Evaluating Mental Patients Utilizing Video Analysis of Facial Expressions

M. Tziomaka, A. Kallipolitis, P. Tsanakas, Ilias Maglogiannis

Evaluating mental patients utilizing video analysis of facial expressions

M. Tziomaka[1] [0000-0002-2171-1069], A. Kallipolitis[1] [0000-0001-9234-0069], P. Tsanakas[2] and I. Maglogiannis[1] [0000-0003-2860-399X]

[1] Department of Digital Systems, University of Piraeus, Greece
[2] National Technical University of Athens, Greece
tziomakamel@unipi.gr
nasskall@unipi.gr
panag@cs.ntua.gr
imaglo@unipi.gr

Abstract. The objective of this work is to put in numbers the degree of symptoms severity based on the social behavior and cognitive functioning of mental patients when conducting a routine conversation with their attending doctor. Examination of patient's facial expression manifestations can be a key indicator towards the quantization of cognitive impairment in respect to receiving external emotion expressions. Recent advancements in computer vision machine and deep learning techniques allow the evaluation and recognition of temporal emotional status through facial expressions. In this context, the paper studies the application of these techniques for the automated recognition of Positive and Negative Syndrome Scale (PANSS) indicators by means of extracting features from patients' facial expressions during video teleconferences. The paper discusses the technical details of the implementations of a video classification methodology for the prediction of schizophrenia symptoms' severity, introduces a novel approach for the interpretation of video classification results and presents initial results where it is demonstrated that the proposed automated techniques can classify to a certain extend specific PANSS indicators.

Keywords: Schizophrenia, Facial Emotion Recognition, Speeded Up Robust Features, Efficient Net, Bag of Visual Words, Interpretability, PANSS.

1 Introduction

Schizophrenia is a severe, complex, heterogenous psychiatric disorder that, apart from the negative effects influencing the life of the patients and its kinsfolk, may have devasting side effects in the financial of balance of a country, accounting for generating large expenditures concerning unemployment, social and health care [1]. It has been estimated that it affects 20 million people worldwide [2], who are 2 or 3 times more likely to die than the general population [3]. Even though the disorder is treatable, a significant percentage of patients fail to receive appropriate care [4], while others fail to be recognized as patients due to important challenges that haunt the respective

diagnosis process. Equally challenging remains the understanding of causes and inner mechanisms that lead to the specific disorder. Except for the diagnosis of a patient with schizophrenia, the effort of clinicians lies as well on the assessment of symptoms' severity when referring to already diagnosed schizophrenic patients due to its importance in effectively treating the disorder. One of the most dominant and well-established procedures in the assessment of symptoms severity as the Positive and Negative Syndrome Scale (PANSS) interview [5]. As its name states, the interview focuses on the evaluation of two types of symptoms, the positive ones that refer to the excessive occurrence of normal functions and the negative ones which correspond to limited occurrence of normal functions. Overall, 30 symptoms are rated on a scale from 1 to 7, resulting to a maximum score of 210 points. The symptoms are divided in three main categories: positive, negative, and general psychopathology symptoms. The main idea that connects facial emotion recognition to schizophrenia and PANSS is hidden in the cognitive impairment of patients to perceive emotional material [6]. This deficiency to decode human emotion leads to the generation of the patient's facial expressions that can be informative of the symptoms' scale. Schizophrenic patients demonstrate discomfort when having to deal with the interpretation of neutral or negative facial expressions. Since this discomfort is evident on their own facial expressions [14], their quantification can discover new knowledge concerning the cognitive impairment on different stages of the disorder.

Driven by the latest advancements in the field of Emotion Artificial Intelligence (EAI) [7] and in a wide range of technological areas, namely Healthcare [11], Augmented Reality [12], Internet of Things [13], Business Analytics [10], Advanced Driver Assistance [9] and Gaming [8] the quantification of facial emotion recognition (FER) has become the heart of many human-centered applications and it is gaining the industry as well due to its overwhelming results. In many cases, these applications utilize the universally recognized basic emotions as defined by Ekman in [15] as a standardized procedure to classify facial expression.

Taking under considerations the achievements in EAI and the need for the quantification of patients' facial expression in order to predict and discover meaningful correlations with PANSS indicators, we are focusing only on certain symptoms that can be directly induced by the facial expression manifestations. The corresponding PANNS items have been dictated by specialized personnel. Therefore, in this paper, we describe the design and implementation of a machine learning methodology, based on handcrafted and learned features, extracted from video teleconferences for the prediction of schizophrenic syndrome's severity that has been assessed through PANSS questionnaires. Our main contribution lies on the effective prediction of specific questionnaire indicators and the introduction of a novel approach for video classification results. The results for some PANSS indicators are encouraging and suggest that automated facial expression recognition can be utilized for the prediction of symptoms severity.

The remainder of this paper is structured in 6 sections, as follows: Section 2 presents the related research works, while Section 3 describes the proposed methodology workflow. Section 4 reports the experiments conducted and the corresponding results.

Section 5 explains the integrated interpretability scheme. Finally, Section 6 concludes the paper.

2 Related Work

As stated earlier, the basic hypothesis behind the effort to discover hidden associations and knowledge lies on the difficulty that schizophrenic patients face when trying to recognize stimuli, mostly negative, [16,17] on other humans' facial expressions as it is directly connected to their cognitive deficiency. In most research works, an attempt has been witnessed to directly measure this difficulty by scoring the patients' answers a) as proposed by the automated tool Emotion Recognition Index (ERI) [18] analysis to find associations with symptoms' severity, functionality and cognitive impairment [19], and b) following the emotion recognition assessment described in [16] to discover impairments in patients at clinical high-risk for schizophrenia before the full expression of psychotic illness [21].

Non-verbal behavior of schizophrenic patients and its relation to symptoms' severity is examined as well in the literature, as it accounts for the expression of 60–65% of social communication [20]. In [22], non-verbal behavior, following the modified version of the Ethological Coding System for Interviews (ESCI), and symptoms' severity, based on three established scales [PANSS, (Clinical Assessment Interview for Negative Symptoms) CAINS and the Calgary scale], were separately evaluated by different specialized personnel to reach the fruitful conclusion that association between negative symptoms and a limited engaging non-verbal behavior truly exists. A different non -verbal approach is proposed in [23] by means of a joystick tracking task to assess the visual motor processing of schizophrenic patients.

Contrary to the basic trend that manually assess the verbal or non-verbal responses of schizophrenic patients to exterior stimuli and driven by the achievements of deep convolution neural networks (DCNN) in the field of Computer Vision, the research work in [24] proposed an automatic methodology for the analysis of patients' facial expressions to estimate symptoms of schizophrenia. The human pose estimation is the starting point through which the face is detected by means of a designated neural network for face detection. The extraction of low-level features in terms of action units in faces with a separate VGG-16 net is followed by the extraction of high-level features concerning each video. The results are promising and verify the automated detection of correlations between patients' facial expression and the corresponding symptoms. Towards the same path of facial expression automated analysis by means of machine learning and deep learning techniques are directed the efforts in [25, 26]. The video samples, through which the automated analysis of facial expressions is performed, are taken during the professional-patient interview for the assessment of symptoms severity and are in certain cases captured in a multiple camera setting or by utilizing special equipment (i.e., depth cameras).

Inspired by the abovementioned research work, we propose a simple, yet efficient architecture for the prediction of symptoms severity based on low-level (frame) and high-level (video) representations extracted from video teleconferences between cli-

nicians and schizophrenic patients. The successful realization of an automated facial expression analysis system for the prediction of schizophrenia symptoms will discharge the psychiatrists from the burden of manually annotating recorded videos, enhance objectivity and reliability at the procedure of symptoms' severity estimation and make the assessment accessible to more patients through better allocation of freed resources and telemedicine.

3 Methodology

3.1 Overview

To classify the videos into the subscales of PANSS, the utilized techniques in this work follow a methodology scheme, which consists of four consecutive stages: Data Preprocessing, Frame Representation, Video Representation and Classification (Fig. 1).

3.2 Data Processing

Initially, each input video is converted into a sequence of RGB frames, with a sampling rate of one frame per second. After the frame extraction, the face region in each frame is detected by utilizing the Multiple Task Cascaded Neural Network (MTCNN) deep learning model [24] and cropping is performed to the dimensions of the detected face. As a result, we obtain a segmented region from each frame, to which the frame is cropped.

3.3 Frame Representation

For the representation of each frame, experiments with two different feature extraction techniques were conducted: the method of bag of visual words (BOVW) and the method of transfer learning, by utilizing a pre-trained convolutional neural network (CNN) as feature extractor.

In the case of utilizing the technique of BOVW, the first step is to detect the interest points of each frame, by utilizing the Speeded Up Robust Features (SURF) algorithm feature detector. The SURF algorithm [26] automatically detects by means of a fast Hessian detector n interest points, where n is the interest points of an image, and describes each one of them by assigning a 64-dimensional vector. After the detection and description of all interest points from all images in the dataset, a collection of 64-dimensional vectors is formed, which is in turn clustered into k groups, where k is a hyperparameter, utilizing the k-means algorithm. The centroid of each cluster represents a visual word, resulting in the formation of a visual vocabulary of k visual words. Lastly, the interest points of each image are assigned to a visual word, and the frequency of each word in an image is computed, thus forming a histogram of k values for each frame.

On the second approach, the pre-trained CNN that works as a feature extractor for the frame representation is the base model of the EfficientNet family of networks, the EfficientNet-B0 [25]. The network's architecture emerged by utilizing the method of multi-object neural architecture search on the ImageNet dataset to optimize accuracy and FLOPS, making it a high quality, yet compact model. Its main building block is the mobile inverted bottleneck convolution (MB Conv), with the depth-wise separable convolution.

Unlike the traditional convolution operation, which applies a 2-D depth filter to directly convolve the input in depth as well, depth-wise separable convolution uses each filter channel only at one input channel. Precisely, it breaks the filter and image into three different channels and applies the corresponding filter to the corresponding channel. Finally, it combines the output by applying a pointwise convolution. The MB Conv Block flips the classic wide – narrow – wide approach, in which skip connections exist between wide parts of the network, to a narrow– wide – narrow approach with skip connections between narrow parts of the network. The first step is a 1x1 convolution, which increases the depth, then follows a depth-wise convolution, and lastly another 1x1 convolution squeezes the network in order to match the initial number of channels for the skip connection. The EfficientNet-B0 not only provides better accuracy, as compared to other state-of-the-art models, but also improves the efficiency of the model by reducing the number of parameters. To leverage the power of the model, the method of transfer learning is employed, with weights pretrained on ImageNet. The pretrained model's last layers for classification are excluded and the layer that is used as a feature extractor, is the last convolutional layer that extracts richer features compared to the lower layers. Lastly, the output of the convolutional layer is flattened to create a single long 1,280-dimensional feature vector. To prepare the images before passing them through the network, all frames were resized according to EfficientNet-B0's input dimensions and normalized by subtracting the mean and dividing by the standard deviation RGB values of the ImageNet dataset, that were used to pretrain the model.

3.4 Video Representation

After the frame representation, the outcome from both approaches is a matrix of m x n for each video, where m is the number of frames in the video and n is the dimension of the feature-extracted vectors (n = k from the BOVW method or n = 1,280 from the transfer learning method). In turn, to represent each video with a vector, the method of BOVW was reapplied to the collection of all n-dimensional vectors of all videos. This collection of all frame representations is standardized and clustered into k' groups, with k' being an additional hyperparameter, as in this step the centroid of each of these clusters represents a visual word for the frame representations. Subsequently, the frame representations of each video are assigned to a visual word, to form a histogram of k' values for each video, which corresponds to a k'-dimensional vector.

3.5 Classification

To classify the video representations by means of the BOVW approach, we utilized three state-of-the-art machine learning models: XGBoost, Random Forest and Support Vector Machines (SVM) with radial basis function (RBF) kernel.

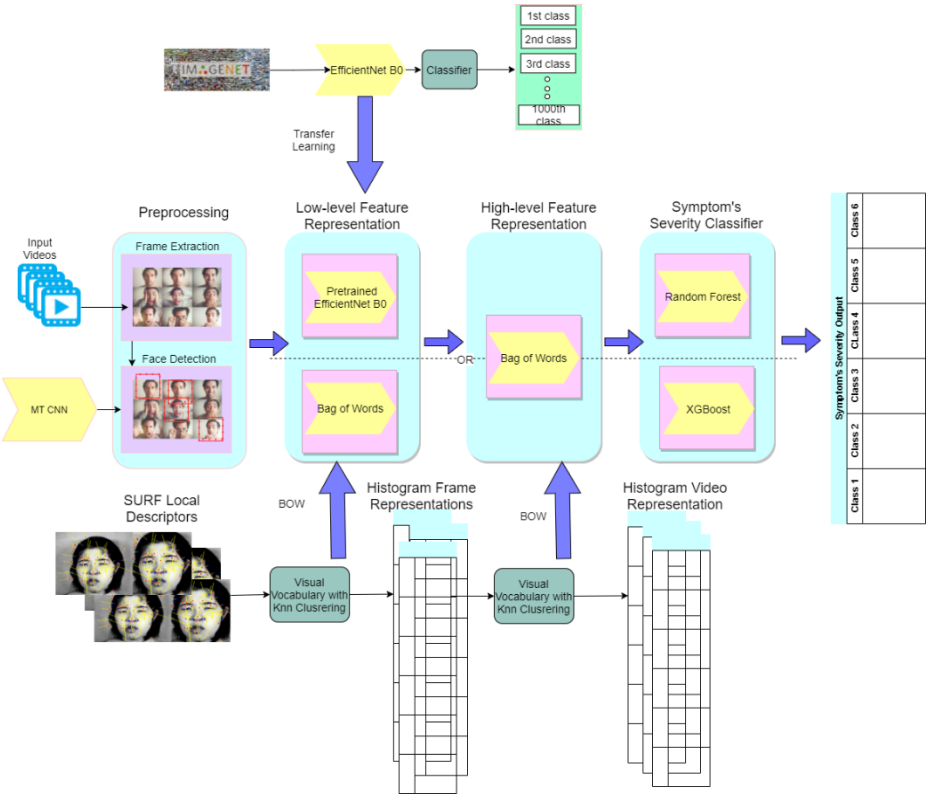

Fig. 1. Overview of the proposed methodology workflow with alternative scenarios.

4 Experimental Results

The verification of the proposed methodology's performance on the task of predicting symptoms' severity on schizophrenic patients was done using a dataset that was organized and manually curated from video-teleconferences between professionals and patients obtained within the framework of the e-prevention project [30]. Patients are diagnosed with the disorder and have already suffered one serious episode. The ground truth concerning the symptoms severity was provided by PANSS questionnaires that were conducted by specialized personnel of the Eginition Hospital. At the time the tests were conducted, the number of videos is 167, corresponding to 22 patients. The videos are captured to two weeks or closer to the PANSS evaluation interview, which is a rather challenging requirement, given that previous work directly

evaluates the video from the PANSS evaluation interview. Furthermore, they were captured in variating lighting and distance conditions and their duration spans from to 30 to 1141 seconds. During the video-teleconferences common everyday questions are asked to form a routine conversation between the individuals.

Regarding to the PANSS items that serve the cause of ground truth for the predictive model, ten were selected due to their correlation with the facial expressions, namely: excitement, hostility (positive items) anxiety, poor impulse, motor retardation, depression, tension (general items), blunted affect, poor rapport, lack of spontaneity, flow of conversation (negative items). For clarification, positive, negative, and general items are the three main scales of PANSS questionnaire. The values in the respective items that correspond to the symptoms' severity can take values from one to six according to Fig.2. For the training of the various configurations of the predictive models a correspondence between videos and PANSS questionnaires was created by matching each video to the most recent questionnaire in a timeframe of two weeks, meaning that the video was captured at most two weeks before the questionnaire. The dataset was split into two subsets, training (70%) and testing (30%). The above-mentioned setting was utilized in different machine learning and deep learning techniques for the prediction of a specific PANSS item at a time.

As already mentioned in the methodology the various combinations of workflows follow the same guideline of a) initially transforming the input videos into cropped images of facial expressions, b) forming low-level representations from frames (facial expressions) and c) forming high-level representations from videos. In this context, we tested the following configurations: BOVW2, Bag of Visual Words (BOVW) for low-level representation and BOVW for high-level representation and EfficientNet to BOVW, EfficientNet for low-level representation and BOVW for high-level representation. For the determination of the number of clusters that would represent the visual vocabularies, we conducted exhaustive grid search resulting in different number of clusters for each case, for which we are presenting in Table 1 the best performing configurations. The performance is measured in terms of balanced accuracy and top-2 accuracy metrics. As far as the balanced accuracy metric is concerned, predictions of tension, hostility and poor impulse control's severities show the most promising results, whereas both configurations fail to predict anxiety and poor rapport. Against our expectations lies the fact that the pretrained Efficient Net extracted features show inferior performance than the SURF features. In some cases, such as the poor impulse control, poor rapport, tension, and excitement items top-2 accuracy supersedes the barrier of 80% and demonstrate big divergence from the balanced accuracy results.

5 Interpretability

Thanks to the simplicity of the proposed methodology, the direct association between the classification results and visual patterns in the video frames can be unveiled. The connection between cause and result is of major importance when developing a machine learning predictive model, since it provides useful insight concerning the reasoning of misclassifications, enhance trust and transparency towards the users and can lead to the discovery on newly breed knowledge through the dictation of patterns

previously unknown to humans. Therefore, the build-in explainability properties in high-stake predictive models (i.e., health-care computer aided diagnosis systems)

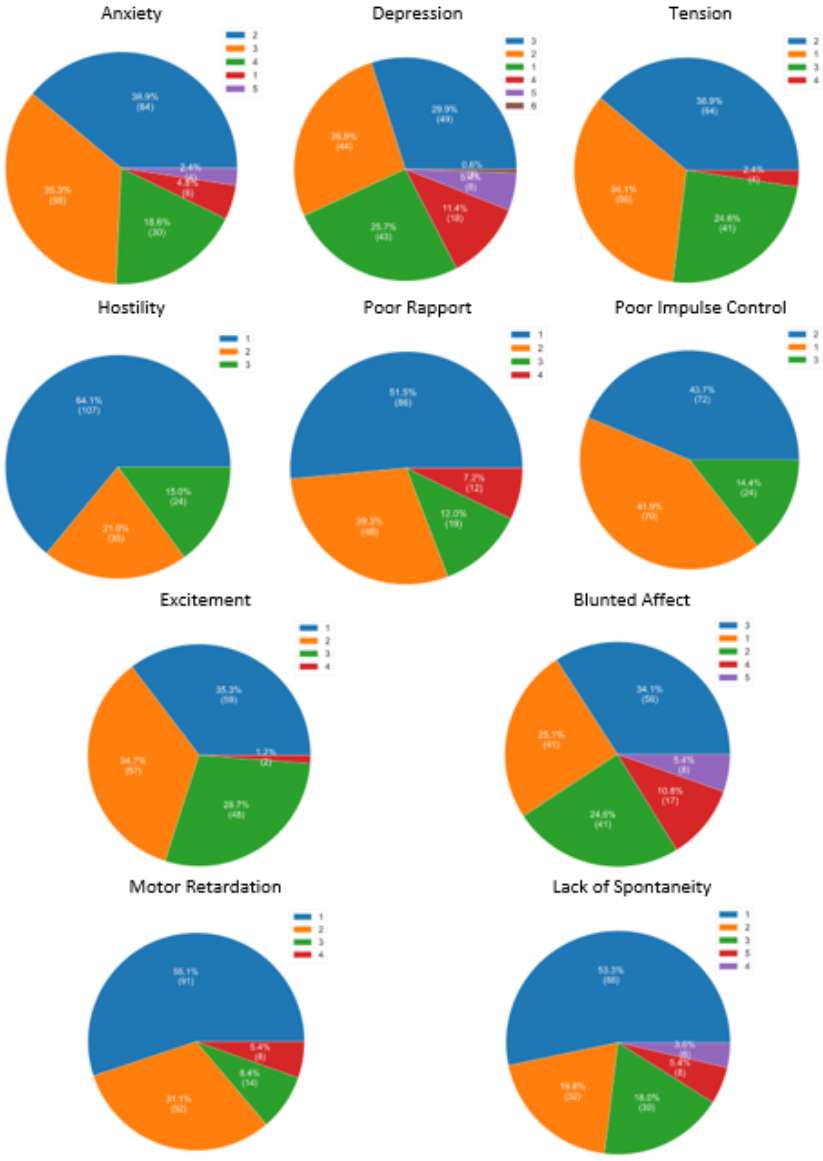

Fig. 2. Ground truth values according to selected PANSS items.

should be an important prerequisite [27]. In our proposed methodology (BOVW^2 configuration), the frame representation, video representation and classification stage are

by design equipped with interpretable properties that can be seamlessly fused to provide measurements concerning the degree of visual patterns' influence to the result. As explained earlier, these measurements are the combination of a three-fold scheme. Firstly, the classifier provides a feature importance mechanism based on the individual inner workings that directly maps the output to those inputs that were mostly influential. By acquiring this feature importance map in the video representation stage, the information of the most important visual words is stored in a k' vector equal to the number of clusters corresponding to the clustering of all frame representations. Since, each frame is already assigned to a cluster, the importance of each frame can be calculated as follows:

$$I_{fr} = \frac{W_{vw}}{D_{fr}} \quad (1)$$

, where I_{fr} is the importance of each frame, W_{fw} is the weight of the video visual word provided by the feature importance mechanism of the classifier and D_{fr} the distance of the frame representation from the assigned cluster in the frame representation clustering. In the same manner, the importance of each SURF keypoint in the cropped frame can be calculated to highlight the visual patterns in the patient's face that lead to the formation of a certain prediction, as described in equation (2):

$$I_k = \frac{I_{fr}}{D_k} \quad (2)$$

, where I_k is the importance of each keypoint whose coordinates are known and D_k the distance of the frame representation from the assigned cluster in the keypoints representation clustering. I_k can be visually observed in the form of a heatmap on the cropped facial expression, but due to data privacy regulations the respective heatmaps cannot be shown in the paper. For a more detailed explanation of the proposed interpretation scheme, the methodology is based in the work presented in [28]. However, a general paradigm of the heatmap on a facial expression provided by the JAFFE dataset [29] is shown in Fig.3. By utilizing the proposed interpretation scheme, useful knowledge can be extracted concerning the most significant frames and the most significant keypoints to the classification result.

Table 1. Classification results for BOVW^2 (left) and Efficient to BOVW (right) configurations based on balanced accuracy (left) and top-2 accuracy (right) metrics. The number of respective classes is shown next to PANSS items.

PANSS items	**Configuration**											
	BOVW^2						**EfficientNet to BOVW**					
	Balanced Accuracy | Top-2 Accuracy											
	RF		**XGB**		**SVC**		**RF**		**XGB**		**SVC**	
Depression (6c)	0.49	0.82	**0.6**	0.85	0.3	0.68	0.47	0.64	0.53	0.64	0.3	0.75
Anxiety (5c)	**0.44**	0.65	0.37	0.69	0.33	0.84	0.26	0.77	0.36	0.77	0.29	0.85

Tension (4c)	**0.7**	0.89	0.68	0.85	0.56	0.62	0.54	0.77	0.61	0.85	0.46	0.81
Poor Rapport (4c)	0.44	0.86	0.49	0.9	0.34	0.68	0.41	0.86	**0.55**	0.82	0.34	0.57
Poor Impulse Control (3c)	**0.72**	1	0.71	0.79	0.39	1	0.43	0.96	0.66	0.82	0.37	0.96
Motor Retardation (4c)	0.4	0.75	**0.6**	0.82	0.36	0.75	0.41	0.82	0.4	0.79	0.32	0.79
Excitement (4c)	0.58	0.78	**0.61**	0.82	0.42	0.78	0.49	0.85	0.47	0.85	0.41	0.67
Hostility (3c)	**0.72**	0.96	0.68	0.93	0.42	0.78	0.42	1	0.6	0.89	0.49	0.79
Blunted Affect (5c)	**0.58**	0.68	0.55	0.64	0.34	0.64	0.34	0.64	0.47	0.68	0.3	0.64
Lack of Spontaneity(5c)	0.49	0.64	**0.65**	0.79	0.4	0.75	0.49	0.64	0.26	0.68	0.43	0.75

6 Conclusion

In this paper a video classification methodology for the prediction of schizophrenia symptoms' severity based on video-teleconferences between doctors and patients was presented. While the ground truth is provided by PANSS that are conducted in a two-weeks period before the videos, the predictive model manages to discover important correlations between facial expressions and specific PANSS items that evaluate symptoms' severity. The classification results show good performance in some cases. Although results are promising, further testing with the utilization of an anticipated larger dataset should be performed to enhance confidence in the presented results. The

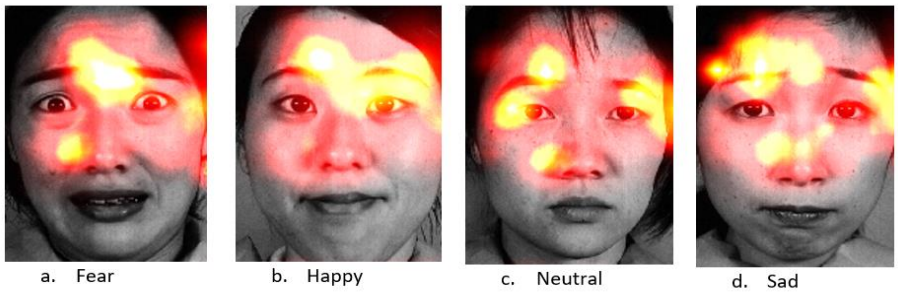

Fig. 3. Keypoint importance heatmaps generated by the proposed methodology on facial expressions depicting four emotions from left to right: a. Fear, b. Happy, c. Neutral, d. Sad.

big divergence between the balanced and top-2 accuracy metrics can be strongly related with the subjectivity of PANSS questionnaire scoring. Future work will be focused on the utilization of transformers that are intended to extract knowledge from large sequences such as videos, the analysis and qualitive results of the interpretation scheme and towards the prediction of potential relapses of the disorder in diagnosed patients. Should public datasets be available, the testing for the generalization properties of our methodology will be conducted. To sum up, the PANSS questionnaire evaluation is a demanding task that requires time and excessive training and on-site

presence. Automated estimation of PANSS items can assist clinician in this arduous process, make it more accessible to patients and provide objectivity and reliability.

Acknowledgment

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T1EDK-02890)

References

1. Owen M.J., Sawa A., Mortensen P.B.: Schizophrenia. Lancet 388(10039), 86-97 (2016).
2. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet (2018).
3. Laursen T.M., Nordentoft M., Mortensen P.B.: Excess early mortality in schizophrenia. Annual Review of Clinical Psychology 10, 425-438 (2014).
4. Lora A et al.: Service availability and utilization and treatment gap for schizophrenic disorders: a survey in 50 low- and middle-income countries. Bulletin World Health Organisation 90(1), 47-54 (2012).
5. Opler, M., Yavorsky, C., Daniel, D. G.: Positive and Negative Syndrome Scale (PANSS) Training: Challenges, Solutions, and Future Directions. Innovations in clinical neuroscience 14(11-12), 77–81 (2017).
6. Kohler C.G., Walker J.B., Martin E.A., Healey K.M., Moberg P.J.: Facial emotion perception in schizophrenia: a meta-analytic review. Schizophr Bull 36, 1009–1019 (2010).
7. Schuller D., Schuller, B. W.: The Age of Artificial Emotional Intelligence. Computer 51 (9), 38-46 (2018).
8. McStay, A., Rosner, G. (2021).: Emotional artificial intelligence in children's toys and devices: Ethics, governance, and practical remedies. Big Data & Society 8 (1), (2021).
9. T. Wilhelm.: Towards Facial Expression Analysis in a Driver Assistance System. 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pp. 1-4, Lille, France (2019).
10. Subhashini, R., Niveditha, P.R..: Analyzing and Detecting Employee's Emotion for Amelioration of Organizations. Procedia Computer Science 48, 530-536 (2015).
11. Alhussein, M.: Automatic facial emotion recognition using weber local descriptor for e-Healthcare system. Cluster Comput 19, 99–108 (2016).
12. Chen C.H., Lee I.J., Lin L.Y.: Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders. Res Dev 36, 396-403 (2015).
13. Xi, Z., Niu, Y., Chen, J., Kan, X., Liu, H.: Facial Expression Recognition of Industrial Internet of Things by Parallel Neural Networks Combining Texture Features. IEEE Transactions on Industrial Informatics 17(4), 2784-2793 (2021).

14. Tse, W. S., Lu, Y., Bond, A. J., Chan, R. C., & Tam, D. W.: Facial emotion linked cooperation in patients with paranoid schizophrenia: A test on the Interpersonal Communication Model. International Journal of Social Psychiatry 57(5), 509-517 (2011).
15. Ekman, P.: Facial expression and emotion. American psychologist, 48(4), 384 (1993).
16. Edwards J., Jackson H. J., Pattison P.E.: Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review. Clin Psychol Rev. 22, 789–832 (2002).
17. Adolphs R., Tranel D.: Impaired judgments of sadness but not happiness following bilateral amygdala damage. J Cogn Neurosci. 16, 453–462 (2004).
18. Suárez-Salazar J.V., Fresán-Orellana A., Saracco-Álvarez R.A.: Facial emotion recognition and its association with symptom severity, functionality, and cognitive impairment in schizophrenia: preliminary results. Salud Mental 43(3), 105-112 (2020).
19. Burgoon, J. K., Guerrero, L. K., & Floyd, K. (2010). Nonverbal communication. (1 ed.) Allyn & Bacon.
20. Amminger, G. P., Schäfer, M. R., Papageorgiou, K., Klier, C. M., Schlögelhofer, M., Mossaheb, N., Werneck-Rohrer, S., Nelson, B., & McGorry, P. D.: Emotion recognition in individuals at clinical high-risk for schizophrenia. Schizophrenia bulletin 38(5), 1030–1039 (2012).
21. Worswick, E., Dimic, S., Wildgrube, C., Priebe, S.: Negative Symptoms and Avoidance of Social Interaction: A Study of Non-Verbal Behaviour. Psychopathology 51(1), 1–9 (2018).
22. Lu, P.Y., Huang, Y.L., Huang, P.C. et al.: Association of visual motor processing and social cognition in schizophrenia. npj Schizophr 7, 21 (2021).
23. Bishay, M., Palasek, P., Priebe, S., & Patras, I.: SchiNet: Automatic Estimation of Symptoms of Schizophrenia from Facial Behaviour Analysis. ArXiv, abs/1808.02531 (2018).
24. Zhang, K., Zhang, Z., Li, Z., & Qiao, Y.: Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters 23, 1499-1503 (2016).
25. Tan, M., Le, Q.V.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. ArXiv, abs/1905.11946 (2019).
26. Bay, H., Tuytelaars, T., Gool, V.G.: Speeded Up Robust Features. Computer Vision and Image Understanding 110 (3), 346-359 (2008).
27. Rudin C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead Nat Mach Intell 1, 206–215 (2019).
28. Kallipolitis, A., Stratigos, A., Zarras, A., & Maglogiannis, I.: Explainable Fully Connected Visual Words for the Classification of Skin Cancer Confocal Images: Interpreting the influence of visual words in classifying benign vs malignant pattern. 11th Hellenic Conference on Artificial Intelligence (2020).
29. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with Gabor wavelets. In Proceedings, Third IEEE International Conference on Automatic Face and Gesture Recognition pp. 200–205. IEEE Computer Society, Nara, Japan (1998).
30. Maglogiannis, I., Zlatintsi, A., Menychtas, A., Papadimatos, D., Filntisis, P.P., Efthymiou, N., Retsinas, G., Tsanakas, P., Maragos, P.: An intelligent cloud-based platform for effective monitoring of patients with psychotic disorders, Proc. 16th International Conference on Artificial Intelligence Applications and Innovations (AIAI-2020), June 2020.