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

***Abstract*—In recent years, the topic of (facial) emotion recognition has gained considerable attention within the field of Artificial Intelligence (AI). However, the 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 a transfer test with three ML models. The findings indicate inconsistencies in data annotations of emotional states, implying label bias and/or ambiguity. Our empirical analysis emphasizes the need for more interdisciplinary research to collect high-quality labeled (training) data. Such research is the foundation for developing more accurate AI-based emotion recognition systems, which are also robust in real-life scenarios.**

***Index Terms*—Emotion Recognition, Facial Expression Recognition, Emotional Artificial Intelligence, Transfer Testing, Data Quality**

I. Introduction

Over the past few 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 [1], [2].

The use of emotion recognition systems has expanded to various fields, including customer service [3], emotional support [4], and personal relationships [5]–[7].

Numerous studies have been conducted to develop emotion recognition technology using various data modalities and classification taxonomies [8], [9]. Facial expression recognition (FER) is one of the most widely used and promising technologies [10], [11], mainly due to the fact that human emotions are strongly conveyed through facial expressions [12], [13]. The use of computerized FER has been the subject of extensive research in recent years [1], [2].

The primary focus of these studies is to enhance the performance of machine learning models and their architectures as well as the overall model performance [14]. 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). As a result, the significance of inconsistencies in data annotation and/or labeling ambiguity of emotional states remains poorly understood [15].

This paper aims to investigate potential discrepancies in data annotations of facial emotional databases by comparing the recognition accuracy of individual emotional states across various facial expression databases. Our goal is to gain a better understanding of how recognition results for different emotions vary across different data sets using various model architectures. Therefore, multiply models were trained on one data set and tested on another data set. This process ic called transfer testing. By applying transfer testing, we investigate potential discrepancies in data annotations and provide new insights for future research directions.

The rest of this paper is organized as follows: In section II, a detailed review of the existing literature on facial emotion recognition is presented, as well as the relevant machine learning techniques and a description of emotional data sets. The research methodology and data used for the systematic analysis will be find in section III. In section IV the accomplished results will be presented and discussed in depth in section V. The conclusions summarizes our findings.

II. Related Work

A. Facial Expression Recognition

FER combines Psychology [12], [16] and Technology (Computer Vision). This interdisciplinary research field aims to infer human's emotional state to gather highly relevant information contained in facial expressions [17], [18].

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 [16]. Ekman defined six basic emotional states, namely anger, fear, disgust, happiness, surprise and sadness [12], [19]. Focusing on machine learning, 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, [15])

SLL describes a multi-class machine learning 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 [15], the other techniques will not be discussed.

Since computers require binary states, research and practice develop machine learning models, which perform well by assigning one emotional class to a facial expression. That is still the main focus of research. By taking a closer look on the models there are certain dependencies between the machine learning approaches, for instance, SLL can be seen as LDL instances [15]. 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

in one facial expressions carries more than one emotional state [15]. From past research is known that certain emotional states can be recognized better than others [20], [21].

These challenges have been investigated on different facial data sets [22]. 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 support of each emotion has influence on the recognition accuracy, the underlying data (label) quality affects this too [22]. In addition, higher image resolution data sets do not necessarily lead to better recognition results [22].

B. Machine Learning Techniques

There are different approaches for FER in Machine Learning. A general machine learning 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 Machine Learning and/or modern Deep Learning methods can be applied within each phase. Conventional Machine Learning methods consist of Support Vector Machine (SVM), Adaptive Boosting (AdaBoost), Random Forest (RF), Decision Tree [15]. Deep Learning models extract automatically relevant facial features during training [23], [24]. In FER Deep Learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) attract attention [15].

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 [25].

A CNN model automatically extracts features. For this reason, separate feature extraction methods like in traditional machine learning algorithms are not necessary [15]. Some different popular CNN architectures are listed below in chronological order: LeNet-5 [25], AlexNet [26], GoogleLeNet [27], VGGNet [28], ResNet [29], Xception [30], SENet [31]. 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.

C. 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 [15]. Databases set up in a controlled environment are for instance The Extended Cohn-Kanade data set (CK+) [32] and The Japanese Female Facial Expression (JAFFE) database [33]. 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 [34] and Real-world Affective Faces Database (RAF-DB) [35]. Most of these databases include six basic emotional states [36], usually adding one neutral facial expression. Therefore, emotional labels can be annotated manually by experts [34], by computers or by a combination of these [37].

III. Methodology

A. Technical Environment

We implemented all machine learning models on our institute server. It runs on Ubuntu 20.04 LTS, including the NVIDIA data science stack [38]. 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.

B. 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 [22]. 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, wich 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 x 48 pixels [37]. 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 x 100 pixels [35]. The mini AffectNet consists of 291,650 only manually annotated images in RGB-color with 224 x 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 [34].

TABLE I
Distribution of Emotional Classes per Data Set

Emotion	FER-2013		RAF-DB		AffectNet	
Pixel Size	48 x 48		100 x 100		224 x 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%)

C. Data Pre-Processing

The pre-processing stage covers typically different methods. For instance, face detection, facial landmark localization, face

normalization and data augmentation [18]. Face localization is the first step. The previously described data sets have already aligned and cropped images. That is why we limit pre-processing 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 I.

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. After that we split it the same ways as the others. By the end of this publication we want to 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, will be evaluated on AffectNet. Since FER2013 only has decolorized images, we turn all images into black and white.

D. Deep Learning Model Architecture

As already mentioned, we implement various CNN models. First of all we need to consider 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 decided to use three models which use stacked CNN layers directly. The first one is AlexNet [26]. The secound model is a standard CNN model based on AlexNet [26]. In the following this CNN model will be called defaultNet. The architecture 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 as 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 model we decided to use VGGNet [28] as our last model.

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 of training epoch with the highest validation accuracy.

IV. Results

In this section, we present emotion recognition accuracy of the seven basic emotional states for every of the three data sets evaluated for the three architectures. 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-score, which represents the harmonic mean of precision and recall. The following results evaluate the trained models on every test set for each one of the three data sets, i.e. the AlexNet which was trained on RAF-DB will be evaluated on FER2013. Table II shows the evaluation on FER2013 for the trained models. Furthermore, the results for RAF-DB are represented in Table III and for AffectNet in IV.

TABLE II
F1-scores on FER2013

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)

For each data set, we run the model 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 with higher occurrence to have lower standard deviations, for instance, happy and neutral in the AffectNet data set for every model architecture. The variation in F1-scores for each trained model on remaining data sets is conspicuous. For better impression on the impact of the model on the results Tabel V 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.

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

V. Discussion

By focusing on the models we conclude that the impact of the models on the result is not the crucial factor. All models

TABLE III
F1-SCORES ON RAF-DB

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)

TABLE IV
F1-SCORES ON AFFECTNET

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 V
ACCURACY AS WEIGHTED AVERAGE

Test Set	Train Sets	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

tend to the same results. VGGNet, the most complex model, tends to have higher accuracy. Obviously the models trained and tested on the same data set provide the best accuracy. The results of our analysis in Table II - IV show that the emotional state happiness is best recognizable in every data set independent of the model while transfer testing on the same data set. Evaluating FER2013 and RAF-DB trained models on these mutual data sets still recognizes happiness the best. AffectNet seems to differ from these two data sets since the recognition ranking vary in order during 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 VI 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 VIII such as for AffectNet in Table VII. The emotional state surprise in the AffectNet data set represents the major exception in the ranking while training and testing on the same data set. Furthermore, the other emotions hardly vary in order in all three data sets for all models. As soon as we evaluate the AffectNet trained models on one of the remaining data sets we get results which highly vary from the patterns.

TABLE VI
RECOGNITION ACCURACY ORDINAL RANKING 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

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

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

Focusing disgust, the F1-scores differences are the worst doing training and testing on the same data set. Therefore,

TABLE VII
RECOGNITION ACCURACY ORDINAL RANKING 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

TABLE VIII
RECOGNITION ACCURACY ORDINAL RANKING 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

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 I. Fear is in AffectNet and RAF-DB underrepresented, too, and accordingly the F1-score difference is high. In FER2013 fear seems to have some label inconsistency, since the corresponding F1-score differences, are always in high position. The strong F1-score variations in certain emotions is a further sign of potential irregularities in the underlying data sets.

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

TABLE IX
F1-SCORE DIFFERENCES PER EMOTIONAL STATE ACROSS ALL MODELS AND ALL DATA SETS

Trained	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 II - IV, we present a ranking for every emotional state based on F1-scores in every data sets among all models. Table X 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 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 among of images. The emotions where RAF-DB has a lower rank have a small number in it. 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.

The low share of disgust might explain the high F1-score differences in Table IX and the generally low F1-scores in Tables II - IV. 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,

TABLE X
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.

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 IX).

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. Overall, 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 [10]. On the other hand, recognition accuracy of individual emotions vary (strongly) between different data sets, here between three facial expression databases, 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 instances 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 instances, reaches the lowest recognition accuracy for three emotions.

Second, reducing the image size and color range of RAF-DB and AffectNet to fit the requires 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 [20], [39]. 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 [15]. 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.

VI. CONCLUSION

In conclusion, this paper presented an approach to detect label inconsistencies in commonly used facial expression datasets by a comparative analysis of a transfer test of three different ML models. Therefore, these 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 models. Initial experiments indicate that the complexity of the used ML architectures does not have a significant impact on the overall performance of the emotion recognition. The portability 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 support for each emotion in it. Furthermore, it seems to be (strongly) influenced by the underlying data (label) quality.

All in all, this leads to several future research directions. First, more empirical analysis is required, comparing more data sets and complex machine learning models. 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 portability of machine learning 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] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, 2020, review on Machine Learning Models.

[2] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: Review and insights," *Procedia Computer Science*, vol. 175, pp. 689–694, 2020, review on Machine Learning Models.

[3] "Affectiva - Humanizing Technology," https://www.affectiva.com/, Nov. 2021.

[4] "Replika," https://replika.com, Nov. 2021.

[5] T. Davenport, A. Guha, D. Grewal, and T. Bressgott, "How artificial intelligence will change the future of marketing," *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 24–42, 2020.

[6] L. Davoli, M. Martalo, A. Cilfone, L. Belli, G. Ferrari, R. Presta, R. Montanari, M. Mengoni, L. Giraldi, E. G. Amparore, M. Botta, I. Drago, G. Carbonara, A. Castellano, and J. Plomp, "On Driver Behavior Recognition for Increased Safety: A Roadmap," *Safety*, vol. 6, no. 4, Dec. 2020.

[7] M.-H. Huang and R. T. Rust, "Artificial intelligence in service," *Journal of Service Research*, vol. 21, no. 2, pp. 155–172, 2018.

[8] ——, "A strategic framework for artificial intelligence in marketing," *Journal of the Academy of Marketing Science*, vol. 49, no. 1, pp. 30–50, Jan. 2021, the following values have no corresponding Zotero field: accession-num: WOS:000585729700001.

[9] J. Marín-Morales, J. L. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, E. P. Scilingo, M. Alcañiz, and G. Valenza, "Affective computing in virtual reality: Emotion recognition from brain and heartbeat dynamics using wearable sensors," *Scientific reports*, vol. 8, no. 1, pp. 1–15, 2018.

[10] A. Generosi, S. Ceccacci, and M. Mengoni, "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)*. IEEE, 2018, pp. 1–6.

[11] J. Shao and Y. Qian, "Three convolutional neural network models for facial expression recognition in the wild," *Neurocomputing*, vol. 355, pp. 82–92, 2019.

[12] C. Darwin, *The Expression of the Emotions in Man and Animals*. John Murray, 1872.

[13] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[14] P. V. Rouast, M. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Transactions on Affective Computing*, 2019, review on Machine Learning Models.

[15] O. S. Ekundayo and S. Viriri, "Facial Expression Recognition: A Review of Trends and Techniques," *Ieee Access*, vol. 9, pp. 136 944–136 973, 2021, overview of FER databases on p. 9
Summary of popular Deep CNN
Summary of experimental results for different models and datasets.

[16] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971, emotion Theory Concepts, for instance ekman and pleasure-arousal-dominance framework (PAD) or newer concepts like Plutchnik model.

[17] A. N. Ekweariri and K. Yurtkan, "Facial expression recognition using enhanced local binary patterns," in *2017 9th International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE, 2017, pp. 43–47.

[18] A. Jaison and C. Deepa, "A Review on Facial Emotion Recognition and Classification Analysis with Deep Learning," *Bioscience Biotechnology Research Communications*, vol. 14, no. 5, pp. 154–161, 2021, methodological Approach NOT to complex.

[19] P. Ekman, "Basic emotions. Handbook of cognition and emotion," *Wiley, New York*, pp. 301–320, 1999.

[20] Y. Khaireddin and Z. Chen, "Facial Emotion Recognition: State of the Art Performance on FER2013," *arXiv:2105.03588 [cs]*, May 2021, comment: 9 pages, 5 figures, 2 tables.

[21] M.-A. Quinn, G. Sivesind, and G. Reis, "Real-time emotion recognition from facial expressions," *Standford University*, 2017.

[22] J. Gebele, P. Brune, and S. Faußer, "Face Value: On the Impact of Annotation (In-)Consistencies and Label Ambiguity in Facial Data on Emotion Recognition," *IEEE 26th International Conference on Pattern Recognition*, 2022.

[23] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, "Adaptive Deep Metric Learning for Identity-Aware Facial Expression Recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul. 2017, pp. 522–531.

[24] H. Yang, U. Ciftci, and L. Yin, "Facial Expression Recognition by De-expression Residue Learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 2168–2177.

[25] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[28] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Apr. 2015.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778.

[30] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7132–7141.

[32] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "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*, Jun. 2010, pp. 94–101, the following values have no corresponding Zotero field:
alt-title: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops.

[33] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, Apr. 1998, pp. 200–205.

[34] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[35] S. Li, W. Deng, and J. Du, "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: IEEE, Jul. 2017, pp. 2584–2593.

[36] P. Ekman and W. V. Friesen, *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*. Ishk, 2003, vol. 10.

[37] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in Representation Learning: A report on three machine learning contests," *arXiv:1307.0414 [cs, stat]*, Jul. 2013, comment: 8 pages, 2 figures.

[38] "NVIDIA Data Science Stack," NVIDIA Corporation, Dec. 2021.

[39] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video," *arXiv:1711.04598 [cs]*, Nov. 2017, comment: 4 pages.