TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Ruslan Kononov 194250IVGM

Evaluation of Facial Emotion Recognition Models for the Potential Deployment in Web-Based Learning Environments

Master's thesis

Supervisor: Sadok Ben Yahia, PhD
Co-supervisor: Silvia Lips, LL.M, MSc

Tallinn 2021

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Ruslan Kononov 194250IVGM

Näo emotsioonide tuvastamise mudelite võimalik rakendamine veebipõhistes õppekeskkondades

Magistritöö

Põhijuhendaja: Sadok Ben Yahia, PhD
Kaasjuhendaja: Silvia Lips, LL.M, MSc

Tallinn 2021

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Ruslan Kononov

06.05.2021

Abstract

The COVID-19 pandemic has had an extraordinarily disruptive impact on the educational systems all over the world. The frequent suspension of classes and the near-total closures of thousands of educational institutions worldwide have created an insuperable impediment to conventional face-to-face classes, making remote learning the only feasible alternative. Unfortunately, learners' facial expressions — a valuable natural source of instant feedback to the learning experience for any educator — are commonly overlooked during online sessions. Although modern Deep Learning (DL) systems classifying human emotions by facial expressions may effectively tackle this challenge, the general reasoning and decision-making of those systems often remain ignored by their evaluators.

This study addresses this problem by proposing a robust, theory-driven, and case-oriented evaluation framework enabling preliminary selection of the facial expression recognition (FER) models that provide accurate, valid and trustworthy information on students' learning experience in a web-based learning environment. Contrary to the existing evaluation approaches, the proposed framework goes beyond conventional performance metrics (e.g., accuracy), shifting the focus to detecting potentially inherent biases, evaluating algorithmic generalisation capabilities and models' interpretability. By following the Design Science (DS) research methodology, this study also includes the controlled experiment that validates the proposed evaluation criteria and serves as an input for a new cycle of the artefact's building and refinement. The proposed evaluation framework is arguably the first documented scientific attempt to enable comprehensive analysis and evaluation of FER models for their eventual deployment in web-based learning environments.

This thesis is written in English and is 116 pages long, including 4 chapters, 52 figures and 10 tables.

Keywords: facial expression recognition (FER), interpretability, explainable artificial intelligence (XAI), data biases, black-box models, evaluation framework.

Annotatsioon

COVID-19 pandeemia on avaldanud erakordselt häirivat mõju haridussüsteemidele kogu maailmas. Sagedane kontaktõppe peatamine ja tuhandete haridusasutuste peaaegu täielik sulgemine kogu maailmas on takistanud tavapärastele näost näkku tundide läbiviimist, muutes distantsõppe ainsaks võimalikuks alternatiiviks. Kahjuks jäetakse online-sessioonide ajal kahe silma vahele õppijate näoilmed, mis on õpetajale väärtuslikuks allikaks õpikogemuse osas kohese tagasiside saamiseks. Ehkki tänapäevased süvaõppe (DL) süsteemid, mis klassifitseerivad inimese emotsioone näoilmete järgi, võivad selle väljakutsega tõhusalt toime tulla, siis jäävad nende süsteemide üldised põhjendused ja otsustused tihti hindajate poolt tähelepanuta.

Selles uurimuses käsitletakse just seda probleemi, pakkudes välja tugeva, teooria- ja juhtumipõhise hindamisraamistiku võimaldades näo tuvastamise (FER) mudelite eelvalikut, mis annab täpset ja usaldusväärset teavet õpilaste õppimiskogemuse kohta veebitundides. Vastupidiselt olemasolevatele hindamiskäsitlustele läheb kavandatav raamistik kaugemale tavapärasest jõudluse mõõdikust (nt täpsus), suunates fookuse võimalike korduvate omaduste tuvastamisele, algoritmide üldistusvõime hindamisele ja mudelite tõlgendatavusele. DS-i uurimismetoodika põhimõtteid järgides hõlmab see uuring ka katset kontrollitud keskkonnas, mis kinnitab pakutud hindamiskriteeriumid ja on sisendiks artefakti ehitamisele ja täiustamisele uue tsükli jaoks. Kavandatav hindamisraamistik on vaieldamatult esimene dokumenteeritud teaduslik katse võimaldada FER-mudelite põhjalikku analüüsi ja hindamist nende võimaliku juurutamise jaoks veebipõhistes õpikeskkondades.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 116 leheküljel, 4 peatükki, 52 joonist, 10 tabelit.

Märksõnad: näoilmete tuvastamine (FER), tõlgendatavus, seletatav tehisintellekt (XAI), andmete eelarvamused, musta kasti mudelid, hindamisraamistik.

List of abbreviations and terms

ADR	Action design research
AI	Artificial intelligence
ANN	Artificial neural network
API	Application programming interface
CNN	Convolutional Neural Network
ConvLSTM	Convolutional Long Short-Term Memory
DARPA	The Defense Advanced Research Projects Agency
DCPN	Deeper Cascaded Peak-piloted Network
DL	Deep learning
DNN	Deep neural network
DR	Design research
EU	The European Union
FACS	Facial Action Coding System
FAN	Frame Attention Network
FER	Facial expression recognition
GDPR	General Data Protection Regulation
GPU	Graphics processing unit
HOG	Histogram of Oriented Gradients
IT	Information technology
LBP	Local Binary Patterns
LBP-TOP	LBP on Three Orthogonal Planes
LIME	Local Interpretable Model-Agnostic Explanations
LPQ	Local Phase Quantisation

LRP	Layer-wise Relevance Propagation
LSTM	Long Short-Term Memory
ML	Machine learning
NMF	Non-Negative Matrix Factorisation
PGS	Peak Gradient Suppression
PPDN	Peak-Piloted Deep Network
RNN	Recurrent Neural Network
RNN	Recurrent neural network
SHAP	SHapley Additive exPlanations
SIFT	Scale-invariant Feature Transform
SpRAy	Spectral Relevance Analysis
SVM	Support Vector Machine
t-SNE	t-Stochastic Neighbourhood Embedding
VOC	Visual Object Classes
XAI	Explainable Artificial Intelligence

Table of contents

List of figures

List of tables

1 Introduction and Problem Statement

In 2020 and 2021, for thousands of education institutions around the world, distance learning became probably the only viable option to guarantee that the fundamental right to education is still indefeasible. Unfortunately, the global transition to remote learning has not been an example of an evolutionary transformation or gradual development of education. On the contrary, the abrupt and hasty decision to shut down education institutions and switch to online learning was a desperate attempt to curb the COVID-19 pandemic. Although distance learning might help reduce social interactions and slow down the spread of the virus, its limitations may negatively affect students' performance and academic success. In particular, learners' emotions — crucial for stimulating attention and triggering the learning process (Linnenbrink-Garcia & Pekrun, 2012) — are commonly overlooked during online sessions.

Given that students' mental health and physiological well-being tend to worsen during the pandemic (Essadek & Rabeyron, 2020; Khan et al., 2020; Jiang, 2020; Kecojevic et al., 2020; Elmer et al., 2020), observing emotions of learners during online classes may be crucial for ongoing efforts of schools and universities to make better-informed decisions on the teaching strategies, modes of interactions and support. Due to recent developments in the field of Deep Learning (DL), a vast range of algorithms are now able to recognise human emotions based on facial expressions with astonishing accuracy (Benitez-Quiroz et al., 2016; Mollahosseini et al., 2016; Lopes et al., 2017; Zeng et al., 2018). However, the decision-making of the DL models often requires comprehensive analysis and evaluation to become reliable, trusted and tailored to the needs of their end-users (Samek et al., 2017; Holzinger et al., 2019; Guidotti et al., 2019; Rothman, 2020). Although some recent studies investigated the application of facial expression recognition (FER) algorithms in e-learning systems (El Hammoumi et al., 2018; Sun et al., 2018; Zhang et al., 2020;), most of them focused on performance and accuracy, vastly overlooking potentially inherent biases, generalisation and interpretability of the proposed models. This study aims to address this gap by proposing a robust, theory-driven and case-oriented evaluation framework enabling preliminary selection of the most

suitable FER models that can guarantee accurate, unbiased, and trustworthy information on students' learning experience for any educator working in a web-based learning environment.

This study's intended audience includes two primary groups. The first group consists of FER algorithm developers and researchers concerned with robust algorithmic comparability and evaluation that go beyond the analysis of conventional performance metrics. The second group includes IT specialists tasked with the selection of a particular FER model for a web-based learning environment in which the risks implied by "black-box" architecture and data-related biases are mitigated.

2 Research Design

As the ultimate goal of this thesis is to design, build and evaluate an innovative IT artefact — i.e., evaluation framework for FER models — this research is oriented towards Design Science (DS) research. In particular, the study follows several guidelines for Design Science in Information Systems Research outlined and proposed by Hevner et al. (2004).

As suggested by the first guideline, the resulting evaluation framework is *designed as a purposeful IT artefact* intended to address the specific organisational problem in the domain of distance learning. |However, as noted in the original paper, the proposed framework may not and should not solve the problem *per se* (Hevner et al., 2004). Instead, the ultimate goal of the resulting IT artefact in DS is to generate knowledge and define practices applicable to the class of field problems that the research problem exemplifies.

As recommended by the second guideline, the research is carried out by ensuring the *relevance* of the proposed evaluation framework *to the specific business problem* — i.e., valid and trustworthy facial expression analysis in web-based learning environments. In particular, the research focuses on the thorough problem formulation and in-depth investigation of the potential solutions that create solid premises for the initial design of the evaluation framework and ensure its relevance to the specific use-case.

Moreover, by following the third guideline of DS methodology, this thesis lays the foundation for *rigorous* and *comprehensive evaluation design*. Given the subjects' novelty, complexity and specificity, the artefact's evaluation is carried out via the controlled experiment. Its primary goal is to study the proposed framework in the controlled environment, discover its potential shortcomings and validate its core components (i.e., evaluation criteria). The overall robustness and replicability of the experiment and subsequent evaluation are ensured by using publicly available algorithms and real training data – i.e., pre-trained FER models and datasets with labelled facial expressions.

Moreover, recognising that the design is inherently an iterative and incremental process, the evaluation results from the controlled experiment also serve as an input for a new cycle of the artefact's building and refinement. On the one hand, it helps the proposed evaluation framework better meet the requirements and constraints of the problem it aims to solve. On the other hand, it views designing and building the IT artefact as a search process whose end goal is to discover an optimal solution. The latter is suggested by the sixth guideline proposed by Hevner et al. (2004) as an effective strategy to create feasible, good designs that can be later implemented in the business environment.

2.1 Research Questions

The core research question of the thesis is as follows:

RQ: What evaluation criteria should be applied to FER models for their further adjustment to and deployment in a web-based learning environment, so that educators can receive valid and trustworthy information on students' learning experience?

As DS methodology suggests, answering this research question and proposing the initial design of the IT artefact (i.e., the evaluation framework) will require capturing the two seemingly conflicting perspectives. On the one hand, the framework should include genuinely technical criteria or requirements that permit comparisons and rigorously demonstrate the quality and efficacy of any FER model considered for the deployment in a web-based environment. For example, this study focuses on performance and interpretability as major technical properties of FER algorithms suggested by the scientific community. On the other hand, the proposed evaluation framework as an IT-artefact should accommodate the needs and limitations imposed by the organisational context – i.e., peculiarities of the teaching-learning process and the role of emotions in this process, the nature of facial expressions and peculiarities of their recognition. Importantly, DS research recognises the inseparable influences mutually exerted by the technical domain and the organisational context. Therefore, the research findings related to the latter are bound to shape selection, use or interpretation of the technical design constructs from the former, and vice versa.

To accomplish the goal of combining technical and non-technical perspectives into the solid and coherent artefact, four sub-questions have been framed. Each sub-question

seeks to describe a particular component of facial expression recognition as a technical or social phenomenon with its respective limitations and possibilities. This information will be used to discover existing as well as to infer new evaluation criteria for FER models in a web-based learning environment.

SQ1: How emotions of learners may influence their learning experience and academic performance in a web-based learning environment?

Based on a thorough literature review, the presumed link between students' emotions and academic performance will be investigated. In particular, objective evidence on the role of specific emotions in the learning process will be summarised to be used for inferring both technical and non-technical evaluation criteria for FER algorithms selection.

SQ2: What is the taxonomy of modern FER algorithms?

Based on a comprehensive literature review, distinct characteristics and generic classification of various FER algorithms will be identified. The discovered advantages and shortcomings of the particular groups of FER algorithms for web-based learning environments will be translated into the specific evaluation and selection criteria.

SQ3: What are the tools and methods of Explainable AI (XAI) suitable for the validation of FER models given their algorithmic complexity and high-dimensional input?

Based on a thorough literature review, general taxonomy of the most relevant XAI tools and methods will be investigated. The research findings in this Section (3) will become a theoretical basis for deciding what XAI techniques can be used to validate FER models' rationale and decision-making for each class of emotions.

SQ4: What are the evaluation criteria currently used to evaluate different FER algorithms?

Based on a comprehensive literature review, this Section (4) will provide an overview of the well-established evaluation metrics and criteria used for the variety of FER approaches and algorithms discovered in Section 2. The identified criteria will become the backbone of the proposed evaluation framework balancing scientifically proven conventional measures with those inferred in Sections 1, 2, 3.

Figure 1 illustrates how different research stages and components defined by the DS methodology were integrated into this research.

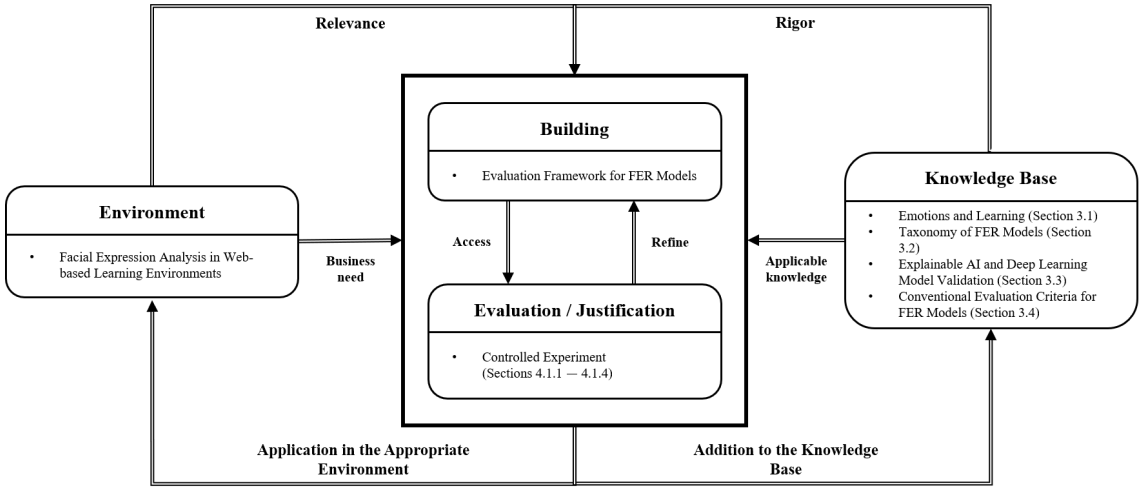

Figure 1. Research framework oriented towards the Design Science research methodology.

2.2 Limitations

This thesis focuses primarily on the thorough problem formulation and in-depth investigation of the potential solutions that create solid theoretical premises for the initial design of the evaluation framework. The research also contains the controlled experiment that attempts to challenge the initial, theory-driven design of the proposed framework and adjust it to the use case based on empirical evidence. However, due to the temporal limitations of this research, this thesis does not include multiple cycles of the framework's design refinement and evaluation that seem indispensable for ensuring the artefact's generality, validity, reliability and utility.

Moreover, the specificity and novelty of the research problem make the involvement of practitioners with the relevant domain expertise extremely valuable and yet challenging. Considering the imposed time limit and complexity of this research, the direct and comprehensive inclusion of the practitioners' perspective fall naturally out of its scope. Therefore, the resulting evaluation framework should be regarded as a generic construct that undoubtedly requires further adjustment to the specific environment and organisational context. Moreover, in the absence of any qualitative or quantitative input from the domain experts, the proposed framework neither includes any scoring nor assigns weights to any particular evaluation criteria. Furthermore, as this research deals with the highly novel and insufficiently studied realm of Explainable AI, even the most

rigorous theories and assumptions ingrained in the evaluation framework should be repeatedly tested and be subject to users' assumptions, expectations, and knowledge.

Additionally, the controlled experiment itself includes several significant limitations. Firstly, considering the absence of funding for the research, the controlled experiment does not include analysis and evaluation of commercial FER models available on the market. Secondly, the controlled experiment does not include any FER models trained to distinguish compound emotions based on Auction Units (AU) determined in FACS. This particular limitation is imposed by the relatively small number of publicly available datasets with the annotated AU. The latter would be an unavoidable impediment to testing the pre-selected FER on different unseen data. Moreover, to enable a profound analysis of the models' decisions via various attribution methods during the controlled experiment, this thesis utilises the DeepExplain[1] framework – the only framework for XAI in Python supporting a wide range of gradient-based and perturbation-based local explanation techniques. As DeepExplain is compatible with Tensorflow and Keras libraries only, the FER models trained and compiled with different libraries (e.g., PyTorch) were dismissed for the experiment. Thus, further research and more comprehensive testing may reveal other weaknesses or ambiguities of the proposed evaluation framework that were unnoticed during the controlled experiment.

Last but not least, the primary goal of the proposed evaluation framework is to provide robust theory-driven and case-oriented criteria enabling effective comparison, filtering, and selection of the most suitable FER models for potential deployment in web-based environments. However, the given framework assumes that even the best performing models with perfect criteria satisfaction need further comprehensive adjustment to specific organisational and user needs. Thus, the proposed evaluation framework may be unsuitable for organisations or individuals looking for fully-fledged FER systems with an option of immediate deployment.

[1] Available at https://github.com/marcoancona/DeepExplain

3 Related Work

3.1 Emotions and Learning

Section 3.1 aims to explore and provide scientific evidence that forms the basis for an answer to the first research sub-question (SQ1). Essentially, this Section investigates the literature on the presumed link between students' emotions and academic performance. The evidence on the role of specific emotions in the learning process discovered in this Section is summarised and translated into specific evaluation criteria in the proposed framework (see Section 4).

The academic community has extensively explored the presumed link between students' emotions and academic performance. Previous studies have shown compelling evidence that emotions do play a crucial role in students' learning outcomes (Linnenbrink-Garcia & Pekrun, 2012). For instance, Pekrun et al. (2011) suggest that negative emotions such as anger, anxiety, shame, hopelessness, and boredom are linked to students' use of learning strategies, self-regulation of learning, and academic performance. Skinner et al. (2008) generally confirm that behavioural and emotional components are positively correlated, therefore, students' disengagement, withdrawal, and academic failure can result from negative emotions associated with learning. Williams et al. (2013) investigate how students' experience of positive emotions in the classroom environment can stimulate and enhance learning. Their findings indicate that students who experience positive emotions during classes are likely to be academically successful, dedicating more resources to studying, attending classes, participating in classroom discussions, and engaging in extracurricular activities. Um et al. (2012) argue that positive emotions can increase motivation, satisfaction, and perception toward the didactic materials.

Some studies also discuss how emotions can affect memory storage and retrieval, attention and problem-solving capabilities. Nielson & Lorber (2009) evaluated that emotional arousal experienced by students after learning new information contributes to better information retention and retrieval. A recent study by Vogel & Schwabe (2016) concludes that negative emotions such as emotional stress or confusion have far-reaching

consequences on students' ability to learn and remember academic material. Timely identification of emotional instability among students may help educators personalise approaches or tailor training programmes to prevent stress-induced impairments.

Several groups of researchers have also investigated how teachers' communication styles and behaviours are associated with the emotions that students experience in the classroom. Maresh (2007) was one of the first researchers who studied how students react to various classroom communication styles. He explores that students experience negative emotions and try to save face by changing majors or avoiding future interactions with the instructor who enacts heartful messages. Mazer et al. (2014) identify that students who deal with teachers lacking in clarity and communication competence are more likely to report an increased level of deactivating emotions such as shame, boredom, hopelessness, and a heightened level of activating emotions like anxiety and anger. Conversely, Skinner et al. (2008) present that in situations when teachers communicate in a supportive manner, learners report higher levels of emotional engagement and lower levels of boredom, frustration, and anxiety. Moreover, Titsworth et al. (2013) demonstrate that active listening, emotional support and clarity in teachers' communication behaviour tend to arouse positive emotions such as enjoyment, hope, and pride.

It is worth mentioning that learning experience during online sessions may show different emotion dynamics when compared to the dynamics resulting from traditional classroom instructions. As such, online or distance learning may differ from the traditional one due to temporal, spatial, and technical conditions. Students who study in web-based learning environments must exercise a higher degree of self-regulation to succeed academically. They have to manage the time, place, and progress of their learning more effectively and to a greater extent than their classroom counterparts (Dabbagh & Kitsantas, 2004). Moreover, as the success in online studies depends largely on self-evaluation and self-management of a learner (Yukselturk & Bulut, 2007), the importance of providing timely and meaningful feedback and encouragement by educators increases (Eom et al., 2006).

Although the difference between online learning and learning in the traditional classroom environment for students is significant, little scientific research has been conducted to show the link between students' emotions and learning outcomes in web-based environments. Artino Jr & Jones II (2012) addressed this research gap by investigating the relations between several achievement-related emotions – i.e., frustration, boredom,

and enjoyment – and some self-regulated learning behaviours – i.e., elaboration and metacognition – in a fully online course. Their findings reveal that online learning environments rely on the same theoretical and empirical evidence from the prior research in traditional classrooms (Pekrun et al., 2006; Pekrun et al., 2002). Additionally, Bosch & D'Mello (2017) call upon broader use of the so-called affect-aware learning technologies in web-based learning environments to detect specific emotional states of learners and identify when an intervention from instructors is needed. In particular, the findings presented in their paper suggest that the above-mentioned technologies should focus on recognising confusion, engagement, frustration, curiosity, and boredom as the most frequent and informative emotional states.

3.2 Facial Expression Recognition

Section 3.2 thoroughly investigates the taxonomy of modern FER algorithms to provide an answer to the second research sub-question (SQ2). In particular, the scientific literature reviewed in this Section identifies distinct characteristics and generic classification of various FER algorithms. The discovered advantages and shortcomings of the particular groups of FER algorithms for a web-based learning environment are translated into specific evaluation criteria in the proposed framework (see Section 4).

According to various studies, nonverbal behaviour and communication are at least as important as verbal (Cuddy et al., 2015; Jacob et al., 2016). Among several nonverbal components, facial expressions may arguably be the most explicit in revealing an individual's actual emotional state. Indeed, a brisk nod, a broad smile or slightly lifted eyebrows can convey precise information about human emotional states, intentions and mood (Ochs et al., 2015). Moreover, ample evidence on the universality of facial expression of emotions (Matsumoto, 2001; Elfenbein & Ambady, 2002) also sparked scientific interest and gained scrupulous attention in the field of computer vision and Machine Learning (ML). As a result, various facial expression recognition (FER) systems have been proposed to encode expression information and features from facial representations.

According to the survey conducted by Li & Deng (2020), most of the FER systems can identify six prototypical emotions initially suggested by Ekman and Friesen in 1971. Those are anger, disgust, fear, happiness, sadness, and surprise. Some of the systems also

include contempt that was also proposed by Ekman & Friesen (1986) as another not culture-specific basic emotion. Although some complex models (Du & Martinez, 2014; Zhang et al., 2017) based on the Facial Action Coding System (FACS) (Ekman, 1997) can distinguish compound emotions, the categorical models that limit human emotions to "discrete basic emotions are still the most popular perspective for FER, due to its pioneering investigations along with the direct and intuitive definition of facial expressions" (Li & Deng, 2020).

Generally, all FER systems can be divided into two groups by the approaches used for feature extraction and classification. The so-called conventional or traditional FER systems have a clear prediction process flow consisting of three distinctive steps: (1) face and facial component detection, (2) feature extraction, and (3) input classification. The major difference between more recent – deep learning (DL) – methods and traditional FER systems lies in the fact that feature extraction schemes and classification methods in those DL models are not handcrafted (Ko, 2018). For instance, the Local Binary Patterns (LBP) (Shan et al., 2009), the LBP On Three Orthogonal Planes (LBP-TOP) (Zhao & Pietikainen, 2007), Local Phase Quantisation (LPQ) (Chan et al., 2009), Histogram of Oriented Gradients (HOG) (Carcagnì & Distante, 2015), or the Non-Negative Matrix Factorisation (NMF) (Buciu & Pitas, 2004) belong to the group of the handcrafted feature descriptors widely used in traditional FER systems. Figure 2 illustrates the differences in training between conventional and DL FER models.

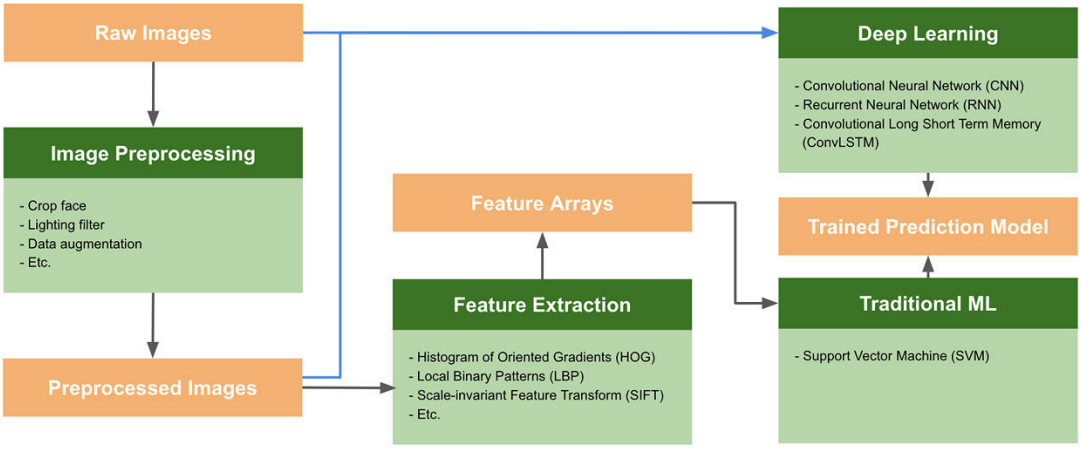

Figure 2. Process flow diagram depicting the training process for conventional and DL FER models (Perez, 2018).

In the meanwhile, due to the dramatic increase in computing power, speed-up gains obtained with powerful GPU cards and recent improvements in deep learning methodology (LeCun et al., 2012), FER based on deep learning techniques have attracted more academic attention. Moreover, since the 2010s, more training data have become available (Li & Deng, 2020). Hence, recent FER systems empowered by multi-layered structure have demonstrated greater accuracy and performance when compared to the conventional FER algorithms (e.g., Szegedy et al., 2015; Jung et al., 2015; Hamester et al., 2015). In addition, as deep neural networks (DNNs) have the ability to extract relevant features from the training data without human tuning, they succeed in learning more implicit patterns of the input and thus perform better in real-world scenarios.

3.2.1 Deep-Learning Based FER Approaches

Most of the state-of-the-art DNN algorithms for FER can be divided into static-based FER and sequence-based FER algorithms. In static-based FER systems, a feature vector contains numerical information about the current input only – i.e., a single image or a video frame with recorded facial expression. Conversely, sequence-based FER systems presume temporal correlation among consecutive frames of a video, so they analyse and recognise the expression based on multiple frames' input. Sequence-based FER systems are also often called dynamic (Dong et al., 2018; Sun et al., 2016; Sun et al., 2020).

Although a comparatively higher number of the deep FER models focus on the analysis of static images (Revina & Emmanuel, 2018), in the real-world scenarios – where learners display facial expressions dynamically, e.g., from slightly subtle to explicit, and vice versa – applying DNNs to video data may provide a far more accurate picture on emotion dynamics to educators. Therefore, video- or sequence-based deep FER algorithms (e.g., Hasani & Mahoor, 2017; Zhang et al., 2019;) may better analyse learners' facial expressions at a particular moment as well as their variations over time. Li & Deng (2018) group sequence-based deep FER algorithms into three categories: FER with *frame aggregation*, *expression intensity-invariant* and *spatiotemporal* FER systems.

In a FER system with *frame aggregation*, each frame of a sequence gets an *n*-class probability vector. These probabilities are then aggregated to create a fixed-length representation for each video. As the length of a sequence cannot be known, two aggregation techniques have been proposed to ensure that a fixed-length video descriptor is generated: frame averaging and frame expansion (Kanou et al., 2013; Kahou et al.,

2016). Bargal et al. (2016) suggest generating a feature vector of the entire video sequence with a statistical encoding module (STAT). In particular, frame aggregation is achieved by computing the mean, variance, minimum, and maximum of each frame's feature vectors.

One obvious limitation of the proposed aggregation methods is that they regard the importance of each frame for FER equally. In reality, a video or an image set may contain frames with facial expressions captured under different conditions (e.g., varying lighting conditions, head poses, camera angles, different occlusions). For better performance, an algorithm should select the more discriminative frames and downweigh the importance of frames with low expression density for the final recognition (Yang et al., 2017). Meng et al. (2019) introduce Frame Attention Network (FAN) that extracts frame-level features from a video and aggregates them via a weighted averaging. Due to discriminative frame selection, fixed-size feature representation of the video fragments may become more accurate and informative. Moreover, computational simplicity and fair accuracy of *frame aggregation* methods make them highly efficient and quite popular.

However, as frame aggregation techniques focus on recognising high-intensity (i.e., peak) expressions and discard lower-intensity (i.e., non-peak) expressions, they may fail to classify the emotions of learners whose facial expressions are less intense. Moreover, as non-peak expressions are more common than peak expressions, it may be even more important to distinguish them for higher accuracy and better discriminative capabilities. Expression intensity-invariant networks address this challenge by training on data samples with different expression intensities indicated in input to exploit the intrinsic correlations between peak and non-peak facial expressions. Zhao et al. (2016) propose a peak-piloted deep network (PPDN) architecture, which is based on a heuristic that peak and non-peak facial expressions from the same subject often show significant visual correlations (e.g., similar face attributes) and can naturally contribute to the recognition of one another. The authors embed the evolution of expressions into the PPDN framework by applying the L2-norm loss function to the feature maps of non-peak and peak expression images. Then, the PPDN is trained with a novel back-propagation algorithm – peak gradient suppression (PGS) – that "drives the feature responses to non-peak expressions towards those of the corresponding peak expressions, while avoiding the inverse". As a result, the proposed algorithm performs intensity-invariant FER by effectively recognising the most frequent non-peak expressions.

Yu et al. (2018) propose a refined version of the PPDN to boost recognition of low-intense expressions and fundamentally improve overall FER accuracy. The authors introduce a deeper cascaded peak-piloted network (DCPN) that takes advantage of a deeper and larger architecture, capturing the subtle details of week (i.e., non-peak) expressions with higher accuracy. A new integration training method – cascaded fine-tuning – is proposed to prevent the enlarged network architecture from overfitting. Nevertheless, even though *expression intensity-invariant networks* may demonstrate adequate performance, they require prior knowledge of expression intensity, which may be quite challenging to collect and label in real-world scenarios.

When compared to *frame aggregation* and *expression intensity-invariant networks*, deep *spatiotemporal networks* can encode temporal dependencies in consecutive frames and learn spatial features along with temporal features. The deep learning community has established several tools that capture both temporal and spatial features for sequence-based FER. The most popular are those enabled by Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and 3D Convolutional Neural Network (Byeon & Kwak, 2014). Ebrahimi Kahou et al. (2015) model the spatiotemporal evolution of facial expressions using an RNN combined with a Convolutional Neural Network. The higher layer representation from the CNN provides structural information of a frame, and the RNN models the spatiotemporal evolution of the structure over time. Kim et al. (2019) propose a spatiotemporal learning method that also incorporates qualities of *expression intensity-invariant* networks. In particular, they suggest using the CNN with representative expression-states (i.e., onset, onset to apex transition, apex, apex to offset transition and offset) to learn spatial feature representation of the facial expressions. In order to capture the facial expression dynamics, temporal feature representation of the facial expression is learned via the LSTM. As a result, the proposed method can generate discriminative spatiotemporal feature representations that improve FER performance at different expression intensities.

Additionally, some researchers expand a 2D structure of CNN to a 3D structure for dynamic FER (Byeon & Kwak, 2014; Ji et al., 2012). Due to 3D convolution, both spatial and temporal features from video data are extracted, thus capturing motion information in video streams. Li et al. (2019) propose a FER model that also builds upon 3D-CNN and additionally incorporates an optical flow for more accurate micro-expression detection in sequence-based FER. Hasani & Mahoor (2017) introduce a 3D Inception-

ResNet architecture accompanied by an LSTM unit. The authors also extract and incorporate facial landmarks, which improve the recognition of subtle changes in the facial expressions in a sequence. As we see, RNN and its variations (e.g., LSTM) along with 3D-CNN are apparently the most well-studied networks for learning spatiotemporal features. Nevertheless, the performance of these networks is somewhat poor. Moreover, training and utilising such extensive and complex networks is computationally expensive.

3.3 Explainable Artificial Intelligence (XAI)

Section 3.3 provides an answer to the third research sub-question (SQ3) by exploring the XAI tools and methods suitable for the validation of FER models given their algorithmic complexity and high-dimensional input. Essentially, this Section identifies the general taxonomy of the most relevant XAI tools and methods and shortly discusses their main advantages and shortcomings. The research findings in Section 3.3 form a theoretical basis for deciding what XAI techniques can be used to validate FER models' rationale and decision-making via the proposed evaluation framework (see Section 4).

The exponential development in machine learning and artificial intelligence (AI) has resulted in numerous complex algorithms that transform, disrupt or give impetus to the emergence of entirely new industries and sectors. As argued earlier, deep learning models for FER have become remarkably accurate and at times successfully surpass human performance. However, the difficulty of explaining the decisions made by any modern DNN has grown proportionally to the progress made. Due to inherent multi-layer architecture and consequent non-linearity, state-of-the-art FER systems become highly non-transparent. Essentially, developers of such FER systems may not determine with certainty what information in the input data makes these algorithms arrive at specific predictions. Thus, these models are typically seen as "black boxes". To address this lack of transparency, a completely new field of research began to emerge – Explainable AI.

3.3.1 Definition and Nature of XAI

Explainable AI (XAI) as a term was first coined and described by Van Lent et al. (2010), who determined that XAI systems "can explain their behaviour either during execution or after the fact". Nevertheless, there is no standard and generally accepted definition of Explainable AI. Sometimes this term may generally refer to some initiatives, projects,

and efforts seeking to achieve higher transparency, interpretability and trust in the AI-enabled systems. According to David Gunning from DARPA (2017) – an organisation that lavishly funds XAI-related research – XAI aims to "produce more explainable models while maintaining a high level of learning performance (i.e., prediction/classification accuracy); and enable human users to understand, appropriately, trust, and effectively manage the emerging generation of artificially intelligent partners".

Unfortunately, along with somewhat ambiguity around the term XAI, there is no universally agreed definition of interpretability. Lipton (2018) suggests that AI interpretability reflects several distinct ideas and concepts that may refer to trust, systems' architecture or extensive explanations behind each prediction. Doshi-Velez & Kim (2017) argue that interpretability in the context of ML is "the ability of a system to explain or to present its predictions and decisions in understandable terms to humans". Therefore, the interpretability of FER in the context of web-based learning environments should be determined by and tailored to its end-users, e.g., educators.

In academic literature, interpretability and explainability are usually used as synonyms. For instance, Doran et al. (2017) use terms understanding, explaining and interpreting often interchangeably. Nevertheless, some researchers emphasise certain differences. Montavon et al. (2018) suggest that an interpretation is the mapping of abstract concepts into a domain that humans can understand (e.g., sequences of words or arrays of pixels), whereas an explanation is a collection of features of the interpretable domain (e.g., specific pixels displayed as a heatmap) that have contributed for a given input to reach a decision (e.g., classification). Rothman (2020) proposes a definition that goes pretty much in line with Montavon et al. (2018). In particular, explaining means making something understandable and plain to see. In contrast, interpreting as a process goes beyond explaining and provides a user with the underlying meaning of classification. In that view, the European Union's General Data Protection Regulation[1] (GDPR) focuses primarily on explainability, as it obliges a business using personal data for automated processing to explain how the system arrives at specific decisions. Freitas (2014) uses comprehensibility as another synonym for interpretability, and Lipton (2018) refers to

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

interpretability as an understanding of the model's inner logic. However, none of the definitions mentioned above is restrictive or specific enough to provide formalisation. Therefore, their use depends on the context, author's preferences and expertise (Došilović et al., 2018).

3.3.2 Rationale Behind Building XAI

As the most recent AI solutions based on neural network architectures are becoming more efficient and accurate, their adoption is inevitably spreading across sectors – from private companies to governments. (Ammanath et al., 2020). For many ML models deployed today – in customer segmentation or personalised recommendation services – interpretability may not be a key requirement as long as these systems perform well and effectively meet business needs. In case of some failures – e.g., Netflix recommends a film that one may not enjoy or a search engine does not seem to find a website of one's interest – the consequences are not disastrous. Conversely, for applications where one's safety or monetary interests are in place, the situation may be quite different. For instance, relying unquestionably on a black-box system's decisions in medical (Reddy et al., 2019), banking (Lui & Lamb, 2018), judiciary (Sourdin, 2018), or autonomous driving (Fagnant & Kockelman, 2015) domains may have detrimental repercussions. Understanding the evidence and underlying logic behind each algorithm's decision becomes essential as the output predictions may not be obviously wrong (Goebel et al., 2018).

The transparency issues related to DL gain even greater significance as some black-box systems can make decisions on possibly unethical grounds, e.g. when they accurately predict a person's weight, health (Kocabey et al., 2017) and sexual orientation (Wang & Kosinski, 2018) based on social media images, ethnicity and intelligence by likes on Facebook (Kosinski et al., 2013), score probability of committing a crime or quitting a job (Zhao et al., 2018). Hence, the lack of transparency in "how" and "why" such algorithms reach decisions may be a limiting or even disqualifying factor for their further use and adoption.

A. XAI for Trust and Confidence.

There is a unanimous consensus within the academic community that the ability to verify an AI system's decisions is an essential prerequisite for fostering trust and

confidence among its users (Hengstler et al., 2016). This assumption is valid in situations where some AI system performs a supportive role (e.g., virtual assistants such as Apple's Siri or automated grammar checker Grammarly) and situations where it takes decisions without a human in the loop (e.g., autonomous driving). In the first case, explanations enable the user to compare different alternatives by describing them in detail and providing justification behind the decision. The primary purpose of providing such explanations is to give the user confidence in the decisions of the system. Notably, the users may not be interested in opening the black box and learning its inner workings; instead, they may need to know why it is a reasonable decision (Pieters, 2011). In the second case, where the user partly or entirely relinquishes control, explanations may help comprehend how and in which scenarios a system reaches a specific decision. For instance, if the explanations confirm that some DNN accurately mimics human logic – i.e., it is commonly accurate whenever humans are accurate – then it can engender trust due to the absence of risks related to relinquishing control. Conversely, if the explanations reveal that a model fails to perform as expected for inputs that humans classify accurately, its adoption or autonomous use should be questioned (Lipton, 2018).

Besides, the actual need in providing explanations may also have a social dimension. Heath & Bryant (2013) suggest that human interactions heavily rely on understanding the rationale behind our decisions. Moreover, trust is an essential prerequisite for overcoming human perceptions of risk and uncertainty to accept novel technologies (Gefen et al., 2003). Thus, making AI systems transparent and interpretable via XAI tools may help gain trust for their further adoption and use. Explanations and extra information provided along with a model's decisions can also ensure and maintain two-way interactions that are crucial for building up trust among its end-users in a gradual manner (Li et al., 2008). Based on the finding of Rempel et al. (1985), Lee & See (2004) suggest that trust stemming from an understanding of the motives of an autonomous system will be less fragile than trust built on the reliability of the system's performance. The scholars also conclude that designing interfaces that provide operators (i.e., users) with information regarding the purpose, process, and performance of automation could help build and maintain trust in the long run.

B. XAI as Legal Prerequisite.

AI is already making its way into every aspect of our life, from boosting our analytical abilities and streamlining business processes to unleashing automation potential in decision-making and production. As ML algorithms take on new roles, they pose numerous questions related to their ethical, social, and economic impact (Mittelstadt et al., 2016). The issues of responsibility gap (Matthias, 2004), opaque determination of liability (Fagnant & Kockelman, 2015; Vladeck, 2014) or concerns about anti-discrimination and fairness (Hajian et al., 2016; Mehrabi et al., 2019; Osoba & Welser, 2017) have sparked heated debate in the academic community. Most of them call for new policies and regulations.

Some of the countries have already introduced new regulations that enshrine a right to receive an explanation for algorithmic decisions. For instance, in April 2016, the member states of the European Union (EU) adopted the GDPR. The Articles 13, 14 and 22 of this Regulation stipulate that, when some automated processing or profiling takes place, a data subject (i.e., a natural person to whom data relates) has a right to receive an explanation of the algorithmic decision and "meaningful information about the logic" (Goodman & Flaxman, 2017). France adopted Loi pour une République numérique[1] (e.g., the Digital Republic Law), which goes beyond the GDPR and provides an explicit framework for explaining decisions made by algorithms. In particular, a data subject can access information about the classification parameters, and where appropriate, their weighting applied to the individual case of the person concerned (Edwards & Veale, 2018). These examples demonstrate that citizens' right to receive an explanation for decisions reached by some AI-powered system implies new design requirements such as human interpretability and transparency. If, as expected, the current trend on a more comprehensive AI regulation persists, there will be a pressing need for almost any algorithm in place to operate within this new legal framework.

C. XAI for Verification and Validation.

Besides social and legal reasons, an explanation can become a useful tool to verify a peculiar algorithmic decision as well as validate an entire AI-empowered system. As argued earlier, understanding the evidence and underlying logic behind each algorithm's

[1] Available at https://www.economie.gouv.fr/republique-numerique

decision becomes essential when testing whether the learned strategy is valid and generalisable or whether the model reaches its decision due to some spurious correlation in the training data. In psychology, the reliance on such spurious correlations is generally called the Clever Hans phenomenon (Pfungst, 1911). The phenomenon owes its name to a horse named Clever Hans that could supposedly perform arithmetic and thus attracted considerable scientific attention in the 1900s. As it turned out later, strikingly high accuracy – roughly 90% of correct answers – was due to Hans' ability to derive the right answers from the questioner's posture and facial expression. Lapushkin et al. (2016) have recently demonstrated analogous behaviour in state-of-the-art AI systems. Due to one of the XAI tools, they showed how the algorithms learned some spurious correlations in the training data, and similarly to Hans, predicted correct answers based on the wrong reasoning. For instance, the authors proved that the PASCAL Visual Object Classes (VOC) competition's winning method often failed to detect the presence of objects of interest in the image but instead utilised correlations or background details to generate correct classification. In particular, the model recognised trains and boats due to the presence of railroads and water on the bottom of the image. Moreover, the algorithm could recognise a horse by the presence of a copyright tag that was discriminative of this class in the training data. Interestingly, both organisers and participants of the challenge had overlooked these tags in the dataset for many years.

Ribeiro et al. (2016), known for proposing Local Interpretable Model-agnostic Explanations (LIME), used Google's pre-trained Inception neural network (Szegedy et al., 2015) to train on an image dataset wolves and huskies. The researchers show that the DNN distinguished the classes "Wolf" and "Husky" mostly by the presence of snow or light background in the image. Another striking research was carried out by Mordvintsev et al. (2015), who discovered that some neural networks trained to classify different images had quite a bit of the information needed to generate respective visual representations of the learnt classes. One particular experiment showed that the DNN learned to recognise dumbbells along with a muscular arm lifting them. The authors demonstrate that the network failed to learn dumbbells as an independent concept and could consequently underperform in real-life scenarios. Overall, these cases support the view that XAI helps detect implicit biases in a model or data. Moreover, they illustrate that explanations provided for a single input image can also reveal the classifier's misbehaviour (e.g., an extreme focus on background details such as snow or a

watermark). Thus, understanding a model's inner workings and logic help validate that a given product fulfils all the goals and expectations. Additionally, it is a powerful tool to identify some room for the algorithm's improvement.

Apart from ensuring that some AI system is the "right" product, the model's verification and validation via XAI tools can also lead to new discoveries. As argued earlier, DNN can considerably exceed human performance in discovering patterns and inter-dependencies. Therefore, explaining and interpreting what features an AI system uses for classification, can be more valuable than the classification itself, because it may reveal information or scientific insights about the previously unnoticed phenomena or dependencies (Samek & Müller, 2019).

3.3.3 Scope of Interpretability

Interpretability can imply understanding an automated agent and its logic of making decisions. However, the scope of a model's interpretability can vary. It can either refer to providing explanations of the entire model behaviour or to understanding a single prediction. In the academic literature, both levels of interpretability have been carefully studied and covered. Thus, scholars tend to distinguish between two subclasses: global interpretability and local interpretability (Adadi & Berrada, 2018; Das & Rad, 2020; Carvalho et al., 2019; Rai, 2020).

A. Local Interpretability

Explaining the reasons for a specific decision implies a major focus on a single input and explicit understanding of a model's reasoning with regards to its particular prediction. For instance, in the context of FER in web-based learning environments, an educator may need a thorough understanding of why a model predicted that certain students experienced fear or disgust at the particular moment of the lecture. Interpretability of individual decisions made by a classifier can be achieved due to local explanation methods. Sometimes they are also known as attribution methods, as they produce an explanation by assigning a scalar attribution value to each input feature of a network (Ancona et al., 2017). Generally, local explanation methods identify what dimensions of a single input data instance have most contributed to a specific DNN's output. Technically, these methods generate an explanation $\boldsymbol{g}$ for the decisions $\boldsymbol{y}$ made by a model $\boldsymbol{f}$ based on a

single input instance $\boldsymbol{x}$. A schematic diagram displaying a high-level design of local explanation methods is illustrated in Figure 3.

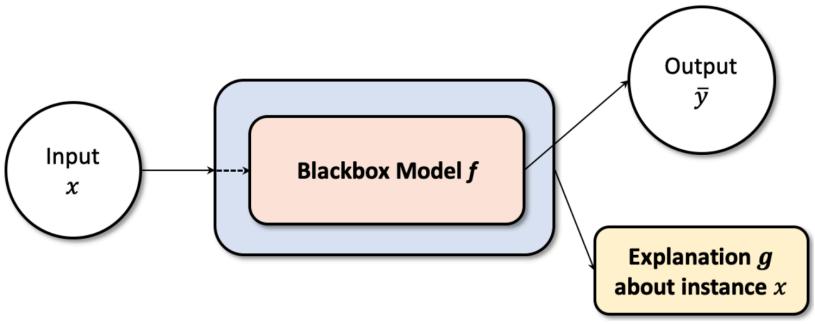

Figure 3. A high-level design of local explanation methods (Das & Rad, 2020).

Das & Rad (2020) conclude that different local explanation methods can expose feature correlations and importance towards output predictions due to a wide variety of techniques: heatmaps, Bayesian techniques, rule-based methods and feature importance matrices. They discovered that the earliest methods tend to provide output explanations in the form of positive real-valued matrices or vectors. Conversely, more recent local explanation methods advance by incorporating attribution maps, graph-based, and game-theory based models that score features based on their positive or negative contribution to the final classification. Specifically, an attributed score or a colour on a heatmap reflects how the particular feature or a data point increases or decreases the probability for a given classification output.

Activation Maximisation. As argued earlier, most of the dynamic deep neural networks used for FER take advantage of CNN. In a CNN, each convolutional layer can have several learned filters that maximise a given hidden unit activation when a similar template pattern is detected in the input image. First convolutional layer learns the high-level features, so it is easy to project – simple multiplication of the learned weights by input pixels generates highly interpretable visualisation. However, subsequent convolutional layers and their filters strongly rely on the outputs of the previous layers, thus, any summarisation and visualisation become particularly challenging.

In 2009, Erhan et al. were arguably among the first who introduced a robust technique that would identify and visualise feature importance of DL models. The proposed method called Activation Maximisation identifies input patterns of images which maximise a hidden unit activation. The rationale behind their idea was that a pattern to which a hidden

unit responds most actively could be a suitable first-order representation of what the unit is doing. The authors thus define maximising the activation of a unit as an optimisation problem. If θ denotes neural network parameters, $z_{i,j}(\theta, x)$ is the activation of a unit i in a layer j. By assuming fixed parameters θ after training the network, an activation map x^* can be generated as shown in Equation (1).

$$x^* = \arg \max_{\mathbf{x}\ s.t.\ \|\mathbf{x}\|=\rho} z_{i,j}(\theta, x) \tag{1}$$

By performing a gradient ascent in the input space, one finds local minima that can be averaged or selected by their maximisation level to generate an explanation map g. The ultimate goal is to minimise the activation maximisation loss that outputs larger filter activations correlated to specific input patterns. Thus, we can compute and highlight layer-wise feature importance with respect to a particular input instance. The activation maximisation method later inspired other researchers who based on its core idea proposed other powerful techniques for local and global interpretability (e.g., Simonyan et al., 2013).

Sensitivity analysis. Sensitivity analysis was one of the first methods for local interpretability that was built upon an idea of using the gradient information to generate attributions to an input image (Simonyan et al., 2013). This method constructs attributions by taking the absolute value of the partial derivative of the target output S_c with respect to the input features x_o. Thus, importance of the corresponding pixels w from the input image x can be calculated as shown in Equation (2).

$$w = \left| \frac{\partial S_c}{\partial x} \right|_{x_0} \tag{2}$$

Intuitively, the gradient's absolute value indicates those input features (i.e., pixels) that need to be changed or perturbed the least to affect the classification score for the class c most. The resulting attribution map would supposedly highlight critical regions in the input space. In practice, the sensitivity analysis does produce an attribution map showing a correlation with regions where the object of classification is present (Simonyan et al., 2013). However, taking the absolute value in the core of the method discards the "direction of the change". In other words, the absolute value does not show how specific pixels contribute to a given prediction score – either positively or negatively. Moreover,

saliency maps are usually rather noisy and, thus, may bring little clarity on the role of specific input regions (Smilkov et al., 2017). Figure 4 is a good illustration of the explanation quality that saliency maps provide.

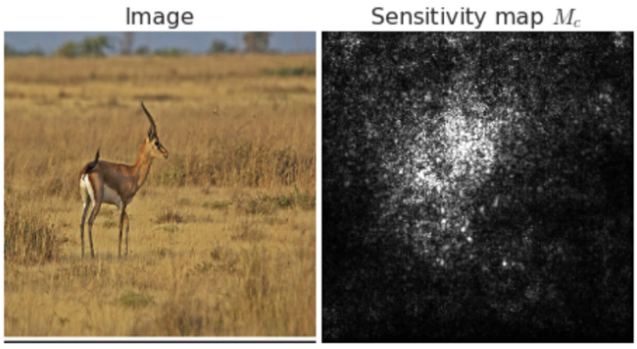

Figure 4. A saliency map generated for the class "gazelle" (Smilkov et al., 2017).

*Gradient * Input*. This method was initially proposed in attempt to improve the sharpness of saliency maps (Shrikumar, 2016). The attribution map R_0^c is constructed by taking the partial derivatives of the output S_c with respect to the input features x_o and multiplying them feature-wise by the input values as seen in Equation (3).

$$R_0^c(x) = \frac{\partial S_c(x)}{\partial x_0} \odot x_0 \tag{3}$$

The intuition for this approach was replicated from linear models – the gradients are regarded as the coefficients of each input feature (e.g., pixel), and the product of the input with a coefficient constitute the total contribution of the input feature to the model's output. Both *Sensitivity analysis* and the *Gradient∗Input* method have obvious shortcomings, as the partial derivative $\partial S_c\,(x)/\partial x_0$ varies not only with x_0 but also with the value of other input features (Ancona et al., 2017). Moreover, visual explanations produced by these two methods can be extremely noisy as DNNs do not filter out irrelevant input features during forward propagation (Kim et al., 2019). When irrelevant input features have positive pre-activation values and consequently pass through an activation function (e.g., the Rectified Linear Unit), they result in nonzero gradients at unimportant regions. Figure 5 clearly shows how both methods tend to highlight obviously irrelevant features for the class predicted.

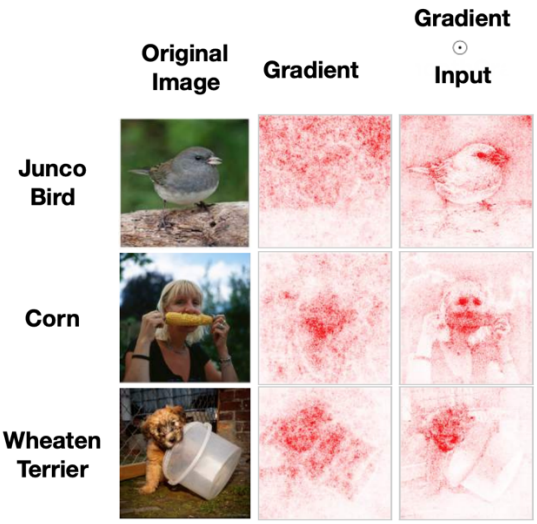

Figure 5. Explanations generated by Sensitivity analysis (i.e., Gradient) and Gradient∗Input method (Adebayo et al., 2018).

Local Interpretable Model-Agnostic Explanations (LIME). In 2016, Ribeiro et al. proposed a novel explanation technique that would explain the predictions of any regressor or classifier in an interpretable and faithful manner, by approximating it locally with a linear regression model. The method has two important characteristics: it is *model-agnostic* and *locally faithful*. The former stands for the capability of the method to explain any black-box model without knowing its inner mechanism of making predictions. The latter, local fidelity, corresponds to generating explanations for the rationale behind an individual prediction based on the understanding of how the model behaves in the vicinity of the instance being predicted. However, the resulting local surrogate model does not imply global fidelity – i.e., features that are locally important may not be relevant in the global context, and vice versa.

Mathematically, the authors define an explanation as a model $g \in G$, where G is a class of potentially interpretable models (e.g. linear regression model), so that g can be readily demonstrated to the user with its textual or/and visual artefacts. A model g seeks to minimise loss L (e.g. mean squared error), which measures how faithful the explanation is to the prediction of the original model f with a proximity measure π_{x_0}. Moreover, it should also minimise its explanation complexity $\Omega(g)$ – selecting the minimum number of the most informative features for a given classification. For instance, for decision trees $\Omega(g)$ may equal to the depth of the tree, whereas for linear regression models, $\Omega(g)$ may equal to the number of non-zero coefficients.

$$\text{explanation}(x_0) = \arg\min_{g \in G} L(f, g, \pi_{x_0}) + \Omega(g) \quad (4)$$

Essentially, in order to learn the local behaviour of f, the LIME algorithm creates a new dataset by sampling instances around an observation x_0. The vicinity of the sampled instances is weighted by π_{x_0}. Importantly, numerical features are selected based on the distribution and categorical variables are picked according to their occurrence. Next, the algorithm feeds new data into original model f and then opts for the defined number of the most informative features based on $\Omega(g)$. Finally, it fits a simple model (e.g., linear regression model) to the permuted data with $\Omega(g)$ features and similarity scores as weights. The weights and selected arguments of the linear model serve as an explanation of a decision x_0. Figure 6 illustrates an explanation generated by the LIME algorithm on a single instance.

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 6. Explanation for image classification made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$) (Ribeiro et al., 2016).

Layer-wise Relevance Propagation (LRP). The method was proposed in 2015 as an explanation technique that operates by propagating the prediction score $f(x)$ for a given input x backwards through the DNN's layers by following specially devised local propagation rules (Bach et al., 2015). Fundamentally, the method utilises the idea of tracing back the individual contributions of input nodes (e.g., pixels) to the final classification. The rule for propagating relevance score $(R_k)_k$ across DNN's layers is presented in Equation (5):

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k \quad (5)$$

In Equation (5), j and k denote neurons at two consecutive layers, and z_{jk} equals to the activation a_j of the neuron j multiplied by the weight w_{jk} between these neurons. In

essence, the variable z_{jk} quantifies the extent to which neuron j has contributed to the relevance of neuron k. The denominator in Equation (5) serves to ensure the conservation property – some value received by a neuron must be propagated backwards in equal amount to the lower level. The propagation terminates once it reaches the input layer. When applying the procedure mentioned above, one can verify the layer-wise conservation property $\sum_i R_j = \sum_k R_k$, and, consequently, the global conservation property $\sum_i R_i = f(x)$. The schematic illustration of the LRP method can be seen in Figure 7.

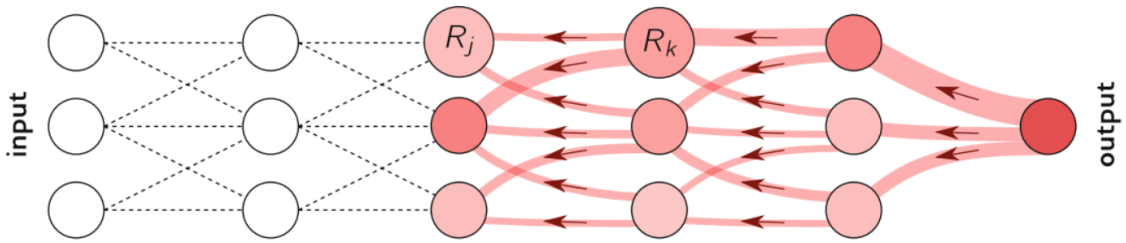

Figure 7. The Layer-wise Relevance Propagation procedure in a DNN (Montavon et al., 2019).

In the original paper, Bach et al. also present some enhancements for the generic LRP procedure. For example, the authors propose the so-called LRP-ε, which modifies the original formula by adding a small positive term ε to the denominator as depicted in Equation (6). Adding ε may prevent $R_{j\leftarrow k}$ from taking unbounded values when the contributions to the activation of neuron k are contradictory or minuscule. As ε grows with each layer, only the most informative nodes survive the "absorption". This simple modification typically results in fewer input features highlighted and, therefore, less noisy explanations.

$$R_j = \sum_k \frac{z_{jk}}{\varepsilon + \sum_j z_{jk}} R_k \tag{6}$$

Another modification called LRP-γ tends to favour positive contributions of specific input features over negative ones (Montavon et al., 2019). The parameter γ defines the extent by how much positive contributions are favoured. As the value of γ grows, negative contributions become less significant and evident. This technique helps provide more stable explanations. Equation (7) stands for mathematical implementation of LRP-γ.

$$R_j = \sum_k \frac{a_j \cdot (w_{jk} + \gamma w_{jk}^+)}{\sum_j a_j \cdot (w_{jk} + \gamma w_{jk}^+)} R_k \tag{7}$$

However, Montavon et al. (2019) suggest using Composite LRP, which means selecting and applying a specific version of the LRP method depending on layers' location in the neural network's structure. For example, the authors argue that generic LRP should be applied to upper layers, LRP-ε to middle layers, and LRP-γ to lower ones.

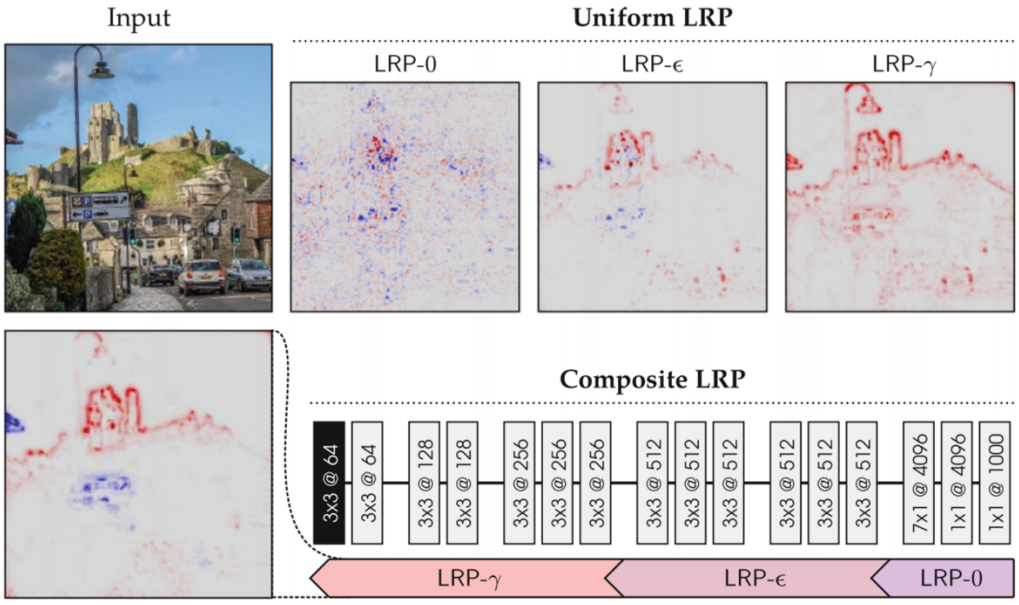

Figure 8. Pixel-wise explanations generated by various LRP procedures for the output class "castle". Red-coloured pixels highlight input features that positively contributed to the current prediction, whereas blue ones decreased the prediction score for this class (Montavon et al., 2019).

SHapley Additive exPlanations (SHAP). In a similar way to LIME or LRP, this local explanation method proposed by Lundberg & Lee (2016) seeks to explain an algorithmic decision by computing the contribution of each input feature of some instance x to the final prediction. The core idea of SHAP is based on approximating Sharpley values introduced and conceptualised by Lloyd Shapley (1953) as an extension of the cooperative game theory. Shapley values express the contribution that a single feature or a group of features has on the output of a model in the presence of multicollinearity. Classic approach of calculating Sharpley values for a linear regression requires retraining the model on all possible feature subsets $S \subseteq F$, where F is the set of all input features. To compute and attribute an importance value to each predictor (i.e., input feature), two models $f_{S\cup\{i\}}$ and f_S are trained with and without the feature respectively. Next, predicted

outputs from the two models are compared on the current input $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$, where x_S stands for the values of the predictors in the set S. As the input feature's withdrawal affects the other features in the model, the preceding differences are calculated for all possible subsets $S \subseteq F \setminus \{i\}$. The computed Shapley values are then used as feature attributions. In essence, they equal to a weighted average of all possible differences as shown in Equation 8.

$$\varphi_j = \sum_{S \subseteq F\setminus\{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f_{S\cup\{i\}}\,(x_{S\cup\{i\}}\,) - f_S(x_S)] \tag{8}$$

However, the total number of input features in DNNs can often amount to thousands; therefore, computing Shapley values for each feature is computationally expensive. As an alternative, the authors suggest approximating Shapley values either with Shapley sampling values (Strumbelj & Kononenko, 2010; Štrumbelj & Kononenko, 2014) or with another approximation method proposed in the paper – Kernel SHAP. Two years later, the authors presented another highly effective implementation of SHAP – the Tree SHAP algorithm – which is specific to tree-based ML models such as decision trees, gradient boosted trees, and random forests (Lundberg et al., 2018).

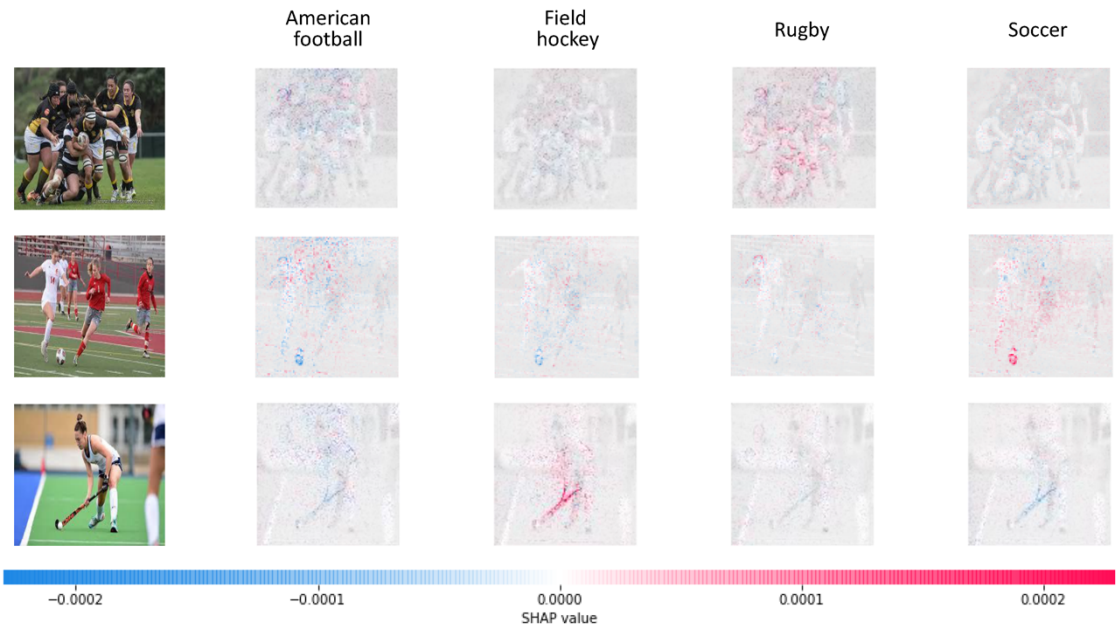

Figure 9. Example of SHAP explanations based on the image classification performed by CNN-TL-DE (Podgorelec et al., 2020)[1].

Although SHAP and LIME provide locally accurate and consistent explanations, they have several significant disadvantages. The quality and complexity of the generated explanations in both methods may heavily rely on the number of input features defined as a hyperparameter. This problem becomes particularly acute when working with image data. A larger number of input features (i.e., superpixels) is likely to result in a more explicit representation of an image's most important areas. However, each new input feature, in turn, severely increases computation time. Moreover, LIME and SHAP are post-hoc explanations techniques that rely on input perturbations. Therefore, they may be susceptible to adversary attacks in which it is possible to generate a biased classifier whose post hoc explanations can be arbitrarily controlled and mask biases (Slack et al., 2020).

B. Global Interpretability

[1] Red pixels indicate areas of the image with a positive contribution towards specific classification output (e.g., "Soccer" or "Rugby"), whereas blue pixels highlight areas of the image that do the opposite. For instance, a hockey stick is apparently the most crucial feature for the third image to be classified as "Field hockey", while the presence of a soccer ball on the second image decreases the probability for classes "Field hockey" and "Field hockey".

Global interpretability refers to understanding the generalised reasoning of a model and its view of the data features. Essentially, this level of interpretability aims to estimate the global impact of input features on the model's decisions in the form of parameters, weights, or rules. One of the most vivid examples of the globally interpretable models is linear regression. Its coefficients are estimates of some unknown population parameters that describe the magnitude of changes in the model's output as a result of one unit of change in its input variables (Neter et al., 1989). Although linear regression models might fail to encode non-linear dependencies, their regression coefficients may still shed some light on global effects or trends in different domains such as sociology, psychology, medicine, and other quantitative research fields. Decision trees (Hastie et al., 2009) or decision rules (Fürnkranz et al., 2012) are other examples of globally interpretable ML models. If-then rules and cut-off point inherent to their architecture enable tracing back the output decision back to the input (i.e., root node). However, as argued earlier, models built with a multi-layered neural network surpass linear regression and decision tree models' performance in most of the cases. Therefore, the need for specific methods explaining the global rationale of black-box models is bound to keep growing.

Generally, methods for global interpretability work on groups of inputs to approximate the overall behaviour of the black-box model as illustrated in Figure 10. In particular, the explanation g describes the aggregated feature attributions of the model f on some set of inputs $\{x_1, x_2, \ldots, x_n\}$. Therefore, methods for global explanations may become particularly helpful in forging a better understanding of the model's behaviour that is tested on a large variety of inputs and previously unseen data. As far as FER models are concerned, global explanations may bring more clarity about the role of the main features and distinct areas of the face in predicting a specific class (i.e., emotion).

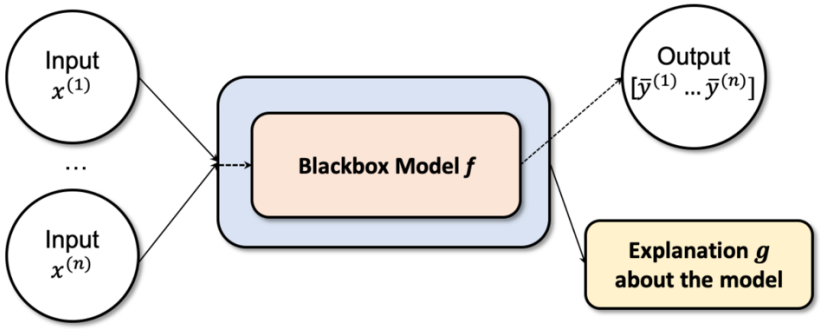

Figure 10. A high-level design of methods for global interpretability (Das & Rad, 2020).

Guidotti et al. (2018) made a comprehensive review of the methods for global interpretability. Unlike the local interpretability methods, the variety of techniques for generating global explanations for the black-box models dealing with image or video data is extremely limited. The authors showed that most of the existing methods attempt to approximate the black box with another globally explainable model (e.g., linear regression, decision tree). These models are often called *global surrogate models*. However, this approach may appear inept at mimicking the entire reasoning of the image or video classification models due to their complexity. Moreover, Rudin (2019) argues that global surrogate models cannot have perfect fidelity with respect to the original black-box model. If their explanations were utterly faithful to what the original model computes, the surrogate model would equal the original one, and one would not need to utilise the original model in the first place. Thus, the author suggests that any explanation method approximating the inner logic of a black-box model can be its inaccurate representation. For instance, a global surrogate model that computes the same output as the original model in 90% of the cases indeed explains the original model most of the time. However, if such a model is correct 90% of the time, it must be wrong 10% of the time. If a hundred out of a thousand hypothetical patients receive a false explanation for their diagnosis, one cannot trust the explanations, and thus one is unlikely to trust the original black box. Therefore, all the global interpretability methods covered here are based on the aggregation of the local explanations generated with the inner workings of the original models.

Class Model Visualisation. Apart from image-specific class saliency visualisation introduced by Simonyan et al. (2013), in the same paper the authors also proposed the method called class model visualisation. Given a learnt classification DNN and a class of interest, the visualisation method is built upon the numeric generation of an image (Erhan et al., 2009). The resulting image appears to be discriminative enough to represent the entire class in terms of the DNN' class scoring.

In Equation 9, $S_c(I)$ denotes the score of the class c, computed at the classification layer of the selected DNN for an input image I. We seek to reconstruct an image with L_2-regularisation λ, such that the score S_c is high:

$$I' = \arg\max_I S_c(I) - \lambda \|I\|_2^2 \tag{9}$$

The procedure is closely connected to the training procedure of the DNN, where the back-propagation algorithm adjusts the model's parameters (i.e., weights). However, in contrast to the classic training procedure, the optimisation is performed with respect to the input image, while the weights determined during the training stage remain constant.

Figure 11. Visualisations of various class appearance models learnt by a multi-layered CNN (Simonyan et al., 2013).

In the paper, the authors exemplify the proposed method with a deep CNN trained on the large-scale ImageNet dataset for image classification (Deng et al., 2012). They initialised the optimisation with the zero image, as the CNN was trained on the pre-processed data with mean centring, and then added a mean image of the training set to receive the final result. Visualisations generated with the class model visualisation method for three different classes are shown in Figure 11.

Spectral Relevance Analysis (SpRAy). Lapuschkin et al. (2019) proposed this method as a technique for efficient investigation of classifier behaviour on large-scale datasets. At its core, SpRAy takes advantage of spectral clustering (Von Luxburg, 2007) and applies it to a dataset of local explanations generated by the LRP algorithm. Due to clustering, the method may identify typical as well as anomalous decision behaviours of the ML model and present the results in a concise and interpretable manner to its end-user. Technically, SpRAy detects various prediction strategies of the classifier based on frequently reoccurring patterns in the heatmaps (e.g., specific input features). As the authors suggest, the identified groups (i.e., clusters) of image features may be meaningful representatives of a particular class. Conversely, these clusters may also be some groupings of co-occurring features learnt by the model but not intended to be discriminative properties of the class and ultimately of the model's reasoning.

The workflow of SpRAy consists of four steps:

1. We compute the relevance maps for some sample of the class objects we want to explain. Theoretically, the relevance maps generated by the LRP algorithm will highlight the most critical input features for the model's classifier.
2. We reduce the dimensionality of the relevance maps and make them uniform in shape and size. The dimensionality reduction makes the subsequent analysis more computationally efficient and statistically tractable.
3. We perform a spectral cluster analysis on the relevance maps. At this stage, we structure the distribution of relevance maps by grouping various classifier behaviours into a finite number of discriminative clusters.
4. We identify potentially interesting clusters with eigengap analysis. The spectral clustering algorithm encodes the cluster structure of the relevance maps with the eigenvalue spectrum. Considerable differences between two successive eigenvalues (i.e., eigengap) are likely to indicate well-separated clusters, including anomalous classification strategies. Therefore, such clusters will require further inspection from the user.

Additionally, the authors suggest visualising the analysis results with t-Stochastic Neighbourhood Embedding (t-SNE). Although this last step is not a part of the proposed method, it may bring more clarity on how SpRAy works. The workflow of the SpRAy analysis is illustrated in Figure 12.

Moreover, in the very paper proposing SpRAy, the authors practically demonstrated how this method helped detect a previously unknown bias in the prediction behaviour of the DNNs trained on the Pascal VOC database (Everingham et al., 2015). In particular, the researchers showed how the analysed multi-label classifier learned that uniform pixel values next to the image borders are discriminative for object class "aeroplane". In fact, the classifier "expected" that any image depicting an aeroplane should have some areas of uniform colour – i.e., sky – at its top and bottom. As Lapuschkin et al. argue, using that model as a predictor for aeroplanes outside the laboratory is likely to result in high rates of false-positive predictions if an input image captures the sky. Thus, the researchers verified the efficacy of SpRAy as an effective method for pinpointing implicit patterns in the model's prediction behaviour.

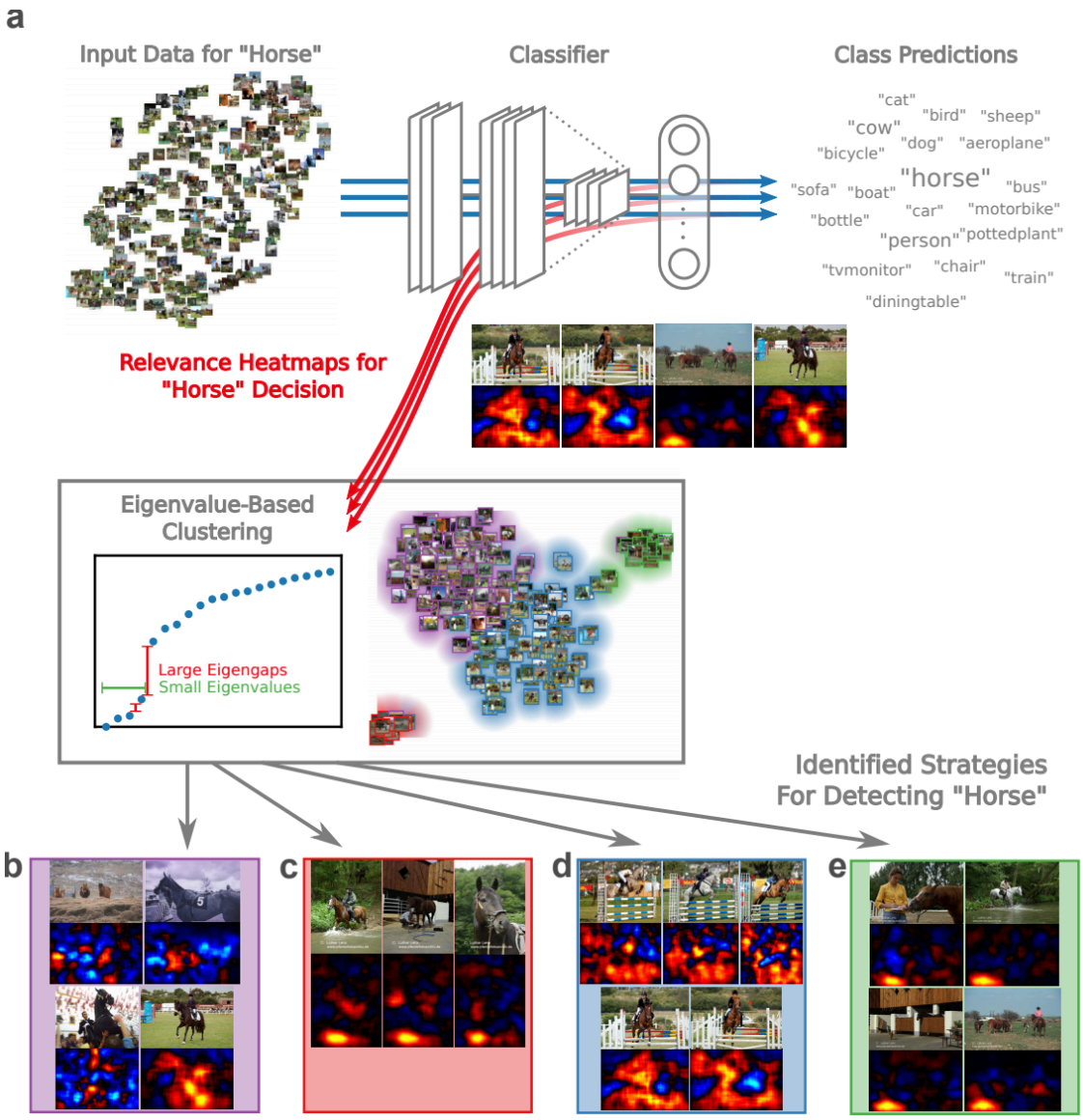

Figure 12. The workflow of SpRAy (a) and four different classification strategies for the class "horse": (b) images with a horse (and rider), (c) portrait-oriented images with a source tag, (d) images depicting wooden hurdles and other elements of horseback riding, and (e) landscape-oriented images with a source tag (Lapuschkin et al, 2019).

3.4 Evaluation Criteria for FER Models

Section 3.4 investigates scientific evidence on the evaluation criteria currently used to evaluate different FER algorithms. Essentially, this Section aims to give an answer to the fourth research sub-question (SQ4). The scientific literature analysed in Section 3.4 provides an overview of the well-established evaluation metrics and criteria used for the variety of FER approaches and algorithms discovered in Section 3.2. The identified criteria form the backbone of the proposed evaluation framework balancing some conventional measures with those inferred in Sections 3.1, 3.2 and 3.3 (see Section 4).

Given different FER approaches and a wide variety of pre-trained models, some objective evaluation criteria are essential to enable unbiased comparisons and robust assessment. Without careful scrutiny of the relative performance of competing algorithms, one may not get the clear picture of their strengths and weaknesses that are essential for making further decisions regarding their deployment. As any IT artefact, FER models can be evaluated in terms of their functionality, completeness, consistency, accuracy, performance, reliability, usability, and other relevant attributes (Hevner et al., 2004). The ISO/IEC 25023:2016 also includes security, portability and maintainability as essential software quality dimensions (ISO/IEC, 2015).

Generally, most of the research papers on FER algorithms focus primarily on the limited number of evaluation criteria such as predictive accuracy or performance (e.g., Revina & Emmanuel, 2018; Saxena et al., 2020). The most likely causes of such a strong emphasis on these particular model's properties can be their unambiguous definition, relatively simple computation and easy comparability (Giraud-Carrier, 1998). Moreover, evaluating security, reliability, usability, or other properties of the FER models mentioned above may not be feasible or reasonable. Principally, in all the reviewed papers, the researchers present some novel technology or method for FER with a primary goal to demonstrate its state-of-the-art accuracy or/and performance on specific datasets. Such a strong emphasis on performance metrics seems logical, given that most academic publications on FER may aim to disseminate and test new scientific ideas rather than propose a fully customised market product.

Moreover, as argued in section 3.3.2, attempts to make highly complex DNNs explainable and comprehensible have gained growing scientific and public interest. Besides multiple examples of XAI tools helping verify, justify and improve black-box models, interpretability has also become a legal prerequisite for the deployment of such models in certain cases. Thus, evaluating the inner logic of an algorithm is nothing but a legal and societal demand for trust and fairness in algorithmic decision-making. Moreover, introducing interpretability as an evaluation criterion may significantly contribute to a more robust assessment of models with no significant difference in predictive accuracy.

Therefore, this section provides an overview of the most popular metrics and criteria used for a FER model evaluation within the scientific community. These evaluation metrics and criteria are grouped into two major clusters: performance-related and interpretability-

related metrics. This particular grouping is chosen according to the model's dimension these metrics aim to evaluate.

3.4.1 Performance-related Metrics

Choosing the right metrics may help address evaluation bias that ultimately arises because of a need to compare different models against each other objectively (Suresh & Guttag, 2019). Arbitrary preference for the particular metrics to report performance may significantly exacerbate this bias. For instance, Suresh et al. (2018) demonstrate how aggregate measures can hide subgroup underperformance. Moreover, considering a single type of metric (e.g., accuracy) can also mask disparities in other types of errors (e.g., false-positive rate). Buolamwini & Gebru (2018) demonstrated how three commercial facial analysis algorithms with error rates less than 1% for lighter male faces under-performed on darker female faces (20.8% − 34.7% error rate). The researchers aimed to emphasise the need for rigorous reporting on the performance metrics to ensure algorithmic fairness and explainability.

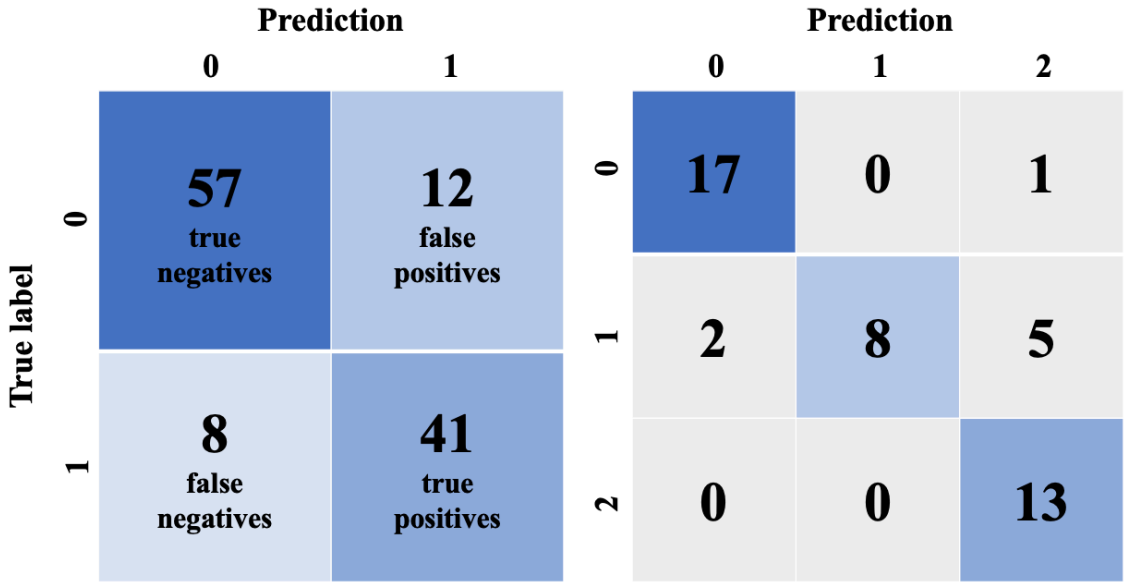

Figure 13. Example of the confusion matrices for the case of binary classification (on the left) and multi-class classification (on the right).

Generally, any classification model can generate four different types of outcomes. Grouping the model's outcomes into distinctive groups helps calculate different performance-related metrics. These four groups are *true positives*, *true negatives*, *false positives*, and *false negatives*. *True positive* is a classification outcome when the model correctly identifies the actual class of the input. For instance, a true positive for any FER

model is an input image of a smiling person classified as "happy". Conversely, *true negatives* occur when the model correctly classifies the negative class – this type of classification outcome is mainly used in binary classification. A *false positive* is an outcome when the model classifies some input instance with a class to which it does not belong. For example, the model assigns class "angry" to an input image of a smiling person. Finally, *a false negative* is a classification outcome when the model classifies an input with some other class. These four types of outcomes are often depicted on a confusion matrix. Figure 13 illustrates an example of two confusion matrices for the case of binary classification and multi-class classification respectively.

Classification Accuracy. This evaluation metric is a common measure of statistical bias quantifying how often some trained classifier is correct. Mathematically, it equals the ratio of all correctly predicted classes to the total number of all predictions (Equation 10).

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{10}$$

This evaluation metric can be reliable only if there is an equal number of samples belonging to each class. In the case of an imbalanced dataset (i.e., unequal distribution of instances belonging to different classes), a misclassified instance of the minor class is unlikely to affect accuracy significantly.

Precision. The precision for a class is the fraction of true positives among all input instances classified as belonging to the positive class (i.e. the sum of true positives and false positives). This evaluation metric is also known as positive predictive value. High precision means that a classification algorithm primarily generates a correct label for a given class. For instance, if some FER model classifies five separate images of a smiling person as "happy" and one image of a crying person as "happy" too, the precision for class "happy" would equal roughly 83%. This value would signify that the model is correct 83% of the time when it predicts happy emotion.

$$Precision = \frac{True\ positives}{True\ positives\ +\ False\ positives} \tag{11}$$

Recall. The recall is the measure quantifying the proportion of true positives that some model correctly identified. For instance, if our FER model has a recall value of 75% for the class "surprised", it identifies 75% of all "surprised" learners. This metric is also known as *true positive rate* or *sensitivity* (Yerushalmy, 1947).

$$Recall = \frac{True\ positives}{True\ positives + False\ negatives} \tag{12}$$

Measuring both precision and recall is crucial, as the former ensures that the model does not overlook instances of a specific class, while the latter shows whether the model has an excessive misclassification. In some domains, these two metrics may be especially significant. For instance, Walsh et al. (2017) proposed an algorithm that demonstrated state-of-the-art accuracy in predicting future suicide attempts. In the paper, the researchers emphasised the importance of the precision and recall values for their model. Moreover, achieving the highest possible value for the latter was particularly critical, as it implied minimising the probability of missing someone who was really considering suicide.

F-score. There can be many situations when both precision and recall are equally important. In such cases, one can use another evaluation metric called the F-score (i.e., the F-measure, the F_1-score). To calculate the F-score, one needs to take the harmonic mean of precision and recall (Equation 13). Its highest possible value is 1, meaning perfect precision and recall, and the lowest possible value is 0 if either the precision or the recall equals zero.

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{13}$$

Another version of the F-score, F_β, may be more useful if one needs to treat recall and precision differently. The F_β uses a real positive factor β, where β can be chosen such that recall is considered β times as significant as precision (Equation 14). Two common values for β are 0.5, which weighs the value of recall lower than precision, and 2, which weighs recall higher than precision.

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall} \quad (14)$$

Although the F-score may be a more informative and rigorous measure of the algorithm's classification performance than accuracy, this metric may be still misleading when testing on datasets with the imbalanced class distribution. Jeni et al. (2013) maintain that any performance metric such as the F-score may reveal more about skew (i.e., imbalance) than about actual algorithmic performance. The researchers also claimed that FER databases that are identical with respect to the action unit intensity or a head pose might yield very different metric values due to differences in class distribution. To minimise biased estimates of performance metrics, Jeni et al. suggest measuring performance both with original and skew-normalised (i.e., balanced) data. Due to this data manipulation, classifiers may become more comparable across different databases.

Generally, Ponce et al. (2006) argue that any performance-related metrics should be only a means, not an end in itself. In particular, the researchers indicated that the scientific community tends to be overly obsessed with maximising the value of a specific metric, forgetting that any performance increase – e.g., in accuracy – by itself is not necessarily a sign of progress toward better generalisation or recognition in the wild. Torralba & Efros (2011) demonstrated that instead of using a particular performance metric as an absolute measure, one might better evaluate classifier's generalisation abilities by calculating its average performance drop from testing on unseen data. Li & Deng (2020) applied a similar approach in the context of FER. However, instead of measuring performance drop, the researchers chose the average cross-dataset performance as a metric to evaluate algorithmic generalisation. The researchers observed a significant drop in performance when testing two pre-trained DL models on different datasets. Specifically, the average in-dataset performance of algorithms trained on different datasets was 63.83%, and it dropped to 39.40% for the average cross-dataset performance.

Liao et al. (2020) identified that understanding the performance of an algorithm may also contribute to the interpretability of its predictions. In particular, based on performance evaluation of some system, one can tell how often it makes mistakes, what kinds of misclassification occur more often, and in which scenarios the system is likely to be incorrect. The answers to these questions can be critical in discovering potential limitations and underlying logic of the model's decision-making.

3.4.2 Interpretability-related Criteria

As argued in Section 3.3.1, because of the subjective and multidimensional nature of interpretability, researchers can agree neither upon its definition nor its measure. This study groups different interpretability-related criteria according to the taxonomy described by Carvalho et al. (2019). In particular, the researchers grouped different methods of achieving interpretability of a model in three major categories: *pre-model*, *in-model* (i.e., intrinsic), and *post-model* (i.e., post-hoc) *interpretability*. Pre-model interpretability is closely related to a comprehensive exploratory analysis of training data. Generally, it aims to summarise major characteristics of a dataset and reveal potential data biases that can eventually affect the model's performance (Tukey, 1977). Intrinsic interpretability refers to models that are inherently globally interpretable (e.g., linear regression). Finally, post-hoc interpretability refers to explanation methods that are applied to the trained model and its prediction outcomes (e.g., local and global interpretability methods). As this study focuses solely on DL algorithms for FER that are not intrinsically interpretable, it suggests grouping the evaluation criteria for the model's interpretability into two clusters: *data-related interpretability criteria* and *post-hoc interpretability criteria*.

A. *Data-related Interpretability Criteria*

The foundation of any classification model is the data chosen to train it. When selected and used arbitrarily, the data can entail undesirable risks and consequences, especially when regarded as a basis for decision-making. Unfortunately, in most research papers on FER, extensive analysis of the training data and its potentially inherent biases seems to be the exception rather than the rule. Even though some research papers might seek to evaluate interpretability of the proposed FER methods, they tend to disregard their data-related interpretability and focus solely on the post-hoc interpretability.

Generally, being aware of the training and test data specifications can be crucial for understanding the limits in any learning method's generalisation abilities (Ponce et al., 2006; Torralba & Efros, 2011). The limits affecting the model' generalisation are often called data biases. Mitchell et al. (2018) decompose the notion of data-related biases into more precise notions of *statistical bias* – i.e. issues regarding non-representative sampling and measurement error – and *societal bias* – i.e. problems with definitions of the social phenomena measured by and represented in the data (see Figure 14). Suresh & Guttag

(2019) extended the notion of data-related biases with *aggregation bias* – i.e., generalising the patterns of a specific subgroup for the entire population or some other weakly related subgroup. Torralba & Efros (2011) also identified *capture bias* – i.e., the way how image or video data are recorded.

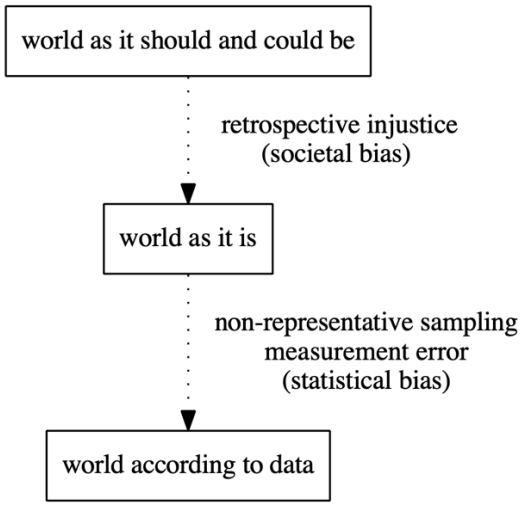

Figure 14. Two components of "biased" data: societal bias and statistical bias (Mitchell et al., 2018).

Sampling Bias. This bias represents a systematic mismatch between the sample used to train and validate the model and the real world as it currently is. Specifically, this bias occurs when some selected data set is not representative of the entire population to which the trained model is applied. Interestingly, Suresh & Guttag (2019) argue that representation bias (i.e., sampling bias) can arise for several reasons. For one thing, the sampling methods can only reach a portion of the target population. For instance, datasets collected from social media can under-represent lower-income or older groups, who are less likely to use social platforms. For another thing, the target population has changed or is distinct from the population used during model training. Data that was representative of students at Harvard 20 years ago will probably not reflect today's population.

Measurement Error. This type of bias occurs when the classification error varies between different population groups (Mitchell et al., 2018). Suresh & Guttag (2019) maintain that this error occurs when data quality varies across groups. Alternatively, they argue that the defined classification task may be an oversimplification – choosing a biased proxy label that only captures a particular aspect of what one aims to measure. A vivid example of the measurement error is the study carried out by Grother et al. (2019), who quantified the accuracy of more than 180 commercial face recognition algorithms for different

demographic groups defined by age, sex, and race or country of birth. They trained and validated these algorithms on four large image datasets collected in U.S. governmental applications that were in operation as of 2019. Their findings concluded that some algorithms misidentified African American and Asian people up to 100 times more likely than white men. Moreover, according to the study, Native Americans had the highest false-positive rate of all ethnicities. The researchers suggest that varying error rate discovered in some algorithms could depend on the data used to train them.

Societal bias. Sometimes training data can be representative and accurate. However, an algorithm may still record objectionable social structures that contradict the decision-maker's goals (Mitchell et al., 2018). For instance, using arrests as a crime measure can introduce statistical bias from measurement error that is differential by race because of a potentially racist policing system (Lum & Isaac, 2016). But even if one could measure crime with perfect accuracy, it would unlikely make data free from "bias" in a normative sense. Any phenomenon such as crime rates may reflect societal bias, including how crime is generally defined. Similarly, any image or video data capturing facial expressions are biased, as its quality depends on how one understands different emotions. Unfortunately, Mitchell et al. (2018) argue that this bias may not have technical solutions at all.

Aggregation Bias. Aggregation bias occurs when one draws false conclusions for a subgroup based on observations made for a different subgroup. Suresh & Guttag (2019) suggest that underlying aggregation bias assumes that the mapping function from inputs to labels is consistent across various groups. However, this often may not be the case. Aggregation bias can result in a not optimal or accurate model for any group or a model that is tuned to some dominant population. For example, training a FER model on a data set with the facial expression of Asian females may lead to poor performance for African people.

Capture bias. Torralba & Efros (2011) argue that this type of bias is inherent to many datasets with labelled images for object detection and classification. In particular, the researchers observed that photographers tended to take pictures of objects for the inspected datasets in similar ways. This bias can potentially occur in the datasets for FER, as some of them contain photos or videos made in a professional studio. For instance, all the photos from JAFFE (Lyons et al., 1998) are frontal-view, while MMI (Pantic et al.,

2005) includes only image and video data recorded with professional studio lighting. These factors may considerably affect emotion recognition in the wild.

B. *Post-hoc Interpretability Criteria*

As demonstrated in Section 3.3, the interpretability of ML models has sparked paramount academic interest and induced extensive scientific research on its very notion, scope and methods. However, despite remarkable progress, there is still no consensus about defining (as shown in Section 3.3.1), quantifying, or measuring ML models' interpretability (Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017; Carvalho et al., 2019). Its different notions, such as simplicity, fairness, simulatability, transparency, or trustworthiness, may be conflated (Lipton, 2016). This problem is further exacerbated by the fact that ML models may have a wide range of stakeholders with varying needs and goals depending on their roles and expertise (Hohman et al., 2019; Tomsett et al., 2018). For instance, the approach suitable for regulatory bodies auditing a case when some black-box system rejected a loan application may significantly differ from the approach that works best for a team of data scientists debugging the model.

Poursabzi-Sangdeh et al. (2018) argue the lack of consensus around defining or quantifying interpretability mentioned above, as well as insufficient scientific evidence for its benefits, make interpretability hard to directly manipulate or measure. The researchers maintain that interpretability is a latent and fundamentally human property. Their argument stems from the fact that the model's interpretability can depend on different manipulable factors such as the number of input features (Ribeiro et al., 2016), the complexity of the model's explanation (Lage et al., 2019), or the user interface (Weitz et al., 2020). Thus, Poursabzi-Sangdeh et al. conclude that interpretability and its measures must be built upon people's behaviour, not by what appeals to intuition.

A vast majority of studies on XAI focus solely on discovering general principles and practices of tailoring explanations generated by various XAI tools to lay-user needs and requirements (e.g., Arrieta et al., 2020; Doshi-Velez & Kim, 2017; Hall et al., 2019; Kulesza et al., 2013; Miller et al., 2017; Miller, 2019). In particular, they primarily focus on an assessment of how good an explanation is.

For example, Lipton (1990) argues that good human-oriented explanations should be *contrastive*. His line of argument relies on the fact that people are usually relatively

interested in real causes why an algorithm reaches a particular prediction. Instead, the researcher suggests that people may prefer to know why the specific prediction was made instead of another. Moreover, Lipton suggests that people may not be specifically interested in all the factors that led to the classification, but the factors could bring about a different (i.e., contrastive) one if input values are changed. Similarly, Wachter et al. (2018) maintain that counterfactual (i.e., contrastive) explanations are intentionally restricted and crafted in such a way as to provide a minimal amount of information capable of altering an algorithmic decision. The researchers also emphasise that such explanations do not require its end-user to understand any of the internal workings of a model to make use of it. Thus, contrastive or counterfactual explanations may be *incomplete*, nevertheless, they require defining a meaningful reference point.

Miller (2018, 2019) argues that explanations for a black-box model's decisions should be *selective*. His suggestion is very much in line with the arguments of Wachter et al. (2018) and Lipton (1990) about the *restrictive* or *incomplete* nature of human-friendly explanations. Miller suggests that people do not expect explanations to cover the actual and complete list of causes of an event. In contrast, they prefer limiting the entire variety of causes to one or two principal ones and regard them as an actual explanation. Thus, explanation methods should ideally provide selected explanations or, at least, explicitly highlight which of their components are the principal causes for a prediction. Apart from the *selectivity* of algorithmic explanations, Miller also claims that they should be presented in the form of a conversation or any other type of human interaction with high awareness of the end-users' *social* context.

Molnar (2020) maintains that the causes that generally have a small probability but did happen should also be included in the explanation. He exemplifies the research of Kahneman & Tversky (1981), who stated that humans tend to focus more on abnormal causes to explain specific events. Molnar also argues that abnormal causes are a great example of a counterfactual, as they can dramatically affect the model's final decision. Thus, including information on abnormalities that possibly impacted algorithmic prediction can appeal to human counterfactual thinking and make explanations look reasonable to the end-user.

Additionally, Molnar (2020) suggests that user-oriented explanations should be *consistent* with the end user's prior beliefs. His assumption is based on the study of Nickerson

(1998), who argued that humans tend to ignore or underestimate evidence that clashes with their prior hypotheses or beliefs. Doshi-Velez & Kim (2017) also argue that human evaluation is essential to assessing interpretability. Specifically, they proposed a taxonomy for the evaluation of interpretability with three different modes of human involvement – *application-grounded*, *human-grounded* and *functionally-grounded* evaluation. Human-grounded evaluation seems particularly appropriate in the case of FER models, as this type of evaluation was proposed for the situations when double-checking model's predictions with simultaneous real-world experiments may be challenging without experts with domain-specific knowledge. As Doshi-Velez & Kim propose, one can present an explanation, an input, and an output to the end-users and ask them to indicate what must be changed to make the model's prediction generate the desired output. Cheng & Bernstein (2015) suggest that end-users select or "nominate" a list of features relevant for a specific classification to check via XAI methods how well the trained model picks up those or similar features. There are multiple examples of research papers in which the authors just followed suit and validated different black-box models according to some ground truth (e.g., Samek et al., 2021; Zhou et al., 2016). Yet, the researchers have failed so far to discover methods or some general rules of how XAI tools can help validate the model.

4 A Proposal of an Evaluation Framework for FER Algorithms in Web-based Learning Environments

As shown in Section 3.1, emotions play a crucial role in students' learning outcomes. They are linked to students' use of learning strategies, self-regulation of learning, and academic performance. Timely and accurate identification of negative emotions can help address students' disengagement, withdrawal, and academic failure. Moreover, negative emotions may help better identify educators who may lack clarity and communication competence. Conversely, positive emotions can signify more chances for academic success, engagement and motivation to learn. Therefore, to maximise the potential advantages of deploying a FER model in a web-based learning environment, *both positive and negative emotions should be identifiable for the FER system in use*.

The reviewed literature in Section 3.2 shows that the scientific community generally agreed that the human face can express seven universal (i.e., not culture-specific) emotions: anger, disgust, fear, happiness, sadness, surprise, and contempt. In the case of the suppressed or inactive facial muscles, the expression is regarded as neutral. Therefore, *the validity of any FER model trained to classify emotions not mentioned above may be questionable*. Moreover, extensive reviews and surveys on FER algorithms mentioned in Section 3.2 reveal superior classification accuracy among DL models. Thus, *the priority among different FER models should be given for the algorithms with a multilayer network*.

In Subsection 3.2.1, the study presents a variety of approaches utilised by DL models for FER in videos. In particular, current literature distinguishes three major methods: frame aggregation, emotion classification with expression intensity-invariant networks, and emotion classification with deep spatiotemporal networks. Although there is no general consensus within the academic community on the superiority of any of these approaches, FER algorithms from the latter group may be considered more promising. Specifically, their advantage lies in their ability to encode temporal dependencies in consecutive frames and learn spatial features along with temporal features. Given that any facial expression is of temporal nature – i.e., it has varying dynamics and intensity in time –

deployment of the FER models classifying on sequences (i.e., videos) rather than discrete frames should be prioritised.

Introduction to Section 3 briefly discusses that the FER models designed as an artificial neural network (ANN) with multiple layers do not provide explicit information on the role of input data' variations for the specific predictions. Therefore, due to the lack of transparency in decision-making, these models are often called black-boxes. Subsections 3.3.1 and 3.3.2 give an overview of Explainable AI (XAI) – a recent research field that attempts to improve the black-box algorithms' interpretability and comprehensibility. Specifically, the reviewed research papers argue that different XAI tools can help engender users' trust and confidence as well as contribute to the model's verification and validation. Moreover, in some cases, the interpretability of ML models has already become a legal prerequisite for their deployment and operation. Therefore, *understanding FER models' rationale and inner workings via XAI methods and tools can be an additional safeguard against potential algorithmic biases, severe public opposition or lawsuits*.

Subsection 3.3.3 delves deeper into the XAI and discusses different scopes of the ML models' interpretability. Based on the overview of different methods, one can conclude that there is no optimal way to explain algorithmic decisions despite a great variety of explanation techniques. Therefore, *when evaluating and explaining a FER model's decisions, one should refrain from overreliance on any particular method*.

In Section 3.4, the literature review reveals that the research community focuses primarily on the limited number of evaluation criteria for FER models. This study outlines two groups of evaluation metrics: performance-related and interpretability related. As for the former group, the F-score is likely to be the most optimal and informative metric for the FER model evaluation among the other metrics described in Subsection 3.4.1. This metric gives two valuable insights into whether the model does not overlook or does not excessively misclassify instances of a specific class (i.e., emotion). Thus, *calculating the F-score can help reveal which emotions the given FER model recognises best and worst*. Moreover, this metric can indicate how often the chosen model makes mistakes, what kinds of misclassification occur more often, and in which scenarios the system is likely to be incorrect. Therefore, *the F-score can help discover potential limitations and underlying logic of the FER model's decision-making*.

Moreover, some papers reviewed in Subsection 3.4.1 suggest using performance metric also as a relative measure of algorithmic generalisation. Specifically, *calculating the classifier's average performance drop on unseen data may provide a more informative metric regarding the FER model's generalisation abilities*.

The literature review in Subsection 3.4.2 shows that the evaluation criteria for any DNN model's interpretability can be grouped into two clusters: data-related interpretability criteria and post-hoc interpretability criteria. The former cluster of criteria summarises the training dataset's major characteristics and reveals potential data biases that can affect the model's performance. In particular, *the training data and the FER model itself should be carefully inspected for the presence of the following biases: sampling bias, aggregation bias, and capture bias*. The list of the biases suggested in this study may not be exhaustive.

As for the post-hoc interpretability criteria, the research findings from different papers suppose it to be a group of abstract and rather case-specific criteria. Most of the studies exemplified in Subsection 3.4.2 maintain that *FER models' explainability should be linked to some ground truth determined by domain experts or the model's users*. Specifically, the ground truth can be expressed in the form of input features that predefined a particular algorithmic decision. Moreover, *different research papers recommend displaying only the most informative and contrastive input features for a particular models' prediction*.

In view of the research findings discovered and outlined above, this study can propose the following selection criteria that can help identify the most suitable FER model for further adjustment to and deployment in a web-based learning environment:

1. The model classifies both positive and negative emotions
2. All the emotions identifiable by the model belong to the seven universal emotions
3. The model is an ANN with multiple hidden layers
4. The model takes a video fragment (i.e., sequence of multiple frames) as an input
5. The model was trained with data without capture bias

6. The model was trained with data without sampling bias

7. The model does not demonstrate aggregation bias

8. The model demonstrates fair generalisation abilities. I.e., the mean performance drop – is the smallest among the other models considered for the selection

9. The model's importance attribution to the specific input features (i.e., particular areas of human face activated for each emotion)[1] coincides with the one defined by humans.

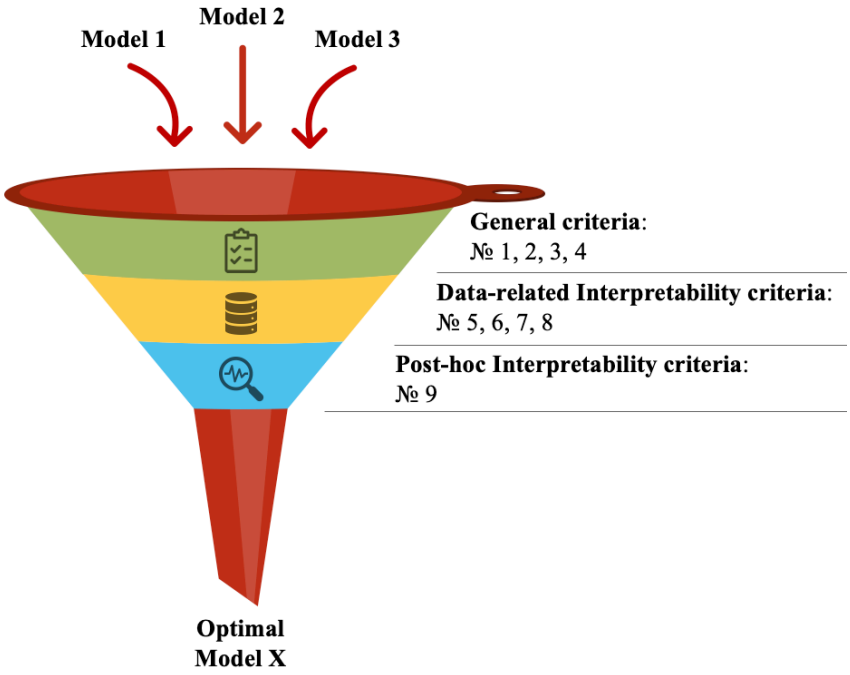

Figure 15. The workflow of the evaluation process of different FER models according to the proposed framework.

[1] The model's importance attribution to the specific input features (i.e., particular areas of human face activated for each emotion) is determined by different explanation methods. This study suggests that an image area can be considered significant for a given prediction if it meets two requirements: 1) it has the above-average concentration of highlighted pixels on the attribution maps or above-average colour intensity on a heatmap; 2) it is highlighted in at least 2/3 of attributions maps generated by different explanation methods for the given prediction.

By following the design science methodology, we will conduct a controlled experiment to test the use of the proposed framework. The experiment should be regarded as an integral part of the artefact's emergence and refinement. It must also be noted that this refinement may result in trivial fixes as likely as substantial changes to the design of the initially proposed evaluation framework and its components (Walls et al. 1992).

4.1 Experimental Use of the Proposed Evaluation Framework

4.1.1 Preliminary Model Selection

To make the experiment resemble the real-world conditions, three different pre-trained FER models from the Papers with Code[1] website were selected. The number of the models eventually analysed during the controlled experiment appeared to be limited due to the following constraints:

- As required in the proposed framework, the FER models classifying emotions other than those regarded as universal were not considered for the controlled experiment.
- In the absence of funding for this research, commercial FER models were not considered for the controlled experiment.
- The FER models trained to distinguish compound emotions based on Auction Units (AU) determined in FACS were not considered for the controlled experiment.
- To ensure variability and more comprehensive evaluation, priority was given to the models trained on different datasets.
- To analyse models' decisions and inner logic via various attribution methods, this study utilises the DeepExplain[2] framework – the only framework for XAI in Python supporting a wide range of gradient-based and perturbation-based local explanation techniques. As DeepExplain is compatible with Tensorflow and

[1] This is a specialised portal providing free and open access to state-of-the-art ML papers, code and evaluation tables. Available at https://paperswithcode.com/

[2] Available at https://github.com/marcoancona/DeepExplain

Keras libraries only, the FER models trained and compiled with different libraries (e.g., PyTorch) were not considered for the controlled experiment.

4.1.2 General Analysis

As a result, only three out of 12 models discovered on the Papers with Code portal were selected for the controlled experiment. Table 1 provides a detailed overview of the shortlisted FER models.

Table 1. Summarised information about FER models selected for the experiment.

	Model 1[1]	**Model 2**[2]	**Model 3**[3]
Architecture	ResNet-50 (He et al., 2016)	Inception-v3 (Szegedy et al., 2016)	MobileNets (Howard et al., 2017)
Input	Image (i.e., a single frame)		
Training dataset	FER-2013 (Goodfellow et al., 2013)		RAF-DB (Li et al., 2017; Li & Deng, 2018)
Classified emotions	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral		
Accuracy	71.25%	63.86%	76.96%
F1-score	0.69	0.60	0.69

As Table 1 shows, all the selected models can be used for real-time emotion analysis. Moreover, all these FER models process video data as a set of independent input images, ignoring potentially high interdependencies between adjacent frames. Thus, *none of the discovered models considers the temporal dynamics of facial expressions*. Moreover, **Model 1** and **Model 2** were trained with the same database. It means that these FER models may have encoded the common data biases inherent to the training dataset. Therefore, it will be particularly intriguing to compare their performance, as any recurrent similarities between **Model 1** and **Model 2** will be a sign of the validity of our assumption. As for the performance metrics, **Model 2** has the lowest accuracy equalling 63.86%,

[1] Available at https://github.com/ivadym/FER

[2] Available at https://github.com/ivadym/FER

[3] Available at https://github.com/Jackli95/real-time_emotion_recognition

whereas **Model 3** has the highest one reaching 76.96%. Interestingly, the difference in accuracy between **Model 1** and **Model 2** makes up almost 6%, however, the computed values of the F_1-score of both models are equal. Thus, the decision not to include accuracy as a metric for FER models' comparison and evaluation seems to be justified.

Based on the information collected in Table 1, we can already state that the selected models have met the three first general criteria determined in the evaluation framework. However, none of them fulfils the fourth criteria related to input specifications. To determine whether the selected FER models meet criteria 5, and 6, we need firstly to scrutinise their training data.

4.1.3 Analysis of Training Data

A. General Description of the FER-2013 Database

As indicated in Table 1, **Model 1** and **Model 2** were trained with the FER-2013 database created by Pierre Luc Carrier and Aaron Courville for the Facial Expression Recognition contest in 2013. In contrast to some other datasets for FER, it is publicly available for download[1]. The FER-2013 dataset was created using the Google image search API to collect images of human faces that matched a set of 184 emotion-related keywords (e.g., "blissful", "enraged"). Additionally, the emotion-related keywords were combined with words related to age, gender or ethnicity, to generate nearly 600 strings for image search queries. OpenCV computer vision library (Bradski & Kaehler, 2008) was used to obtain bounding boxes around each face in the first 1000 images returned for each query. Human labellers filtered out incorrectly labelled images, corrected the cropping, and removed duplicate data. Finally, cropped images were resized to 48x48 pixel resolution and converted to grayscale. Ian Goodfellow and Mehdi Mirza prepared a subset of the images for the contest mentioned above and mapped the emotion keywords into the seven basic emotion categories. The resulting FER-2013 dataset contains 35887 images – 28709 images for the training set, 3589 images for the validation set, and 3589 images for the test set respectively. The distribution of the emotions in all the sets is near-identical. Figure 16 displays randomly sampled 12 images from the FER-2013 database.

[1] Available at https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge

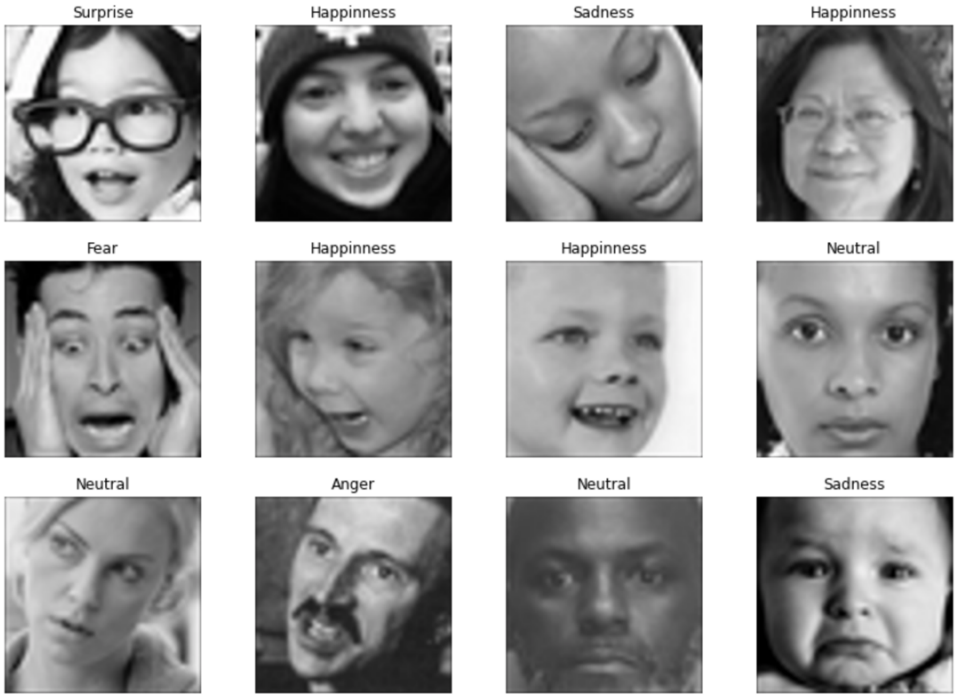

Figure 16. Random sample of images from the FER-2013 database.

B. General Description of the RAF-DB Database

As shown in Table 1, Model 3 was trained with the RAF-DB database. The curators of the database – i.e., Shan Li, Weihong Deng, and JunPing Du – used a data collection strategy similar to the one implemented for FER-2013. Firstly, they determined a set of emotion-related keywords (e.g., "smile", "cry", "scared", "frightened", "expressionless") to enable automatic image search via Flickr API on the Internet. Human annotators filtered out the resulting set of images according to the seven basic emotions. Finally, the images were aligned and resized to 100x100 pixel resolution. As a result, the original dataset contains 29672 real-world facial images captured in an unconstrained environment. However, only 15339 images are available for the general public – 12271 images in the training set and 3068 in the test subset. The distribution of the emotions in both sets is near-identical. Moreover, the database creators also specify some metadata. In particular, the age of the subjects depicted on the RAF-DB images varies from 0 to 70 years old. In the full version of the database, 52% are female, 43% are male, and 5% are undefined. As for racial distribution, most of the database is comprised of Caucasians (i.e., white) – 77%. Asians and African-Americans make up 15% and 8% respectively. Figure 17 displays randomly sampled 12 images from the RAF-DB database.

Figure 17. Random sample of images from the RAF-DB database.

C. Capture Bias

The general analysis of both training datasets showed that images in FER-2013 and RAF-DB were collected in the unconstrained environment without applying any specific filtering. Moreover, the randomly sampled examples confirmed that the captured facial expressions in both databases greatly vary in terms of camera angles, head poses, lighting conditions and occlusions. Thus, there is sufficient evidence to suppose that the possibility of a specific capture bias in FER-2013 or RAF-DB is likely to be insignificant.

D. Sampling Bias

As far as the sampling bias is concerned, we can look at it at least from two different perspectives. On the one hand, we can regard it as a highly skewed distribution of classes (i.e., emotions) in the training set. On the other hand, we can interpret it as an uneven distribution of images in terms of age, gender or ethnicity of people they display. In light of this newly discovered ambiguity, this specific criterion of the evaluation framework should be specified. Firstly, we define *sampling bias* as an uneven distribution of target classes (i.e., emotions). Secondly, we refer to the lack of representativeness of the training data regarding human biological characteristics – e.g., age, gender, ethnic diversity – as to *representation bias*. As the latter may be equally important as the former for the

classifier's generalisation ability, we add the *representation bias* to the framework as a stand-alone evaluation criterion.

However, discovering representation bias in the dataset may pose additional challenges. For example, given 28709 images in the FER-2013 training set, figuring out the actual distribution of data instances in terms of human biological characteristics may be technically burdensome. Thus, when precise information about the above-mentioned characteristics is not available, we suggest random sampling with statistically determined size to check manually if the training data *generally* reflect any ethnic, gender, and age diversity. When there is a possibility to determine the actual value distribution quantitatively, training data should *closely* reflect the specified characteristics of a target population at present. Nevertheless, the question of "how close" should be determined by an organisation or institution selecting a model.

As for the *sampling bias*, Figure 18 vividly demonstrates the uneven distribution of target classes in the FER-2013 dataset. In particular, the class "Happiness" alone makes up 25% of the image set. Obviously, **Model 1** and **Model 2** demonstrate the highest values of the F_1-score for this particular emotion – 0.89 and 0.87 respectively (see Figure 29 and Figure 30 in Appendix 2). Moreover, the class "Disgust" is significantly underrepresented, accounting for only 1.5% of the overall data. Yet, both FER models succeed in classifying it more accurately than "Fear" or "Sadness". Thus, we may assume that poor classification

accuracy for some emotions may potentially result from less intense facial muscle activation rather than underrepresentation.

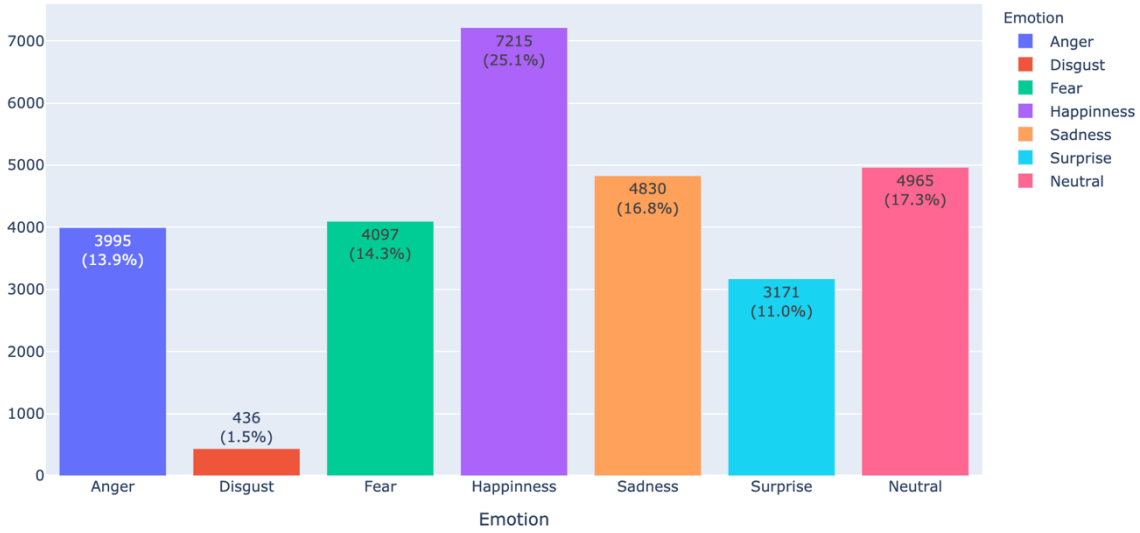

Figure 18. Distribution of classes in the FER-2013 training set.

Surprisingly, the class "Happiness" in the RAF-DB dataset also ranks first by constituting roughly 39% of the training set. And similarly to **Model 1** and **Model 2**, **Model 3** recognises this class with the utmost accuracy (see Figure 31 in Appendix 2). Nevertheless, images depicting emotions such as "Fear", "Disgust", and "Anger" altogether fall short of making up even a quarter of the RAF-DB database (see Figure 19). Interestingly, despite considerable underrepresentation of the latter class, **Model 3** classifies it as accurately as more numerous "Sadness" or "Neutral". Moreover, **Model 3** also demonstrates decent accuracy for the class "Surprise". This category's F_1-score is the second highest despite its relatively moderate fraction in the dataset.

Thus, we can conclude that, in our case, class imbalance does necessarily result in poor classification for the underrepresented groups. It may imply that looking at the distribution of classes in the training dataset does not always help predict classifier performance for certain categories. To refute or confirm this observation, we will later test each model on unseen data. During each test, we will document three best and three worst recognisable classes for each model. If the models similarly classify the same classes in other databases, we, therefore, may suggest that sampling bias in a multiclass FER training set should be a minor concern during the evaluation.

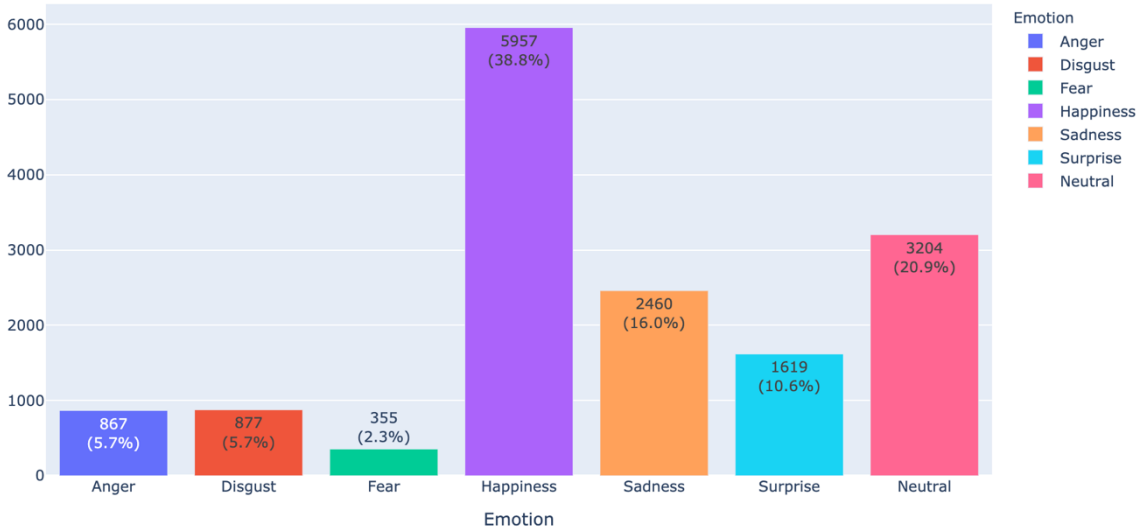

Figure 19. Distribution of classes in the RAF-DB training set.

E. Representation Bias

As argued earlier, we define *representation bias* as the lack of representativeness of the training data regarding biological characteristics of the individuals that these data depict – e.g., age, gender, ethnic diversity. Thus, during this experiment, we will analyse the distribution of data subjects for each training set according to two human biological characteristics: race and sex. We will consider that the training data is likely to have representation bias if it does not reflect the racial and sex distribution of the target population. To make the controlled experiment resemble real-world evaluation, we will use the 2019 enrolment data by race and sex provided by Yale University[1].

It is important to note that we intentionally omit the age distribution analysis of the training sets for two reasons. Firstly, the FER-2013 database documentation does not provide any information related to the data subjects' age. Secondly, determining human age based on a single input image with varying occlusions, head poses and camera angles seems at best unrealistic.

E.1 Racial Distribution of Data Subjects

[1] Yale University / The Office of Institutional Research (OIR). (2019). *2019-2020 Factsheet*. https://www.yale.edu/about-yale/yale-facts

As far as the racial distribution in FER-2013 is concerned, the original paper describing the dataset does not provide any information about the biological characteristics of the depicted people. Given the limited human resources and time constraints for this research, a human check is a troublesome option. To address this challenge, we will take advantage of the Deepface package – a lightweight face recognition and facial attribute analysis framework in Python (Serengil & Ozpinar, 2020). Based on single image input, the Deepface inbuilt DL models can predict a person's gender and race. Therefore, we will utilise this framework to get the approximate racial distribution of data subjects in the FER-2013 database. It is important to stress that the predicted values are just *rough approximations* with a high chance of considerable deviation from the actual numbers.

Figure 20 illustrates the distribution of the predicted categories. Importantly, the original model classifies images of individuals by six categories. Besides those displayed in Figure 20, the model predicted Indian and Middle Eastern ethnic groups. However, to enable comparability of this distribution with the Yale original data, we merged "White" and "Middle Eastern" into a single "White" class, and we added data instances of "Indian" to the "Asian" class. These aggregations are carried out according to the U.S. Guidance on Maintaining, Collecting, and Reporting Racial and Ethnic Data[1]. Figure 33 in Appendix 4 displays the original distribution of the predicted races.

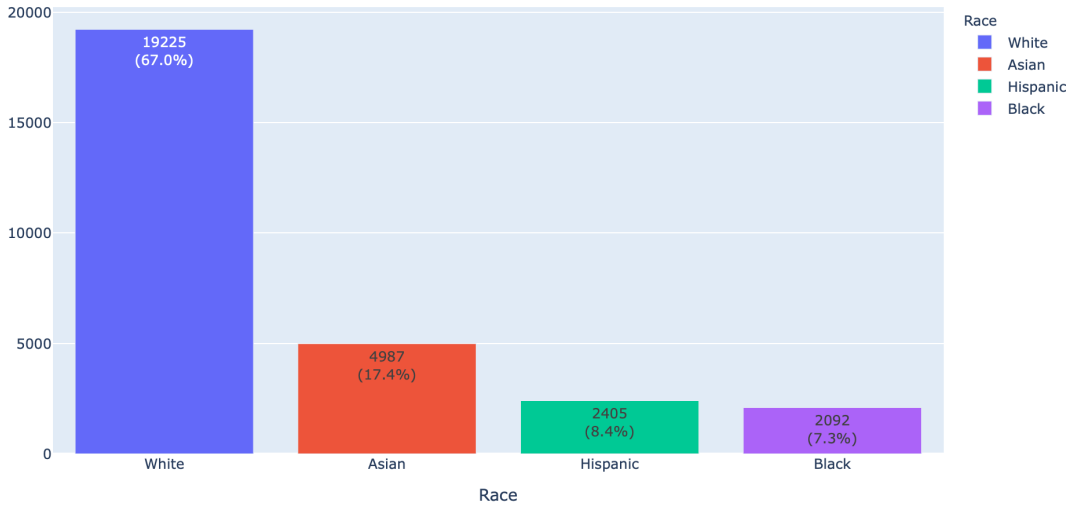

Figure 20. Predicted racial distribution in the FER-2013 training set.

[1] Final Guidance on Maintaining, Collecting, and Reporting Racial and Ethnic Data to the U.S. Department of Education. (2007). Federal Register Volume 72, Issue 202. Available at: https://www.govinfo.gov/app/details/FR-2007-10-19

One can see that FER-2013 has a vast diversity of races whose facial expressions are captured in this training set. Figure 20 clearly shows that the predominant group is White. Asians are the second biggest race present in the FER-2013 training set. Hispanic and Black comprise relatively equal fractions. Most importantly, racial categories present in the FER-2013 dataset represent approximately 90% of all students at Yale University. Moreover, racial distribution in the analysed dataset is similar to the one at Yale. Nevertheless, the trustworthiness of the predicted numbers is questionable, as the actual accuracy of the prediction model is unknown.

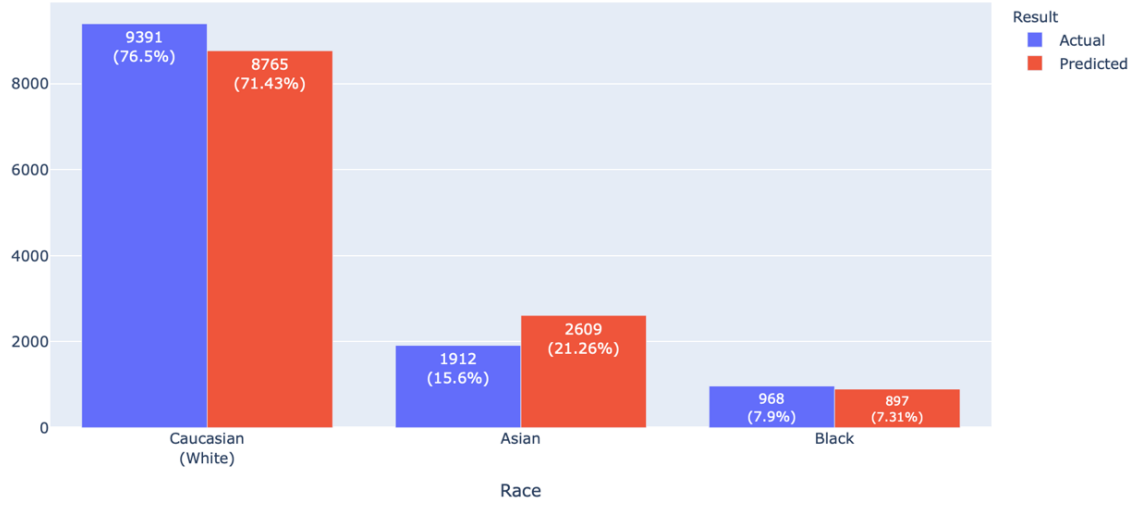

Figure 21. Actual vs predicted racial distribution in the RAF-DB training set.

As for the racial distribution in RAF-DB, Figure 21 depicts actual numbers for the training set as indicated in the database documentation. Interestingly, RAF-DB seems to reflect the diversity of students enrolled in Yale University to a lesser extent than FER-2013. Generally, it includes the images of individuals representing three major races: Caucasian (i.e., White), Asian, and African-American (i.e., Black). As the data were annotated by volunteers, it is impossible to find out whether they followed any objective criteria for racial classification. For example, it may be the case that annotators did not differentiate people of Hispanic origin as a separate category and instead added those to the "White" class. Given that images were collected in the wild without any filtering related to human ethnicity, the aforementioned annotation strategy seems particularly probable. To check this assumption, we utilised the Deepface framework to compare the predicted distribution of classes by race with the actual one in the RAF-DB database. We used the same aggregation techniques applied to FER-2013. Additionally, following our assumption, we add the predicted instances of the "Hispanic" class to the "Caucasian"

(i.e., White) category. Figure 21 also illustrates the predicted distribution of data subjects by race in the RAF-DB database. Moreover, Figure 34 in Appendix 4 demonstrates the predicted racial distribution in RAF-DB before aggregation.

As shown in Figure 21, the difference between the actual and predicted class fractions after aggregation is not critical. Therefore, we can suggest that RAF-DB images may be racially more diverse than the annotators decided them to be. However, the result of our controlled experiment does not provide sufficient and indubitable evidence to either confirm or refute our hypothesis. Provided that the exemplified races in RAF-DB represent roughly 80% of all the students at Yale University, representation bias regarding the ethnic set-up of this database is obvious and yet not critical.

E.2 Sex Ratio of Data Subjects

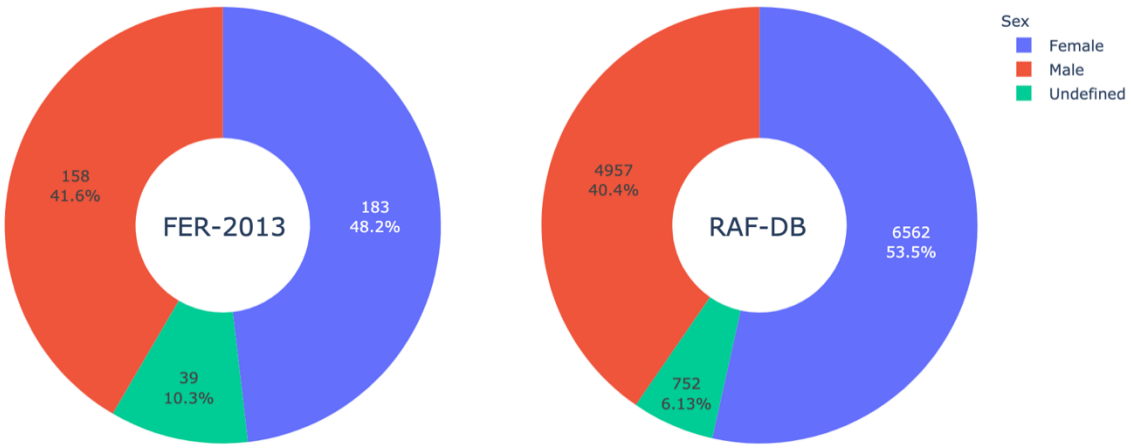

Figure 22. Sex ratio in the FER-2013 and RAF-DB training set.

As mentioned earlier, the paper proposing and describing the FER-2013 database does not contain any information related to the people depicted in the images. However, determining the sex of a person does not require any domain-specific knowledge, as in the case of classifying individuals into ethnic groups. Therefore, we determine the sex ratio in the FER-2013 dataset via sampling and manual annotation. As we aspire to make the sample closely represent the database's overall population in terms of sex, our sample size equals 380, ensuring a 95% confidence level and 5% precision. The author of the paper carried out the initial annotation. After that, each of the three volunteers additionally checked the annotated images. The images that created any ambiguity among the volunteers and the author were classified as "Undefined". The list of the volunteers and

their contact information can be found in Appendix 5. Figure 22 displays the sex ratio among data subjects in the FER-2013 sample and RAF-DB training set.

As one can see, the fraction of male subjects in FER-2013 is likely to be slightly smaller than females, accounting for approximately 42%. Those images grouped into the "Undefined" category mostly depict babies or children whose sex differences are not yet conspicuous. Most importantly, the numbers on the graph demonstrate that the sex ratio in the FER-2013 database with high confidence may correspond to the one at Yale University (see Figure 32 in Appendix 3).

The graph on the right in Figure 22 also displays the sex ratio in the RAF-DB training dataset. The exact fractions were calculated based on the detailed information provided by the database annotators for each image. As one can see, the population distribution by sex in RAF-DB is almost similar to FER-2013. Notably, the fraction of females in this dataset is the same as at Yale University, although the proportion of males is 6% less. However, the observed differences in both datasets are, in fact, negligible. Thus, we can conclude that the representation bias regarding the sex ratio in FER-2013 and RAF-DB is absent or insignificant.

F. Aggregation Bias

As argued in Section 3.4.2, aggregation bias occurs when one draws false conclusions for some population group based on observations made for a different group. This bias can also be present in the model tuned to the quantitatively dominant class in the training dataset. For example, in our case, the analysed models may potentially perform varying accuracy levels depending on learners' ethnicity, gender, or age. However, to detect anomalous model's performance with regards to a specific population group requires testing with thoroughly profiled and annotated data.

To check the presence of aggregation bias in the analysed models, we will use the RAF-DB test set, which contains 3068 images annotated by sex, race, and age. As argued earlier, determining the age of a person based on a single image may be quite unrealistic, so the quality of annotations made by volunteers is likely to be questionable. Thus, during the experiment, we will focus on the models' performance for different gender and ethnic groups, ignoring the age factor. The proportions of those groups closely resemble their

distribution in the RAF-DB training set (see Figure 21 and Figure 22). Figure 23 and Figure 24 depict the results of the testing.

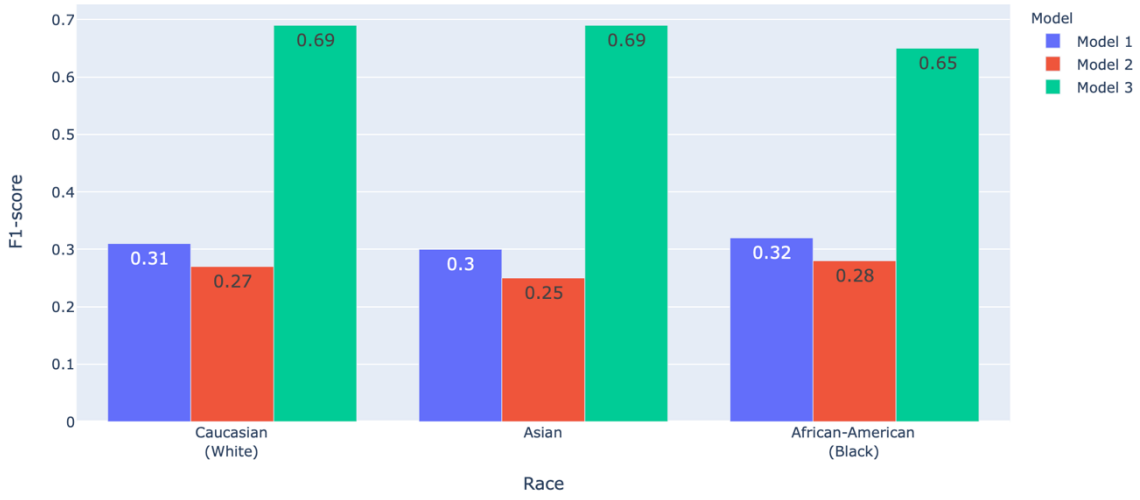

Figure 23. Performance of the analysed models for the different ethnic groups – Caucasian, Asian, and African-American.

Figure 23 vividly shows that each model demonstrates similar accuracy levels (measured as the F_1-score) for different ethnic groups. (see Figures 35, 36 and 37 in Appendix 6 for detailed classification results). Figure 24 shows similar results. The models classify people of different sexes with almost the same accuracy. Thus, based on the presented evidence, we may suppose that **Model 1**, **Model 2**, and **Model 3** are unlikely to demonstrate aggregation bias regarding gender and ethnicity of the data subjects (i.e., learners) during the classification.

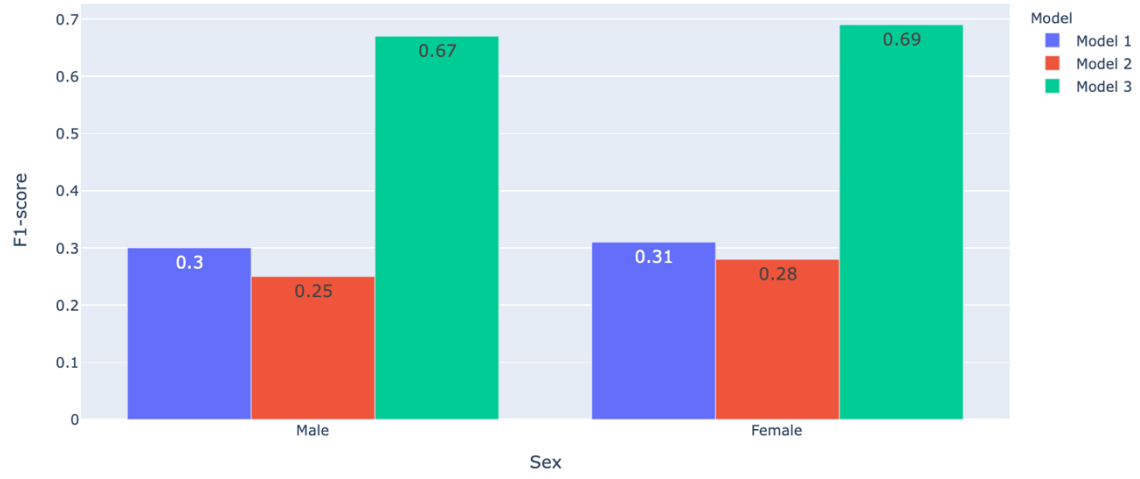

Figure 24. Performance of the analysed models for the different gender groups – males and females.

G. Performance drop

As argued in Section 3.4.1, evaluating the classifier's generalisation abilities by calculating its average performance drop may be a simple yet very informative metric. Therefore, we proposed it in the evaluation framework as a separate criterion. However, as also shown in Section 3.4.1, algorithmic performance can be measured via different metrics. For example, Torralba & Efros (2011) chose accuracy as a performance metric. Provided that a different metric had been selected, the calculated values of the performance drop might have shown different results. Therefore, to make the proposed framework more objective, we calculated the performance drop of the analysed models based on two metrics – accuracy and the F_1-score.

Since the calculation of the performance drop requires testing on unseen data, we used five different FER datasets. Additionally, we tested **Model 1** and **Model 2** on the RAF-DB test set and **Model 3** on the FER-2013 test set respectively. The complete list of the datasets used in this part of the controlled experiment is provided below. As FER-2013 and RAF-DB were thoroughly described earlier in this section, they are not mentioned in the list.

1) **CK+** (Lucey et al., 2010). The database consists of images depicting 182 adults between the ages of 18 and 50 years. 69% of the subjects are female, whereas 31% are male. The paper proposing the dataset also indicates that 81% of adults are Euro-American, 13% are Afro-American, and 6% belong to other ethnic groups. The images presented in CK are 2105 fragmented frontal-view video sequences recorded in a studio with uniform lighting conditions. This study uses a fragment of the original database consisting of 1520 images representing six basic facial expressions (i.e., anger, disgust, fear, happiness, sadness, surprise) and a neutral face.

2) **JAFFE** (Lyons et al., 1998). The database consists of 219 frontal-view facial expression images recorded from ten subjects in a studio with uniform lighting conditions. All the subjects are Japanese females. The hair of the subjects was intentionally tied to expose all expressive zones of the face. The database includes three or four examples of each female expressing one of the six basic facial expressions (i.e., anger, disgust, fear, happiness, sadness, surprise) and a neutral face.

3) **iSAFE** (Singh & Benedict, 2019). The dataset consists of image sequences of 44 volunteers of Indian origin whose age ranges from 17 to 22 years. Out of 44 recorded subjects, 25 are male, whereas 19 are female. All the volunteers were placed in a brightly lit room and shown a set of carefully pre-selected videos aimed to induce particular emotions. A camera placed in front of the subjects recorded their facial response throughout the experiment. The subjects' recorded facial expressions were self-annotated and later cross-annotated by a professional annotator according to the six basic facial expressions. The recorded videos were analysed and trimmed to include only the segments with identifiable facial expressions. This study uses a fragment of the original database consisting of 9046 images depicting the facial expressions of 22 subjects.

4) **IMPA-FACE3D** (Mena-Chalco et al., 2011). The dataset is composed of 38 subjects – 22 men and 16 women whose age ranges from 19 to 65 years. The images of the subjects were captured in a frontal position in a room with controlled lighting conditions. The subjects were not allowed to wear eyeglasses or other objects that could modify their facial appearance. However, no restrictions on clothing or hairstyle were imposed. This study uses a fragment of the original database consisting of 252 images depicting the facial expressions of 36 subjects. All the images represent one of the six basic facial expressions (i.e., anger, disgust, fear, happiness, sadness, surprise) or a neutral face. Moreover, in the following, this study will refer to this database as to **FacesDB**.

5) **AffectNet** (Mollahosseini et al., 2017). This database is arguably the largest database of images depicting human facial expressions collected in the wild. The images were automatically collected by querying emotion-related keywords from three search engines – i.e., Google, Bing, and Yahoo. In addition, the keywords were combined with words related to gender, age, or ethnicity. Besides English, the search queries were also written in Spanish, Portuguese, German, Arabic and Farsi. The OpenCV computer vision library was used to obtain bounding boxes around each face returned by the query. Human annotators labelled a total of 450,000 images into both discrete categorical and continuous dimensional models. The former includes the following facial expressions: neutral, happy, sad, surprise, fear, anger, disgust, contempt, none, uncertain, and non-face. This study uses a fragment of the original database consisting of 3500 images depicting one

of the six basic facial expressions (i.e., anger, disgust, fear, happiness, sadness, surprise) or a neutral face.

Table 2 shows the results of the testing with the datasets described above. The numbers in Table 2 represent the performance of the analysed models that is measured as accuracy. Table 3 also demonstrates the performance of the pre-selected models, however, their classification accuracy is measured as the F1-score. In addition, Table 9 in Appendix 7 demonstrates the top three best and worst identifiable facial expressions by each model for a respective test set.

Table 2. Performance (measured as the F_1-score) of the selected FER models on different datasets. The numbers marked with an asterisk (i.e., '*') stand for the original/initial classification accuracy (also measured as the F_1-score) for a respective model on the test set.

Dataset / Model	**FER-2013**	**RAF-DB**	**CK+**	**JAFFE**	**iSAFE**	**FacesDB**	**AffectNet**
Model 1	**0.69***	0.31	0.61	0.41	0.17	0.23	0.26
Model 2	**0.61***	0.27	0.56	0.23	0.18	0.18	0.24
Model 3	0.45	**0.69***	0.67	0.33	0.30	0.45	0.45

Table 3. Performance (measured as accuracy) of the selected FER models on different datasets. The numbers marked with an asterisk (i.e., '*') stand for the original/initial classification accuracy of a respective model on the test set.

Dataset / Model	**FER-2013**	**RAF-DB**	**CK+**	**JAFFE**	**iSAFE**	**FacesDB**	**AffectNet**
Model 1	**71.25***	56.06	72.43	43.66	32.85	26.19	30.26
Model 2	**63.86***	51.33	70.26	29.58	33.43	25.0	28.89
Model 3	51.52	**76.96***	79.34	38.03	46.09	45.63	40.53

As we used data from Table 2 and Table 3 as an input to calculate the mean performance drop of each model, we will not comment on it. Nevertheless, the data presented in Table 4 and Table 5 are of primary interest for this research. The former shows the mean performance drop calculated with data from Table 2, whereas the latter illustrates the values of the mean performance drop calculated with data from Table 3.

Table 4. The mean performance drop of the analysed FER models. The performance drop was calculated based on the data presented in Table 2.

Dataset / Model	FER-2013	RAF-DB	CK+	JAFFE	iSAFE	FacesDB	AffectNet	Mean Drop
Model 1	**X**	-55.1	-11.6	-40.6	-75.4	-66.7	-62.3	**-52.0%**
Model 2	**X**	-55.7	-8.2	-62.3	-70.5	-70.5	-60.7	**-54.65%**
Model 3	-34.8	**X**	-2.9	-52.2	-56.5	-34.8	-34.8	**-36.0%**

Table 5. The mean performance drop of the analysed FER models. The performance drop was calculated based on the data presented in Table 3.

Dataset / Model	FER-2013	RAF-DB	CK+	JAFFE	iSAFE	FacesDB	AffectNet	Mean Drop
Model 1	**X**	-21.3	+1.7	-38.7	-53.9	-63.2	-57.5	**-38.8%**
Model 2	**X**	-19.6	+10.0	-53.7	-47.7	-60.9	-54.8	**-37.8%**
Model 3	-33.1	**X**	+3.1	-50.6	-40.1	-40.7	-47.3	**-34.8%**

The data in both tables vividly demonstrates that none of the analysed models has perfect generalisation capability. Each of them showed more than a 30% drop in classification accuracy. Provided that the test datasets have an uneven distribution of classes, the mean performance drop that takes accuracy as an input metric fails to provide a robust measure of the models' performance for the minor classes, especially if those classes tend to have a higher classification error. Thus, based on the experiment's results demonstrated above, we can suppose that the mean performance drop suggested by the framework as an evaluation criterion should take the F1-score as an input metric rather than accuracy.

Additionally, even though Model 1 and Model 2 were trained with the same data set, the information presented in Table 2 and Table 4 provides convincing evidence that the latter performs worse. Therefore, the decision-making of Model 2 regarding the classification of the universal facial expressions will not be analysed in Section 4.1.4.

4.1.4 Analysis of Post-hoc Interpretability

As discovered in Subsection 3.4.2, the interpretability of the analysed model should be linked to some ground truth determined by domain experts or the model's users. Besides, the last evaluation criterion in the proposed framework specifies that in the context of FER, the ground truth for each class can be a group of particular areas of a human face activated for each facial expression. In addition, the framework requires that the importance attribution of the analysed FER model should coincide with the one defined

by humans. Specifically, this study suggests that an image area can be considered significant for a given prediction if it meets two requirements:

1) The area of interest in the image has the above-average concentration of highlighted pixels on the attribution maps or above-average colour intensity on a heatmap

2) The area of interest in the image is highlighted in at least 2/3 of attributions maps generated by different explanation methods for the given prediction.

As shown in Section 3.3.3, the research community has proposed numerous attribution methods that can explain the logic behind the model's decisions either globally or locally. As there is a considerably wider variety of local rather than global interpretability methods, this study will focus on the former group. Nevertheless, it is important to note that each local interpretability method may have some disadvantages and controversies related to its fidelity, noisiness or accuracy. To reduce the risk of over-reliance on any particular method, the proposed framework suggests analysing models' decisions and inner logic via several different attribution methods. Thus, as a part of the controlled experiment, this study will utilise the following methods: Gradient∗Input (Shrikumar, 2016), Integrated Gradients (Sundararajan et al., 2017), DeepLIFT (Shrikumar et al., 2017), ε-LRP (Bach et al., 2015) and LIME (Ribeiro et al., 2016). This study used the DeepExplain framework in which all of the before-mentioned methods, except for LIME, had been implemented. Moreover, the study took advantage of the Python implementation of LIME provided on its author's GitHub account[1].

As for ground truth, this study relied on the description of the seven basic facial expressions provided on the official website[2] of Paul Ekman – the very researcher who discovered micro-expressions and provided substantial and ample evidence for their universality. As an example, Figure 25 demonstrates the description and mapping for "Anger". The rest of the Figures describing the movement and mapping of different facial regions for each universal expression can be found in Appendix 8.

[1] Available at https://github.com/marcotcr/lime

[2] Available at https://www.paulekman.com/universal-emotions/

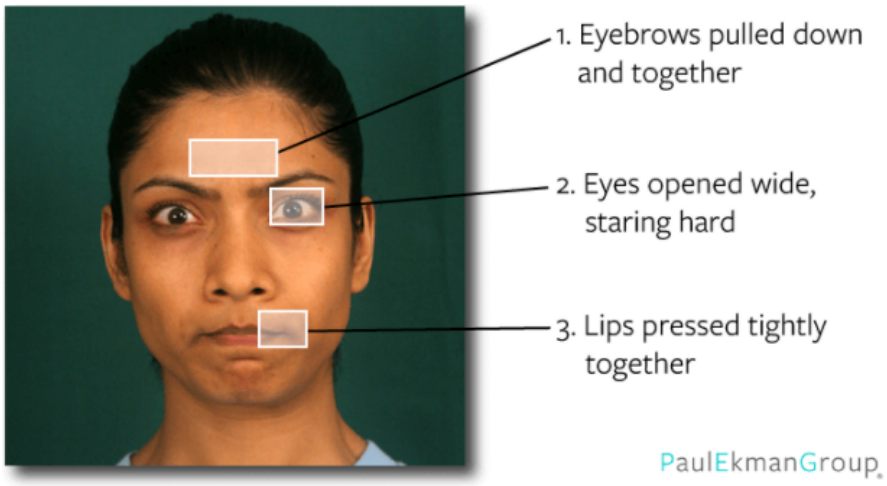

Figure 25. The face of anger[1].

In fact, we analysed the hidden logic of **Model 1** and **Model 3** by selecting six images from the CK+ database. Each image represents one of the six universal facial expressions – i.e., anger, disgust, fear, happiness, sadness and surprise. The decision to select CK+ as a test database was made for two reasons. Firstly, the images from that database can be considered as "unseen data" for both models. Therefore, none of the models had any advantage when given data from CK+. Secondly, among all the datasets mentioned and used in this study, CK+ is arguably the most widely-used database for the development, validation and testing of FER algorithms within the scientific community.

Figure 26 and Figure 27 illustrate the attribution maps generated for the input image of the class "Angry" for **Model 1** and **Model 3** respectively[2]. In particular, as determined by Ekman, red rectangles outline areas of the human face that are discriminative for this particular emotion (see Appendix 8). Additionally, the concentration of crimson pixels in some areas of this face signifies the importance of these regions for the given facial expression to be classified as angry. It must be noted, however, that a heatmap generated by LIME shows a positive contribution of the specific image fragments with different

[1] Available at https://www.paulekman.com/universal-emotions/what-is-anger/

[2] The rest of the attribution maps generated for **Model 1** and **Model 3** can be found in Appendix 9 and Appendix 10 respectively.

shades of blue, not crimson. Crimson colour, in this case, highlights image areas that decrease the probability of the given facial expression being classified as angry.

Generally, Figure 26 demonstrates that **Model 1** follows human logic, as most of the pixels or highlighted fragments are concentrated within the red rectangles. In contrast, attribution maps generated for **Model 3** (see Figure 27) are slightly nosier – i.e., a significant number of highlighted pixels fall outside the outlined areas. Nevertheless, even though importance attribution by **Model 1** may seem more reasonable to humans, none of the analysed models demonstrated perfect alignment with the human theoretical underpinnings.

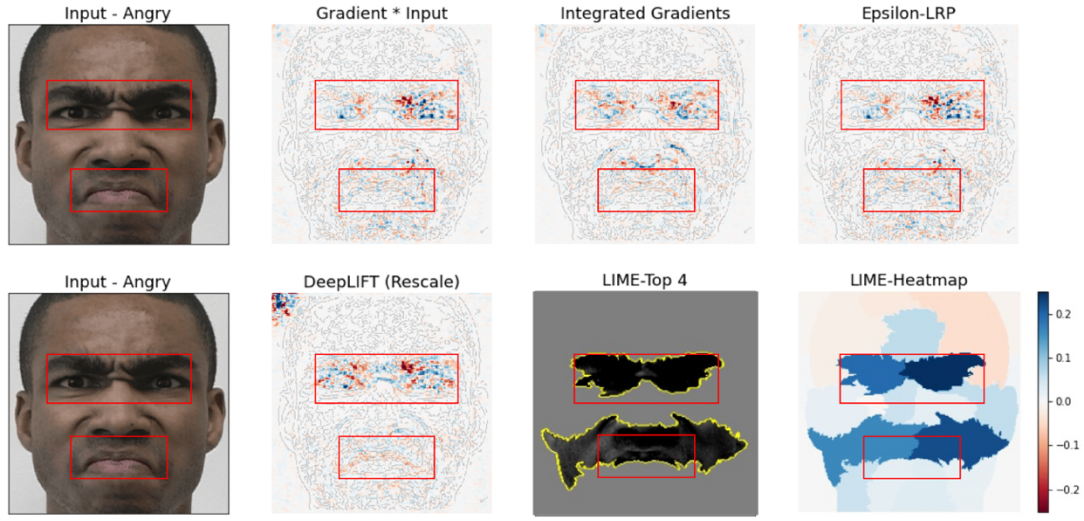

Figure 26. Attribution maps generated for the input image of the class "Angry" (Model 1).

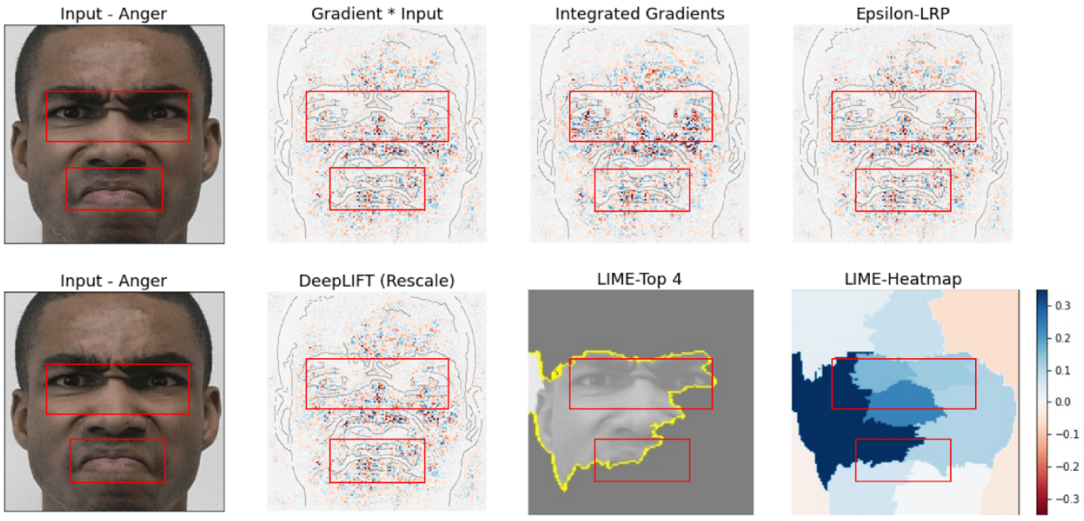

Figure 27. Attribution maps generated for the input image of the class "Angry" (Model 3).

Moreover, despite the minimal number of input images analysed with the attribution maps during the experiment, one can already observe that both models tend to consider image fragments beyond the area of a human face as relevant for the classification – e.g., Figures 41, 42, 44 and 51 in Appendix 9 and 10. Such behaviour may signify the potential presence of biases inherited from the training data. For example, the training set could contain watermarked images for specific classes so that the presence or absence of these patterns could have become an extra feature affecting the model's classification. As Lapuschkin et al. (2019) demonstrated, SpRAy could be most helpful in this case, as it proved to be an effective method for pinpointing implicit and anomalous patterns in the model's prediction behaviour. Although this study did not apply any global explanation methods to the models' inner workings due to temporal constraints, the controlled experiment vividly demonstrated that an evaluation of any DNN without analysing its global logic should be considered incomplete.

All in all, Table 6 shows the aggregated results of the models' evaluation during the controlled experiment. Each column informs of the model's compatibility with one specific evaluation criterion presented in the framework. As one can see, all three models demonstrate relatively good results. Their training data do not seem to have any notable biases and thus closely resemble real-world setting. When judged by information from the first eight columns, **Model 1** and **Model 2** are equally suitable for deployment in a web-based learning environment. However, the values of the mean performance drop tilt this balance in favour of **Model 3**, indicating its more advanced generalisation capabilities. Unfortunately, due to the reasons outlined above, drawing conclusions on the given models' interpretability does not seem feasible within the scope of this controlled experiment.

Table 6. Aggregated results of the controlled experiment.

Criteria / Model	**Accuracy (The F_1-score)**	**Positive & Negative Emotions Identifiable**	**Universality of Identifiable Emotions**	**Multi-layer ANN**	**Video Sequence as Input**	**No Capture Bias in Training Data**	**No Sampling Bias in Training Data**	**No Representation Bias in Training Data**	**No Aggregation Bias for Different Demographic Groups**	**Mean Performance Drop, %**	**The Models Considers Relevant Areas of Human Face**
Model 1	0.69	+	+	+	–	+	+	+	+	52.0	NA

Criteria / Model	Accuracy (The F_1-score)	Positive & Negative Emotions Identifiable	Universality of Identifiable Emotions	Multi-layer ANN	Video Sequence as Input	No Capture Bias in Training Data	No Sampling Bias in Training Data	No Representation Bias in Training Data	No Aggregation Bias for Different Demographic Groups	Mean Performance Drop, %	The Models Considers Relevant Areas of Human Face
Model 2	0.60	+	+	+	–	+	+	+	+	54.65	NA
Model 3	0.69	+	+	+	–	+	+	+	+	36.0	NA

4.1.5 Summary of the Controlled Experiment and Discussions

The controlled experiment demonstrated that the proposed framework ensures a comprehensive multidimensional evaluation of any FER model considered for deployment in a web-based learning environment. The variety of criteria can give a better understanding of the model's strengths and pinpoint its potential weaknesses. Moreover, the controlled experiment showed that the proposed evaluation order can help filter out the most unsuitable FER models at the very beginning and allocate more time to a thorough analysis of the most relevant ones.

Additionally, the experiment proved to be an effective technique to discover several ambiguities and limitations of the initially determined criteria. First of all, the sixth criteria regarding the absence of sampling bias in its initial form proved to be somewhat vague in practice. Therefore, we suggested that the absence of sampling bias in the training data should be regarded as a balanced distribution of the model's target classes (i.e., emotions). Secondly, by narrowing down the definition of sampling bias, it was crucial to ensure that the age, gender and ethnic diversity of the analysed people are duly considered during the evaluation. As those biological characteristics may be equally important as the distribution of target classes for the classifier's generalisation capability, we modified the framework by adding the representation bias as a stand-alone evaluation criterion. Moreover, taking actual statistics of Yale University as a baseline for comparison justified the decision to include representation bias as another crucial criterion for the models' all-around evaluation.

Additionally, the controlled experiment verified the validity of applying the idea of performance drop to FER models. As expected, all the pre-selected models demonstrated a significant drop in classification accuracy when tested with "unseen" data. Moreover, we empirically demonstrated that using the F-score rather than accuracy as a metric for performance drop calculation may provide a more informative and robust measure of algorithmic generalisation capabilities.

Finally, conducting the controlled experiment helped reinforce the line of reasoning that local explanation methods alone are not sufficient to evaluate a model's interpretability. Although local explanations provide valuable insight into the algorithmic logic for particular cases, they do not reveal the models' decision-making and potential inherent biases on a general level.

Based on the outcomes of the experiment outlined above, the thesis proposes the modified version of the evaluation framework for FER models. Its evaluation criteria now look as follows:

1. The model classifies both positive and negative emotions
2. All the emotions identifiable by the model belong to the seven universal emotions
3. The model is an ANN with multiple hidden layers
4. The model takes a video fragment (i.e., sequence of multiple frames) as an input
5. The model was trained with data without capture bias
6. The model was trained with data without sampling bias. I.e., the training data did not have any severe imbalances in terms of its target classes (i.e., emotions).
7. The model was trained with data without representation bias. I.e., the training data generally reflects the age, gender and ethnic diversity of the target population.
8. The model does not demonstrate aggregation bias.
9. The model demonstrates fair generalisation abilities. I.e., the mean performance drop – which is calculated based on the F-score for each test dataset – is the smallest among the other models considered for the selection

10. The model's importance attribution to the specific input features (i.e., particular areas of human face activated for each emotion) coincides with the one defined by humans. This rule applies to local and global explanations equally.

Figure 28 presents an updated visualisation of the evaluation process' workflow suggested by the proposed framework.

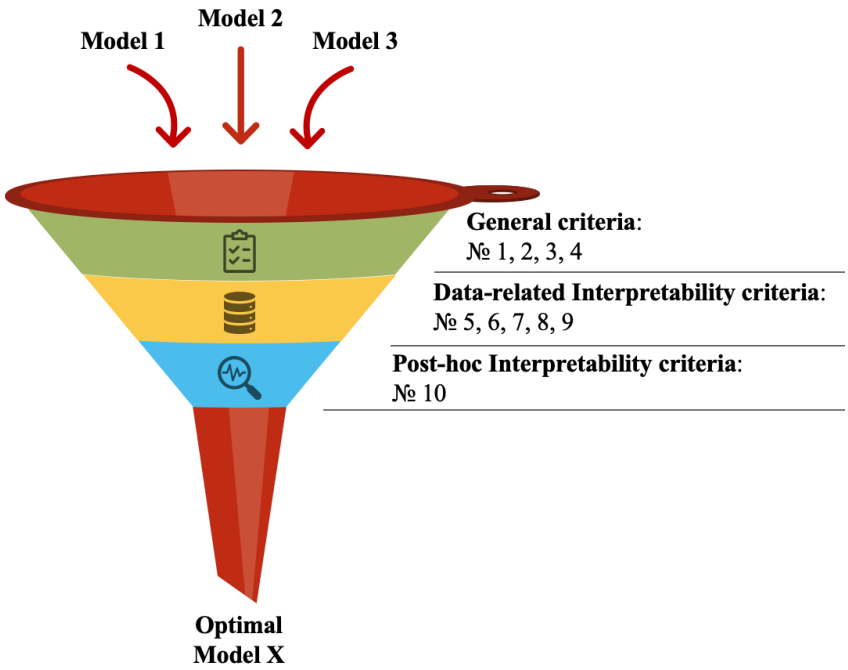

Figure 28. The workflow of the evaluation process of different FER models suggested by the modified version of the initially proposed framework.

The proposed framework is the first documented attempt to suggest generic evaluation criteria for FER model evaluation and selection that go beyond performance metrics. Moreover, the framework includes interpretability as an essential requirement for model validation and deployment in real-time.

Summary

As a result of the COVID-19 pandemic, the frequent suspension of classes and the near-total closures of thousands of educational institutions worldwide create an insuperable impediment to conventional face-to-face classes and thus make remote learning the only feasible alternative. Unfortunately, learners' facial expressions — a valuable natural source of instant feedback to the learning experience for any educator — are commonly overlooked during online sessions. Although modern DL systems classifying human emotions by a single facial expression may be crucial for improving the online learning experience, the general reasoning and decision-making of those systems often remain ignored during their evaluation and deployment.

This study addresses this problem by proposing a robust, theory-driven, and case-oriented evaluation framework enabling preliminary selection of the FER models that provide accurate, valid and trustworthy information on students' learning experience in a web-based learning environment. Contrary to the existing evaluation approaches, the proposed framework goes beyond conventional performance metrics (e.g., accuracy), shifting the focus to detecting potentially inherent biases, evaluating algorithmic generalisation capabilities and models' interpretability. By following the DS research methodology, this study also includes the controlled experiment that validates the proposed evaluation criteria and serves as an input for a new cycle of the artefact's building and refinement. The overall robustness and replicability of the experiment are ensured by using publicly available algorithms and real training data – i.e., pre-trained FER models and datasets with labelled facial expressions.

Despite the multiple limitations discussed in this study, the proposed evaluation framework is arguably the first documented scientific attempt to enable comprehensive analysis and evaluation of FER models for their eventual deployment in web-based learning environments.

Future work

Although this research and its findings lay the solid groundwork for comparing and evaluating different FER models for a web-based learning environment, the limitations outlined in Section 2.2 make further research particularly crucial for expanding and strengthening its initial results. First and foremost, the proposed evaluation framework requires extensive recurrent testing and evaluation with the involvement of the practitioners — i.e., FER researchers, developers and educators. Their feedback is bound to challenge and verify the assumptions and findings presented in this thesis. Moreover, future experiments and testing should also cover commercial FER systems (e.g., Amazon Rekognition API[1]). Irrespective of the outcome, such testing is likely to result in interesting findings with a high utility for the research community and potentially significant implications for the FER market players.

Additionally, as more and more human interactions occur on the web, models for facial expression recognition have enormous potential for deployment in use cases going beyond the scope of the current study. For instance, FER algorithms may be helpful during negotiations with a business partner, bring additional insights during the presentation of a new product, or provide an instant evaluation of the user experience with intelligent virtual assistants. Moreover, some researchers have already demonstrated the tremendous importance of emotional awareness and facial expression recognition in the context of semi-autonomous or fully autonomous vehicles (Gressenbuch & Bergemann; Izquierdo-Reyes et al., 2018). Thus, a constantly growing variety of use cases for the application of FER models create a pressing need for further research about their robust comparison and evaluation in each given example.

[1] Available at https://aws.amazon.com/rekognition/?nc=sn&loc=1&blog-cards.sort-by=item.additionalFields.createdDate&blog-cards.sort-order=desc

References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160.
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, *31*, 9505-9515.
3. Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2017). Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*.
4. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.
5. Artino Jr, A. R., & Jones II, K. D. (2012). Exploring the complex relations between achievement emotions and self-regulated learning behaviors in online learning. *The Internet and Higher Education*, *15*(3), 170–175.
6. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, *10*(7), e0130140.
7. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K. R. (2010). How to explain individual classification decisions. *The Journal of Machine Learning Research*, *11*, 1803-1831.
8. Bargal, S. A., Barsoum, E., Ferrer, C. C., & Zhang, C. (2016). Emotion recognition in the wild from videos using images. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 433–436. https://doi.org/10.1145/2993148.2997627
9. Benitez-Quiroz, C., Srinivasan, R., & Martinez, A. (2016). EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016*, 5562-5570.
10. Bosch, N., & D'Mello, S. (2017). The affective experience of novice computer programmers. *International journal of artificial intelligence in education*, *27*(1), 181-206.
11. Bradski, G., & Kaehler, A. (2008). *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc.
12. Buciu, I., & Pitas. (2004). Application of non-negative and local non-negative matrix factorization to facial expression recognition. *Proceedings of the 17th*

International Conference on Pattern Recognition, 2004. ICPR 2004, 1, 288-291 Vol.1.

13. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77–91.
14. Byeon, Y.-H., & Kwak, K.-C. (2014). Facial expression recognition using 3d convolutional neural network. *International Journal of Advanced Computer Science and Applications*, *5*(12).
15. Carcagnì, P., Del Coco, M., Leo, M., & Distante, C. (2015). Facial expression recognition and histograms of oriented gradients: a comprehensive study. *SpringerPlus*, *4*(1), 645.
16. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.
17. Chan, C. H., Kittler, J., Poh, N., Ahonen, T., & Pietikäinen, M. (2009). (Multiscale) local phase quantisation histogram discriminant analysis with score normalisation for robust face recognition. *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, 633–640.
18. Cheng, J., & Bernstein, M. S. (2015). Flock: Hybrid crowd-machine learning classifiers. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 600–611.
19. Cuddy, A. J., Wilmuth, C. A., Yap, A. J., & Carney, D. R. (2015). Preparatory power posing affects nonverbal presence and job interview performance. *Journal of Applied Psychology*, *100*(4), 1286.
20. Dabbagh, N., & Kitsantas, A. (2004). Supporting self-regulation in student-centered web-based learning environments. *International Journal on E-learning*, *3*(1), 40-47.
21. Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.
22. Deng, J., Berg, A., Satheesh, S., Su, H., Khosla, A., & Li, F. F. (2012). Large scale visual recognition challenge. http://www.image-net.org/challenges/LSVRC/ 2012/.
23. Dong, J., Zheng, H., & Lian, L. (2018). Dynamic facial expression recognition based on convolutional neural networks with dense connections. In *2018 24th International Conference on Pattern Recognition (ICPR)* (pp. 3433-3438). IEEE.
24. Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*.
25. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
26. Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 0210–0215.

27. Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. Communications of the ACM, 63(1), 68-77.
28. Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, *111*(15), E1454-E1462.
29. Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., & Pal, C. (2015). Recurrent neural networks for emotion recognition in video. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 467–474.
30. Edwards, L., & Veale, M. (2018). Enslaving the algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?. IEEE Security & Privacy, 16(3), 46-54.
31. Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, *17*(2), 124.
32. Ekman, Paul, and Wallace V. Friesen. "A new pan-cultural facial expression of emotion." *Motivation and emotion* 10.2 (1986): 159-168.
33. Ekman, R., What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA, 1997.
34. El Hammoumi, O., Benmarrakchi, F., Ouherrou, N., El Kafi, J., & El Hore, A. (2018). Emotion Recognition in E-learning Systems. *2018 6th International Conference on Multimedia Computing and Systems (ICMCS), 2018*, 1-6.
35. Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, *128*(2), 203.
36. Elmer, T., Mepham, K., Stadtfeld, C. (2020) Students under lockdown: Comparisons of students' social networks and mental health before and during the COVID-19 crisis in Switzerland. PLOS ONE, 15(7), e0236337.
37. Eom, S. B., Wen, H. J., & Ashill, N. (2006). The determinants of students' perceived learning outcomes and satisfaction in university online education: An empirical investigation. *Decision Sciences Journal of Innovative Education*, *4*(2), 215-235.
38. Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, *1341*(3), 1.
39. Essadek, A., & Rabeyron, T. (2020). Mental health of French students during the Covid-19 pandemic. *Journal of Affective Disorders, 277*, 392-393.
40. Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, *111*(1), 98-136.
41. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2011). The pascal visual object classes challenge 2012 (voc2012) results (2012). In *URL http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html*

42. Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, *77*, 167-181.
43. Freitas, A. A. (2014). Comprehensible classification models: A position paper. *ACM SIGKDD Explorations Newsletter*, *15*(1), 1–10.
44. Fürnkranz, J., Gamberger, D., & Lavrač, N. (2012). Rule Learning in a Nutshell. In *Foundations of rule learning* (pp. 19-55). Springer Science & Business Media.
45. Gefen, Karahanna, & Straub. (2003). Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly*, *27*(1), 51. https://doi.org/10.2307/30036519
46. Giraud-Carrier, C. (1998). Beyond predictive accuracy: what. In Proceedings of the ECML-98 Workshop on Upgrading Learning to Meta-Level: Model Selection and Data Transformation (pp. 78-85).
47. Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., & Holzinger, A. (2018). Explainable AI: The new 42? *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 295–303.
48. Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Zhou, Y. (2013). Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing* (pp. 117-124).
49. Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., & Lee, D.-H. (2013). Challenges in representation learning: A report on three machine learning contests. *International Conference on Neural Information Processing*, 117–124.
50. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, *38*(3), 50-57.
51. Gregor, S. (2006). The nature of theory in information systems. *MIS quarterly*, 611-642.
52. Gressenbuch, L., & Bergemann, S. (2019, June 28). *Emotional awareness in autonomous driving — Challenges, Approaches & Vision.* Lecture presented at Seminar Emotional awareness in autonomous driving SS2019 in Technical University of Munich, Munich. Retrieved from https://www.in.tum.de/fileadmin/w00bws/i06/Teaching/SS19/EAAD_Group_B_presentation.pdf
53. Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face recognition vendor test part 3: Demographic effects*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8280
54. Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and Counterfactual Explanations for Black Box Decision Making. IEEE Intelligent Systems, 34(6), 14-23.
55. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1-42.

56. Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, *2*(2).
57. Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2125–2126. https://doi.org/10.1145/2939672.2945386
58. Hall, M., Harborne, D., Tomsett, R., Galetic, V., Quintana-Amate, S., Nottle, A., & Preece, A. (2019). A systematic method to understand requirements for explainable AI (XAI) systems. *Proceedings of the IJCAI Workshop on EXplainable Artificial Intelligence (XAI 2019), Macau, China*.
59. Hamester, D., Barros, P., & Wermter, S. (2015). Face expression recognition with a 2-channel convolutional neural network. In *2015 international joint conference on neural networks (IJCNN)* (pp. 1-8). IEEE.
60. Hasani, B., & Mahoor, M. H. (2017). Facial expression recognition using enhanced deep 3D convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 30–40.
61. Hasani, B., & Mahoor, M. H. (2017). Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 790-795). IEEE.
62. Hastie, T., Tibshirani, R., & Friedman, J. (2009). Additive models, trees, and related methods. In *The Elements of Statistical Learning* (pp. 295-336). Springer, New York, NY.
63. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
64. Heath, R. L., & Bryant, J. (2013). Human communication theory and research: Concepts, contexts, and challenges. Routledge.
65. Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, *105*, 105–120. https://doi.org/10.1016/j.techfore.2015.12.014
66. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75-105.
67. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.
68. Hohman, F., Head, A., Caruana, R., DeLine, R., & Drucker, S. M. (2019). Gamut: A design probe to understand how data scientists understand machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.
69. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1312.

70. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv Preprint ArXiv:1704.04861*.
71. ISO/IEC. (2015). ISO/IEC 25023:2016 Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of system and software product quality. https://www.iso.org/standard/35747.html
72. Izquierdo-Reyes, J., Ramirez-Mendoza, R. A., Bustamante-Bello, M. R., Pons-Rovira, J. L., & Gonzalez-Vargas, J. E. (2018). Emotion recognition for semi-autonomous vehicles framework. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 12(4), 1447-1454.
73. Jacob, H., Kreifelts, B., Nizielski, S., Schütz, A., & Wildgruber, D. (2016). Effects of emotional intelligence on the impression of irony created by the mismatch between verbal and nonverbal cues. *PloS one*, *11*(10), e0163211.
74. Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data–recommendations for the use of performance metrics. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 245–251.
75. Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221–231.
76. Jiang, R. (2020). Knowledge, attitudes and mental health of university students during the COVID-19 pandemic in China. *Children and Youth Services Review, 119*, Children and youth services review, December 2020, Vol.119.
77. Jung, H., Lee, S., Yim, J., Park, S., & Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 2983-2991).
78. Kahneman, D., & Tversky, A. (1981). *The simulation heuristic*. Stanford Univ Ca Dept Of Psychology.
79. Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., & Boulanger-Lewandowski, N. (2016). Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, *10*(2), 99–111.
80. Kanou, S. E., Ferrari, R. C., Mirza, M., Jean, S., Carrier, P.-L., Dauphin, Y., Boulanger-Lewandowski, N., Aggarwal, A., Zumer, J., Lamblin, P., Raymond, J.-P., Pal, C., Desjardins, G., Pascanu, R., Warde-Farley, D., Torabi, A., Sharma, A., Bengio, E., Konda, K. R., … Bengio, Y. (2013). Combining modality specific deep neural networks for emotion recognition in video. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction - ICMI '13*, 543–550. https://doi.org/10.1145/2522848.2531745
81. Kecojevic, A., Basch, C. H., Sullivan, M., & Davi, N. K. (2020). The impact of the COVID-19 epidemic on mental health of undergraduate students in New Jersey, cross-sectional study. PLOS ONE, 15(9), e0239696. https://doi.org/10.1371/journal.pone.0239696

82. Khan, A., Sultana, M., Hossain, S., Hasan, M., Ahmed, H., & Sikder, M. (2020). The impact of COVID-19 pandemic on mental health & wellbeing among home-quarantined Bangladeshi students: A cross-sectional pilot study. *Journal of Affective Disorders, 277*, 121-128.
83. Kim, B., Seo, J., Jeon, S., Koo, J., Choe, J., & Jeon, T. (2019). Why are saliency maps noisy? Cause of and solution to noisy saliency maps. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 4149–4157.
84. Kim, D. H., Baddar, W. J., Jang, J., & Ro, Y. M. (2019). Multi-Objective Based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition. *IEEE Transactions on Affective Computing*, *10*(2), 223–236. https://doi.org/10.1109/TAFFC.2017.2695999
85. Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *sensors*, *18*(2), 401.
86. Kocabey, E., Camurcu, M., Ofli, F., Aytar, Y., Marin, J., Torralba, A., & Weber, I. (2017). Face-to-BMI: Using Computer Vision to Infer Body Mass Index on Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), Article 1. https://ojs.aaai.org/index.php/ICWSM/article/view/14923
87. Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110
88. Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, 3–10.
89. Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S., & Doshi-Velez, F. (2019). An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*.
90. Lapuschkin, S., Binder, A., Montavon, G., Muller, K. R., & Samek, W. (2016). Analyzing classifiers: Fisher vectors and deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2912-2920).
91. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, *10*(1), 1-8.
92. LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K. R. (2012). Efficient backprop. In *Neural networks: Tricks of the trade* (pp. 9-48). Springer, Berlin, Heidelberg.
93. Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
94. Lethbridge, T. C., Sim, S. E., & Singer, J. (2005). Studying software engineers: Data collection techniques for software field studies. *Empirical software engineering*, *10*(3), 311-341.

95. Li, J., Wang, Y., See, J., & Liu, W. (2019). Micro-expression recognition based on 3D flow convolutional neural network. *Pattern Analysis and Applications*, *22*(4), 1331–1339. https://doi.org/10.1007/s10044-018-0757-5
96. Li, S., & Deng, W. (2018). Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, *28*(1), 356–370.
97. Li, S., & Deng, W. (2020). A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing*.
98. Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2852–2861.
99. Li, X., Hess, T. J., & Valacich, J. S. (2008). Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems*, *17*(1), 39–71.
100. Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.
101. Linnenbrink-Garcia, L., & Pekrun, R. (2011). Students' emotions and academic engagement: Introduction to the special issue. *Contemporary Educational Psychology, 36*(1), 1-3.
102. Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, *27*, 247-266.
103. Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, *16*(3), 31-57.
104. Lopes, A., De Aguiar, E., De Souza, A., & Oliveira-Santos, T. (2017). Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. *Pattern Recognition, 61*, 610-628.
105. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). 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). IEEE.
106. Lui, A., & Lamb, G. W. (2018). Artificial intelligence and augmented intelligence collaboration: regaining trust and confidence in the financial sector. *Information & Communications Technology Law*, *27*(3), 267-283.
107. Lum, K., & Isaac, W. (2016). To predict and serve?. *Significance*, *13*(5), 14-19.
108. Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *ArXiv Preprint ArXiv:1802.03888*.
109. Lundberg, S., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.
110. Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding facial expressions with gabor wavelets. *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 200–205.
111. Mahendran, A., & Vedaldi, A. (2016). Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, *120*(3), 233-255.

112. Maresh, M. M. (2007). Hurt feelings in the classroom: Facework in student communicative responses to college instructors' hurtful messages. In *annual meeting of the National Communication Association, Chicago, IL*.
113. Matsumoto, D. (2001). Culture and emotion. *The handbook of culture and psychology*, 171-194.
114. Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, *6*(3), 175-183.
115. Mazer, J., Mckenna-Buchanan, T., Quinlan, M., & Titsworth, S. (2014). The Dark Side of Emotion in the Classroom: Emotional Processes as Mediators of Teacher Communication Behaviors and Student Negative Emotions. *Communication Education, 63*(3), 149-168.
116. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.
117. Mena-Chalco, J. P., Velho, L., & Junior, R. C. (2011). 3D human face reconstruction using principal components spaces. *Proceedings of WTD SIBGRAPI Conference on Graphics, Patterns and Images*, *6*.
118. Meng, D., Peng, X., Wang, K., & Qiao, Y. (2019). Frame attention networks for facial expression recognition in videos. *ArXiv:1907.00193 [Cs]*. http://arxiv.org/abs/1907.00193
119. Miller, T. (2018). Contrastive explanation: A structural-model approach. *arXiv preprint arXiv:1811.03163*.
120. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.
121. Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*.
122. Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2018). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*.
123. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 205395171667967. https://doi.org/10.1177/2053951716679679
124. Mollahosseini, A., Chan, D., & Mahoor, M. (2016). Going deeper in facial expression recognition using deep neural networks. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV),* 1-10.
125. Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18-31.
126. Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.
127. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. R. (2019). Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, 193-209.

128. Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011
129. Mordvintsev, A., Olah, C., & Tyka, M. (2015). Inceptionism: Going deeper into neural networks.
130. Neter, J., Wasserman, W., & Kutner, M. H. (1989). Applied linear regression models.
131. Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175-220.
132. Nielson, K. A., & Lorber, W. (2009). Enhanced post-learning memory consolidation is influenced by arousal predisposition and emotion regulation but not by stimulus valence or arousal. *Neurobiology of learning and memory*, *92*(1), 70-79.
133. Ochs, M., Niewiadomski, R., & Pelachaud, C. (2015). 18 Facial Expressions of Emotions for Virtual Characters. In *The Oxford Handbook of Affective Computing* (p. 261). Oxford University Press, USA.
134. Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, 13.
135. Osoba, O., & Welser, W. (2017). An intelligence in our image: The risks of bias and errors in artificial intelligence. RAND Corporation.
136. Pantic, M., Valstar, M., Rademaker, R., & Maat, L. (2005). Web-based database for facial expression analysis. *2005 IEEE International Conference on Multimedia and Expo*, 5-pp.
137. Papernot, N., & McDaniel, P. (2018). Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*.
138. Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational psychology review*, *18*(4), 315-341.
139. Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary educational psychology*, *36*(1), 36-48.
140. Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational psychologist*, *37*(2), 91-105.
141. Perez, A. (2018, August 30). *Recognizing human facial expressions with machine learning*. ThoughtWorks. https://www.thoughtworks.com/insights/articles/recognizing-human-facial-expressions-machine-learning
142. Pfungst, O. (1911). Clever Hans:(the horse of Mr. Von Osten.) a contribution to experimental animal and human psychology. Holt, Rinehart and Winston.
143. Pieters, W. (2011). Explanation and trust: What to tell the user in security and AI? *Ethics and Information Technology*, *13*(1), 53–64.

144. Podgorelec, V., Pečnik, Š., & Vrbančič, G. (2020). Classification of Similar Sports Images Using Convolutional Neural Network with Hyper-Parameter Optimization. *Applied Sciences*, *10*(23), 8494.
145. Ponce, J., Berg, T. L., Everingham, M., Forsyth, D. A., Hebert, M., Lazebnik, S., Marszalek, M., Schmid, C., Russell, B. C., & Torralba, A. (2006). Dataset issues in object recognition. In *Toward category-level object recognition* (pp. 29–48). Springer.
146. Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). Manipulating and measuring model interpretability. *ArXiv Preprint ArXiv:1802.07810*.
147. Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137-141.
148. Reddy, S., Fox, J., & Purohit, M. P. (2019). Artificial intelligence-enabled healthcare delivery. *Journal of the Royal Society of Medicine*, *112*(1), 22–28. https://doi.org/10.1177/0141076818815510
149. Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of personality and social psychology*, *49*(1), 95.
150. Revina, I. M., & Emmanuel, W. S. (2018). A survey on human face expression recognition techniques. *Journal of King Saud University-Computer and Information Sciences*.
151. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). ‘ Why should I trust you?’ Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
152. Rothman, D. (2020). Hands-On Explainable AI (XAI) with Python. Packt Publishing.
153. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.
154. Samek, W., & Müller, K.-R. (2019). Towards explainable artificial intelligence. In *Explainable AI: interpreting, explaining and visualizing deep learning* (pp. 5–22). Springer.
155. Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proceedings of the IEEE*, *109*(3), 247-278.
156. Samek, W., Wiegand, T., & Müller, K. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models.
157. Saxena, A., Khanna, A., & Gupta, D. (2020). Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, *2*(1), 53-79.
158. Serengil, S. I., & Ozpinar, A. (2020). LightFace: A Hybrid Deep Face Recognition Framework. *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 1–5.

159. Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, *27*(6), 803-816.
160. Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, *2*(28), 307-317.
161. Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *International Conference on Machine Learning*, 3145–3153.
162. Shrikumar, A., Greenside, P., Shcherbina, A., & Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.
163. Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
164. Singh, S., & Benedict, S. (2019). Indian Semi-Acted Facial Expression (iSAFE) Dataset for Human Emotions Recognition. *International Symposium on Signal Processing and Intelligent Recognition Systems*, 150–162.
165. Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2009). A Motivational Perspective on Engagement and Disaffection: Conceptualization and Assessment of Children's Behavioral and Emotional Participation in Academic Activities in the Classroom. *Educational and Psychological Measurement*, *69*(3), 493–525.
166. Skinner, E., Furrer, C., Marchand, G., & Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic?. *Journal of educational psychology*, *100*(4), 765.
167. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186.
168. Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.
169. Sourdin, T. (2018). Judge v. robot: Artificial intelligence and judicial decision-making. University of New South Wales Law Journal, 41(4), 1114-1133.
170. Štrumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, *11*, 1–18.
171. Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665.
172. Sun, A., Li, Y., Huang, Y.-M., & Li, Q. (2018). The exploration of facial expression recognition in distance education learning system. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11003*, 111-121.
173. Sun, B., Wei, Q., Li, L., Xu, Q., He, J., & Yu, L. (2016). LSTM for dynamic emotion and group emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 451-457).

174. Sun, X., Xia, P., & Ren, F. (2020). Multi-attention based Deep Neural Network with hybrid features for Dynamic Sequential Facial Expression Recognition. *Neurocomputing*.
175. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*, 3319–3328.
176. Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.
177. Suresh, H., Gong, J. J., & Guttag, J. V. (2018). Learning tasks for multitask learning: Heterogenous patient populations in the icu. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 802–810.
178. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.
179. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.
180. Taylor, S. (2019). The Psychology of Pandemics: Preparing for the Next Global Outbreak of Infectious Disease.
181. Titsworth, S., McKenna, T. P., Mazer, J. P., & Quinlan, M. M. (2013). The bright side of emotion in the classroom: Do teachers' behaviors predict students' enjoyment, hope, and pride?. *Communication Education*, *62*(2), 191-209.
182. Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*.
183. Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *CVPR 2011*, 1521–1528.
184. Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2, pp. 131-160).
185. Um, E., Plass, J. L., Hayward, E. O., & Homer, B. D. (2012). Emotional design in multimedia learning. *Journal of educational psychology*, *104*(2), 485.
186. Van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the national conference on artificial intelligence* (pp. 900-907). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.
187. Vladeck, D. C. (2014). Machines without principals: liability rules and artificial intelligence. *Wash. L. Rev.*, *89*, 117.
188. Vogel, S., & Schwabe, L. (2016). Learning and memory under stress: implications for the classroom. *npj Science of Learning*, *1*(1), 1-10.
189. Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*(4), 395-416.
190. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology (Harvard JOLT), 31(2),* 841-888.

191. Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an information system design theory for vigilant EIS. *Information systems research*, *3*(1), 36-59.
192. Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, *5*(3), 457-469.
193. Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, *114*(2), 246–257. https://doi.org/10.1037/pspa0000098
194. Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2020). "Let me explain!": exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces*, 1-12.
195. Williams, K., Childers, C., & Kemp, E. (2013). Stimulating and Enhancing Student Learning Through Positive Emotions. *Journal of Teaching in Travel & Tourism,13*(3), 209-227.
196. Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., & Hua, G. (2017). *Neural Aggregation Network for Video Face Recognition*. 4362–4371. https://openaccess.thecvf.com/content_cvpr_2017/html/Yang_Neural_Aggregation_Network_CVPR_2017_paper.html
197. Yerushalmy, J. (1947). Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques. *Public Health Reports (1896-1970), 62*(40), 1432-1449. doi:10.2307/4586294
198. Yu, Z., Liu, Q., & Liu, G. (2018). Deeper cascaded peak-piloted network for weak expression recognition. *The Visual Computer*, *34*(12), 1691–1699. https://doi.org/10.1007/s00371-017-1443-0
199. Yukselturk, E., & Bulut, S. (2007). Predictors for student success in an online course. *Journal of Educational Technology & Society*, *10*(2), 71-83.
200. Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., & Dobaie, A. (2018). Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing (Amsterdam), 273*, 643-649.
201. Zhang, L., Jiang, M., Farid, D., & Hossain, M. A. (2013). Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. *Expert Systems with Applications*, *40*(13), 5160-5168.
202. Zhang, S., Pan, X., Cui, Y., Zhao, X., & Liu, L. (2019). Learning affective video features for facial expression recognition via hybrid deep learning. *IEEE Access*, *7*, 32297-32304.
203. Zhang, Z., Li, Z., Liu, H., Cao, T., & Liu, S. (2020). Data-driven Online Learning Engagement Detection via Facial Expression and Mouse Behavior Recognition Technology. *Journal of Educational Computing Research,58*(1), 63-86.
204. Zhao, G., & Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, *29*(6), 915-928.
205. Zhao, X., Liang, X., Liu, L., Li, T., Han, Y., Vasconcelos, N., & Yan, S. (2016). Peak-Piloted Deep Network for Facial Expression Recognition. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 425–

442). Springer International Publishing. https://doi.org/10.1007/978-3-319-46475-6_27

206. Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018). Employee turnover prediction with machine learning: A reliable approach. *Proceedings of SAI Intelligent Systems Conference*, 737–758.

207. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2921–2929.

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I Ruslan Kononov

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis “Evaluation of Facial Emotion Recognition Models for the Potential Deployment in Web-Based Learning Environments”, supervised by Sadok Ben Yahia and Silvia Lips
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

06.05.2021

1 The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 – Performance Metrics of the Tested Models

```
Accuracy:
0.7124547227640011


Report:
              precision    recall  f1-score   support

       Anger       0.64      0.64      0.64       491
     Disgust       0.73      0.65      0.69        55
        Fear       0.58      0.49      0.53       528
  Happinness       0.88      0.91      0.89       879
     Sadness       0.58      0.60      0.59       594
    Surprise       0.81      0.80      0.80       416
     Neutral       0.68      0.74      0.71       626

    accuracy                           0.71      3589
   macro avg       0.70      0.69      0.69      3589
```

Figure 29. Performance of Model 1 (the FER-2013 test set).

```
Accuracy:
0.6386179994427417


Report:
              precision    recall  f1-score   support

       Anger       0.56      0.55      0.56       491
     Disgust       0.74      0.36      0.49        55
        Fear       0.49      0.38      0.43       528
  Happinness       0.85      0.89      0.87       879
     Sadness       0.47      0.50      0.48       594
    Surprise       0.76      0.75      0.75       416
     Neutral       0.58      0.66      0.62       626

    accuracy                           0.64      3589
   macro avg       0.64      0.58      0.60      3589
```

Figure 30. Performance of Model 2 (the FER-2013 test set).

```
Accuracy:
0.7695567144719687


Report:
              precision    recall  f1-score   support

    Surprise       0.93      0.71      0.81       329
        Fear       0.47      0.69      0.56        74
     Disgust       0.69      0.31      0.42       160
   Happiness       0.96      0.82      0.89      1185
     Sadness       0.55      0.96      0.70       478
       Anger       0.67      0.79      0.72       162
     Neutral       0.78      0.69      0.73       680

    accuracy                           0.77      3068
   macro avg       0.72      0.71      0.69      3068
```

Figure 31. Performance of Model 3 (the RAF-DB test set).

Appendix 3 – Ethnic- and Gender-related Statistics at Yale University

Table 7. Total enrolment to Yale University by race in Fall 2019. The data does not include international students[1].

Total University Enrollments* (% of non-International):		**International Students:**
Black or African American:	7.7%	
American Indian/Alaska Native:	0.4%	International Students: 21%
Asian:	19.3%	Countries Represented: 120
Native Hawaiian or other Pacific Islander:	0.1%	
Hispanic of any race:	13.3%	Countries most represented: China,
White:	52.7%	Canada, India, South Korea, United Kingdom,
Two or more:	6.5%	and Germany.
Race/ethnicity unknown:	1.0%	

Table 8. Total enrolment to Yale University by sex in Fall 2019. The data excludes international students[1].

Degrees conferred: (Between July 1, 2018 and June 30, 2019)	**Men**	**Women**	**Total**	**% International**
Bachelors:	724	683	1,407	11%
Masters & Post-Masters Certificates:	1,237	1,414	2,651	34%
Research & Scholarship:	242	188	430	31%
Professional Practice:	157	148	305	9%
Doctorates:	399	336	735	20%
Total:	**2,360**	**2,433**	**4,793**	**21%**

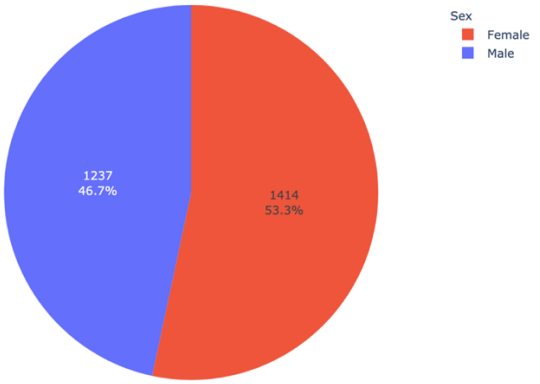

Figure 32. Distribution of people with conferred degrees from Yale University between July 2018 and June 2019 by sex (non-Doctorates). The graph illustrates the numbers indicted in Table 3.

[1] Yale University / The Office of Institutional Research (OIR). (2019). *2019-2020 Factsheet*. https://www.yale.edu/about-yale/yale-facts

Appendix 4 – Predicted Racial Distribution in FER-2013

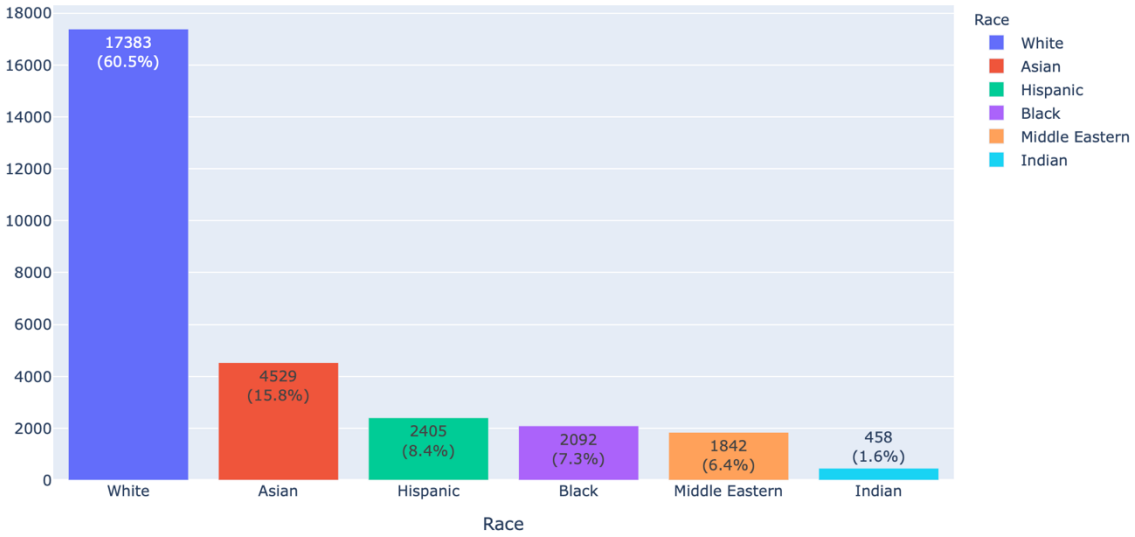

Figure 33. Predicted racial distribution of the data subjects in FER-2013 before aggregation.

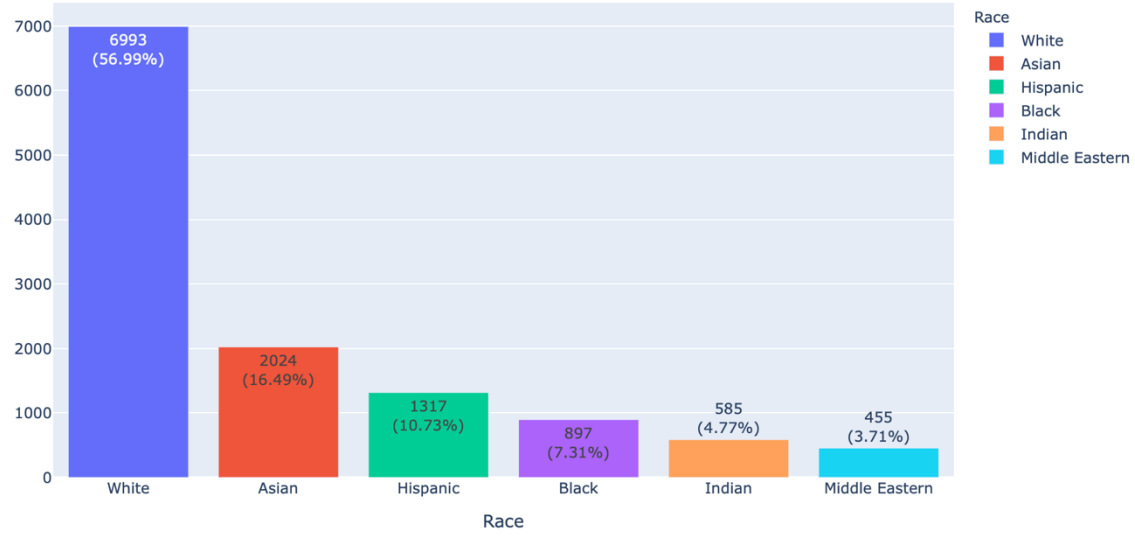

Figure 34. Predicted racial distribution of the data subjects in RAF-DB before aggregation.

Appendix 5 – List of Independent Annotators

Table 9. Independent evaluators of gender-related annotations for the RAF-DB sampled images.

Annotator's name	**Contact info**
Halyna Kiryk	halynaskiryk@gmail.com
Hanna Khudyk	anna.hudyk@gmail.com
Valentyna Tsap	valentyna.tsap@taltech.ee
Yevdokymova Oleksandra	sevdokimova99@gmail.com

Appendix 6 – Accuracy of the Pre-selected Models on Different Population Groups

Accuracy:
0.5092073658927142

Report:

	precision	recall	f1-score	support
Anger	0.10	0.13	0.11	138
Disgust	0.00	0.00	0.00	39
Fear	0.09	0.10	0.09	69
Happinness	0.89	0.79	0.84	429
Sadness	0.43	0.70	0.53	147
Surprise	0.00	0.00	0.00	114
Neutral	0.59	0.53	0.56	313
accuracy			0.51	1249
macro avg	0.30	0.32	0.30	1249
weighted avg	0.52	0.51	0.51	1249

Accuracy:
0.5981481481481481

Report:

	precision	recall	f1-score	support
Anger	0.08	0.06	0.06	159
Disgust	0.04	0.03	0.03	35
Fear	0.05	0.06	0.05	89
Happinness	0.94	0.83	0.89	712
Sadness	0.47	0.73	0.57	239
Surprise	0.00	0.00	0.00	45
Neutral	0.63	0.55	0.58	341
accuracy			0.60	1620
macro avg	0.32	0.32	0.31	1620
weighted avg	0.63	0.60	0.61	1620

Figure 35. Performance of Model 1 for the different gender groups – males (on the left) and females (on the right).

Accuracy:
0.44595676541232987

Report:

	precision	recall	f1-score	support
Anger	0.07	0.09	0.08	138
Disgust	0.00	0.00	0.00	39
Fear	0.04	0.04	0.04	69
Happinness	0.77	0.75	0.76	429
Sadness	0.30	0.52	0.38	147
Surprise	0.00	0.00	0.00	114
Neutral	0.55	0.46	0.50	313
accuracy			0.45	1249
macro avg	0.25	0.27	0.25	1249
weighted avg	0.45	0.45	0.44	1249

Accuracy:
0.5679012345679012

Report:

	precision	recall	f1-score	support
Anger	0.12	0.08	0.10	159
Disgust	0.00	0.00	0.00	35
Fear	0.03	0.02	0.02	89
Happinness	0.86	0.81	0.83	712
Sadness	0.39	0.58	0.47	239
Surprise	0.00	0.00	0.00	45
Neutral	0.55	0.56	0.56	341
accuracy			0.57	1620
macro avg	0.28	0.29	0.28	1620
weighted avg	0.57	0.57	0.56	1620

Figure 36. Performance of Model 2 for the different gender groups – males (on the left) and females (on the right).

Accuracy:
0.7405924739791834

Report:

	precision	recall	f1-score	support
Surprise	0.90	0.62	0.74	138
Fear	0.49	0.69	0.57	39
Disgust	0.63	0.28	0.38	69
Happiness	0.96	0.79	0.87	429
Sadness	0.47	0.93	0.63	147
Anger	0.71	0.82	0.76	114
Neutral	0.76	0.72	0.74	313
accuracy			0.74	1249
macro avg	0.70	0.69	0.67	1249
weighted avg	0.79	0.74	0.75	1249

Accuracy:
0.7820987654320988

Report:

	precision	recall	f1-score	support
Surprise	0.98	0.75	0.85	159
Fear	0.44	0.69	0.54	35
Disgust	0.72	0.33	0.45	89
Happiness	0.97	0.84	0.90	712
Sadness	0.52	0.97	0.68	239
Anger	0.58	0.71	0.64	45
Neutral	0.81	0.68	0.74	341
accuracy			0.78	1620
macro avg	0.72	0.71	0.69	1620
weighted avg	0.83	0.78	0.79	1620

Figure 37. Performance of Model 3 for the different gender groups – males (on the left) and females (on the right).

Appendix 7 – The Top Identifiable Facial Expressions by Each Model

Table 10. The top three best and worst identifiable facial expressions by each model for a respective dataset.

Dataset / Model	**FER-2013**	**RAF-DB**	**CK**	**JAFFE**	**iSAFE**	**FacesDB**	**AffectNet**
Model 1	Happiness, Sadness, Neutral --- Fear, Sadness, Anger	Happiness, Sadness, Neutral --- Surprise, Disgust, Fear	Surprise, Happiness, Neutral --- Anger, Fear, Sadness	Happiness, Surprise, Fear --- Disgust, Anger, Sadness	Happiness, Sadness, Neutral --- Surprise, Disgust, Fear	Happiness, Neutral, Sadness --- Surprise, Disgust, Anger	Happiness, Sadness, Neutral --- Disgust, Surprise, Anger
Model 2	Happiness, Surprise, Neutral --- Fear, Sadness, Disgust	Happiness, Neutral, Sadness --- Disgust, Surprise, Fear	Happiness, Neutral, Surprise --- Disgust, Fear, Sadness	Surprise, Happiness, Sadness --- Disgust, Neutral, Anger	Happiness, Neutral, Sadness --- Disgust, Surprise, Fear	Happiness, Neutral, Sadness --- Anger, Surprise Fear	Happiness, Neutral, Sadness --- Disgust, Surprise, Anger
Model 3	Happiness, Surprise, Sadness --- Fear, Disgust, Anger	Happiness, Surprise, Neutral --- Disgust, Fear, Sadness	Happiness, Surprise, Neutral --- Anger, Sadness Fear	Happiness, Fear, Sadness --- Disgust, Anger, Neutral	Happiness, Sadness, Surprise --- Fear, Neutral, Anger	Happiness, Neutral, Fear --- Anger, Disgust, Surprise	Happiness, Sadness, Neutral --- Anger, Disgust, Fear

Appendix 8 – Universal Facial Expressions

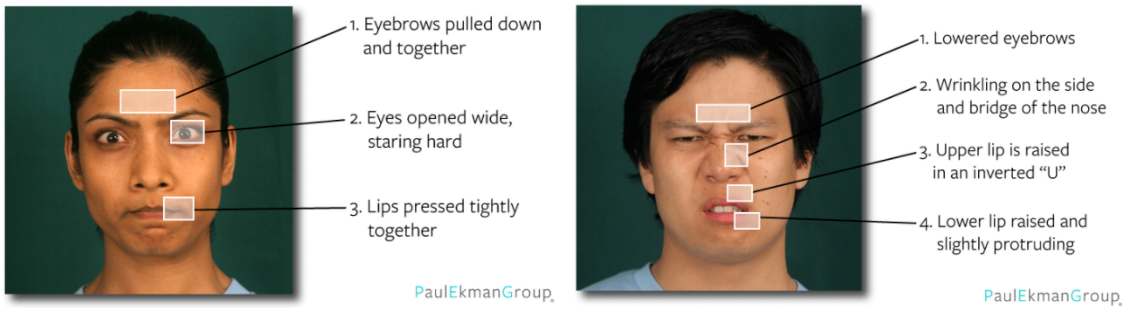

Figure 38. The face of anger (on the left) and the face of disgust (on the right)[1].

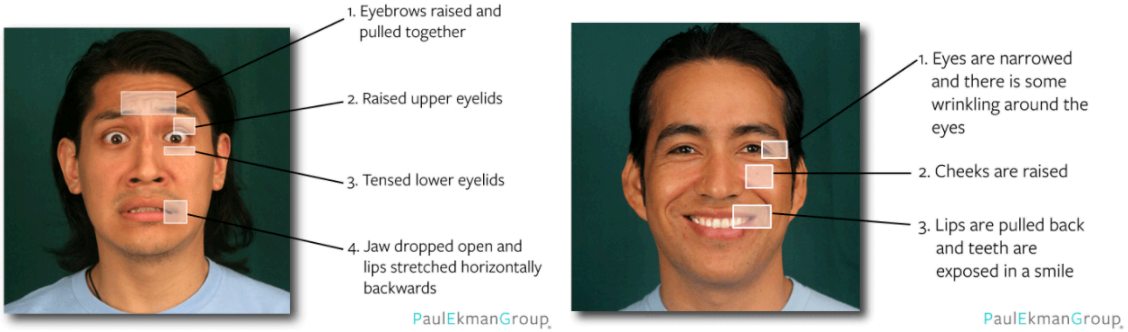

Figure 39. The face of fear (on the left) and the face of happiness (on the right)[1].

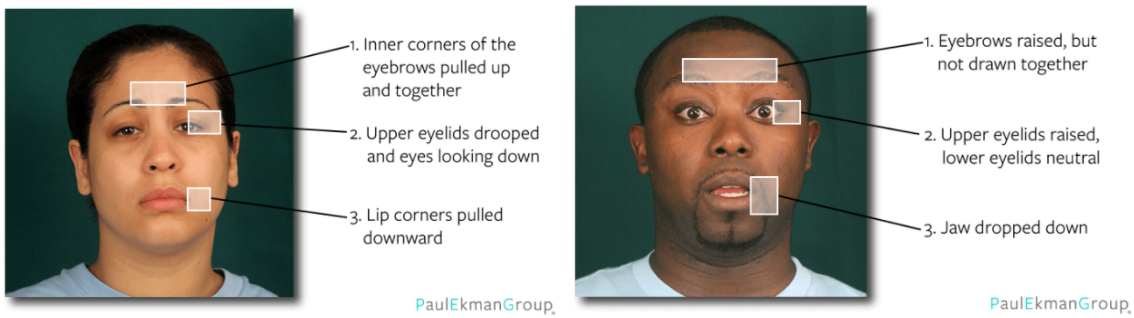

Figure 40. The face of sadness (on the left) and the face of surprise (on the right)[1].

[1] Available at https://www.paulekman.com/universal-emotions/

Appendix 9 – Attribution maps (Model 1)

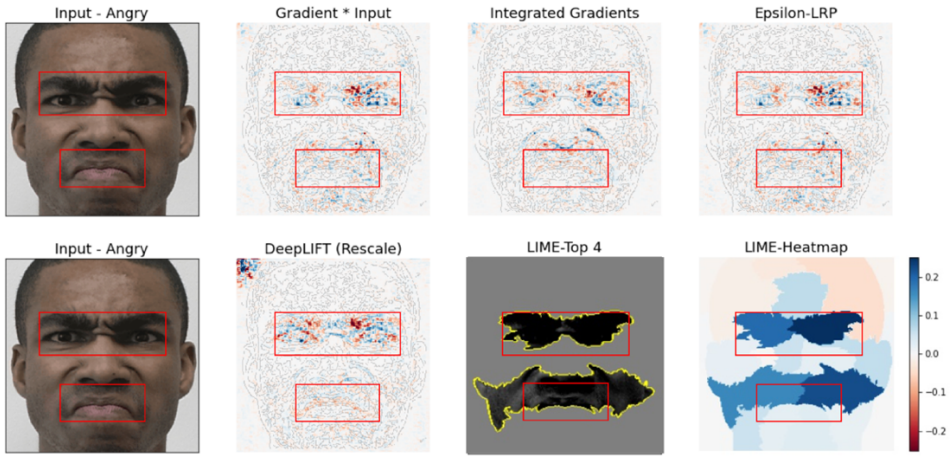

Figure 41. Attribution maps generated for the input image of the class "Angry" (Model 1).

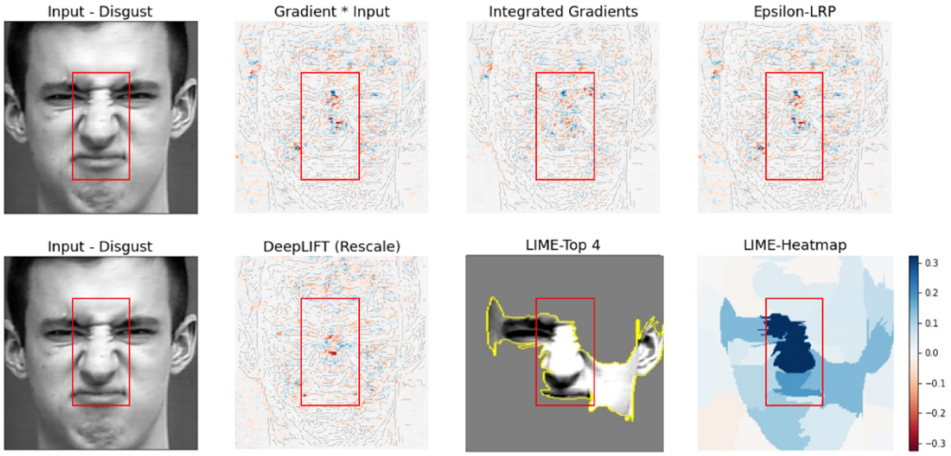

Figure 42. Attribution maps generated for the input image of the class "Disgust" (Model 1).

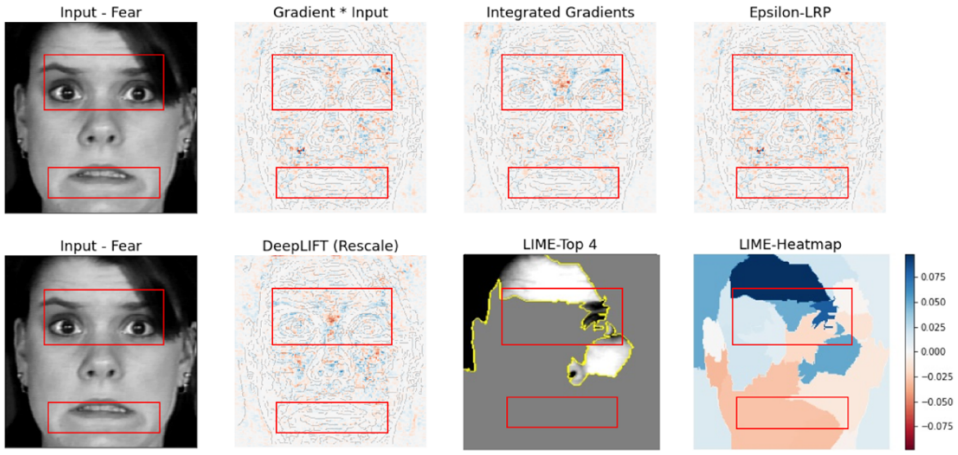

Figure 43. Attribution maps generated for the input image of the class "Fear" (Model 1).

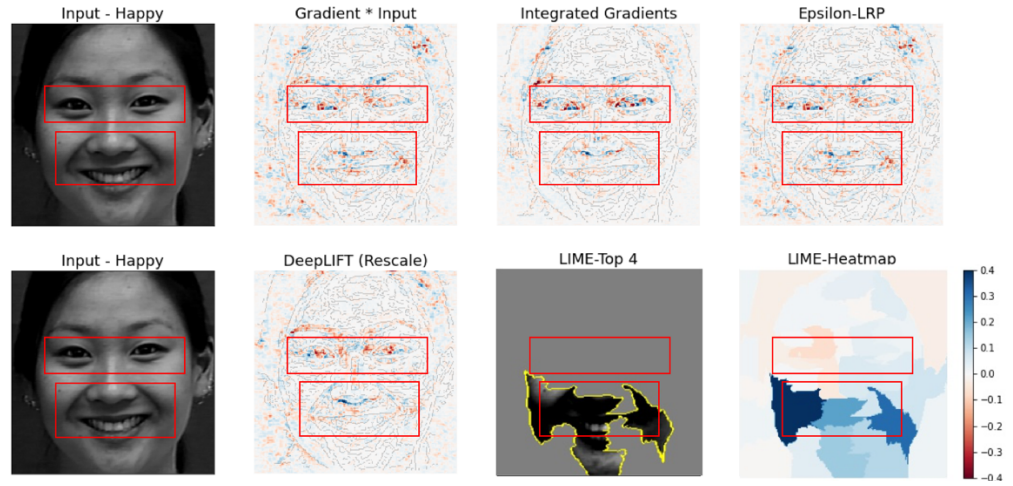

Figure 44. Attribution maps generated for the input image of the class “Happiness” (Model 1).

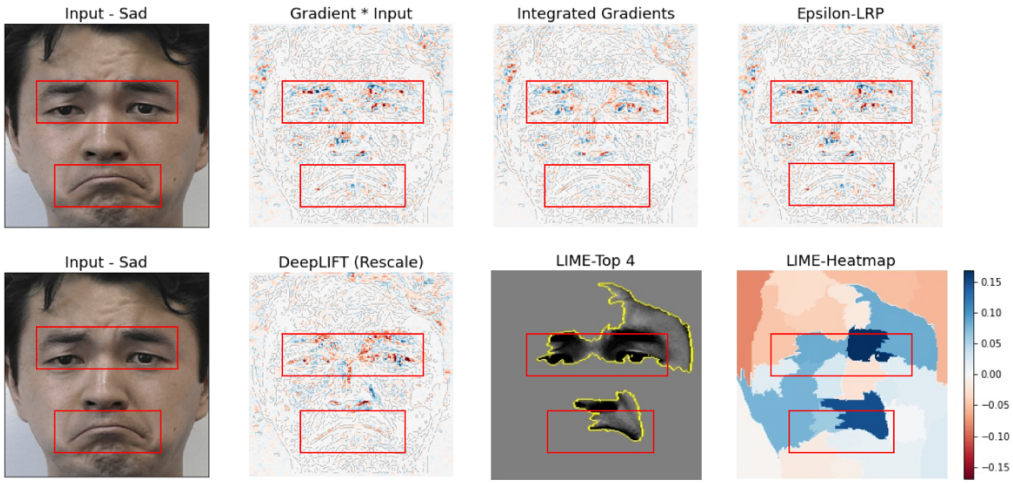

Figure 45. Attribution maps generated for the input image of the class “Sadness” (Model 1).

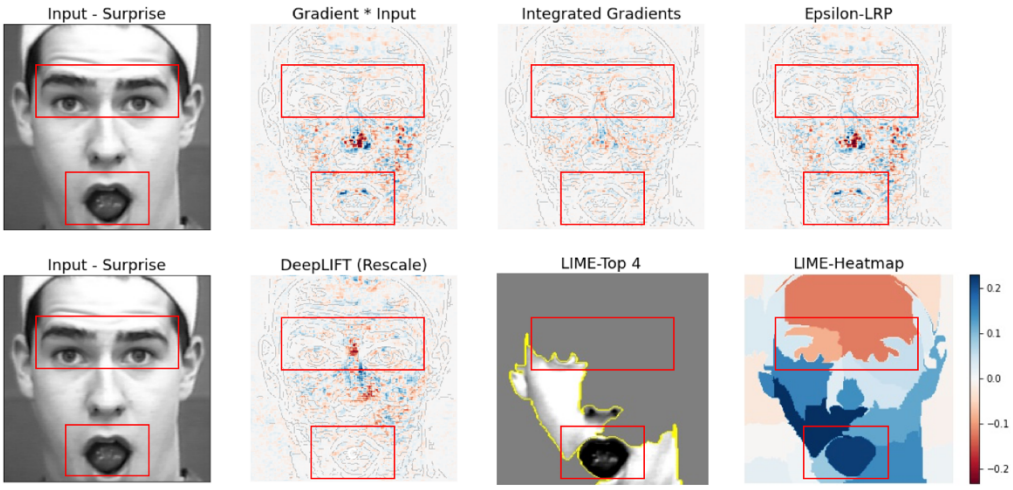

Figure 46. Attribution maps generated for the input image of the class “Surprise” (Model 1).

Appendix 10 – Attribution maps (Model 3)

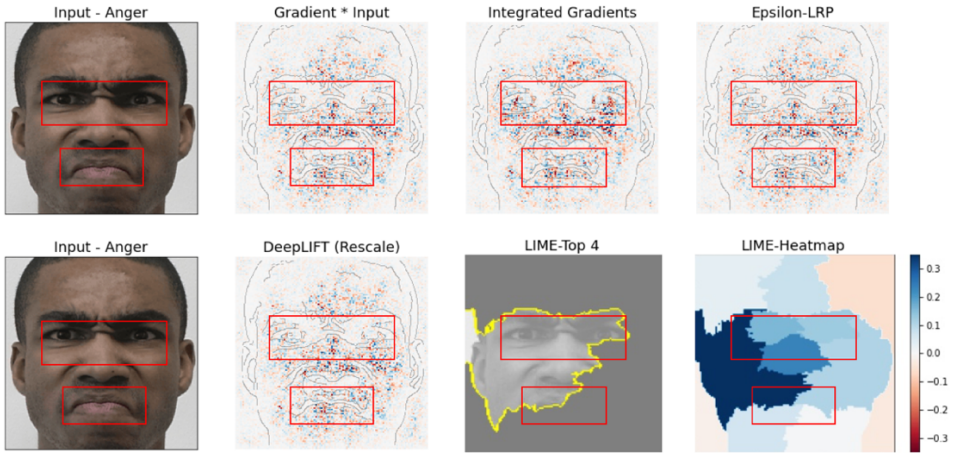

Figure 47. Attribution maps generated for the input image of the class “Angry” (Model 3).

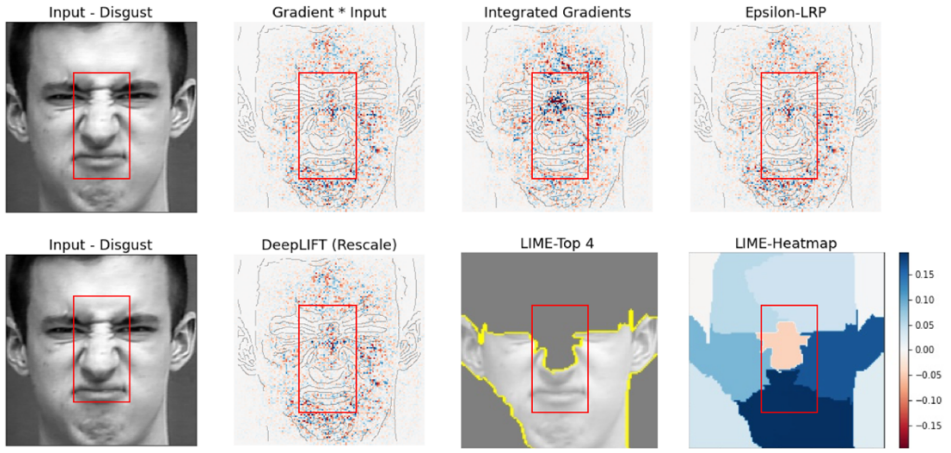

Figure 48. Attribution maps generated for the input image of the class “Disgust” (Model 3).

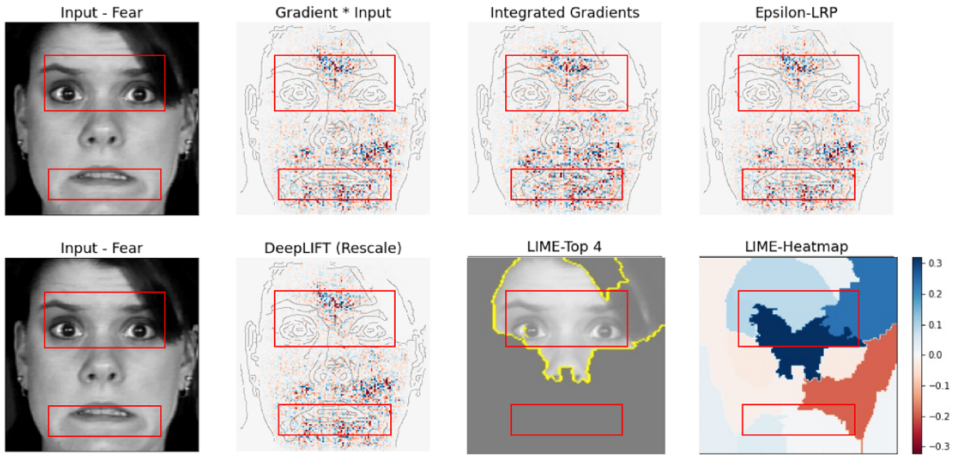

Figure 49. Attribution maps generated for the input image of the class “Fear” (Model 3).

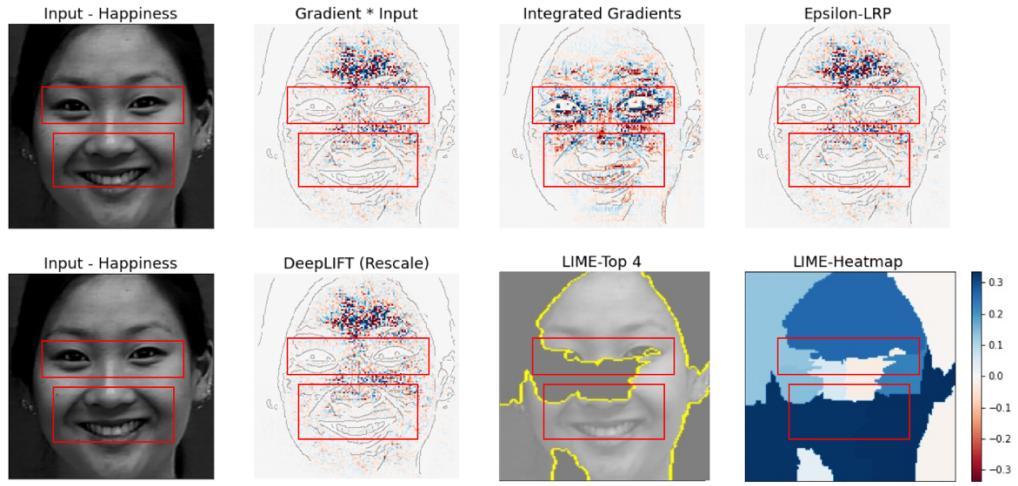

Figure 50. Attribution maps generated for the input image of the class “Happiness” (Model 3).

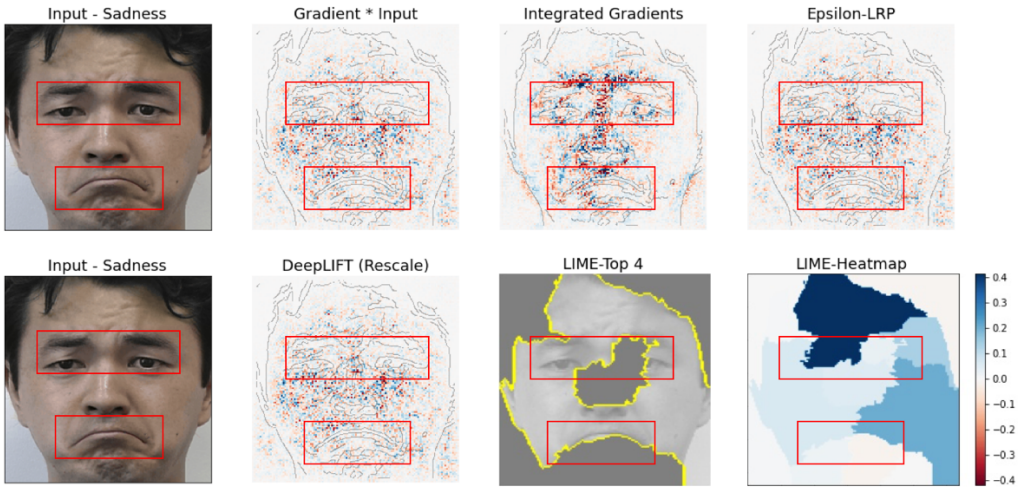

Figure 51. Attribution maps generated for the input image of the class “Sadness” (Model 3).

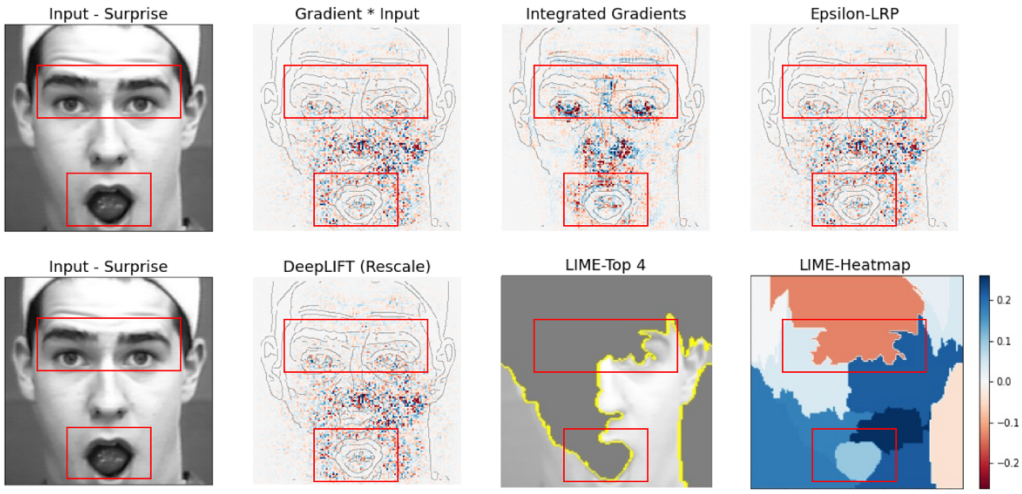

Figure 52. Attribution maps generated for the input image of the class “Surprise” (Model 3).