BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 66, No. 6, 2018
DOI: 10.24425/bpas.2018.125933

Application of convolutional neural networks with anatomical knowledge for brain MRI analysis in MS patients

B. STASIAK[1]*, P. TARASIUK[1], I. MICHALSKA[2], and A. TOMCZYK[1]

[1]Institute of Information Technology, Lodz University of Technology, Wolczanska 215, 90-924 Lodz, Poland
[2]Department of Radiology, Barlicki University Hospital, Kopcinskiego 22, 91-153 Lodz, Poland

Abstract. In this paper we consider the problem of automatic localization of multiple sclerosis (MS) lesions within brain tissue. We use a machine learning approach based on a convolutional neural network (CNN) which is trained to recognize the lesions in magnetic resonance images (MRI scans) of the patient's brain. The training images are relatively small fragments clipped from the MRI scans so – in order to provide additional hints on location of a given clip within the brain structures – we include anatomical information in the training/testing process. Our research has shown that indicating the location of the ventricles and other structures, as well as performing brain tissue classification may enhance the results of the automatic localization of the MS-related demyelinating plaques in the MRI scans.

Key words: multiple sclerosis, convolutional neural networks, skull stripping, ventricular system.

1. Introduction

Convolutional neural networks (CNNs) are biologically-inspired machine learning tools, which have been gaining much attention recently. Due to their unique architectural properties and processing principles, they are especially suited for automatic image analysis, classification and recognition. Envisioned and designed in their basic principles as early as in the 1980s [1], they reached their true potential with the advent of efficient GPU implementations, which allowed them to solve real-life pattern recognition problems with impressive effectiveness.

Apart from the general task of object classification [2], scientific or medical data analysis may also be done with the CNNs, provided that enough training data is available [3–5]. In these applications we often need not only the classification but also the precise localization/segmentation of anatomical structures or tissue lesions. In general, the approaches found in the literature may be roughly divided in two basic groups: patch-based and whole image-based. In the first case, we try to classify individual regions of the image (in particular, the regions representing the neighborhood of a given pixel). This may be considered a modified sliding window technique with CNN as a classifier. For training we also use patches cut from the training images, manually segmented by an expert, instead of the whole images. Such a method was used, for example, in the segmentation of anatomical regions in MRI images [3]. In the second group, the proposed solutions are based on a "fully convolutional" approach [6]. In this case, the whole image is given as an input and an image of the same size is obtained at the output of the neural network. The output image presents the detected regions of interests of the pre-defined type (or types). To obtain this, a two-stage architecture of the neural network is usually used. At the first stage, standard convolutional and pooling layers are used to reduce the size of the resulting feature maps, and then – at the second stage – some upscaling (deconvolutional) layers are added to enlarge and combine these maps to obtain the image of a proper size. Such a fully convolutional network is trained using whole images without the need of cutting them into patches. This kind of approach was successfully used e.g. in the analysis of transmitted light microscopy images [4] and MRI prostate examinations [5]. The latter approach is particularly interesting since it operates directly on 3D data (3D MRI sequences) processed by a CNN by means of 3D convolution operation.

The solution proposed in [7] combines the features of both aforementioned groups of approaches. On the one hand, the CNN is trained to act as a non-linear filter capable of detecting regions of interest in the images of arbitrary size (so that the output is the image of the same size as the input). In this case, however, no pooling is used and, consequently, no upscaling is required. On the other hand, such a network may be trained with smaller patches without the necessity of processing the whole images during the learning phase. This provides an additional advantage, as more representative patches may be selected for training, which is especially important when the regions to be detected are sparse within the source images.

In this work, we adopted the approach presented in [7] to automatically detect the demyelinating plaques in brain MRI (Magnetic Resonance Imaging) scans in multiple sclerosis (MS) patients. We have applied the same neural net model and the same set of images but with additional information based on brain tissue classification results and known location of some anatomical structures, manually annotated by radiologists. In this way we were able to assess the influence of domain knowledge involved in the recognition process and compare

*e-mail: bartlomiej.stasiak@p.lodz.pl

Manuscript submitted 2017-09-26, revised 2018-04-09 and 2018-05-15, initially accepted for publication 2018-05-16, published in December 2018.

the results with the previous, pure image-based, example-driven approach.

The rest of the paper is structured as follows. In the next section the detailed goals of the present study are formulated, along with the medical background and the testing material characteristics. Section 3 presents related works while Section 4 recapitulates the fundamental facts about the convolutional neural networks and their application in image analysis, classification and recognition. In the main part of the paper (Section 5) we present the details of our CNN architecture, the experiment design and methodology, as well as data preparation/postprocessing algorithms utilizing the additional information on tissue type and anatomical structures. The results of the experimental validation are presented in Section 6 and summarized in Section 7.

2. Material and objectives

Multiple sclerosis (MS) is the most common chronic autoimmune disease of the central nervous system (CNS), leading to neurological disability manifested by a broad range of signs and symptoms [8, 9]. The underlying mechanism of the MS is the destruction of the myelin sheaths of the CNS nerves by patient's own immune system. The areas of the white matter where the layer of myelin has been damaged are called demyelinating plaques and the whole process is known as demyelination. The diagnosis of the disorder is made by the combination of clinical findings, the examination of cerebrospinal fluid (CSF) and MRI of the central nervous system. In patients with clinical symptoms suggesting MS, the brain MR imaging can show multifocal white matter lesions which are plaques of demyelination. It should be noted that the process of demyelination is not specific to MS only – it can be a part of many other disorders. The diagnosis of MS is more likely if the plaques are distributed in some typical areas in the brain such as: around the lateral ventricles (periventricular), especially while they are orientated perpendicularly to the long axis of the ventricles, in the corpus callosum, along the boundary between the white matter and cortex, in the cerebral and cerebellar peduncles, in pons and medulla oblongata. The most useful MRI scans for identifying white matter lesions are T2-weighted images (T2WI) – particularly FLAIR sequences (fluid-attenuated inversion recovery). In these images the demyelinating areas are hyperintense and hence they are easily detected within the normal white matter (Fig. 1). In T2WI sequences both cerebrospinal fluid and white matter lesions are hyperintense, so the contrast between them is rather poor. In FLAIR sequences, the signal of the CSF is attenuated, which improves the detection of the white matter lesions, especially in the periventricular distribution [10].

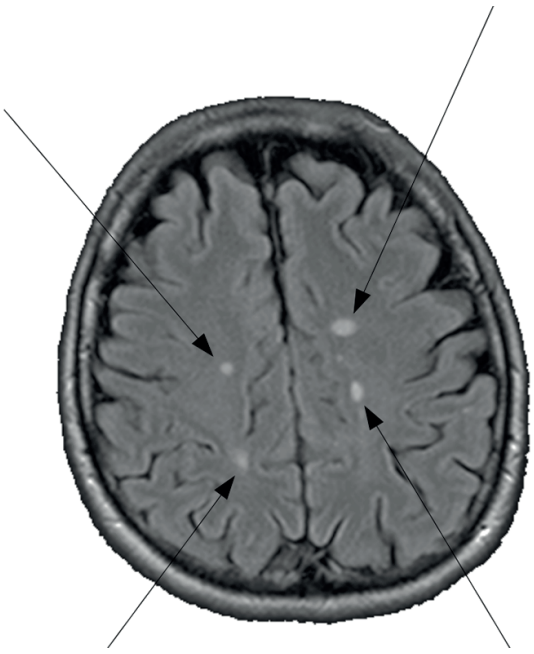


Fig. 1. Example of a MRI scan with four distinct MS lesions indicated

This paper concerns the detection of demyelinating lesions on MR scans of the brain (FLAIR sequences in the axial plane). The magnetic resonance images, obtained with a 1.5 Tesla scanner, represent slices of thickness between 3 mm and 5 mm. The patient population consisted of hundred people (fifty men and fifty women) of different age groups (between 19 and 66 years old). The study has taken into consideration only patients with confirmed diagnosis of MS. The severity of the disease varied from newly diagnosed to longstanding disorders.

As mentioned above, the plaques in multiple sclerosis may be located in some characteristic areas within the cerebral white matter. Therefore, in the present research we have decided to include the information about location of some of these structures in the training dataset. These structures, including lateral ventricles, corpus callosum, cerebral peduncles, cerebellar peduncles, pons and medulla oblongata, were manually annotated by specialists. However, it should be noted that it is also possible to do it automatically with the use of specialized tools for brain MRI data analysis, after proper registration of the scans and setting appropriately the 3D coordinate system. In particular, other properly trained convolutional neural networks could be of use here, leading to a hierarchical system imitating – to a certain degree – the process of human image analysis.

The annotations, used in our experiment as an extension to the input images, may be considered a source of additional domain knowledge which is expected to produce a better and more effective model. However, it is not the only source of external information that we have examined. In a separate, second experiment we applied a tissue classification procedure, based on deformable surface models [11] to further refine the obtained results. The main goal of tissue classification was to segment the regions containing the nervous tissue, excluding the skull bones, sinuses, eyeballs and other structures irrelevant for MS diagnostics. The main motivation here was the fact that in the FLAIR MRI scans used in our experiments, the MS-specific demyelinating plaques appear brighter than the surrounding tissue, which is unfortunately also characteristic for some other types of tissue, e.g. for orbital fat or for most bone structures. Elimination of these regions from further analysis provided the ability to significantly reduce the number of false alarms induced by "bright" structures located outside the brain in MRI images, as demonstrated in Sect. 6.

3. Related work

The problem of automatic localization of multiple sclerosis lesions in MRI scans have been studied for decades, with a variety of tools and methods of image segmentation and analysis. Processing brain MRI data involves both low-level tools, such as gradient operators or local thresholding as well as high-level, anatomically motivated techniques, including e.g. brain surface modeling with B-splines [12]. Many approaches are based on some form of initial segmentation of the MRI images and classification of the tissue type, typically including white matter (WM), gray matter (GM), cerebro-spinal fluid (CSF) and components of the ventricular system. The detection of MS lesions in [13] is based on FLAIR image thresholding, with a preliminary processing step involving brain tissues segmentation with a variant of expectation-maximization (EM) algorithm. In [14] the tissue types are not classified, but a brain extraction tool (BET) is used to discard non-relevant areas and then a dictionary is constructed to enable sparse coding of individual parts (patches) of the MRI scans. Several other approaches involving machine learning have been proposed, such as [15], where a combination of genetic algorithm (GA) and Support Vector Machine classifier (SVM) is applied to analyze feature vectors based on texture descriptors: co-occurrence matrix (GLCM) and gray-level run length (GLRLM) matrix. The features are computed on the basis of several MRI modalities combined together by volumetric wavelet fusion. In this approach 3D brain representation is used with several additional operations, including brain extraction, registration, segmentation [16], and filtering [17]. Among other machine learning techniques, neural networks and convolutional neural networks in particular are an interesting tool for MS lesion detection and segmentation [18], which we will cover further in more detail. A comparison of results of the above approaches with our outcomes as well as discussion of similarities and differences between methods will be presented in Section 6.3.

4. Theoretical background

Neural networks have been well known for decades as effective, biologically-inspired tools for solving various machine-learning problems. Having gone a long way from the initial concepts and simplified models [19] they are experiencing now their renaissance due to the computational potential of modern graphics processing units (GPUs) enabling efficient training of "deep" neural architectures with many hidden layers, modeling complex dependencies inherent to real-world problems [20, 21]. The biological inspirations are especially important [22] in architectural and functional principles of the *convolutional neural networks* (CNNs), where the analogies to some elements of a human visual system led to a significant reduction of the connections between layers and extensive weights sharing [1, 23]. CNNs are capable of performing visual information analysis (e.g. due to the fact that they are translation-independent by design) in a way resembling the hierarchical processing of images performed by a human brain. Outputs of the hidden layers (*convolutional layers*) are called *feature maps* [20, 24], since they actually describe locations of certain features of the image on different levels of abstraction. The CNN input is usually just a raw digital image, with optional very basic preprocessing (scaling, normalization, etc.) [21].

The typical application area of the convolutional networks is image classification – e.g. the state-of-the-art solutions to the ImageNet Large Scale Visual Recogintion Challenge, ILSVRC [2] are based on CNNs [21, 25, 26]. In this case the CNN integrates two elementary steps of a usual pattern recognition system: the feature-extraction and classification. However, it may also be used just as general-purpose feature extractor [27] or as a tool for locating individual objects within the image [28, 29]. Some deep and complex CNNs trained for ILSVRC were also successfully applied as a part of larger solution to other image recognition problems [30].

In a typical approach the CNN is expected to perform some dimensionality reduction of the input data, so as to reduce the size of feature maps in the consecutive hidden layers. This reduced data representation is then fed to a general-purpose classifier, such as a multilayer perceptron (MLP). Using MLP is especially convenient, because – being itself a neural network – it may be implemented just as a set of additional, densely connected neural layers appended to the CNN and trained jointly (CNN + MLP network as a whole) with a gradient optimization technique [20].

In the object localization task the expected output of the network is a feature map itself. In this case we do not usually use the classifier module (such as the MLP), limiting the neural architecture to convolutional layers only. The output of the network may be of the same size as the input, which provides the ability to relate the positions of the detected objects to the original (input) image directly. The network may be trained so that each output value represents the likelihood that the corresponding pixel of the input image belongs to an object of a given type. This is an approach used in [7], which we also adopt here, as described in Sect. 5.

4.1. CNN fundamentals. The construction of a convolutional neural network is based on a sequence of consecutive convolutional layers, where each layer computes its output on the basis of 2D convolution of its input. Both input and output data may be interpreted here as images, usually multi-channel ones (Fig. 2). Let p and q denote the number of channels of the input and output images of a convolutional layer, respectively. The input image may then be defined as a tuple of matrices $A_1 \ldots A_p$ of a fixed $n_a \times m_a$ size, where each matrix represents a single channel. Typically, for RGB images we have $p = 3$ in the first layer (note, however, that for the other layers p may take arbitrary values). The neural weights of the layer define the so-called *filters* (playing the role of the *convolution kernels*), which are collected in *filter groups*. Each filter group is defined as a tuple of p matrices of $n_f \times m_f$ size, where each matrix represents a single filter $F_{i,j}$ for $i = 1 \ldots q, j = 1 \ldots p$. The output of the convolutional layer is a tuple of *feature maps* $M_1 \ldots M_q$ defined, for each $i = 1 \ldots q$, as:

$$M_i = Z_i + \left(\sum_{j=1}^{p} \right) A_j * F_{i,j}. \tag{1}$$

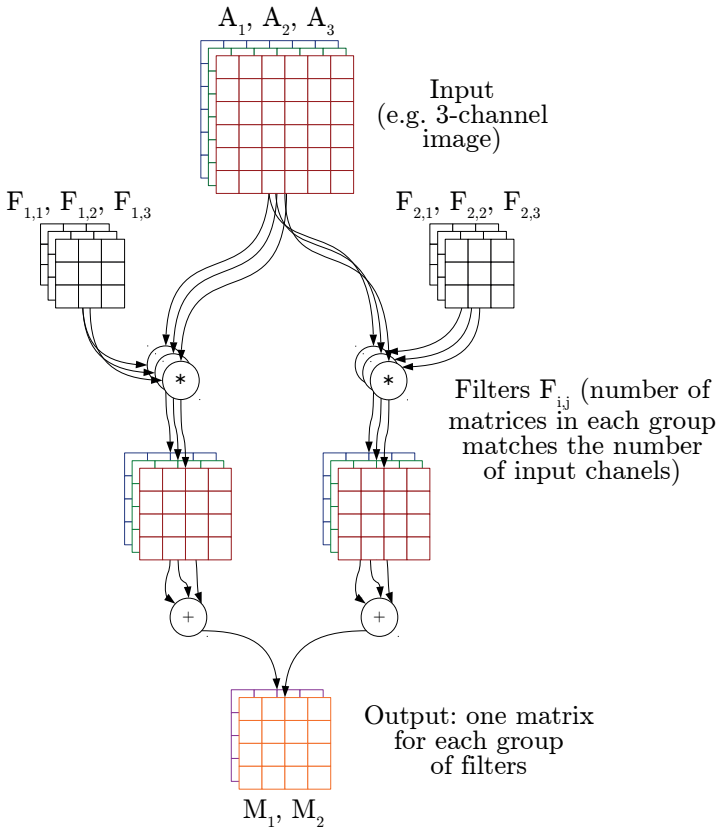


Fig. 2. Internal structure and operation of a convolutional layer [7]. In this example, a 3-channel input (A_1, A_2, A_3) is processed by 2 groups of filters $F_{i,j}$ (3 filters in each group). Convolution results produced by each filter group are summed up. Each sum is a separate output matrix (output channel) – in this case: M_1, M_2

In the formula above Z_i is a bias matrix of the same size as M_i. Matrix convolution $A_j * F_{i,j}$ is a matrix of elements $(A_j * F_{i,j})_{r,c}$ for $r = 1 \ldots (n_a) - (n_f) + 1$, $c = 1 \ldots (m_a) - (m_f) + 1$ such that:

$$(A_j * F_{i,j})_{r,c} = \left(\sum_{d_n=0}^{n_f-1}\right)\left(\sum_{d_m=0}^{m_f-1}\right) = \\ = s(F_{i,j})_{(n_f-d_n),(m_f-d_m)} \cdot (A_j)_{(r+d_n),(c+d_m)} \cdot \tag{2}$$

It should be noted that the size of the resulting M_i matrices is $n_a - n_f + 1 \times m_a - m_f + 1$, so it basically differs from the input image size for any size of the filters other than trivial 1×1. If we want to keep the size constant throughout the processing (which is actually the case in our present study), we can use zero-padding of the A_j to increase the size of the input data to $(n_a + n_f - 1) \times (m_a + m_f - 1)$ *before* the convolution. On the other hand, a typical approach employed to radically modify the output size (usually to reduce it, as required in most classification problems) is to use some kind of pooling after the convolutional layers. Max-pooling or average pooling (implemented in separate *pooling layers* inserted between the convolutional layers), reducing the output matrix size by a certain factor, is often applied [24]. However, as mentioned above, in the present study we do not use the pooling layers at all.

Convolution of a matrix with a fixed filter $F_{i,j}$ is linear (and so is the whole layer), hence – in order to effectively process the output data with the next convolutional layer – it is reasonable to use some non-linearity between the consequent layers. The obvious solution is to apply a non-linear activation function element-wise. While a sigmoid-like function is known to work, the modern approach is to use ReLU (rectified linear unit) [21] or PReLU (parametrized extension of ReLU) [31].

Each element of the output of the convolutional layer is a result of processing some $n_f \times m_f$ rectangles picked from each A_j. For the first feature map, $n_f \times m_f$ is a size of *visual field* [22]. For further layers, the size of visual fields could be easily calculated by tracking down the range of CNN input pixels affecting each output element. Should the network consist of convolutional layers and element-wise operations only, the visual field size would be $n_z \times m_z$ where $n_z = (n_{f_1} + \ldots + n_{f_t}) - t + 1$ and $m_z = (m_{f_1} + \ldots + m_{f_t}) - t + 1$. In these formulas t denotes the number of convolutional layers and $n_{f_w} \times m_{f_w}$ is w-th layer filter size for $w = 1 \ldots t$ [7].

5. Method

5.1. CNN Architecture. We have applied a 6-layer convolutional neural network that proved successful in [7], with square convolution kernels and stride equal to 1 in all layers. As mentioned above, the specificity of this network lies in the lack of pooling layers and we also do not apply fully-connected layers, which are typically used in most classification problems. The structure of our network is as follows:

- Layer 1: 20 filter groups $(n_f \times m_f = 5 \times 5$, padding: $2 \times 2)$
- Layer 2: 20 filter groups $(n_f \times m_f = 7 \times 7$, padding: $3 \times 3)$
- Layer 3: 40 filter groups $(n_f \times m_f = 9 \times 9$, padding: $4 \times 4)$
- Layer 4: 60 filter groups $(n_f \times m_f = 7 \times 7$, padding: $3 \times 3)$
- Layer 5: 20 filter groups $(n_f \times m_f = 5 \times 5$, padding: $2 \times 2)$
- Layer 6: 1 filter group $(n_f \times m_f = 5 \times 5$, padding: $2 \times 2)$.

After every layer, except for the last one, a non-linear activation function (parametric rectified linear unit, PReLU) is used. After the last layer we applied the unipolar sigmoid activation, as the goal of the training of the network was to generate binary (0–1) output. In particular, we expected the output value of 1 for every pixel within a MS lesion and the output value of 0 for every pixel within a normal tissue region. This was a natural consequence of our supervised training scheme, in which every input brain scan was accompanied by the target image (expected at the output of the network) displaying the demyelinating plaques annotated by a specialist as white regions against the black background (Fig. 3). It should be noted that the padding size in each convolutional layer was set so that the output image dimensions were identical to those of the input image and that our CNN guarantees the strict correspondence between the location of the regions of interest (demyelination plaques) in the input images and location of the corresponding annotations in the target output images. The lack of densely connected MLP layers in our model means also that the proposed CNN works

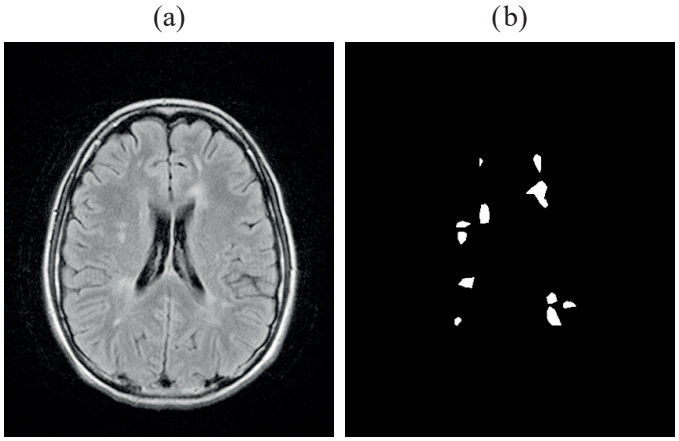


Fig. 3. An example from the training set: a) the input image (MRI scan); b) the target output image (the specialist's annotations)

as a kind of image filter and it can basically process an image of any size, provided that it is reasonably bigger that the visual field size (33×33 in our case – cf. Sect. 4.1). We actually make use of this feature in the training phase, where we cut input/target images into small tiles for effective training, as described in Sect. 5.4. On the other hand, the already trained network is used in the testing phase to process whole full-size scans without the need of any structural modifications.

5.2. Experiments. As indicated at the end of Sect. 2, in this study we use additional sources of anatomical information, which must be properly handled in the experiment design and in the neural models applied. We conducted 2 separate experiments to test the influence of the known location of large-scale anatomical structures (experiment 1) and the influence of known tissue types (experiment 2) on the detection of demyelinating plaques.

In the first experiment the anatomical annotations available were divided into two general types: those related to the ventricular system (the lateral ventricles) and those related to some characteristic areas of the brain tissue (corpus callosum, cerebral and cerebellar peduncles, pons and medulla oblongata). We decided to encode both types of annotations on two separate image channels, which – combined with the original MRI scan – formed a 3-channel RGB input image, as demonstrated in Fig. 4. From the structural point of view, this required only increasing, from 1 to 3, the number of filters in each of the 20 filter groups in the first convolutional layer (cf. Fig. 2).

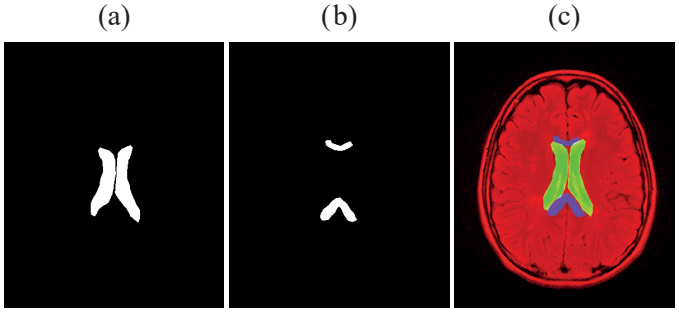


Fig. 4. Anatomical data for the image in Fig. 3: a) the lateral ventricles; b) corpus callosum (genu and splenium); c) the resulting 3-channel input image

In the second experiment, we used the standard one-channel input (MRI scan only, Fig. 3a), but the output was further filtered with a special brain-tissue mask, computed separately for each particular input image. This approach was therefore based on an additional post-processing phase, while the general structure of the CNN remained the same (except of the first layer, processing one- instead of 3-channel input images).

5.3. Evaluation of the results. Having discussed the input data of the CNN, we should now consider its output and the ways we could interpret it in relation to our processing goal. The last layer contains only one group of filters, which means that the network output is a single-channel image I_{out}. We expect that – in the course of training – this image will get as close to the target image I_{targ} (the radiologist's annotations, Fig. 3b) as possible. This is explicitly expressed in the learning objective defined as minimization of mean square error (MSE), or difference, between I_{out} and I_{targ}. However, it is unrealistic to expect the ideal, binary output image and MSE of zero (such a case would in fact suggest heavy over-fitting and poor generalization properties of the obtained model). Instead, we get a real-valued output image with individual elements *close to* one (for MS lesions) and *close to* zero (for normal tissue), which needs thresholding with a threshold $T \in (0, 1)$ in order to convert it into binary image I_T, directly comparable with the ground-truth image I_{targ}.

The comparison itself is based on counting the number of *true positive* (*TP*) pixels, i.e. such pixels that:

$$TP : \big(I_T(x,y) = 1\big) \wedge \big(I_{targ}(x,y) = 1\big),$$

where x, y are the pixel coordinates.
We also count the number of *false positive* (*FP*) and *false negative* (*FN*) pixels:

$$FP : \big(I_T(x,y) = 1\big) \wedge \big(I_{targ}(x,y) = 0\big),$$

$$FN : \big(I_T(x,y) = 0\big) \wedge \big(I_{targ}(x,y) = 1\big),$$

in order to compute the standard measures of binary classification quality – precision and recall:

$$\text{Precision} = \frac{TP}{TP + FP};$$

$$\text{Recall} = \frac{TP}{TP + FN}.$$

The precision is hence defined as the proportion of the number of *TP* pixels (correctly reported within the lesion areas) to all of the *actually detected* pixels, while the recall is the proportion of *TP* to all the pixels that *should be reported*.

It should be noted that a low value of threshold T, used to compute I_T, maximizes the recall and a high threshold maxi-

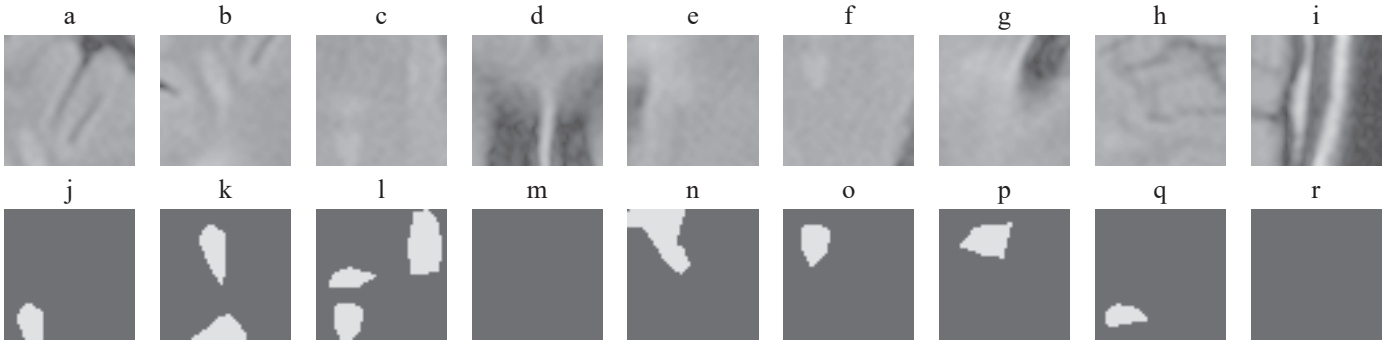


Fig. 5. Examples of training tiles cut from the image in Fig. 3. Top: Input tiles cut from Fig. 3a; Bottom: Corresponding target tiles cut from Fig. 3b. Note, that tiles (d)/(m) and (i)/(r) do not contain any lesions

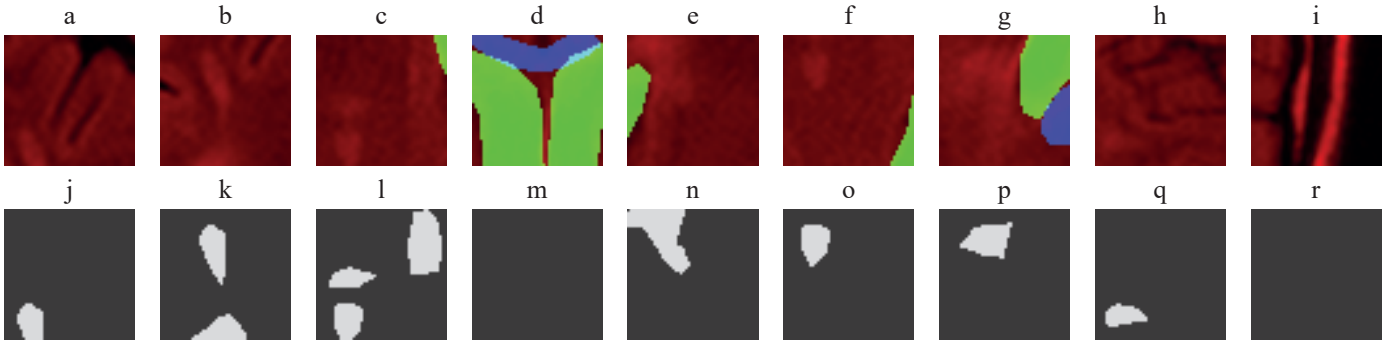


Fig. 6. Examples of training tiles cut from the image in Fig. 4 (3-channel version of Fig. 5). Top: Input tiles cut from Fig. 4c; Bottom: Corresponding target tiles (the same as in Fig. 5)

mizes the precision. An extremely low threshold would render all the pixels positive, yielding 100% recall and close-to-zero precision, while an extremely high threshold would do the opposite. Therefore, a standard approach employed to obtain representative results, applied also in the present study, is to compute the harmonic mean of precision and recall, known as *F-measure*:

$$F = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

The value of F-measure is used in the evaluation of the obtained results to find the appropriate threshold value T. We search through all possible threshold values, recording the resulting F-measure values for the training images. The threshold maximizing the F-measure is used to compute the final results on a separate set of testing images.

5.4. Dataset Preparation. For all the experiments we have used the same dataset as in [7], extended in the current study by additional anatomical information. The 96 patients available[1] were split at random into the training set (77 patients) and the testing set (19 patients). Each patient was represented by a set of MR scans of the size 448×512 pixels, out of which only the scans containing plaques of demyelination were considered. As a result, the testing set contained 242 scans and the training set was based on 982 scans. These 982 scans were not, however, used directly but they were cut into tiles of 50×50 pixels, as mentioned in Sect. 5.2, and only some of these tiles were selected for inclusion into the final training set. Out of the total number of 7856 selected training tiles, approximately 2/3 contained MS lesions, and the remaining ones were included as negative examples, to make the trained model more robust.

In Fig. 5 some tiles cut from the training image from Fig. 3a are presented. These are only examples, presenting exactly a half of all the 18 tiles cut from this particular input image, with some duplicate lesions and two "negative example" tiles. It is worth noting that the original full scan (Fig. 3a), while not used directly for training, is applied for determination of the optimal threshold T, as explained at the end of Sect. 5.3.

5.5. The annotated anatomical structures (experiment 1). The tiles, as presented in Fig. 5, are used directly to train the CNN in experiment 2. However, in experiment 1 we use 3-channel images, additionally containing the annotated anatomical structures, so also the training tiles have three channels, as demonstrated in Fig. 6. These training tiles are obtained in a natural way, by cutting full scan images (Fig. 4c), so that the resulting dataset corresponds exactly to the 1-channel version. Apart from the necessary CNN modification, explained in

[1] four patients had been removed from the original dataset of 100 subjects, due to data format issues

Sect. 5.2 (extending the input neural layer to handle 3-channel images) all further processing in experiment 1 follows the scheme proposed in [7].

5.6. Tissue classification and segmentation (experiment 2). This experiment is based on exactly the same CNN structure and on the same (1-channel) input data as proposed in [7]. The only difference is that the output images obtained from the network in the testing phase are additionally post-processed with a domain-specific segmentation procedure.

As a segmentation method we used an algorithm specialized for nervous tissue extraction from head MRI scans, based on deformable surface models [11], embedded in Slicer 3D platform [32, 33]. This method consists of several consecutive processing stages, realized individually for each patient:

1. Construction of a 3D representation of the image data on the basis of all the available head MRI scans of the patient;
2. Computation of some basic data statistics (histogram) and determination of the starting region in the form of an ellipsoid contained completely within the brain area;
3. Initial tissue classification within the starting region (gray matter, white matter, cerebro-spinal fluid) by fuzzy clustering analysis [34];
4. Tissue classification on the whole image;
5. Iterative deformation of the surface of the starting region on the basis of tissue classification results and constraints based on the smoothness of the resulting surface.

As a result of the segmentation, we obtained the outer surface of the brain, which was then projected onto the planes defined by consecutive MRI scans. This finally led to obtaining the contour around the region occupied by the nervous tissue in every input image (Fig. 7a). The masks corresponding to these contours needed further processing to correct some imperfections of the segmentation process. A sequence of morphological operations was applied here, as a result of which the final masks were obtained, as presented in Fig. 7b:

1. Erosion – in order to remove sporadically appearing thin protrusions extending far beyond the brain region;
2. Closing – in order to eliminate frequently appearing holes within the brain region (Fig. 7a);
3. The second erosion – in order to decrease slightly the mask region to eliminate the risk of covering the skull bones surrounding the brain.

From these three steps, the second (closing operation) is the most crucial one – a big structuring element was applied, to guarantee that the resulting mask would cover the whole region of the nervous tissue, even if only a part of it was obtained after the segmentation. The basic motivation here was the need to decrease the risk of not detecting the demyelinating plaques in case they would be located in the areas not covered by the mask. On the other hand, a big structural element might result in an undesired growth of the mask area, which could in turn cover the "false positive", bright regions of the image. These regions are, however, located outside the brain region in most MRI scans, so – considering the specificity of the closing operation – this risk was not too severe and it was additionally reduced by the third morphological operation (the second erosion).

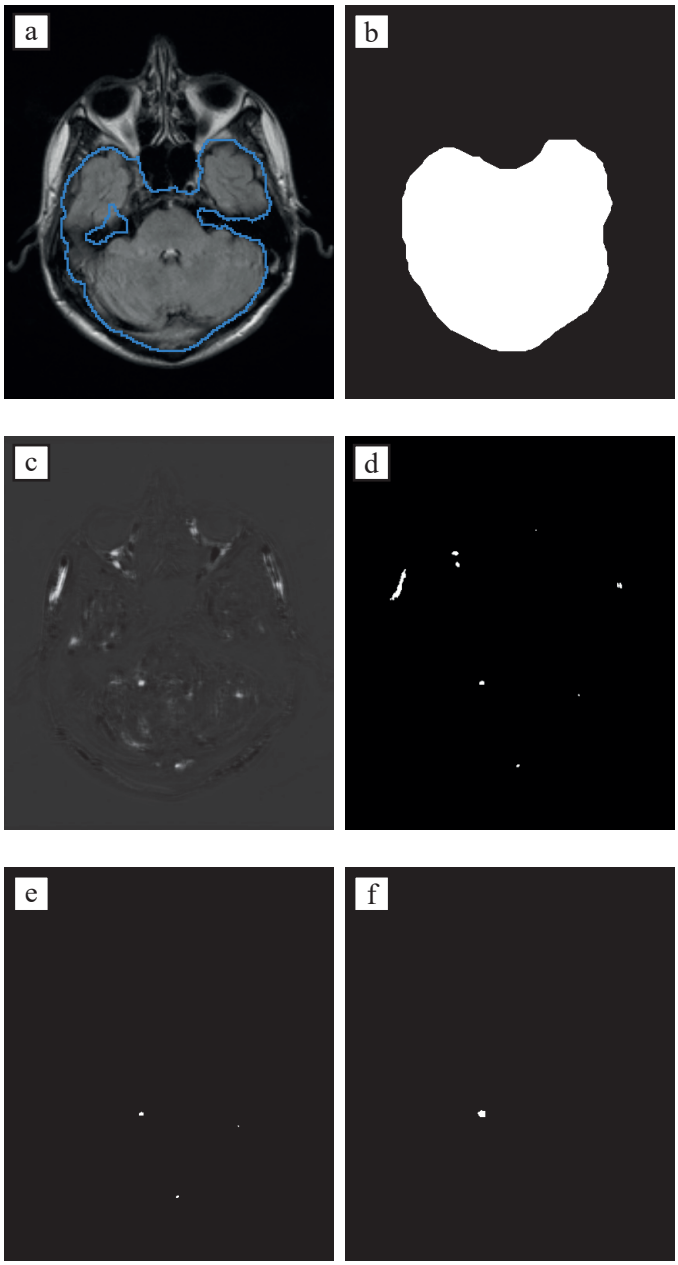


Fig. 7. Experiment 2 – example images: a) input image with segmented brain (contour); b) contour-based mask after morphological operations; c) raw CNN output image (no thresholding); d) output image after thresholding; e) output image after thresholding and filtration with mask (b); f) expected (ground–truth) output image

The masks thus obtained were used for the filtration of the CNN results in the testing phase (on the testing set). After the analysis of the characteristics of the available dataset, we decided not to use them during the training phase. This decision was motivated by the need to maintain high robustness of the network to "false positive" areas of high brightness which, notwithstanding the segmentation methods applied, might occasionally appear in the testing set. Limiting the training data to nervous tissue only would inevitably deteriorate the generalization properties of the network in this regard.

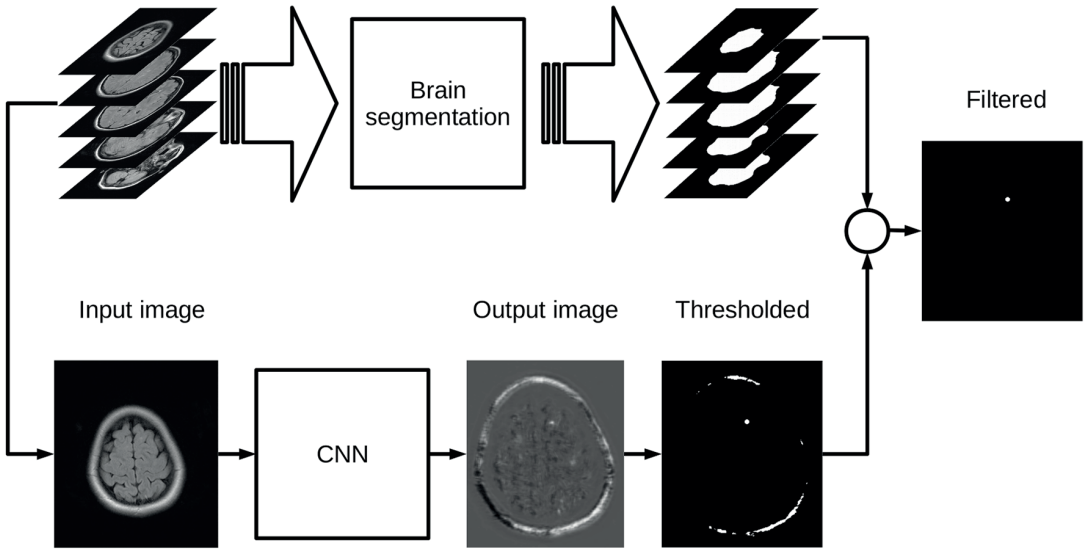


Fig. 8. Testing procedure in experiment 2; note, that: – the CNN is already trained as in [7];
– brain segmentation is obtained for all the MRI scans of a given (test) patient

Summarizing, the process of testing our neural network in experiment 2 comprised the following steps (Fig. 8):

1. Forward propagation of the input image (the whole single MRI scan) through the network;
2. Thresholding of the obtained output image (Fig. 7c) with threshold T resulting in a binary image I_T (Fig. 7d);
3. Filtration, i.e. removing all white areas located outside of the mask;
4. Comparing of thus obtained result (Fig. 7e) with the expected target output, defined by a human specialist (Fig. 7f).

6. Experimental evaluation

The experiments were done with Caffe deep learning framework [35] on a cluster node with Tesla K80A GPU accelerator. The training set of 7856 tiles was fed to the network in mini-batches of 100 tiles each. Mean square error (Euclidean loss) between the network outputs and the ground-truth target images was used as the optimization efficacy measure, in accordance to Sect. 5.3.

6.1. Experiment 1. In order to increase the reliability of the obtained results we repeated the whole training process 10 times. Apart from the 3-channel images with the annotated anatomical structures, we have also rerun the original experiment from [7] based on 1-channel MRI scans. For comparison purposes, the training in this case was also repeated 10 times. The results are presented in Table 1.

As may be observed, the influence of adding the structural, anatomical knowledge, although visible, is in fact quite limited. There are several possible reasons for this unspectacular outcome. First, although the information about localization of a ventricle or a peduncle may increase the probability of nearby MS lesion detection, still the training tiles are quite small

Table 1
Experiment 1 – the results

	1-channel [7] mean results from 10 tests	3-channels (Fig. 6) mean results from 10 tests
Precision:	51.27% ($\sigma = 1.32\%$)	53.00% ($\sigma = 1.58\%$)
Recall:	54.28% ($\sigma = 0.66\%$)	55.66% ($\sigma = 0.76\%$)
F-measure	52.73% ($\sigma = 0.93\%$)	54.29% ($\sigma = 0.94\%$)

(50×50 pixels out of full scan images of 448×512), which limits the clues on their global position that might be found. Mixing in one channel the information concerning the cerebral and cerebellar peduncles, and also genu and splenium of the corpus callosum, medulla and pons, perhaps did not help as well. Moreover, the dark regions of the lateral ventricles are already quite characteristic in the original MRI scans, so putting this part of the additional information in a separate channel was probably mostly redundant.

However, it should be noted that the observed increase of the F-measure value is statistically significant. The training process was very stable in terms of repeatability (standard deviation of F-measure values below 1%) and exactly the same conditions were held for both datasets (1- and 3-channel images). As the obtained results are quite sensitive to the moment in which the training is stopped, as demonstrated in [7], we applied here exactly the same number of epochs for both datasets and for all the tests (training sessions)[2]. More precisely, each training lasted 1250 epochs, which means that 9 820 000 tiles were used (1250 epochs × 7856 tiles in the dataset). This number of ep-

[2] we observed that the training time was slightly longer for the 3-channel images (ca. 911 vs 908 minutes for the 1-channel variant), due to the increased number of neural weights in the first convolutional layer

ochs was close to optimal in [7], although it should be noted that for the extended 3-channel dataset, a longer training procedure would perhaps be beneficial. In this context, the comparison of the range of the results obtained here for the original 1-channel database[3] (51.48% – 54.03% in 10 repetitions) and for the 3-channel version (53.09% – 56.18% in 10 repetitions) clearly supports the conclusion that the anatomical information encoded in the separate channels of the input image does have a positive impact on our detection process.

6.2. Experiment 2. In order to precisely measure the significance of tissue classification, we did not actually repeat the training process – we only applied the masks to filter the output images obtained in [7]. The original, reference results (without the filtering) are presented in the first column of Table 2. The values in the middle column present the result of our mask-based filtration process described in Sect. 5.6.

As may be seen, an increase of the F-measure value by almost 4% was achieved, which resulted from a significant increase of precision. This effect is quite natural, when we take into account that the irrelevant areas (not containing lesions) have been filtered out now, while practically all properly detected lesions have been preserved (recall drop by as little as 0.01%).

The obtained result proves the effectiveness of the filtration, although – considering how much additional information about the analyzed areas is included – one might expect a more significant enhancement of the result. Therefore, we analyzed the influence of the value of the threshold T on the obtained results. In the middle column of Table 2 we present the results obtained for the same value of $T = 0.56$ as in [7]. However, as the filtration of the resulting images with the use of the masks obtained on the basis of nervous tissue segmentation introduces quite significant changes, we computed new threshold, maximizing F-measure on filtered training images (Table 2, the third column).

As one could expect, the threshold value thus obtained was a bit smaller ($T = 0.52$), which led – in the testing phase – to the increase of the recall value by ca 3%, while the precision dropped by over 3%, though still remaining slightly higher with respect to the results of [7]. Nevertheless, it should be noted that despite these changes, the final F-measure value remained practically on the same level (0.05% increase).

Continuing the search for the explanation of the observed moderate enhancement of the results, we computed the statistics of the changes introduced by the filtration operation in the whole testing set. It appeared that using the additional information from the brain tissue classification process was significant in only 30% of the testing images. In the remaining 70% of cases the resulting images after thresholding did not contain "false positive" areas detected outside the brain region, so the filtration did not introduce any significant changes in their case. Therefore, the final conclusion from the experiment is that the applied CNN in conjunction with a proper training and good choice of the training examples is an effective classifier of demyelinating plaques, robust to "false positive" areas appearing in the analyzed images even without the application of the additional anatomical knowledge about the tissue type. On the other hand, using this knowledge still leads to some improvements of the obtained results, which offers a significant potential for future research.

Table 2
Experiment 2 – the results. The numbers in the 2nd, 3rd and 4th row refer to the total number of pixels in the final output images

	No Segmentation [7]	No Segmentation (threshold found without segmentation)	Segmentation
Threshold T	0.56	0.56	0.52
TP + FN	143 207	143 207	143 207
TP + FP	140 339	121 750	136 903
TP	77 405	77 388	81 888
Precision: $\frac{TP}{TP + FP}$	55.16%	63.56%	59.81%
Recall: $\frac{TP}{TP + FP}$	54.05%	54.04%	57.18%
F-measure	54.60%	58.42%	58.47%

6.3. Comparison with other approaches. Considering our results in a broader context, we should note the fact, that the considered problem is generally a difficult one and the results we have obtained are on par with up-to-date studies. The direct comparison is difficult due to different methodologies, evaluation procedures and the MRI data itself. The quality of the material and the coherence of the annotations may also vary significantly, as shown in [13], who report the mean results of MS lesion segmentation for 45 patients from three different hospitals as being equal to 22%, 43% and 44%, respectively. These figures refer to Dice similarity coefficient (DSC) which is computed identically to the F-measure[4] we report in Tables 1 and 2. In this context we may regard our results as being definitely superior.

Similarly moderate results are reported in [14] (DSC equal to 29%) and in [36] (DSC equal to 31%). Only in some recent works based on machine learning, significantly better results are reported, e.g. DSC of 62.7% in [18] and 68% in [15].

The convolutional neural network used in [18] yielded results better by 4.23% than ours (DSC: 62.7% vs. 58.47). However, they used incomparably richer training data, including

[3] even if we consider the best result of 54.60% reported for the 1-channel case in [7]

[4] other alternative names of the same measure include: Sørensen–Dice index, F1-score and Czekanowski's index.

four MRI scan types[5] (T2-weighted FLAIR, T1-weighted MPRAGE, T2-weighted and Proton Density weighted), all three possible projections[6] (axial, coronal and sagittal) and repeated scans for different time points (4–6 time points). They also used several independent convolutional models, all of which used a fully-connected layer at the end of the architecture, which is a significant difference between their approach and ours.

The approach proposed by [15] is based on a complex procedure, in which two optimization/classification methods (SVM/GA) are used for analysis of feature vectors, including i.a. GLCM/GLRLM texture descriptors, computed from volumetric wavelet fusion of several MRI modalities. Three-dimensional data are preprocessed with several specialized tools, including BET-FSL for registration, realignment and brain extraction [16] and nonlinear anisotropic diffusion filter [17]. In this context, our approach may be considered a simple and straightforward one, offering good segmentation capabilities without the need of costly preprocessing and feature extraction.

Finally, let us note that inter-variability between experts, reported by [15], in terms of the Dice coefficient, was ca 25%, ca 70% and ca 75% for three different brain MRI databases. This gives a practical assessment of the upper limit of the MS lesion segmentation capabilities of any classifier.

7. Conclusions and Future Work

In this paper, we tested two methods of anatomical data inclusion within our CNN-based demyelinating plaques detection procedure. Both methods – the first one, based on embedding the anatomical data in the input images and the second one, based on post-processing and filtration of the output images – increased the detection rate, although neither of them proved groundbreaking.

The obtained results may be interpreted in two ways. Firstly, we can conclude that the presented convolutional neural network is robust enough to yield good results which simply cannot be significantly boosted, considering the inevitable uncertainty of the annotation process and the quality of the available MRI scans. On the other hand, some enhancements of the results were actually achieved, indicating the benefits that might be expected if more information on the anatomical context was available during the training/testing process.

How to provide this information is still an open question. The method applied in experiment 1 is quite natural for a convolutional network, designed to process multi-channel images. The drawbacks include probably too generalized treatment of the available data (only two additional channels applied) and the laborious, manual annotation process. The latter issue may be resolved in a similar way as in experiment 2, i.e. via specialized tools capable of registering the medical imagery data and atlas-based segmentation of anatomical structures. Putting thus obtained data into several separate channels of the input images may improve the training and increase the lesion detection capabilities of the network.

It should be noted that the two experiments reported in this paper can be combined. This could help to further enhance the overall detection outcome, as the additional information used in each experiment is independent of the other. However, taking into account the presented results one cannot expect the F-measure value significantly higher than 60%. On the other hand, we are aware that the applied methodology of evaluation is partially misleading – we are currently expecting the CNN to recreate the exact shape of the specialist's annotations, down to single pixels and we simply report the number of the pixels that do not match. With account being taken of the unavoidable inaccuracy and uncertainty of the annotation process, it is completely unrealistic to expect a 100% detection rate. Giving more priority to the evaluation of *the number of lesions* that have been found, instead of the number of the matching pixels, would probably be more clinically justified and would generate better-looking detection scores. The problems concerning the formulation of more flexible evaluation measures and the general reliability of the manual annotations used for training are among the open issues that we will try to resolve in the future work.

Acknowledgements. This project has been funded with support from National Science Centre, Republic of Poland, decision number DEC-2012/05/D/ST6/03091.
Authors would like to express their gratitude to the Department of Radiology of Barlicki University Hospital in Lodz for making head MRI sequences available.

References

[1] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics* 36, 193–202, 1980.
[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in CVPR09, 2009.
[3] A. de Brebisson and G. Montana, "Deep neural networks for anatomical brain segmentation," *ArXiv e-prints* 1502.02445, Feb. 2015.
[4] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," *ArXiv e-prints* 1606.04797, June 2016.
[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *ArXiv e-prints*, 1505.04597, May 2015.
[6] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *ArXiv e-prints*, vol. 1605.06211, May 2016.
[7] B. Stasiak, P. Tarasiuk, I. Michalska, A. Tomczyk, and P.S. Szczepaniak, "Localization of demyelinating plaques in MRI using convolutional neural networks," in *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017)* 55–64, 2017.
[8] R. Milo and E. Kahana, "Multiple sclerosis: Geoepidemiology, genetics and the environment," *Autoimmunity Reviews* 9 (5) A387–A394, 2010.

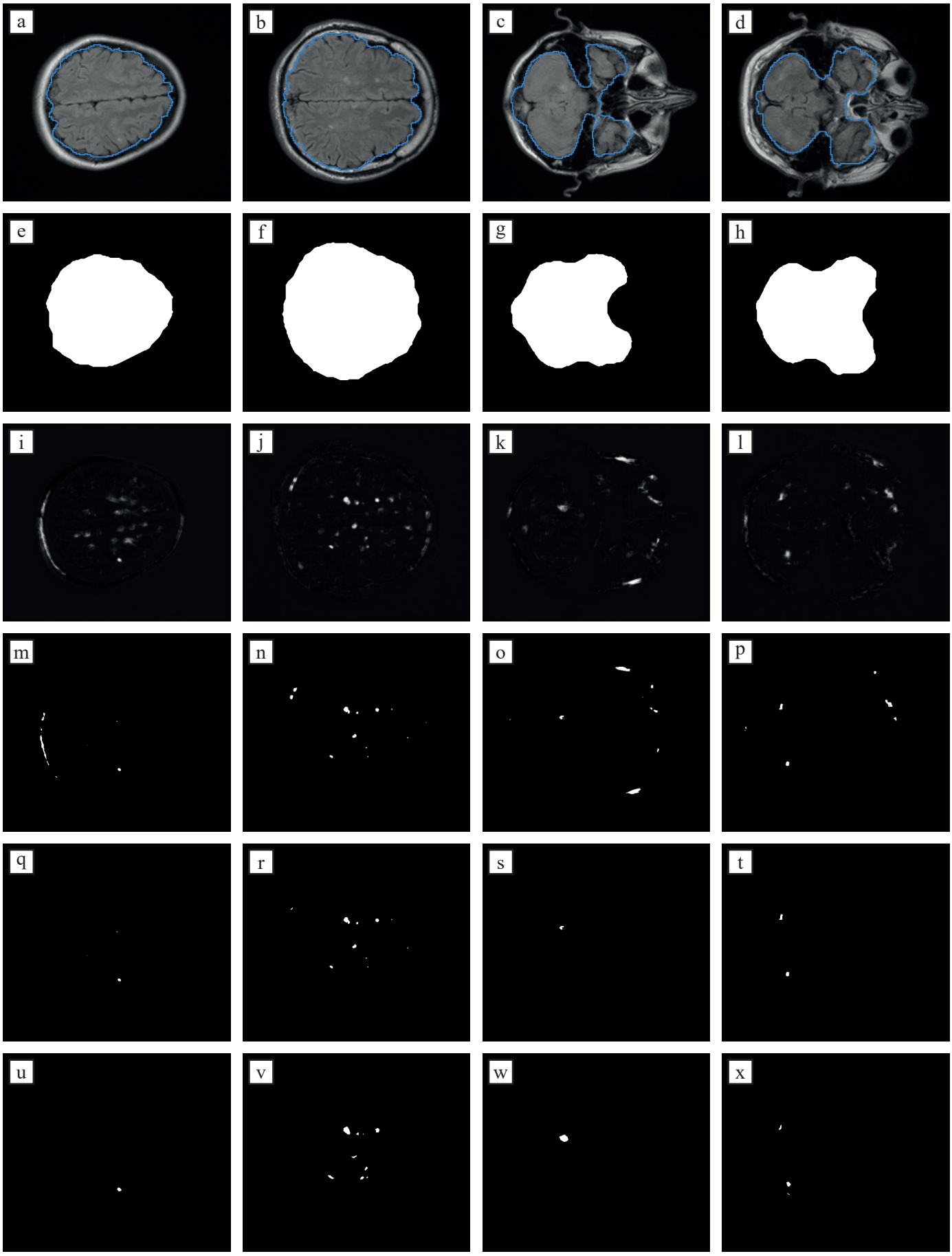
[5] we use only one modality – the FLAIR MRI scans.
[6] we use only axial plane.

[9] K. Berer and G. Krishnamoorthy, "Microbial view of central nervous system autoimmunity," *FEBS Letters* 588 (22), 4207–4213, 2014.
[10] P.M. Parizel, L. van den Hauwe, F. De Belder, J. Van Goethem, C. Venstermans, R. Salgado, M. Voormolen, and W. Van Hecke, *Magnetic Resonance Imaging of the Brain,* 107–195. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.
[11] X. Tao and M.-C. Chang, "A skull stripping method using deformable surface and tissue classification," 2010.
[12] I. Kapouleas, "Automatic detection of white matter lesions in magnetic resonance brain images," *Computer Methods and Programs in Biomedicine* 32 (1), 1–35, 1990.
[13] M. Cabezas, A. Oliver, E. Roura, J. Freixenet, J.C. Vilanova, L. Ramió-Torrentà, Àlex Rovira, and X. Lladó, "Automatic multiple sclerosis lesion detection in brain MRI by flair thresholding," *Computer Methods and Programs in Biomedicine*, 115 (3 147–161, 2014.
[14] N. Weiss, D. Rueckert, and A. Rao, "Multiple sclerosis lesion segmentation using dictionary learning and sparse coding," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013* (K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, eds.), 735–742, Springer Berlin Heidelberg, 2013.
[15] O. Ghribi, L. Sellami, M.B. Slima, C. Mhiri, M. Dammak, and A.B. Hamida, "Multiple sclerosis exploration based on automatic MRI modalities segmentation approach with advanced volumetric evaluations for essential feature extraction," *Biomedical Signal Processing and Control* 40, 473–487, 2018.
[16] FMRIB, "FMRIB centre, Nuffield Department of Clinical Neurosciences, University of Oxford, Brain extraction tool (BET)," 2012.
[17] G. Gerig, O. Kubler, R. Kikinis, and F. Jolesz, "Nonlinear anisotropic filtering of MRI data," 11 (2), 221–232, 1992.
[18] A. Birenbaum and H. Greenspan, "Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks," in *Deep Learning and Data Labeling for Medical Applications* (G. Carneiro, D. Mateus, L. Peter, A. Bradley, J.M.R.S. Tavares, V. Belagiannis, J.P. Papa, J.C. Nascimento, M. Loog, Z. Lu, J.S. Cardoso, and J. Cornebise, eds.), (Cham), pp. 58–67, Springer International Publishing, 2016.
[19] W.S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, 5 (4), 115–133, 1943.
[20] D.C. Cireşan, U. Meier, J. Masci, L.M. Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence – Volume Volume Two*, IJCAI'11 1237–1242, 2011.
[21] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.
[22] D.H. Hubel and T.N. Wiesel, "Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat," *Journal of Neurophysiology* 28, 229–289, 1965.
[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE* 2278–2324, 1998.
[24] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time-series," in *The Handbook of Brain Theory and Neural Networks* (M. A. Arbib, ed.), MIT Press, 1995.
[25] M.D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *CoRR* abs/1311.2901, 2013.
[26] T.V. Nguyen, C. Lu, J. Sepulveda, and S. Yan, "Adaptive non-parametric image parsing," *CoRR* abs/1505.01560, 2015.
[27] K.R. Mopuri and R.V. Babu, "Object level deep feature pooling for compact image representation," *CoRR* abs/1504.06591, 2015.
[28] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network.," *Neural Networks*, 16 (5–6), 555–559, 2003.
[29] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," *CoRR* abs/1412.1283, 2014.
[30] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing* 54 (12), 7405–7415, 2016.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *CoRR* abs/1502.01852, 2015.
[32] R. Kikinis, S. D. Pieper, and K. G. Vosburgh, *3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support* 277–289. New York, NY: Springer New York, 2014.
[33] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion- Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J. Miller, S. Pieper, and R. Kikinis, "3D slicer as an image computing platform for the quantitative imaging network," *Magnetic Resonance Imaging* 30 (9), 1323–41, 2012.
[34] M.-C. Chang and X. Tao, "Subvoxel segmentation and representation of brain cortex using fuzzy clustering and gradient vector diffusion," 2010.
[35] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.
[36] R. Mechrez, J. Goldberger, and H. Greenspan, "Patch-based segmentation with spatial consistency: Application to MS lesions in brain MRI," *Journal of Biomedical Imaging* 2016, 3:3–3:3, Jan. 2016.

8. APPENDIX

On the following page we demonstrate some more results of experiment 2. Each column presents a separate MRI scan, while each row presents the results of the separate stages of processing:

- Contour obtained on the basis of brain segmentation, i.e. the "raw" result of the segmentation algorithm a – d;
- Mask obtained on the basis of brain segmentation and the morphological operations e – h;
- Resulting images obtained at the output of the neural network, without thresholding i – l;
- Result of thresholding the network output without filtration m – p;
- Result of thresholding the network output and filtration on the basis of the masks q – t;
- Expected (ground–truth) output, i.e. the specialist's annotations u – x.


a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t
u
v
w
x