Facial Expression Recognition System for Stress Detection with Deep Learning

José Almeida[1] and Fátima Rodrigues[2][a]
[1]Polytechnic of Porto, School of Engineering, Rua Dr. António Bernardino de Almeida, Porto, Portugal
[2]Interdisciplinary Studies Research Center (ISRC), Polytechnic of Porto, School of Engineering, Porto, Portugal

Keywords: Stress Detection, Emotion, Facial Expression Classification, Convolutional Neural Networks.

Abstract: Stress is the body's natural reaction to external and internal stimuli. Despite being something natural, prolonged exposure to stressors can contribute to serious health problems. These reactions are reflected not only physiologically, but also psychologically, translating into emotions and facial expressions. Based on this, we developed a proof of concept for a stress detector. With a convolutional neural network capable of classifying facial expressions, and an application that uses this model to classify real-time images of the user's face and thereby assess the presence of signs of stress. For the creation of the classification model was used transfer learning together with fine-tuning. In this way, we took advantage of the pre-trained networks VGG16, VGG19, and Inception-ResNet V2 to solve the problem at hand. For the transfer learning process two classifier architectures were considered. After several experiments, it was determined that VGG16, together with a classifier based on a convolutional layer, was the candidate with the best performance at classifying stressful emotions. The results obtained are very promising and the proposed stress-detection system is non-invasive, only requiring a webcam to monitor the user's facial expressions.

1 INTRODUCTION

Demanding jobs are a significant cause of stress in people. Situations like frequent exposure to danger, short deadlines, rigorous tasks or even repetitive tasks are some stress originators.

Nearly one in three workers in Europe and the United States report that they are affected by stress at work. Work-related stress, depression, and anxiety can result in reduced work performance and absenteeism, costing an estimated 3% to 4% of gross national product (Dewa & Hoch, 2015). Also, about 61% of European institutions participating in the 2019 EU-OSHA study (EU-OSHA, 2019) reported that a reluctance to talk openly about these issues seems to be the main difficulty for addressing psychosocial risks.

There is evidence that stress conditions are both preventable and treatable in the workplace and that workers who receive treatment are more likely to be more productive (Carolan et al., 2017). Hence, non-intrusive stress sensing tools that continuously monitor stress levels, with a minimal impact on workers' daily lives, could be used to automatically initiate stress-reduction interventions. In stressful work settings, these applications could not only lead to more timely and reduced-cost interventions, but also to more productive environments where workers could better manage their workload.

1.1 Facial Expressions and Stress

Every day, people communicate with each other, not only verbally, but also with gestures and facial expressions. Often these gestures and facial expressions are automatic, and the transmitter does not even realize that he/she is executing them. This unintentional information is the primary way to know the transmitter's emotions in a non-invasive way.

There are facial expressions that express the same emotion universally. These are called universal facial expressions of emotion. There is ample consensus in the scientific community (Ekman, 2016) about the universality of five emotions: anger, disgust, fear, happiness and sadness. In this same study, scientists also agreed on the relationship between emotions and

[a] https://orcid.org/0000-0003-4950-7593

Almeida, J. and Rodrigues, F.
Facial Expression Recognition System for Stress Detection with Deep Learning.
DOI: 10.5220/0010474202560263
In *Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1*, pages 256-263
ISBN: 978-989-758-509-8

moods. Like other emotions and moods, stress can also be manifested from facial expressions, gestures, and even from the voice. Stress does not have an universal facial expression of emotion however, there are studies (Dinges et al., 2005), (Lerner et al., 2007) where the feeling of stress was validated by the increase in cortisol levels and cardiac activity, that confirm the relationship between facial expressions and stress. In these studies, the negative emotions anger, disgust, and fear were unquestionably related with stress.

1.2 State of the Art in Stress Detection

Automatic stress detection has been studied for many years. From some intrusive approaches, such as saliva or blood tests, to less intrusive approaches, with the collection of images.

One of those works are the study (Gao et al., 2014) where a camera mounted inside the dashboard of a car collected images of the driver face for the detection of stress. Using Support Vector Machines (SVMs) trained in public facial expression datasets, the collected images were then classified into one of six facial expressions. An algorithm would subsequently count the classifications within a time window. If the number of anger and disgust classifications exceeds a specific threshold, the driver would be considered under stress. The best classifier obtained was trained, not only with the images from the public datasets but also with images of the subjects posing for the stress classification. In this way, the models could adapt to the way these subjects show their facial expressions. They obtained an accuracy of 90.5% for the stress classification.

In another work (Maaoui et al., 2015), was developed a system that collects Remote Photoplethysmography (rPPG) signals using the computer's webcam for the detection of stress. The rPPG signals were then translated into a sinusoidal wave, which represents the heart rate. An SVM classifier presented the best results with 94.40% accuracy.

The group (Giannakakis et al., 2016) developed a system capable of detecting stress/anxiety emotional states through video-recorded facial cues. Applying a multitude of techniques, such as Active Appearance Models, Optical Flow, and rPPG they extracted the most relevant features that were used for the stress classification. The best classification accuracy of 87.72% was obtained with a K-NN classifier.

In the study (Viegas et al., 2018) was proposed a system capable of detecting signs of stress through Facial Action Units (FAUs) extracted from videos. They performed binary classification using several simple classifiers on FAUs extracted in each video frame and were able to achieve an accuracy of up to 74% in subject independent classification and 91% in subject dependent classification.

2 SOLUTION DESIGN

The system use case is to detect and notify the user that he shows signs of stress using facial expressions detected only with video images. The advantage of using video for personal stress detection is the easy accessibility to webcams when working with computers. The program will run in the background, monitoring the user's facial expressions, and will notify him/her when (s)he is showing signs of stress. In Figure 1 are presented the various modules composing the system.

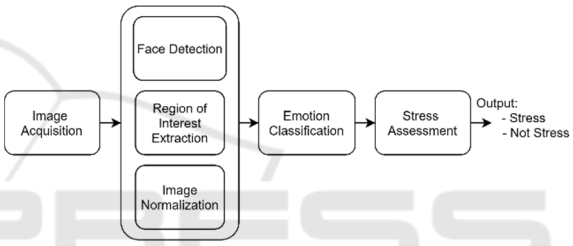

Figure 1: Overview of the modules composing the stress detection system.

The system's first module will be responsible for capturing images in real-time, from the computer's webcam, and sending these images to the second module. In the second module will be determined the user's face's location, using a Haar-like feature selection technique and cropped. The face will then be resized to 299x299 pixels and normalized, by dividing the value of all pixels by 255 so that all have values between 0 and 1. The Emotion Classification module, consisting of a trained classification model, will classify the face and return a list of seven probability scores, one for each facial expression. The facial expression with the highest probability will then be fed to the fourth module (stress assessment). The various classifications made over time will be recorded and based on the parameters provided to the program, it will then determine whether the user is under stress.

The program only requires three parameters:

1) the "frequency" of the program, which determines the time interval, in milliseconds, between each image extraction from video's webcam, and consequently, the remaining modules' execution;

2) the "time window" in seconds, which indicates the period of past classifications that will be considered in the stress assessment process;

3) a "threshold" that indicates the percentage of negative emotions needed to determine, within the time window, if the user is under stress.

For example, if the time window is 900s (or 15 min.) and the threshold is 75%, it means that if in the last 15 minutes, 75% or more of the classifications are for stressful emotions, then the system determines that the user shows signs of stress and it will be displayed a notification alerting for that fact. If the user wishes, he has the possibility to confirm this notification. Also, if the user permits, the images collected by the webcam will be saved on disk and those positively classified by the model will be labelled with stress and its timestamp. The goal to collect these images is twofold, first to create a real dataset of images labelled with stress/non stress, as far as we know, from the research carried out, there aren´t any public dataset with such characteristics accessible; and also make these images available to be analysed by experts.

3 EXPERIMENTATION

For the development of this project, was follow the CRISP-DM methodology (Wirth & Hipp, 2000) that provides guidelines for data mining and machine learning process in general.

3.1 Business and Data Understanding

To tackle the stress detection problem there were two possibilities. Address the problem with data classified as stress or non-stress (preferable approach), or with data classified with the seven facial expressions. Unfortunately, datasets classified directly as stress or non-stress are few and not publicly available. Therefore, tacking in consideration the direct correlation between specific facial expressions and stressful situations, two datasets were requested.

The Karolinska Directed Emotional Faces (KDEF) (Lundqvist et al., 1998) is a dataset created by the Karolinska Institute, bringing together 4900 images of seven human facial expressions, 700 for each one. The facial expressions collected coincide with the five facial expressions of emotion accepted by the scientific community, plus neutral and surprise. Despite we will consider these seven emotions, only anger, disgust and fear will be associated with stress, since only these are accepted as related to stress. Although complete and well uniformed, KDEF is a very homogeneous dataset, where the subjects are all of the same age group, same race, without any facial modification such as glasses or beard, and the images were all captured in a controlled environment

As such, to counteract this homogeneity, it was decided to obtain the CK+ dataset and also to create a new dataset with images collected by us from the internet (here called Net Images), to train the models with more heterogeneous data. With the use of these three datasets, it is expected to obtain a more realistic training and assessment of the models and hopefully closer to the real world.

The CK+ dataset is the result of an extension of the CK dataset (Kanade et al., 2000) that aimed to promote research into automatically detecting individual facial expressions (Lucey et al., 2010). The CK+ dataset has a total of 327 sequences of images and the number of each facial expression in the dataset varies a little presenting a distribution of 45 sequences of anger, 18 of contempt, 59 of disgust, 25 of fear, 69 of happiness, 28 of sadness and 83 of surprise.

The dataset created by us, Net Images, consists of twenty images for each emotion (same emotions as KDEF) in a total of 140 images. These images were collected from searches on two search engines (Google and DuckDuckGo) and free stock images sites (unsplash.com; pexels.com; shutterstock.com; freepik.com). Then, were selected images where the face was visible, and the emotion was unmistakably present. We tried to obtain very heterogeneous images, from people of different ages, races, and with and without a beard and the same for glasses. Images with watermarks, with visible image edition and in which the facial expression could be interpreted as a mix of emotions, as mentioned in (Ekman & Friesen, 2003) were avoided.

3.2 Data Preparation

After collecting the datasets, the images were prepared for training the models. One of the first changes was to adapt KDEF to the problem. Given that this project will use the user's computer webcam to capture user's facial expressions, it was decided to use only the half-left profile, straight, and half right profile images to train the neural network. The first reason was that the full profile images contain little information about facial expressions, which can hinder learning for the neural network. Furthermore, by default, the webcams will be pointing at the user's face from the front, capturing mainly the straight and

half profiles. With this adaptation, the dataset was reduced to 2940 images, 420 for each emotion.

In the CK+ dataset, the sequences' images were reorganized for two reasons. It had the contempt facial expression, which was not present in the other datasets, and it does not have the separation of the neutral facial expression images. Therefore, the reorganization consisted of:

- Extracting the first image of every sequence (which is always a neutral facial expression) to form a new neutral class;
- Elimination of all images except the last six images of each sequence, since the intermediate images do not have very pronounced facial expression.

After this reorganization, the CK+ includes 270 images of anger, 354 of disgust, 150 of fear, 414 of happiness, 309 of neutral, 168 of sadness, and 490 of surprise, in a total of 2155 facial expressions.

Once the dataset was adapted, all the images were cropped around the face and scaled to 299 by 299 pixels. Contrary to the KDEF dataset, the distribution of classes in CK+ is quite different with a certain imbalance between classes.

Lastly, following the hold-out method the training, validation, and test sets were created. The datasets were merged and divided into 80% training, 10% validation and 10% test. But for the partitioning of the datasets, it was considered the works of (Gao et al., 2014; Viegas et al., 2018) where the models showed an ability to adapt to people's faces, or even to the way they express their emotions. Therefore, for the separation of the datasets in training, validation and test data, the divisions were made in such a way that images of a specific person only existed in one of the datasets. As a result, the models will be trained and tested, not only with different images but with different persons, as images of the same person only exists in one of the datasets. Table 1 presents the number of facial expressions in each dataset, train, test and validation.

Table 1: Number of each facial expression in train, validation and test datasets.

Facial expression	Train	Test	Validation
anger	568	68	74
disgust	632	80	80
fear	469	56	62
happiness	682	86	86
neutral	597	76	74
sadness	490	56	62
surprise	730	92	92

3.3 Modelling

For the classification of facial expressions convolutional neural networks (CNN) were explored. As there are not many images to train those CNNs, it was also decided to use pre-trained neural networks and apply transfer learning.

In all, were selected three neural networks created based on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015). The selected neural networks were the VGG16 and VGG19 proposed in (Simonyan & Zisserman, 2015) and the InceptionResNetV2 (Szegedy et al., 2016). VGG16 and VGG19 were selected because they are very simple and straightforward architecturally and are often referred to as the main networks used for transfer learning. InceptionResNetV2 was also selected as it is mentioned in (Hung et al., 2019) presenting good results in a facial expression classification task.

The two pre-trained networks VGG16 and VGG19 differ only in the number of layers of the convolution base. The convolution base is made of convolutional and max-pooling layers. The convolutional layers use a 3x3 pixels filter, padding and stride of 1 pixel. For the max-pooling layers were used a 2x2 pixels window and a stride of 2. In all layers is used the ReLU activation function, except for the Dense layer output of the classifier that uses Softmax.

At an architectural level, Inception-ResNet V2 is a more complex network than the VGG, wider, with filters of different sizes, and organized in different types of blocks. Inception-ResNet modules are only made of convolution layers, while reduction modules are made of both convolution layers and max-pooling layers in order to reduce the image size across the network. The activation function used in the network is ReLU, with the exception of the output layer that uses Softmax and some layers in the Inception-ResNet modules that do not use activation function.

Due to the size of the train dataset, it was necessary during the training process applied data augmentation to the training images. This data augmentation consisted of:

- rotations of up to 20 degrees;
- 10% and 15% translations for width and height, respectively;
- brightness changes between 0.2 and 1;
- zoom-out up to 10% and zoom-in up to 20%;
- horizontal flips.

No data augmentation was made to the images in the validation or test set.

3.3.1 Transfer Learning

Transfer learning reuses characteristics learned in solving a general problem as a starting point for solving another problem. With this technique, it is possible to leverage learning to solve a problem in fewer iterations than those that would be necessary without the previous knowledge. The most common case found in computer vision is the use of pretrained networks, which are trained on a large benchmark dataset.

These pre-trained networks are divided into two parts, convolutional base and classifier. The convolutional base, commonly composed of stacks of convolutional and pooling layers, aims to generate features from the image. This process is called feature extraction. The classifier, often composed of fully connected layers, classifies the image based on the features extracted by the convolutional base.

The typical transfer learning workflow in computer vision is:

- Select a pre-trained network that solves a problem similar to the one intended to be solved;
- Replace the classifier with a new one to be trained in the new dataset;
- Freeze the convolutional base and train the neural network in the new dataset.

The first step taken in creating the models was the definition of classifiers architecture for the transfer learning process. Were tried two different approaches, one classifier based on a global average pooling (GAP) layer and a second with a convolution layer. The two architectures can be seen in Figure 2.

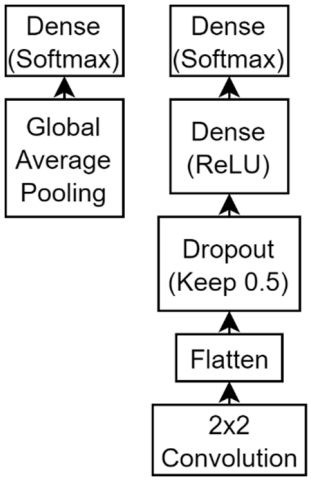

Figure 2: Classifier architecture of the two tried approaches.

For the Global Average Pooling approach was used only a GAP layer together with a fully connected layer with seven neurons and Softmax as the activation function. With this architecture, there are no hyperparameters to define.

The second approach is made up of a convolution layer, followed by a flatten layer, that reshapes all the filters in a single array of one dimension, a 50% dropout to reduce overfitting along with two fully connected layers. The activation function ReLU composes the first fully connected layer and the last, which serves as the output layer that uses a softmax function with seven neurons. For this approach, there are a set of hyperparameters that need to be selected. Therefore, were trained a set of different classifiers to select the most suitable values.

Essentially, we tried to determine the best value for the number of filters in the convolution layer and the number of neurons for the penultimate densely connected layer. Giving four different configurations:

- A: 64 filters with 256 neurons;
- B: 64 filters with 512 neurons;
- C: 128 filters with 256 neurons;
- D: 128 filters with 512 neurons.

To understand the best configuration for each of the pre-trained networks, four configurations were trained and tested with each one of them. For the training of these classifiers, it was necessary to select an optimizer. Initially, was tried the optimizer Adam, however, sporadically, he lost what he had learned so far and dropped to high loss values. Therefore, it was decided to use the Mini-batch Gradient Descent, which corresponds to Keras' SGD, with a momentum of 0.9 and Nesterov Accelerated Gradient active.

After training the various models with only transfer learning, that is, with the training only habilitated to the classifier layers, the following results were obtained with the validation images.

Table 2: Accuracy of the pre-trained networks for each one of the configurations.

Configurations	VGG16	VGG19	Inception-Resnet V2
A	0.7412	0.6770	0.8132
B	0.7198	0.6868	**0.8249**
C	**0.7665**	0.7062	0.8152
D	0.7626	**0.7257**	0.7996

As can be seen, for VGG networks, the configurations with more filters, C and D, were those that presented the highest performance, being the opposite for the Inception ResNet V2, where the configuration with 64 filters, showed some advantage. Based on the results presented in the Table 2, the models to be created with the pre-trained network are: VGG16 with configuration C; VGG19

with configuration D; and Inception ResNet V2 with configuration B.

3.3.2 Fine Tuning

For the fine-tuning process, first it is necessary to determine how much of the convolutional base to train. It is essential to bear in mind two factors, the similarity between the target and source domain, as well as the size of the dataset to classify (Marcelino, 2018).

Even though that the source domain, the ILSVRC, presents images for more than 1000 different classes, none of them refers to people or their facial expressions, our target domain. Therefore, it will be necessary to consider the two domains as non-similar, justifying the use of fine-tuning. Regarding the size of the target domain's dataset, we will have to consider our train dataset a small dataset, so it is necessary to fine-tune most of the pre-trained network's layers.

Since both networks are structured in blocks, the division of the layers to train followed that structure. For the VGG were fine-tuned the last three convolution blocks, conv5, conv4, and conv3, and for the Inception-ResNet V2, were fine-tuned the Inception-resnet-C, Reduction-B, and Inception-resnet-B.

3.3.3 Created Models

Since were selected three pre-trained neural networks and defined two different classification architectures, six classification models were trained:

- **VGG16:** The pre-trained network VGG16, with the convolutional layer classifier, following the configuration C.
- **VGG19:** The pre-trained network VGG19, with the convolutional layer classifier, following the configuration D.
- **IRNV2:** The pre-trained network Inception-ResNet V2, with the convolutional layer classifier, following the configuration B.
- **VGG16 GAP:** The pre-trained network VGG16, with the global average pooling layer classifier.
- **VGG19 GAP:** The pre-trained network VGG19, with the global average pooling layer classifier.
- **IRNV2 GAP:** The pre-trained network Inception-ResNet V2, with the global average pooling layer classifier.

The implementation was made with Keras and the models were trained and validated at the end of each epoch with images from the validation dataset. Once trained and fine-tuned for the best possible result in the validation data, they were evaluated with the test data.

All models were trained on Google Colab, with the free version that allowed access to Nvidia K80s GPUs, 12 GB of RAM, 68 GB of disk space and runs of up to 12 hours. The used version of Keras was 2.3.0 and Python 3.

3.4 Evaluation

To evaluate the performance of the models, confusion matrices were used. Once in possession of this source of information, it is then possible to extract several metrics to evaluate the models. Two kind of models will be evaluated: multi-class classification to predict the seven facial expressions and binary classification to classify stress/non-stress. The distribution of the seven facial expressions in the three datasets doesn't differ too much, so the accuracy and F1-score (harmonic mean of precision and recall) metrics will be used to evaluate the models.

3.4.1 Multi-class Evaluation

Once trained and fine-tuned with the validation dataset for the best possible result, the six models were tested with confusion matrixes presenting each one of the seven facial expressions.

In Table 3 the overall accuracy and F1-score metrics of each of the six multi-class models are presented.

Table 3: Multi-class classification overall models metrics.

Models	Accuracy	F1-score
VGG16	89.6%	89.7%
VGG19	89.4%	89.5%
IRNV2	82.8%	82.4%
VGG16 GAP	84.5%	84.6%
VGG19 GAP	86.2%	86.3%
IRNV2 GAP	81.5%	81.4%

Analysing the performances obtained, clearly the GAP network is superior to Inception-Resnet V2 and for both of these approaches, the combination with convolution layer works better than with the global average pooling layer. The difference between the performance of the two best models, VGG16 and VGG19, is not significant, so the VGG16 will be used since it is simpler than VGG19.

The multi-class VGG16 model has an overall accuracy of 89.6%. The performance of the model for each one of the seven facial expressions is shown in Table 4.

Table 4: Performance metrics for the seven facial expressions with the VGG16 model.

Facial Expression	Precision	Recall	F1-Score	Support
anger	90.6%	78.4%	84.1%	74
disgust	97.3%	90.0%	93.5%	80
fear	77.1%	87.1%	81.8%	62
happiness	100%	93.0%	96.4%	86
neutral	82.1%	93.2%	87.3%	74
sadness	86.2%	90.3%	88.2%	62
surprise	92.5%	93.5%	93%	92

Happiness, disgust and surprise are the facial expressions better recognized with very high F1-score (more than 92%). Concerning the three emotions related with stress, fear is the facial expression with lower accuracy and F1-score. This may be due to less examples associated with this emotion.

3.4.2 Binary Evaluation

Despite the models having been trained for multiclass classification, stress detection is a binary problem, as such, it is necessary to evaluate the performance of the models for classification between stress and non-stress. For this binary evaluation, the test images were separated between stress and non-stress, relating anger, disgust, and fear with stress and the remaining as non-stress.

The VGG16 model presented an accuracy of 92.1%. Table 5 presents the precision, recall and F1-score metrics of the VGG16 model for stress/non-stress classification.

Table 5: Performance metrics for stress/non stress with the VGG16 model.

	Precision	Recall	F1-Score	Support
stress	91.8%	88.4%	90.1%	216
not stress	92.2%	94.6%	93.4%	314

Precision is a measure of confirmation, when the classifier indicates stress, how often it is in fact correct. So, 91.8% of precision means 8.2% of persons were flagged as stress were in fact not. Recall is a measure of utility, how much the model finds stress persons. Most stress persons are in fact tagged (we have high recall – 88.4%) and precision is emphasized over recall, which is appropriate for the stress application.

4 CONCLUSIONS

We developed a system capable of capturing real-time images of the user's face and, using a facial expression classifier, assess if the user presents signs of stress, notifying him/her in such case.

For the creation of the classification model was used transfer learning together with fine-tuning. The pre-trained networks VGG16, VGG19, and Inception-ResNet V2 were considered with two different classifier architectures.

Despite the best model presenting an accuracy of 92.1%, this was based on the association between facial expressions and stress, to overcome the lack of a dataset directly classified in terms of stress or non-stress. This relationship must be supported with more data and case studies. Thus, as future work, we will collect feedback from users concerning the alerts that our system sends them. This will permit to validate if the system correctly classifies stressful situations, thereby increasing the confidence in the association between negative emotions and stress.

Another proposal for future work is the improvement of the classification models, training them with more data collected from our users. The migration of the classification module to a server where it could take advantage of centralized processing with graphics cards (reducing the impact on users' device) will also be considered.

ACKNOWLEDGEMENTS

The authors would like to thank to the creators of KDEF and CK+ for giving us access to use these datasets.

REFERENCES

Carolan, S., Harris, P. R., & Cavanagh, K. (2017). Improving Employee Well-Being and Effectiveness: Systematic Review and Meta-Analysis of Web-Based Psychological Interventions Delivered in the Workplace. *Journal of Medical Internet Research*, *19*(7), e271. https://doi.org/10.2196/jmir.7583

Dewa, C., & Hoch, J. (2015, July). *Barriers to Mental Health Service Use Among Workers With Depression and Work Productivity*. Journal of Occupational and Environmental Medicine; J Occup Environ Med. https://doi.org/10.1097/JOM.0000000000000472

Dinges, D. F., Rider, R. L., Dorrian, J., McGlinchey, E. L., Rogers, N. L., Cizman, Z., Goldenstein, S. K., Vogler, C., Venkataraman, S., & Metaxas, D. N. (2005). Optical computer recognition of facial expressions associated

with stress induced by performance demands. *Aviation, Space, and Environmental Medicine*, *76*(6 Suppl), B172-182.

Ekman, P. (2016). What Scientists Who Study Emotion Agree About. *Perspectives on Psychological Science*, *11*(1), 31–34. https://doi.org/10.1177/1745691615596992

Ekman, P., & Friesen, W. V. (2003). *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*. ISHK.

EU-OSHA. (2019). *Third European Survey of Enterprises on New and Emerging Risks* [Reports]. European Agency for Safety and Health at Work. https://osha.europa.eu/en/publications/third-european-survey-enterprises-new-and-emerging-risks-esener-3/view

Gao, H., Yüce, A., & Thiran, J.-P. (2014). Detecting emotional stress from facial expressions for driving safety. *2014 IEEE International Conference on Image Processing (ICIP)*, 5961–5965. https://doi.org/10.1109/ICIP.2014.7026203

Giannakakis, G., Pediaditis, M., Manousos, D., Kazantzaki, E., Chiarugi, F., Simos, P. G., Marias, K., & Tsiknakis, M. (2016). Stress and anxiety detection using facial cues from videos. *Biomedical Signal Processing and Control*, *31*, 89–101. https://doi.org/10.1016/j.bspc.2016.06.020

Hung, J. C., Lin, K.-C., & Lai, N.-X. (2019). Recognizing learning emotion based on convolutional neural networks and transfer learning. *Applied Soft Computing*, *84*, 105724. https://doi.org/10.1016/j.asoc.2019.105724

Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 46–53. https://doi.org/10.1109/AFGR.2000.840611

Lerner, J. S., Dahl, R. E., Hariri, A. R., & Taylor, S. E. (2007). Facial Expressions of Emotion Reveal Neuroendocrine and Cardiovascular Stress Responses. *Biological Psychiatry*, *61*(2), 253–260. https://doi.org/10.1016/j.biopsych.2006.08.016

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. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska Directed Emotional Faces (KDEF), CD-ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9*.

Maaoui, C., Bousefsaf, F., & Pruski, A. (2015). Automatic human stress detection based on webcam photoplethysmographic signals. *Journal of Mechanics in Medicine and Biology*, *16*. https://doi.org/10.1142/S0219519416500391

Marcelino, P. (2018). *Transfer learning from pre-trained models*. https://towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv: 1409.1556 [Cs]*. http://arxiv.org/abs/1409.1556

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *ArXiv:1602.07261 [Cs]*. http://arxiv.org/abs/1602.07261

Viegas, C., Lau, S.-H., Maxion, R., & Hauptmann, A. (2018). Towards Independent Stress Detection: A Dependent Model Using Facial Action Units. *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, 1–6. https://doi.org/10.1109/CBMI.2018.8516497

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining*. 29–39.