2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-1179-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/EMBC46164.2021.9631065

Solving the problem of imbalanced dataset with synthetic image generation for cell classification using deep learning*

David Kupas, Balazs Harangi

Abstract — **The low number of annotated training images and class imbalance in the field of machine learning is a common problem that is faced in many applications. With this paper, we focus on a clinical dataset where cells were extracted in a previous research. Class imbalance can be experienced within this dataset since the normal cells are in a great majority in contrast to the abnormal ones. To address both problems, we present our idea of synthetic image generation using a custom variational autoencoder, that also enables the pretraining of the subsequent classifier network. Our method is compared with a performant solution, as well as presented with different modifications. We have experienced a performance increase of 4.52% regarding the classification of the abnormal cells.**

Clinical Relevance — **We extract images from cervical smears, using digitized Pap test. When working with these kinds of smears, a single one can contain more than 10,000 cells. Examination of these is done manually by going over each cell individually. Our main goal is to make a system that can rank these samples by importance, thus making the process easier and more effective. The research that is described in this paper gets us a step closer to achieving our goal.**

I. Introduction

The Papanicolaou smear test [1] - also known as the Pap test - is a cervical screening method, where cells are collected from the outer opening of the cervix. Using special scanners (3DHistech Pannoramic 1000), digital images are collected from Pap smears. By examining the cells on these smears, an early stage of cervical cancer can be detected.

The cell examination process is an extremely exhaustive task, in which cytological experts manually check each cell on the extracted smear in order to locate the abnormal ones. As such, this is a time-consuming process and consequently an expensive task. Our main goal is to develop an automatic screening system that can extract the cells on the smears and respectively separate them into normal and abnormal classes. The main aim of this system is to sort the smears on the basis of their severity, using machine learning-based automatic screening. In this way, this system will be able to highlight those cases where immediate intervention and possible second grading are required. This is a challenging task in which we are faced with many different difficulties that we overcome using our proposed solutions and some machine learning methods.

The extracted smears may contain over 10,000 cells for each patient. The main difficulty is that only a low amount of these cells is abnormal in the case of a diseased patient. This means that if we want to classify the extracted cells by using an approach based on machine learning, the dataset that we use for training would be highly imbalanced while also having a low number of images.

When working with machine learning algorithms, the class imbalance problem is a difficult obstacle to overcome. Generally, an even distribution between classes is preferred when considering binary classification. Many solutions were proposed for solving this problem, such as oversampling; undersampling, and cost-sensitive learning [2]. Furthermore, a viable method is to generate synthetic data points for the minority class, as also done in [3].

In this paper, we address the problem of an imbalanced training set and propose a solution, that uses a variational autoencoder model in order to generate synthetic, abnormal cells, thus balancing out the dataset. The same convolutional neural network architecture was used to extract the most important features of images at the encoder part of the variational autoencoder and to classify the cell images. The weights learned from training the autoencoder were also used, to initialize the DenseNet classifier network. In this way, we overcome the problem of a small training set and managed to get a higher classification performance despite the imbalance regarding the classes.

The rest of this article is organized as follows: in the next section, we present the available dataset in detail, the required preprocessing steps and the specific numerical data for each set used in the training and testing process. The following section is about the classification of each cell into two classes, highlighting the model architecture that we used and the training process. After this, we briefly describe the methods used to overcome the class imbalance problem. Finally, every result is thoroughly presented, finishing with the conclusions regarding the proposed methods.

II. Dataset

The input data in our case is the digitalized image that has been collected using the special microscopes. An example of this can be seen in Fig. 1. In our previous work [4], we focused on the segmentation of cell groups on these smears. In order to achieve that, first, we needed to make smaller, 2,000 x 2,000

* This research was supported by the ÚNKP-19-2-I-DE-345 and the ÚNKP-20-5-DE-31 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund. Moreover, the research was supported in part by the Janos Bolyai Research Scholarship of the Hungarian Academy of Sciences, and the GINOP-2.2.1-18-2018-00012 supported by the European Union, co-financed by the European Social Fund.

David Kupas, and Balazs Harangi are with the Department of Computer Graphics and Image Processing, Faculty of Informatics, University of Debrecen, Kassai Road 26, Debrecen, Postcode 4028 (e-mail: kupas.david@inf.unideb.hu, harangi.balazs@inf.unideb.hu)

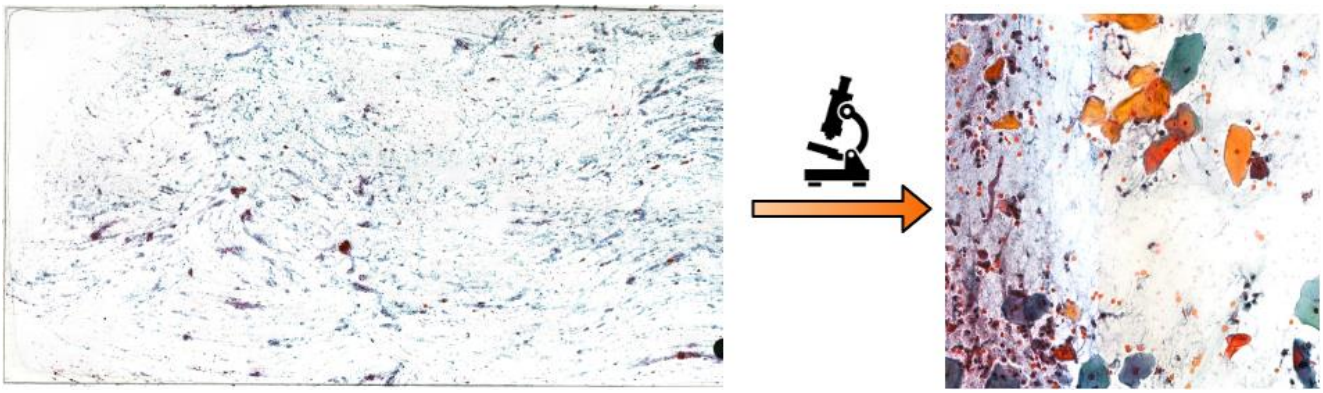

Figure 1. A treatable sized (2,000 x 2,000) sub-image is extracted from the digital smear.

sized images. Then we used an ensemble system, applying methods based on fully convolutional networks [5] and super-pixel-based segmentation [6]. In this way, we achieved an algorithm, which can accurately segment cell regions from the input image. In order to make the data more heterogeneous, many smears were processed from over 100 patients. Using randomly selected outputs of this segmentation, we can focus on the classification of each cell into normal and abnormal classes respectively.

In order to build the required dataset, after the histological examination of the cells, an annotation process was done by a total of 3 experts. Since the number of cell images that we plan to annotate is high, every expert received different images to make the process less time-consuming. They labeled each individual cell into two distinct classes based on whether they were normal or abnormal. In this way, a total of 2,527 cells were annotated. These are made up of 2,164 normal and 363 abnormal cells. Before training any model, we split these images into training and test sets. The training set consists of 1,727 normal and 294 abnormal cells, while the test set is made of 437 normal and 69 abnormal cells (~20% of the whole dataset). The training set was used to train the models, and also the test set was used to validate our model and measure model performance. Some examples of normal and abnormal cells can be seen in Fig. 2.

III. Classification of Cells

By experimenting with different convolutional neural network architectures, we used a model based on the widely used DenseNet architecture [7]. Our intention with this network is to obtain a benchmark result, meaning that we have only used the original dataset for the training process, without

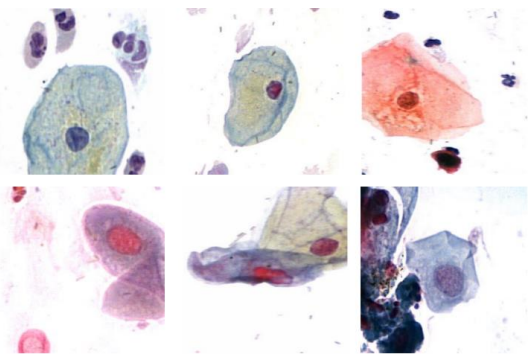

Figure 2. Sample normal images are presented in the first row and abnormal in the second one.

addressing the problem of an imbalanced dataset. Originally, the DenseNet model is used to classify images in 1,000 classes. Since in our case a binary classification is needed, we have truncated the top layers and attached the following layers to the base DenseNet model: global average pooling layer, dense layer with 1,024 neurons, ReLU activation function, fully connected layer with two output neurons, and softmax activation function. The training was done with random weight initialization, over 20 epochs, using Adam optimizer with a learning rate of 1e-5.

In order to be able to compare our proposed method with existing solutions regarding the dataset imbalance problem, we used the DenseNet based network and trained it with a custom modified loss function. The widely used cross-entropy loss can be customized by adding a weight mask and setting it in a way that the network is penalized more when misclassifying abnormal cells. For this, we used our own implementation based on the Real-World-Weight Cross-Entropy (rwwce) [8], which can be described with the following formula:

$$J = -\frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M}\left[w_{fn}^{k}\cdot y_{m}^{k}\cdot\log\left(h_{\theta}(x_{m},k)\right)\right. \\ \left. +\sum_{\substack{k'=1\\k'\neq k}}^{K} w_{fp}^{k,k'}\cdot y_{m}^{k}\cdot\log\left(1-h_{\theta}(x_{m},k')\right)\right] \quad (1)$$

where M is the number of training examples. K denotes the number of classes, also y_m^k is the target label for training example m regarding class k. h_θ stands for the model with neural network weights of θ. x_m indicates the input for training example m. w_{fn}^k is the cost of a false negative over a true positive, and finally $w_{fp}^{k,k'}$ denotes the cost of a false positive of class k' over a true negative, when the true positive is k.

With the goal of further increasing classification performance, we propose to balance out the dataset in order to make an equal distribution between normal and abnormal cells. For this, we used a variational autoencoder [9] trained on cells from the training set. This will be elaborated additionally in the next section.

IV. Solution for the Imbalanced Dataset Problem

Since we have an imbalanced dataset, meaning that we have only 294 abnormal and 1,727 normal cells in the training set, our goal is to generate synthetic abnormal cell images to make them equal to the number of normal cells. The intention is to reduce the bias towards the majority class and increase classification performance. It is important to note here, that the main goal of the screening process is to find all abnormal cells in case of the patients. The identification of normal cells is not enough to determine whether a patient is healthy or not. Keeping this in mind, we are mainly focusing on the performance regarding the correct classification of the abnormal cells in the test set. Naturally, normal cells cannot be neglected, meaning that if we concentrate only on the abnormal ones, we could easily manage to find all of them – but at a high cost. In the sense that in this case the normal cells would be classified with a very low performance, thus rendering the result irrelevant.

For the purpose of synthetic image generation, we have used a variational autoencoder-based solution, which is used to learn the latent space representation of the cell images. Then we sampled a random point from this space, which was decoded into a synthetic image afterward. The architecture of the model network can be split into two main parts: the encoder, and the decoder. In our case, the first part (encoder) has the same architecture as the DenseNet model. This is done with the aim to pretrain the classification network to overcome the problem caused by the low number of training images. After encoding the images using this architecture, we can get the mean and variance vectors of the latent space. In order to maintain the spatial organization of the preceding layer outputs, we have slightly modified the traditional variational autoencoder architecture: instead of using flatten layers, for this, we used two convolutional layers. In this way, we get two vectors with the size of 5x5x256. A custom sampling layer was then added to the network, which uses these vectors to define a latent space and sample a point from it. As the decoder part of the network, we use the following block six times: convolutional layer, batch normalization, leaky relu, upsampling. As the result of the decoder, we have an output of the exact shape as the input, meaning that by using the network we deconstruct and then rebuild an image, which is based on sampling from the learned latent space. We trained the network using the training set over 300 epochs. The traditional variational autoencoder dual loss function was applied, meaning the combination of the reconstruction loss and the regularization loss [10].

After successfully training the variational autoencoder, we used it to generate synthetic images. This was done with the usage of solely the abnormal cell images from the original training set. In this way, we generated 1,433 images that are added to the abnormal cells in the training set. Consequently, obtaining a balanced dataset that has in total 3,454 images, with equal distribution among the classes. Some examples of synthetic abnormal images can be seen in Fig. 3. Note that these images appear blurry. In order to nullify the potential error that this could make (i.e. neural network picking blurry images as abnormal), the test set does not have any synthetic or blurry data.

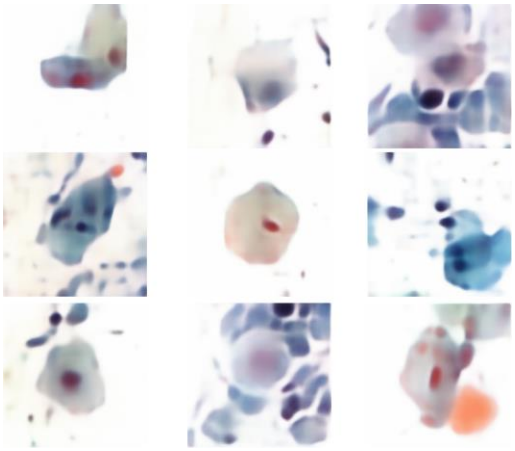

Figure 3. Generated abnormal synthetic images.

This extended training set was used for training the classifier network, which was also fine-tuned with the usage of the weights of the encoder part of the variational autoencoder. Hopefully, these pretrained weights will provide a better starting point than the random initialized ones, so we can enhance the classifier network. After the pretrained weights initialization, the DenseNet network was additionally trained with the custom weighted loss function.

V. Results

Each evaluation mentioned in this section was measured using the same test set as presented in Section II. The result of the overall evaluation can be seen in Table I., where the values, regarding the different measurements, mean the average and standard deviation derived from 5-fold runs. For the discussion of the results, two standard metrics were selected: recall and precision. In order to focus on the performance regarding the abnormal cells, we also used the balanced accuracy [11] to show a complex view about the performance and to avoid misleading simple accuracy in case of an imbalanced test set. This can be formulated as follows:

$$\frac{1}{C} \times \sum_{i=1}^{C} \frac{p_i}{n_i} \tag{2}$$

TABLE I. Evaluation of Different Setups

	Normal		**Abnormal**		**Total**
	Precision	***Recall***	***Precision***	***Recall***	***Balanced Accuracy***
$\text{DenseNet}_{\text{orig}}$	0.924 ± 0.005480	0.970 ± 0.007070	0.726 ± 0.041590	0.484 ± 0.053670	0.7150 ± 0.021210
$\text{DenseNet}_{\text{rwwce}}$	0.934 ± 0.008940	0.924 ± 0.020740	0.546 ± 0.047750	0.596 ± 0.054130	0.7580 ± 0.022800
$\text{DenseNet}_{\text{rwwce}}$ trained on extended dataset	0.944 ± 0.005477	0.924 ± 0.026077	0.588 ± 0.069785	0.650 ± 0.044159	0.7880 ± 0.028284
pre-$\text{DenseNet}_{\text{rwwce}}$ trained on extended dataset	**0.954 ± 0.011402**	**0.888 ± 0.030332**	**0.602 ± 0.203887**	**0.750 ± 0.100000**	**0.8032 ± 0.020729**

where C is the number of classes, p_i denotes the number of correct predictions of class i, and n_i stands for the number of data items in the dataset for class i.

As a first scenario, we present the results made by the core DenseNet model, using the original dataset, with random weight initialization and without the weighted loss function (*DenseNet$_{orig}$*). This will be used as the benchmark performance. As it can be seen, the precision and recall of the normal cells are quite high, however, the problem is that the abnormal cells are misclassified – as the balanced accuracy suggests. The same model was trained with the custom weighted loss function according to formula (1) (*DenseNet$_{rwwce}$*). Based on these results, we can note that the weighting method can help to increase performance regarding the abnormal class, and balanced accuracy has increased to 0.7580 from 0.7150. In the meantime, the recall of the normal class and the precision of the abnormal class have lowered to 0.924 and 0.546 respectively. We also present the results of the DenseNet model, trained using the extended dataset, where synthetic images were also added with the usage of the variational autoencoder (*DenseNet$_{rwwce}$ trained on extended dataset*). In this case, we got a higher overall performance: a value of 0.7880 balanced accuracy, as well as a higher, 0.650 value of the recall of the abnormal class.

Finally, we also comment on the result where we apply pretrained weights for the classifier network, meaning that we use the weights from the encoder part of the variational autoencoder in order to initialize the classifier network. With the classifier network fine-tuned in this way and the usage of the custom weighted loss function (*pre-DenseNet$_{rwwce}$ trained on extended dataset*), we got the overall best performance with a balanced accuracy value of 0.8032. It can also be noted that the recall of the abnormal class is the highest in this case with a value of 0.750. In order to validate also the different model performances regarding the found abnormal cells in the test set, confusion matrices are shown in Fig. 4. As it can be seen, our proposed solution found the most number of abnormal cells.

VI. Conclusion

In this paper, we have presented a custom generative network-based approach to expand the minority class using a clinical dataset. We have used a DenseNet based model with the original dataset in order to get the benchmark results. Furthermore, we have described our method of using a variational autoencoder for synthetic image generation purposes as well as using the weights from the encoder to fine-tune the classifier network. We have also done an exhaustive testing of different modifications.

Results show, that using the extended dataset did increase the overall performance of the model. Also expanding the dataset with synthetic images that were generated using a variational autoencoder, together with the custom weighted loss function and the fine-tuning method, in which the DenseNet classifier is pretrained using the weights from the encoder, did significantly improve the results. Overall, this method had the best performance, especially regarding the abnormal cells of the dataset, which is of most importance when dealing with this kind of screening process.

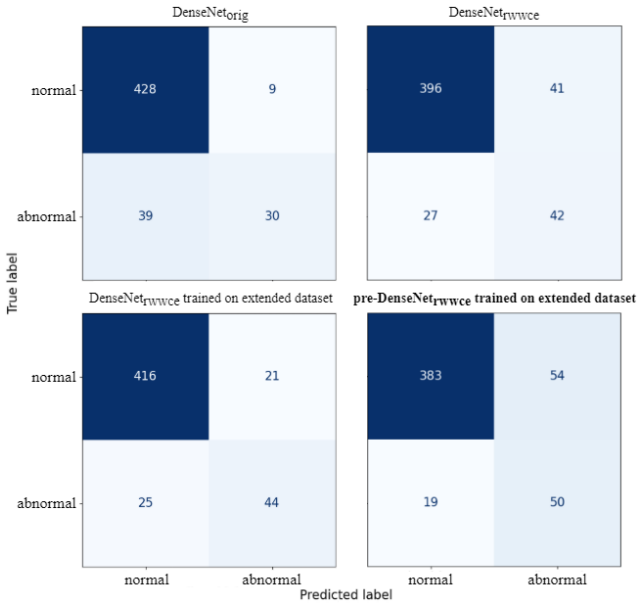

Figure 4. Confusion matrices of all tested methods.

References

[1] G. N. Papanicolaou, and H. F. Traut, "The diagnostic value of vaginal smears in carcinoma of the uterus**this study has been aided by the commonwealth fund. Presented before the new york obsterical society, march 11, 1941." *American Journal of Obsterics and Gynecology*, vol. 42, no. 2, pp. 193–206, March 1941.

[2] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," *2016 International Joint Conference on Neural Networks (IJCNN)*, Vancouver, BC, Canada, 2016, pp. 4368–4374.

[3] S. Sharma, C. Bellinger, B. Krawczyk, O. Zaiane and N. Japkowicz, "Synthetic Oversampling with the Majority Class: A New Perspective on Handling Extreme Imbalance," 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 2018, pp. 447–456, doi: 10.1109/ICDM.2018.00060.

[4] B. Harangi, J. Toth, G. Bogacsovics, D. Kupas, L. Kovacs and A. Hajdu, "Cell detection on digitized Pap smear images using ensemble of conventional image processing and deep learning techniques," 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA), Dubrovnik, Croatia, 2019, pp. 38–42, doi: 10.1109/ISPA.2019.8868683.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* June 2015.

[6] Z. Lu, G. Carneiro, and A. P. Bradley, "Automated nucleus and cytoplasm segmentation of overlapping cervical cells," *International Conference on Medical image Computing and Computer-Assisted Intervention,* Springer, 2013, pp. 452–460.

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[8] Y. Ho, and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, pp. 4806–4813, Dec 2019.

[9] D.J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *International conference on machine learning*, pp. 1278-1286, June 2014

[10] D.P. Kingma, and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[11] A. Gupta, N. Tatbul, R. Marcus, S. Zhou, I. Lee and J. Gottschlich, "Class-Weighted Evaluation Metrics for Imbalanced Data Classification", unpublished.