A Convolutional Neural Network for the Automatic Diagnosis of Collagen VI-related Muscular Dystrophies

Adrián Bazaga[a,b,c], Mònica Roldán[d,e], Carmen Badosa[f], Cecilia Jiménez-Mallebrera[f,g,*], Josep M. Porta[a,*]

[a] *Institut de Robòtica i Informàtica Industrial, UPC-CSIC, 08028 Barcelona, Spain*
[b] *Department of Genetics, University of Cambridge, Cambridge, United Kingdom*
[c] *STORM Therapeutics Ltd, Cambridge, United Kingdom*
[d] *Unitat de Microscòpia Confocal, Servei d'Anatomia Patològica, Institut Pediàtric de Malalties Rares, Hospital Sant Joan de Déu, 08950 Barcelona, Spain*
[e] *Institut de Recerca Sant Joan de Déu, Hospital Sant Joan de Déu, 08950 Barcelona, Spain*
[f] *Neuromuscular Unit, Institut de Recerca Sant Joan de Déu, Hospital de Sant Joan de Déu, 08950 Barcelona, Spain*
[g] *Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Instituto de Salud Carlos III (ISCIII), Unidad 703*

Abstract

The development of machine learning systems for the diagnosis of rare diseases is challenging, mainly due to the lack of data to study them. This paper surmounts this obstacle and presents the first Computer-Aided Diagnosis (CAD) system for low-prevalence collagen VI-related congenital muscular dystrophies. The proposed CAD system works on images of fibroblast cultures obtained with a confocal microscope and relies on a Convolutional Neural Network (CNN) to classify patches of such images in two classes: samples from healthy persons and samples from persons affected by a collagen VI-related muscular distrophy. This fine-grained classification is then used to generate an overall diagnosis on the query image using a majority voting scheme. The proposed system is advantageous, as it overcomes the lack of training data, points to the possibly problematic areas in the query images, and provides a global quantitative evaluation of the condition of the patients, which is fundamental to monitor the effectiveness of potential therapies. The system achieves a high classification performance, with 95% of accuracy and 92% of precision on randomly selected independent test images, outperforming alternative approaches by a significant margin.

Keywords: Convolutional neural networks, Deep learning, Classification, Computer aided diagnosis, Confocal microscopy images.

1. Introduction

Deficiencies in the structure of collagen VI are a common cause of neuromuscular diseases with manifestations ranging from the Bethlem myopathy to the severe Ullrich congenital muscular dystrophy. Their symptoms include proximal and axial muscle weakness, distal hyperlaxity, joint contractures, and critical respiratory insufficiency, requiring assisted ventilation and resulting in a reduced life expectancy. Moreover, the

*Corresponding authors.
Email addresses: ar989@cam.ac.uk (Adrián Bazaga), mroldanm@sjdhospitalbarcelona.org (Mònica Roldán), mcbadosa@fsjd.org (Carmen Badosa), cjimenezm@fsjd.org (Cecilia Jiménez-Mallebrera), porta@iri.upc.edu (Josep M. Porta)

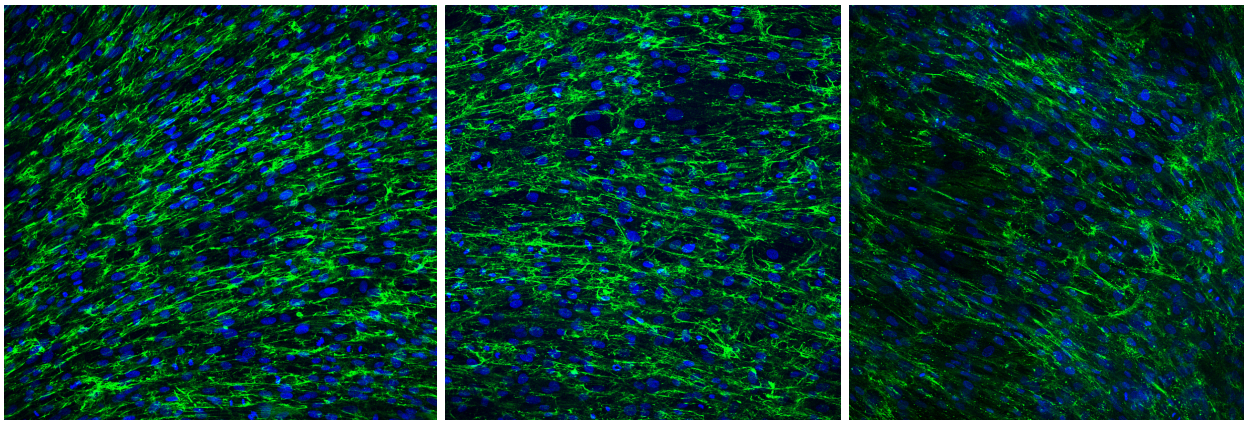

Figure 1: Confocal microscopy images of fibroblast cultures. Left: A control sample. Center: Sample from a patient with the Bethlem myopathy. Right: Sample from a patient of the Ullrich muscular dystrophy. In the three images, the network of collagen is shown in green and the fibroblast nuclei in blue.

skin and other connective tissues where collagen VI is abundant are also affected [1, 2]. As described in the Online Mendelian Inheritance in Man (OMIM) [3] database entries 254090 and 158810, the collagen VI structural defects are related to mutations of three main genes, namely genes COL6A1, COL6A2, and COL6A3. However, despite the implementation of the next generation of gene sequencing technologies, the diagnosis is still challenging. This happens in general for diseases caused by dominant mutations, where there is not a complete absence of a main protein, and when the effect of a genetic variant on the protein structure may not be evident. Thus, before any genetic analysis, the standard technique for the diagnosis of collagen VI-related dystrophies is the analysis of images of fibroblast cultures [4] (see Fig. 1). Several aspects of the images, such as the orientation of the collagen fibers, the distribution of the collagen network, and the arrangement of cells in such network, are taken into account by the specialists to identify potential patients. However, this evaluation is only qualitative, and the regulatory agencies would not approve any treatment (such as, for instance, gene editing via the CRISPR technology) without an objective methodology to evaluate its effectiveness [5]. Thus, there is an imperative need for accurate methodologies to quantitatively monitor the effects of any possible new therapy.

This paper presents the first Computer-Aided Diagnosis (CAD) system for dystrophies caused by defects in the structure of collagen VI. The addressed problem entails classifying confocal microscopy images from fibroblast cultures in two classes: samples obtained from healthy persons (labeled as *control*) and samples obtained from persons affected by any kind of collagen VI-related muscular distrophy (labeled as *patient*). The proposed classification system relies on a Convolutional Neural Network (CNN) trained from cases labeled by specialists. However, as in the case of all rare diseases, there is a severe lack of training data. To address this issue, we propose a patch-based data-augmentation technique similar to the one described in [6]. Thus, the available labeled images are split in non-overlapping patches and the CNN is trained on them. In the diagnostic step, the CAD system classifies the patches of the query images using the CNN, and provides an overall diagnostic using majority voting.

This paper is structured as follows. Section 2 frames the contribution of this paper in the context of CAD systems for the classification of medical images, and Section 3 introduces the basic methods used to develop the CAD system presented in this paper, including the image acquisition procedure, the proposed data augmentation method, and the CNN used to classify the patches of the input images. Next, Section 4 compares the performance of the proposed system with alternative approaches and, finally, Section 5 summarizes the paper and points to issues deserving further attention

2. Related work

The automatic classification of medical images is traditionally divided into two steps. In the first one, relevant features are extracted from the image. In the second step, these features are used to group the input images in classes, which in our case are the control and patient classes.

Several methodologies have been proposed to identify relevant features in medical images. For instance, in [7, 8] a Gaussian filterbank is used to find perturbations in chest X-ray images. However, in our case, the disease causes a global disorder of the collagen network structure, rather than local perturbations on it. In [9, 10], the authors use a Principal Component Analysis (PCA) technique to extract relevant features from Single-Photon Emission Computed Tomography (SPECT) or Positron Emission Tomography (PET) images. These features, however, are linearly related to the input image and such a simple transformation is not expected to provide relevant features in our case. In other approaches, Local Binary Patterns (LBP) [11, 12] and Gray-Level Co-occurrence Matrices (GLCM) [13, 14] techniques have been used to generate features to characterize medical images. None of these techniques, though, are specific for the problem tackled in this paper and, as we show in Section 4, they lead to a sub-optimal classification performance. Some procedures have been proposed to extract features from collagen images [15, 16]. However, none of such techniques are specially tailored for the defects in collagen VI and, consequently, they do not provide adequate insights for our problem, as also confirmed by the experiments in Section 4. Thus, in the particular problem addressed in this paper, the selection of relevant features to characterize the images is a challenge in itself.

In a second step, a classical approach would use the computed features to classify the images. Since there is not a single method adequate for all kinds of problems, different types of classifiers have been used on features extracted from medical images, including naive Bayes classifiers [17], Support Vector Machines (SVM) [18, 19], Random Forests (RF) [20, 21], and Convolutional Neural Networks (CNN) [22–28]. CNNs are an evolution of the classical Artificial Neural Networks (ANN). However, there are two main differences between ANNs and CNNs. First, CNNs are invariant to translations in the image due to shared parameters between the image convolutions. Second, new types of layers are used in the CNNs, such as the so-called max-pooling layers, which enlarge the receptive field of the convolutions and also contribute to the translation invariance. Such improvements, together with the availability of specialized hardware and software for training CNNs have established them as the *de-facto* standard tool for image processing, provided that enough training data is available [29]. A key advantage of the CNNs over the rest of the approaches is that they can directly

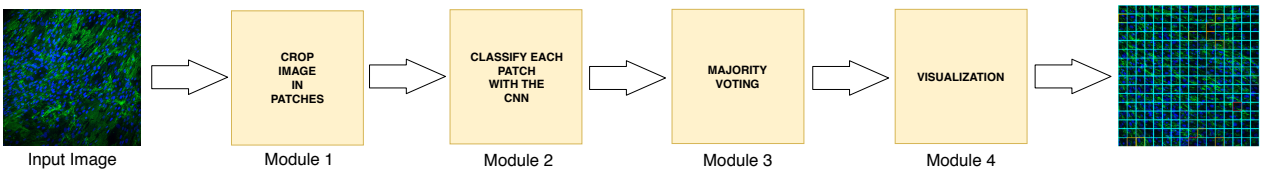

Figure 2: Overview of the proposed automatic diagnosis system using a majority voting on the individual patches decisions of the CNN model. The system also provides a detailed visualization of the diagnosis.

operate on the input images, coping with the possible noise in the images, taking care of identifying relevant features, and performing the classification relying on them [30, 31]. This is the fundamental reason to apply this technique to the problem addressed in this paper.

3. Materials and Methods

Figure 2 provides a global view of the proposed CAD system. The system is divided into four modules. The first module receives a full image and splits it into non-overlapping patches of 64x64 pixels. The second module is formed by a CNN classification model, that reads the patches and generates an independent prediction for each one of them. The third module receives the local decisions for each patch and takes a global decision using majority voting (i.e., an image is said to be of a control/patient if most of the patches are classified as such). The last module visualizes the input image and the decision for each patch represented in a color code (see Fig. 3). In this regard, cyan is used to frame patches with more than 90% probability of belonging to the control class, steel blue is used for patches with probability between 70% and 90%, yellow for patches with probability between 50% and 70%, orange for patches with probability between 30% and 50%, and finally red for patches with less than 30% of probability of belonging to the control class. This color code offers the possibility of easily spotting potentially relevant areas in the image. The system also provides the overall decision on the image and a global score computed as the percentage of patches classified as control in the image. This score enables to track the evolution of a patient and, thus, it offers a tool to assess whether a particular treatment is effective. Next, we describe in detail the three key elements of the proposed CAD system: the process to obtain the input images, the image patch generation to augment the number of training examples, and the CNN used to classify these patches.

3.1. Image acquisition

To obtain the training images, samples from the forearm were obtained from patients, as well as from age-matched controls. Primary fibroblasts cultures were established using standard procedures [4]. Patient and control samples were treated in parallel, with 25 μg/mL of L-ascorbic acid phosphate magnesium (Wako Chemicals GmbH, Neuss, Germany) for 24 hours. After that time, cells were fixed with 4% paraformaldehyde in phosphate-buffered saline solution. Collagen VI was detected by indirect immunofluorescence using a monoclonal antibody (MAB1944, Merck, Germany) and fibroblast nuclei were stained using 4,6-diamidino-2-phenylindole (Sigma Chemical, St. Louis, USA).

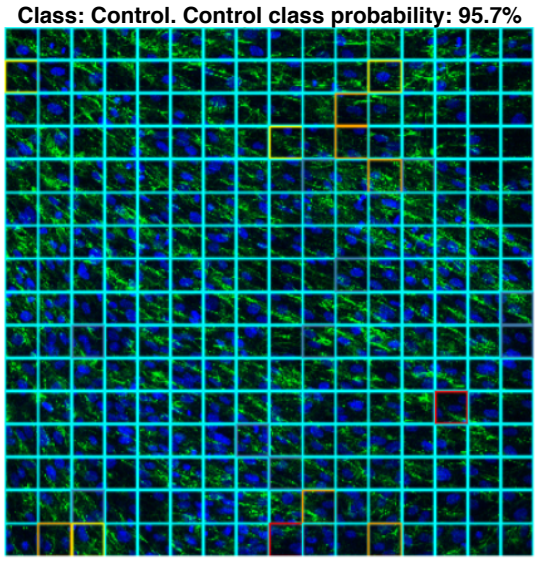

Figure 3: Visualization of the diagnosis of a given fibroblast culture image. Each patch of the image is colored according to its probability of belonging to the control class. The system also gives an overall score computed as the percentage of patches classified as control in the image.

Images of the samples with 1024x1024 pixels were obtained with a Leica TCS SP8 X White Light Laser confocal microscope with hybrid spectral detectors (Leica Microsystems, Wetzlar, Germany), an HCX PL APO 20x/0.75 dry objective, and the confocal pinhole set to 1 Airy unit. Collagen VI was excited with an argon laser (488 nm) and detected in the 500-560 nm range. Nuclei were excited with a blue diode laser (405 nm) and detected in the 420–460 nm range. Appropriate negative controls were used to adjust confocal settings to avoid non-specific fluorescence artifacts. The detector gain and offset values were adjusted to use the entire dynamic rate of the detector (12 bits), but avoiding oversaturated voxels. Sequential acquisition settings were used to avoid inter-channel cross-talk. Ten sections of each sample were acquired every 1.5 μm along the focal axis (Z-stack) and combined into an integrated intensity projection to form a single image.

Finally, the training images were labelled either as control or patient by a team of expert doctors. For this reason, we can safely assume that there is no noise or incorrectness in the labels of the data used for training the proposed system.

3.2. Image patch generation

Since the image acquisition procedure is complex and time-consuming, and we are dealing with a rare disease, the available samples are limited. For this reason, to generate enough data to train the CNN, we propose the data augmentation scheme described next.

The data is augmented by splitting the available full-size images into small, non-overlapping patches. Each patch is used as an independent input to train the CNN model. In our approach, the patch is of size 64x64 pixels, since it is a particularly relevant window size for CNN classification models on similar images [32]. In our case, the patches capture the main features of the input images, i.e., a significant portion of the collagen

network and several nuclei or parts of them.

Since our input images are of 1024x1024 pixels, 256 patches are generated from each image. Each patch is further transformed to get even more variations of the data. The patches are rotated clockwise by 90, 180, 270 and 360 degrees and every rotated patch is flipped horizontally. Thus, eight different variations of each original patch can be obtained and, consequently, each input image generates 2048 training inputs for the CNN. Since initially we have 276 images, with the proposed data augmentation process, the training and testing sets include, respectively, 56320 and 14336 patches, without taking into account the rotated and mirrored patches. The patches are normalized so that they have zero mean and unit variance. Normalization reduces the effect of possible noise in the images and improves the learning process avoiding its premature convergence.

3.3. A CNN to classify image patches

A CNN is composed of a set of simple computational elements, called neurons, arranged in layers and connected to elements in the previous layers. The layers are evaluated in sequence, from the input layers directly connected to the image pixels to the output decision layers. The computational elements are fixed and, consequently the parameters encoding the learned function are the weights associated with the connections between the neurons in the different layers of the network.

Three different kinds of layers are typically present in a CNN: convolutional layers, pooling layers, and fully-connected layers [33]. Convolutional layers handle noise and extract features applying convolutional kernels on the input image. Formally, the value y of a given kernel is computed as

$$y = \sum_{n=1}^{N} I_n k_n, \tag{1}$$

where I_n is a pixel in the image, k_n is n-th kernel weight, and N is the kernel size, i.e., the amount of pixels in the image affected by the kernel. Convolutions are typically applied on squared areas around selected pixels in the input image and, thus, the result of this operation, is a new image, Y, which is called a feature map. Often, the output of a kernel is passed through an activation function. For instance, in our approach, each kernel is followed by a rectified linear unit (ReLU) activation function,

$$f(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases}, \tag{2}$$

which eliminates negative inputs, introducing non-linearities in the CNN with a low computational cost. The horizontal/vertical distance (in pixels) between the centers of two consecutive kernels is known as the stride of the kernel. Strides larger than one are typically used to downsample a feature map, i.e., to generate higher-level features. Alternatively, the pooling layers down-sample the input by summarizing a patch in the image with a single value. For instance, in our work, we use the max-pooling operation, where the output is the maximum value among the inputs.

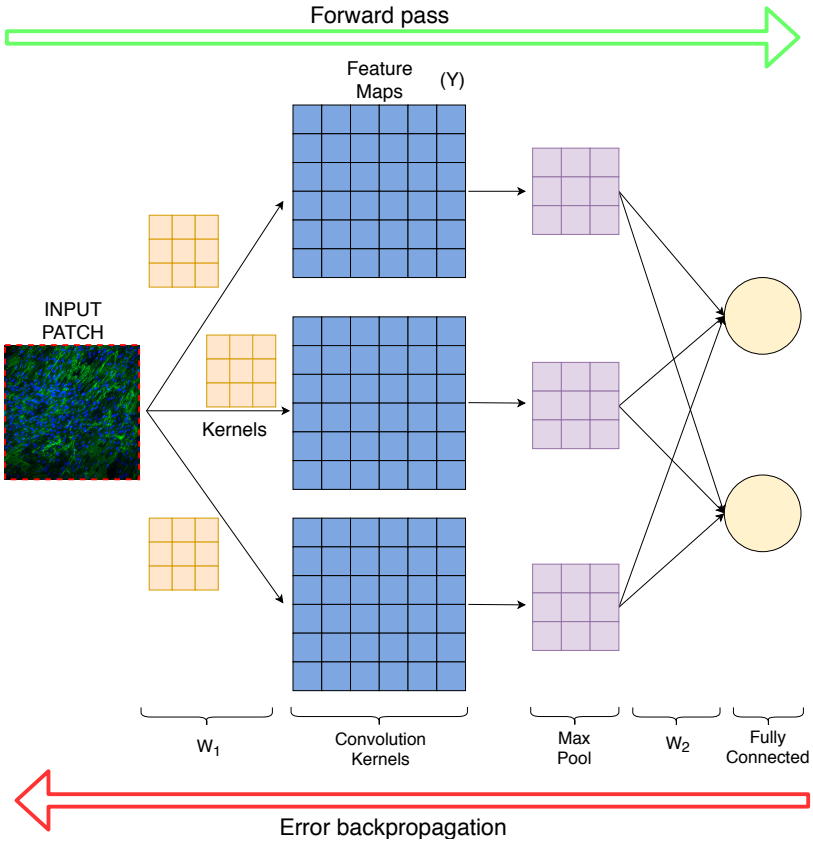

Figure 4: Process of training of a CNN. A forward pass through the CNN calculates the activation of neurons. The error between the output and the known ground-truth is sent back through the CNN to compute the gradient of the error with respect to the sets of weights (W_1 and W_2 in this illustration). Then, gradient descent is used to update the weights.

Finally, the fully-connected layers are typically placed at the end of the CNN, and provide the classification decision by applying an activation function. In our case, we use the sigmoid activation function for binary classification, defined as

$$\sigma(x) = \frac{1}{1+e^{-x}}, \tag{3}$$

where x is a weighted sum of inputs and where the output is a number between 0 and 1.

As depicted in Fig. 4, a CNN learns comparing its outputs with the ones given in a labeled dataset. More formally, we want to make the output o_i and the target t_i to be as similar as possible for each input instance $i = 1,...,n$, where n is the number of instances, i.e., we want to minimize the error

$$E(W_k) = \frac{1}{2}\sum_{i=1}^{n} ||o_i(W_k) - t_i||^2, \tag{4}$$

with W_k denoting the weights of the CNN determining the output at iteration k. The weights are usually initialized with random values, that are updated to minimize the aforementioned error with a gradient descent rule as

$$W_{k+1} = W_k - \alpha \frac{\partial E(W_k)}{\partial W_k}, \tag{5}$$

Table 1: Details of the CNN architecture proposed to classify the 64x64 pixels image patches.

Layer	Type	Number of neurons (output layer)	Kernel size	Stride
1	Convolution	64x64x128	3x3	1
2	Max Pooling	32x32x128	2x2	2
3	Convolution	32x32x64	3x3	1
4	Max Pooling	16x16x64	2x2	2
5	Convolution	16x16x32	3x3	1
6	Max Pooling	8x8x32	2x2	2
7	Fully Connected	150	-	-
8	Fully Connected	2	-	-

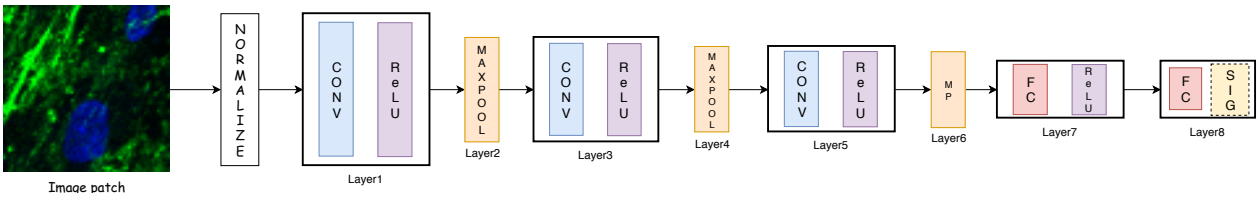

Figure 5: Schematic representation of the proposed CNN architecture.

where α is the learning rate determining the intensity of the gradient descent. The backpropagation algorithm [34] is an efficient way to compute the gradient of the error with respect to the weights, and it is the approach used in this work.

For binary classification tasks, as the one considered in this paper, the available data is usually split into two randomly-selected sets: a training set typically containing about 80% of the data and a test set, containing the remaining data. A particular batch of training inputs is used at each iteration and the process is repeated until all the training inputs are used. The iteration over all the available training set is known as an epoch. After each epoch, the performance of the CNN is evaluated using a validation set formed by 10% of the training data. Finally, after training for a certain number of epochs, the CNN generalization capability is evaluated using the test set, which was not used for training. Typically, cross-validation is also used to ensure the robustness of the evaluation results with respect to any possible bias or noise in the input data. In this procedure, the training process is repeated with different splits of the data in training and test sets, and the average accuracy is calculated as an indicator of the model generalization capabilities.

The evaluation with a test set not used for training and the cross-validation procedure reduce the chances of overfitting, which is a major issue happening when a CNN perfectly classifies the training data but is unable to correctly classify unseen cases. Dropout is another mechanism to prevent overfitting and to deal with the possible noise in the input data [35]. With this mechanism, some neurons are deactivated with a given probability. These elements do not participate in the activation of posterior neurons nor in the error

correction, i.e., they are virtually removed from the CNN. The benefit is that the network becomes less sensitive to specific weights.

We experimentally identified a CNN architecture suitable for the addressed problem. In this experimental selection, two concepts were considered: the size of the network (with a small/large number of convolution kernels) and the abstraction level (with increasing/decreasing sizes of the subsequent layers of the CNN). Experimentally, small CNNs with a decreasing number of features provided the best results. A key characteristic of the selected architecture is that many low-level features are learned in the first layers and few high-level discriminating features of the images are generated in deeper layers. The proposed architecture is detailed in Table 1 and is similar to the one in [6], which has already been proven to be particularly adequate for image classification. However, our network is smaller, since the classification task addressed here is simpler than the one addressed in [6].

Figure 5 provides a schematic representation of the best CNN architecture found in the selection process. In this architecture, the image patch is passed through three convolutional layers (layers 1, 3, and 5), which include the application of ReLU activation functions after the kernel computation. The first convolution layer defines 128 feature maps of size 64x64, the second one defines 64 feature maps of size 32x32, and the last one defines 32 feature maps of size 16x16. The reduction in the size of the feature maps is obtained with max-pooling layers (layers 2, 4, and 6), with a stride of 2 following the convolution layers. After the feature generation layers, the classification is implemented with two fully-connected layers (layers 7 and 8). The first one has 150 neurons followed by a ReLU activation function and trained with a dropout mechanism with probability 0.5. The second fully connected layer has 2 neurons, whose output is truncated into a single binary output by a sigmoid activation function to provide the final classification.

4. Results

The proposed CNN-based system was implemented using the Python programming language and the Keras 2.2 library [36] with TensorFlow 0.19 backend [37]. The code is available at [38]. The CNN was trained and tested on a workstation with an Intel Core i7-7700HQ processor and 16 GB of RAM and all the processing was performed in a NVIDIA GeForce GTX 1050 GPU. The batch size is set to 32 and the training is executed while the performance of the classification of the test set improves. On average, each epoch took about 2 minutes to complete and less than 10 epochs are necessary to converge. For training the neural network we use the Adam optimizer [39], with a learning rate of $\alpha = 0.01$, an exponential decay rate for the first moment estimates of $\beta_1 = 0.9$, an exponential decay rate for the second moment estimates of $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ to prevent any division by zero. In order to evaluate the different models, we carried out a 5-fold cross-validation.

Figure 6 gives the confusion matrix of the system diagnosis performance on the 64x64 patches from the test set. In this confusion matrix, the number of patches correctly and incorrectly classified as belonging to the control class is in the first row. From a total of 8115 control patches in the test set, 7585 are true

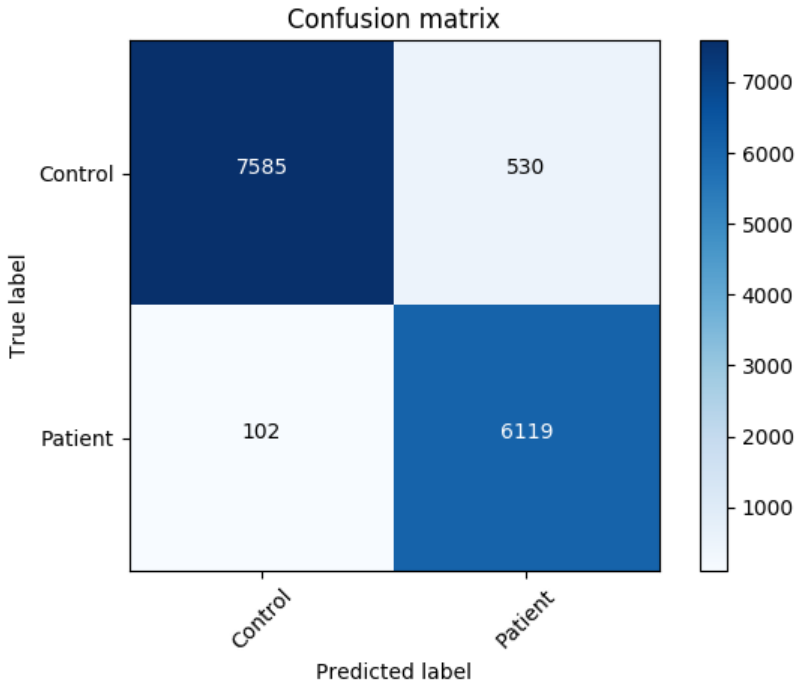

Figure 6: Confusion matrix of the test set for the model trained with the 64x64 image patches.

Table 2: Classification performance of the different approaches.

Method	Accuracy	Precision	Sensitivity	Specificity	F_1 score
RF (LBP)	0.6086	0.6022	0.5470	0.5521	0.6072
SVM (LBP)	0.6186	0.6560	0.5211	0.5041	0.6056
RF (FOS+GLCM)	0.6683	0.6619	0.6067	0.6232	0.6330
SVM (FOS+GLCM)	0.6783	0.6957	0.5736	0.6011	0.6287
CNN (Ours)	0.95	0.92	0.98	0.93	0.95

negatives (t_n), i.e., the inputs correctly classified as control, and 530 are false positives (f_p), i.e., the inputs incorrectly classified as patient. The classification of patients is given in the second row, where 102 are false negatives (f_n), i.e., the inputs incorrectly classified as control, and 6119 are true positives (t_p), i.e., the inputs correctly classified as patient. In the majority voting system, the accuracy is perfect and, thus, the confusion matrix is trivial and not given here.

Although there are no alternative approaches specifically designed for the problem tackled in this paper, for the sake of comparison we implemented four alternative approaches. Two of them rely on features obtained with the Local Binary Patterns (LBP) method, and the other two use the First-Order Statistics (FOS) and the Gray-Level Co-occurrence Matrix (GLCM) collagen-specific features described in [15]. These two types of features are used to train classifiers based on a Random Forest (RF) and a Support Vector Machine (SVM) with Radial Basis Function (RBF) kernels. We found the best hyperparameters for the RF and SVM models

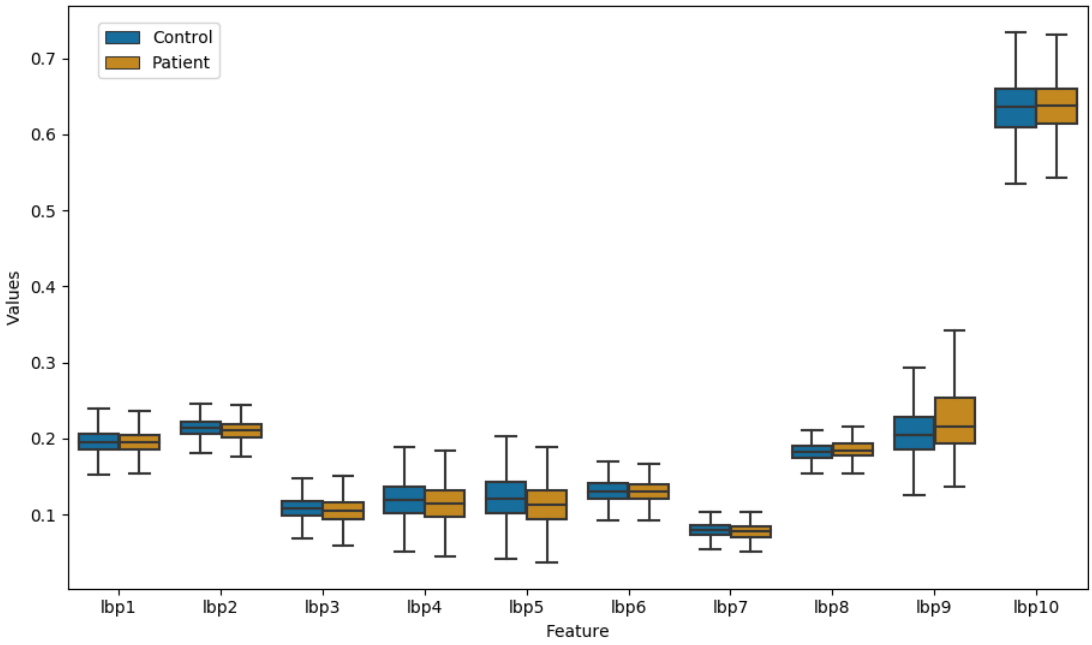

Figure 7: Distribution of the LBP features for control and patient classes.

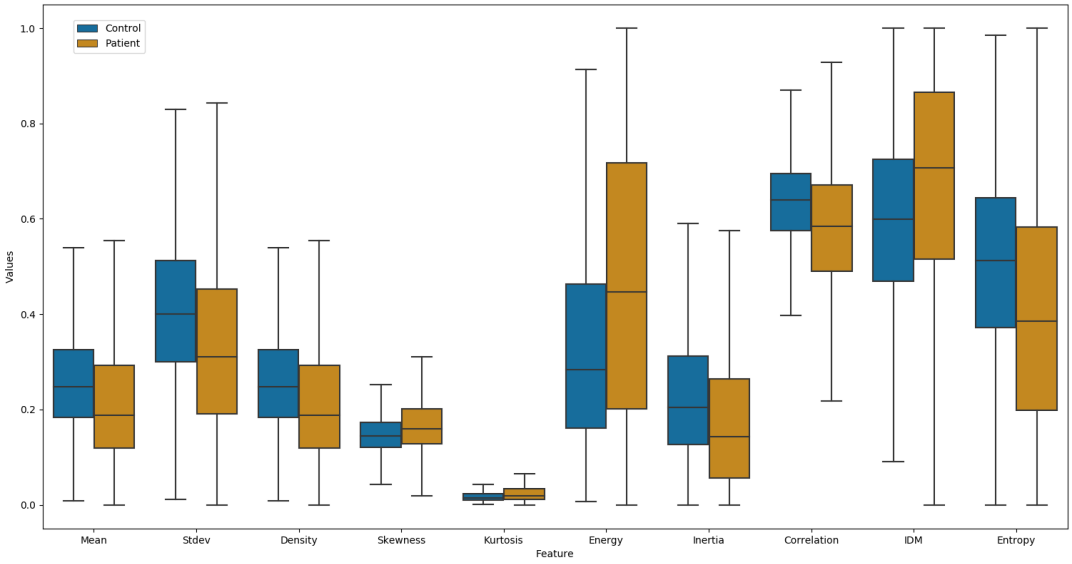

Figure 8: Distribution of the FOS and GLCM features for control and patient classes.

through Bayesian optimization [40] with 5-fold cross-validation. For the RF, we tuned two parameters: the number of decision trees, n_t, and the number of features in each decision tree, n_f. The optimal parameters found were $n_t = 350$ and $n_f = 7$ for the LBP features and $n_t = 400$ and $n_f = 8$ for the FOS and GLCM features. For the SVM, two main parameters were taken into account in the optimization process, namely the cost, C, and the free parameter of the RBF kernel, γ. The best parameters found were $C = 0.1$ and $\gamma = 0.01$, both for the LBP features and the FOS and GLCM features.

In Table 2, the accuracy, A, the precision, P, the sensitivity, S, the specificity, Sp, and the F_1 score are

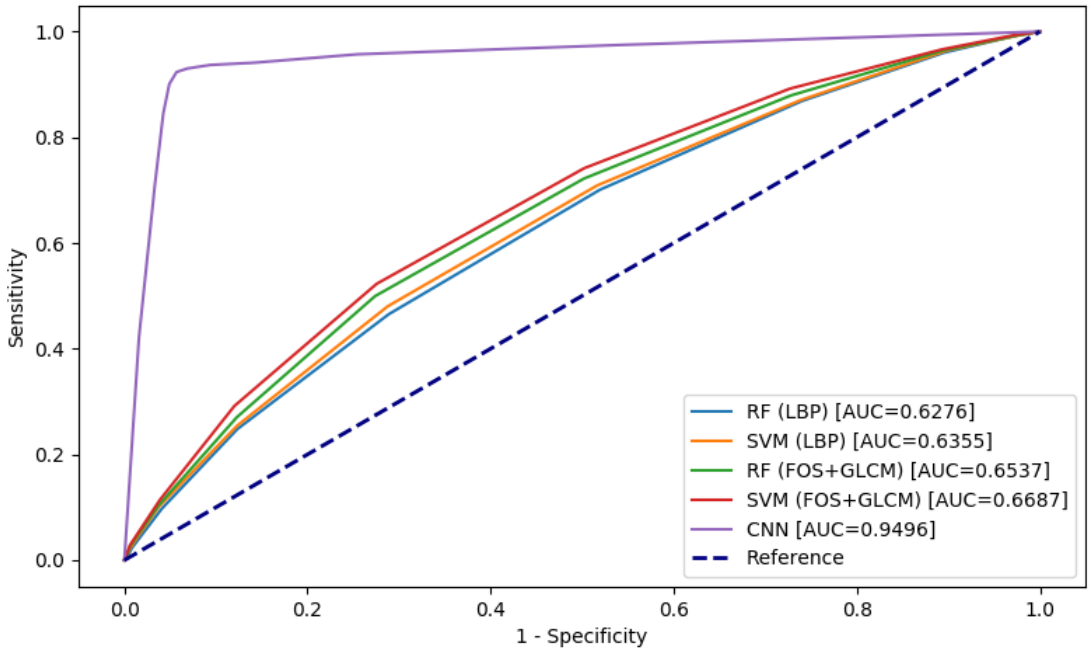

Figure 9: Receiver operating characteristic curve (ROC) for the classification methods compared in this paper. The performance of each method is given by the area under the curve (AUC).

used to assess the performance of the different approaches at the level of patch classification. The accuracy,

$$A = \frac{t_p + t_n}{t_p + f_p + t_n + f_n}, \tag{6}$$

refers to the correct classification rate, defined as the ratio of correctly classified cases respect to the total number of cases. The precision,

$$P = \frac{t_p}{t_p + f_p}, \tag{7}$$

is the ratio of inputs correctly classified as patient respect to the total number of inputs (correctly or incorrectly) classified a patient. The sensitivity,

$$S = \frac{t_p}{t_p + f_n}, \tag{8}$$

gives the ratio of samples correctly classified as patient respect to total inputs that are actually in the patient class. The specificity,

$$Sp = \frac{t_n}{t_n + f_p}, \tag{9}$$

gives the ratio of samples correctly classified as control respect to total inputs that are actually in the control class. Finally, the F_1 metric

$$F_1 = \frac{2\,P\,S}{P + S}, \tag{10}$$

is the weighted harmonic mean of precision and sensitivity.

From the results in Table 2, it is clear that the performance is rather independent of the used classifier since different classifiers get similar accuracy when applied on the same set of features. Classifiers using

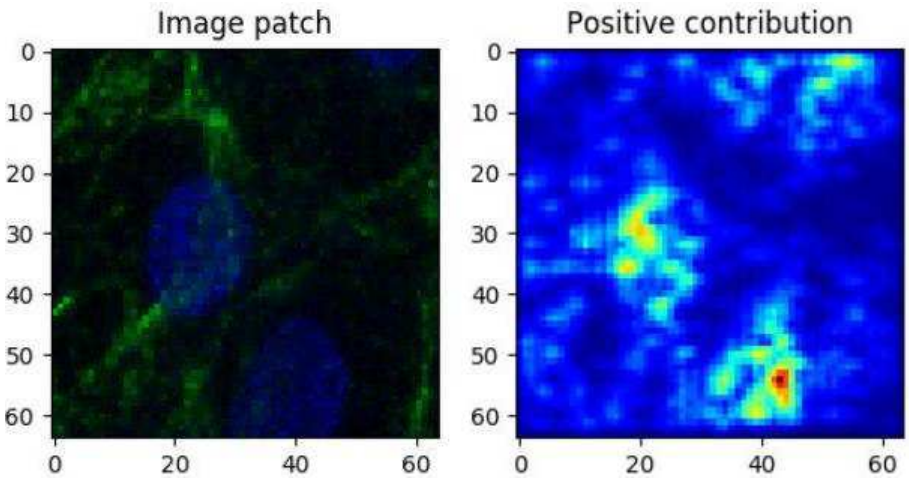

Figure 10: Left: Visualization of an image patch. Right: The saliency map with respect to the control class (in red the pixels with higher contribution to the classification).

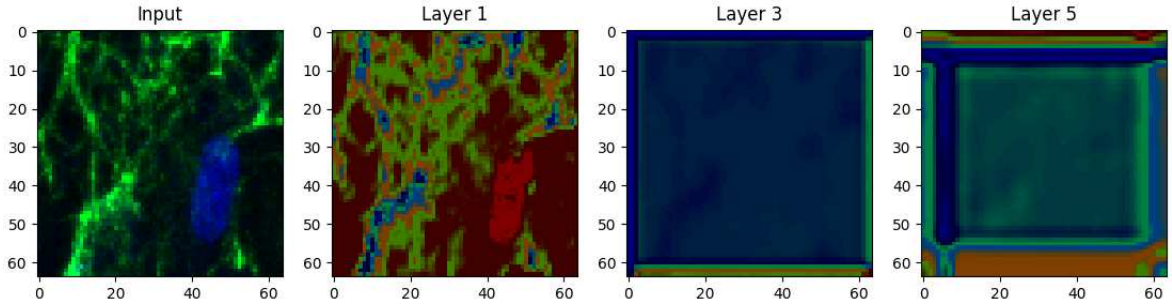

Figure 11: Visualization of the activation of each convolutional layer with respect to the patient class.

collagen-specific features outperform those relying on features obtained with LBP, which are not specific for the problem at hand. The poor discriminative power of the LBP features is explained in Fig. 7, where the differences of the feature distributions across classes are shown to be minimal. The distribution differences between classes for the collagen-specific features based on FOS and GLCM are shown in Fig. 8. In this case, the differences are larger, but not enough to facilitate a correct classification. In contrast, the approach proposed in this paper generates features that capture the latent patterns in the images and, consequently, it outperforms the alternative approaches (Fig. 9). It achieves an accuracy of 0.95, a precision of 0.92, a sensitivity of 0.98, a specificity of 0.93, and an F_1 score of 0.95. This outstanding performance at the level of patches is the basis of the perfect classification obtained with the majority voting scheme.

As just described, CNNs can provide very accurate results, but they are often seen as black boxes, since it is hard to find out what are the features and the classification criteria learned from the data. However, recent works try to interpret the CNNs after the training phase [41]. Following this approach, we provide a visualization of two relevant parts of the CNN. First, Fig. 10 shows the contribution of each pixel in the image to the classification as control class. The network focuses on the fibroblast nuclei and on the regions without collagen VI. Note that the intensity on the later is lower, but their extension is larger and, thus, its

actual contribution to the classification is also significant. In second term, Fig. 11 shows the saliency maps of the convolutional layers with respect to the patient class, where the saliency is the gradient of the output with respect to the corresponding convolutional layer. At early layers, the network learns to detect low-level features, such as the presence of fibroblasts or the lack of collagen VI, confirming our previous analysis. In posterior layers, the features become more abstract and, thus, more difficult to relate with elements in the input images.

5. Conclusions

This paper describes a system for the computer-aided diagnosis of muscular dystrophies caused by deficiencies in the structure of collagen VI. The proposed system relies on a deep convolutional neural network and on a data augmentation scheme to handle the problem of lack of data typical of rare diseases. The proposed system is capable of achieving perfect results in the diagnosis task on the available testing set. This provides a solid tool to automatically and accurately track the effectiveness of potential therapies for the recovery of collagen VI in patients. Furthermore, we visualized the contribution of each pixel to the classification as well as the features learned by the CNN in each convolutional layer through saliency maps. This identifies the primary features discovered by the CNN. Our current research endeavors focus on extending this analysis to clarify the abstract features and the classification criteria identified by the proposed CNN. Moreover, we are studying the potential application of image-based computer-aided diagnosis procedures to other rare diseases related to deficiencies in collagen VI or other extracellular matrix proteins.

Ethics Statement

This study was carried out in accordance with the recommendations of the Fundación Sant Joan de Déu Ethics Committee. Written informed consent was obtained from patients and/or their parents or guardians in accordance with the Declaration of Helsinki. The protocol was approved by the Fundación Sant Joan de Déu Ethics Committee.

Acknowledgements

Adrián Bazaga was supported by a JAE-intro scholarship granted by the Spanish Council of Scientific Research, and acknowledges funding from Innovate UK under grant number KTP011266. Josep M. Porta is supported by the Spanish Ministry of Economy and Competitiveness under project DPI2017-88282-P.

Cecilia Jiménez-Mallebrera and Carmen Badosa are funded by the Health Institute 'Carlos III' (ISCIII, Spain) and the European Regional Development Fund (ERDF/FEDER), 'A way of making Europe', grants references PI16/00579, PI19/00122 and CP09/00011, and Fundación Noelia.

Author Contributions

Adrián Bazaga – Conceptualization; Investigation; Formal analysis; Methodology; Software; Validation; Visualization; Writing - original draft; Writing - review & editing.

Mònica Roldán – Data curation; Resources; Writing - original draft; Writing - review & editing.

Carmen Badosa – Data curation; Writing - review & editing.

Cecilia Jiménez-Mallebrera – Data curation; Resources; Writing - original draft; Writing - review & editing.

Josep M. Porta – Conceptualization; Formal analysis; Methodology; Supervision; Funding acquisition; Writing - original draft; Writing - review & editing.

References

[1] A. Nadeau, M. Kinali, M. Main, C. Jimenez-Mallebrera, A. Aloysius, E. Clement, B. North, A. Y. Manzur, S. A. Robb, E. Mercuri, F. Muntoni, Natural history of Ullrich congenital muscular dystrophy, Neurology 73 (1) (2009) 25–31.
URL https://doi.org/10.1212/WNL.0b013e3181aae851

[2] S. R. Lamandé, J. F. Bateman, Collagen VI disorders: Insights on form and function in the extracellular matrix and beyond, Matrix Biology 71-72 (2018) 348–367.
URL https://doi.org/10.1016/j.matbio.2017.12.008

[3] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, A. Hamosh, OMIM.org: Online mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders, Nucleic Acids Research 43 (D1) (2014) D789–D798.
URL https://doi.org/10.1093/nar/gku1205

[4] C. Jimenez-Mallebrera, M. Maioli, J. Kim, S. Brown, L. Feng, A. Lampe, K. Bushby, D. Hicks, K. Flanigan, C. Bonnemann, C. Sewry, F. Muntoni, A comparative analysis of collagen vi production in muscle, skin and fibroblasts from 14 Ullrich congenital muscular dystrophy patients with dominant and recessive col6a mutations, Neuromuscular Disorders 16 (9) (2006) 571–582.
URL https://doi.org/10.1016/j.nmd.2006.07.015

[5] K. Anthony, V. Arechavala-Gomeza, L. E. Taylor, A. Vulin, Y. Kaminoh, S. Torelli, L. Feng, N. Janghra, G. Bonne, M. Beuvin, R. Barresi, M. Henderson, S. Laval, A. Lourbakos, G. Campion, V. Straub, T. Voit, C. A. Sewry, J. E. Morgan, K. M. Flanigan, F. Muntoni, Dystrophin quantification, Neurology 83 (22) (2014) 2062–2069.
URL https://doi.org/10.1212/WNL.0000000000001025

[6] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems,

Vol. 1, 2012, pp. 1097–1105.
URL http://dl.acm.org/citation.cfm?id=2999134.2999257

[7] A. M. Schilham, B. van Ginneken, M. Loog, A computer-aided diagnosis system for detection of lung nodules in chest radiographs with an evaluation on a public database, Medical Image Analysis 10 (2) (2006) 247–258.
URL https://doi.org/10.1016/j.media.2005.09.003

[8] N. Singh, S. Hamde, Tuberculosis detection using shape and texture features of chest X-rays, in: Innovations in Electronics and Communication Engineering, 2019, pp. 43–50.
URL https://doi.org/10.1007/978-981-13-3765-9_5

[9] M. López, J. Ramírez, J. Górriz, I. Álvarez, D. Salas-González, F. Segovia, R. Chaves, P. Padilla, M. Gómez-Río, Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease, Neurocomputing 74 (8) (2011) 1260 – 1271.
URL https://doi.org/10.1016/j.neucom.2010.06.025

[10] D. Blum, I. Liepelt-Scarfone, D. Berg, C. la Fougère, T. Gasser, M. Reimold, Improving statistical analysis in brain imaging: an alternative way to use principal component analysis, Journal of Nuclear Medicine 59 (supplement 1) (2018) 275.
URL http://jnm.snmjournals.org/content/59/supplement_1/275.abstract

[11] L. Nanni, A. Lumini, S. Brahnam, Local binary patterns variants as texture descriptors for medical image analysis, Artificial Intelligence in Medicine 49 (2) (2010) 117 – 125.
URL https://doi.org/10.1016/j.artmed.2010.02.006

[12] J. S. Athertya, G. S. Kumar, J. Govindaraj, Detection of modic changes in MR images of spine using local binary patterns, Biocybernetics and Biomedical Engineering 39 (1) (2019) 17 – 29.
URL https://doi.org/10.1016/j.bbe.2018.09.003

[13] N. Zulpe, V. Pawar, GLCM textural features for brain tumor classification, International Journal of Computer Science 9 (3) (2012) 354–359.
URL https://www.ijcsi.org/papers/IJCSI-9-3-3-354-359.pdf

[14] Z. Abbas, M. Rehman, S. Najam, S. M. Danish Rizvi, An efficient gray-level co-occurrence matrix (GLCM) based approach towards classification of skin lesion, in: Amity International Conference on Artificial Intelligence (AICAI), 2019, pp. 317–320.
URL https://doi.org/10.1109/AICAI.2019.8701374

[15] L. B. Mostaço-Guidolin, A. C.-T. Ko, B. X. Fei Wang, M. Hewko, G. Tian, A. Major, M. Shiomi, M. G. Sowa, Collagen morphology and texture analysis: from statistics to classification, Scientific Reports 3

(2013) 2190.
URL http://doi.org/10.1038/srep02190

[16] M. Sun, A. B. Bloom, M. H. Zaman, Rapid quantification of 3D collagen fiber alignment and fiber intersection correlations with high sensitivity, PLOS ONE 10 (7) (2015) e0131814.
URL https://doi.org/10.1371/journal.pone.0131814

[17] S. Doyle, M. Feldman, J. Tomaszewski, A. Madabhushi, A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies, IEEE Transactions on Biomedical Engineering 59 (5) (2012) 1205–1218.
URL https://doi.org/10.1109/TBME.2010.2053540

[18] I. El-Naqa, Yongyi Yang, M. N. Wernick, N. P. Galatsanos, R. M. Nishikawa, A support vector machine approach for detection of microcalcifications, IEEE Transactions on Medical Imaging 21 (12) (2002) 1552–1563.
URL https://doi.org/10.1109/TMI.2002.806569

[19] B. Srinivas, G. Sasibhushana Rao, Performance evaluation of fuzzy C means segmentation and support vector machine classification for MRI brain tumor, in: Soft Computing for Problem Solving, 2019, pp. 355–367.
URL https://doi.org/10.1007/978-981-13-1595-4_29

[20] A. Lebedev, E. Westman, G. V. Westen, M. Kramberger, A. Lundervold, D. Aarsland, H. Soininen, I. Kloszewska, P. Mecocci, M. Tsolaki, B. Vellas, S. Lovestone, A. Simmons, Random forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness, NeuroImage: Clinical 6 (2014) 115 – 125.
URL https://doi.org/10.1016/j.nicl.2014.08.023

[21] A. R. Chowdhury, T. Chatterjee, S. Banerjee, A random forest classifier-based approach in the detection of abnormalities in the retina, Medical & Biological Engineering & Computing 57 (1) (2019) 193–203.
URL https://doi.org/10.1007/s11517-018-1878-0

[22] Y. Wang, Y. Chen, N. Yang, L. Zheng, N. Dey, A. S. Ashour, V. Rajinikanth, J. M. R. Tavares, F. Shi, Classification of mice hepatic granuloma microscopic images based on a deep convolutional neural network, Applied Soft Computing 74 (2019) 40 – 50.
URL https://doi.org/10.1016/j.asoc.2018.10.006

[23] B. E. Olivas-Padilla, M. I. Chacon-Murguia, Classification of multiple motor imagery using deep convolutional neural networks and spatial filters, Applied Soft Computing 75 (2019) 461 – 472.
URL https://doi.org/10.1016/j.asoc.2018.11.031

[24] J. M.-T. Wu, M.-H. Tsai, Y. Z. Huang, S. H. Islam, M. M. Hassan, A. Alelaiwi, G. Fortino, Applying an ensemble convolutional neural network with Savitzky–Golay filter to construct a phonocardiogram prediction model, Applied Soft Computing 78 (2019) 29 – 40.
URL https://doi.org/10.1016/j.asoc.2019.01.019

[25] L. Tian, R. Yuan, An automatic end-to-end pipeline for CT image-based EGFR mutation status classification, in: SPIE Medical Imaging, 2019, p. 10949.
URL https://doi.org/10.1117/12.2512465

[26] S. Basheera, M. S. S. Ram, Classification of brain tumors using deep features extracted using CNN, Journal of Physics: Conference Series 1172 (2019) 012016.
URL http://doi.org/10.1088/1742-6596/1172/1/012016

[27] J. Bullock, C. Cuesta-Lázaro, A. Quera-Bofarull, Xnet: a convolutional neural network (CNN) implementation for medical X-ray image segmentation suitable for small datasets, in: SPIE Medical Imaging, 2019, p. 10953.
URL https://doi.org/10.1117/12.2512451

[28] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, S. Shetty, End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography, Nature Medicine 25 (2019) 954–961.
URL https://doi.org/10.1038/s41591-019-0447-x

[29] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60 – 88.
URL https://doi.org/10.1016/j.media.2017.07.005

[30] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.
URL https://doi.org/10.1109/5.726791

[31] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556. arXiv:1409.1556.
URL https://arxiv.org/abs/1409.1556

[32] F. A. Spanhol, L. S. Oliveira, C. Petitjean, L. Heutte, Breast cancer histopathological image classification using convolutional neural networks, in: International Joint Conference on Neural Networks (IJCNN), 2016, pp. 2560–2567.
URL https://doi.org/10.1109/IJCNN.2016.7727519

[33] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
URL https://www.deeplearningbook.org

[34] S. Hung, H. Adeli, Parallel backpropagation learning algorithms on CRAY Y-MP8/864 supercomputer, Neurocomputing 5 (6) (1993) 287–302.
URL https://doi.org/10.1016/0925-2312(93)90042-2

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958.
URL http://jmlr.org/papers/v15/srivastava14a.html

[36] F. Chollet, et al., Keras, https://keras.io (2015).

[37] M. Abadi, et al., TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org (2015).
URL http://tensorflow.org/

[38] A. Bazaga, Collagen VI-related muscular dystrophies diagnosis sofware, https://github.com/AdrianBZG/Muscular-Dystrophy-Diagnosis (2019).

[39] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980. arXiv:1412.6980.
URL http://arxiv.org/abs/1412.6980

[40] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, in: Advances in Neural Information Processing Systems 25, 2012, pp. 2951–2959.
URL https://dl.acm.org/citation.cfm?id=2999464

[41] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, CoRR abs/1312.6034. arXiv:1312.6034.
URL http://arxiv.org/abs/1312.6034