Convolutional Neural Networks for False Positive Reduction of Automatically Detected Cilia in Low Magnification TEM Images

Anindya Gupta[1(✉)], Amit Suveer[2], Joakim Lindblad[2,3], Anca Dragomir[4], Ida-Maria Sintorn[2,5], and Nataša Sladoje[2,3]

[1] T.J. Seebeck Department of Electronics,
Tallinn University of Technology, Tallin, Estonia
anindya.gupta@ttu.ee
[2] Department of IT, Centre for Image Analysis,
Uppsala University, Uppsala, Sweden
{amit.suveer,joakim.lindblad,
ida.sintorn,natasa.sladoje}@it.uu.se
[3] Mathematical Institute,
Serbian Academy of Sciences and Arts, Belgrade, Serbia
[4] Department of Surgical Pathology,
Uppsala University Hospital, Uppsala, Sweden
anca.dragomir@igp.uu.se
[5] Vironova AB, Stockholm, Sweden

Abstract. Automated detection of cilia in low magnification transmission electron microscopy images is a central task in the quest to relieve the pathologists in the manual, time consuming and subjective diagnostic procedure. However, automation of the process, specifically in low magnification, is challenging due to the similar characteristics of non-cilia candidates. In this paper, a convolutional neural network classifier is proposed to further reduce the false positives detected by a previously presented template matching method. Adding the proposed convolutional neural network increases the area under Precision-Recall curve from 0.42 to 0.71, and significantly reduces the number of false positive objects.

Keywords: Convolutional neural network · Primary Ciliary Dyskinesia · Template maching · Transmission electron microscopy

1 Introduction

Primary Ciliary Dyskinesia (PCD) is a rare genetic disorder resulting in dysfunctional cilia - the hairlike structures protruding from certain cells. Dysfunctionality of cilia can result in severe chronic respiratory infection, and infertility in both genders. To diagnose the disorder, pathologists examine the morphological appearance of cilia (∼220–250 nm) using transmission electron microscopy

P. Sharma and F.M. Bianchi (Eds.): SCIA 2017, Part I, LNCS 10269, pp. 407–418, 2017.
DOI: 10.1007/978-3-319-59126-1_34

(TEM). Qualitative analysis of cilia in the TEM images is still largely subjective and manual diagnosis is laborious, monotonous, and hugely time consuming (diagnosis takes ca. two hours per sample). An expert pathologist has to zoom in and out at locations of cilia which possibly exhibit structural information necessary for correct diagnosis. Navigation through the huge search space, together with change of magnification, is very demanding. Hence, there is an inevitable requisite for the automation of the cilia detection and diagnosis process. However, it is not feasible to acquire images which cover the whole sample at a magnification that allows structural analysis; such an acquisition would take tens of hours. Furthermore, objects of interest are rare, very small, and not spreading over more than a couple of percents of the total sample. Locating these regions of interest at low magnification, and acquiring high magnification images only at selected locations, would therefore be highly beneficial.

Automated detection of cilia structures (of a quality sufficient for diagnosis) at low magnification is a challenging task due to (1) their similar characteristics with the large number of non-cilia structures, and (2) variance in the size, shape and appearance of the individual cilia structures. The task becomes more complicated also due to noise and the non-homogeneous background at low magnification, see Fig. 1.

Lately, availability of large amounts of data and strong computational power have rapidly increased the popularity of machine learning approaches (deep learning). Convolutional neural networks (CNN) [10] have outperformed the state-of-the-art in many computer vision applications [8]. Similarly, the applicability of

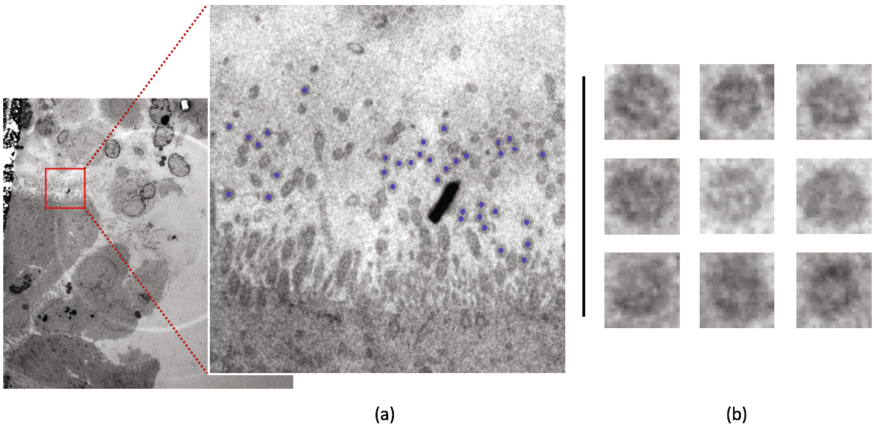

Fig. 1. (a) Low magnification TEM image of 4096×4096 pixels utilized for training purpose with the magnified view of 350×350 pixel bounding box (marked in red) with indicated ground truth marked by an expert pathologist. Here, cilia candidates marked with blue dots are of the suitable quality. (b) Some examples of patches extracted by previously reported method [15], the first and second rows contain true positives (TP) whereas patches in the third row are false positives (FP). Note the high similarity between the classes, this makes the problem a serious challenge. (Color figure online)

CNN is also investigated in the medical image analysis field [1,11]. In particular, their capability to learn discriminative features while trained in a supervised fashion makes them useful for automated detection of structures in, e.g., electron microscopy images. For instance, Ciresan *et al.* [5] reported a CNN model to segment the neuronal membranes in electron microscopy images; in [19], a CNN with autoencoder for automated detection of nuclei in high magnification (HM) microscopy images was employed.

Previously, a template matching (TM) method to detect cilia candidates in low magnification TEM images was proposed [15]. Considering that we aim at locating regions highly populated by good quality cilia, for further HM image acquisition and analysis, it is crucial that the identification of such regions is not misled by a large number of false positives (FP). In the current work, we aim at improving the performance by incorporating a dedicated CNN model in the cilia detection scheme with the special focus on reducing the number of FP. A performance benchmark for the proposed model is presented, and independent validation on an additional image is performed.

2 Image Data

Two low magnification (LM) TEM images from different patients, each with ca. 200 cilia structures, are used for training and independent validation purposes. Both images are acquired with a FEI Tecnai G2 F20 TEM and a bottom mounted FEI Eagle 4K × 4K HR CCD camera, resulting in 16-bit gray scale TIFF images of size 4096 × 4096 pixels.

For each LM image field, a set of mid magnification (MM) images are acquired, where the ground truth, i.e., true cilia candidates of promising quality for diagnosing at HM (not dealt with in this paper), are manually marked by an expert pathologist (author AD). Some examples of extracted patches of marked cilia candidates are shown in Fig. 1(b). The field of view (FOV) for a MM (2900×) image is 15.2 µm and for a LM (690×) image, it is 60.6 µm.

3 Method

The overall detection workflow consists of two stages: (1) Template matching as described in [15], and (2) further FP reduction using a 2-D CNN model, which is the core of this paper.

3.1 Initial Candidate Detection

Template matching based on normalized cross-correlation (NCC) and a customized synthetic template is used to detect the initial cilia candidates. The cross correlation image is thresholded at a suitable threshold, followed by area filtering and position filtering, meaning that only the best hit in a local region is kept as a candidate [15].

3.2 Data Partitioning and Augmentation

For each candidate position, we extracted patches of 23×23 pixels centered at a given position $p = (x, y)$. The patch size was chosen in order to contain a cilia object (∼19–20 pixels diameter), and some local background around the cilia instances (∼3 pixels) to include sufficient context information.

A training set of cilia, as well as non-cilia candidates, was extracted from the training image based on ground truth markings made by our expert pathologist (author AD), in MM images covering the same area of the sample. All true cilia (a total of 136) regardless of their match score, i.e., their NCC values, were chosen. A set of 272 non-cilia candidates was extracted from different NCC levels in order to represent non-cilia objects with high similarity to good cilia (136 randomly chosen non-cilia objects with NCC values ≥ 0.5) as well as non-cilia objects more different from true cilia (136 randomly chosen objects with NCC threshold values between 0.2 and 0.5).

While training a CNN model, an imbalanced dataset can mislead the optimization algorithm to converge to a local minimum, wherein the predictions can be skewed towards the candidates of the majority class, resulting in an overfitted model. To avoid overfitting, candidates from both classes (i.e. cilia and non-cilia) are augmented. Augmentation on test data has shown a considerable improvement in terms of robustness of the system, as it, if designed properly for the problem at hand [3].

Prior to the augmentation step, the candidates are randomly divided into training, validation and test sets. The training set consists of 82 cilia and 164 non-cilia candidates whereas the validation and test sets, each consists of 27 cilia and 54 non-cilia candidates. The candidates are augmented using affine transformations (rotation, scaling and shear) and bilinear interpolation. Horizontal flipping is applied to the cilia candidates to balance the sets. A fully automated script is created to perform the combination of seven random angular rotations (0–360°), six random scalings within ±10% range and five random shearings within 5% range in both x- and y- directions, resulting in 1050 augmented variations for each candidate. The augmentation scheme is applied separately for each subset to ensure independency of the training set from the validation and test sets.

3.3 2-D CNN Configuration

The architecture of the proposed CNN model is initially derived from the LeNet architecture [9]. The motivation behind this choice is its efficiency, as well as lower computational cost compared to the architectures such as Alexnet [8] and VGGnet [13]. These models have extended the functionality of LeNet into a much larger neural network with often better performance but at a cost of a massive increase in number of parameters and computational time. Training of such large networks is still difficult due to the lack of powerful ways to regularize the models and large feature sizes in many layers [16]. Hence, we decided to empirically modify the LeNet architecture to fit our application.

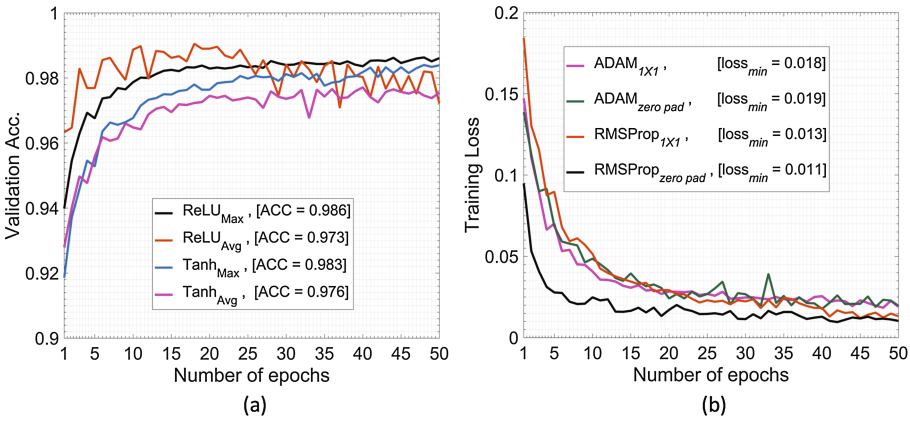

Fig. 2. Performance curves of different configuration: (a) validation accuracy for different activation functions and pooling layer combinations; (b) training loss for different optimizers with zero-padding and kernel of 1×1.

In our modified architecture, the default activation function i.e., hyperbolic tangent (tanh) [18] is replaced with Rectified linear units (ReLU) [12]. In comparison to the tanh, the constant gradient of ReLUs results in faster learning and also reduces the problem of vanishing gradient. We also implemented the maxpooling layer instead of average pooling as subsampling layer [8]. A comparative performance of both activation functions with different subsampling layers are shown in Fig. 2(a). The figure shows the accuracy for each configuration at different number of epochs. It is noticeable that the performance is better when ReLU was configured with maxpooling layer, resulting in higher accuracy after 50 epochs.

We also compared the usability of zero-padding and 1×1 convolution filters (as suggested in [16]) for two different optimizers, Adam [7] and RMSProp [17]. A kernel of size 1×1 in the first convolutional layer reduced the number of parameters (difference of 1 120 parameters compared to the zero-padding), thus keeping the computations reasonable. Comparatively, in either configuration, RMSProp with zero-padding resulted in a better training loss, as shown in Fig. 2(b). We thus, selected the configuration with minimum training loss. Moreover, several parameters (number of layers, kernel size, training algorithm, and number of neurons in the dense layer) were also experimentally determined.

In the proposed CNN classifier, the input patches are initially padded with a three pixels thick frame of zeros in order to keep the spatial sizes of the patches constant after the convolutional layers, as well as to keep the border information up to the last convolutional layer. Next, two consecutive convolutional layers and subsampling layers are used in the network. The first convolutional layer consists of 32 kernels of size $6 \times 6 \times 1$. The second convolutional layer consists of 48 kernels of size $5 \times 5 \times 32$. The subsampling layer is set as the maximum values in non-overlapping windows of size 2×2 (stride of 2). This reduces the size of

the output of each convolutional layer by half. The last layer is a fully connected layer with 20 neurons followed by a softmax layer for binary classification. ReLU are used in the convolutional and dense layers, where the activation y for a given input x is obtained as $y = max\ (0,\ x)$. The architecture of the proposed CNN model is shown in Fig. 3.

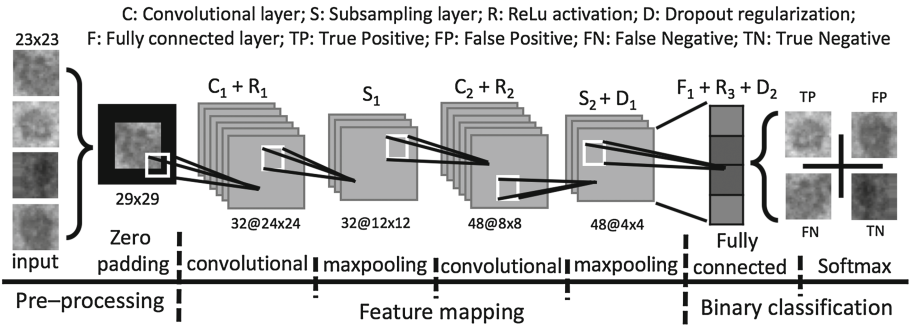

Fig. 3. An overview of the proposed CNN model.

3.4 Network training

The training of the classifier was performed in a 5-fold cross-validation scheme. For each fold, the candidates were randomly split into five blocks to ensure that each set was utilized as test set once. The distribution of candidates in each fold was kept as shown in Table 1.

Table 1. The number of cilia and non-cilia candidates in the different sets. Candidates marked in bold are finally utilized for building the model.

Set	Training	Validation	Test
Cilia	82	27	27
Aug (cilia)	172 364	56 754	56 754
Non-cilia	164	54	54
Aug (non-cilia)	172 364	56 754	56 754
Final set	**344 728**	**113 508**	**113 508**

On the given training dataset, RMSProp [17] is used to efficiently optimize the weights of the CNN. RMSProp is an adaptive optimization algorithm, which normalizes the gradients by utilizing the magnitude of recent gradients. The weights are initialized using normalized initialization as proposed in [6] and updated in a mini-batch scheme of 128 candidates. The biases were initialized with zero and learning rate was set to 0.001. A dropout of 0.5 is implemented

as regularization, on the output of the last convolutional layer and the dense layer to avoid overfitting [14]. Softmax loss (cross-entropy error loss) is utilized to measure the error loss. The CNN model is implemented using theano backend in Keras [4]. The average training time is approximately 48 s/epoch on a GPU GeForce GTX 680.

4 Experimental Results and Discussion

The performance of the proposed CNN model was evaluated in terms of *Precision*, *Recall*, *Area under the Precision-Recall curve (AUC)*, and *F-score*, defined as:

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN},$$

$$F\text{–}score = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \qquad AUC = \int_0^1 P(r)dr.$$

The AUC is the average of precision $P(r)$ over the interval $(0 \leq r \leq 1)$, and $P(r)$ is a function of recall r. Additionally, for different NCC threshold levels, the Free-response Receiver Operating Characteristic (FROC) curve [2] was utilized to measure the sensitivities at a specific number of false positives per image. The FROC curve is an extension of the receiver operating characteristic (ROC) curve, which can be effective when multiple candidates are present in a single image. It plots the Recall (Sensitivity) against the average number of false positives per images. FROC is more sensitive at detecting small differences between performances and has higher statistical discriminative power [2].

4.1 Quantitative results

Figures 4(a) and (b) show the precision-recall curves corresponding to cilia detection for the CNN classifier applied after thresholding the template matching at different NCC levels (0.2, 0.3, 0.4, and 0.5), as well as the detection when using only template matching (which includes NCC thresholding at 0.546), as proposed in [15], for the training and test image, respectively. In the figures, the AUC is also stated. The results show that adding a CNN classifier significantly improves the AUC to 0.82 and 0.71 compared to the AUC of 0.48 and 0.42, for both the training and test image, respectively, at an NCC threshold level of 0.5.

The FROC curve for the proposed CNN applied to the training and test images when the template matching result was thresholded at different NCC levels (0.2, 0.3, 0.4, and 0.5) is shown in Fig. 5(a)–(b). This corresponds to the sensitivity of the classifier against total number of FP per image.

A classification confusion matrix is also shown in Table 2. The matrix shows the performance of the classifier for both the training and test image, in terms of TP (true positive), FP (false positive), FN (false negative), and TN (true negative), at equal error rate. At an NCC threshold level 0.5, the template matching method detected 212 (73 cilia and 139 non-cilia) candidates as potential cilia candidates. Amongst these, in the Table 2(A), the proposed CNN classifier correctly

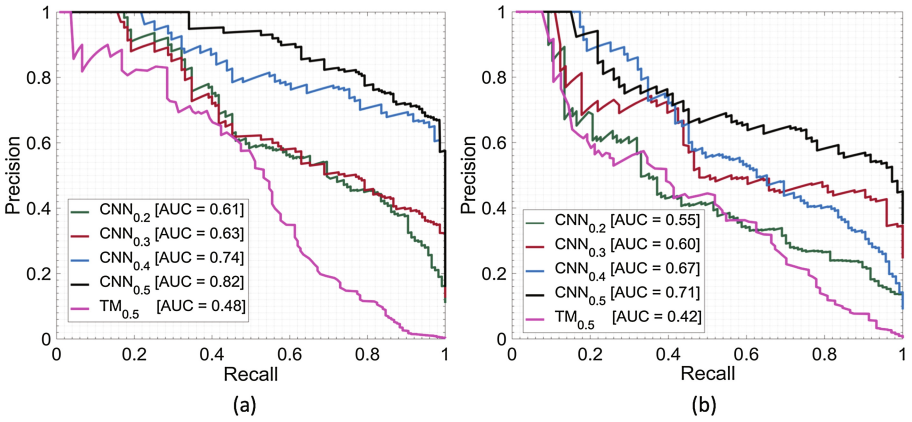

Fig. 4. Precision-recall curves of the CNN classifier at different NCC threshold levels shown together with the AUC for the template matching approach(TM) [15] for (a) training, (b) test images

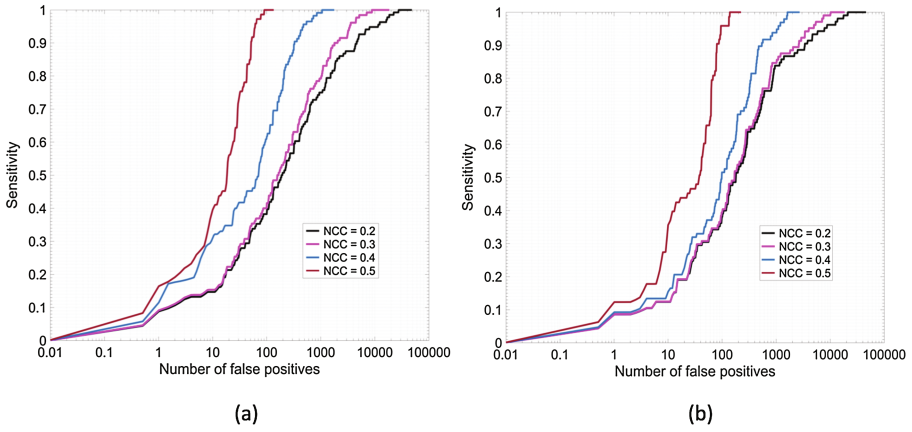

Fig. 5. FROC curves of the CNN classifier for (a) training image (b) test image at different NCC threshold levels. The number of FP are shown on a logarithmic scale.

classified 47 (TP) out of 73 (TP+FN) cilia candidates whereas from the set of 139 (FP+TN) non-cilia candidates, 26 non-cilia candidates (FP) were wrongly classified as cilia candidates by our proposed CNN classifier. We observe, in the training image, at equal error rate (Table 2(A)), the classifier also performed well when tested with the candidates extracted at an NCC threshold level of 0.4, but it eventually underperformed for the test image. The achieved results led us to finally conclude that the proposed CNN model yields a stable performance if it is incorporated with the candidates extracted at an NCC threshold level of 0.5. This observation is supported by the F-Score curves, shown in Fig. 6. Comparatively for

Table 2. Classification matrix of the CNN classifier at different NCC threshold levels for: (A) training image and (B) test image; at equal error rate.

A: Training image (Equal error rate)									
		0.2		**0.3**		**0.4**		**0.5**	
TP	FP	51	85	50	80	64	51	47	26
FN	TN	85	48 004	80	18 035	51	1 113	26	113
B: Test image (Equal error rate)									
		0.2		**0.3**		**0.4**		**0.5**	
TP	FP	38	67	37	66	37	60	37	36
FN	TN	67	45 926	66	18 348	60	2 658	36	188

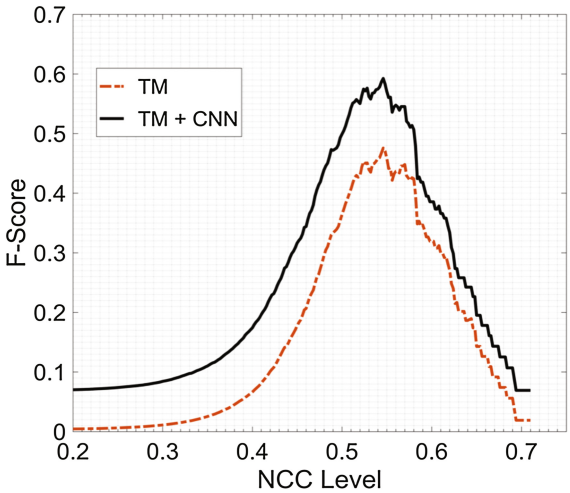

Fig. 6. F-score curves, for the test image, showing the improvement in overall performance by adding a CNN classifier with template matching approach(TM) [15] at different NCC threshold levels

the test image, at an NCC level of 0.546 (as suggested in [15]), the proposed CNN model increases the overall F-Score from 0.47 to 0.59.

4.2 False positive reduction results

Detection results of the proposed CNN model on a ROI of 650×650 for the test LM TEM image, at an NCC level of 0.5, are shown in Fig. 7(c)–(d). Figure 7(c) shows the detection results of the initial candidate detection step (template matching method, [15]) whereas Fig. 7(d) shows the improved results achieved by incorporating the proposed CNN model as an FP reduction step. In these images, the blue circles, red crossed circles, and green squares represent the candidates that have been correctly detected (TP), the candidates that have been

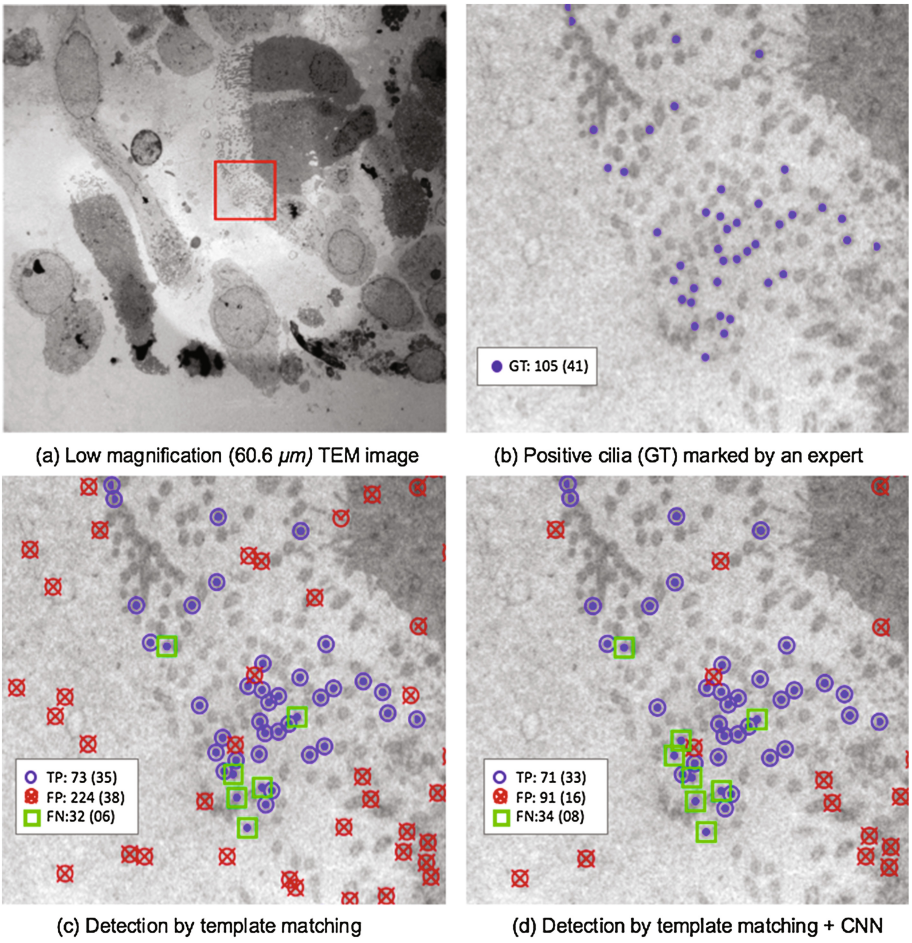

Fig. 7. Illustration of cilia detection results. (a) The 4096×4096 test image, (b) a 650×650 example subregion of the test image, (c) same subregion after initial template matching method, and (d) after proposed CNN classifier. The numbers are given for the whole image and for the ROI is in parenthesis. Here, blue circles, red crossed circles, and green squares represent the TP, FP, and FN, respectively. (Color figure online)

erroneously detected as cilia (FP), and the cilia that were missed with respect to the manually ascertained ground truth delineations and initial detection step (FN), respectively. These results show the potential of our CNN model for cilia detection in low magnification TEM images.

Examples of classified candidate image patches in the test image are shown in Fig. 8. The images marked in the first row are the TP and FP candidates from both methods (i.e., TM and CNN). In the second row, TP candidates detected by TM but erroneously classified as FN by CNN; and FP candidates detected by TM, which are successively classified as TN by proposed classifier.

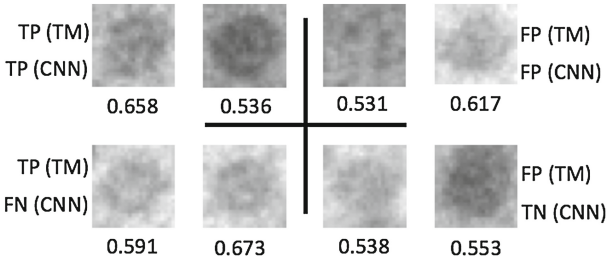

Fig. 8. Examples of candidates (with their corresponding NCC values) detected or missed by the proposed CNN model in the test image at an NCC level of 0.5. The first row shows TP's and FP's of both methods. The second row shows TP and FP candidates which are missed and successively classified by the CNN method, respectively.

5 Conclusion

In this paper, a CNN classifier is presented as a false positive reduction step for automated detection of cilia candidates in low magnification TEM images. The results suggest that adding a CNN classifier as a FP reduction step certainly improves the performance and results in an increased F-Score from 0.47 to 0.59. It was also investigated whether utilizing a CNN classifier as an additional refinement step would allow for using a lower NCC threshold in order to not discard true cilia objects in the template matching step. This was however, not found to be practically suitable as lowering the NCC threshold increases the number of candidates to analyze tremendously while only rather few additional true candidates are detected. It will be interesting in the future to develop and investigate a CNN model for the whole automated cilia detection problem, without relying on a first template matching step. This is currently not possible as it requires more training and test data.

Acknowledgments. The work is supported by Skype IT Academy Stipend Program, EU Institutional grant IUT19-11 of Estonian Research Council and the Swedish Innovation Agency's MedTech4Health program grant no. 2016-02329. J. Lindblad and N. Sladoje are supported by the Ministry of Education, Science, and Technological Development of the Republic of Serbia through projects ON174008 and III44006.

References

1. Brosch, T., Yoo, Y., Li, D.K.B., Traboulsee, A., Tam, R.: Modeling the variability in brain morphology and lesion distribution in multiple sclerosis by deep learning. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8674, pp. 462–469. Springer, Cham (2014). doi:10.1007/978-3-319-10470-6_58
2. Chakraborty, D.: A status report on free-response analysis. Radiat. Prot. dosimetry **139**, 20–25 (2010)

3. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: British Machine Vision Conference (BMVC) (2014)
4. Chollet, F.: Keras (2015). https://github.com/fchollet/keras
5. Ciresan, D., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in Neural Information Processing Systems, pp. 2843–2851 (2012)
6. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Aistats, vol. 9, pp. 249–256 (2010)
7. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR) (2015)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
9. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
10. LeCun, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
11. Li, R., Zhang, W., Suk, H.-I., Wang, L., Li, J., Shen, D., Ji, S.: Deep learning based imaging data completion for improved brain disease diagnosis. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8675, pp. 305–312. Springer, Cham (2014). doi:10.1007/978-3-319-10443-0_39
12. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: 27th International Conference on Machine Learning, pp. 807–814 (2010)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR) (2015)
14. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
15. Suveer, A., Sladoje, N., Lindblad, J., Dragomir, A., Sintorn, I.M.: Automated detection of cilia in low magnification transmission electron microscopy images using template matching. In: 13th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 386–390. IEEE (2016)
16. Szegedy, C., Liu, W., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
17. Tieleman, T., Hinton, G.: Lecture 6.5-RmsProp: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for ML (2012)
18. Vogl, T.P., Rigler, A., Zink, W., Alkon, D.: Accelerating the convergence of the back-propagation method. Biol. Cybern. **59**(4–5), 257–263 (1988)
19. Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., Madabhushi, A.: Stacked saparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. IEEE Trans. Med. Imag. **35**(1), 119–130 (2016)