduce imbalance. However, the provided information on the pre-segementation is insufficient to reproduce the experiments. Further, they only include this pre-segmented subset in their test sets, which inhibits comparability. Ideally, this should be avoided by either using the whole database for testing, or at least by providing enough details to make the experiments reproducible.

Another wide-spread pitfall is to apply supervised learning methods before cross validation, i.e. to use labels of test samples in "pre-processing" steps such as feature selection, dimension reduction, or parameter selection. E.g. Mavadati *et al.* [16] use supervised manifold learning to reduce the dimensionality prior to cross-validation, which leads to results with a significant positive bias (see Hastie *et al.* [8]) and prevents any rational comparison.

Another issue that complicates comparisons is that some authors evaluate their models with subject overlap between training and test sets [1, 7]. This should be avoided.

4 Experiments

Automatic Recognition System We conduct our experiments with a person-independent, fully automatic recognition system that consists of the following pipeline: (1) The face and facial landmarks are detected and tracked with OpenCV and IntraFace [29]. (2) The facial landmarks are registered with a mean face by applying an affine transform and minimizing the mean squared error. We register the texture with the same affine transform into a 200 x 200 pixel image with a between-eye distance of about 100 pixel (see Fig. 1 for some examples). (3) We use the aligned landmark coordinates (98 dim.) and concatenated uniform local binary pattern histogram features (5,900 dim.) extracted from the patches of a regular 10 x 10 grid. We use these features due to good performance, fast extraction, and manageable dimensionality. The features are standardized (z-transform) to ensure similar numeric ranges. (4) We apply one of several machine learning methods. Most experiments use support vector regression (SVR) with a linear kernel ($C = 1$ and $\varepsilon = 0.1$) as implemented in LIBSVM [3], either directly or as the base model in an ensemble (see Sec. 2.2). We prefer regression methods as they provide continuous output, which has more potential for analyzing dynamics and detecting micro-expressions. For comparison, we also conduct some experiments with EasyEnsemble [12], one of the state-of-the-art methods for imbalanced classification problems. It combines AdaBoost ensembles with bagging and use the C4.5 decision tree as the base classifier. We use the parameters proposed by the authors with their published implementation and apply the one-vs-rest multiclass strategy with maximum decision score fusion. To keep the experiments feasible regarding training time, we train with a *maximum of 2,500 samples per model*, which are randomly selected if necessary. Due to training time it was also not possible to optimize parameters for each of the trained models, but we conducted several preliminary experiments to select the parameters.

Experimental Procedure We evaluate the predictive performances through 10-fold cross validation, in which we ensure that there is *no subject overlap* between training and test set. Each sample of the respective dataset is used for testing in exactly one fold. For modeling we select AUs that were shown by at least 25% of the subjects in the particular dataset. This

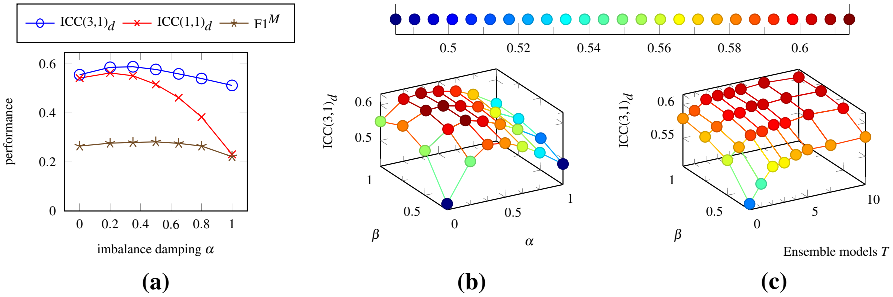


Figure 3: Cross-validated performances on Bosphorus dataset (mean of 26 AUs). MIDRUS with (a) single SVR, (b) SVR ensemble with fixed T, and (c) SVR ensemble with fixed α.

way, in the Bosphorus database we model 26 out of 33 AUs; in the DIFSA and UNBC-McMaster databases all AUs are modeled. For some samples, the fully automatic landmark detection failed. We exclude these samples from our experiments (Bosphorus 0.1%, UNBC-McMaster 0.1%, and DISFA 0.3% of the samples). For most of the experiments, we report the ICC(3,1)$_d$ measure, as it most widely used. See the supplemental material for the results with other measures.

4.1 MIDRUS Parameters

We analyze the influence of the tuning parameters α, β, and T by varying them on the Bosphorus Database. First, we consider a single SVR that is trained with a training set sampled with MIDRUS. We fix $\beta = 1$ and vary α. Fig. 3a plots the performance in the three measures that we found to be most useful in Sec. 3.1. As evident in the plot, neither using all samples ($\alpha = 0$, imbalanced), nor balancing the majority class with the second most frequent class ($\alpha = 1$, 'balanced') yields optimal performance. Higher performance can be gained with the proposed idea of damping ($0 < \alpha < 1$). The optimal α depends on the measure and the data; so it is a typical tuning parameter that should be optimized for each given application during cross-validation. For some qualitative results, see supplemental material.

In a second experiment we trained an ensemble of $T = 10$ SVR models, each with a training set that was independently sampled using MIDRUS. Fig. 3b shows the results of varying α and β. Regarding α, the results are similar to Fig. 3a, which confirms the usefulness of imbalance damping. Further, the plot shows that it is reasonable to chose $\beta < 1$ with an ensemble. Dropping some of the minority class' samples for the individual ensemble models increases variability between them. This way, $\beta < 1$ can improve performance while reducing training time. In a third experiment we fix $\alpha = 0.5$ and vary T and β (see Fig. 3c). If we compare the single SVR ($T = 0$) with the SVR Ensembles ($T \geq 1$), the performance benefit of the ensemble is apparent. With $T \geq 4$ and $\beta \geq 0.75$ the performance does not changes significantly on this dataset. A lower β or T results in lower performance, because more training samples remain unused.

Dataset Measure	**Bosphorus** ICC(3,1)$_d$	**DISFA** ICC(3,1)$_d$	**UNBC** PCC$_c$
Mavadati [17]		0.235	
Kaltwang [11]			0.306
EasyEnsemble	0.340	0.362	0.301
SVR Ensemble	🟦 0.553	🟦 0.346	🟦 0.286
	🟥 0.533	🟥 0.412	🟥 0.301
	🟩 **0.603**	🟩 **0.439**	🟩 **0.311**

🟦 imbalanced 🟥 ‘balanced’ 🟩 proposed: MIDRUS

(a)

		Testing: ICC(3,1)$_d$		
		Bosphorus	DISFA	UNBC
Training	Bosphorus	0.603*	0.401	0.201
	DISFA	0.515	0.439*	0.220
	UNBC	0.346	0.279	0.294*

* Obtained by 10-fold cross validation.

(b)

Figure 4: **(a)** Comparison of methods. Mean cross validated performances of 26 AUs (Bosphorus dataset [21]), 12 AUs (DISFA [16]), resp. 10 AUs (UNBC-McMaster [15]). **(b)** Mean cross-dataset performance of the seven common AUs (SVR Ensemble with $\alpha = 0.5$).

4.2 Comparison of Methods

In Fig. 4a we compare several methods on three databases. On the Bosphorus dataset, EasyEnsemble is clearly outperformed by our proposed SVR Ensemble. MIDRUS ($\alpha = 0.5$, $\beta = 1$) improves performance compared to using the originally imbalanced data (equivalent to $\alpha = 0$) and to balancing it with the second most frequent class (‘balanced’, equivalent to $\alpha = 1$). On the DISFA database all ensemble methods outperform the person-independent modeling results reported by Mavadati and Mahoor [17]. MIDRUS ($\alpha = 0.5$) also performs best on DISFA, but the advantage over EasyEnsemble is lower than for Bosphorus. DISFA contains spontaneous expressions with more low intensity AU occurrences, which might be easier to detect with non-linear classifiers like EasyEnsemble than with the linear SVR ensemble. On the UNBC-McMaster dataset we compare to Kaltwang *et al.* [11] (Relevance Vector Regression on imbalanced data), but do not observe clear benefits of the ensemble methods. This is probably caused by differences in the early stages of the recognition pipeline. Kaltwang uses the manually labeled landmarks provided with the database and align the faces with piece-wise affine transform for each triangle of the face mesh. In contrast, we use a fully automatic landmark detector and a much simpler alignment (one affine transform for the whole image), which is less suited for out-of-plane head poses that occur frequently in UNBC-McMaster. A better face alignment would probably improve the results obtained with ensemble methods. Nevertheless, for the tuned parameters ($\alpha = 0.9$ and $\beta = 1$), MIDRUS still slightly outperforms the results of Kaltwang. See supplemental material for more details.

4.3 Cross-Database Performance

To evaluate generalization performance beyond the scope of a database, we consider the seven AUs coded across all three databases (AU 4, 6, 9, 12, 20, 25, and 26). For these AUs, MIDRUS SVR Ensembles ($\alpha = 0.5$ and $\beta = 1$) are trained on each dataset and tested with both other datasets. The results are reported in Fig. 4b (together with the cross-validation results of the previous section).

Independent of the training set, performance is the best for testing with Bosphorus. This is plausible, as Bosphorus is a non-spontaneous dataset, i.e. high intensities (exaggerated expressions) occur more often, which are easier to classify. Nevertheless, training with Bosphorus still yields quite good results on the spontaneous datasets. If we train with DISFA, performances increase on the spontaneous datasets, probably due to more low intensity samples and occurrence of out-of-plane head poses. Out-of-plane poses are very common in UNBC-McMaster, which may be a reason for the lower performance (especially due to our unsuited face registration). Training on UNBC leads to models that are more appropriate for out-of-plane head poses, which are less common in DISFA and do not occur in Bosphorus; so those models perform poorly on these datasets.

5 Conclusion

Both, too strong data imbalance and too rigorous under-sampling lead to suboptimal performance. We propose MIDRUS, a method that reduces imbalance to achieve a compromise, and propose to apply it within an ensemble. For AU intensity estimation, experiments on three databases shows the superior performance of our method compared to state-of-the-art approaches. In contrast to most ensemble methods, our MIDRUS ensemble directly suits multiclass problems and can be used with various classification and regression models. To our best knowledge, it is the first method that re-balances the class distribution with such flexibility. It allows the integration of over-sampling (e.g. sampling with repetition or SMOTE [4]), which we will address in future works along with comparisons to other datasets (outside the facial expression domain). We will also work on more advanced face registration methods.

Acknowledgment

This work was supported by Transregional Collaborative Research Centre SFB/TRR 62 funded by the German Research Foundation (DFG).

References

[1] Deniz Bingöl, Turgay Celik, Christian W. Omlin, and Hima B. Vadapalli. Facial action unit intensity estimation using rotation invariant features and regression analysis. In *International Conference on Image Processing (ICIP)*, pages 1381–1385. IEEE, 2014.

[2] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, August 1996. ISSN 0885-6125, 1573-0565. doi: 10.1007/BF00058655.

[3] Chih-Chung Chang and Chih-Jen Lin. LIBSVM – A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011. URL http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[4] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligent Research*, 16:321–357, 2002.

[5] Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6(1):1–6, June 2004. ISSN 1931-0145. doi: 10.1145/1007730.1007733.

[6] P. Ekman, W. Friesen, and J. Hager. *Facial Action Coding System: The Manual on CD ROM*. A Human Face, 2002.

[7] Jeffrey M. Girard, Jeffrey F. Cohn, and Fernando De la Torre. Estimating smile intensity: A better way. *Pattern Recognition Letters*, 2014. ISSN 0167-8655. doi: 10.1016/j.patrec.2014.10.004.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*, volume 2, chapter The Wrong and Right Way to Do Cross-validation, pages 245–247. Springer, 2009.

[9] L.A. Jeni, J.M. Girard, J.F. Cohn, and F. De la Torre. Continuous AU intensity estimation using localized, sparse facial feature space. In *10th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–7, April 2013. doi: 10.1109/FG.2013.6553808.

[10] László A. Jeni, Jeffrey F. Cohn, and Fernando De La Torre. Facing imbalanced data - recommendations for the use of performance metrics. In *Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 245–251. IEEE, 2013.

[11] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. Continuous Pain Intensity Estimation from Facial Expressions. In *Advances in Visual Computing*, number 7432 in Lecture Notes in Computer Science, pages 368–377. Springer Berlin Heidelberg, January 2012. ISBN 978-3-642-33190-9, 978-3-642-33191-6.

[12] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[13] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on System, Man and Cybernetics*, 39(2):539–550, 2009.

[14] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141, 2013.

[15] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 57–64, 2011.

[16] Seyed Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[17] S.M. Mavadati and M.H. Mahoor. Temporal Facial Expression Modeling for Automated Action Unit Intensity Measurement. In *International Conference on Pattern Recognition (ICPR)*, pages 4648–4653, 2014. doi: 10.1109/ICPR.2014.795.

[18] Zuheng Ming, Aurélie Bugeau, Jean-Luc Rouas, and Takaaki Shochi. Facial Action Units Intensity Estimation by the Fusion of Features with Multi-kernel Support Vector Machine. In *11th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, Ljubljana, Slovenia, May 2015.

[19] O. Rudovic, V. Pavlovic, and M. Pantic. Context-Sensitive Dynamic Ordinal Regression for Intensity Estimation of Facial Action Units. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 37(5):944–958, 2015. doi: 10.1109/TPAMI.2014.2356192.

[20] Georgia Sandbach, Stefanos Zafeiriou, and Maja Pantic. Markov random field structures for facial action unit intensity estimation. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 738–745. IEEE, 2013.

[21] Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus Database for 3d Face Analysis. In *Biometrics and Identity Management*, number 5372 in Lecture Notes in Computer Science, pages 47–56. Springer Berlin Heidelberg, 2008.

[22] Arman Savran, Bulent Sankur, and M. Taha Bilge. Regression-based intensity estimation of facial action units. *Image and Vision Computing*, 30(10):774–784, October 2012. ISSN 0262-8856. doi: 10.1016/j.imavis.2011.11.008.

[23] Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[24] Muhammad Atif Tahir, Josef Kittler, and Fei Yan. Inverse random undersampling for class imbalance problem and its application to multi-label classification. *Pattern Recognition*, 45:3738–3750, 2012.

[25] Fernando De la Torre and Jeffrey F. Cohn. Facial Expression Analysis. In *Visual Analysis of Humans*, pages 377–409. Springer London, January 2011.

[26] M. Valstar, J. Girard, T. Almaev, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and J. Cohn. Fera 2015-second facial expression recognition and analysis challenge. *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

[27] Michel Valstar. Automatic Facial Expression Analysis. In *Understanding Facial Expressions in Communication*, pages 143–172. Springer India, January 2015.

[28] Philipp Werner, Ayoub Al-Hamadi, Robert Niese, Steffen Walter, Sascha Gruss, and Harald C. Traue. Towards Pain Monitoring: Facial Expression, Head Pose, a new Database, an Automatic System and Remaining Challenges. In *Proc. BMVC*, pages 119.1–119.13. BMVA Press, 2013. doi: 10.5244/C.27.119.

[29] Xuehan Xiong and Fernando De la Torre. Supervised Descent Method and its Applications to Face Alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, 2013. doi: 10.1109/CVPR.2013.75.

[30] Shuang Yang, Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Personalized Modeling of Facial Action Unit Intensity. In *Advances in Visual Computing*, number 8888 in Lecture Notes in Computer Science, pages 269–281. Springer International Publishing, December 2014. ISBN 978-3-319-14363-7, 978-3-319-14364-4.

[31] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. BP4D-Spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32 (10):692–706, 2014. ISSN 0262-8856. doi: 10.1016/j.imavis.2014.06.002.