Applying Transfer Testing to Identify Annotation Discrepancies in Facial Emotion Data Sets

Sarah Dreher(✉), Jens Gebele, and Philipp Brune

Neu-Ulm University of Applied Sciences, Wileystraße 1, 89231 Neu-Ulm, Germany
{Sarah.Dreher,Jens.Gebele,Philipp.Brune}@hnu.com

Abstract. The field of Artificial Intelligence (AI) has a significant impact on the way computers and humans interact. The topic of (facial) emotion recognition has gained a lot of attention in recent years. Majority of research literature focuses on improvement of algorithms and Machine Learning (ML) models for single data sets. Despite the impressive results achieved, the impact of the (training) data quality with its potential biases and annotation discrepancies is often neglected. Therefore, this paper demonstrates an approach to detect and evaluate annotation label discrepancies between three separate (facial) emotion recognition databases by Transfer Testing with three ML architectures. The findings indicate Transfer Testing to be a new promising method to detect inconsistencies in data annotations of emotional states, implying label bias and/or ambiguity. Therefore, Transfer Testing is a method to verify the transferability of trained ML models. Such research is the foundation for developing more accurate AI-based emotion recognition systems, which are also robust in real-life scenarios.

Keywords: Emotion Recognition · Facial Expression Recognition · Emotional Artificial Intelligence · Transfer Testing · Data Quality · Transferability

1 Introduction

Over the past years, there has been significant interest in AI-based emotion recognition both in research and in practical applications. This technology enables machines to identify the emotional state of humans [26,32].

The use of emotion recognition systems has expanded to various fields, including customer service [1], emotional support [3], and human-computer relationships [6,7,18].

Numerous studies have been conducted to develop emotion recognition technology using various data modalities and classification taxonomies [19,31]. Facial expression recognition (FER) is one of the most widely used and promising technologies [14,37], mainly due to the fact that human emotions are strongly conveyed through facial expressions [5,40]. The use of computerized FER has been the subject of extensive research in recent years [26,32].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
S. Bouzefrane et al. (Eds.): MSPN 2023, LNCS 14482, pp. 157–174, 2024.
https://doi.org/10.1007/978-3-031-52426-4_11

The primary focus of these studies is to enhance the performance of ML models and their architectures as well as the overall model performance [36]. Although research has mostly centered on individual or a limited number of data sets, not much attention has been given to the underlying data (quality) which affects the transferability. As a result, the significance of inconsistencies in data annotation and/or labeling ambiguity of emotional states remains poorly understood [11] and the models perform badly on additional data sets.

To assess the impact of inconsistencies and/or labeling ambiguity and increase the transferability past research uses Transfer Learning. Transfer Learning uses pre-learned knowledge from a task to improve the performance of a related task [34]. In the context of FER, Transfer Learning is used to improve the performance of a model on a new data set by using a pre-trained model on a different data set [25]. However, Transfer Learning does not provide any information about the quality of the data annotations.

Therefore, in this paper we propose a new method called Transfer Testing to get further insights into quality of the data annotations. This extends and builds on the results some of the authors presented in [13], as we discuss later in Sect. 5. In [13], the same data sets were examined for inconsistencies in data annotation and/or labeling ambiguity. As outlined in Sect. 5, our earlier paper revealed label similarities in RAF-DB and FER2013, whereas AffectNet annotations differed from the other data sets. In the present paper, these earlier results are extended by using more than one ML architecture, as well as by, Transfer Testing to evaluate across databases, while the methodology is not changed.

Transfer Testing is a method to verify the transferability of trained ML models. It is a systematic approach to detect and evaluate annotation label discrepancies between separate (facial) emotion recognition databases. The findings indicate Transfer Testing to be a new promising method to detect inconsistencies in data annotations of emotional states, implying label bias and/or ambiguity. Such research is the foundation for developing more accurate AI-based emotion recognition systems, which are also robust in real-life scenarios.

Our goal is to gain a better understanding of how transferability is affected by different data sets annotations using various model architectures. Therefore, multiply models were trained on one single data set and tested on both other data set. This process is called Transfer Testing. By applying Transfer Testing, we investigate potential discrepancies in data annotations and provide new insights for future research directions in the field of transferability.

The rest of this paper is organized as follows: In Sect. 2, a detailed review of the existing literature on facial emotion recognition is presented, as well as the relevant ML techniques and a description of emotional data sets. The research methodology and data used for the systematic analysis is shown in Sect. 3. In Sect. 4 the accomplished results are presented and discussed in depth in Sect. 5. The conclusions summarize our findings.

2 Related Work

2.1 Facial Expression Recognition

FER combines Psychology [5,9] and Technology (Computer Vision). This interdisciplinary research field aims to infer human's emotional state to gather highly relevant information contained in facial expressions [12,20].

Most research on facial emotion recognition is based on Paul Ekman's work. He claims, different cultural backgrounds do not affect dependencies between certain facial expressions and human emotional states [9]. Ekman defined six basic emotional states, namely *anger*, *fear*, *disgust*, *happiness*, *surprise* and *sadness* [5,8]. Focusing on ML, emotion recognition can be differentiated into following four different tasks: Single Label Learning (SLL), SLL Extension (extended by Intensity Estimation), Multi-Label Learning (MLL) and Label Distribution Learning (LDL) [11].

SLL describes a multi-class ML problem. Based on the highest likelihood one emotional class is identified from several possible emotional states in a facial expression. Since, this study focuses on limitations directly linked to SLL [11], the other techniques are not discussed.

Since computers require binary states, research and practice develop ML models, which perform well by assigning one emotional class to a single facial expression. That is still the main focus of research. By taking a closer look on the ML task, there are certain dependencies between the ML approaches, for instance, SLL can be seen as LDL instances [11]. This work deals with a two-sided aspect of data annotations in SLL tasks. Firstly, data annotations (labels) can either be manually or automatically generated which can lead to inconsistencies/biases. Secondly, recent research claims that one facial expression can carry more than one emotional state [11]. From past research is also known that certain emotional states can be recognized better than others [21,35].

These challenges have already been investigated on different facial data sets by some of the authors [13]. By exploring various data sets with one basic CNN model, past research came to the conclusion that not only the size of the data set nor the share of each emotion has influence on the recognition accuracy, the underlying data (label) quality affects this too [13]. In addition, higher image resolution data sets do not necessarily lead to better recognition results [13]. Furthermore, latest research has successfully developed models for FER which discount annotations and still achieve impressive results [25]. These models transform the given emotion into a neutral expression in order to reconstruct the emotion on this basis.

2.2 Machine Learning Techniques

There are different approaches for FER in ML. A general ML process consists of up to three phases. First preprocessing phase, second feature extraction phase, which can be optional, and third emotion recognition or rather classification phase. Different conventional ML and/or modern Deep Learning methods can be

applied within each phase. Conventional ML methods consist of Support Vector Machine (SVM), Adaptive Boosting (AdaBoost), Random Forest (RF), Decision Tree [11]. Deep Learning models extract automatically relevant facial features during training [28,41]. In FER Deep Learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) attract attention [11].

A CNN has multiple layers similar to a Deep Neural Network (DNN). A CNN contains convolutional layer(s), pooling layer(s), dense layer(s) and fully connected layer(s). The convolutional layer(s) train the relevant features, starting from low-level features in early layers, up to high-level (abstract) features. The following pooling layer(s), aggregate information and thereby reduce computational complexity [24].

A CNN model automatically extracts features. For this reason, separate feature extraction methods like in traditional ML algorithms are not necessary [11]. Some different popular CNN architectures are listed below in chronological order: LeNet-5 [24], AlexNet [23], GoogleLeNet [39], VGGNet [38], ResNet [16], Xception [4], SENet [17]. The architectures have evolved and got more complex over the time. Further on, convolutional layers have been stacked directly and inception modules, residual learning (with skip connections) and depthwise separable convolution layer have been developed.

2.3 Emotional Facial Databases

Previous research in FER has led to a lot of facial databases. These differ based on data type (static, sequential), data dimension (two-dimensional, three-dimensional), data collection environment (controlled, uncontrolled), and number of facial expressions [11]. Databases set up in a controlled environment are for instance The Extended Cohn-Kanade data set (CK+) [29] and The Japanese Female Facial Expression (JAFFE) database [30]. Since systems based on these data sets reach only lower performance in real-world scenarios, research demanded for databases collected in an uncontrolled setting. Examples are AffectNet [33] and Real-world Affective Faces Database (RAF-DB) [27]. Most of these databases include six basic emotional states [10], usually adding one neutral facial expression. Therefore, emotional labels can be annotated manually by experts [33], by computers or by a combination of these [15].

3 Methodology

3.1 Technical Environment

We implemented all ML models on our institute server. It runs on Ubuntu 20.04 LTS, including the NVIDIA data science stack [2]. The server has two NVIDIA A40 Graphics Processing Units. The code is developed in Python, using Jupyter Notebook as integrated development environment, and made use of these Python frameworks: NumPy, Matplotlib, Pandas, Scikit-Learn, Keras and TensorFlow.

3.2 Data Collection

We consider three different data sets which contain the six basic emotions (*anger*, *fear*, *disgust*, *happiness*, *surprise* and *sadness*) with an added neutral facial expression [13]. In addition, to receive a sufficient quantity of data, as well as a more realistic and representative data, we excluded databases with a size of less than 10,000 instances and/or the ones collected in a controlled environment. The remaining data sets are FER2013, RAF-DB and AffectNet with eight labels (the so-called Mini Version).

FER2013 contains 35,887 gray images, which are automatically cropped, labeled and then cross-checked by experts. It has seven emotional classes and all images are resized to a format of 48×48 pixels [15]. RAF-DB on the other hand has 15,339 aligned colorful RGB-images. All images were manually annotated by about 40 experts and aligned to a size of 100×100 pixels [27]. The mini AffectNet consists of 291,650 only manually annotated images in RGB-color with 224×224 pixels each. The emotional state contempt was removed in addition to leave us with the same seven emotions as in the other data sets [33].

Table 1. Distribution of Emotional Classes per Data Set

Emotion	FER-2013		RAF-DB		AffectNet	
Pixel Size	48×48		100×100		224×224	
Anger	4,953	(14%)	867	(6%)	25,382	(9%)
Disgust	547	(2%)	877	(6%)	4,303	(1%)
Fear	5,121	(14%)	355	(2%)	6,878	(2%)
Happiness	8,989	(25%)	5,957	(39%)	134,915	(47%)
Sadness	6,077	(17%)	2,460	(16%)	25,959	(9%)
Surprise	4,002	(11%)	1,619	(11%)	14,590	(5%)
Neutral	6,198	(17%)	3,204	(21%)	75,374	(26%)
Total	35,887	(100%)	15,339	(100%)	287,401	(100%)

3.3 Data Pre-processing

The pre-processing stage covers typically different methods. For instance, face detection, facial landmark localization, face normalization and data augmentation [20]. Face localization is the first step. The previously described data sets have already aligned and cropped images. That is why we limit preprocessing to data normalization and augmentation.

To have equal conditions for the comparison, we resize the images of RAF-DB and AffectNet to the pixel size of FER2013. Since we combine normalization and data augmentation method, we divide each pixel by 255, which results in a range from 0 to 1 for each pixel. The total distribution of the emotional classes is presented in Table 1.

The pixel size is similar on the three different data sets. Most emotional states are sufficiently well represented in all data sets, with a few exceptions. Every data set is split into one training and one test set, with ratios of 80 percent to 20 percent. 30 percent of the training set are used as validation set. By stratifying the splits we keep the proportions of each emotional class equal in training, validation and test set. Since AffectNet is provided with a small test set, we first combine training and test set. Afterwards we split it in the same ways as the others. By the end of this publication we address differences in annotation and label ambiguity between the three data sets. Therefore, we use the trained models on each data set and evaluate these on the other two data sets, i.e. AlexNet, which was trained on RAF-DB, is evaluated on AffectNet. Since FER2013 only has decolorized images, we turn all images into black and white.

3.4 Deep Learning Model Architecture

As already mentioned, we implement various CNN architectures. First of all we need to emphasize that our aim is to compare the emotion recognition accuracy for individual emotional states in different data sets, which does not require beating a certain performance threshold. Hence, we have decided to use three architectures which use stacked CNN layers directly. The first one is AlexNet [23]. The second architecture is a standard CNN based on AlexNet [23]. In the following this CNN architecture is called defaultNet. This defaultNet is the same architecture used in [13]. The architecture of defaultNet consists of four blocks, each block contains two convolutional layers followed by one pooling layer. In each convolutional layer, we chose the padding option same and ReLu activation function. The pooling layer uses max pooling, which generally performs better than average pooling. After a stack of these four blocks, the output is flattened and then two dense layers including dropout follow. In the end, we classify between seven possible emotional states. To consider a more complex architecture we decided to use VGGNet [38] as our last architecture.

For training of our models we define 50 epochs and a batch size of 128 for every data set, in order to have the same amount of weight updates. However, the steps per epoch differ due to the different size of the data sets. Furthermore, we use Adam Optimizer starting with a learning rate of 0.0001. This learning rate is dynamic because it is automatically reduced during training, if validation accuracy does not improve for three epochs in a row. At the end, we use on each architecture the model with the highest validation accuracy during training.

3.5 Transfer Testing

All architectures defined in Sect. 3.4 are trained on each data set. After training, the models are initially evaluated on the test set of the data set they were trained on. In addition, we evaluate the trained models on the test sets of the other two data sets. This process is called Transfer Testing. A graphical explanation of the whole process is shown in Fig. 1.

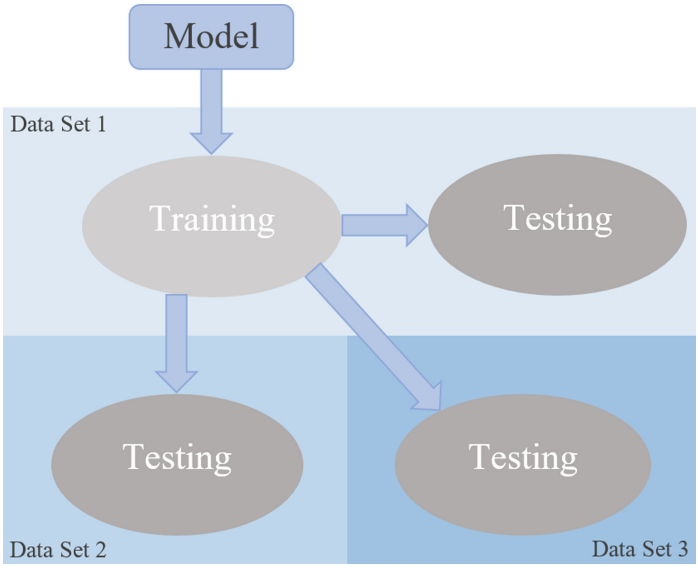

Fig. 1. Evaluation process of one architecture.

Figure 1 show the evaluation of one ML model on the three data sets. This process is done for all three model architectures (AlexNet, defaultNet and VGGNet) among all data sets. For instance, AlexNet is trained on RAF-DB. This trained AlexNet is evaluated on the test set of RAF-DB, FER2013 and AffectNet. This process is repeated for every architecture and each data set. The results are presented in the next Section.

4 Results

In this section, we present emotion recognition accuracy of the seven basic emotional states for every model architecture evaluated on the three data sets. The outcome metrics are limited to precision, recall and F1-score as these are relevant to answering our research question(s). Due to class imbalances, overall accuracy is not very meaningful. Our main focus of the analysis is on the F1-scores, which represents the harmonic mean of precision and recall. The following results evaluate the three trained models on every test set of one of the three data sets, i.e. the AlexNet which was trained on RAF-DB is evaluated on FER2013 and AffectNet. Therefore, the results contain F1-scores of normal testing and Transfer Testing. Table 2 shows the evaluation of the three trained models on the test set of FER2013. Furthermore, the results for RAF-DB are represented in Table 3 and for AffectNet in Table 4.

For each data set, we run the models five times in order to address random model initialization. Additionally, the corresponding standard deviation is shown in brackets for every metric. There is a general tendency for emotional classes

Table 2. Evaluation od Models trained on FER2013 for all three data sets

Count	Emotion	AlexNet	defaultNet	VGGNet
991	Anger-FER	0.34 (± 0.03)	0.49 (± 0.01)	0.46 (± 0.01)
	Anger-RAF	0.16 (± 0.01)	0.16 (± 0.01)	0.12 (± 0.01)
	Anger-Aff	0.19 (± 0.01)	0.17 (± 0.01)	0.12 (± 0.01)
109	Disgust-FER	0.04 (± 0.08)	0.05 (± 0.10)	0.27 (± 0.11)
	Disgust-RAF	0.02 (± 0.01)	0.00 (± 0.01)	0.02 (± 0.01)
	Disgust-Aff	0.02 (± 0.00)	0.01 (± 0.00)	0.02 (± 0.00)
1,024	Fear-FER	0.34 (± 0.02)	0.38 (± 0.02)	0.41 (± 0.02)
	Fear-RAF	0.08 (± 0.04)	0.05 (± 0.02)	0.14 (± 0.03)
	Fear-Aff	0.12 (± 0.04)	0.18 (± 0.01)	0.14 (± 0.03)
1,798	Happiness-FER	0.66 (± 0.01)	0.78 (± 0.01)	0.80 (± 0.01)
	Happiness-RAF	0.43 (± 0.00)	0.50 (± 0.02)	0.52 (± 0.01)
	Happiness-Aff	0.03 (± 0.02)	0.04 (± 0.01)	0.52 (± 0.01)
1,216	Sadness-FER	0.35 (± 0.01)	0.45 (± 0.01)	0.46 (± 0.02)
	Sadness-RAF	0.25 (± 0.02)	0.27 (± 0.02)	0.28 (± 0.01)
	Sadness-Aff	0.01 (± 0.01)	0.02 (± 0.00)	0.28 (± 0.01)
800	Surprise-FER	0.64 (± 0.00)	0.71 (± 0.01)	0.73 (± 0.01)
	Surprise-RAF	0.10 (± 0.02)	0.05 (± 0.01)	0.06 (± 0.03)
	Surprise-Aff	0.00 (± 0.00)	0.00 (± 0.00)	0.06 (± 0.00)
1,240	Neutral-FER	0.44 (± 0.02)	0.53 (± 0.01)	0.52 (± 0.01)
	Neutral-RAF	0.24 (± 0.02)	0.26 (± 0.03)	0.20 (± 0.03)
	Neutral-Aff	0.07 (± 0.02)	0.07 (± 0.01)	0.20 (± 0.01)

with higher occurrence to have lower standard deviations, for instance, *happiness*, *sadness* and *neutral*. The variation in F1-scores for each trained model on remaining data sets is conspicuous. F1-scores on AffectNet tend to be the lowest, except for the ones trained on AffectNet. For better impression on the impact of the model architectures on the results Table 5 displays the accuracy of each model on every data set. In the table we use the weighted average measured on the quantity of images for each emotion. This means that we first consider the support of each emotion into account and then take the average of the five training cycles.

In the next section, we discuss results, similarities and differences in the recognition accuracy of emotional states and work out possible reasons for this.

5 Discussion

By focusing on the models, we conclude that the impact of the architecture on the result is not the crucial factor. All model architectures tend to the same results. VGGNet, the most complex architecture, tends to have higher accuracy

Table 3. Evaluation od Models trained on RAF-DB for all three data sets

Count	Emotion	AlexNet	defaultNet	VGGNet
173	Anger-FER	0.01 (± 0.01)	0.00 (± 0.00)	0.00 (± 0.00)
	Anger-RAF	0.43 (± 0.03)	0.56 (± 0.02)	0.53 (± 0.03)
	Anger-Aff	0.00 (± 0.00)	0.00 (± 0.00)	0.02 (± 0.04)
175	Disgust-FER	0.07 (± 0.03)	0.05 (± 0.02)	0.05 (± 0.02)
	Disgust-RAF	0.10 (± 0.07)	0.25 (± 0.05)	0.33 (± 0.03)
	Disgust-Aff	0.07 (± 0.03)	0.07 (± 0.01)	0.08 (± 0.03)
71	Fear-FER	0.00 (± 0.01)	0.00 (± 0.00)	0.00 (± 0.02)
	Fear-RAF	0.22 (± 0.03)	0.32 (± 0.08)	0.29 (± 0.05)
	Fear-Aff	0.04 (± 0.01)	0.03 (± 0.01)	0.03 (± 0.00)
1,192	Happiness-FER	0.71 (± 0.01)	0.79 (± 0.01)	0.79 (± 0.01)
	Happiness-RAF	0.83 (± 0.01)	0.87 (± 0.01)	0.88 (± 0.01)
	Happiness-Aff	0.01 (± 0.01)	0.01 (± 0.00)	0.00 (± 0.00)
492	Sadness-FER	0.29 (± 0.01)	0.38 (± 0.01)	0.38 (± 0.03)
	Sadness-RAF	0.50 (± 0.02)	0.54 (± 0.02)	0.57 (± 0.03)
	Sadness-Aff	0.01 (± 0.01)	0.01 (± 0.00)	0.02 (± 0.01)
324	Surprise-FER	0.11 (± 0.02)	0.11 (± 0.02)	0.13 (± 0.02)
	Surprise-RAF	0.63 (± 0.01)	0.67 (± 0.01)	0.67 (± 0.02)
	Surprise-Aff	0.21 (± 0.02)	0.20 (± 0.02)	0.21 (± 0.04)
641	Neutral-FER	0.40 (± 0.06)	0.48 (± 0.03)	0.41 (± 0.02)
	Neutral-RAF	0.61 (± 0.01)	0.68 (± 0.01)	0.67 (± 0.01)
	Neutral-Aff	0.06 (± 0.02)	0.09 (± 0.01)	0.07 (± 0.03)

scores. Obviously the models trained and tested on the same data set provide the best accuracy.

Doing testing and training on the same data set confirms the findings of the previous work [13] for all three architectures. New insights on the overall performance, as well as the emotional states itself, is discussed on the basis of Transfer Testing in the following.

The results of our analysis in Tables 2, 3 and 4 show that the emotional state *happiness* is best recognizable in every data set, independent of the architectures, while testing on the same data set. Using Transfer Testing of FER2013 and RAF-DB trained models on these mutual data sets, still *happiness* is detected best. AffectNet seems to differ from these two data sets, since the recognition ranking vary in order during Transfer Testing on FER2013 or RAF-DB. *Fear* and *disgust* are the most difficult emotional states to recognize in all data sets and for all models except for AffectNet trained ones.

Table 6 illustrates a ranking of recognition accuracy for every emotional state based on F1-score on FER2013. The same information for RAF-DB is shown in Table 8 such as for AffectNet in Table 7. The emotional state *surprise* in the

Table 4. Evaluation od Models trained on AffectNet for all three data sets

Count	Emotion	AlexNet	defaultNet	VGGNet
5,076	Anger-FER	0.14 (± 0.02)	0.12 (± 0.00)	0.10 (± 0.01)
	Anger-RAF	0.13 (± 0.01)	0.15 (± 0.01)	0.11 (± 0.01)
	Anger-Aff	0.41 (± 0.02)	0.54 (± 0.01)	0.53 (± 0.01)
861	Disgust-FER	0.01 (± 0.00)	0.01 (± 0.00)	0.00 (± 0.00)
	Disgust-RAF	0.03 (± 0.00)	0.04 (± 0.00)	0.05 (± 0.01)
	Disgust-Aff	0.00 (± 0.00)	0.00 (± 0.00)	0.07 (± 0.07)
1,376	Fear-FER	0.05 (± 0.00)	0.05 (± 0.00)	0.04 (± 0.00)
	Fear-RAF	0.04 (± 0.00)	0.06 (± 0.01)	0.06 (± 0.01)
	Fear-Aff	0.22 (± 0.02)	0.26 (± 0.03)	0.33 (± 0.03)
26,983	Happiness-FER	0.00 (± 0.01)	0.00 (± 0.00)	0.00 (± 0.00)
	Happiness-RAF	0.04 (± 0.01)	0.00 (± 0.00)	0.04 (± 0.03)
	Happiness-Aff	0.85 (± 0.00)	0.89 (± 0.00)	0.90 (± 0.00)
5,192	Sadness-FER	0.14 (± 0.01)	0.11 (± 0.02)	0.14 (± 0.01)
	Sadness-RAF	0.07 (± 0.04)	0.09 (± 0.02)	0.12 (± 0.01)
	Sadness-Aff	0.32 (± 0.05)	0.49 (± 0.01)	0.49 (± 0.02)
2,918	Surprise-FER	0.04 (± 0.00)	0.04 (± 0.00)	0.04 (± 0.00)
	Surprise-RAF	0.04 (± 0.00)	0.04 (± 0.00)	0.03 (± 0.00)
	Surprise-Aff	0.28 (± 0.03)	0.42 (± 0.01)	0.42 (± 0.02)
15,075	Neutral-FER	0.12 (± 0.03)	0.17 (± 0.01)	0.16 (± 0.03)
	Neutral-RAF	0.23 (± 0.03)	0.18 (± 0.03)	0.18 (± 0.04)
	Neutral-Aff	0.62 (± 0.00)	0.68 (± 0.00)	0.68 (± 0.00)

Table 5. Accuracy as Weighted Average

Test Set	Models trained on	AlexNet	defaultNet	VGGNet
FER2013	FER2013	0.47	0.56	0.58
	RAF-DB	0.24	0.25	0.26
	AffectNet	0.06	0.07	0.07
RAF-DB	FER2013	0.42	0.49	0.47
	RAF-DB	0.63	0.69	0.70
	AffectNet	0.05	0.05	0.05
AffectNet	FER2013	0.06	0.07	0.07
	RAF-DB	0.10	0.07	0.09
	AffectNet	0.65	0.72	0.72

AffectNet data set represents the major exception in the ranking for traditional training and testing. Furthermore, the other emotions hardly vary in order on all three data sets for all architectures. As soon as we evaluate the AffectNet trained models using Transfer Testing we get results which highly vary from the other patterns.

Table 6. Recognition Accuracy Ordinal Ranking for Models trained on FER2013

Trained	Rank	AlexNet	defaultNet	VGGNet
FER	1	Happiness	Happiness	Happiness
	2	Surprise	Surprise	Surprise
	3	Neutral	Neutral	Neutral
	4	Sadness	Anger	Anger
	5	Fear	Sadness	Sadness
	6	Anger	Fear	Fear
	7	Disgust	Disgust	Disgust
RAF	1	Happiness	Happiness	Happiness
	2	Sadness	Sadness	Sadness
	3	Neutral	Neutral	Neutral
	4	Anger	Anger	Fear
	5	Surprise	Surprise	Anger
	6	Fear	Fear	Surprise
	7	Disgust	Disgust	Disgust
Aff	1	Anger	Fear	Fear
	2	Fear	Anger	Anger
	3	Neutral	Neutral	Neutral
	4	Happiness	Happiness	Happiness
	5	Disgust	Sadness	Sadness
	6	Sadness	Disgust	Disgust
	7	Surprise	Surprise	Surprise

Table 7. Recognition Accuracy Ordinal Ranking for Models trained on AffectNet

Trained	Rank	AlexNet	defaultNet	VGGNet
FER	1	Anger	Neutral	Neutral
	2	Sadness	Anger	Sadness
	3	Neutral	Sadness	Anger
	4	Fear	Fear	Fear
	5	Surprise	Surprise	Surprise
	6	Disgust	Disgust	Disgust
	7	Happiness	Happiness	Happiness
RAF	1	Neutral	Neutral	Neutral
	2	Anger	Anger	Sadness
	3	Sadness	Sadness	Anger
	4	Fear	Fear	Fear
	5	Surprise	Disgust	Disgust
	6	Happiness	Surprise	Happiness
	7	Disgust	Happiness	Surprise
Aff	1	Happiness	Happiness	Happiness
	2	Neutral	Neutral	Neutral
	3	Anger	Anger	Anger
	4	Sadness	Sadness	Sadness
	5	Surprise	Surprise	Surprise
	6	Fear	Fear	Fear
	7	Disgust	Disgust	Disgust

The results from trained AffectNet models lead to a totally new order in the recognition ranking. *Fear* is better recognizable whereas *happiness* is not the best emotion to recognize. These outliers in the comparative ranking are the first indications of data inconsistencies.

Table 8. Recognition Accuracy Ordinal Ranking for Models trained on RAF-DB

Trained	Rank	AlexNet	defaultNet	VGGNet
FER	1	Happiness	Happiness	Happiness
	2	Neutral	Neutral	Neutral
	3	Sadness	Sadness	Sadness
	4	Surprise	Surprise	Surprise
	5	Disgust	Disgust	Disgust
	6	Anger	Anger	Anger
	7	Fear	Fear	Fear
RAF	1	Happiness	Happiness	Happiness
	2	Surprise	Neutral	Surprise
	3	Neutral	Surprise	Neutral
	4	Sadness	Anger	Sadness
	5	Anger	Sadness	Anger
	6	Fear	Fear	Disgust
	7	Disgust	Disgust	Fear
Aff	1	Surprise	Surprise	Surprise
	2	Disgust	Neutral	Disgust
	3	Neutral	Disgust	Neutral
	4	Fear	Fear	Fear
	5	Happiness	Happiness	Anger
	6	Sadness	Sadness	Sadness
	7	Anger	Anger	Happiness

Furthermore, it is worth taking a closer look on F1-score intervals at every emotional state. There are differences between the best and worst F1-score for every emotional state in the data sets across all model architectures. The difference in F1-scores are presented in Table 9.

Focusing on *disgust*, the F1-scores differences are the worst for traditional training and testing. Therefore, we can assume that this emotion has the highest label inconsistency. This is also influenced by the low share of this emotion in every data set, see Table 1. *Fear* is underrepresented in AffectNet and RAF-DB as well, and accordingly the F1-score difference is higher. In FER2013 *fear* seems to have some label inconsistency, as the corresponding F1-score differences are always high. The strong F1-score variations in certain emotions is a further sign of potential irregularities in the underlying data sets.

Table 9 leads to the conclusion that trained models on AffectNet tend to worst F1-score ranges among all data sets. This confirms the point that AffectNet differs from the other data sets with reference to label inconsistency and annotation.

Table 9. F1-score Differences for every Emotional State across all architectures and all data sets

Model trained on	Emotion	Max F1-score differences on		
		FER2013	RAF-DB	AffectNet
FER	Anger	0.15	0.01	0.04
	Disgust	0.23	0.02	0.01
	Fear	0.07	0.00	0.01
	Happiness	0.14	0.08	0.00
	Sadness	0.11	0.09	0.03
	Surprise	0.09	0.02	0.00
	Neutral	0.09	0.08	0.05
RAF	Anger	0.04	0.13	0.04
	Disgust	0.02	0.23	0.02
	Fear	0.09	0.10	0.02
	Happiness	0.09	0.05	0.04
	Sadness	0.03	0.07	0.05
	Surprise	0.05	0.04	0.01
	Neutral	0.06	0.07	0.05
Aff	Anger	0.07	0.02	0.13
	Disgust	0.01	0.01	0.07
	Fear	0.06	0.01	0.11
	Happiness	0.49	0.01	0.05
	Sadness	0.27	0.01	0.17
	Surprise	0.06	0.01	0.14
	Neutral	0.13	0.03	0.06

In accordance to the ranking in Tables 2, 3 and 4, we present a ranking for every emotional state based on F1-scores in every data sets among all models. Table 10 indicates best recognition accuracy, considering the average of F1-scores across all models. All data sets have the best F1-scores across the emotions while training and testing on the same data set, except *disgust* in trained AffectNet. Due to the stratified split into training, validation and test data, class imbalances are present. A reason for *disgust* being more recognizable on RAF-DB for trained AffectNet models is the fact that this class has a very low share in every data set, but is more present in RAF-DB. For trained FER2013 the recognition accuracy in RAF-DB is better than for AffectNet despite the fact RAF-DB

has the smallest amount of images. The emotions where RAF-DB ranks have a smaller share. Focusing on RAF-DB trained models, the ranking for *disgust*, *fear* and *surprise* are the worst on FER2013. Even FER2013 has a higher percentage on the data set for these emotions, AffectNet has a higher recognition accuracy. *Happiness* and *neutral* have higher appearance on AffectNet while the accuracy level is lower. This is an indicator of label inconsistency on this emotions in AffectNet. Overall, this leads to the assumption that RAF-DB has the lowest data inconsistencies, while FER2013 and AffectNet have higher ones.

Table 10. Recognition Accuracy F1-score Ranking

Emotion	FER2013 with trained			RAF-DB with trained			AffectNet with trained		
	FER	RAF	Aff	FER	RAF	Aff	FER	RAF	Aff
Anger	I	II	III	III	I	II	II	III	I
Disgust	I	III	III	III	I	I	II	II	II
Fear	I	III	III	III	I	II	II	II	I
Happiness	I	II	III	II	I	II	III	III	I
Sadness	I	II	II	II	I	III	III	III	I
Surprise	I	III	III	II	I	II	III	II	I
Neutral	I	II	III	II	I	II	III	III	I

The low share of *disgust* might explain the high F1-score differences in Table 9 and the generally low F1-scores in Tables 2, 3 and 4. However, the emotional states *anger* and *fear* also have comparatively small shares, but significantly lower F1-score differences and relatively good F1-scores for all models. Emotional states with lower proportions can also achieve quite satisfactory recognition accuracy, for instance, *surprise* on FER2013 and RAF-DB. Beyond, *happiness* with its high appearance in all data sets has small accuracy levels on FER2013 for trained AffectNet models (see Table 9).

The emotional classes are largely equally distributed across all data sets. In all three data sets, *happiness* is the emotional state with the highest share followed by *neutral* and *sadness*. For this reason, we believe that a comparison without further adjustment of the class weights in the training set is valid. However, we are also aware that our analysis has limitations and suggest future research considering class imbalances. This could help to understand the potential impact of class imbalances on our findings. Generally speaking, our analysis provides convincing evidence that recognition accuracy of individual emotional states differs. On the one hand, between individual emotional states, which is known from previous studies as well [14]. On the other hand, recognition accuracy of individual emotions vary (strongly) between different data sets while the image features (e.g. size, color) and training parameters (e.g. epochs) are kept constant. The used model architectures slightly vary in the accuracy, therefore the results lead to the same conclusions. AffectNet data set indicates higher (F1-score)

differences. The least data inconsistencies can be assumed on RAF-DB. Our findings imply data inconsistencies and/or label ambiguity. Possible reasons for these variations in emotional data can be multifactorial. Three potential factors follow.

First, the number of total support and the proportion of emotional classes tend to have an influence on the recognition accuracy of emotional states. This does not apply to all emotional states. The AffectNet data set with most support, reaches the lowest recognition accuracy for three emotions.

Second, reducing the image size and color range of RAF-DB and AffectNet to fit the size of FER2013 could potentially lead to losses of information content. However, interestingly initial experiments without pixel reduction showed the opposite. The AffectNet data set with the highest image resolution and detail information, had generally the lowest recognition accuracy scores.

Third, our findings show that certain emotions, i.e., disgust and *fear*, have lower recognition accuracy. This is in line with previous publications [21,22]. It is worth mentioning that emotions can have different intensities. Plus, differences between certain emotions are not very obvious. Some emotions are very similar and their expressions can be closely related to other emotions. Recently, research has also questioned whether it is valid to assume that facial expressions only contain single emotional states [11]. As consequence, data annotations can be biased and/or incorrect. This can be possible for images carrying higher information content, which leads to higher ambiguity, variability and variance. Therefore, manual image annotation is more difficult and subject to a higher error rate.

The previous analysis clearly shows that Transfer Testing reveals differences between the databases. Through these findings, trained models can be tested for their transferability to other databases. However, further research on Transfer Testing is necessary to determine a measure of transferability and to apply this to real-world scenarios. Overall, the transfer test is a valid method for detecting inconsistencies and/or label ambiguity.

6 Conclusion

In conclusion, this paper presents a comparative analysis to detect label inconsistencies in three commonly used facial expression datasets by Transfer Testing with three different ML model architectures. For this, the data sets have been processed using the same resolution to classify the contained facial images with respect to the expressed emotions. To eliminate possible influences of model architectures, we considered three different types of architectures. Our experiments indicate that the complexity of the ML architectures does not have a significant impact on the overall performance. The transferability among the data sets, on the other hand, deserved a closer look. By Transfer Testing, the presented results demonstrate that recognition accuracy is influenced by the size of the data set and the share for each emotion in it. Furthermore, transferability seems to be (strongly) influenced by the underlying data (label) quality. Transfer testing shows the existence of label biases and/or ambiguity. Furthermore, Transfer Testing shows that the transferability is decisively influenced by this.

All in all, this leads to several future research directions. First, more empirical analysis is required, comparing more data sets. Class imbalances should also be taken into account. Second, investigations are necessary to understand why certain emotions have low recognition accuracy and possible solutions for this challenge. Third, based on our results, it is necessary to investigate a potential relationship between annotation inconsistencies and transferability of ML architectures. Fourth, research is required to minimize and/or distinguish data annotation inconsistencies and label ambiguity as well as the implications which entail with each of them.

AI-based emotion recognition is in general a promising technique for applications. Nonetheless, our results show that AI needs to be applied with great care. On the one hand we should always critically reflect its outcomes, and on the other hand its data input (quality).

References

1. Affectiva - Humanizing Technology (2021). https://www.affectiva.com/
2. NVIDIA Data Science Stack. NVIDIA Corporation (2021)
3. Replika (2021). https://replika.com
4. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
5. Darwin, C.: The Expression of the Emotions in Man and Animals. John Murray, London (1872)
6. Davenport, T., Guha, A., Grewal, D., Bressgott, T.: How artificial intelligence will change the future of marketing. J. Acad. Mark. Sci. **48**(1), 24–42 (2020)
7. Davoli, L., et al.: On driver behavior recognition for increased safety: a roadmap. Safety **6**(4) (2020). https://doi.org/10.3390/safety6040055
8. Ekman, P.: Basic emotions. In: Handbook of Cognition and Emotion, pp. 301–320. Wiley, New York (1999)
9. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124 (1971)
10. Ekman, P., Friesen, W.V.: Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues, vol. 10. ISHK (2003)
11. Ekundayo, O.S., Viriri, S.: Facial expression recognition: a review of trends and techniques. IEEE Access **9**, 136944–136973 (2021). https://doi.org/10.1109/ACCESS.2021.3113464
12. Ekweariri, A.N., Yurtkan, K.: Facial expression recognition using enhanced local binary patterns. In: 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 43–47. IEEE (2017)
13. Gebele, J., Brune, P., Faußer, S.: Face value: on the impact of annotation (in-)consistencies and label ambiguity in facial data on emotion recognition. In: IEEE 26th International Conference on Pattern Recognition (2022)
14. Generosi, A., Ceccacci, S., Mengoni, M.: A deep learning-based system to track and analyze customer behavior in retail store. In: 2018 IEEE 8th International Conference on Consumer Electronics-Berlin (ICCE-Berlin), pp. 1–6. IEEE (2018)
15. Goodfellow, I.J., et al.: Challenges in Representation Learning: A report on three machine learning contests. arXiv:1307.0414 (2013)

16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
17. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745
18. Huang, M.H., Rust, R.T.: Artificial intelligence in service. J. Serv. Res. **21**(2), 155–172 (2018)
19. Huang, M.H., Rust, R.T.: A strategic framework for artificial intelligence in marketing. J. Acad. Mark. Sci. **49**(1), 30–50 (2021). https://doi.org/10.1007/s11747-020-00749-9
20. Jaison, A., Deepa, C.: A review on facial emotion recognition and classification analysis with deep learning. Biosci. Biotechnol. Res. Commun. **14**(5), 154–161 (2021). https://doi.org/10.21786/bbrc/14.5/29
21. Khaireddin, Y., Chen, Z.: Facial Emotion Recognition: State of the Art Performance on FER2013. arXiv:2105.03588 (2021)
22. Knyazev, B., Shvetsov, R., Efremova, N., Kuharenko, A.: Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. arXiv:1711.04598 (2017)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural. Inf. Process. Syst. **25**, 1097–1105 (2012)
24. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791
25. Li, J., Nie, J., Guo, D., Hong, R., Wang, M.: Emotion Separation and Recognition from a Facial Expression by Generating the Poker Face with Vision Transformers. http://arxiv.org/abs/2207.11081
26. Li, S., Deng, W.: Deep facial expression recognition: a survey. IEEE Trans. Affect. Comput. **13**(3), 1195–1215 (2020)
27. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp. 2584–2593. IEEE (2017). https://doi.org/10.1109/CVPR.2017.277
28. Liu, X., Kumar, B.V.K.V., You, J., Jia, P.: Adaptive deep metric learning for identity-aware facial expression recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 522–531 (2017). https://doi.org/10.1109/CVPRW.2017.79
29. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94–101 (2010). https://doi.org/10.1109/CVPRW.2010.5543262
30. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with Gabor wavelets. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205 (1998). https://doi.org/10.1109/AFGR.1998.670949
31. Marín-Morales, J., et al.: Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. Sci. Rep. **8**(1), 1–15 (2018)
32. Mellouk, W., Handouzi, W.: Facial emotion recognition using deep learning: Review and insights. Procedia Comput. Sci. **175**, 689–694 (2020)

33. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (2017)
34. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010). https://doi.org/10.1109/TKDE.2009.191
35. Quinn, M.A., Sivesind, G., Reis, G.: Real-time emotion recognition from facial expressions. Standford University (2017)
36. Rouast, P.V., Adam, M., Chiong, R.: Deep learning for human affect recognition: insights and new developments. IEEE Trans. Affect. Comput. **12**(2), 524–543 (2019)
37. Shao, J., Qian, Y.: Three convolutional neural network models for facial expression recognition in the wild. Neurocomputing **355**, 82–92 (2019)
38. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 (2015)
39. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
40. Tian, Y.I., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell. **23**(2), 97–115 (2001)
41. Yang, H., Ciftci, U., Yin, L.: Facial expression recognition by de-expression residue learning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2168–2177 (2018). https://doi.org/10.1109/CVPR.2018.00231