Description of Changes

We thank the reviewers for their valuable comments and comprehensive feedback, which we thoroughly addressed in the revised version. In the following, we outline the changes we made to the paper in order to incorporate the reviewers' feedback.

1) Associate Editor

1.1) The associate editor asked for further clarification on the conclusion about the correlation between accuracy and image resolution.

On page six, in Section VI, in the third paragraph, we replaced our original strong claim about this prevalent relationship and added instead, that initial experiments indicate that data sets with higher image resolution does not necessarily lead to better emotion recognition results. In the next paragraph, we explain that further investigations are necessary to understand whether there is a potential relationship between image resolution and recognition accuracy.

1.2) The editor asked for reasons why we excluded a validation set and why there was one included in the code.

Originally, we believed our experiments can be conducted without a validation set. However, after reading the reviews we understood that adding a validation set leads to more meaningful results, even though our work does not concentrate on hyperparameter (tuning). As a consequence, we have repeated our experiments for all three data sets and included a validation set. On page 4, in the first paragraph we updated the methodological part respectively. The validation set has now a 30% share of the original training set. The initial split for training (80%) and test set (20%) remained the same. The new results do not contradict our original statements and findings. There are still similar differences in recognition accuracy of emotions between data sets, indicating data inconsistencies/label ambiguity. Consequently, we have updated the results in Table III to VII.

In the code, the actual test set was simply incorrectly referred to as the validation set.

1.3) The associate editor asked for explanations why we did not address class imbalances

On page five, in the last paragraph, we additionally explained that the emotional classes are largely equally distributed across all data sets and happy, neutral, and sad, in that order, are the most frequent emotional states in each data set. Nevertheless, we are aware that our comparing experiments have limitations. For this reason, we have additionally pointed out that future research considering class imbalances is needed, see second last paragraph, on page six.

1.4) The editor asked to address the reviewers' comments on consistency indicator(s) and the difference in meaning between data inconsistencies and label ambiguity.

In Section I, on page one, in the last three paragraphs, we added definitions of label ambiguity and data inconsistency and explained why we use these terms synonymously in this work. In addition, on page six, in Section VI, in the second last paragraph we highlighted that future research is necessary to understand the role of each of these two aspects on the overall emotion recognition accuracy.

In the following we address the other still remaining comments / questions from the reviewers.

2) Reviewer 4

2.1) Reviewer 4 expected solutions to the problem of label ambiguity. Moreover, the reviewer expected further analysis on the reasons why certain emotional categories are difficult to recognize.

The aim of our paper is first to provide evidence that there is label ambiguity and/or data inconsistencies. We have added the need to investigate in the future reasons why certain emotions are difficult to recognize and possible solutions to this problem to page six, in Section VI, in the second last paragraph.

3) Reviewer 5

3.1) Reviewer 5 asked how we accounted for differences in image resolution/quality between data sets that may require more training or more complex CNN architectures. Since these could also be factors responsible for differences in emotion recognition accuracy.

While repeating our experiments with the validation set again (1.2), we also resized the images of RAF-DB and AffectNet data set to the pixel size of FER2013 data set (48px * 48px). In Section III, page 3, in the last paragraph, we updated the methodology respectively. The results after resizing do not contradict our original statements and findings. Consequently, we have updated the results in Table II to VII. Through resizing we now can exclude that these differences are related to image resolution, CNN architecture complexity and number of training epochs. We run all experiments with 50 training epochs, excluding early stopping. The updated methods can be found in the last two paragraphs of Section III. [Remark: Among all three data sets, training, monitoring validation accuracy, converged fastest for AffectNet (~17 epochs), followed by FER2013 (~30 epochs) and RAF-DB (~33 epochs)]

3.2) Reviewer 5 asked also to define data inconsistency and label ambiguity. Plus, to explain why a small F1-score interval for FER2013 data set could not also be an indicator of data consistency.

For the definition of the terms, please refer to 1.4).

We believe that F1-score intervals between data sets may also indicate consistencies. However, the F1-score interval ranks vary between lowest and 2nd lowest ranked emotional state (see Table V). The FER2013 data set has the lowest interval difference only between highest and second lowest emotion, not between highest and lowest. In our view, F1-score intervals can be meaningful indicators of inconsistency or consistency when we consider both interval dimensions together. First, F1-score intervals for data sets and second, for individual emotional classes. After the results are not consistent in this respect, we infer rather inconsistencies than consistency.

4) Reviewer 7

4.1) Reviewer 7 requested further explanation of the initially excluded validation set.

Please refer to 1.2).

4.2) Reviewer 7 addressed the topic of class imbalances.

Please refer to 1.3).

4.3) Reviewer asked why we did not consider image resizing for a better comparison.

Please refer to 3.1).

Face Value: On the Impact of Annotation (In-)Consistencies and Label Ambiguity in Facial Data on Emotion Recognition

Jens Gebele
Technology-Transfer-Centre Günzburg
Neu-Ulm University of Applied Sciences
Neu-Ulm, Germany
Jens.Gebele@hnu.de

Philipp Brune
Technology-Transfer-Centre Günzburg
Neu-Ulm University of Applied Sciences
Neu-Ulm, Germany
Philipp.Brune@hnu.de

Stefan Faußer
Technology-Transfer-Centre Günzburg
Neu-Ulm University of Applied Sciences
Neu-Ulm, Germany
Stefan.Fausser@hnu.de

***Abstract*—Artificial Intelligence (AI)-based emotion recognition using various kinds of data has attracted vast attention in recent years. Impressive results have been achieved, but only recently the influence of the training data with its potential biases and variations in annotation quality are discussed. Still, the majority of the research literature focuses on improving machine learning techniques and model performance using single data sets. Literature on the impact of training data remains scarce. Therefore, in this paper we investigate the influence of the training data on the accuracy of recognizing emotional states in facial expressions by a comparative evaluation, using multiple established facial image databases. Results reveal inconsistencies in the data annotations as well as ambiguities in the emotional states expressed. Thus, they allow to critically discuss data quality of the training data, contributing to a more in-depth understanding of previous emotion recognition approaches, and improving the design of more transparent AI solutions.**

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

I. Introduction

In recent years, AI-based emotion recognition has received large attention in research as well as in practice. Basically, it enables machines to determine the emotional state of humans [1], [2].

In practice, we encounter more and more emotion recognition systems in different application areas. For instance, the company Affectiva provides technological solutions to detect and analyze customers' emotions while they consume media content like movies, video clips or advertisement [3]. Another example is the application Replika, an emotional chatbot, that enables meaningful interactions based on human's feelings, emotions and experiences [4]. In the context of marketing, machines with such capabilities enable empathic automatic conversations with customers [5]–[7].

In this way, virtual agents can comfort angry customers, provide emotional support, adapt to human's expression style, or even establish personal relationships [8], [9]. This raises human-machine interaction to a new level and one step closer to emotional artificial intelligence.

Existing research covers computerized emotion recognition based on different data modalities, technologies, and emotion classification taxonomies [9], [10]. One common emotion classification concept suggests the following six basic emotional states: anger, fear, disgust, happiness, surprise, and sadness [11]. To recognize these states, emotion recognition systems commonly use visual, audio, text, and/or bio-physiological data modalities [5], [10], [12], [13].

Of these technologies, facial expression recognition (FER) belongs to the most common and promising approaches [14], [15]. The reason is that facial expressions contain very strong signals of human's emotional state [16], [17]. The role of computerized facial emotion recognition has been extensively studied in recent years [1], [2].

These studies mainly focus on improving machine learning models and architectures and overall model performance, respectively [18]. However, research has been mainly concentrated on single data sets only or very few different data sets and has paid less attention to the underlying data (quality). In consequence, the role of (in-)consistencies in data annotation and/or (label) ambiguity of emotional states are still poorly understood [19].

Therefore, in this paper the recognition accuracy of individual emotional states for different facial expression databases is investigated in depth to better understand how recognition results for individual emotions vary across different data sets. For this, we treat both terms, data annotation inconsistencies and label ambiguity synonymously because differentiation between these terms requires further research.

Data inconsistency defines errors in annotations of facial expressions, which means they have been incorrectly labeled. Label ambiguity refers to the theory that facial expressions carry more than one emotional state at the same time [19].

Potential deviations provide information that at least one of the two aspects is present. Thus, we advance the theoretical understanding of facial emotion recognition systems

Manuscript 479 submitted to 26th International Conference on Pattern Recognition. Received April 16, 2022.

and provide new insights for future research directions in this area. To the best of our knowledge, this is the first work that systematically evaluates recognition accuracy of individual emotions in different data sets.

The rest of this paper is organized as follows: In section II, the existing literature on facial emotion recognition is reviewed in detail, and the relevant state-of-the-art machine learning techniques and emotional data sets are described. Section III presents the research methodology and data used for the systematic analysis. The obtained results are illustrated in section IV, and discussed in depth in section V. We conclude with a summary of our findings.

II. Related Work

A. Facial Expression Recognition

FER is an interdisciplinary research field which combines Psychology [16], [20] and Technology (Computer Vision). The aim is to infer human's emotional state from facial expressions because these contain highly relevant information [21], [22]. Past research resulted in various machine learning models and facial databases.

Most of the research on facial emotion recognition is based on Paul Ekman's work, which supports the idea that dependencies between certain facial expressions and human emotional states are not affected by different cultural backgrounds [20]. According to him, there are six basic emotional states, namely anger, fear, disgust, happiness, surprise and sadness [16], [11]. Ekman elaborated also a Facial Action Coding System (FACS) which describes emotional states by facial movements, so called Action Units (AUs) [23]. From a machine learning perspective, emotion recognition can be differentiated into the following four different tasks:

- Single Label Learning (SLL)
- SSL Extension (extended by Intensity Estimation)
- Multi-Label Learning (MLL)
- Label Distribution Learning (LDL)

SLL describes a multi-class machine learning problem in which one emotional class is identified from several possible emotional states in a facial expression based on the highest likelihood. The second type, the SSL Extension, considers in addition to the emotional class, its prevalent intensity. This can be measured as the difference between the neutral facial expression and the present facial expression.

MLL refers to the prediction of one or more possible emotional states from a single facial expression. The assumption here is that a facial expression can carry several emotional states at the same time. LDL infers an emotion with corresponding proportion of its occurrence in an image [19]. Despite recent developments in FER, computerized emotion detection remains still a challenging task for machines [21]. In addition, all of the before mentioned four machine learning types have their limitations [19].

This study concentrates on limitations directly linked to SLL. The reason is that facial emotion recognition represents basically a SLL task, as computers require binary states. As a consequence, research and practice aim to build machine learning models that perform well by assigning one emotional class to a facial expression. This still enjoys the highest research attention. Moreover, there are certain dependencies between the machine learning approaches, for instance, SLL and MLL can be seen as LDL instances [19]. Our work deals with a two-sided aspect of data annotations in SLL tasks. Firstly, data annotations (labels) can have inconsistencies/biases as they are either manually or automatically generated. Secondly, recent research claims that facial expressions contain more than one emotional state [19]. To address these challenges, we investigate different facial data sets in order to find out to what extent the recognition accuracy of single emotions differs. The results of this analysis provide interesting insights for future research. From past research is known that certain emotional states can be recognized better than others [24], [25]. A systematic investigation allows to estimate to what extent both aspects described above are predominant.

B. Machine Learning Techniques

FER uses different machine learning techniques to achieve successful inference of emotional states from facial expressions. In general, a machine learning process consists of up to three distinct phases. First preprocessing phase, second feature extraction phase, which can be optional, and third emotion recognition or rather classification phase. Within each phase different conventional Machine Learning and/or modern Deep Learning methods can be applied. Conventional Machine Learning methods comprise Support Vector Machine (SVM), Adaptive Boosting (AdaBoost), Random Forest (RF), Decision Tree [19]. Nowadays, Deep Learning models are state-of-the-art. They are capable of extracting automatically relevant facial features while training [26], [27]. Among Deep Learning models, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) enjoy the greatest interest in FER [19].

In particular CNN belong to the most popular deep learning classifiers in Computer Vision and are conceptually based on Artificial Neural Network (ANN). A CNN consists of multiple layers similar to a Deep Neural Network (DNN). (If an ANN contains at least two hidden layers, it is referred to as a DNN.) The main architectural elements of a CNN are convolutional layer(s), pooling layer(s), dense layer(s) and fully connected layer(s). The input layer, for instance, is a colored image with three channels for red, green and blue (RGB). The task of the convolutional layer(s) is to train the relevant features, starting from low-level features in early layers, up to high-level (abstract) features in higher levels of the stack. The pooling layer(s) follow(s) convolutional layer(s), aggregate information and thereby reduce computational complexity [28].

A CNN model, similar to a DNN model, automatically extracts features. For this reason, separate feature extraction methods like in traditional machine learning algorithms are not mandatory [19]. In the past, different popular CNN architectures have been developed. Some of them are listed below in chronological order: LeNet-5 [28], AlexNet [29], Google-

LeNet [30], VGGNet [31], ResNet [32], Xception [33], SENet [34]. Over the years, the architectures have evolved and got more complex. For instance, the amount of layers increased to more than 150 (ResNet-152). Further on, convolutional layers have been stacked directly. Plus, inception modules, residual learning (with skip connections) and depthwise separable convolution layer have been introduced.

C. Emotional Facial Databases

Previous research in FER has led to a number of different facial databases. These can differ based on data type (static, sequential), data dimension (two-dimensional, three-dimensional), data collection environment (controlled, uncontrolled), and number of facial expressions [19]. Databases collected in a controlled environment are for instance The Extended Cohn-Kanade data set (CK+) [35] and The Japanese Female Facial Expression (JAFFE) database [36]. Typically, systems based on these data sets reach only lower performance in real-world scenarios. As a consequence, research demand increased for databases collected in an uncontrolled setting, so called in-the-wild scenario. Examples are AffectNet [37] and Real-world Affective Faces Database (RAF-DB) [38]. In most of these databases six basic emotional states are present [39]. In addition, one neutral facial expression is usually added. That said, emotional labels can be annotated manually by experts [37], by computers or by a combination of both [40].

III. Methodology

A. Technical Environment

We implemented all machine learning models on our institute server, which runs on Ubuntu 20.04, including the NVIDIA data science stack [41]. The server is equipped with two NVIDIA A40 Graphics Processing Units. We developed the code in Python, used Jupyter Notebook as integrated development environment, and made use of these Python frameworks: NumPy, Matplotlib, Pandas, Scikit-Learn, Keras and TensorFlow.

B. Data Collection

Initially, we reviewed over 40 different facial expression databases. Of these, we considered those containing six basic emotions (anger, fear, disgust, happiness, surprise and sadness) [11] and one neutral expression. Following that, we excluded multimodal and three-dimensional data sets, as we only focus on facial two-dimensional data sets. This left us with the following databases, for which we then requested access:

- FER2013 [40]
- The Cohn Kanade (CK) Dataset [42]
- CK+ [35]
- Acted Facial Expressions in the Wild (AFEW) [43]
- Static Facial Expressions in the Wild (SFEW) [44]
- JAFFE [36]
- Karolinska Directed Emotional Faces (KDEF) [45]
- Averaged Karolinska Directed Emotional Faces [46]
- RAF-DB with Basic Emotions [38]
- AffectNet (Mini Version) [37]

These databases still differ in size, data type (static, sequential) and data collection environment (controlled, uncontrolled). In the next step, we excluded databases with a size of less than 10,000 instances and/or the ones collected in a controlled environment. The reason is, we aim to have a sufficient quantity of data, more realistic and representative data, as well.

As a result of that, the following three databases remained. First, the FER2013 data set, which consist of 35,887 gray images. These were automatically cropped, labeled and then cross-checked by experts. This data set contains seven emotional classes and all images are resized to a format of 48 x 48 pixels [40].

Second, the RAF-DB, with basic emotions, which contains in total 15,339 aligned colorful RGB-images. These were manually annotated by approximately 40 experts and aligned to a size of 100 x 100 pixels [38].

Third, Affect Net with eight labels (the so-called Mini Version). The smaller version consists of only manually annotated images in RGB-color. In total, 291,650 images with a size of 224 x 224 pixels each. We removed the emotional state contempt, leaving us with the same seven emotions as in the other data sets [37].

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
Angry	4,953	867	25,382
Disgust	547	877	4,303
Fear	5,121	355	6,878
Happy	8,989	5,957	134,915
Sad	6,077	2,460	25,959
Surprise	4,002	1,619	14,590
Neutral	6,198	3,204	75,374
Total	35,887	15,339	287,401

C. Data Pre-Processing

The pre-processing stage covers typically different methods, among them are, for instance, face detection, facial landmark localization, face normalization and data augmentation [22]. The first step is face localization. The images of the previously described data sets had already been provided in an aligned and cropped form. As a result, we limit pre-processing to data normalization and augmentation. Face localization and facial landmarks are not considered further.

In addition, we resize the images of RAF-DB and AffectNet to the pixel size of FER2013 (48px * 48px), in order to have equal conditions for the comparison. We combine normalization and data augmentation method. For this purpose, we divide on the fly each image pixel by 255, which results in a range of values per pixel from 0 to 1. The total distribution of the emotional classes in the data sets is presented in Table I.

The pixel size is similar across the three different data sets. Most emotional states are sufficiently well represented in the three datasets, with a few exceptions. We split every data set into one training and one test set, with ratios of 80 percent to 20 percent. In addition, we use 30 percent of the training set as validation set. The splits are all stratified in order to keep the proportions of each emotional class equal in training, validation and test set. AffectNet is provided with a small test set. For this reason, we first combine training and test set. Then we split them again to have the same ratios for training, validation and test data as for the other two data sets.

D. Deep Learning Model Architecture

As already mentioned, we implement a standard CNN model. Our aim is to conduct a systematic comparative analysis which considers the overall model performance only indirectly. Our goal is to compare the emotion recognition accuracy for individual emotional states in different data sets. This does not require to reach a certain performance threshold. For this reason, the CNN architecture is somewhat based on AlexNet [29], which was the first model that stacked CNN layers directly. Our final CNN architecture is structured in Table II.

TABLE II
CNN MODEL ARCHITECTURE USING THE EXAMPLE OF RAF-DB

Layer	Output Shape	Parameters
Conv2D	(None, 48, 48, 32)	2,432
Conv2D	(None, 48, 48, 32)	25,632
MaxPooling2D	(None, 24, 24, 32)	0
Conv2D	(None, 24, 24, 64)	18,496
Conv2D	(None, 24, 24, 64)	36,928
MaxPooling2D	(None, 12, 12, 64)	0
Conv2D	(None, 12, 12, 128)	73,856
Conv2D	(None, 12, 12, 128)	147,584
MaxPooling2D	(None, 6, 6, 128)	0
Conv2D	(None, 6, 6, 256)	295,168
Conv2D	(None, 6, 6, 256)	590,080
MaxPooling2D	(None, 3, 3, 256)	0
Flatten	(None, 2,304)	0
Dense	(None, 128)	295,040
Dropout	(None, 128)	0
Dense	(None, 64)	8,256
Dropout	(None, 64)	0
Dense	(None, 7)	455

Basically, 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.

For the training of our model 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 the model of the 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. The outcome metrics are limited to precision, recall and F1-score as these are relevant to answering our research questions(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. Table III represents the final averaged results of five training runs.

TABLE III
CLASSIFICATION REPORT

Emotion	Precision	Recall	F1-Score	Count
Angry-FER	0.49 (±0.02)	0.51 (±0.01)	0.50 (±0.01)	991
Angry-RAF	0.55 (±0.02)	0.52 (±0.02)	0.53 (±0.01)	173
Angry-Aff	0.59 (±0.03)	0.50 (±0.04)	0.54 (±0.01)	5,076
Disgust-FER	0.35 (±0.44)	0.04 (±0.08)	0.07 (±0.13)	109
Disgust-RAF	0.51 (±0.09)	0.19 (±0.09)	0.26 (±0.10)	175
Disgust-Aff	0.00 (±0.00)	0.00 (±0.00)	0.00 (±0.00)	861
Fear-FER	0.42 (±0.02)	0.33 (±0.05)	0.37 (±0.04)	1,024
Fear-RAF	0.68 (±0.17)	0.23 (±0.11)	0.31 (±0.14)	71
Fear-Aff	0.50 (±0.06)	0.20 (±0.03)	0.28 (±0.02)	1,376
Happy-FER	0.78 (±0.01)	0.79 (±0.01)	0.79 (±0.01)	1,798
Happy-RAF	0.85 (±0.02)	0.89 (±0.01)	0.87 (±0.01)	1,192
Happy-Aff	0.88 (±0.01)	0.91 (±0.01)	0.89 (±0.00)	26,983
Sad-FER	0.45 (±0.02)	0.50 (±0.03)	0.47 (±0.02)	1,216
Sad-RAF	0.58 (±0.02)	0.54 (±0.06)	0.56 (±0.04)	492
Sad-Aff	0.57 (±0.02)	0.44 (±0.03)	0.50 (±0.01)	5,192
Surprise-FER	0.71 (±0.01)	0.70 (±0.01)	0.70 (±0.00)	800
Surprise-RAF	0.67 (±0.01)	0.66 (±0.03)	0.67 (±0.02)	324
Surprise-Aff	0.50 (±0.04)	0.36 (±0.03)	0.42 (±0.01)	2,918
Neutral-FER	0.51 (±0.02)	0.56 (±0.02)	0.53 (±0.01)	1,240
Neutral-RAF	0.61 (±0.03)	0.74 (±0.03)	0.67 (±0.01)	641
Neutral-Aff	0.62 (±0.02)	0.75 (±0.03)	0.68 (±0.00)	15,075

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. The deliberate stratified split into training, validation and test set leads to class imbalances within every data set. However, in our view this is the best approach for a valid comparison in terms of data quality. In the next section, we discuss results, similarities and differences in the recognition accuracy of individual emotional states and work out possible reasons for this.

V. DISCUSSION

The results of our analysis in Table III show that the emotional state happiness is best recognizable in every data

set. Fear and disgust are the most difficult emotional states to recognize in all data sets. Table IV illustrates a ranking of recognition accuracy for every emotional state in every data set based on F1-score. The emotional state surprise in the AffectNet data set represents the major exception in the comparative ranking. The other emotions hardly vary in order in all three data sets. The finding of one outlier in the comparative ranking is the first indication of data inconsistencies.

TABLE IV
RECOGNITION ACCURACY ORDINAL RANKING

Rank	FER2013	RAF-DB	AffectNet
1	Happiness	Happiness	Happiness
2	Surprise	Surprise	Neutral
3	Neutral	Neutral	Angry
4	Angry	Sadness	Sadness
5	Sadness	Angry	Surprise
6	Fear	Fear	Fear
7	Disgust	Disgust	Disgust

In addition to the ordinal ranking, we observe differences in the F1-score intervals between the best and worst recognizable emotional state. In the FER2013 data set F1-scores range from 0.79 (highest) to 0.07 (lowest). For RAF-DB, F1-scores vary between 0.87 (highest) and 0.26 (lowest). The AffectNet data set F1-scores range from 0.89 (highest) to 0.00 (lowest).

These differences are less pronounced when the last emotion class of the ranking is removed. It is quite plausible to do so, since the last ranked class is also among the classes that are proportionally least present in each data set. Even then, AffectNet has still the widest F1-score interval from 0.89 (highest) to 0.28 (lowest) between the best and the second worst recognizable emotional state.

The F1-score intervals are illustrated in Table V. The proportions of all emotional states are presented in Table VIII. We believe that differences in F1-score intervals are an additional indicator of data inconsistencies in facial data.

TABLE V
F1-SCORE INTERVALS BETWEEN HIGHEST AND (2ND) LOWEST RANKED EMOTIONAL STATE

Data Set	F1-score Interval Highest to Lowest	F1-score Interval Highest to Second Lowest
FER2013	0.79 - 0.07	0.79 - 0.37
RAF-DB	0.87 - 0.26	0.87 - 0.31
AffectNet	0.89 - 0.00	0.89 - 0.28

Furthermore, it is worth taking a closer look at every emotion spectrum separately. There are differences between the best and the worst F1-score for every emotional state in the three data sets, as well. Highest F1-score variances occur in disgust and surprise. From this we can assume that these two emotions have the highest label inconsistency.

In case of disgust, this may be, among other things, due to the fact that this class has a very low share in every data set, which can be seen in Table VIII. Whereas the occurrence of surprise is quite comparable to the best result, which is anger, and the third best result, which is fear. Table VI provides an overview. The strong F1-score variations in certain emotions is a further sign of potential irregularities in the underlying data sets.

TABLE VI
F1-SCORE DIFFERENCES PER EMOTIONAL STATE ACROSS ALL DATA SETS

Data Set	Rank	Max F1-score differences
Angry	1	0.04
Sad	2	0.09
Fear	3	0.09
Happy	4	0.1
Neutral	5	0.15
Disgust	6	0.26
Surprise	7	0.28

In accordance with the main ranking in Table III, we present also a ranking for every emotional state based on F1-scores in every of the three data sets. Table VII indicates best recognition accuracy in RAF-DB data set, as no emotional state contains the worst rank. FER2013 has the best F1-score for two emotions and the worst F1-score for four emotions. AffectNet has in three out of seven emotions the lowest F1-score and in three the best rank. Overall, this leads to the assumption that RAF-DB has the lowest data inconsistencies, while FER2013 and AffectNet have higher inconsistencies.

TABLE VII
RECOGNITION ACCURACY F1-SCORE RANKING

Emotion	FER-2013	RAF-DB	AffectNet
Angry	III.	II.	I.
Disgust	II.	I.	III.
Fear	I.	II.	III.
Happy	III.	II.	I.
Sad	III.	I.	II.
Surprise	I.	II.	III.
Neutral	III.	II.	I.

In this context, it is useful to look at the proportions of each emotional class in the total number of instances in each data set. Due to the stratified split into training, validation and test data, class imbalances are present, illustrated in Table VIII.

At first glance, the low share of disgust might explain the high F1-score differences in Table VI and the generally low F1-scores in Table III. However, the emotional states anger and fear also have comparatively small shares, but significantly lower F1-score differences and relatively good F1-scores. Moreover, emotional states with lower proportions can also achieve quite satisfactory recognition accuracy, for instance, surprise in the FER2013 and RAF-DB data set.

The emotional classes are largely equally distributed across all data sets. In all three data sets, Happy is the emotional state with the highest share followed by Neutral and Sad. 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.

TABLE VIII
PROPORTION OF EMOTIONAL STATES IN EVERY DATA SET

Emotion	FER-2013	RAF-DB	AffectNet
Angry	14%	6%	9%
Disgust	2%	6%	1%
Fear	14%	2%	2%
Happy	25%	39%	47%
Sad	17%	16%	9%
Surprise	11%	11%	5%
Neutral	17%	21%	26%

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 [14].

On the other hand, recognition accuracy of individual emotions vary (strongly) between different data sets, here between three facial expression databases, while the model architecture, image size and the number of training epochs is kept constant.

In particular, the AffectNet data set indicates F1-score differences compared to both other data sets, which tend to have more in common. The RAF-DB data set can be assumed to contain the least data inconsistencies. These findings imply data inconsistencies and/or label ambiguity.

Possible reasons to explain these variations in emotional facial data can be multifactorial. We present three potential factors in the following.

First, the number of total instances and the proportion of emotional classes tend to have an influence on the recognition accuracy of emotional states. However, in this analysis this does not apply to all emotional states. Moreover, the AffectNet data set with most instances, reached lowest recognition accuracy for three emotional classes.

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

Third, our findings show that certain emotions, for instance, disgust and fear, have basically lower recognition accuracy. This is in line with previous publications [24], [47]. In that respect, it is worth mentioning that emotions can have different intensities. Plus, differences between certain emotions are not very obvious. Some emotions are very similar to each other and their expressions can be closely related to other emotions. Lately, research challenges also whether it is valid to assume that facial expressions contain only single emotional states at one time [19].

As a consequence, data annotations can be biased and/or incorrect. This can be possible, in particular, for images carrying higher information content, which leads to higher ambiguity, variability and variance in facial expression images. Therefore, manual image annotation of emotional classes is more difficult and subject to a higher error rate.

VI. CONCLUSION

In conclusion, in this paper the quality and impact on emotion recognition of three commonly used facial expression data sets have been investigated by a comparative analysis. Therefore, all the data sets have been processed using the same resolution and the same Convolutional Neural Network (CNN) model to classify the contained facial images with respect to the expressed emotions.

The presented results demonstrate that the recognition accuracy of automated emotion recognition depends not only on the size of the data set and the support for each emotion in it, but also seems to be (strongly) influenced by the underlying data (label) quality.

In addition, initial experiments indicate that higher image resolution data sets do not necessarily lead to better recognition results. A possible interpretation is, that more details contained in higher-resolutions images lead to a larger ambiguity in the expressed emotions and, therefore, a less definite classification by the human labelers. However, this needs to be further investigated in depth.

All in all, there are several future research directions, that we outline in the following. First, empirical analysis that compare more data sets and different machine learning models are required. For this, class imbalances should also be taken into account. Second, investigations are necessary to understand why certain emotions are difficult to recognize (and/or annotate) and which possible solutions can be provided to this challenge. Third, based on our initial experiments, it is necessary to investigate a potential relationship between image resolution and detection accuracy. Fourth, research is required to explore methods how data annotation inconsistencies and label ambiguity can be distinguished and what implications each of them has.

Nevertheless, as AI-based emotion recognition is in general a promising technique for applications, e.g. for analyzing consumers' perceptions in sales and marketing, our results make it clear that AI needs to be applied with great care here and we should always critically reflect its outcomes and in particular also 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] K. French, "Your new best friend: AI chatbot," 2018, the following values have no corresponding Zotero field:
number: 31.05.2021.

[9] M.-H. Huang and R. T. Rust, "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.

[10] 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.

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

[12] K. T. Johnson and R. W. Picard, "Advancing Neuroscience through Wearable Devices," *Neuron*, vol. 108, no. 1, pp. 8–12, 2020.

[13] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. Rahaman, and T. Zia, "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals," *Journal of Network and Computer Applications*, vol. 149, p. 102447, 2020.

[14] 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.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] E. Friesen and P. Ekman, "Facial action coding system: A technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, p. 5, 1978.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.

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

[42] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. IEEE, 2000, pp. 46–53.

[43] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol,"

in *Proceedings of the 16th International Conference on Multimodal Interaction*, 2014, pp. 461–466.

[44] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expressions in tough conditions: Data, evaluation protocol and benchmark," in *1st IEEE International Workshop on Benchmarking Facial Image Analysis Technologies BeFIT, ICCV2011*, 2011.

[45] D. Lundqvist, A. Flykt, and A. Öhman, "The Karolinska Directed Emotional Faces (KDEF), 1998," *Department of Neurosciences Karolinska Hospital: Stockholm, Sweden*, 1998, belongs to KDEF-original data set.

[46] D. Lundqvist and J. E. Litton, "The averaged Karolinska directed emotional faces-AKDEF," *AKDEF CD ROM. Psychology section, Karolinska Institutet, Stockholm*, 1998, belongs to AKDEF-original data set.

[47] 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.