Diagnostic and Interventional Imaging (2019) **100**, 251—257

ORIGINAL ARTICLE /*Research and new developments*

Detecting abnormal thyroid cartilages on CT using deep learning

M. Santin[a], C. Brama[a], H. Théro[a], E. Ketheeswaran[a], I. El-Karoui[a], F. Bidault[b], R. Gillet[c], P. Gondim Teixeira[c], A. Blum[c,*]

[a] *Kernix Software, 6, rue Lalande, 75014 Paris, France*
[b] *Radiology Department, Gustave Roussy, IR4M, CNRS, Université Paris Sud-Paris Saclay, 94805 Villejuif, France*
[c] *Service d'imagerie Guilloz, avenue de Lattre de Tassigny, CHRU de Nancy, 54000 Nancy, France*

KEYWORDS
Thyroid cartilage;
Artificial intelligence (AI);
Deep learning;
Post-mortem computed tomography (CT);
Larynx

Summary
Purpose: The purpose of this study was to evaluate the performance of a deep learning algorithm in detecting abnormalities of thyroid cartilage from computed tomography (CT) examination.
Materials and methods: A database of 515 harmonized thyroid CT examinations was used, of which information regarding cartilage abnormality was provided for 326. The process consisted of determining image abnormality and, from these preprocessed images, finding the best learning algorithm to appropriately characterize thyroid cartilage as normal or abnormal. CT images were cropped to be centered around the cartilage in order to focus on the relevant area. New images were generated from the originals by applying simple transformations in order to augment the database. Characterizations of cartilage abnormalities were made using transfer learning, by using the architecture of a pre-trained neural network called VGG16 and adapting the final layers to a binary classification problem.
Results: The best algorithm yielded an area under the receiving operator characteristic curve (AUC) of 0.72 on a sample of 82 thyroid test images. The sensitivity and specificity of the abnormality detection were 83% and 64% at the best threshold, respectively. Applying the model on another independent sample of 189 new thyroid images resulted in an AUC of 0.70.
Conclusion: This study demonstrates the feasibility of using a deep learning-based abnormality detection system to evaluate thyroid cartilage from CT examinations. However, although promising results, the model is not yet able to match an expert's diagnosis.

* Corresponding author.
E-mail address: alain.blum@gmail.com (A. Blum).

https://doi.org/10.1016/j.diii.2019.01.008
2211-5684/

The detection of thyroid cartilage invasion is important for choosing the appropriate care for laryngeal and hypopharyngeal cancers. In early stage lesions with no cartilage invasion, preserving the larynx is a primary goal. In disease with focal cartilage invasion, function-preserving partial laryngectomy or chemoradiotherapy has been introduced. An advanced stage lesion with apparent cartilage invasion may require total laryngectomy. The presence of neoplastic cartilage invasion may also affect the response to radiation therapy, leading to a higher rate of tumor recurrence [1—4]. Another application for thyroid cartilage determination is in forensic imaging. In cases of hanging, the type of thyroid cartilage fracture depends on the completeness of body suspension and ligature knot location [5—7].

The best technique to evaluate thyroid cartilage is still a matter of debate but computed tomography (CT) is considered the most effective imaging tool [3,8]. CT demonstrates erosions and sclerotic changes indicative of tumor invasion. However, due to its complex three-dimensional shape and its numerous normal variants, thyroid cartilage requires a long and thorough analysis by experienced imaging specialists [9—13].

Radiology has been marked in recent years by a rise in artificial intelligence (AI), a problem-solving approach involving the use of highly complex methods to replicate specific cognitive functions based on computer technologies and algorithmic techniques [14,15]. AI using deep learning techniques, such as convolutional neural networks, has received extensive attention after demonstrations that it could perform at least as well as humans in imaging-classification tasks [16,17]. AI might also provide a comprehensive and clear analysis of thyroid cartilage, which could help radiologists evaluating it.

The purpose of this study was to evaluate the performance of a deep learning algorithm in detecting abnormalities of thyroid cartilage from CT examination.

Materials and methods

Context

Data analysis took place during the Data Challenge organized during the 2018 JFR (''Journées Françaises de Radiologie''). Various subjects were proposed, but the authors opted to participate in the challenge using the detection of abnormal thyroid cartilages from CT data. The model presented here had the highest performance of all participants in predicting thyroid cartilage abnormality on an independent testing dataset. This paper addresses the deep learning algorithm methodology used in order to achieve classification of thyroid cartilages.

Database characteristics

The database contained 515 images of CT examinations of the thyroid, as well as a CSV file containing a label associated with each image (normal or abnormal cartilage). Gathered from 35 different radiology services across France, one horizontal cut was available for each patient; and all the examinations were harmonized as much as possible. During the entire challenge, only 326 of the 515 images were available to the various teams. The 189 remaining images have never been shared with the participants and were used by the organizers to independently evaluate the models developed. Among the 326 images, 30 contained abnormal thyroid cartilages (9.2%). Image files were in NIfTI format, had been resampled to have 1 mm pixels, and were recentered to all be the same size. The data were anonymized such that no clinical information regarding the population was available. The precise characterization of the origin of the lesions was thus impossible on CT examinations; as mentioned above, abnormalities arise from tumoral invasion, inflammation of autoimmune diseases, or fractures due to trauma.

Image processing

To focus on the most important parts of the images and feed relevant data to the algorithm, the images were centered around the cartilages using fixed coordinates (all the images were centered similarly). Three hundred twenty-six images were given as an initial database. They were divided into training and test sets containing 244 and 82 images, respectively, and transformed to JPEG format. The training set contained 24 abnormal cartilages, and the test set contained 6 to respect the initial proportions. Thus, the algorithm learned on a sample of the data; and its performance could be tested on images it had never seen, reducing possible overfitting. The model could then be evaluated before being submitted for final results. Since 244 is a very low number of images to train a deep learning algorithm, they were used to generate new ones and improve the performance and robustness of the model. Six new images were generated for every image in the training sample by applying various transformations on the original images, including rotations and translations (Fig. 1). These could then be fed to the algorithm to trick it into believing it is seeing new data, making it more robust in handling new types of images. Our final training sample contained 1,708 images. During the final evaluation, the various teams' models were then evaluated on 189 new images.

Image classification

The detection of abnormality in the thyroid cartilage was achieved using transfer learning from a pre-trained convolutional neural network (CNN) algorithm devoted to image classification. Since those algorithms are extremely expensive to train from scratch in terms of computational power and size of the dataset, transfer learning enables expansion of the algorithm for specific use on a small dataset. It starts with a CNN model using pre-trained weights and retraining it with the CT scans of thyroids and the appropriate classification (normal/abnormal).

The original model used in this work was the Visual Geometry Group 16 (VGG16), developed at Oxford University in 2015, which has been trained and evaluated on the ImageNet collection and composed of everyday-object pictures [18,19]. This model converts an input image into the probability of belonging to a class of the ImageNet categories. An image is scanned through several filters (convolutional layers), which detect the essence of each category, in order to extract the relevant features for the final classification.

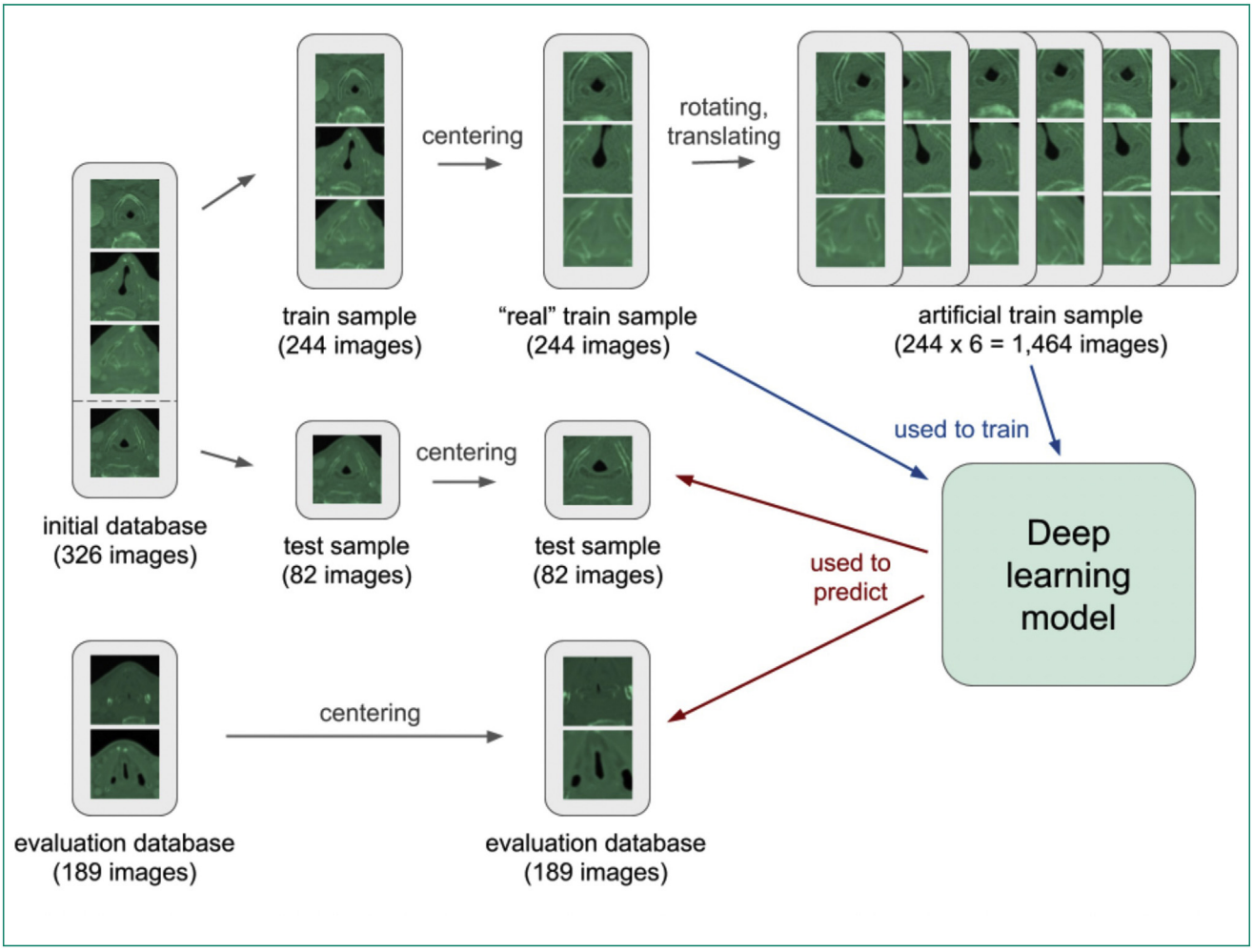

Figure 1. Diagram shows image processing and classification workflow.

The preference for VGG16 was a compromise between high-performance level for image classification tasks and a light structure in comparison to more recent CNN architectures, such as Inception V3 or ResNet-50 [20,21].

Whereas CT images of thyroid cartilages are different from the ImageNet dataset, CNN architectures such as VGG16 are able to extract the main features and shapes of a picture, whatever the content. The first layers are more generic and can be kept as-is, while deep layer weights must be retrained in order to achieve a good classification of normal/abnormal thyroid cartilage. In the present case, those deep layers are expected to extract the features that focus on cartilage shape and abnormalities once the retraining is achieved.

The VGG16 model is composed of five blocks, which combine two-dimensional (2D) convolutions, 2D pooling, and max-padding layers, finishing with a dense layer that operates the final classification through 1000 different categories proposed by the ImageNet dataset [18,19]. Thanks to its simplicity, the VGG16 model is well-adapted for transfer learning for a small dataset, with a relatively reasonable number of trainable parameters (approximately 14 million). To achieve the binary classification (normal/abnormal thyroid cartilage), the final dense layers of VGG16, designed for 1000 categories, are removed and replaced with a 2D global average pooling followed by two fully connected layers with random initialization. This final architecture was chosen in order to slowly reduce the high-dimensional space of the last convolutional layer ($7 \times 7 \times 512$) to a 2D binary classification space (Fig. 2).

To train the model, the last two convolutional blocks and the dense layers were trained, whereas the parameters of the first three convolutional blocks were fixed to the original weights of the VGG16 model [19]. This allowed to keep the more generic features of the VGG16 model included in the first layers and adapt the model to the CT images through a limited number of trainable parameters (Fig. 2).

The framework has been implemented using Python 3.6. The loading of NIfTI images and the preprocessing have been done using the libraries Nibabel (version 2.3.0), Numpy (1.15.2) and openCV (version 3.4.3.18). The deep learning part was built with the Keras package (version 1.0.6) using TensorFlow in back end (version 1.11.0). For the CNN training, a stochastic gradient descent was used as an optimizer (learning rate = 0.01 and momentum = 0.9). All computations were carried out on a personal computer with a 2.2 GHz Intel

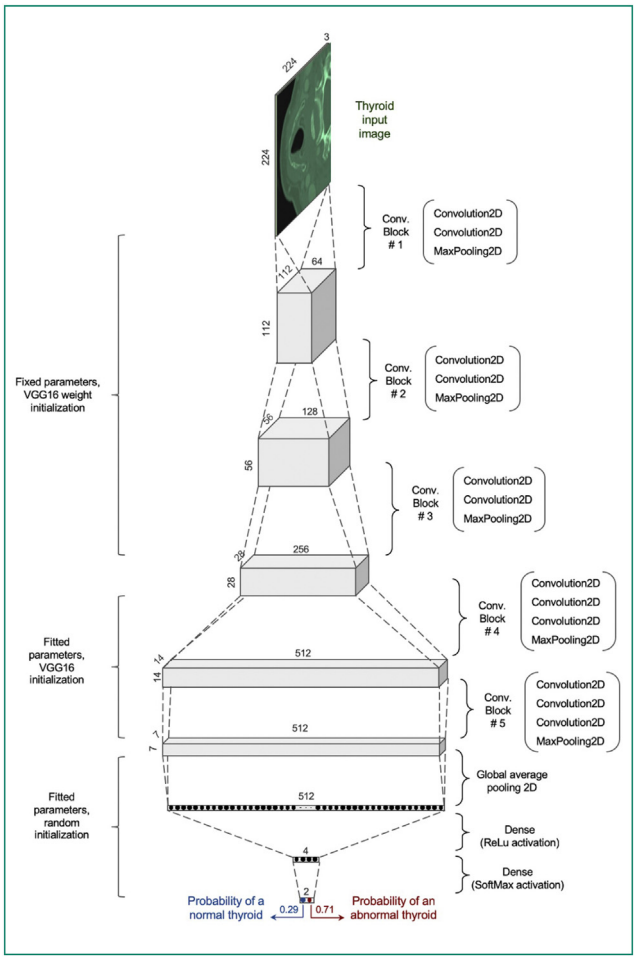

Figure 2. Diagram shows the tuned VGG16 model used in this study. Grey parallelepipedic blocks represent the image tensor, starting with dimensions $224 \times 224 \times 3$ (224 pixels wide and three colors) and ending at $7 \times 7 \times 512$ at the end of the VGG16 convolutional neural network. These layers start from the weight trained from ImageNet. Only the last two blocks (#4 and #5) were trainable. Black dots represent the final layers, adapted for a binary classifier by adding two dense layers at the end of the model. These layers are randomly initialized and trainable.

Core i7. The time required for an epoch was between 5 and 8 minutes, and about 8 epochs were enough to reach the final score evaluated on the test dataset.

Statistical analysis

Performances of the model were evaluated using receiving operator characteristic (ROC) curve, which highlights the prediction ability of a binary classifier. The ROC curve represents the true positive rate (TPR), or the sensitivity, as a function of the false-positive rate (FPR), which is equivalent to 1 — specificity. These rates were computed from the probability of being abnormal, predicted for each image, and their true label. The area under the ROC curve (AUC) indicates the quality of the prediction.

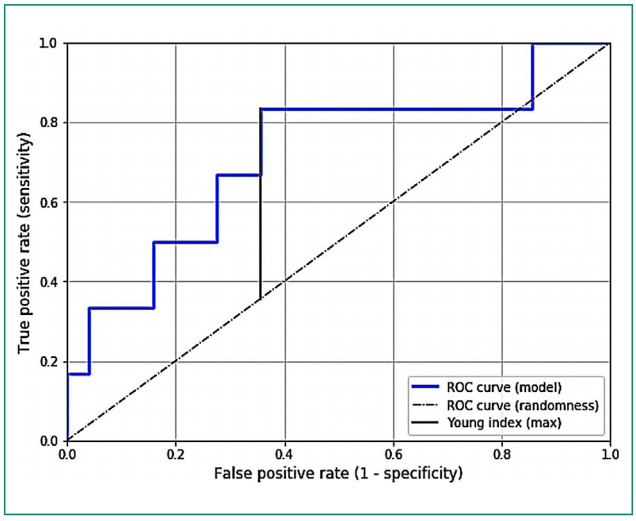

Figure 3. Graph shows ROC deduced from the model applied on the 82 images of the test dataset (76 negative and 6 positive labels). The optimal value (according to the maximum value of the Youden index, illustrated by the continuous grey vertical line) of the sensitivity and specificity are 83% and 64%, respectively. The dashed line represents the ROC curve of a pure random model. The value of the area under the ROC curve (AUC) is 0.72.

Results

Fig. 3 shows the ROC curve of the model presented in this work applied on the 82 images (76 normal and 6 abnormal thyroid) of the test dataset. Those images were independent from the training dataset and, therefore, were never seen by the CNN architecture. Deduced from this ROC curve, the AUC value was 0.72 while the optimal value of the specificity and the sensitivity were 64% and 83%, respectively, as per the Youden index [22].

Despite the low volume of the test dataset and therefore the complexity of obtaining accurate statistics, a similar AUC value was recovered independently on the dataset kept by the challenge organizers, which was not shared to the participants. Indeed, applying the model on those 189 new thyroid images resulted in an AUC score of 0.70, demonstrating the reliability of the results. As the precise scores associated with the validation images were not communicated, it was not possible to compute sensitivity and specificity for this dataset.

A comparison of the classification predicted by the algorithm and the true state of four cartilage images is presented in Fig. 4.

Discussion

In this study, the performance of a deep learning algorithm to detect the abnormality of a thyroid cartilage from CT examination was reflected by a 0.72 AUC score, and sensitivity and specificity of 83% and 64%, respectively.

Thyroid cartilage is a hyaline cartilage structure, which forms the bulk of the front wall of the larynx. It is comprised of a left and a right lamina converging anteriorly, the lower two thirds of which fuse in the midline, constituting the laryngeal prominence. Immediately above it, the

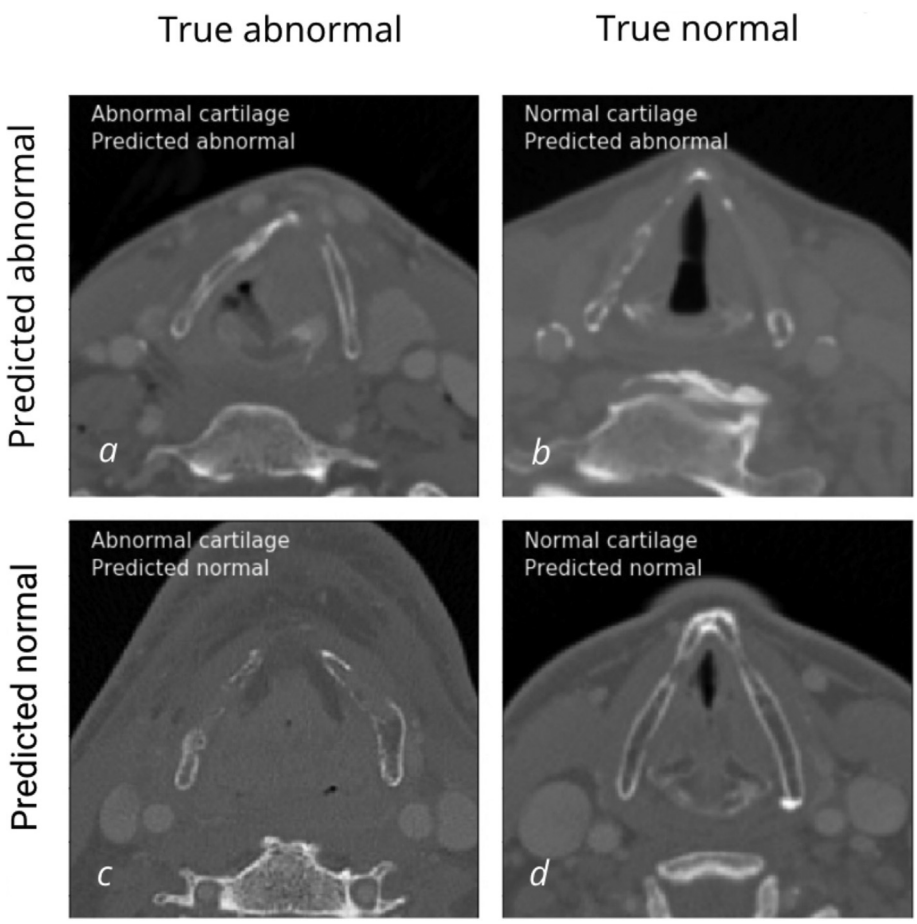

Figure 4. Panel shows the four possibilities for a classifier: a: abnormal cartilage predicted as abnormal (true-positive); b: normal cartilage predicted as abnormal (false-positive); c: abnormal cartilage predicted as normal (false-negative), and d: normal cartilage predicted as normal (true-negative). Figures a and d represent correct classifications, while figures b and c represent misclassification. Partial-volume effect due to the complex shape of the cartilage as well as variation in cartilage ossification might explain the false-positive and false-negative results. More precisely, in figure b, the algorithm may have been fooled by the natural asymmetry of an ossification. The difficulty to propose a diagnostic, even for a radiologist, may explain the algorithm misclassification in figure c.

laminæ are separated by a V-shaped notch; the superior thyroid notch. The laminæ are irregularly quadrilateral in shape, and their posterior angles are prolonged into processes termed the superior and inferior cornua or horns. Calcification and ossification of laryngeal cartilage are part of the normal aging process and are usually detected after the second decade of life [6,9]. Ossification usually starts after age 20 at the posteroinferior aspect of the laminae and inferior cornua, and spreads upward and anteriorly. It is unusual to see isolated superior laminae or superior cornua ossification. Normal variants include agenesis of the thyroid horns, presence of a triticeal cartilage, ectopic horns, lateral thyrohyoid ossification, and terminal segmentation of the thyroid horns [9,12]. The complex shape of the thyroid cartilage, the variation in ossification, and the presence of normal variants make the imaging evaluation of the structure quite difficult for radiologists. The same problem applies to deep learning algorithms, which can have difficulties identifying clear features from one image to another.

The best imaging technique for the evaluation of thyroid cartilage is a matter of debate. CT is considered to have low sensitivity and high specificity for assessing cartilage invasion, whereas magnetic resonance imaging (MRI) has high sensitivity but lower specificity [3,23,24]. In our study, the performance of a deep learning algorithm to detect the abnormality of a thyroid cartilage from a CT scan examination was not leveling up to the expertise of a radiologist. Becker et al. reported that the selection of appropriate diagnostic criteria of neoplastic invasion of laryngeal cartilage could lead to a sensitivity of 91% with a specificity of 68% or an overall specificity of 79% associated with a sensitivity of 82% [25]. A systematic review of CT detection of cartilage invasion in laryngeal carcinoma found that CT imaging is a suitable tool to assess laryngeal cartilage invasion [23]. Dhoot et al. evaluated thyroid cartilage invasion by cancer and found the detection rate of CT was 98%. CT achieved a sensitivity of 91% and a specificity of 75% [3].

The characteristics of the VGG16 model might explain the differences between the performances of a deep learning algorithm and the radiologist to detect the abnormality of a thyroid cartilage from CT examination. The methodological bases of this study could also explain some differences in performance. Of note, when reading a case, the radiologist has at his disposal all the native images of the larynx as well as the multiplanar reconstructions with both the soft tissue and bone windows, which allow the depiction of the

tumor and are, therefore, mandatory for thyroid cartilage analysis. Indeed, in most situations, cartilage invasion is visible in contact with the tumor [25]. In our study, only one axial image with bone windowing was available so that key information was lacking. Using multiplanar reconstruction of the cartilage could improve the performance by providing more information. A study has recently shown the use of three-dimensional images of knee MRIs for the detection of abnormalities [26]. However, this approach requires huge computational power and resources.

In our study, the database used to train and test the model was restrained and harmonized, and the variety of the sources was unknown. This means that the model's capacity to generalize is difficult to estimate, and it might behave very erratically on completely independent data. The algorithm should also be tested on post-mortem CT examinations performed for medicolegal purposes. Indeed, in forensic sciences, the evaluation of thyroid cartilage is of great importance in determining the cause of death. Despite its limitation, our model has the potential to be greatly improved, for example, through image segmentation to locate the cartilage on the images. This would allow the algorithm to focus on the relevant features to detect an abnormality. Nonetheless, it would require a significant amount of work to label the data and create a segmentation system based on shapes or intensity, which would then need to be evaluated. Image processing could also be used to improve image quality (e.g., increase contrast) and therefore improve classification results. Another potential improvement to the model would be to have access to a bigger dataset: at least 1000 images would be much more effective for the model training, by feeding the neural network with a wider type of images. Other neural network structures could also be tested, notably Inception V3 or ResNet-50 [20,21], which have achieved higher performance on the ImageNet classification. However, those models are more complex with an important number of parameters (more than 25 million); therefore, developing such models makes more sense with larger databases.

The algorithm is not yet able to be autonomous or match a professional's evaluation, but could act as a facilitator for radiologists presented with obvious diagnoses, allowing them to focus on more difficult cases. As opposed to replacing the radiologists, deep learning can become a tool for medical professionals to improve productivity and decrease workload and mistakes. The goal is a system whereby decisions can be utilized but are consistently evaluated and checked by the medical expert to make it progressively more efficient. This can be done through the growing collaboration between data scientists and radiologists, which has become more prevalent in many hospitals and is the basis of this project.

In conclusion, this project demonstrates the feasibility of using a deep learning algorithm to assist radiologists in detecting abnormal thyroid cartilages from CT examinations due, for instance, to tumoral invasion or traumas in the case of forensic science. Although the results presented show real potential, it is clear that the model is not yet as effective as a radiologist. This is a consequence of the small dataset used for training the model, and is partly due to the complex shape of the thyroid cartilage resulting in very disparate CT images. To be used in clinical practice, the algorithm needs more data and improvement. This can be accomplished through a collaboration between data scientists and radiologists who understand its deficiencies and aim to make the algorithm more effective.

Funding

This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Contribution of authors

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

Disclosure of interest

The authors declare that they have no competing interest.

References

[1] Becker M, Zbaren P, Casselman JW, Kohler R, Dulguerov P, Becker CD. Neoplastic invasion of laryngeal cartilage: reassessment of criteria for diagnosis at MR imaging. Radiology 2008;249:551—9.
[2] Castelijns JA, Becker M, Hermans R. Impact of cartilage invasion on treatment and prognosis of laryngeal cancer. Eur Radiol 1996;6:156—69.
[3] Dhoot NM, Choudhury B, Kataki AC, Kakoti L, Ahmed S, Sharma J. Effectiveness of ultrasonography and computed tomography in assessing thyroid cartilage invasion in laryngeal and hypopharyngeal cancers. J Ultrasound 2017;20:205—11.
[4] Kuno H, Onaya H, Iwata R, Kobayashi T, Fujii S, Hayashi R, et al. Evaluation of cartilage invasion by laryngeal and hypopharyngeal squamous cell carcinoma with dual-energy CT. Radiology 2012;265:488—96.
[5] Garetier M, Deloire L, Dedouit F, Dumousset E, Saccardy C, Ben Salem D. Postmortem computed tomography findings in suicide victims. Diagn Interv Imaging 2017;98:101—12.
[6] Naimo P, O'Donnell C, Bassed R, Briggs C. The use of computed tomography in determining developmental changes, anomalies, and trauma of the thyroid cartilage. Forensic Sci Med Pathol 2013;9:377—85.
[7] Zatopkova L, Janik M, Urbanova P, Mottlova J, Hejna P. Laryngohyoid fractures in suicidal hanging: a prospective autopsy study with an updated review and critical appraisal. Forensic Sci Int 2018;290:70—84.
[8] Blum A, Kolopp M, Teixeira PG, Stroud T, Noirtin P, Coudane H, et al. Synergistic role of newer techniques for forensic and postmortem CT examinations. AJR Am J Roentgenol 2018;211:3—10.
[9] Advenier AS, De La Grandmaison GL, Cavard S, Pyatigorskaya N, Malicier D, Charlier P. Laryngeal anomalies: pitfalls in adult forensic autopsies. Med Sci Law 2014;54:1—7.
[10] Aramaki T, Ikeda T, Usui A, Funayama M. Age estimation by ossification of thyroid cartilage of Japanese males using Bayesian analysis of postmortem CT images. Leg Med (Tokyo) 2017;25:29—35.
[11] Beitler JJ, Muller S, Grist WJ, Corey A, Klein AM, Johns MM, et al. Prognostic accuracy of computed tomography findings

for patients with laryngeal cancer undergoing laryngectomy. J Clin Oncol 2010;28:2318—22.
[12] Castan Senar A, Dinu LE, Artigas JM, Larrosa R, Navarro Y, Angulo E. Foreign bodies on lateral neck radiographs in adults: imaging findings and common pitfalls. Radiographics 2017;37:323—45.
[13] Pinheiro J, Cascallana JL, Lopez de Abajo B, Otero JL, Rodriguez-Calvo MS. Laryngeal anatomical variants and their impact on the diagnosis of mechanical asphyxias by neck pressure. Forensic Sci Int 2018;290:1—10.
[14] Ambroise Grandjean G, Hossu G, Bertholdt C, Noble P, Morel O, Grange G. Artificial intelligence assistance for fetal head biometry: assessment of automated measurement software. Diagn Interv Imaging 2018;99:709—16.
[15] Blum A, Zins M. Radiology: is its future bright? Diagn Interv Imaging 2017;98:369—71.
[16] Liu F, Zhou Z, Samsonov A, Blankenbaker D, Larison W, Kanarek A, et al. Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. Radiology 2018;289:160—9.
[17] Prevedello LM, Erdal BS, Ryu JL, Little KJ, Demirer M, Qian S, et al. Automated critical test findings identification and online notification system using artificial intelligence in imaging. Radiology 2017;285:923—31.
[18] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Inter J ComputVision 2015;115:211—52.
[19] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition; 2014.
[20] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition; 2016. p. 770—8 [Proceedings of the IEEE conference on computer vision and pattern recognition].
[21] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision; 2016. p. 2818—26 [Proceedings of the IEEE conference on computer vision and pattern recognition].
[22] Youden WJ. Index for rating diagnostic tests. Cancer 1950;3:32—5.
[23] Adolphs AP, Boersma NA, Diemel BD, Eding JE, Flokstra FE, Wegner I, et al. A systematic review of computed tomography detection of cartilage invasion in laryngeal carcinoma. Laryngoscope 2015;125:1650—5.
[24] Kuno H, Sakamaki K, Fujii S, Sekiya K, Otani K, Hayashi R, et al. Comparison of MR imaging and dual-energy CT for the evaluation of cartilage invasion by laryngeal and hypopharyngeal squamous cell carcinoma. AJNR Am J Neuroradiol 2018.
[25] Becker M, Zbaren P, Delavelle J, Kurt AM, Egger C, Rufenacht DA, et al. Neoplastic invasion of the laryngeal cartilage: reassessment of criteria for diagnosis at CT. Radiology 1997;203:521—32.
[26] Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Med 2018;15 [e1002699].