IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 September 2023, accepted 31 October 2023, date of publication 8 November 2023,
date of current version 14 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3331315

RESEARCH ARTICLE

Handcrafted Features Can Boost Performance and Data-Efficiency for Deep Detection of Lung Nodules From CT Imaging

PANAGIOTIS GONIDAKIS[1,2], ALEXANDER SÓÑORA-MENGANA[3], BART JANSEN[1,2], AND JEF VANDEMEULEBROUCKE[1,2,4]

[1]Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), 1050 Brussel, Belgium
[2]IMEC, 3001 Leuven, Belgium
[3]Centro de Biofísica Médica, Universidad de Oriente, Santiago de Cuba 90500, Cuba
[4]University Hospital UZ Brussel, Vrije Universiteit Brussel (VUB), 1090 Brussel, Belgium

Corresponding author: Panagiotis Gonidakis (panagiotis.gonidakis@vub.be)

ABSTRACT Convolutional neural networks have been widely used to detect and classify various objects and structures in computer vision and medical imaging. Access to large sets of annotated data is commonly a prerequisite for achieving good performance. Before the deep learning era, systems based on handcrafted features were employed, which typically required less annotated data but also reached inferior performance. In this work, we investigate the benefit of combining deep learning using a convolutional neural network (CNN), with handcrafted features for lung nodule detection from CT imaging. We investigate three fusion strategies with increasing complexity, and evaluate their performance for varying amounts of training data. Our results indicate that combining handcrafted features with a 3D CNN approach significantly improves lung nodule detection performance in comparison to an independently trained CNN model, regardless of the fusion strategy. Comparatively larger increases in performance were obtained when less training data was available. The fusion strategy in which features are combined with a CNN using a single end-to-end training scheme performed best overall, allowing to reduce training data by 33% to 43%, while maintaining performance. Among the investigated handcrafted features, those that describe the relative position of the candidate with respect to the lung wall and mediastinum, were found to be of most benefit.

INDEX TERMS 3D CNN, convolutional neural networks, deep learning, data augmentation, false positive reduction, handcrafted features, lung cancer, pulmonary nodule detection.

I. INTRODUCTION

Deep learning models have shown tremendous progress in image classification, object detection, and segmentation, both in the domains of computer vision and medical imaging. Considering the case of image classification, convolutional neural networks (CNN) constitute the most widely used type of network. The initial convolutional layers are able to discover and extract discriminating feature maps, while the later fully connected layers of the architecture handle their classification and the output of the system. In general, a prerequisite for their success is large amounts of annotated data for training and tuning the deep learning models.

In various fields, including medical image analysis, data and in particular annotations are costly and time-consuming to come by. Reducing the amount of needed annotated samples for a deep learning model to reach a certain performance is an active area of research, which has led to various promising techniques currently receiving a high amount of attention in the domain, including but not limited to transfer learning, self-supervision, and weakly or unsupervised methods.

Non-deep learning machine learning methods often require comparatively less labelled data to reach their optimal

The associate editor coordinating the review of this manuscript and approving it for publication was Jinhua Sheng.

accuracy, but their performance is often below that of a deep learning system. In this case, so-called handcrafted features are carefully selected to accurately represent the available data, after which generic classifiers are employed to categorize these representations. While such approaches may require less annotated data, the achievable performance is usually inferior to that of CNNs, explaining the popularity of the latter approaches when sufficient data is available.

Combining handcrafted features and deep learning using CNNs for image classification can potentially improve performance, depending on the amount of available annotated data. In this work, we aim to investigate the benefit of such an approach for the case of automated lung nodule detection from CT imaging. The task has been extensively studied, achievable performance is well established, and a large public dataset is available. For this purpose, we first compare three approaches for combining handcrafted features and a CNN classification network. Next, we assess the influence of the amount of annotated data available for training on the classification performance. Finally, we perform an initial analysis on which handcrafted features have the largest benefit to classification performance.

A. LUNG NODULE DETECTION AND THE LUNA16 DATASET

Lung nodule detection from CT thoracic imaging is a challenging application in medical imaging. The task is of high clinical importance because pulmonary lung nodules may indicate early stages of lung cancer, and early detection is the primary lever to improve patient survival [1]. Automated approaches to lung nodule detection have been studied extensively in the field of medical image analysis [2], [3], benefiting from the availability of the high-quality, public LIDC-ICRI dataset [4].

In 2016, the LUNA16 Challenge [5] was held, enabling the scientific community to objectively compare approaches to detect and classify lung nodules. The challenge featured two tracks: one for full lung nodule detection systems, taking as an input thoracic CT imaging; and a second track for false positive reduction in which coordinate locations of candidate nodules were given. The associated dataset was based on 888 scans of the LIDC-ICRI dataset, and participants were asked to submit their prediction for the full dataset by performing 10-fold cross-validation over predefined folds.

At the time of the LUNA16 challenge workshop (April 2016), Dou et al. [6] presented the system that reached the highest performance for false positive reduction track in terms of the employed competition metric. Their approach comprised three networks that were independently trained with varying patch sizes, the outputs of which were fused by calculating the weighted average of the predictions. LUNA16 initially continued to accept challenge submissions, and numerous new studies further improved the classification score. In January 2018, submissions to the online challenge platform were no longer accepted by the organizers, as they suspected authors overfitted through excessive use of cross-validation, and newly reached performances were overly optimistic.

B. PREVIOUS WORK ON COMBINING HANDCRAFTED FEATURES WITH DEEP LEARNING

Our aim is to combine CNNs with handcrafted features extracted from the same imaging source. There have been numerous works in which CNNs have been combined with other features, in a range of application domains. In the medical field, such features often describe information obtained from other, non-imaging sources such as demographic data or blood analysis results. In general, they are referred to as tabular or structured data. Terminology for referring to different fusion approaches was found to be inconsistent across domains and authors.

To facilitate the discussion of previous work in the field, we chose to adopt the terms early, intermediate and late fusion, as used in a recent review on combining multi-modal data for precision healthcare [7]. Early fusion takes place at the data input level, mapping multiple sources to the same information space. Several authors have found it to be unsuitable for combining imaging data with structured data, as the preprocessing methods most suited for each type of data are different [8].

Late fusion merges the final predictions or decisions obtained for each single type of data, and can be seen as a form of ensemble learning. This type of fusion is most straightforward, and typically used when the features from both sources have been optimized separately. Combining corresponding models to make the final decision typically offers better performance over individual models.

Intermediate fusion, also referred to as joint fusion, covers a wide range of architectures in which fusion takes place at a feature level. It is generally a multistep approach, involving stages for feature extraction and selection from each source, and fusion models which combine the features to reach a final decision. Stages can be trained independently or jointly depending on whether the combined features capture different aspects of the image.

Considering previous work in the fields of video classification [9], human activity recognition [10], emotion recognition [11] and medical imaging [7], a universal optimal choice of fusion does not exist. The choice seems to be dependent on the type of data and the considered task.

Approaches for combining deep learning with handcrafted image features have been proposed in both computer vision and medical imaging. For the application of lung nodule detection, Li et al. [12] proposed a technique where a CNN model was combined with a set of handcrafted features using intermediate fusion. More specifically, the feature values in the output layer of the CNN were combined with 29 handcrafted features representing intensity, geometry and texture characteristics. A feature selection method, the sequential forward selection [13] method coupled with

an SVM, was employed to choose the final feature set and classification. Several CNN architectures were compared. The proposed fusion was found to outperform all individual CNN approaches. The authors hypothesized the fusion with handcrafted features reduced the need for large sets of training data, but this was not experimentally verified.

Considering other applications, notable approaches include the following. Wang et al. [14], presented a cascaded strategy consisting of a light CNN model and a system based on handcrafted features describing morphology, color, and texture. Two sub-systems individually performed classification and a probability output was calculated by taking a weighted average. For cases where the system was uncertain, a second-stage classifier was employed, based on the concatenation of the handcrafted and CNN-derived features. The final decision was made by thresholding. This approach was applied to detect mitosis in breast cancer pathology images and it was found to be sufficiently accurate and fast for clinical use. In this case, we observe that initially, a late fusion was employed, and followed by an intermediate one for specific cases.

Similarly, Hansley [15] employed a two-stage ear recognition framework, and handcrafted features improved a CNN-based ear descriptor. In this framework, a scoring system was implemented to combine the outputs of the two systems, which were trained individually. Late fusion schemes such as taking the sum, min, or max were investigated and eventually, the sum rule was found to work the best.

Kashif et al. [16] concluded that tumor cells could be more accurately detected when the input of a spatially constrained CNN (SC-CNN) is combined with RAW image data and handcrafted features. An early fusion methodology was followed where texture and color characteristics were combined with raw image intensities and forwarded as input to the SC-CNN. They observe an improvement compared to previously implemented CNN systems, which have only automatically learned features as input.

Nanni et al. [17] performed an extensive analysis by concatenating the output of a layer of the CNN with handcrafted features and then forwarding it to an SVM classifier. More specifically, firstly, a CNN-based model was trained. Afterwards, the last feature map before decision-making was extracted and concatenated with a vector containing handcrafted features. In the end, an SVM had to classify the resulting vector. Their proposed framework outperformed other state-of-the-art implementations after testing it on various datasets, including images from computer vision like paintings and pictures of smoke, but also medical and sub-cellular images.

Georgescu et al. [18] presented an intermediate fusion strategy to recognize facial expressions. CNN learned features were merged with the handcrafted ones and they were encoded by the bag-of-visual-words method. This way, the information was presented by (words) representing clusters of the initial data. The classification task was performed on the visual words using a local SVM performing better than a global one.

In previous work, we explored the benefit of fusing the prediction of a CNN, with architecture inspired by Dou et al. [6], with that of a system based on hand-crafted features [19]. Several classifiers were compared to optimally combine the predictions of the hand-crafted and CNN system in a late fusion scheme. The fused prediction was found to increase performance in all cases, with the random forest classifier leading to the best results.

The previously mentioned works demonstrate the potential of handcrafted features when combined with CNNs, despite authors employing varying fusion strategies. Research on the performance of different fusion strategies has not been reported. To the best of our knowledge, the influence of the amount of training data has not yet been investigated. In this work, we aim to explore fusion strategies that combine a CNN architecture with handcrafted features for varying amounts of training samples.

II. METHODOLOGY

We make use of LUNA16 dataset and we explore various late and intermediate fusion strategies to detect lung nodules on CT images by combining CNN-learned features with handcrafted ones. We aim to design the most effective strategy to perform this fusion, keeping in mind that high-performance rates on a limited amount of training annotated data are highly appreciated. At the same time, investigating which handcrafted features benefit more can be helpful to our design decisions.

As a starting point, we use two previously presented frameworks that classify suspicious CT areas as lung nodules to perform false positive reduction. We used a light 3D CNN model, previously proposed within our group, for which we investigated extensive data augmentation methods to boost its performance and obtain state-of-the-art performance [20]. Next, we adopted a lung nodule detection system based on hand-crafted features [19], shown to perform well at the LUNA16 Challenge [5].

For the experimental set-up, we chose to differentiate from the LUNA16 competition guidelines and avoid performing excessive 10-fold cross-validation. Instead, we split the LUNA16 dataset into training, validation, and testing sets. The training hyper-parameters are determined based on the performance obtained on the validation set and the results are reported by examining the fixed testing set.

We follow the second track of the LUNA16 Challenge. Therefore, having a long list of candidate locations, we aim to give a probability of being a nodule to each suspicious area. This list consists of 551,065 candidates, computed based on three existing candidate detection systems chosen by the competition of which 1,120 represent lung nodules.

We tackle this binary classification task by building various false positive reduction frameworks making use of 3D CNNs, handcrafted features, and combinations of them.

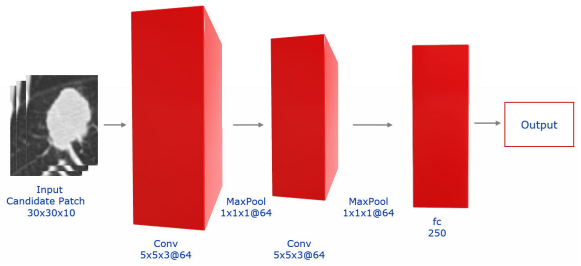

FIGURE 1. The proposed light 3D CNN consists of two convolutional and two max-pooling layers. A fully-connected layer realizes the decision-making.

These approaches are described in detail in the following sections.

A. FALSE POSITIVE REDUCTION USING A CNN

The employed architecture for CNN classification was inspired from the work of Dou et al. [6], where three individually trained networks were trained to classify a single candidate. Each of these networks was trained on a different 3D patch size, and their combination was found to reach high classification rates. As a downside, the approach involves a high computational effort, requiring three networks of which especially the one for large patch sizes is very computationally demanding.

We adapted and tuned the architecture with the aim to simplify the network and facilitate training while retaining its performance. To this end, we retained the best-performing middle patch size by fine-tuning the hyperparameters of the data augmentation and the network topology. A max-pooling layer was introduced after each convolutional layer whereas only one max-pool layer was present in the initial topology. In return, we employed one less convolutional layer.

The adopted architecture can be seen in figure 1. The input of this CNN is a 3D patch cropped from the original CT image. Its center is calculated by the previously mentioned candidate detector. To facilitate the convergence of the network, we clip the pixel intensities to the interval $(-1000, 400)$ Houndsfield Units (HU) and normalize them to the range of (0,1). The new network consists of 1,1 million trainable parameters, which is comparable but somewhat lower than other well-known 3D networks. This makes it easier and faster to be trained in various settings. This network topology simplicity can lead to lower performance but state-of-the-art results can be reached if a higher data augmentation rate is chosen, as previously demonstrated in previous work [20].

B. FALSE POSITIVE REDUCTION USING HANDCRAFTED FEATURES

The second approach to performing the false positive reduction is a handcrafted feature (HC) system. The employed set of features is inspired from the work Tan et al. [21]. This set of features has been carefully selected to be invariant to orthogonal transformations such as translations and rotations. To achieve this, some of them are computed in a 3D gauge coordinates system. Classical geometric features describing shape and location, and local grey-value and texture descriptors complete the list of the employed features.

TABLE 1. The set of 45 handcrafted features used by the HC system uses to discriminate the lung nodules from false positive areas extracted by the candidate detector.

	Feature	Notes
1	Volume	Number of voxels in the resampled image.
2	$\min_{\text{diam}}$	Largest major axis size of the bounding ellipsoid.
3	$\max_{\text{diam}}$	Shortest minor axis size of the bounding ellipsoid.
4	Compactness1	(*)Volume/ $\prod_{i=1}^{3}(\text{dim}_i)$.
5	Compactness2	Volume/$(\max_{\text{diam}})^3$.
6	Elongation factor	$\max_{\text{dim}}/\min_{\text{dim}}$.
7	3D ellipsoid	True for 3D bounding ellipsoid, False otherwise.
8	$D_{\text{centroid to lung wall}}$	Distance to lung wall.
9	$D_{\text{centroid to slice center}}$	Distance to the center of centroid's holding slice.
10-21	$E(L_{uu})$ and $E(L_{vv})$	(**) Averages calculated at scales $\sigma = 1$ and 2.
22-24	$E(z_{\text{nod}})$	(**) Average of z_{nod}.
25-27	$E(z_{\text{vessel}})$	(**) Average of z_{vessel}.
28-30	max(DNG)	(**) Average of maximum of DNG.
31-45	Grey-value features	(**)(***)Statistical measures

(*)$\prod_{i=1}^{3}(\text{dim}_i)$: Product of all the major axis of the bounding ellipsoid.
(**) Applied in combination with spherical kernels of radius 1 and 3 pixels.
(***) Mean, median, maximum, minimum and standard deviation.

Table 1 summarises the 45 handcrafted features, which can be divided into three categories: the geometric descriptors which characterize the shape and location of the candidate area(1-9), the gauge derivative invariant features (10-21) and the regional descriptors (22-45) [21]. Classification is obtained using a linear SVM classifier. The C parameter, which controls the misclassification rate, is set to 50.

Features are extracted over the segmented region of the potential nodule. The segmentation is performed by applying three different procedures involving filtering, thresholding and mathematical morphology operations to account for the different characteristics of isolated, juxtavascular and juxtapleural nodules. Three partially overlapping segmentations are obtained by applying the three procedures in parallel, and the final segmentation is found by merging all three using a logical OR operation. In case the segmentation does not lead to a segmented volume, all feature values are set to zero. We refer to the work of Tan et al. [21] for a more detailed description of employed lung nodule segmentation.

C. FALSE POSITIVE REDUCTION COMBINING LEARNED AND HANDCRAFTED FEATURES

The focus of this work is to fuse the 3D CNN network and the handcrafted feature system and evaluate its performance when trained with varying amounts of training data. We propose three different fusion strategies termed: S_1 - Prediction fusion, S_2 - Fusion of independent features, and S_3 - Joint training fusion. Their main difference is the point where the information is combined and the way training is performed.

S_1 is inspired by our previous work on late fusion [19], where preliminary results indicated a clear potential. S_2 and S_3 describe two intermediate fusion techniques, one with a multi-step training scheme and the other with one joint, end-to-end training scheme. Early fusion, in which the unstuctured image data is directly combined with

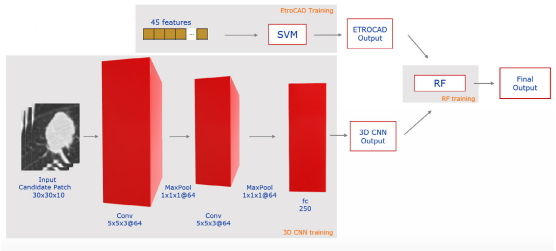

FIGURE 2. Fusion S_1: The prediction fusion. The handcrafted features are led to an SVM and the 3D image patch to a light 3D CNN. Then the two probabilities are fed to a random forest classifier.

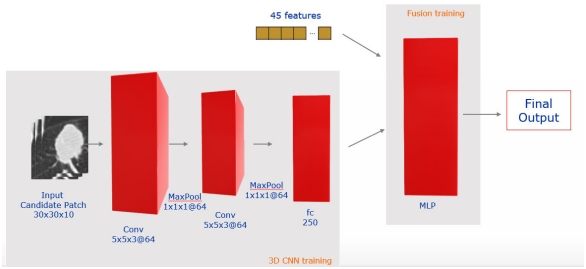

FIGURE 3. Fusion S_2: Fusion of independent features. A late image feature map is concatenated with the 45 handcrafted features and are fed to an MPL which consists of two fully connected layers.

structured hand-crafted features, was not explored in this study. Preliminary experiments on such a fusion strategy did not lead to improved performance, in line with results reported by others [8].

1) S_1: PREDICTION FUSION

The most straightforward approach to combine the CNN and the handcrafted features is a late fusion in which we train and run the systems independently and classify the candidate samples using the combinations of the final output predictions of both systems. We term this fusion strategy S_1, Prediction fusion.

On one hand, the HC system is employed making use of an SVM classifier as described before. In parallel, the 3D CNN is trained using patches of the same training set. The predicted outputs of these models are combined using a fusing classifier for which we adopted a random forest (RF) (Figure 4), which was found to perform best for this task among a number of tested classifiers [19]. The maximum number of trees is set to be 20 and the maximum depth 10.

We note that S_1 consists of three independent training stages: two for the independent systems and one for the prediction fusing classifier, i.e. the RF.

2) S_2: FUSION OF INDEPENDENT FEATURES

One could opt to fuse the two systems at an intermediate point, meaning at the (latent) feature level rather than combining the two final predictions. The first approach we investigated was to concatenate the handcrafted features of the HC system with the learned features obtained before the first fully connected layer of the CNN system. Such a strategy may benefit from additionally exploiting relations between the two subsets of features.

To this end, we design a multilayer perceptron (MLP) consisting of two fully connected layers to combine and fuse the two feature sets. The first fully connected layer has a ReLu activation function whereas the second one has a softmax (Figure 3).

The strategy denoted S_2 comprises only two training stages: one for the CNN system and one for the fusion classifier. The fusion takes place in an earlier stage than S_1 and the HC features are introduced to the system without any preprocessing.

3) S_3: SINGLE TRAINING FUSION

In the case of S_2, the features of the CNN are derived from the 3D image input without taking into account the presence of the 45 handcrafted features. An alternative intermediate fusion approach, denoted S_3 investigates the benefit of training the CNN while having access to the handcrafted features via a connected branch. This can potentially trigger the CNN to extract complementary information with respect to that provided by the handcrafted features and result in a unified network with improved performance.

This strategy was implemented as a single network comprising of two input branches. The first branch is an MLP designed to preprocess the 45 handcrafted features. This MLP consists of two fully connected layers with Relu activation functions introducing a non-linearity to the handcrafted features before the concatenation with the feature maps of the CNN. The second branch is a 3D CNN and accepts the volumetric data i.e. the input candidate 30 $\times$ 30x10 patches. This branch is identical to the previously employed 3D CNN model. The resulting features of the two branches are concatenated and further processed by one fully connected layer. The entire network is trained in an end-to-end fashion (Figure 4).

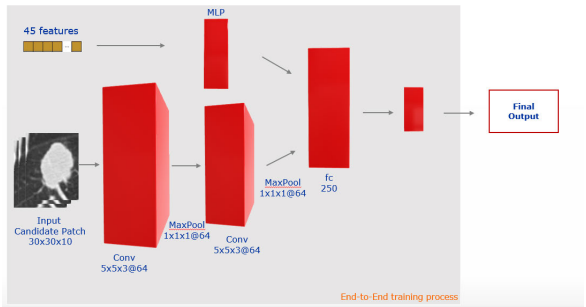

FIGURE 4. Fusion S_3: Single training fusion. The two modalities are fed to a single 3D CNN with two branches. The first branch receives the 45 handcrafted features and consists of two fully-connected layers and the second one receives the 3D image data.

D. FUSION WITH SELECTED FEATURES

In a final set of experiments, we explore which subset of the 45 handcrafted features (see Table 1) leads to the largest performance boost. The first set of experiments includes only the features 8 and 9, which correspond to the distance of the lung wall and the distance to the center of the centroid's holding slice. The second set includes all the geometric descriptors (features 1-9) and lastly, the third set makes use of all the remaining features (features 10-45). For each fusion strategy, considering each set of features described above, we evaluated the performance when using 40% of training data. For the best-performing fusion strategy, we further evaluated the performance for all amounts of training data.

The three new sets of experiments were designed based on the hypothesis that the position of the suspicious area regarding the closest lung wall and the centroid of the lung is predictive and plays an important role in classifying a candidate as a lung nodule or as a false positive sample. In addition, such information is not extracted by processing patches of the CT image using the proposed 3D CNN architecture.

E. THE TRAINING AND TESTING PROCESS

Our training and testing procedure was aimed at ensuring generalisability and reproducibility of the results. The LUNA16 challenge was conducted by performing 10-fold cross-validation and reporting the average performance over the folds. Hyper-parameter tuning through excessive cross-validation could lead to optimistic results. For this reason, we decided to deviate from the LUNA16 training scheme.

Instead, we split the LUNA16 set into a training, validation, and testing set. The training parameters like dropout, learning rate, and the number of training epochs were determined based on the performance obtained on the validation set, and for cases where the tuning procedure proved to be challenging, a subset of the training set was rotated with the validation set. The testing set was only used to report the final performance and never for any tuning processes.

The LUNA16 dataset consists of 10 randomly divided subsets. We reserved two specific subsets for testing (the 2nd and the 7th subset), judged to be sufficiently large and representative. Then for each trained model, we randomly select one subset to represent the validation set to tune any hyper-parameters. Finally, the training set consists of the remaining subsets and can have up to seven. These can not be used for testing or validation purposes.

Extensive data augmentation was employed to augment the amount of training samples, balance the highly imbalanced dataset and improve performance. We previously investigated the impact of data augmentation in detail [20] and adopted the settings which were found to perform best for this work. This included rotations of 90, 180 and 270 degrees, translations of one and two voxels in all dimensions, and flipping along all axes in 3D. Random combinations of these transformations were used to create 500 unique samples from each positive sample, i.e. a patch of the CT image containing a true lung nodule. Because of the high amount of negative samples found by the candidate detector, negative samples were not augmented.

The layer weights of the CNNs were randomly initialized using Xavier initialization, and an Adam optimizer was employed. Each model's training hyperparameters were optimized independently based on the performance of the validation set. To this end, we experimented with learning rates ranging from $10^{-}3$ to $10^{-}5$, and batch sizes ranging from 32 to 128, following powers of two. The dropout rate was fixed at 10%. The amount of training epochs was manually determined based on the observation of the training and validation loss curves.

We investigated the performance when trained with varying amounts of annotated data. From the full LUNA16 dataset containing 10 subsets, the maximum amount of subsets that was used for training was 7, corresponding to 70% of the full dataset. One subset was allocated for validation and two for testing. For each fusion strategy, seven different percentages of labelled data were considered (with incremental steps of 10%). Five models were trained with randomly initialized weights for each configuration and the amount of annotated training data was considered.

F. COMPUTATIONAL DETAILS

The described frameworks were implemented using Keras with the TensorFlow backend. As an indication of the model and its computational complexity, we list the amount of trainable parameters for each of the compared approaches (Table 3) and give indicative measures of the training times.

On an NVIDIA 2080 RX card, one training epoch of the 3D CNN with batch size 32 using 70% of training data takes 4 hours. In comparison, a 3D implementation of a VGG16 [22] requires 20 hours for the same settings. On top of this, we require two hours to fully train the RF classifier for S_1 fusion and 5 hours to train the MLP for S_2 fusion. Finally, one training epoch training the CNN for S_3 fusion takes approximately the same time as the 3D CNN.

The extraction of the handcrafted features is a computationally light process compared to training 3D CNNs. Extracting the 45 handcrafted features for the whole LUNA16 dataset takes two hours using a middle-range CPU, without any GPU-accelerated algorithms. Training an SVM on these hc features, like in the case of the S_1, requires the same amount of time.

G. VALIDATION METRICS

We adopted the Competition Performance Metric (CPM) used for the LUNA16 Challenge to evaluate the detection and classification rate. To calculate this, the Free-Response Receiver Operating Characteristic (FROC) curve is drawn. Each point of this curve is obtained by selecting those candidates whose probability of being a lung nodule is above

a threshold t, and we calculate the sensitivity and the average number of false positives per scan. The process is repeated for all t that produce a unique point on the FROC curve. The CPM is obtained by averaging the sensitivity at seven false positive rates(FPs): 1/8, 1/4, 1/2, 1, 2, 4, and 8 FPs per scan. Note that the CPM can be seen as a rough approximation of the AUC, in which more weight is given to low false positive rates. We report the average CPM over the five repeated experiments and its standard deviation.

Statistical analysis was performed using the non-parametric Freidman's test [23] since we have repeated measurements following non-normal distribution. This way we investigate if the results are significant and whether it is safe and meaningful to extract any conclusions.

III. RESULTS

The FROC curves for the independent systems and the three fusion strategies when trained with 40% of the data are shown in Figure 5. Table 3 summarizes their performances for all amounts of training data in terms of the CPM and sensitivity for a mean of 0.5 false positives per scan. The performance of the systems as a function of the amount of training data is visualized in Figure 6.

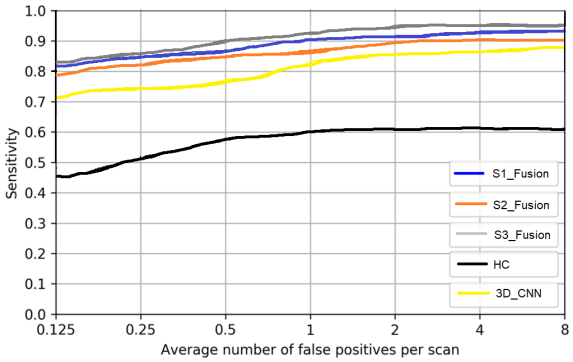

FIGURE 5. Free-response receiver operating characteristic (FROC) curves for models trained with 40% of the available labelled data.

A. PERFORMANCE OF THE INDEPENDENT SYSTEMS

The performance of the HC model is overall poor. Even when considering the full amount of training data (70% of the labelled data), this system would be impractical to use in a clinical environment since the number of false positives is high. Remarkably, its performance varies less than 5 percent points in terms of the CPM when lowering the training data from 50% to 10%.

As expected, the 3D CNN outperforms the HC system for most tested sizes of the training set. When given enough training data, i.e. 60% or more of the full dataset, it reaches more than 90% in terms of CPM. Its performance decreases rapidly with the amount of training data, losing 30 points in terms of CPM when lowering the training data from 50% to 10%. When using only 10% of the labelled data, 3D CNN scores lower than the HC model. At 60% of data used for

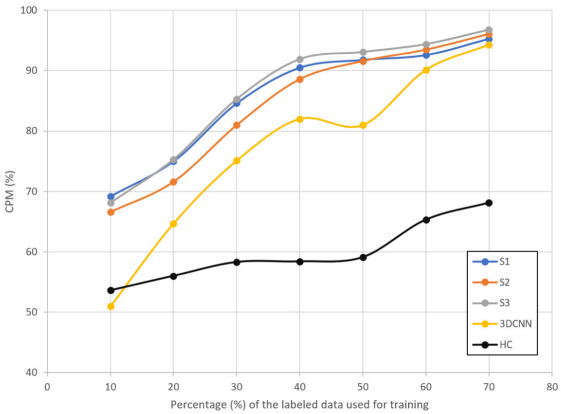

FIGURE 6. LUNA16 competition performance metric (CPM) as a function of the amount of training data employed, expressed as a percentage of total amount of labelled data, for the two independent systems, 3D CNN and HC and three fusion strategies.

training, the performance of the 3D CNN was found to be slightly lower than at 50%. We suspect this to be due to the randomly selected folds for training and validation being less representative of the test set.

For both models, we observed small differences in performance between the five repetitions, which can be seen from the variance reported in Table 3.

The performance of the trained models using the maximum amount of training data (70%) is compared to the state-of-art in Table 2, where we list the official results of LUNA16 challenge[1] along with our results. Note that direct comparison is not possible due to the use of different training and testing protocols. LUNA16 results were obtained after 10-fold cross-validation, thereby each time training on 90% of the data and testing on 10%. Our results were obtained by training on 70% of the data, tuning hyperparameters on 10% and testing on a fixed set of 20% of the data.

Despite these differences, it can be seen that the implemented 3D CNN approach leads to comparable performance with respect to the top-ranking frameworks of the competitive challenge. It would rank fourth on the LUNA16 challenge, achieving only 2.5 percentage points less in terms of CPM than the top-performing algorithm. The result indicates that the employed architecture, despite its simplicity, has sufficient complexity to achieve state-of-the-art results on this task when given sufficient training data.

B. PERFORMANCE OF THE COMBINED SYSTEMS

All fusion strategies were found to offer increased performance over the 3D CNN, at any of the amounts of training data tested, and the observed differences were found to be significant in all cases (p<0.025). The performance increase with respect to the 3D CNN was found to increase as the amount of training data is reduced, going from 1-2 points when using 70% of the data, to 10 points at 40% training data,

[1] https://luna16.grand-challenge.org/Results/

TABLE 2. Comparing our approaches with the official Top10 LUNA16 results for the classification track[1] (Track 2) updated until December 2017.
*** CPM: LUNA16 Competition Performance Metric.**
Results obtained by averaging those of a 10-fold cross-validation, each time training on 90% of labelled data and using 10% for testing according to the LUNA16 guidelines.
**** Average over five repetitions, training on 70% of labelled data, validation on 10% and testing on 20%, following the experimental set-up followed in this work.**

LUNA16 Rank	Framework	CPM*
1	PATech	96.8*
	Our S_3 fusion	96.8**
2	LUNA16FONOVACAD	96.6*
3	IHPC$_z$kj	96.5*
	Our S_2 fusion	96.1**
	Our S_1 fusion	95.3**
	Our 3D CNN	94.3**
4	MILAB$_C$oncatCAD	94.0*
5	JianPeiCAD	91.6*
6	qfpxfd	91.3*
7	GIVECAD	91.2*
8	CUMedVis [6]	90.8*
9	MILAB$_R$esCAD	88.9*
10	JianPeiCAD	88.9*
...	...	...
26	Nav	73.4*
	Our HC	68.1**

up to 15 percentage points in terms of CPM at 10% training data.

Performance differences between the fusion strategies were overall small, remaining within a few percent points in terms of CPM, and varied depending on the amount of training data used. When all available training samples were employed, single training fusion (S_3) performed best, equalling the best result reported at the LUNA16 challenge (Table 2), closely followed by feature fusion (S_2). As the training data reduced, the performance of all fusion models decreased in a comparable fashion. The S_2 model was most affected by the reduced training, performing worse when using 50% of the data or less for training. Simple prediction fusion (S_1) was found to be slightly less affected by the decrease in data, performing comparable or best for models trained on 30% of the data or less.

Variance between the five repetitions was low for all approaches but tended to increase as the training data was reduced.

C. FUSION OF SELECTED FEATURES

The performances in terms of CPM for the fusion models using a subset of handcrafted features are summarized in Figure 7 and Table 4. When comparing the different fusion methods trained with 40% of the data, S_3 performs best for all sets of selected features (Table 4). As expected, the best performance overall is obtained when using all handcrafted features. Fusion with geometric features 1-9 was found to lead to a larger increase in performance than fusion with non-geometric features (features 10-45). In fact, fusion with later features only leads to slight increases in performance over the 82% in terms of CPM obtained for the 3D CNN.

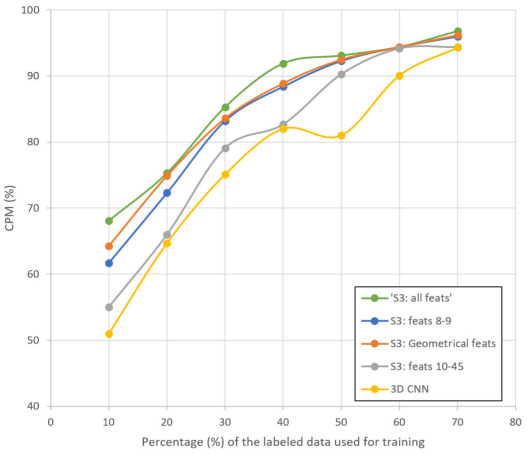

FIGURE 7. LUNA16 competition performance metric (CPM) as a function of the amount of training data employed, for models trained using S_3 fusion strategy. Each model was trained in combination with a different subset of handcrafted features to investigate which features contribute to the performance gain with respect the indepenedent 3D CNN system.

When considering the performance of S_3 fusion with the selected features, for the different amounts of training data (Figure 7), the same trend can be observed. Considering all features performs best, closely followed by fusion with all geometrical features. Non-geometric features perform worse than geometric, and the difference in terms of CPM was found to be significant for all amounts of training data (p<0.025), except when using 60% of the data for training ($p > 0.179$).

Considering only features 8 and 9, which describe the distance of the candidate to the lung wall and the slice center, respectively, leads to comparable increase in performance to using all geometric features. When using 50% or more, no significant differences in performance were found ($p > 0.179$). When using 40% of the data for training or less, difference in performance was found to be significantly different ($p < 0.025$) but remained limited to three percentage points. These results indicate that features describing candidate location capture the most predictive geometric information, and are the main drivers of the performance increase with respect to the 3D CNN.

We obtained comparable, low variances for the five repeatedly trained models for each case.

IV. DISCUSSION

The employed 3D CNN achieved good performance when trained on 70% of the full dataset, reaching 94.3% in terms of the LUNA16 competition metric. Direct comparison to the results obtained during the LUNA16 challenge is not possible because of the differences in training and testing data. That being said, the obtained results are comparable to those reported by Setio et al. [5] and reported on the challenge website (Table 2). Based on this, we consider 70% of the

TABLE 3. The average of the LUNA16 competition performance metric (CPM) along with the standard deviation and the sensitivity calculated for 0.5 mean false positives per scan (Se) for the two independent systems, 3D CNN and HC, and the three fusion strategies in regard to the amount of training annotated data employed.

Percentage of data used for training	3D CNN		HC		S_1 Fusion		S_2 Fusion		S_3 Fusion	
	CPM	Se	CPM	Se	CPM	Se	CPM	Se	CPM	Se
70	94.3 ± 0.05	91.7	68.1 ± 0.1	69.1	95.3 ± 0.3	92.1	96.1 ± 0.2	92.0	96.8 ± 0.2	93.1
60	90.1 ± 0.2	87.2	65.3 ± 0.2	65.6	92.6 ± 0.3	88.5	93.5 ± 0.1	89.4	94.4 ± 0.2	91.7
50	81.0 ± 0.3	78.1	59.1 ± 0.3	59.8	91.8 ± 0.2	87.4	91.6 ± 0.2	87.7	93.1 ± 0.1	90.5
40	82.0 ± 0.4	77.5	58.4 ± 0.3	57.4	90.5 ± 0.3	86.3	88.6 ± 0.4	84.3	91.9 ± 0.3	89.4
30	75.1 ± 0.3	72.3	58.3 ± 0.2	55.6	84.6 ± 0.5	81.7	81.0 ± 0.5	77.6	85.3 ± 0.4	83.3
20	64.0 ± 1.0	63.4	56.0 ± 0.5	56.4	75.0 ± 0.7	71.3	71.6 ± 0.6	67.7	75.3 ± 0.4	73.4
10	51.0 ± 2.0	50.2	53.6 ± 0.4	50.1	69.2 ± 0.6	64.2	66.6 ± 0.7	64.3	68.1 ± 0.6	66.5
Amount of training parameters	1.1M		46 (SVM)		1.2M + 46 (SVM)		1.5M		1.7M	

TABLE 4. CPM (LUNA16 Competition Performance Metric) for each fusion method when using 40% of training data, and considering a different subset of features are included. In comparison, the 3D CNN without fusion achieves 82% when using the same amount of training data.

Selected features	S_1	S_2	S_3
All 45 features	90.5	88.6	91.9
Features [8, 9]	85.3	84.9	88.4
Geometrical features [1-9]	85.9	85.2	88.9
Non-geometrical features [10-45]	82.5	82.4	82.7

training data sufficient to reach close to optimal results with the used 3D CNN approach.

When fusing handcrafted features with the 3D CNN approach, regardless of the fusion strategy, we observed a performance increase, even when considering all data available for training. The results are inline with literature [15], [16], [17], [18]. Two equivalent viewpoints can be adopted to explain these outcomes. Handcrafted features carry complementary, predictive information to the learned features of the CNN. Alternatively, one could state that the combination of the CNN and HC systems allows to reduce the (stochastic and deterministic) errors made by the individual systems.

When considering less training data, fusing the CNN approach with handcrafted features leads to a comparatively higher increase in performance. We hypothesize that for lower amounts of training data, the representations learned by the deep learning approach are suboptimal and generalize poorly. The HC system, characterized by significantly lower complexity, also performs worse when considering less training data, but is less affected. The combination of both, regardless of the fusion strategy, offers an important benefit when lower amounts of training data are available.

This benefit becomes substantial when considering the results obtained for S_3 when using 40% of the available data (CPM of 91.9%), which are competitive to the state-of-the-art and comparable to that of the 3D CNN using 60% to 70% of the data. Put differently, for the studied case of lung nodule detection, enhancing a 3D CNN by intermediate fusion with handcrafted features in a joint training scheme (S_3), allows for a reduction of 33% to 43% of the amount of training data, while maintaining performance compared to a regular CNN approach. This result encourages further research on this approach for applications in which low amount of annotated training data is available.

Overall, the differences in performance of the investigated fusion strategies were small. The observed behaviour for varying amount of training data does align with intuition when considering the complexity of the approaches in terms of the amount of parameters (Table 3). Single training fusion (S_3), the most complex approach, and employing a single training scheme while taking into account the handcrafted features, performs best when sufficient training data is available. Prediction fusion (S_1), the simplest of the studied strategies, performs comparably or better when few training data is available. For the studied case, feature fusion (S_2), in which learned and handcrafted features are combined by a separately trained fusion model did not show a benefit over S_1 or S_3, for any of the data availability settings.

When further investigating which features led to the largest benefit when combined with a 3D CNN, geometric features were found to be of most value. In fact, including non-geometrical features did not lead to notable performance increases for any of the fusion strategies. Geometric features, on the other hand, did improve results with respect to the 3D CNN, and the improvement was considerably larger for S_3. We hypothesize that the joint training scheme is more suited to exploit the complementary information given by geometric features.

In particular, the features describing the distance of the nodule candidate to the lung wall and the center of the slice proved to be of importance. Such location information can not be reliably extracted using the patch-based processing of the employed 3D CNN architecture, potentially explaining the improved performance when combining these features with the CNN.

The location of nodule candidates has previously been described as a relevant predictor in medical literature. McWilliams et al. [24] reported a higher probability of lung

nodules in upper lobes with respect to lower lobes. Our results are inline with those reported by Song et al. [25], who included a feature describing the location in terms of closeness to vessels, pleural wall or isolated in the parenchyma. More recently, inclusion of anatomical location in a deep segmentation framework was found to improve performance when applied to ascites in the pelvic region. Our results indicate that the value of location encoding in deep learning seems promising and merits further research for lung nodule detection but also other tasks in medical imaging.

V. CONCLUSION

In this work, we investigated the benefit of combining a CNN-based classification network with handcrafted features for the task of lung nodule detection. Different fusion strategies were investigated and their performance was evaluated when considering varying amounts of training data. Combining handcrafted features with a 3D CNN approach was found to improve detection performance, regardless of the fusion approach. Comparatively larger increases in performance were obtained when less training data was available. Among the investigated handcrafted features, those that describe the relative position of the candidate with respect to lung wall and mediastinum, were found to be of most benefit.

REFERENCES

[1] G. P. LeMense, E. A. Waller, C. Campbell, and T. Bowen, "Development and outcomes of a comprehensive multidisciplinary incidental lung nodule and lung cancer screening program," *BMC Pulmonary Med.*, vol. 20, no. 1, p. 115, Apr. 2020.

[2] P. Monkam, S. Qi, H. Ma, W. Gao, Y. Yao, and W. Qian, "Detection and classification of pulmonary nodules using convolutional neural networks: A survey," *IEEE Access*, vol. 7, pp. 78075–78091, 2019.

[3] S. Dodia and P. A. Mahesh, "Recent advancements in deep learning based lung cancer detection: A systematic review," *Eng. Appl. Artif. Intell.*, vol. 116, Nov. 2022, Art. no. 105490.

[4] S. G. Armato et al., "The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans," *Med. Phys.*, vol. 38, no. 2, pp. 915–931, Feb. 2011.

[5] A. A. A. Setio et al., "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," *Med. Image Anal.*, vol. 42, pp. 1–13, Dec. 2017.

[6] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1558–1567, Jul. 2017.

[7] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, and Y. Luo, "Multimodal machine learning in precision health: A scoping review," *NPJ Digit. Med.*, vol. 5, no. 1, p. 171, Nov. 2022.

[8] C. Cui, H. Yang, Y. Wang, S. Zhao, Z. Asad, L. A. Coburn, K. T. Wilson, B. A. Landman, and Y. Huo, "Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: A review," *Prog. Biomed. Eng.*, vol. 5, Apr. 2023, Art. no. 022001.

[9] Z. Wang, K. Kuan, M. Ravaut, G. Manek, S. Song, Y. Fang, S. Kim, N. Chen, L. F. D'Haro, L. A. Tuan, H. Zhu, Z. Zeng, N. M. Cheung, G. Piliouras, J. Lin, and V. Chandrasekhar, "Truly multi-modal YouTube-8M video classification with video, audio, and text," 2017, *arXiv:1706.05461*.

[10] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, "Glimpse clouds: Human activity recognition from unstructured feature points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 469–478.

[11] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "AVEC 2017: Real-life depression, and affect recognition workshop and challenge," in *Proc. 7th Annu. Workshop Audio/Visual Emotion Challenge*, Oct. 2017, pp. 3–9.

[12] S. Li, P. Xu, B. Li, L. Chen, Z. Zhou, H. Hao, Y. Duan, M. Folkert, J. Ma, S. Huang, S. Jiang, and J. Wang, "Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features," *Phys. Med. Biol.*, vol. 64, no. 17, Sep. 2019, Art. no. 175012.

[13] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, Dec. 1997.

[14] H. Wang, A. C. Roa, A. N. Basavanhally, H. L. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, and A. Madabhushi, "Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features," *J. Med. Imag.*, vol. 1, no. 3, pp. 1–8, 2014.

[15] E. E. Hansley, M. P. Segundo, and S. Sarkar, "Employing fusion of learned and handcrafted features for unconstrained ear recognition," *IET Biometrics*, vol. 7, no. 3, pp. 215–223, May 2018.

[16] M. N. Kashif, S. E. A. Raza, K. Sirinukunwattana, M. Arif, and N. Rajpoot, "Handcrafted features with convolutional neural networks for detection of tumor cells in histology images," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 1029–1032.

[17] L. Nanni, S. Ghidoni, and S. Brahnam, "Handcrafted vs. non-handcrafted features for computer vision classification," *Pattern Recognit.*, vol. 71, pp. 158–172, Nov. 2017.

[18] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019.

[19] A. Sóñora-Mengan, P. Gonidakis, B. Jansen, J. Garcia-Naranjo, and J. Vandemeulebroucke, "Evaluating several ways to combine handcrafted features-based system with a deep learning system using the LUNA16 challenge framework," *Proc. SPIE*, H. K. H. Maciej and A. Mazurowski, Eds. vol. 11314, 2020, pp. 900–906.

[20] P. Gonidakis, B. Jansen, and J. Vandemeulebroucke, "Artificially augmenting data or adding more samples? A study on a 3D CNN for lung nodule classification," in *Proc. SPIE*, H. K. H. Maciej and A. Mazurowski, Eds. vol. 11314, 2020, pp. 565–570.

[21] M. Tan, R. Deklerck, B. Jansen, M. Bister, and J. Cornelis, "A novel computer-aided lung nodule detection system for CT images," *Med. Phys.*, vol. 38, no. 10, p. 5630, 2011.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[23] M. Friedman, "A comparison of alternative tests of significance for the problem of M rankings," *Ann. Math. Statist.*, vol. 11, no. 1, pp. 86–92, 1940.

[24] A. McWilliams, "Probability of cancer in pulmonary nodules detected on first screening CT," *New England J. Med.*, vol. 369, no. 10, pp. 910–919, 2013.

[25] Y. Song, W. Cai, Y. Wang, and D. D. Feng, "Location classification of lung nodules with optimized graph construction," in *Proc. 9th IEEE Int. Symp. Biomed. Imag. (ISBI)*, May 2012, pp. 1439–1442.

PANAGIOTIS GONIDAKIS received the B.S. degree in physics from the Department of Physics, University of Patras, Greece, in 2012, and the M.S. degree in electronics and communications from the University of Patras, in 2015. He is currently pursuing the Ph.D. degree with the Department of Electronics and Information (ETRO), Vrije Universiteit Brussel (VUB), Belgium.

From 2015 to 2017, he was a Researcher with ETRO, VUB. His current research interest includes machine learning algorithms mainly in the domain of medical imaging.

ALEXANDER SÓÑORA-MENGANA received the B.S. degree in electronic and telecommunications engineering from Universidad de Oriente (UO), Santiago de Cuba, Cuba, in 1998, the M.S. degree in biomedical engineering from Universidad Central Martha Abreu de Las Villas, Villa Clara, in 2002, the B.S. degree in computer sciences from UO, in 2009, and the Ph.D. degree in technical sciences/engineering sciences from UO/Vrije Universiteit Brussel, in 2021.

From 2000 to 2016, he was a Researcher with Centro de Biofsica Mdica (CBM), UO. He has been working on hardware design for NMR equipment and biomedical signal acquisition and processing. Since 2016, he has been the Head of the Department of Bioengineering, CBM. Also, he is an Assistant Professor with FITIB, UO. His current research interests include machine learning technology and embedded solutions for medical applications.

BART JANSEN received the M.Sc. degree in computer science and the master's degree in AI from Vrije Universiteit Brussel (VUB), in 2001 and 2003, respectively, and the Ph.D. degree in computer science from the AI-Laboratory (promotor Luc Steels), VUB, in 2005.

Since 2006, he has been with the Department of Electronics and Informatics (ETRO), VUB. In 2016, he was a Professor and has been a Professor (10% appointment) with IMEC, since 2020. He is currently a Professor with ETRO, VUB. His current interests include developing image and signal processing methods and artificial intelligence methods for a variety of applications in the broad biomedical engineering domain, but mainly focusing on rehabilitation engineering.

JEF VANDEMEULEBROUCKE received the master's degree in electronic engineering from the University of Ghent, Belgium, and the Ph.D. degree in numerical optimization techniques from the Federal University of Santa Catarina (UFSC), Florianpolis, Brazil, in 2010, with a focus on lung motion estimation and modeling for image-guided radiation therapy, performed in collaboration with the Creatis Laboratory, University Lyon 1, France, and the Center for Machine Perception, Czech Technical University, Prague, Czech Republic. He specialized in artificial intelligence in Granada, Spain, for one year. He is currently an Associate Professor of medical image analysis with the Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Belgium. He is also an Affiliated Researcher with IMEC, an international research and innovation hub in nanoelectronics and digital technologies. His current research interests include medical image analysis for applications in computer-aided diagnosis and image-guided interventions, with a particular focus on thoracic, whole-body and dynamic imaging for oncology, and musculoskeletal pathologies.