An Autoencoder Convolutional Neural Network Framework for Sarcopenia Detection Based on Multi-frame Ultrasound Image Slices

Emmanuel Pintelas, Ioannis E. Livieris, Nikolaos Barotsis, George Panayiotakis, Panagiotis Pintelas

An autoencoder convolutional neural network framework for Sarcopenia detection based on multi-frame ultrasound image slices

Emmanuel Pintelas[1], Ioannis E. Livieris[1], Nikolaos Barotsis[2], George Panayiotakis[2], and Panagiotis Pintelas[1]

[1] Department of Mathematics, University of Patras, Patras, Greece.
e.pintelas@upatras.gr, livieris@upatras.gr, ppintelas@gmail.com
[2] Department of Medicine, University of Patras, Patras, Greece.
nbarotsis@icloud.com, panayiot@upatras.gr

Abstract. Multi-Frame classification applications are constituted by instances composed by a package of image frames, such as videos, which frequently require very high computational re-sources. Furthermore, when the input instances contain a large proportion of noise, then the incorporation of noise filtering pre-processing techniques are considered essential. In this work, we propose an AutoEncoder Convolutional Neural Network model for Multi-Frame input applications. The AutoEncoder model aims to reduce the huge dimensional size of the initial instances, compress useful information while simultaneously remove the noise from each frame. Finally, a Convolutional Neural Network classification model is applied on the new transformed and compressed data instances. As a case study scenario for the proposed framework, we utilize Ultrasound images (image slices/frames extracted from every patient via a portable ultrasound device) for Sarcopenia detection. Based on our experimental re-sults the proposed framework outperforms traditional approaches.

Keywords: AutoEncoders · Deep learning · Transfer learning · Computer vision · Image classification · Sarcopenia detection · Ultrasound

1 Introduction

Multi-Frame (MF) input data mining applications require a significant amount of computational resources. A MF instance can be considered as a package of images arranged in time or space, such as videos or image slices extracted from 3D objects. One fast and light approach solution for solving such MF problems could be via the utilization of a CNN model applied only on one image slice/frame per MF instance. Nevertheless, by such an approach it is obvious that significant information from the initial MF instance is lost. Another common and efficient approach in terms of computation speed and accuracy lies on utilizing a CNN, which takes as input every image frame via a Multi-Channel (MC) input layer [19]. Nevertheless, when the initial dataset contains a large

ratio of noisy instances, the incorporation of pre-processing noise filtering techniques is considered essential [18].

In order to reduce the high dimensional input size of such applications and create an efficient prediction framework in terms of accuracy, robustness, and computation cost, the incorporation of dimensionality reductions algorithms becomes imperative. AutoEncoders (AE) constitute a popular dimensional reduction algorithm proved to filter out noise from an initial feature space; thus, create more robust final feature representations [18].

In this work, we propose an AutoEncoder Convolutional Neural Network framework (AE-CNN) for noisy MF input applications. More specifically, an AutoEncoder model will transform every input image frame into a new compressed image representation. Then, a concatenation layer will create a new composite image constituted by all the encoded image frames. As a result, the initial MF input instance is transformed into a compressed 2D flattened image. Finally, this flattened image is used as input into a common CNN prediction model. As a case study scenario, we utilize MF Ultrasound images (image slices/frames extracted from every patient via an ultrasound device) for Sarcopenia detection.

Ultrasound scanning for sarcopenia diagnosis is an emerging imaging tool which recently has attracted a lot of interest [1, 2, 8, 10, 12, 17]. Compared to traditional Computed Tomography (CT) scans and Magnetic Resonance Imaging (MRI), which are considered the gold standards in the detection of low muscle mass in sarcopenia, the low cost and the portability characteristic of ultrasound device are two of the greatest advantages of this diagnostic method. Full-body Dual Energy X-ray Absorptiometry (DXA) is widely used for the diagnosis of sarcopenia. However, DXA machines are not portable, limiting their use in the community, the measurements might be influenced by hydration status and the full-body scan exposes the subject to ionizing radiation. Contrarywise, ultrasound is a widely available, non-ionizing, low-cost and portable imaging modality, which could allow the large-scale screening of the population. A common disadvantage of MRI, CT and DXA is the training requirements for the staff, whereas the acquisition of ultra-sound images for muscle measurements can be done by minimally qualified staff and eventually be augmented by computer aided systems. Recently in [2], the authors developed a non-automatic sarcopenia detection system with 4 degrees of freedom to scan the human thigh with ultrasound probe and determine if he/she has sarcopenia by inspecting the length of muscle thickness in the thigh by ultrasound image.

Nevertheless, in the literature there is a great lack of research works, which actually managed to efficiently incorporate artificial intelligence methodologies, in order to make the ultrasound diagnosis for sarcopenia fully automatic. This is probably due to the fact that there is lack of medical data instances for efficiently training a DL model (such over of 100 instances) and because such images have a high proportion of noise comparing to CT scans, which is in general a more powerful but with a high cost diagnostic tool.

The main contributions of this work are summarized as follows: First, we managed to develop a fully automatic Ultrasound scanning detection system for

sarcopenia detection based on DL algorithms. Second, we propose the idea of combining an AutoEncoder and a CNN prediction model as a general tool for removing noise from instances and compressing the huge dimensional size of the initial MF input instances, without degrading the final prediction performance. As a result, the proposed prediction model will be an efficient way for addressing noisy MF applications which require also high computational resources.

The remainder of this paper is organized as follows: Section 2 describes the proposed framework and Section 3 presents the utilized dataset. Section 4 presents our experimental results and Section 5 summarizes our conclusive remarks.

2 Proposed framework

A high-level description of the proposed framework is presented in Figure 1. For every input case, the ultrasound device extracts high frequency spatial multiple image frames. In order to build a fast and computationally efficient total prediction framework, we subsampled the initial high frequency MF input and created a lower frequency MF input composed by 9 frames per instance. It is worth mentioning that a lower number of frames (1 and 4) was leading to performance degradation while a higher number $(16, 25, 36, \dots)$ was leading to a rapid increase in terms of computational costs.

More specifically, an AutoEncoder (AE) model is trained via the whole set of single-frame images. When AE training procedure finishes, the proposed framework is composed by three main phases as presented in Figure 1

In first phase, the trained AE model is applied on every instance's frame ($\mathbf{Fr_i}$) (nine frames in total per input case) in order to compress useful information and also remove the noise from each frame, creating $H \times W$ image representations $\mathbf{F_i}$ (where H and W is the height and the width size of every image). In second phase, the output images of the AE model are concatenated together via a Concatenation Layer (CL) building a 2D Flattened Image (FI) $3H \times 3W$. Finally, in the third phase, the flattened images are utilized for training a common CNN classification model.

Figure 2 presents three additional different approaches in which our proposed model (AE-FI-CNN) will be compared with and are described in brief as follows: By removing the AE from the AE-FI-CNN model, then the FI is created via the raw unprocessed ultrasound frames. This means that the CNN model will be trained with the raw FI (FI-CNN) as presented in Figure 2(a). Furthermore, by removing the AE and CL and replacing them with a Multi-Channel (MC) layer (MC-CNN), then the CNN model can be trained via the whole set of raw MF instances at once (each channel will take as input each frame) as presented in Figure 2(c). It is worth mentioning that the utilization of a MC-CNN model is the main baseline approach for solving such MF applications [19]. Finally, by interjecting the AE into the MC-CNN topology then the MC-CNN will be trained via the AE's pre-processed frames (AE-MC-CNN) as presented in Figure 2(b).

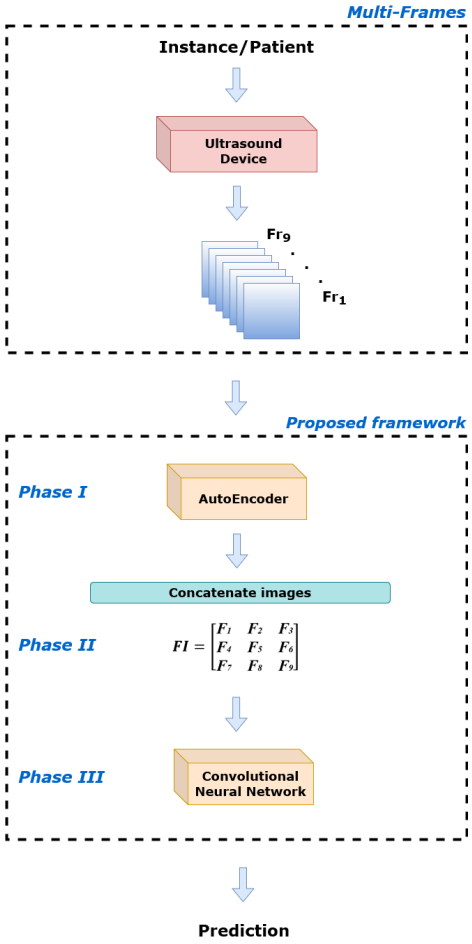


Fig. 1. Main pipeline of proposed Multi-Frame AutoEncoder Convolutional Neural Network model applied on the Ultrasound Sarcopenia detection problem.

2.1 AutoEncoders

AutoEncoders (AE) are unsupervised dimensionality reduction techniques constituted by artificial neural networks able to create a compressed knowledge representation of the original input. They can be used for reducing and compressing the dimension size of an initial feature space removing also noise and thus extracting and composing robust features [18].

More specifically, AE are composed by two main neural networks subcomponents, called *Encoder* and *Decoder*. The Encoder component is responsible for encoding an initial input feature space into a lower dimension one, while the Decoder component is responsible for reconstructing the compressed feature space

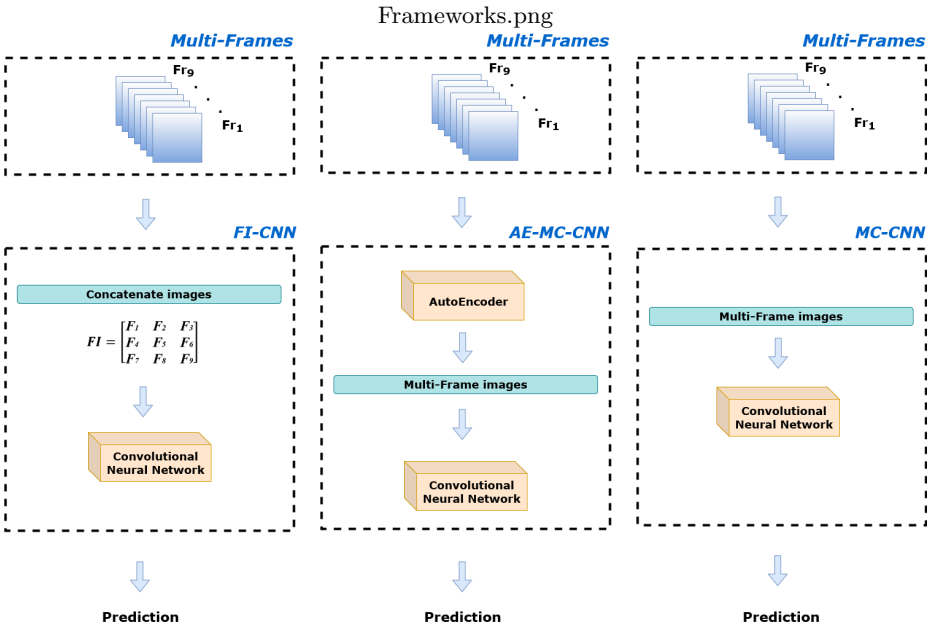


Fig. 2. Baseline approaches: (a) FI-CNN, (b) AE-MC-CNN and (c) MC-CNN multi-frame frameworks

into the initial one. Finally, the Encoder's output can be used as a robust and compressed feature representation for feeding a new classification model.

Figure 3 presents the architecture of the utilized AE topology of the proposed prediction framework. The Encoder component (left symmetric part of the whole topology) is constituted by three batches of blocks. The first block is composed by a 2D convolutional layer of 32 filters with 3×3 kernel size, a batch normalization layer, a ReLU activation layer, and a 2D max pooling layer with 2×2 kernel size. The second block is composed by a 2D convolutional layer of 8 filters with 3×3 kernel size, a batch normalization layer, a ReLU activation layer, and a 2D max pooling layer with 2×2 kernel size. The third block is composed by a 2D convolutional layer of 1 filter with 3×3 kernel size, a batch normalization layer and a Sigmoid activation layer. In the Decoder (right symmetric part of the topology), the main difference is that the max pooling layer is replaced by a 2×2 upsampling layer.

We recall that the output of the Encoder is the compressed image representation $\mathbf{F_i}$, which is used as input into the CNN classification model. The Decoder was only utilized during the initial AE training procedure. During inference mode, the Decoder is totally discarded since in our proposed framework we utilized only the Encoder component, in order to compress the initial multi-frame images and filter out the noise.

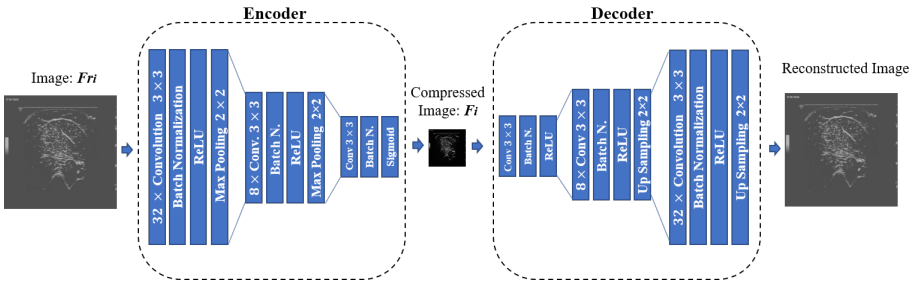


Fig. 3. The utilized AutoEncoder's architecture

2.2 Convolutional Neural Networks and Transfer Learning

Convolutional Neural Networks (CNNs) are a type of artificial neural networks mainly constituted by convolutional layers which are responsible for extracting and creating features from raw data matrices (such as images). They have proved to be very efficient feature extractors achieving remarkable performance especially in image classification problems [6, 9, 11].

Nowadays, the state of the art approach for solving automatic image recognition tasks is mainly based on utilizing pre-trained CNNs models [16]. These models can be used for inheriting their feature extraction knowledge into a small neural network without the need of training a new CNN model from scratch. This procedure is also known as Transfer Learning. Some widely used and efficient pre-trained CNN topologies are ResNet [5], Xception [4], and DenseNet [7].

ResNet (also called Residual Network), is a pretrained CNN model composed by residual blocks of 3×3 and 1×1 convolution filters and identity connections which transfer their input directly to the end of each residual block. This type of connections is well known for addressing the degradation problem, which is mainly caused by exceedingly large network depths.

Xception is another pretrained CNN model, which relies solely on depth-wise separable convolution layers. Xception architecture is based on the hypothesis that the mapping of cross-channels and spatial correlations in the feature maps of convolutional neural networks can been entirely decoupled [4].

DenseNet is a modified version of ResNet and is implemented by dense blocks connecting each layer to every other layer in a feedforward way. This leads to plenty of advantages such as parameter efficiency, feature reuse and implicit deep supervision. DenseNets have proved to be superior comparing to most state-of-the-art CNNs [7].

3 Dataset

The utilized dataset was taken from the Rehabilitation Department of Patras University Hospital. It has exams for Sarcopenia detection based on multi-

frame image slices taken for every patient/subject using the GE Logiq P9 ultrasound system equipped with the ML6-15 linear array transducer (GE Healthcare GmbH, Freiburg, Germany). More specifically, the dataset is composed by 1100 cases, from various muscle groups, of which 17% were detected with sarcopenia. The ultrasound images were acquired according to the protocol described by Barotsis et al. [3].

4 Experimental results

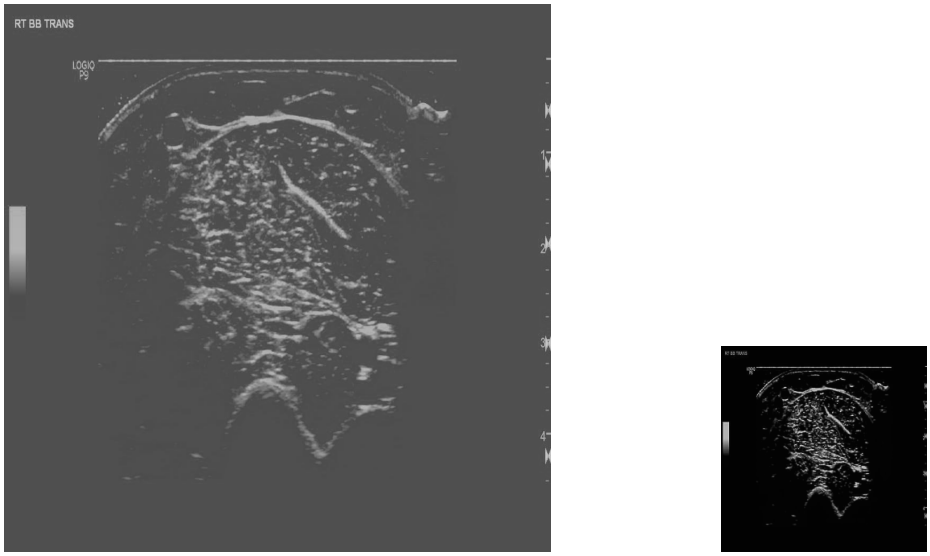
In this section, we present our experimental results regarding the proposed framework for sarcopenia detection based on ultrasound multi-frame image slices. In order to provide a higher visibility of the utilized shortcuts in this work, we summarized them in Table 1.

Model	Acronyms
Auto Encoder	AE
Flattened Image	FI
Multi-Frame	MF
Multi-Chanel	MC
Concatenation Layer	CL
Convolutional Neural Network	CNN

Table 1. List of acronyms and abbreviations for the utilized prediction models

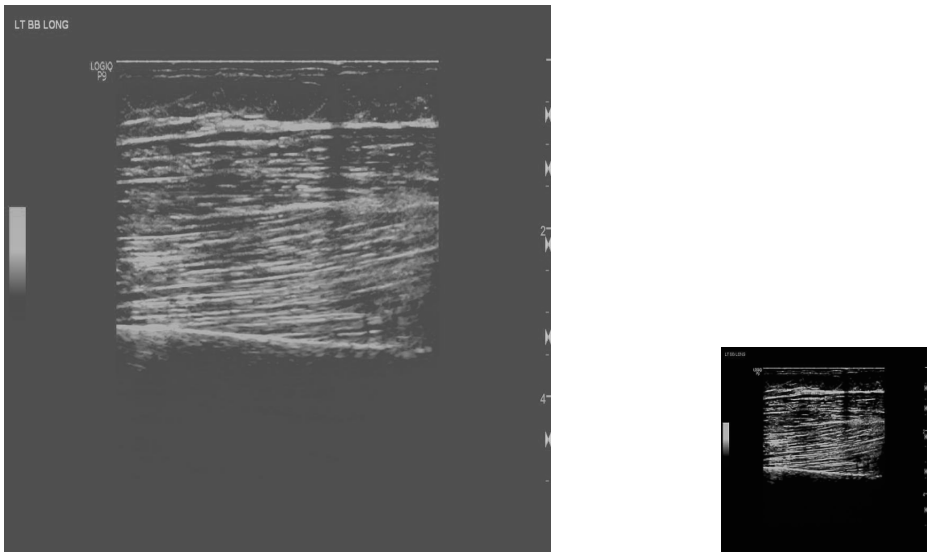
We utilized as CNN models a custom architecture, ResNet, Xception, and Dense-Net. Table 2 presents our experimental results for every utilized model. The Multi-Channel CNN (MC-CNN) models refer to the CNN models applied on the raw multi-frame inputs, the AutoEncoder Multi-Channel CNN (AE-MC-CNN) models refer to the CNN models applied on the transformed multi-frame inputs via the AE model, the FI-CNN models refer to the CNN models applied on the Flattened Images (FI), while finally the AE-FI-CNN (it is the proposed prediction frame-work) refer to the CNN models applied on the transformed flattened images via the AE model. Finally, Figures 4 and 5 present two examples with the original image with dimension 1024×1024 and the corresponding output of the encoder (encoded) with dimension 256×256.

The validation of our experimental simulations was based on the following performance metrics: Accuracy (Acc), Area Under the Curve (AUC) Geometric Mean (GM), F_1-score (F_1), Sensitivity (Sen), and Specificity (Spe) [15]. Additionally, we splitted the initial dataset into a 75% and 25% ratio for creating a training and a test set respectively maintaining also the same balance ratio of the corresponding classes.



(a) Original (b) Encoded

Fig. 4. Example of the application of encoder (a) original image with dimension 1024x1024 and (b) encoded image with dimension 256x256



(a) Original (b) Encoded

Fig. 5. Example of the application of encoder (a) original image with dimension 1024x1024 and (b) encoded image with dimension 256x256

Based on our experimental results, we are able to conclude that the proposed AE-FI-CNN framework significantly outperforms every other approach for every utilized CNN topology (Custom CNN, ResNet, Xception, and DenseNet). More specifically, the AE-FI-DenseNet model managed to achieve the higher overall performance.

It is also revealed that the incorporation of the AE model managed to drastically improve the overall performance of every CNN model for both approaches (MC and FI) especially for the Sensitivity score. However, it leaded to a slight performance degradation for the Specificity score. The Sensitivity score, for this specific application case study scenario measures the accuracy for the sarcopenia cases. At this point, it is essential to mention that since this case study is based on a medical application, the Sensitivity score is considered as the most important metric; this is because it is vital for the patients that the model will correctly classify the sarcopenia instances. On the other hand, the Specificity score, for this specific application case study scenario measures the accuracy for classifying correctly the healthy subjects, which is in general a much less significant matter. It is obvious that the misclassification of the healthy subjects is not as vital as the misclassification of the sarcopenia ones.

Model	Acc	AUC	GM	F_1	Sen	Spe
MC-CUSTOM CNN	83.71%	0.662	33.24	0.185	0.109	0.987
FI-CUSTOM CNN	84.85%	0.739	33.47	0.196	0.109	1.000
AE-MC-CUSTOM CNN	46.75%	0.252	28.86	0.089	0.152	0.531
AE-FI-CUSTOM CNN	87.09%	0.675	67.04	0.545	0.457	0.955
MC-RESNET	80.04%	0.639	32.48	0.156	0.109	0.942
FI-RESNET	77.03%	0.667	56.98	0.354	0.370	0.853
AE-MC-RESNET	80.04%	0.766	73.64	0.518	0.620	0.835
AE-FI-RESNET	81.16%	0.816	74.23	0.532	0.630	0.848
MC-XCEPTION	75.67%	0.546	49.83	0.283	0.283	0.853
FI-XCEPTION	65.20%	0.490	51.99	0.266	0.370	0.710
AE-MC-XCEPTION	79.67%	0.682	45.28	0.267	0.217	0.915
AE-FI-XCEPTION	84.40%	0.784	71.24	0.543	0.543	0.906
MC-DENSENET	83.03%	0.709	43.99	0.281	0.196	0.960
FI-DENSENET	77.47%	0.655	44.61	0.247	0.217	0.888
AE-MC-DENSENET	84.14%	0.612	29.87	0.157	0.087	0.996
AE-FI-DENSENET	90.70%	0.802	76.72	0.684	0.587	0.973

Table 2. Performance results of utilized models

At this point, it is worth mentioning that the performance metrics AUC and GM as well as the balance between Sen and Spe present the information provided by a confusion matrix in compact form; hence, they constitute the proper metrics to evaluate the ability of model of not overfitting the training data.

5 Conclusions and future work

In this work, we proposed an AutoEncoder Convolutional Neural Network framework (AE-CNN) for Sarcopenia detection based on multi-frame ultrasound image slices. An AutoEncoder model was used for transforming every initial input image frame into a new compressed image representation, filtering out the noise. Next, a new composite flattened image was constructed via the concatenation of every AE transformed image frame and finally it was fed into a CNN final classification model. The experimental results revealed the efficiency of the proposed framework since it managed to significantly outperform every other approach for every utilized CNN architecture and thus establish it as a very promising tool for automatically detecting Sarcopenia from Ultrasound scans. Such technique might allow a fast and large-scale screening of the population for the early detection in sarcopenia. Additionally, it might lead to a fast, easy and user-independent method for the follow-up of patients suffering from sarcopenia and the assessment of the efficacy of therapeutic interventions.

It is also revealed that the incorporation of the AE component for removing the noise and compressing every image frame managed to drastically increase the performance results of the final CNN model, comparing to the common approach which utilize the image frames in their raw form. Finally, the proposed FI construction from the multi-frame input instance led also to a performance increase comparing to the MC approach [19] (feeding of every image frame into a different CNNs input channel).

Nevertheless, one minor limitation of our approach is that the AE initial training requires a significant amount of computational resources. However, from our perspective, the main limitation is that the proposed framework is inherently Black Box model meaning that no explanation and interpretation of the model's final prediction mechanism can be given. Explainability and Interpretability are two exceedingly significant issues, especially in medical applications [13, 14].

In future work, we intent to include other information such as gender, age and also incorporate ensemble learning philosophy, such as averaging and stacking [20] into the proposed AutoEncoder CNN model in order to further improve the performance results. Finally, we also aim to incorporate transfer learning into an intrinsic interpretable machine learning model, such as a linear Logistic Regression model, in order to provide a significant degree of explainability [13, 14] into the whole proposed framework.

References

1. Albano, D., Messina, C., Vitale, J., Sconfienza, L.M.: Imaging of sarcopenia: old evidence and new insights. European radiology **30**(4), 2199–2208 (2020)
2. Barotsis, N., Galata, A., Hadjiconstanti, A., Panayiotakis, G.: The ultrasonographic measurement of muscle thickness in sarcopenia. A prediction study. European journal of physical and rehabilitation medicine (2020)
3. Barotsis, N., Tsiganos, P., Kokkalis, Z., Panayiotakis, G., Panagiotopoulos, E.: Reliability of muscle thickness measurements in ultrasonography. International Journal of Rehabilitation Research **43**(2), 123–128 (2020)

4. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Hemanth, D.J., Estrela, V.V.: Deep learning for image processing applications, vol. 31. IOS Press (2017)
7. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
8. Katakis, S., Barotsis, N., Kastaniotis, D., Theoharatos, C., Tsiganos, P., Economou, G., Panagiotopoulos, E., Fotopoulos, S., Panayiotakis, G.: Muscle type and gender recognition utilising high-level textural representation in musculoskeletal ultrasonography. Ultrasound in medicine & biology **45**(7), 1562–1573 (2019)
9. Kim, J., Nguyen, A.D., Lee, S.: Deep CNN-based blind image quality predictor. IEEE transactions on neural networks and learning systems **30**(1), 11–24 (2018)
10. Kim, Y.J., Kim, S., Choi, J.: Sarcopenia detection system using RGB-D camera and ultrasound probe: System development and preclinical in-vitro test. Sensors **20**(16), 4447 (2020)
11. Lu, L., Wang, X., Carneiro, G., Yang, L.: Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics. Springer (2019)
12. Mombiela, R.M., Vucetic, J., Rossi, F., Tagliafico, A.S.: Ultrasound biomarkers for sarcopenia: What can we tell so far? In: Seminars in Musculoskeletal Radiology. vol. 24, pp. 181–193. Thieme Medical Publishers (2020)
13. Pintelas, E., Liaskos, M., Livieris, I.E., Kotsiantis, S., Pintelas, P.: Explainable machine learning framework for image classification problems: Case study on glioma cancer prediction. Journal of Imaging **6**(6), 37 (2020)
14. Pintelas, E., Livieris, I.E., Pintelas, P.: A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability. Algorithms **13**(1), 17 (2020)
15. Raschka, S.: An overview of general performance metrics of binary classifier systems. arXiv preprint arXiv:1410.5330 (2014)
16. Shao, L., Zhu, F., Li, X.: Transfer learning for visual categorization: A survey. IEEE transactions on neural networks and learning systems **26**(5), 1019–1034 (2014)
17. Stringer, H.J., Wilson, D.: The role of ultrasound as a diagnostic tool for sarcopenia. The Journal of frailty & aging **7**(4), 258–261 (2018)
18. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. pp. 1096–1103 (2008)
19. Zhao, M., Chang, C.H., Xie, W., Xie, Z., Hu, J.: Cloud shape classification system based on multi-channel CNN and improved FDM. IEEE Access **8**, 44111–44124 (2020)
20. Zhou, Z.H.: Ensemble methods: foundations and algorithms. CRC press (2012)