Facial Analysis and Binary Classification of Acutely Sick Patients

Malina Chichirau (S3412768)   Andrei Voinea (S3754243)

{m.chichirau, c.a.voinea}@student.rug.nl
University of Groningen
9747 AG, Groningen, The Netherlands

Abstract

It has been suggested that people have an innate ability to discriminate between sick and healthy individuals, even non-experts can identify sick people above chance level. The hypothesis that algorithms can also be trained to perform a similar classification is studied in the present paper, by implementing Deep Learning Convolutional Neural Networks to analyze the phenotype of four facial regions (i.e. eyes, nose, mouth and skin). Due to ethical concerns, the training data was simulated by applying standardized make-up on a small number of participants to mimic the features of critically ill patients. The models were externally validated using photographs of actually sick people. The results suggest that the models can identify sick patients above chance level, however the generalization power of the networks is limited by the size and the nature of the data set available.

1 Introduction

Infectious diseases can be intercepted by humans by coming into contact with an infected person or animal. A severe condition can lead to life-threatening complications, such as sepsis [1], and has thus been speculated that humans evolved mechanisms that enable them to identify the signs of acute illness. Avoiding infected individuals minimises the risk of contamination, which allowed humans to survive various viruses or bacteria [2]. Axelsson et.al. [3] demonstrated that people with no previous training were able to identify the photos of sick patients above chance level. Their study induced a sick state by injecting healthy volunteers with LPS [1].

Axelsson et al. [3] further described the features that the participants most strongly associated with acute illness. Their observations include an overall tired appearance, pale skin, skin texture, swollen face, pale lips and hanging eyelids. Such features can also be identified by computer vision, one example being a Deep Learning model trained to identify multiple genetic disorders based on images of patients [4]. The authors note that the model provided more accurate answers compared to experts for certain syndromes. Unlike human experts who encounter a limited number of patients, a Deep Learning model can be trained using a large number of examples. Furthermore, it can learn to identify rare disorders or new signs of illness, possibly unbiased by previous research or literature.

In certain cases, patients can be misdiagnosed and not receive the proper care in time [5]. A computer model could continuously monitor the state of a patient and/or help junior doctors provide a verdict to each case. As seen in the study conducted by Gurovich et al. [4], a model can identify syndromes with minor discrepancies from normal. Therefore, the current study aims to train a Deep Learning model for the binary classification of acutely sick patients.

Our first research question probes whether a Convolutional Neural Network (CNN) model can distinguish critically ill patients from healthy people. We hypothesise that, by allowing the model to analyze the features identified by Axelsson et al. [3], it can accurately identify acutely ill patients. As such, we will train four individual CNNs for each feature (eyes, nose, mouth and skin) and one CNN that concatenates the previous networks (see Section 2.3 for further details). Thus, by comparing the accuracy

[1] Escherichia coli endotoxin, Lot HOK354, CAT number 1235503, United States Pharmacopeia, Rockville, MD, USA

of the individual networks with the accuracy of the concatenated networks, the face regions which are better indicators of illness can be identified.

Due to privacy concerns and the ethics behind using photographs of acutely ill patients to train the model on, the training data will use photographs of healthy individuals wearing makeup. This was applied in a manner that mimics the features of sick patients, as described by Axelsson et al. [3]. Thus, a secondary research question inquires whether a model trained using such data can successfully generalize its predictions on actual photographs of acutely ill patients. Consequently, the model will be validated using the photographs provided by Axelsson et.al. [3]. Given that acute illness is common enough that non-experts can discriminate between sick and healthy individuals [3], we hypothesize that a Deep Learning model can be trained to reach at least a similar accuracy as non-experts. Furthermore, we aim to compare the face regions used by the Deep Learning model to identify sick patients with the ones reported by the participants of Axelsson et.al. [3].

2 Methods

2.1 Data Collection

Two photographs were taken of each of the 26 participants (15 males), aged between 18 - 30 years (mean 24.1), recruited among fellow students. One of the photographs represented the *healthy state* taken without any make-up, and the other represented the "sick state" where standardized make-up was applied. The make-up was designed to follow the observations of Henderson et.al. [6], Saralidou et.al. [7] and Axelsson et.al. [3] on face features, skin colouration and facial expressions of sick patients. The photographs were taken in a standardized environment (a grey background, LED light), using an iPhone 8 camera (4032 x 3024 pixels), standardized settings (ISO 22, RAW, AF, S1/40, MF: 0,9 and AWB in the Halide app).

The external validation data set consisted of the photographs used by Axelsson et al. [3]. They provided 44 images, where each participant was photographed in two conditions: placebo (healthy) and sick (2 hours and 10 minutes after being injected with LPS). Axelsson et al. recruited 22 participants (13 males) aged between 19 - 34 years (mean 23.4). Due to differences in facial hair between the two conditions or hairstyles covering a large portion of the face, the experimenters excluded 12 images. For the purpose of the present experiment, 6 photographs were also excluded (3 in the sick condition and 3 in the healthy condition) due to similar reasons, especially since the training data was produced such that male participants have no facial hair. However, it is unknown whether the same subjects were excluded from the experiment of Axelsson et.al. [3].

2.2 Data Preprocessing

Since the photographs of Axelsson et.al. [3] were acquired only after collecting the simulated data set, the photographs differ in certain aspects. In the simulated data, the features associated with the sick condition were more accentuated than in the validation data. Furthermore, the lighting was brighter in the validation data set, as the photographs were plainer. In the simulated data set, the lighting was dimmer, the shadows and contrasts being harsher. In order to even out the differences between the data sets, all photographs in the simulated set were brightened (gamma = 1.3). Furthermore, as mentioned before, 6 of the validation photographs (3 sick and 3 healthy) were excluded due to facial hair and excessive hair covering the face. The photographs in both data sets were resized to 128 x 128 pixels and the main features (eyes, nose, mouth and skin) were extracted separately using computer vision algorithms, as shown in Figure 1.

A Haar cascade facial classifier [8], using filters provided by the *OpenCV* library [9] was used to identify the entire face region in an image. The face features were identified using a facial landmark detector present in the *dlib* library [10], obtained by training a shape predictor on a labeled dataset [11]. The eyes, nose and lips were extracted by calculating the minimum circle enclosing the 2D set of points representing each feature (given by the facial landmark detector). Finally, the skin area was extracted by removing the eyes and lips regions (since they offered a different colouration) and everything outside the jaw region. Any other background and hair were removed by first extracting out certain color ranges (between HEX #000000 and #646464; #a0a0a0 and #aaaaaa, which were selected based on observation) and converting images to gray-scale and thresholding in the range #5a5a5a and #969696. The removed

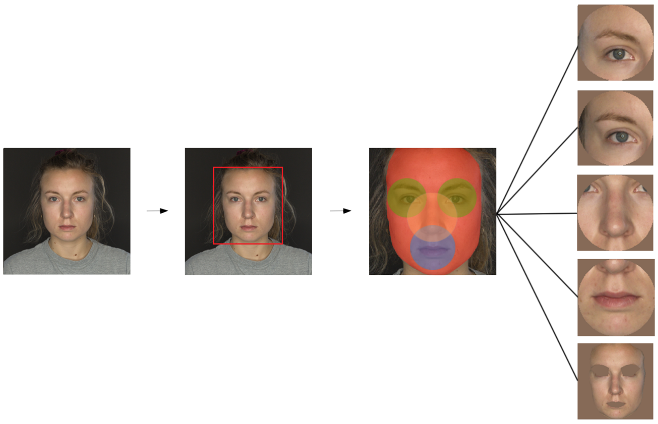


Figure 1: Feature extraction

regions were replaced with the dominant color calculated from each face region, ensuring that no other noise is passed down through the CNNs.

In an attempt to extend the size of the simulated data, all extracted fragments (except the eyes, which were already flipped) were mirrored. Thus, for each CNN, there were 104 fragments available as training data.

2.3 Deep Learning

To identify whether a facial region presents cues of sickness, a CNN was trained for each feature using Keras with a Tensorflow backend [12]. The input for the individual networks is represented by a 128 by 128 pixels RGB image, which is convolved with a convolution kernel of size (3,3) after adding padding, resulting in 128 output filters. We use a rectified linear unit (ReLU) as our activation function, the output being normalized and scaled through a layer of Batch Normalization. The subsequent layers progressively downsample the image data through groups of convolution layers (without padding), batch normalization and max pooling layers with a pool size of (2, 2). As seen in Figure 2, the final downsampling layer uses an average pooling layer (with the same pooling size), to smooth the resulting filters. Finally, the output is flattened, resulting in a tensor of length 288. This is further downsampled through two other fully-connected layers, each having a drop-out layer (30% and, respectively, 50% chance to discard data). The final layer is a fully-connected one of size 1, the activation function being changed to the sigmoidal function, ensuring that the output is a value between 0 and 1.

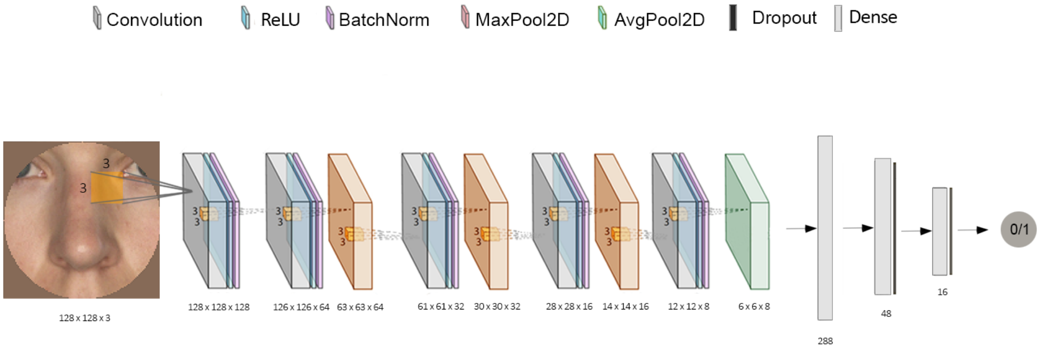


Figure 2: Architecture of a single CNN.

The second architecture employed in this study is a stacked ensemble represented by the previously mentioned CNNs (as seen in Figure 3). The networks had their final layer removed and each tensor of size 16 were concatenated, resulting in a tensor of size 64. The data is, once again, gradually downsampled through four fully-connected layers using ReLU (of size 32, 16, 4 and, respectively, 1). The final activation function is changed to the sigmoid function to ensure a value between 0 and 1. Through the previously mentioned concatenation, we expect that the stacked ensamble can learn to prioritize information coming from phenotypes (e.g. pale lips in rapport to the skin tone) that are easily recognized.

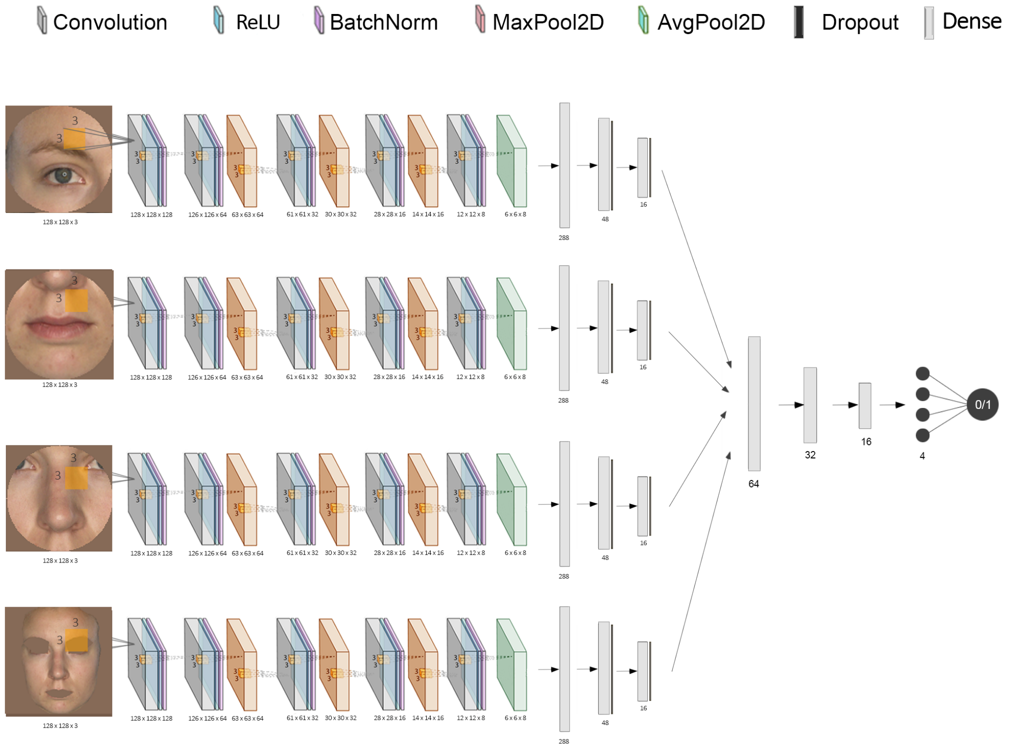
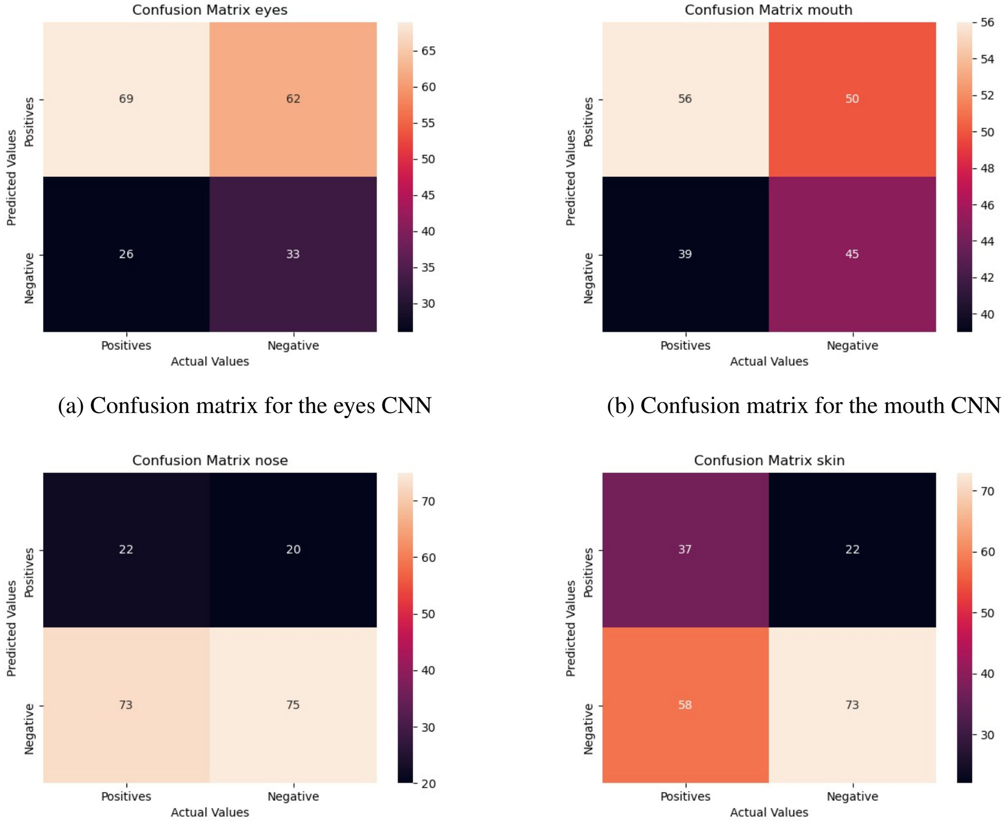


Figure 3: Architecture of the stacked CNN.

Both the CNNs and the stacked network use an Adam optimizer (adaptive moment estimation) with an initial learning rate of 0.001 and default Tensorflow values for $beta1 = 0.9$, $beta2 = 0.999$ and $epsilon = 10^{-8}$. The Adam optimizer allows for different learning rates to be computed and used for each parameter (each weight in the network), which enables a faster and finer tuning of the weights. All models used a binary cross-entropy loss function. In order to avoid over-fitting the models to the training data, model checkpoints were used to save the model with the best testing accuracy during training. Furthermore, to prevent redundantly training the networks past the over-fitting point, early stopping was also implemented. Thus, the number of training epochs was set to 50 (since over-fitting occurred rather early due to the small training set) and the patience parameter was set to 5, such that after 5 epochs with no improvements in the validation accuracy the training stopped. In practice, the training of most networks stopped after around 20 turns. Both the model checkpoints and the early stopping features were implemented using the Keras API in Tensorflow [12]. The batch size was set to 1, thus using online training to adjust the weights after each example. Experimentally, this batch sized allowed for training the networks longer before over-fitting to the training data (by avoiding local minima of the loss function through more noisy updates of the weights); and thus achieving a higher testing accuracy.

3 Results

The CNNs were trained using a 5 fold cross-validation. Thus, in each fold, the best model with regard to testing accuracy (on a fifth of the training data) was saved. The saved models were then used to make predictions on the validation data, which consists of a subset of the photographs used by Axelsson et.al. [3]. Therefore, all the results that will be reported are achieved on an external validation data set.



(a) Confusion matrix for the eyes CNN

(b) Confusion matrix for the mouth CNN

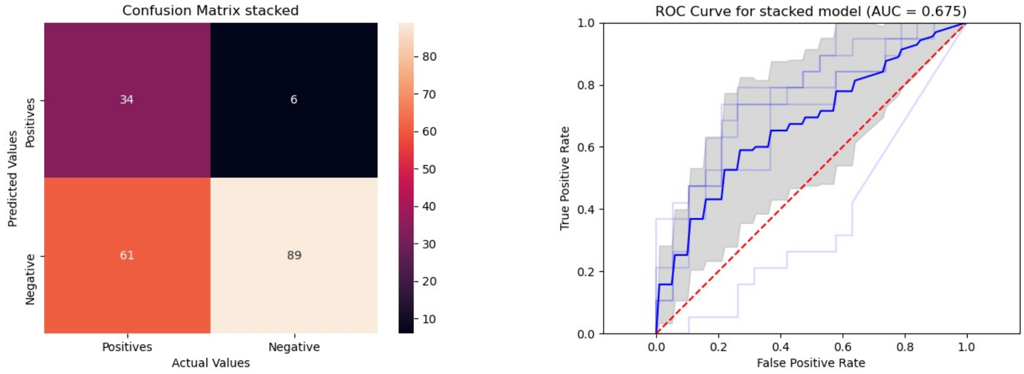
(c) Confusion matrix for the nose CNN

(d) Confusion matrix for the skin CNN

Figure 4: Confusion matrices for the individual CNNs which analyse the eyes, mouth, nose and skin.

In order to check the classification made by the CNNs for the external validation, the confusion matrix of each CNN were plotted. Given that a different model was saved in each cross-validation fold, the confusion matrices aggregate the predictions made by all saved models. The matrices are plotted in Figure 4 and in Figure 5a, for the stacked CNN. Overall, the matrices indicate that all CNNs managed to identify correctly between 23% and 72% of the sick individuals (nose and eyes CNNs, respectively) and between 35% and 93% of the healthy individuals (eyes and stacked CNNs, respectively). This suggests that in the case of all face regions, there are visible signs of illness that can be successfully detected by a CNN. However, none of the networks were able to achieve both a high sensitivity and a high specificity. For instance, the CNNs for nose (Figure 4c), for skin (Figure 4d) and the stacked CNN (Figure 5) had a specificity over 76% but sensitivities below 39%. On the other hand, the CNNs for eyes (Figure 4a) and mouth (Figure 4b) had a sensitivity over 59%, but specificity below 47%. Given that the validation set was balance with regard to the number of sick and healthy patients, accuracy is a reliable indicator of the classification performance. Therefore, the validation accuracy for the eyes is 53%, for the mouth is also 53%, for the nose is 51% and for the skin is 57% and for the stacked ensemble is 64%.

Figures 6 and 5b illustrate the Receiver Operating Characteristic (ROC) curve for each CNN. It is created by plotting the false positive rate against the true positive rate at different threshold settings. The dotted red line represents the *line of identity*, corresponding to a discrimination by chance between sick and healthy patients. The blue line represents the average ROC curve over the 5 cross-validation



(a) Confusion matrix for the CNN ensemble

(b) ROC curve plot for the CNN ensemble

Figure 5: Confusion matrices and ROC curve plot for the ensemble of CNNs that analyse individual features.

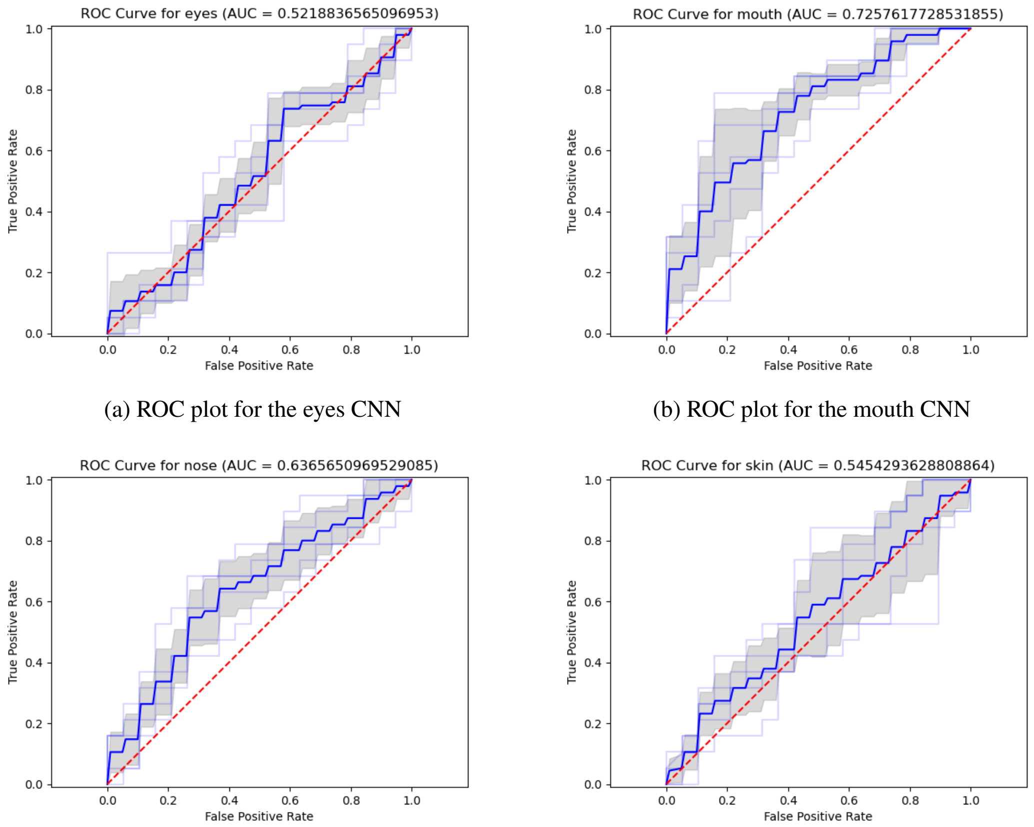
folds, while the grey lines are the ROC curves of individual folds. The area colored in grey is the confidence interval of the curve, given by the mean and one standard deviation. As suggested by the confusion matrices - which only illustrate the predictions for a fixed threshold between the classes - no CNN achieves both a high sensitivity and a low false positive rate (equivalent to a high specificity). In fact, the average ROC curve of the eyes CNN (Figure 6a) and the average ROC curve of the skin CNN (Figure 6d) are close to the the red line, meaning that the sensitivity and the false positive rate are similar and the classification is done nearly at chance level. The corresponding AUC scores are 0.52 and 0.54 for eyes and skin, respectively, meaning that the CNNs are very slightly more likely to identify a sick patient as sick than to identify a healthy patient as sick. Better AUC scores were achieved by the nose, the mouth and the stacked CNNs 0.63, 0.67 and 0.72, respectively. As observed in Figure 5b, unlike the other CNNs, the stacked ensemble is not entirely stable, given by the large confidence interval. This could be a result of the lower efficiency of the skin and eyes networks, which could provide unreliable information to the stacked network.

4 Discussion

This study aims to test whether a convolutional neural network model can be trained to discriminate between critically ill patients and healthy people. For this purpose, four CNN were trained, each targeting a different face feature among: eyes, nose, skin and mouth. Additionally, a stacked model was trained, which combined the four CNNs in order to consider all face regions. In order to analyse the correctness of the classifications made by the CNNs, the confusion matrices (Figures 4 and 5a) and the ROC curves (Figures 6 and 5b) were plotted.

The results suggest that the trained CNNs are able to classify correctly photographs of sick and healthy people above chance level. It was hypothesized that the CNNs would be able to discriminate between sick and healthy patients at least as well as non-experts could. However, compared to the classification made by non-expert, there tends to be a larger trade-off between sensitivity and specificity. Thus, the networks with a high sensitivity tend to have a low specificity and vice-versa, whereas non-experts achieved above chance levels of both sensitivity and specificity.

Interestingly, both the networks and the participants of Axelsson et.al.[3] were better at identifying healthy people rather than sick people. In the case of CNNs this could be explained by a bias in the training data; however the CNNs were trained on an equal number of examples of sick and healthy people. Nonetheless, as previously noted in the Data Pre-Processing section, the sick features were more prominent in the simulated training data than in the validation data set. Thus, it was expected that some sick individuals with less emphasized sick features would be classified incorrectly. Although, those photographs were not considered to present sick people even by expert supervisors of the experiment, they were not excluded as they representative of the natural variation of the sick features between people. Axelsson et.al.[3] reported the paleness of the lips to the feature most strongly associated with sickness,



(a) ROC plot for the eyes CNN

(b) ROC plot for the mouth CNN

(c) ROC plot for the nose CNN

(d) ROC plot for the skin CNN

Figure 6: ROC plots for the individual CNNs which analyse the eyes, mouth, nose and skin.

followed by an overall sick appearance, pale skin and hanging eyelids. The CNNs for eyes and mouth indeed had a high sensitivity, correctly classifying at least half of the sick patients. However, given that they had a low specificity, it cannot be inferred whether the paleness of the lips or the hanging eyelids were correctly interpreted by the models. Furthermore, it is not possible to make such assumptions of the exact features used by the CNNs for classification.

It is important to note that the main limitations of the experiment were the size of the data set and the validity of using simulated data. It is difficult to asses the individual impact of these limitations on the overall results, since they are obviously connected. On one hand, the simulated data did not seem to hinder the generalizability of the classification to the external validation data, since the models managed to achieve accuracy values higher than 50% (and even 64% in the case of the stacked CNN). However, the simulated data is inherently biased, even in the case of using a standardized procedure. Arguably, the rigor and the standardization of the procedure might have reduced the generalizability of the models, as individuals whose sick features are naturally more discrete were under-represented (as pointed out in the previous paragraph). On the other hand, the small size of the data set prevented the possibility of tuning the hyper-parameters of the models on a holdout subset of the data and overall reduced the generalizability power of the CNN models by not providing enough examples.

A possibly less biased option for simulating a data set of sick people would be by using a Generative Adversarial Network (GAN) which could transfer typical sick features on a photograph of a healthy person. This was in fact attempted in the early stage, but at that stage the validation photographs were not acquired yet, and the simulated data was insufficient to train such a network. Furthermore, many databases of face photographs are much more ethnically diverse than both the validation and the simulated data sets which are both strictly Caucasian. This aspect could potentially interfere in generating realistic photographs of sick non-Caucasian people.

In conclusion, the Convolutional Neural Network models were successfully trained to discriminate between sick and healthy patients slightly above chance level, but not quite managed to match the classification performance of non-experts - although some similar trends emerged such as being better at identifying healthy people as healthy rather than sick people as sick. This was in part due to the limited data available and the inherent bias of the simulated data set, which failed to consider some of the natural variation in the features of sick individuals.

References

[1] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J. D. Chiche, C. M. Coopersmith, R. S. Hotchkiss, M. M. Levy, J. C. Marshall, G. S. Martin, S. M. Opal, G. D. Rubenfeld, T. van der Poll, J. L. Vincent, and D. C. Angus, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.

[2] M. Schaller and J. H. Park, "The behavioral immune system (and why it matters)," *Current Directions in Psychological Science*, vol. 20, no. 2, pp. 99–103, 2011.

[3] J. Axelsson, T. Sundelin, M. J. Olsson, K. Sorjonen, C. Axelsson, J. Lasselin, and M. Lekander, "Identification of acutely sick people and facial cues of sickness. proceedings," *Biological Sciences*, vol. 285, no. 1870, p. 20172430, 2018.

[4] Y. Gurovich, Y. Hanani, O. Bar, G. Nadav, N. Fleischer, D. Gelbman, L. Basel-Salmon, P. M. Krawitz, S. B. Kamphausen, M. Zenker, L. M. Bird, and K. W. Gripp, "Identifying facial phenotypes of genetic disorders using deep learning," *Nature medicine*, vol. 25, no. 1, pp. 60–64, 2019.

[5] D. Massey, L. M. Aitken, and W. Chaboyer, "What factors influence suboptimal ward care in the acutely ill ward patient?," *Intensive and Critical Care Nursing*, vol. 25, no. 4, pp. 169 – 180, 2009.

[6] A. J. Henderson, J. Lasselin, M. Lekander, M. J. Olsson, S. J. Powis, J. Axelsson, and D. I. Perrett, "Skin colour changes during experimentally-induced sickness," *Brain, behavior, and immunity*, vol. 60, pp. 312–318, 2017.

[7] G. Sarolidou, J. Axelsson, T. Sundelin, J. Lasselin, C. Regenbogen, K. Sorjonen, J. N. Lundström, M. Lekander, and M. J. Olsson, "Emotional expressions of the sick face," *Brain, Behavior, and Immunity*, vol. 80, pp. 286–291, 2019.

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," vol. 1, pp. I–I, 2001.

[9] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[10] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[11] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing (IMAVIS), Special Issue on Facial Landmark Localisation "In-The-Wild"*, 2016.

[12] F. Chollet *et al.*, "Keras," 2015.

[13] H. Chang, J. Lu, F. Yu, and A. Finkelstein, "Pairedcyclegan: Asymmetric style transfer for applying and removing makeup," pp. 40–48, 2018.

Task Distribution

Data collection was entirely executed by four medical students, as part of their Bachelor Project. In that stage, our involvement was limited to consulting on the quality of the photographs necessary. Later, we were also involved in assisting the students interpret the preliminary results achieved for their thesis.

An attempt was made to develop a Generative Adversarial Network that would increase the size of the initial data set through feature transfer, as seen in the work of Chang et al. [13]. Andrei was in charge of developing it and Malina was in charge of implementing the data analysis through the CNNs. However, as the data set was so limited, the development of a GAN was no longer possible. Before arriving at the training data set, attempts were made to use CNNs for style transfer and to augment the available photographs using noise, blur, etc, which were performed by Malina. However, these methods were not used in this version of the project, but we are planing to use the style transfer algorithm for a future study, using the validation data set. Consequently, Andrei handled the data pre-processing, repurposing some of the code from the GAN and partly establishing the model architecture for the stacked CNN.

As each image had four facial regions extracted, methods to import the data properly were required. Malina focused on functions that would read the data, normalize it and group it appropriately to train the stacked ensemble.

Running the experiments was a shared task, at times we tried different approaches and parameters, Malina focused on the CNNs and Andrei on the stacked model. The resulting data was plotted by Malina as confusion matrices and Andrei wrote the code to plot the mean ROC curves.

In writing the report, Malina focused on the Introduction and the Discussion, Andrei wrote the Methods and we were both involved in writing the Results section.