Deep Oversampling Technique for 4-level Acne Classification in Imbalanced Data

Tetiana Biloborodova[1], Mark Koverha[2], Inna Skarga-Bandurova[3], Yelyzaveta Yevsieieva[4], Illia Skarha-Bandurov[5]

[1] G.E. Pukhov Institute for Modelling in Energy Engineering, 15 General Naumov Str. Kyiv, 03164, Ukraine

[2] Department of Computer Science and Engineering, Volodymyr Dahl East Ukrainian National University, 43 Donetska Street, Severodonetsk, 93400, Ukraine

[3] School of Engineering, Computing and Mathematics, Oxford Brookes University, Wheatley Campus, Oxford, OX33 1HX, UK

[4] School of Medicine, V. N. Karazin Kharkiv National University, 4 Svobody Square, Kharkiv 61002, Ukraine

[5] Luhansk State Medical University, 32 Budivel'nykiv Street, Rubizhne, 93012, Ukraine

Abstract. The current technological horizon in the Internet of Things, computer vision, deep learning and healthcare systems makes it possible to monitor some pre-existing conditions as acne vulgaris, automate the assessment of the acne severity by photo and monitor the skin health by specialists. Remote analysis of the skin condition and automatic image classification has several challenges. One of the most critical problems is imbalanced data raised because the number of clinical cases for each acne grade differs. This paper proposes a deep oversampling technique for 4-level acne classification that enables to deal with imbalanced datasets. The method was validated using several criteria. The experimental results obtained for imbalanced data sets revealed that the acne classification via proposed deep oversampling outperforms benchmark approaches.

Keywords: acne classification, remote monitoring, deep learning, oversampling technique, imbalanced data

1 Introduction

Social distancing and quarantine are steps applied worldwide to minimize social interaction between people and thus reduce coronavirus (COVID-19) transmission. Due to unprecedented distancing measures, frequent visits to the doctor without urgent need are undesirable. However, some physiological processes in the body require regular monitoring to assess the condition and receive further recommendations for care. Acne, also known as acne vulgaris, is one of these processes. Acne vulgaris is a long-term inflammatory skin disease that manifests itself in various skin lesions. Modern advances in the Internet of Things technologies, computer vision, deep learning and healthcare systems make it possible to monitor the skin condition, automate the assessment of the severity of acne and remotely monitor the state of the skin by specialists.

The main goal in the long-term perspective is to assess a response to acne therapy using remote skin monitoring. One of the most important components of remote skin condition monitoring is the stage of automatic recognition and assessment of the severity of acne using a photo of the patient taken at home. To provide patients with a reliable remote solution, we have analyzed tools and techniques for automated grading of acne vulgaris [1, 2].

A recent study on skin assessment for acne developed by a joint team from Microsoft and Nestlé Skin Health [3] aimed to solve the problem of using unlabeled image data and reducing noise in image data. In this study, the team conducted experiments to detect and classify acne using the convolutional neural network (CNN) image classification and regression. The RMSE of the recognition model is 0.482. The qualitative and quantitative analysis of the skin condition using different deep learning techniques allows us to identify the following problems in-home monitoring, which wield significant influence on recognition quality and, as a result, on the level of adoption of this technology:

- Image Variety: a face photo can be taken in different conditions (resolution, angle, light, distance, etc.) and other volumes that turn into data misbalancing.
- Data complexity: several criteria are used to assess the skin condition (oily skin, acne disease, skin radiance), each of which requires an individual approach.
- Preparation of marked images for training the model for specific criteria. Research in the field of image recognition shows that to this moment, high model accuracy can only be achieved using labelled images to train the model.
- Quality of recognition: Without considering the class imbalance problem, learning algorithms or constructed models can be overwhelmed by the majority class and can ignore the minority class.
- Acne grading scales variety: Medical aspect of this problem is in a wide variety of the grading scales; to this moment, more than 25 systems exist for evaluating acne severity. Moreover, acne measurers are divided into two camps, the graders and the lesion counters [4].

Grading is widely used for diagnosing acne severity. It is based on observations of the dominant lesions, evaluating the presence or absence of inflammation and estimating the extent of involvement. We analyzed several grading scales developed by US Food and Drug Administration IGA (Investigator's Global Assessment) [5], American Academy of Dermatology [6], France GEA (Global Acne Evaluation) [7], Ukraine [8], Japan [9]. As it was mentioned in [6], a unified grading system may be very helpful, but at present, there is no universal approach for grading and assessing severity of acne. Most of the scales [5-9] suggest 4-level acne severity rate. For this reason, in this study we also use a 4-type classification.

The majority of the acne cases are graded as a mild, while only about 1% have severe acne rate [10-12]. Class distribution in this case is not uniform that can negatively affect the model performance and lead to incorrect classification. Most contemporary works in class imbalance concentrate on imbalance ratios ranging from 1:4 up to 1:100 [13].

2 State-of-the-art

There are several approaches to deal with imbalanced datasets. In [14], an algorithm for assessing the severity of acne was presented. Photos were annotated according to IGA [5], and CNN was used for classification. The models were trained separately on three different image sizes. In [14], data imbalance was noted as one of the limitations, which did not allow obtaining a classification accuracy of more than 67%. The paper [15] proposes a system for assessing the general acne severity by counting local skin lesions. At the first stage, the system extracts labels, and then their distribution is used to evaluate the overall acne severity level. All experiments were conducted on their own hand-annotated imbalanced dataset. The authors have obtained a classification accuracy equal to 84.11%. In the study [16], the authors used CNN to construct binary and multi classifiers on two balanced datasets. An increase in the data has been carried out and obtained accuracy for any class does not exceed 81%. The authors of [17] proposed a new CNN-based AcneNet that demonstrated a very high classification accuracy equal to 99.44%, given a balanced dataset to train the proposed model. CNN from GoogLeNet is used in [18]. In their study, data augmentation was applied. As a result, 325 image areas were obtained from 164 source images and used for further classification. The achieved classification accuracy is 85.6%. A higher recognition quality (up to 99.44% accuracy) is achieved for balanced datasets for all these papers. This fact determines the relevance of studies in the direction of improving classification performance by eliminating the negative impact of the imbalanced distribution of data.

The palette of the techniques used for handling imbalanced datasets includes three levels: data-layer, algorithm-layer, and evaluation layer. The methods used at the data layer mostly deal with resampling and are more general because they do not rely on the chosen classifier. There are two basic techniques in this layer, they are oversampling and undersampling [19]. Oversampling comprises from elimination of the negative impact of uneven distribution by creating new samples from the minority class while undersampling reduces negative effect of uneven distribution by discarding internal samples in the majority class. Hybrid methods utilize combinations of these two for data balancing. At first glance, the oversampling and undersampling seem functionally equivalent, since they both resize the original dataset and can provide the same proportion of balance. However, removing objects from the majority class can cause the classifier to miss important concepts related to the majority class [20]. Thus, in order to avoid the data loss, especially when dataset is not huge enough, it makes sense to solve data imbalance using oversampling.

3 Methodology

In this section, we describe our approach for automatically recognition of the severity of acne based on deep learning with series of pre-processing techniques formulated below.

Following [20] we assume imbalanced dataset S with m objects $|S| = m$, as a pair $S = \{(x_i, y_i)\}$, $i = 1, ..., m$, where $x_i \in X$ denotes the instance in the n-dimensional feature space $X=\{f_1, f_2, ..., f_n\}$, and $y_i \in Y = \{1, \ldots, C\}$ denotes the class identifier label associated with x_i. in particular, when C = 2 we deal with binary classification task. In addition, we define subsets $S_{min} \subset S$ and $S_{maj} \subset S$, where S_{min} is the set of minority objects of classes in S, and S_{maj} is the set of objects of prevailing classes in S, so that $S_{min} \cap S_{maj} = \{\Phi\}$ and $S_{min} \cup S_{maj} = \{S\}$. Based on these assumptions the problem of acne recognition can be formulated as a 4-type classification task.

Thereafter, the multiclass classification problem is transformed into a regression task [21]. This transformation helps to reduce possible subjectivity in the assessment of the acne severity at the stage of data annotation. Since the acne levels class have a meaningful ordering, we assign the numeric values from 0 to C to each of the corresponding C classes labels. Formally, it presented as follow $y \subseteq S = \{1, \ldots, C\} \xrightarrow{transformation} y \in \mathbb{R}^n$. The values of the class variable are transformed into numbers. A numeric target is corresponding to a level of acne severity.

Thresholding is also used to transform back a regression prediction into a classification prediction $\hat{y} \in \mathbb{R}^n \xrightarrow{transformation} \hat{y} \subseteq S$, where $\hat{y}$ is a prediction.

The proposed methodology includes the several steps, as shown in Fig. 1. Firstly, we extract patches from original images of human faces using one of the two pre-trained models. Then data augmentation, feature extraction, and data oversampling are conducted. Finally, results obtained after oversampling are fed to CNN for model training and evaluation.

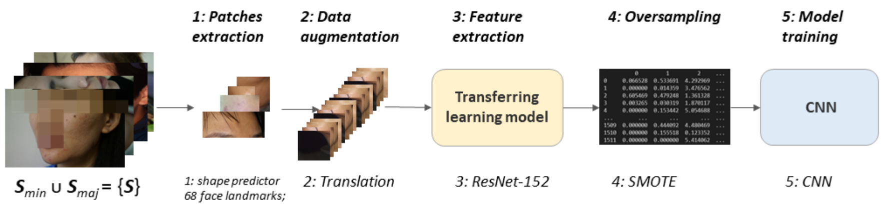

Fig. 1. Overview of the proposed approach

Step 1: Patches extraction. At the first stage, preliminary processing of images is carried out for capturing skin areas from the face image. The severity of acne does not depend on the location; it depends on the volume and severity of the lesion on the patient's face. From this prospective, the separate patches of the face image should give results that are more informative relatively to the entire face. The result of this step is a dataset S consisting of m patches extracted from facial images. Each patch inherits the image class label.

Step 2: Data augmentation. CNNs are spatially sensitive, which leads to insufficient recognition quality when using a limited number of images for network training. That is, in the case when the test images have new acne lesions that did not occur in the training images, the trained model is not able to generalize the training results on the test data. To overcome this issue, we use translation of an image section. Translation involves moving an image around the x and y axes in specified directions by a specified number of pixels. The sliding translation was used to enlarge these areas of the images. The result of this step is a dataset S consisting of an increased number of patches. Each translated patch inherits the class label of the original patch that was used for augmentation.

Step 3: Feature extraction. We utilised the transferring learning paradigm [22] and a pre-trained ResNet-152 to extract features from training set of acne images. The result of this step is a dataset $S = \{(x_i, y_i)\}$ consisting of the features extracted using the trained model, where x_i is the vector of extracted features of the patch m_i, and y_i is the class label which denotes the severity of acne associated with x_i.

Step 4. Oversampling. We use oversampling to balance the number of dataset objects for each class. Formally, oversampling can be represented as follows. Any objects generated from the dataset S are denoted as E, with disjoint subsets of E_{min} and E_{maj} representing the minority and majority of the E objects, respectively, whenever they are applied. The random oversampling process is implemented by adding a set E selected from the minority class: for a set of randomly selected minority examples in S_{min}, increase the original set S by replicating the selected examples and adding them to S. Thus, the number of typical examples in S_{min} increases by $|E|$, and the balance of the class distribution of S is adjusted accordingly. This provides a mechanism for changing the degree of balance in the distribution of classes to any desired level. The result of this step is a dataset $S = \{(x_i, y_i)$ consisting of extracted and generated features, where x_i is a vector of extracted and generated features of patches m_i, and y_i is the class label associated with x_i.

Step 5: Model training. The extracted and generated features are passed to train a CNN model to classify acne severity. At this stage, the classification task was changed into a regression task by defining acne severity grades as integer equivalents. It was done to reduce possible subjectivity in the assessment of the acne severity at the stage of data annotation. The inverse transformation was done using [0.5, 1.5, 2.5] as the edge list. To divide the underlying truth and predicted severity levels into categorical severity levels, we used severity label < 0.5 to assign level 0, a severity label in the range 0.5 - 1.5 to assign level 1, labelled severity in the 1.5-2.5 range to assign level 2, and labelled > 2.5 to assign level 3.

Step 6: Model evaluation . The test pipeline includes the extraction of facial image patches, feature extraction using the ResNet-152, patch classification from trained CNN, obtaining the average predicted value of the acne severity class from the face image, and assessing the quality of recognition. Model evaluation is implemented on test data. In this step, each patch of the face image was evaluated, and the average rate for all patches of the same face was assigned to the image class.

4 Results

For this study the ACNE04 [15] the open dataset was used. The ACNE04 includes 1457 face images and expert annotations according to the Japanese rating scale. The dataset has the following acne severity annotations: level 0 – Mild, level 1 – Moderate, level 2 – Severe, level 3 – Very severe. All images were taken at an angle of approximately 70 degrees from the front of the patient and manually annotated by specialists. The distribution of objects in accordance with the class is presented in Table 1.

Table 1. Image class distribution in ACNE04 [14] used for experiments

Acne severity	Class	Dataset	Train	Test
level 0 Mild	0	513	410	103
level 1 Moderate	1	633	506	127
level 2 Severe	2	507	146	362
level 3 Very severe	3	108	103	6
Total:		1457	1165	292

We used 1165 images to train the model and 292 images to test the model which corresponds to a distribution of 80% for training and 20% for testing. In accordance with a four-point scale for assessing the severity of acne, the images are labeled as follows. The images are annotated as follows: level 0 Mild assigned to class 0, level 1 Moderate - to class 1, level 2 Severe - to class 2, level 3 Very severe - to class 3. A study by Microsoft [3] was used as a benchmark for the experiment.

At the first stage, the procedure of image preprocessing was carried out. This step utilizes two pre-trained models. The first one is shape_predictor_68_face_landmarks model [23], which returns 68 characteristic points for each person given the bounding box of their face. This facial landmark model detects landmarks of frontal borders or faces slightly away from the camera while two open eyes can be observed in the image. Then the patches of the face, cheeks and chin are extracted from the original images and stored as image files in the main directory. In case when the face is very away from the camera and both eyes were not recognized in the image, shape_predictor_68_face_landmarks cannot reveal the boundaries of the face. When this occurs, an alternative, the One Eye model [24], was used. When the One Eye model identifies an eye in the image, areas of the skin of the face and cheeks are cropped from the original image and stored as separate image files (patches) in the main directory. In all scenarios, when the selection of patches from the photo is successful, the only selected patches with reference to the original image are used for further steps. Otherwise, if neither of these models cannot process the image, the original image was used entirely for further processing and analysis. A fragment of a set of dependencies and binding patches to the original image is given below.

After preprocessing stage, we obtained 3806 image patches. For data augmentation, sliding translation of patches was used. All images were normalized to 224 by 224 pixels. Further, the feature extraction from each patch was carried out using the ResNet-152 model [25]. Each patch is bound to the original image and thus the extracted features inherit the dependencies of the patches. At the next step we used Synthetic Minority Oversampling Technique (SMOTE) [26]. The total number of samples for each class was fitted to the most numerous class in augmented dataset. Information about the data distribution at each stage is summarized in Table 2.

Table 2. Data distribution for each preprocessing step

Class	raw data	patched data	augmented data	oversampled data
0	410	1367	3556	4333
1	506	1716	4333	4333
2	146	476	1843	4333
3	103	247	1514	4333
Total	1165	3806	11246	17332

Data generated at the oversampling stage were used to train a CNN classifier. Object classes were defined as integers.

5 Discussion

The results of comparison between the proposed approach and baseline methods without oversampling are shown in Table 3.

Table 3. The results of experiments with and without oversampling

Metrics	without oversampling	with oversampling	delta,%
RMSE	0.422356	0.397419	5.904261
MAE	0.325285	0.297179	8.640423
MedAE	0.256841	0.223299	13.05944
EV	0.826736	0.873874	5.7017
MAPE	0.199264	0.171855	13,75512
R^2	0.826682	0.873646	5.68102
MPD	0.102844	0.087945	14.48699
MGD	0.070191	0.061218	12.78369

Resulting RMSE, MAE, MedAE, MAPE, MPD, MGD criteria show a smaller error, while the EV, R^2 criteria show a higher value using the proposed approach, which indicates a higher quality of the model using the proposed approach. We also compared the performance of the proposed approach with other studies aimed at determining the severity of acne from a photo. The criteria RMSE, MAE, MedAE, R^2 can only be compared between models whose errors are measured in the same units, i.e. using the same data.

A set of metrics for evaluating the model quality as *Accuracy*, *error rate* (*ER*) were computed after converting the results back from continuous to discrete scale using [0.5, 1.5, 2.5] as a list of edges for separating true labels and predicted ones. The *Precision*, *Recall*, *F1-score* metrics were used for evaluating the classification performance of each class. The results are shown in Table 4.

Table 4. Evaluation of classification result

Metrics	Without oversampling (%)	With oversampling (%)
Precision	82.25	82.25
Recall	82.25	84.75
F1-score	82.5	85
Accuracy	80	85
ER	20	15

The oversampling made it possible to achieve a higher accuracy for each class and in general for the results of classification.

Comparison of the obtained results with the benchmark models discussed above is presented in Table 5.

Table 5. Comparison of acne classification studies

Approach for acne classification	Balanced / Imbalanced	Accuracy (%)	ER (%)
Junayed et al. [16]	Balanced	99.44	0.56
Lim et al. [13]	Imbalanced	67	33
Wu et al. [14]	Imbalanced	84.11	15.89
Ours	Imbalanced	85	15

As can be seen, the proposed approach showed the highest accuracy and the lowest error rate in comparison with studies with imbalanced data, however, it could not surpass the results for works where input data was initially balanced, which sounds natural. At the same time, it can be argued that for current task, the proposed approach has shown a significant advantage for acne classification in conditions of imbalanced data which in turn shape expectation to use this technique for remote self-assessment.

6 Conclusion

The paper presents an approach for assessing the severity of acne from a photograph under conditions of imbalanced data. The proposed technique includes image preprocessing, data augmentation, oversampling, feature extraction, training and model evaluation. To reduce the subjectivity of assessing the severity of acne by different experts at the stage of image annotation, the classification task is transformed to a regression task by defining acne severity classes as integer equivalents. When obtaining results on test or new unlabelled data, the predicted values obtained for each sample are reduced to the values of the acne severity classes using the described scaling. To assess the quality of the model, we used different criteria that show a smaller error and a higher value comparing benchmark models, which indicates a higher quality of the model using the proposed approach. The model showed the highest accuracy for studies using imbalanced data and the lowest error rate. Thus, we can talk about the advantage of using the proposed technique for image analysis in conditions of imbalanced data and could suggest it for further investigation.

References

1. Arhubdulal/DermaVigil. Github.com. [Online] https://github.com/arhubdulal/DermaVigil.
2. TroveSkin. Your Smart All-In-One Skincare Tracker. Troveskin.com. [Online] https://www.troveskin.com/.
3. Zhao, T., Zhang, H. and Spoelstra, J. (2019). A Computer Vision Application for Assessing Facial Acne Severity from Selfie Images. *arXiv preprint arXiv*:1907.07901.
4. Witkowski, J.A. and Parish, L.C. (2004). The assessment of acne: an evaluation of grading and lesion counting in the measurement of acne. *Clinics in dermatology,* 22(5), 394-397. doi: 10.1016/j.clindermatol.2004.03.008. PMID: 15556725.
5. Nast, A., Dreno, B., Bettoli, V., Degitz, K., Erdmann, R., Finlay, A.Y., Ganceviciene, R., Haedersdal, M., Layton, A., López-Estebaranz, J.L. and Ochsendorf, F. (2012). European evidence-based (S3) guidelines for the treatment of acne, 1-29. doi: 10.1111/j.1468-3083.2011.04374.x.
6. Zaenglein, A.L., Pathy, A.L., Schlosser, B.J., Alikhan, A., Baldwin, H.E., Berson, D.S., Bowe, W.P., Graber, E.M., Harper, J.C., Kang, S. and Keri, J.E. 2016. Guidelines of care for the management of acne vulgaris. *Journal of the American Academy of Dermatology*, 74(5), pp.945-973.
7. Dreno, B., Poli, F., Pawin, H., Beylot, C., Faure, M., Chivot, M., Auffret, N., Moyse, D., Ballanger, F., Revuz, J., (2011). Development and evaluation of a Global Acne Severity scale (GEA scale) suitable for France and Europe. *Journal of the European Academy of Dermatology and Venereology*, 25(1), 43-48.
8. ACNE: Evidence-based clinical guideline (2017), Ukraine. [Online] https://www.dec.gov.ua/wp-content/uploads/2019/11/akn_akne.pdf.
9. Hayashi, N., Akamatsu, H., Kawashima, M., and Acne Study Group (2008). Establishment of grading criteria for acne severity. *The Journal of Dermatology*, 35(5), 255–260.

10. Heng, A.H.S., Say, Y.H., Sio, Y.Y., Ng, Y.T. and Chew, F.T. (2021). Epidemiological Risk Factors Associated with Acne Vulgaris Presentation, Severity, and Scarring in a Singapore Chinese Population: A Cross-Sectional Study. *Dermatology*, 1-10.
11. Tayel, K., Attia, M., Agamia, N. and Fadl, N. (2020). Acne vulgaris: prevalence, severity, and impact on quality of life and self-esteem among Egyptian adolescents. *Journal of the Egyptian Public Health Association*, 95(1), 1-7.
12. Durai, P.C.T. and Nair, D.G. (2015). Acne vulgaris and quality of life among young adults in South India. *Indian journal of dermatology*, 60(1), 33-40.
13. Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence,* 5, 221–232. https://doi.org/10.1007/s13748-016-0094-0
14. Lim, Z.V., Akram, F., Ngo, C.P., Winarto, A.A., Lee, W.Q., Liang, K., et al. (2020) Automated grading of acne vulgaris by deep learning with convolutional neural networks. *Skin Research and Technology*, 26(2), 187-192. doi: 10.1111/srt.12794.
15. Wu, X., Wen, N., Liang, J., Lai, Y.K., She, D., Cheng, M.M. Yang, J. (2019). Joint acne image grading and counting via label distribution learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10642-10651, doi: 10.1109/ICCV.2019.01074.
16. Shen, X., Zhang, J., Yan, C., Zhou, H., (2018). An automatic diagnosis method of facial acne vulgaris based on convolutional neural network. *Scientific reports*, 8(1), 1-10
17. Junayed, M.S., Jeny, A.A., Atik, S.T., Neehal, N., Karim, A., Azam, S., Shanmugam, B. (2019). AcneNet-A deep CNN based classification approach for acne classes. In *2019 12th International Conference on Information & Communication Technology and System (ICTS)*, 203-208. IEEE. doi:10.1109/icts.2019.8850935
18. Alarifi, J.S., Goyal, M., Davison, A.K., Dancey, D., Khan, R. and Yap, M.H., (2017). Facial skin classification using convolutional neural networks. In *International Conference Image Analysis and Recognition,* 479-485. Springer, Cham.
19. Yijing, L., Haixiang, G., Xiao, L., Yanan, L., Jinling, L. (2016). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems,* 94, 88–104. doi:10.1016/j.knosys.2015.11.013
20. He, H., Garcia, E.A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. doi:10.1109/tkde.2008.239.
21. Osojnik, A., Panov, P. & Džeroski, S. (2017). Multi-label classification via multi-target regression on data streams. *Machine Learning*, 106(6), 745–770. https://doi.org/10.1007/s10994-016-5613-5.
22. Shorten, C., Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1). doi:10.1186/s40537-019-0197-0.
23. Davisking/dlib-models. Github.com. [Online] https://github.com/davisking/dlib-models
24. OpenCV. Github.com. [Online] https://github.com/opencv/opencv/blob/master/data/haarcascades/haarcascade_eye.xml.
25. The Microsoft Cognitive Toolkit. Cntk.ai. [Online] https://www.cntk.ai/Models/Caffe_Converted/ResNet152_ImageNet_Caffe.model
26. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.