Transfer learning with pre-trained deep convolutional neural networks for the automatic assessment of liver steatosis in ultrasound images

Elena Codruța Constantinescu¹, Anca-Loredana Udriștoiu², Ștefan Cristinel Udriștoiu², Andreea Valentina Iacob², Lucian Gheorghe Gruionu³, Gabriel Gruionu³, Larisa Sândulescu¹, Adrian Săftoiu¹

¹Research Center of Gastroenterology and Hepatology, University of Medicine and Pharmacy Craiova, ²Faculty of Automation, Computers and Electronics, University of Craiova, Craiova, ³Faculty of Mechanics, University of Craiova, Craiova, Romania

Abstract

Aim: In this paper we proposed different architectures of convolutional neural network (CNN) to classify fatty liver disease in images using only pixels and diagnosis labels as input. We trained and validated our models using a dataset of 629 images consisting of 2 types of liver images, normal and liver steatosis. Material and methods: We assessed two pre-trained models of convolutional neural networks, Inception-v3 and VGG-16 using fine-tuning. Both models were pre-trained on ImageNet dataset to extract features from B-mode ultrasound liver images. The results obtained through these methods were compared for selecting the predictive model with the best performance metrics. We trained the two models using a dataset of 262 images of liver steatosis and 234 images of normal liver. We assessed the models using a dataset of 70 liver steatosis images and 63 normal liver images. Results. The proposed model that used Inception v3 obtained a 93.23% test accuracy with a sensitivity of 89.9% and a precision of 96.6%, and areas under each receiver operating characteristic curves (ROC AUC) of 0.93. The other proposed model that used VGG-16, obtained a 90.77% test accuracy with a sensitivity of 88.9% and a precision of 92.85%, and areas under each receiver operating characteristic curves (ROC AUC) of 0.91. Conclusion. The deep learning algorithms that we proposed to detect steatosis and classify the images in normal and fatty liver images, yields an excellent test performance of over 90%. However, future larger studies are required in order to establish how these algorithms can be implemented in a clinical setting.

Keywords: fatty liver disease; NAFLD; convolutional neural networks; deep learning

Introduction

With the rapid growth of the population and the increasing rates of obesity and diabetes, fatty liver disease (FLD) has become the most common liver disease in Western countries, non-alcoholic fatty liver disease (NAFLD) affecting around 25% of the global adult population [1]. Liver steatosis is defined as the accumulation of more than 5% of fat in the hepatocytes [2]. In the natural evolution of NAFLD, if no early strategies are adopted, the disease can progress to steatohepatitis (NASH) which is a risk factor for liver fibrosis, cirrhosis and even hepatocellular carcinoma [3]. In order to confirm the diagnosis, to quantify liver steatosis and stage NAFLD, liver biopsy is still considered the gold standard. However, it is an invasive procedure which carries a high risk of serious complications and, therefore, it cannot be routinely used for the evaluation and follow-up of NAFLD [4].
Non-invasive imaging methods used for the diagnosis of FLD include ultrasonography, computed tomography, magnetic resonance and magnetic resonance-based fat quantification techniques. Conventional B-mode ultrasonography is mostly preferred for screening and for the initial assessment of FLD due to the wide availability and low costs [5]. Although a normal ultrasound image cannot rule out liver steatosis, the overall sensitivity and specificity of ultrasound are comparable to those of histology, increasing with the degree of steatosis [5-7]. However, FLD can be diagnosed on ultrasound based on some suggestive parameters: diffuse hyperechoic structure (“bright liver”), deep beam attenuation, liver-to-kidney contrast, bright vessel walls [8,9]. Taking into consideration the public health burden given by NAFLD, there is a great need for the development of a non-invasive approach for the correct diagnosis and to determine the progression of the liver disease in order to establish an early treatment.

Over time, in order to improve the diagnostic accuracy of FLD and to make a quantitative assessment of steatosis, several computer-aided methods were investigated, including hepato-renal index (HRI), gray level co-occurrence matrix (GLCM), artificial intelligence (AI) methods such as traditional machine learning (ML) algorithms and deep learning (DL) algorithms. One of the most popular types of DL algorithms and widely used in medical imaging are convolutional neural networks (CNN). CNN consist of a series of layers (convolutional, pooling and fully connected) that can perform end-to-end supervised learning in order to identify patterns, lines or edges while some of the layers that are “hidden” can conduct unsupervised learning tasks. However, CNN architecture consists of many convolutional layers, which means that a great deal of data is needed for the training set in order to acquire competent accuracy [10]. To overcome this issue, data augmentation techniques have been proposed to reduce overfitting. In order to increase the efficiency of CNN, transfer learning has become very popular because it considerably reduces the training time and requires less data to train, while increasing the classification performance. Up to date, the research in clinical application of AI in liver imaging is small and limited, and more studies are required to demonstrate the efficiency of DL algorithms [11] but the importance of CNN in the detection of liver masses has been demonstrated in previous studies [12-14]. Moreover, several small studies used DL with B-mode ultrasound images for the diagnosis and classification of FLD [15,16].

The aim of our study was to analyse the test performance of recent DL algorithms for the ultrasound evaluation of liver steatosis.

Materials and methods

Pre-trained convolutional neural network implementation and parameter details

Very deep networks are prone to overfit and large convolution operations are computationally expensive. The Inception network, on the other hand, is complex and uses many improvements to obtain better performance, both in terms of speed and accuracy. VGG16, another CNN architecture which is considered to be an excellent convolution neural model, is a large network, comprising approximately 138 million parameters [18].

In this study, we used fine-tuning which is a type of transfer learning. We applied fine-tuning to the pre-trained DL models Inception v3 and VGG-16. The top layers of the pre-trained models are replaced by new fully connected layers with random parameters, while the layers below are kept frozen. The new fully connected layers can learn patterns from previous learned convolutional layers, because a very small learning rate is utilised. By applying fine-tuning, liver steatosis could be recognised even if our networks, Inception v3 and VGG-16 were not originally trained. This method achieved higher accuracy than feature extraction [17,18]. After rescaling the images and using image augmentation, we flowed them in batches of 16. In order to reduce overfitting and improve the model’s ability to generalize, we used dropout, activity regularization and kernel regularization (see: https://github.com/keras-team/keras).

The generic architecture of the proposed methods based on transfer learning with fine-tuning can be observed in figure 1.

Inception v3 model

We used Inception v3 model pre-trained on ImageNet dataset to extract high-level features from B-mode ultrasound liver images [17]. Our proposed model included all the layers in the Inception v3 network, except for the last fully connected layer, making all the layers non-trainable in order to reduce the overfitting. At the top of
the network we added two fully connected layers, one with 1024 hidden units and ReLU activation and the other was the last sigmoid dense layer for classification. Our optimizer was RMSprop with a learning rate of 0.0001 (see: https://github.com/keras-team/keras).

**VGG-16 model**

Another model that we used was the VGG-16 network pre-trained on ImageNet [18]. The model we used included all the layers in the network, except for the last fully connected layer. At the top of the network we added three fully connected layers with 512, 128, and 64 hidden units and ReLU activation. The last layer is the sigmoid dense layer used for classification. Our optimizer was Adam with a learning rate of 0.0001 [19].

**Patient selection and image dataset**

The dataset contains images from 60 patients, 30 patients with hepatic steatosis and 30 healthy subjects, both male and female, age ranging from 18 to 92. The participants were selected by an expert hepatologist from the University of Medicine and Pharmacy, Research Center of Gastroenterology and Hepatology, Craiova from the outpatient clinic of a private healthcare network. Exclusion criteria included excessive alcohol consumption or >20 g ethanol per day, history or clinical and/or laboratory evidence of liver disease, hepatotoxic medication use. Written informed consent was obtained from all patients. The study is non-interventional and all procedures are used in the current daily practice.

The diagnosis of NAFLD was based on the imaging studies (B-mode ultrasound and ARFI elastography) along with the clinical exclusion of the patients who consume > 20 g ethanol per day. The B-mode ultrasound images were collected by the same physician and all examinations were performed on Siemens Acuson NX3 ultrasound machine. The ultrasound diagnosis of steatosis was indicated by diffuse hypechoic structure, deep beam attenuation and increased liver echogenicity in relation to the right kidney. Participants were evaluated with pSWE, an alternative method for the non-invasive evaluation of liver fibrosis that uses acoustic radiation force impulse imaging (ARFI) to induce tissue dislocation. The examinations were performed by another physician on a Siemens Acuson S2000 machine. All the participants had ten valid measurements performed in the right liver lobe with IQR/M ≤30% and a median value ranging from to 1.8 to 3.7 kPa.

The database was created by cropping non-overlapping patches (fig 2) and by specific transformation such as resizing and shifting, obtaining 629 grayscale liver images with a resolution of 75x75. Out of the 629 images, 496 were used for training and 133 for testing. For the training dataset, we used 262 images with liver steatosis and 234 normal liver images. For the test dataset we used 70 liver steatosis images and 63 normal liver images.

**Statistical analysis**

The medical diagnosis metrics used to analyse the performance of our CNN models were: sensitivity, precision, test accuracy and area under curve (AUC).

The other metrics used to evaluate the classifier quality were the receiver operating characteristic (ROC) metric and Precision-Recall (PR). While precision measures the relevancy of the result, recall measures how many relevant results are returned.

**Results**

We used two pre-trained CNN models to detect the characteristic features in each liver image. Our results show a final diagnosis accuracy of 93.23% and the AUC of 0.93 for the Inception v3 model. By comparison, when using the VGG-16 model, the final diagnosis accuracy was 90.77% and the area under curve of 0.91. All metrics computed for Inception v3 vs. VGG-16 are summarized in Table I and showed the improvements achieved by using the Inception v3 model.

The ROC curves for the two models (Inception v3 vs. VGG-16) show the rate of false positive is near to zero while the rate of true positive is between 0.9 and 1 (fig 3a). The precision-recall curves of the two models (Inception v3 vs. VGG-16) demonstrate a better precision and recall of the Inception v3 model (fig 3b). High scores

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>AUC</th>
<th>Sensitivity</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception v3</td>
<td>93.23%</td>
<td>0.93</td>
<td>88.9%</td>
</tr>
<tr>
<td>VGG-16</td>
<td>90.77%</td>
<td>0.91</td>
<td>88.9%</td>
</tr>
</tbody>
</table>

Table I. Evaluation metrics for the classification of non-alcoholic fatty liver disease using the two pre-trained CNN models.
Elena Codruța Constantinescu et al  
Transfer learning with pre-trained deep convolutional neural networks

Transfer learning with pre-trained deep convolutional neural networks for both precision and recall, indicate that the classifier was returning accurate results, along with returning a majority of all positive results, respectively.

Discussion

By analysing the results, our study demonstrated that the use of DL methods with transfer learning and fine tuning had significant effects on the detection of liver steatosis by classifying images into normal and fatty liver cases with high accuracy and precision.

In the last 2 years, several studies addressed the potential of DL for the detection and evaluation of focal liver lesions, the detection of fibrosis associated with viral hepatitis, the detection of NAFLD, to assess chronic liver disease and to facilitate and predict treatment response [20-23].

Yasaka et al [12] proposed a method based on CNN for the differentiation of liver masses in CT, obtaining the median accuracy of 84% for test data and the AUC of 0.92. Ben-Cohen et al proposed a liver metastases detection method in CT [13]. This study reported a true positive rate of 94.6% with a 2.9 false positive result per case. Trivizakis et al developed methods based on 3-D CNN and 2-D CNN for cancer classification with application to MRI liver tumor [14]. The classification performance was 83% for 3-D CNN vs 69.6% and 65.2% for 2-D CNNs. Moreover, a study [15] classified ultrasound liver images of patients who were to undergo bariatric surgery using HI, GLCM and CNN algorithm, with liver biopsy as reference. The results showed that the Inception-Resnet-v2 pre-trained model was efficient and operator-independent, with higher accuracy (96.3%) than HI and GLCM. Recently, Santhosh and Rajalakshmi developed a DL method validating the usefulness of CNNs for accurate classification of fatty liver using ultrasound images. They obtained a 91.37% accuracy for detecting liver steatosis [16].

Also, our results support the robustness of CNNs with transfer learning to detect liver steatosis with high accuracy, precision and sensitivity even though we used a small image dataset from 60 patients. We compared the results obtained by the two powerful networks (Inception v3 and VGG-16), both with high accuracy of 93.23% and 90.77% respectively. As can be observed by the obtained results, the classification metrics of Inception v3 showed a slightly improvement, with a higher precision in classifying the two types of images.

Considering the high classification accuracy obtained, future potential advantages of an automatic diagnosis of FLD using AI algorithms could include (1) an accurate and rapid diagnosis for an initial assessment of a patient with risk factors, thus, making the screening and/or monitoring efficient; (2) could provide assistance for remote or isolated areas in the absence of any other diagnostic imaging resource as well as assisting sonographers in decision making.

On the other hand, the proposed methods have several limitations. Firstly, the dataset size is relatively small and in order to improve the accuracy of DL network and to validate the generality of the proposed methods, a larger sample size is needed. This was also confirmed in other studies that compared DL to other algorithms. Secondly, the performance of our diagnosis methods was not compared with experts. This will be part of a future study. Third, to have more accurate data, US diagnostics should be supported by liver biopsy as the gold standard, even though studies show that the specificity and sensitivity of ultrasound are similar to histology, 93.6% and 84.8% respectively [4]. Moreover, the values we obtained using the pre-trained CNN were calculated as means of images that were manually selected by a physician, and this could have had a significant influence on the results.
A possible direction for future studies would be to optimize the selection process for the region of interest from B-mode ultrasound images.

In conclusion, taking into consideration the global health burden of FLD, in particular NAFLD, there is a pressing need to have an accurate, fast and operator-independent ultrasound diagnosis. The DL algorithms that we proposed to detect steatosis and classify the images in normal and fatty liver images, yields an excellent test performance of over 90%. However, future larger studies are required in order to establish how these algorithms can be implemented in a clinical setting.

Acknowledgement: The research leading to these results received funding from the Competitiveness Operational Program 2014-2020 under the project P_37_357 “Improving the research and development capacity for imaging and advanced technology for minimal invasive medical procedures (iMTECH)” grant, Contract No. 65/08.09.2016, SMIS-Code: 103633. This work was also partially supported by the grant POCU 380/6/13/123990, co-financed by the European Social Fund within the Sectorial Operational Program Human Capital 2014-2020.

Conflict of interest: none

References