

Review

# Artificial Intelligence in Thyroid Field—A Comprehensive Review

Fabiano Bini <sup>1,\*</sup>, Andrada Pica <sup>1</sup>, Laura Azzimonti <sup>2</sup>, Alessandro Giusti <sup>2</sup>, Lorenzo Ruinelli <sup>3,4</sup>, Franco Marinozzi <sup>1</sup> and Pierpaolo Trimboli <sup>5,6</sup>

- <sup>1</sup> Department of Mechanical and Aerospace Engineering, Sapienza-University of Rome, 00184 Rome, Italy; andrada.pica@uniroma1.it (A.P.); franco.marinozzi@uniroma1.it (F.M.)
- <sup>2</sup> Dalle Molle Institute for Artificial Intelligence (IDSIA), Università della Svizzera Italiana (USI), Scuola Universitaria Professionale della Svizzera Italiana (SUPSI), Polo Universitario Lugano-Campus Est, 6962 Lugano-Viganello, Switzerland; laura.azzimonti@idsia.ch (L.A.); alessandro.giusti@idsia.ch (A.G.)
- <sup>3</sup> Information and Communications Technology, Ente Ospedaliero Cantonale, 6500 Bellinzona, Switzerland; lorenzo.ruinelli@eoc.ch
- <sup>4</sup> Clinical Trial Unit, Ente Ospedaliero Cantonale, 6500 Bellinzona, Switzerland
- <sup>5</sup> Servizio di Endocrinologia e Diabetologia, Ospedale Regionale di Lugano e Mendrisio, Ente Ospedaliero Cantonale, 6900 Lugano, Switzerland; Pierpaolo.Trimboli@eoc.ch
- <sup>6</sup> Faculty of Biomedical Sciences, Università della Svizzera Italiana (USI), 6900 Lugano, Switzerland
- \* Correspondence: fabiano.bini@uniroma1.it

**Simple Summary:** The incidence of thyroid pathologies has been increasing worldwide. Historically, the detection of thyroid neoplasms relies on medical imaging analysis, depending mainly on the experience of clinicians. The advent of artificial intelligence (AI) techniques led to a remarkable progress in image-recognition tasks. AI represents a powerful tool that may facilitate understanding of thyroid pathologies, but actually, the diagnostic accuracy is uncertain. This article aims to provide an overview of the basic aspects, limitations and open issues of the AI methods applied to thyroid images. Medical experts should be familiar with the workflow of AI techniques in order to avoid misleading outcomes.

**Abstract:** Artificial intelligence (AI) uses mathematical algorithms to perform tasks that require human cognitive abilities. AI-based methodologies, e.g., machine learning and deep learning, as well as the recently developed research field of radiomics have noticeable potential to transform medical diagnostics. AI-based techniques applied to medical imaging allow to detect biological abnormalities, to diagnostic neoplasms or to predict the response to treatment. Nonetheless, the diagnostic accuracy of these methods is still a matter of debate. In this article, we first illustrate the key concepts and workflow characteristics of machine learning, deep learning and radiomics. We outline considerations regarding data input requirements, differences among these methodologies and their limitations. Subsequently, a concise overview is presented regarding the application of AI methods to the evaluation of thyroid images. We developed a critical discussion concerning limits and open challenges that should be addressed before the translation of AI techniques to the broad clinical use. Clarification of the pitfalls of AI-based techniques results crucial in order to ensure the optimal application for each patient.

**Keywords:** thyroid neoplasm; medical imaging; artificial intelligence; machine learning; deep learning; radiomics; prediction; diagnosis



**Citation:** Bini, F.; Pica, A.; Azzimonti, L.; Giusti, A.; Ruinelli, L.; Marinozzi, F.; Trimboli, P. Artificial Intelligence in Thyroid Field—A Comprehensive Review. *Cancers* **2021**, *13*, 4740. <https://doi.org/10.3390/cancers13194740>

Academic Editor: Kennichi Kakudo

Received: 31 August 2021

Accepted: 20 September 2021

Published: 22 September 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The role of medical imaging in the clinical workflow has noticeably increased from a mainly diagnostic tool up to a central contribution in early detection, diagnosis, treatment planning and monitoring of diseases [1–4]. Medical imaging provides information concerning the characteristics of human tissues in a non-invasive, repeatable manner and

became a routine practice in clinical care [2]. In recent decades, the innovations in this field concerned both devices, i.e., hardware, and analysis tools used in medical imaging. In the clinical practice, the main use of medical images corresponds with qualitative assessment of the anatomical area. Images, in addition, are characterized also by a high quantity of numerical information and recently, a quantitative evaluation has been developed in order to identify possible correlations between the numerical data contained in the digital images and the pathophysiology of the tissue [3]. The quantitative analysis has the aim to achieve information from standard-of-care images, e.g., ultrasound imaging (US), computer tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET), which are not easily quantifiable by means of naked-eye observations for clinical outcomes [5,6].

Analysis of image features in the context of medical imaging is an emerging field of study but extensive literature already exists [7–9]. In the majority of earlier works, the image features are analyzed with the aim of detection and diagnosis of abnormal regions within human tissues [10–12]. These applications are often referred as computer-aided detection (CADe) and computer-aided diagnosis (CADx) systems [3]. The output of the CAD analysis is used by the expert clinicians as a second opinion in detecting lesions or making diagnosis and aims at improving the accuracy of the diagnosis and reducing the time for image interpretation [6].

Recently, a further detailed extension associated with quantitative analysis of medical images has led to the emergence of radiomics as a new field of medical research [1,2]. Radiomics aims at extracting numerous quantitative descriptors with the purpose of achieving more useful information of tissue lesion and response of treatment in order to be used for personalized medicine [1,2,13]. It is worth noticing that standardization of the procedure is still under development, as thoroughly discussed in [14].

The above-mentioned approaches apply methodologies from the artificial intelligence (AI) field to achieve a partial or full automation of various steps of the process concerning the analysis of medical images [6]. Thorough understanding of their working principle is necessary in order to develop efficient predictive models and personalization treatment. This review article aims to highlight strengths and limitations of the different AI-based techniques applied for the evaluation of the pathophysiological state of the thyroid.

## 2. Artificial Intelligence in Medical Imaging

Artificial intelligence (AI) is a term coined by McCarthy and colleagues [15,16] in 1950s referring to a branch of computer sciences in which mathematical algorithms attempt to perform tasks that normally require human cognitive abilities [8]. Applications of AI have witnessed unprecedented growth in recent decades due to the enhancement of computational power and availability of large dataset. In the medical field, AI can use complex algorithms to develop models with the scope of improving diagnostic accuracy, prognosis, and medical image interpretation [17]. We discuss in the following two different machine learning (ML) methodologies adopted to perform medical imaging analysis.

### 2.1. Machine Learning

Machine learning (ML), a term first coined by Arthur Samuel [18], is a field of AI in which the computer is trained to perform tasks by learning from example data and make predictions based on its exposition to previous samples [4]. In medical imaging analysis, ML algorithms are crucial components of both CAD systems and radiomics studies.

ML algorithms are generally divided into supervised and unsupervised learning methods. Supervised learning requires a labelled dataset, i.e., a set of input data with their corresponding output (labels) that is used to identify a function linking inputs to outputs [19]. Unsupervised learning operates on an input dataset without the need of labels. This ML algorithm searches for patterns that can separate input data into subsets with similar characteristics [7]. In this review article, we focus on supervised learning since it is the most common approach applied to medical images analysis [20].

In medical applications, input data include medical images or clinical data, while the output label can be the differentiation of malignant from benign nodules, the classification of images into diagnostic categories or the response treatment, e.g., recurrence, survival. The output of the predictive model leads to a subsequent distinction of ML problems: classification and regression. In classification tasks, the model performs a decision among a small and discrete set of choices, i.e., binary classification, e.g., identifying a tumor as malignant or benign. Regression models refers to the estimation of continuous output variables, e.g., assessment of disease severity [20].

Historically, ML algorithms were applied in CAD systems for classifications purposes [20]. Subsequently, this method was used as a step of radiomics analysis. In this section we describe the workflow of the ML algorithm with classification task frequently encountered in the CAD framework [10,21,22] (Figure 1). A supervised ML model is composed of two phases, i.e., training and application phase (Figure 1a). In the training phase, a set of input images with their corresponding class labels are used to train the predicting model. From the input image, a region of interest (ROI) is delineated manually or semi-automatically by expert clinicians. Subsequently, a set of image features, e.g., morphological and grey level-based features, are extracted. Differently from other methods that will be discussed subsequently in this work, in ML algorithms of CAD systems, the extraction and selection of image features are performed manually by the expert. It represents a crucial step in order to identify the significant variables that can be correlated with the medical endpoint. In CAD applications, the features used in the analysis are those closely associated with what clinicians use in their diagnosis of the lesions [4]. Subsequently, the features are entered as input to the ML algorithm to train the model.

Examples of typical feature-based supervised learning algorithms are logistic regressions, support vector machine, random forests and neural networks [23]. As an example of these feature-based ML algorithms, we focus here on the support vector machine (SVM) method, which is commonly used in biomedical binary classification problems [17,24]. Overall, SVM (Figure 1b) is a binary classifier that aims to identify the decision boundary, or hyperplane, that maximizes the separating margin between two classes [4,25].

For instance, let consider  $N$  training samples  $\{(x_i, y_i)\}_{i=1}^N$  of input features  $x$  and their corresponding class or label  $y_i \in \{-1; +1\}$  where  $y_i = -1$  indicates the class with malignant samples and  $y_i = 1$  indicates the class with benign samples. In the simplest case, there exists a function  $f(x)$ :

$$f(x) = \beta \cdot x + \beta_0, \quad (1)$$

with  $\beta$  and  $\beta_0$ —decision boundary parameters such that  $f(x) \geq 0$  for  $y_i = +1$  and  $f(x) < 0$  for  $y_i = -1$ .

This means that the training samples from the two classes are separated by the hyperplane  $f(x) = \beta \cdot x + \beta_0 = 0$ . The margin  $m$ , i.e., the distance between a class and the decision boundary, is set to be inversely proportional to the decision boundary parameter, i.e.,  $m = \frac{1}{\|\beta\|}$ .

In order to identify the hyperplane that maximizes the separating margin between the two classes, SVM solves the following optimization problem that aims to minimize the cost function  $J(\beta, \zeta)$  with respect to  $\beta, \zeta$  [24,26]:

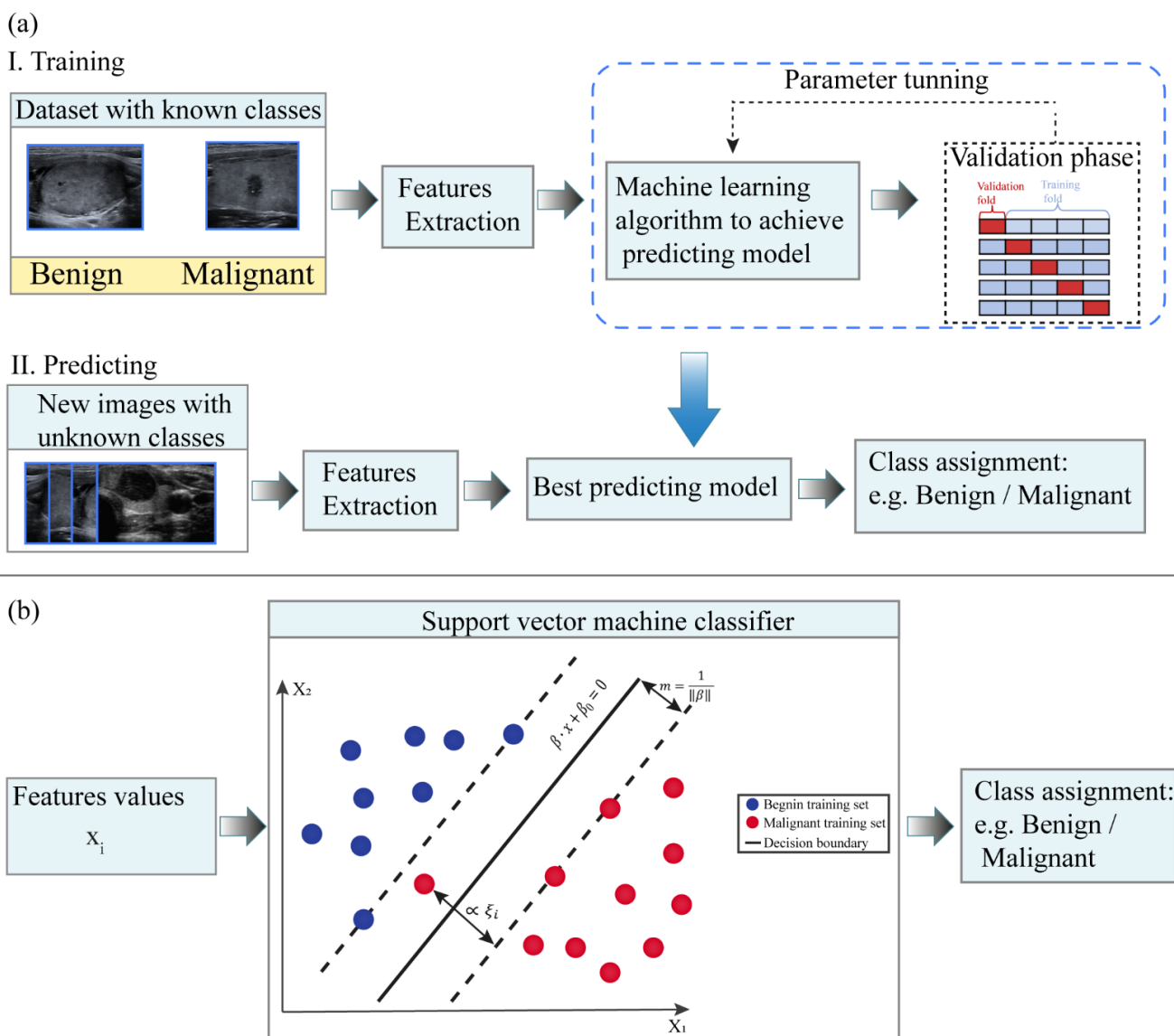
$$\min_{\beta, \zeta} J(\beta, \zeta) = \min_{\beta, \zeta} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^N \zeta_i \quad (2)$$

subject to the constraints  $y_i (\beta \cdot x + \beta_0) \geq 1 - \zeta_i, \zeta_i \geq 0, i = 1, \dots, N$ . In Equation (2),  $C > 0$  is a penalty parameter to control the tolerance error  $\zeta_i$  allowed for each sample being on the wrong side of the margin.

From Equation (2), it can be noticed that the minimization of the parameter  $\beta$  increases the separation between the two classes and improves generalizability of the classifier, while minimization of second term of Equation (1) improves fitting accuracy [4].

Subsequently, in the testing phase, the trained classifier is used to characterize new input data with unknown label (test set).

It is worth pointing out that the decision function of the classifier is fully specified by the training set, while the test set is only used to evaluate the performance of the model. On one hand, to obtain a model that well-performs when applied to new data, the training dataset should be sufficiently large. On the other hand, to obtain robust and reliable evaluation of the performance of the model, the test set should be sufficiently large. Frequently, since this condition is difficult to achieve in the medical field by simply splitting the available data in training and test set, a k-fold cross-validation framework [7] is usually adopted. K-fold cross validation consists of partitioning the dataset into k subsets of equal size. The model is trained on (k - 1) datasets while one subset is retained for model test. The process is repeated k times with each subset used once as test dataset [20]. The overall performance of the model is then assessed for example as the average performance over the k repetitions.



**Figure 1.** (a) Schematic flowchart of the machine learning model implementation and application for medical images classification purposes. (b) Example of the support vector machine (SVM) classification with a hyperplane that maximizes the separating margin  $m$  between the two classes.

Feature-based ML algorithms are suitable for medical image analysis since predictive models can be developed from small datasets [7]. Moreover, these methods are usually interpretable and can provide insights on the reasons why a certain class is predicted. Nonetheless, some initial steps of the process, as the definition of the features to be extracted from images and the selection of the medical region of interest has to be performed by experts. In addition, it should be taken into account that all supervised ML methods could be affected by overfitting, i.e., the predicting model learns exactly the training set but fails to fit new data from the test set [20]. However, it is possible to mitigate this issue by adopting a cross-validation set-up and by reducing the number of features used by the model by means of feature selection methods.

## 2.2. Deep Learning

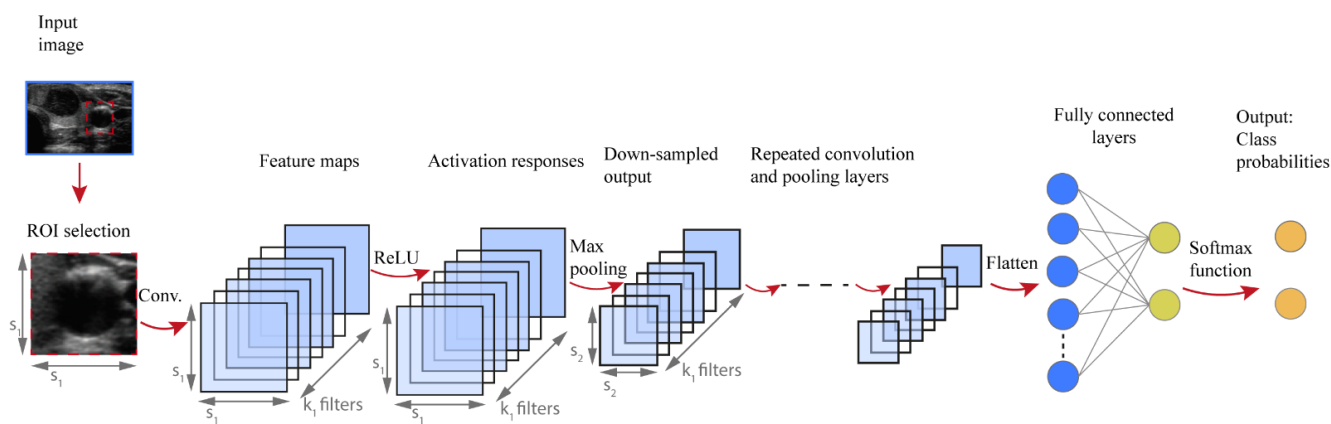
Deep learning (DL), a term coined in 1986 by Rina Dechter [27], is a new class of ML methods developed through the advancement of artificial neural networks which were considered as artificial representations of the human neural architecture [23]. DL relies on networks of computational units, i.e., neural units arranged in layers that gradually extract higher level features from input data, e.g., image. These structures learn discriminative features from data automatically, allowing to approximate complex nonlinear relationship with outstanding performance [27,28]. Differently from traditional feature-based ML approaches, DL is able to achieve diagnosis automation, avoiding human intervention [29]. In medical applications, DL algorithms are implemented for detection and characterization of tissue lesions as well as for the analysis of disease progression [27,28].

While several DL architectures have been developed, this article focuses on convolutional neural networks (CNNs), introduced by LeCun [30]. CNNs are typically applied for image recognition and computer vision applications because they preserve spatial relationships in 2D data, and therefore outperform other architectures on image pattern recognition. More specifically, the input of a CNN is arranged in a grid structure and processed through convolution and pooling layers that preserve these relationships. The final layers are typically fully connected and can be conceived as a multi-layer perceptron classifier on the features automatically extracted by the convolutional part. The network is trained to identify patterns in a set of labelled training data and the outputs are compared with the actual labels. During training the network parameters are tuned until the patterns identified by the network represent good predictions for training data. The network is then used to make predictions on new data in the test set [31].

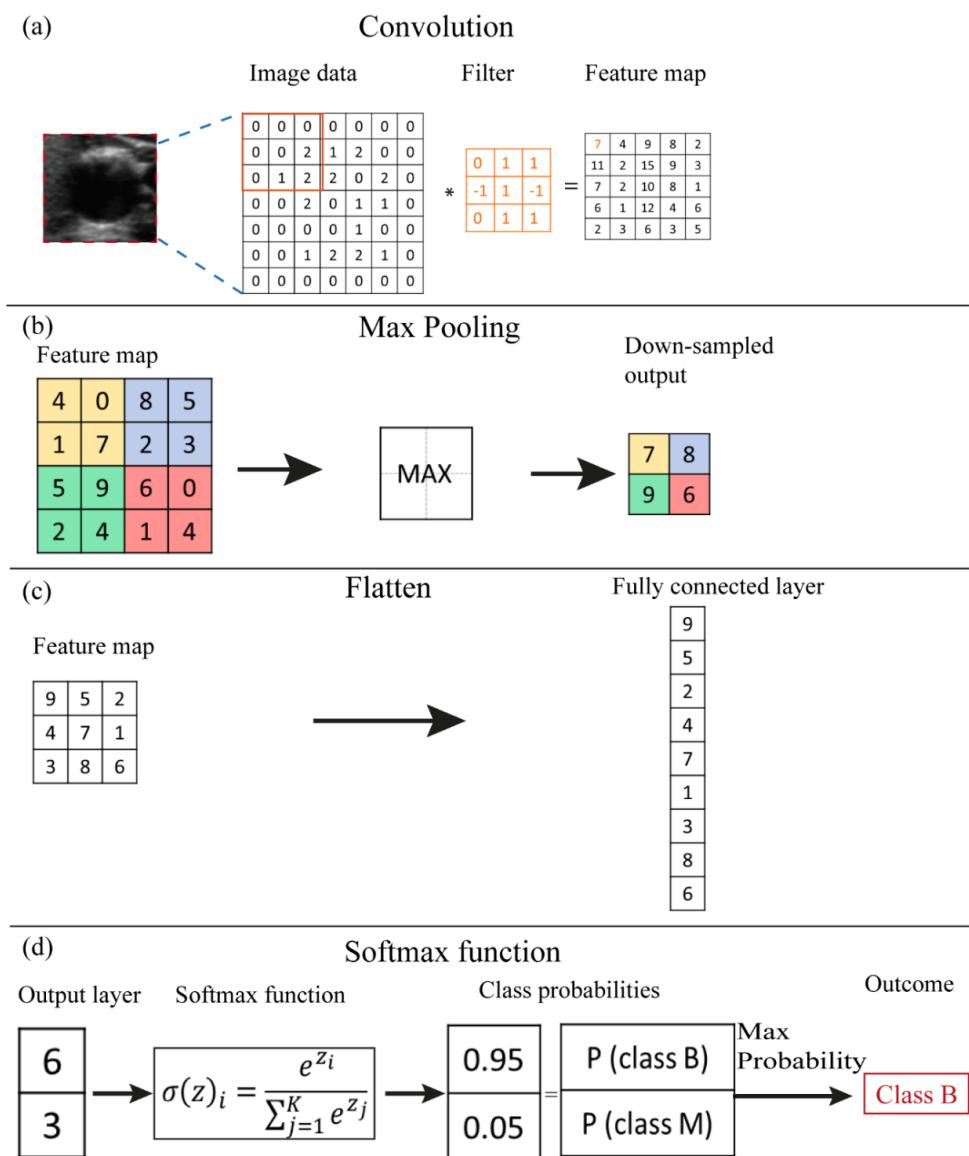
Figure 2 shows a typical architecture of CNN developed to perform classification tasks. The input of the CNN algorithm is represented by numerical data of the selected ROI from the medical image. Firstly, a convolutional step is considered which contains a set of filters, e.g.,  $k_1$  in Figure 2. Thus, a convolution is performed between each filter and the input of the layer, e.g., image data. A convolution is a space-invariant linear operation on 2D grids and is equivalent to applying a filter to an image. The filter slides over the input image, its values are multiplied with the image pixel values and then summed to determine the value in the corresponding position of the output feature map. An example of a convolution operation is reported in Figure 3a. The number and size of filters are CNN hyperparameters and are typically not optimized during training. More and larger filters lead to more powerful network with more parameters to optimize, which increases the risk of overfitting [32]. The convolutional process in every convolutional layer is expressed mathematically as follows:

$$X_k^\ell = \sigma\left(W_k^{\ell-1} * X^{\ell-1} + b_k^\ell\right) \quad (3)$$

where  $X_k^\ell$  is the new feature map,  $\sigma(\cdot)$  is an element-wise nonlinear activation function,  $W$  is the filter values,  $b_k^\ell$  is a bias parameter and the symbol  $*$  indicates a convolutional operator.



**Figure 2.** Schematic flowchart of the deep learning model implementation and application for medical images classification purposes.



**Figure 3.** (a) Numerical example of the functions that compose the CNN architecture: (a) convolution, (b) max pooling, (c) flattening, (d) softmax function.



Subsequently, an activation function is applied element-by-element to the calculated output of the convolution prior to using the map as an input to the next layer of the network. Rectified linear unit (ReLU) is one of the most used activation functions, and has been empirically found to accelerate the convergence of the learning procedure [28]. It is linear for positive inputs, mapping them unchanged to the next layer, while it blocks negative values. Mathematically, ReLU is expressed as follows [28]:

$$f(x) = \max(0, x) \quad (4)$$

where  $x$  is an activation value achieved from the previous layer.

Some CNN architectures also consider pooling operations, whose effect is to down-sample the feature maps. This operation considers small regions of the input map and outputs a single number for each region, e.g., the maximum value as illustrated in Figure 3b. It reduces the dimensions of the feature map and decreases the number of pixels to be processed in the next layers of the network [33]. Conceptually, as we progress deeper in the network, neuron activation values represent progressively higher-level and larger-scale visual patterns in the input, and therefore require lower spatial resolution.

The final part of the CNN architecture is characterized by a fully connected layer, i.e., each neural unit of the actual layer is connected to every neural unit in the successive layer (Figure 2). Firstly, the feature map is flattened into a column vector (Figure 3c) and then connected to one or more fully connected layers. The output nodes of the last fully connected layer can be regarded as a vector of unnormalized probabilities [28].

The softmax function is a function applied to the last fully connected layer of the CNN in order to transform the  $k$  real values of the vector into values in the range (0;1) so that can be assumed as probabilities (Figure 3d). The relation is as follows [28,33]:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (5)$$

where the  $z_i$  values are the elements of the fully connected layer and the denominator represents the normalization term.

The output layer of the CNN considered is constituted by neural units which indicate the probabilities for each class.

The analysis of the available literature shows an increasing interest on applying DL architecture for medical image analysis. It is worth mentioning that for systems in which the set of visual features is well defined, simpler feature-based ML techniques, such as SVM algorithms, are easier, more interpretable and more effective [28].

The main limitation to the use of DL consists of the large datasets required to train the model [34]. Compared with publicly available datasets in other areas, the current availability of medical US datasets is still limited [34]. To face the data requirements, several studies [33,35] considered pre-trained CNN architectures developed with trainings on ImageNet, a large labelled collection of low-resolution color photographs. To date, DL architectures pre-trained on high resolution medical images are not available. Therefore, a large dataset of medical images is a mandatory step to enhance CNNs performance [34].

### 3. Radiomics

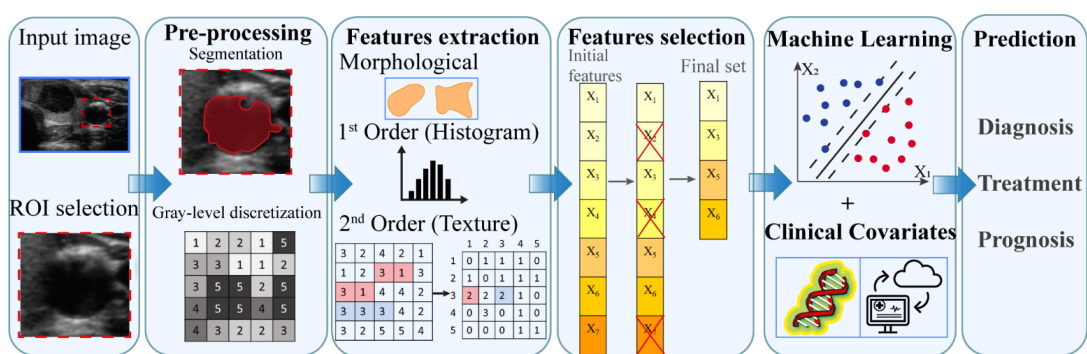
Radiomics is an emerging field that uses automated high-throughput extraction algorithms to achieve large amounts (200+) of quantitative features from medical images [1,2]. Radiomics is also indicated as quantitative imaging [36] which can be applied to any image generated in the clinical setting. It can be performed on subregions of a tumor, metastatic lesions and in normal tissues. The term feature represents a descriptor of an image, of tumor or healthy tissue, such as parameters derived from image grayscale intensity or shape [37].

Radiomics has its roots on computer-aided diagnosis systems [38], although methodological workflow and applications are distinct [2]. It concerns the extraction of quantitative

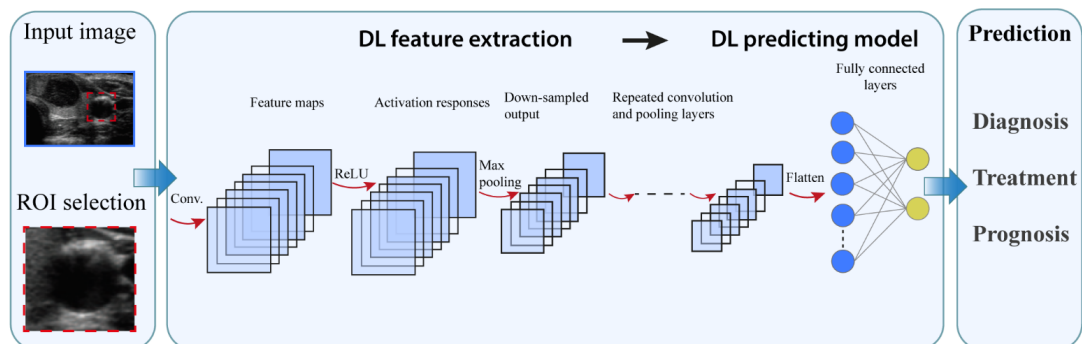
features from medical images that subsequently are related to biological endpoints and clinical outcomes [39]. Radiomics makes use of digital data stored in those images to develop diagnostic, predictive or prognostic models to support clinical decisions and optimize personalized treatment planning. The main difference with CAD systems consists of the relationship that radiomics has to identify between the current characteristics of the tissue lesion and its temporal evolution in the perspective of a personalization of the therapy [38].

Radiomics involves several processes, each with its own critical aspects that need to be taken into account. Two workflows can be implemented to perform radiomic studies in function of the AI technique adopted (Figure 4): (i) conventional or ML-based radiomics where the features to be extracted are predefined and (ii) DL-based radiomics where the features are not predefined but automatically extracted from the underlying data [6,7].

### I. Conventional radiomics



### II. Deep Learning based radiomics



**Figure 4.** Schematic flowchart of radiomics approach.  $X_i$  represent the feature extracted from the image data.

The main aspects of the conventional radiomics workflow concerns: image acquisition, data selection, feature extraction and selection and the development of predictive model [1,36]. From medical image such as US, CT, MR and/or PET images, the region of interest (ROI) is selected and subsequently the lesion is manually segmented, i.e., delineated with computer-assisted contouring, by an experienced clinician [7]. Subsequently, image data undergoes preprocessing operations, e.g., gray-level discretization, which enable a higher reproducibility of results [6]. The extraction of quantitative imaging features involves descriptors of spatial relationships between the various intensity level, heterogeneity patterns, shape and relations of the tissue lesion with surrounding tissues. A feature selection procedure is then performed to identify the most relevant predictive features [7,24]. The collection of features which hold prognostic or predictive value represent a feature signature, frequently indicated also as quantitative imaging biomarkers. The selected features are then analyzed to develop classified models to predict outcomes either alone or in combination with additional information, such as demographic, clinical, comorbidity or genomic data [1,3].



Segmentation represents a crucial subprocess of radiomics since many extracted features may depend on the segmented region. In several radiomics studies the ROI is manually delineated by experts [21,40–42]. A number of algorithms has been developed for semi-automatic segmentation [22]. Region growing-based algorithm and grey-scale threshold-based methods are frequent techniques applied for ROI definition. However, manual delineation by an expert is considered the gold standard though is subjected to inter-observer variability and is a time-consuming task [37]. To avoid possible bias, evaluation by multiple clinicians or a combination of multiple algorithms could be considered [43].

Typically, radiomics features are divided into [2,6,44]:

1. Morphological, that are based on the geometric properties of the ROI, e.g.: volume, maximum surface area, maximum diameter.
2. First-order statistics or histogram based, which describe, through histograms, the distribution of grayscale intensity without concern for spatial relationships within the ROI. For instance, calculated features are grey level mean, maximum, minimum and percentiles.
3. Second-order statistics or textural features, that represent statistical relationship between the intensity levels of neighboring pixels within the ROI that allow to quantify image heterogeneity, e.g., absolute gradient, grey level co-occurrence matrix (GLCM) grey level run-length matrix (GLRLM), grey level size zone matrix (GLSZM) and grey level distance zone matrix (GLDZM). For instance, GLCM indicates the number of times the same combination of intensity occurs in two pixels separated by a specific distance  $\delta$  in a known direction.
4. Higher-order statistics features, which are computed after the application of mathematical transformation and filters that lead to highlighting repeated patterns, histogram-oriented patterns or local binary patterns, e.g., wavelet or Fourier transforms.

Accurate definitions of radiomics features are provided in the image biomarker standardization initiative (IBSI) [14].

The radiomic features are subjected to a subsequent feature selection to prevent overfitting, improve learning accuracy and reduce computation time. The selection process should eliminate unreliable, not informative or redundant features. The selection methods can be divided into three classes: (i) filter methods which assess the usefulness of a given feature with various statistical tests for their correlation with the outcome variable [2,7]; (ii) wrapper method which uses an external classifier algorithm to score different subsets of features based on their classification performance; (iii) embedded method where the selection is intrinsic to the model training, i.e., features are selected to optimize the performance of the implemented learning algorithm. Filter methods are simple and computationally efficient, but consider features as independent and any interaction between them is ignored [24]. Wrapper methods reduce the risk of overfitting but are computationally intensive [7,24]. Embedded methods are computationally more efficient since the selection procedure is part of the training process [7,24]. A frequent embedded algorithm with good performance used in radiomics studies is the least absolute shrinkage and selection operator (LASSO) [7,24].

Subsequently, the selected features are used to implement a mathematical model in order to predict the established medical endpoints. Regarding the choice of modelling methodology, the identification of a suitable method depends on several factors as sample size or study endpoint [36]. It is advantageous to include in the model information beyond radiomics, e.g., clinical data and/or other “-omic” information, e.g., genomic data [45]. The integration of data from multiple sources, e.g., medical imaging, disease risk factors, therapy procedures and follow up data, in the mathematical model will facilitate the development of a personalized treatment.

As previously mentioned, the target of the radiomics studies can be either a present characteristic, e.g., tumor phenotype, or a future prediction, e.g., treatment response. Usually, radiomics studies make use of the feature-based ML algorithms that are also considered in CAD systems. By means of feature-based ML methods, the relationship

between input data, e.g., selected radiomics features and target outcome, is determined by means of training examples. SVM is one of the most successfully applied algorithms.

DL-based radiomics allows to automatically extract imaging features and achieve the predicted outcome. In fact, the different components of the DL architecture perform all the processing steps described in the ML-based model, including feature extraction, selection and predicting model implementation. CNNs is the most common architecture used in radiomics studies and its characteristics have been previously described in Section 2.2.

Validation is a crucial component of the workflow of both conventional and DL-based radiomics. Ideally, the trained model should be tested in cross-validation or on an external, independent dataset before being applied on the new dataset [38].

#### 4. AI and Radiomics in Thyroid Diseases

Ultrasound imaging is the recommended method for early detection and diagnosis of thyroid lesions due to its economy, effectivity and absence of radiation [46–49]. It is widely accepted as the first imaging modality for thyroid disease, for instance by American and European associations of endocrinology [50]. AI applications in the medical field are of increasing interest since they represent a possible approach to reduce the number of invasive clinical procedures [36].

Mainly, AI algorithms have been implemented for the classification of thyroid nodules, i.e., differentiating among benign or malignant state [9,10,21,22,33,41,51–56]. The outcomes of these studies are compared with the diagnosis of radiologists with different levels of experience. Research comparing the diagnostic ability between feature-based ML and DL algorithms is limited in the literature, but interesting outcomes are provided in [22]. Overall, an improvement emerged in terms of both specificity and accuracy in DL studies [57,58] with respect to feature-based ML classical applications [22], mostly determined by the capacity of DL of capturing complex patterns. In some studies [57–59], DL algorithms show accuracy values in line with those of radiologists. In addition, Jin et al. [20] also pointed out that the use of AI algorithms was useful to junior radiologists allowing a noticeable improvement of their diagnostic performance, reaching values of accuracy similar to those of intermediate-level radiologists. Studies of interest concerning the application of feature-based ML methods and DL algorithms are described in Tables 1 and 2, respectively. Tables were organized according to the publication time, in a decreasing order.

**Table 1.** Machine learning (ML)-based studies.

Study	Description	Cohort	Method	Performance
Zhao et al., 2021 [21]	Classification Benign/malignant thyroid nodules US	106 patients	SVM	Accuracy: 82% Sensitivity: 91% Specificity: 78%
Park et al., 2019 [22]	Classification Benign/malignant thyroid nodules US	286 patients	SVM	Accuracy: 75.9% Sensitivity: 90.4% Specificity: 58.8%
Zhang et al., 2019 [51]	Classification Benign/malignant thyroid nodules US	826 patients	SVM	Accuracy: 83% Sensitivity: 86.1% Specificity: 82.7%
Yoo et al., 2018 [41]	Classification Benign/malignant thyroid nodules US	50 patients	SVM	Accuracy: 84.6% Sensitivity: 80% Specificity: 88.1%
Chang et al., 2016 [10]	Classification Benign/malignant thyroid nodules US	118 patients	SVM	Accuracy: 98.3% Sensitivity: N/A Specificity: N/A

Abbreviations: US—ultrasound; SVM—support vector machine; N/A—not available.

**Table 2.** Deep learning (DL) studies.

Study	Description	Cohort	Method	Performance
Kim et al., 2021 [59]	Malignancy risk thyroid modules	757 patients	CNN	Accuracy: 85.1% Sensitivity: 81.8% Specificity: 86.1%
Wu et al., 2021 [52]	Classification Benign/malignant thyroid nodules US	1396 patients	CNN	Accuracy: 82% Sensitivity: 85% Specificity: 78%
Jin et al., 2020 [11]	Classification Benign/malignant thyroid nodules US	695 patients	CNN	Accuracy: 80.3% Sensitivity: 80.6% Specificity: 80.1%
Liang et al., 2020 [9]	Classification Benign/malignant thyroid nodules US	221 patients	CNN	Accuracy: 75% Sensitivity: 84.9% Specificity: 69%
Buda et al., 2019 [57]	Nodule detection Predict malignancy Risk level stratification	1230 patients	CNN	Accuracy: N/A Sensitivity: 87% Specificity: 52%
Ko et al., 2019 [54]	Classification Benign/malignant thyroid nodules US	519 patients	CNN	Accuracy: 87.3% Sensitivity: 90% Specificity: 82%
Park et al., 2019 [22]	Classification Benign/malignant thyroid nodules US	286 patients	CNN	Accuracy: 86% Sensitivity: 91% Specificity: 80%
Wang et al., 2019 [33]	Classification Benign/malignant thyroid nodules US	276 patients	CNN	Accuracy: 90.3% Sensitivity: 90.5% Specificity: 89.91%
Li et al., 2018 [55]	Classification Benign/malignant thyroid nodules US	17 627 patients	CNN	Accuracy: 86% Sensitivity: 84% Specificity: 87%
Chi et al., 2017 [58]	Classification Benign/malignant thyroid nodules US	592 patients	CNN	Accuracy: 96.3% Sensitivity: 82.8% Specificity: 99.3%
Ma et al., 2017 [56]	Classification Benign/malignant thyroid nodules US	4782 patients	CNN	Accuracy: 83% Sensitivity: 82.4% Specificity: 84.9%

Abbreviations: US—ultrasound; CNN—convolutional neural network; N/A—not available.

Radiomics is considered a promising method to be encompassed in the pipeline of precision medicine on the basis of specific characteristics of the patient [2]. Whilst the first AI approach to the medical imaging, i.e., CAD system, is focused on the differentiation among benign and malignant thyroid lesions, radiomics extends the analysis to prognosis and response to treatment evaluation [1]. In fact, [42,60,61] implemented radiomics models that analyze the risk stratification and predict the aggressiveness of the thyroid carcinoma with high values of accuracy, i.e., roughly 85 percent. Radiomics analysis has the potential to determine tumor phenotypes or the presence of gene mutations [62,63]. Furthermore, several studies have investigated by means of radiomic features the occurrence of metastases [64] or disease-free survival [65]. It also emerged that radiomics studies aimed at performing classification tasks regarding the nature of thyroid nodules are characterized by minor accuracy with respect to classical ML approach [66]. It is worth pointing out that although radiomics has been applied for several anatomical areas, research concerning thyroid lesions is relatively limited. Studies of interest concerning radiomics applications for thyroid lesions are described in Table 3, organized according to the publication time, in a decreasing order.

Table 3. Radiomics studies.

Study	Description	Cohort	Method	Performance
Park et al., 2021 [60]	Classification: Benign/malignant thyroid nodules 730 features extracted and 66 selected US	1609 patients	ML-based radiomics	Accuracy: 77.8% Sensitivity: 70.6% Specificity: 79.8%
Peng et al., 2021 [67]	Classification Benign/malignant thyroid nodules US	8339 patients	DL-based radiomics	Accuracy: 89.1% Sensitivity: 94.9% Specificity: 81.2%
Wang et al., 2021 [42]	Evaluation of extrathyroidal extension (ETE) in patients with papillary thyroid carcinoma; 479 features extracted; 10 features selected US	132 patients	ML-based radiomics	Accuracy: 83% Sensitivity: 65% Specificity: 74%
Wei et al., 2021 [61]	Evaluation of extrathyroidal extension (ETE) in patients with papillary thyroid carcinoma MRI	102 patients	ML-based radiomics	Accuracy: 79% Sensitivity: 75% Specificity: 80%
Zhao et al., 2021 [21]	Classification Benign/malignant thyroid nodules US	106 patients	ML-based radiomics	Accuracy: 75.5% Sensitivity: 69.7% Specificity: 78.1%
Guo et al., 2020 [64]	Prediction of thyroid cartilage invasion from Laryngeal and hypopharyngeal squamous cell carcinoma; 1029 features extracted; 30 features selected CT images	265 patients	ML-based radiomics	Accuracy: 90% Sensitivity: 80.2% Specificity: 88.3%
Kwon et al., 2020 [62]	Predict the presence or absence of BRAF proto-oncogene, serine/threonine kinase (BRAF) mutation in papillary thyroid cancer US	96 patients	ML-based radiomics	Accuracy: 64.3% Sensitivity: 66.8% Specificity: 61.8%
Wang et al., 2020 [66]	Classification Benign/malignant thyroid nodules US	1040 patients	ML-based radiomics	Accuracy: 66.8% Sensitivity: 51.2% Specificity: 75.8%
Zhou et al., 2020 [40]	Classification Benign/malignant thyroid nodules US	1734 patients	DL-based radiomics	Accuracy: 97% Sensitivity: 89.5% Specificity: 84.1%
Gu et al., 2019 [63]	Evaluating immunohistochemical characteristics in patients with suspected thyroid nodules CT images	103 patients	ML-based radiomics	Accuracy: 84% Sensitivity: 93% Specificity: 73%
Park et al., 2019 [65]	Estimate disease free survival rate in patients with papillary thyroid carcinoma; 730 features extracted and 40 selected US	768 patients	ML-based radiomics	Accuracy: 77% Sensitivity: N/A Specificity: N/A

Abbreviations: US—ultrasound; MRI—magnetic resonance imaging; CT—computer tomography; ML—machine learning; DL—deep learning; N/A—not available.

## 5. Discussion

Medical images provide a comprehensive view of the tumor and its environment, and they can be used to improve the diagnostic accuracy of early lesions, to classify benign from malignant tissues and to define risk and improve therapy [43,68]. Imaging is a non-invasive method and with no risk of the infections or the complications that accompany biopsies [2]. In recent decades, images have been converted into quantitative data and subsequently analyzed with AI tools.

Intratumoral heterogeneity and modifications over time are common features of neoplasms [43]. Samples of tumor acquired through biopsy may fail to represent the variations within the tumor. In addition, AI methods, analyzing the overall image of the lesion, have the potential to capture tumor heterogeneity and could represent an intermediate step between imaging and biopsy [28,36]. Nonetheless, it is worth pointing out that AI systems learn on a case-by-case basis. AI algorithms are implemented considering gold standards of pathological diagnosis that are hard to identify in every patient, due to inter-variability among subjects. Moreover, as it emerged from the overview of the AI methods, the predicting model is developed on the basis of a finite training dataset. Thus, since human tissues are characterized by high heterogeneity and variability inter- and intra-subjects, no finite training set can fully represent the variety of cases that might occur in the clinical practice. Extensive research is still required to improve the generalizability and accuracy of AI-based models. From this perspective, the standalone use of AI applications for diagnosis should be still avoided in the clinical practice. In fact, to this date, several studies [7,20,28,43] recommend that the lesion evaluation should be achieved from a combination between the clinician evaluation and ML or DL outcome. Moreover, it is worth noticing that most AI-based studies focused on thyroid pathologies are performed using retrospectively collected data [9,11,33,40,42,51,55,60–63,65–67]. Conversely, studies that prospectively evaluate AI predictive models concerning thyroid disease diagnosis are limited in the literature [22,41]. In retrospective studies, cohorts are selected among patients with definitive diagnosis achieved mainly through histopathological examination. As highlighted by Wu et al. [69], evaluations should include more prospective studies on medical AI models to reduce risk of overfitting and enhance accuracy of the clinical outcomes.

AI methods are based on the analysis of image features in order to develop predictive models. Differentiating benign and malignant thyroid nodule is mainly achieved from ML-based studies. The most used US features adopted by ML algorithms for thyroid investigations were size, shape, margin, composition echogenicity, as defined by the thyroid imaging reporting and data system (TI-RADS) classification [10,21,22,51]. According to an analysis of the available literature, the TI-RADS approach allows a good discrimination among benign and malignant thyroid nodules. However, the inclusion of additional features, e.g., calcifications, internal content, can represent a factor that improves accuracy [70].

Radiomics studies were applied also to other thyroid pathologies, e.g., extrathyroidal extension (ETE) in patients with papillary thyroid carcinoma (PTC) [42,61], thyroid cartilage invasion from laryngeal and hypopharyngeal squamous cell carcinoma [64]. In these studies, the extracted features derive from morphological, first order statistics, textural and higher order statistics groups. Wang and colleagues [42] highlighted that improvement of ETE diagnosis is achieved when features related to PTC heterogeneity are taken into account. Similarly, in [64] Guo et al. studied thyroid cartilage invasion from laryngeal and hypopharyngeal squamous cell carcinoma and showed that tumor invasiveness can be investigated considering features related to tumor heterogeneity. Furthermore, Kwon et al. [62] highlight that BRAF mutation may be investigated with histogram-based and textural features that reflect echogenicity and heterogeneity of the region of interest, respectively.

Several studies also performed comparison between the performance of AI-based models and that of expert clinicians. The available data in literature mostly report that the performance of DL algorithms is similar to that of healthcare professionals. As discussed by [20,67], AI applications may improve the accuracy of thyroid diagnosis diseases, especially for junior radiologists. In fact, interpretation of medical images highly depends on the experience level of clinicians. For instance, for junior radiologists the sensitivity is reported in a range between 40 percent and 100 percent while the specificity spans between 50 percent and 100 percent. It was observed that the use of AI algorithms to achieve a second opinion on the characterization of thyroid lesions can improve the accuracy of junior radiologists from roughly 82 percent to 87 percent [67]. Moreover, Peng and co-workers [67] highlighted that taking into account the outcomes of AI as a second opinion has reduced fine needle aspiration procedures by 27 percent and the number of missed malignancies of roughly 2 percent.

Furthermore, the experience level of the clinicians has an important impact also on the performance of the AI-based methods. The input data of the AI algorithms is the ROI selected by the expert. It is commonly accepted that image acquisition and segmentation are critical subprocesses due to inter-operator variability. Recent studies [8,28] suggest that semi- or fully automated methods could improve algorithm performance, but currently the manual segmentation performed by experts continues to be the main method adopted. For instance, most of the ML-based studies applied to the thyroid are performed considering a manual segmentation of the ROI [21,41]. In addition, the ML-based investigations reported in [10,22] have introduced a semi-automatic method that is characterized by an initial automatic selection of a box region and subsequently by a manual contouring performed by expert clinicians. Conversely, the studies that applied DL algorithms to thyroid imaging considered a manual selected box around the region under investigation [9,11,52,54]. Furthermore, it is worth pointing out that radiomics studies are based on a manual contouring along the borders of the thyroid tumor [60–62] or slightly within the borders of the tumor to avoid artifacts [64].

To date, most studies highlight that the main limitation of AI algorithms is the reduced dataset used for predictive model development and validation. Ideally, independent training and validation datasets, composed of data images achieved with different US equipment and from multiple centers, i.e., multicenter training cohorts, allow to optimally develop the predicting model, avoiding overfitting and enhancing generalizability and model performance [67].

For instance, in radiomics studies, Gilies and coworkers [43] provide an empirical rule concerning the size of the dataset in order to avoid overfitting. It is suggested that almost 10–15 patients are needed for each examined radiomic feature. Thus, also features selection represents a crucial step during the evaluation.

AI methods represent a powerful approach that in future may assist clinicians in diagnostic decisions [22,71], while combined with other “-omic” data as occur in radiomics analysis may improve the risk factor analysis for personalized estimation of disease-free survival. As mentioned, AI methods could be also applied to contribute to treatment planning. For instance, radiomics combined with other clinical parameters may help to predict which patients are likely to have a satisfactory response to emerging therapies as high-intensity focused ultrasound (HIFU), that allows the thermal tissue treatment and the consequent reduction in thyroid nodule volume by directing energy inside the target zone with non-invasive instruments [72–74].

Several efforts are performed to increase the availability of open access database of labeled medical images that will help to train the predictive models developed with AI techniques. However, pitfalls and limitations associated with the AI approach should be considered, especially related to the difficulty to achieve a generalizable model in order to ensure optimal application for each patient.

With regard to the application of the AI in the daily practice of the clinical medicine, beyond the hype around these technologies, the financial investment is pouring and



brand-new products started flowing into the market. As of early 2020, there were 64 FDA-approved AI-ML medical device and algorithms, many of which are already integrated into clinical care. Remarkably, 21 were related to Radiology [75]. Nonetheless, recent literature reviews report that the impact is still minimal as the majority of the AI-ML studies are retrospective in nature, deviate from existing reporting standards and often outline proof-of-concept approach [76].

From the pure clinical standpoint, all these findings should be interpreted according to the routine clinical practice. In fact, US is recognized as the most relevant imaging procedure for the assessment of thyroid nodule and almost all thyroid patients are managed according to US features of their thyroid gland. This worldwide diffused approach is based on the high sensitivity and specificity of US in discriminating malignant from benign thyroid lesions. Further improvement of US performance by AI remains however desirable [77,78]. In addition, a not negligible number of thyroid goiters are incidentally discovered during other imaging evaluations (i.e., CT, MR, PET/CT) of patients with non-thyroid indication [79]. While the performance of these imaging procedures is poor or suboptimal to identify malignant and benign nodules among adrenal thyroid incidentalomas, a significant effort should be made in the future to improve their capability to initially select patients requiring an urgent or not endocrinological evaluation combined with in-office US examination.

## 6. Conclusions

The evaluation of images has a central role in the clinical workflow. It is worth highlighting that image interpretation requires deductive reasoning, using knowledge of pathological processes, integration from prior examination and investigations and consultation with other physicians. To date, AI techniques can be an integral part of the procedure, but cannot emulate the overall process.

A further approach to improve the assessment of medical images can be represented by the integration of AI-based models with mixed reality tools. The authors retain that in-depth analysis should be performed to analyze the potential of mixed reality within the diagnostic workflow.

**Author Contributions:** Conceptualization, F.B. and P.T.; methodology, F.B., A.P., L.A., A.G., L.R., F.M. and P.T.; software, A.P.; validation, F.B., A.P., L.A., A.G., L.R., F.M. and P.T.; formal analysis, F.B. and A.P.; investigation, F.B., A.P., L.A., A.G., L.R. and P.T.; resources, P.T.; data curation, F.B. and A.P.; writing—original draft preparation, F.B. and A.P.; writing—review and editing, F.B., A.P., L.A., A.G., L.R., F.M. and P.T.; visualization, F.B. and A.P.; supervision, F.B., F.M. and P.T.; project administration, P.T.; funding acquisition, P.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **2012**, *48*, 441–446. [[CrossRef](#)] [[PubMed](#)]
2. Aerts, H.J.W.L.; Velazquez, E.R.; Leijenaar, R.T.H.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **2014**, *5*, 4006. [[CrossRef](#)] [[PubMed](#)]
3. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **2016**, *278*, 563–577. [[CrossRef](#)]
4. *Machine Learning in Radiation Oncology*; El Naqa, I.; Li, R.; Murphy, M. (Eds.) Springer: Cham, Switzerland, 2015. [[CrossRef](#)]
5. Lohmann, P.; Bousabarah, K.; Hoevels, M.; Treuer, H. Radiomics in radiation oncology—basics, methods, and limitations. *Strahlenther. Onkol.* **2020**, *196*, 848–855. [[CrossRef](#)] [[PubMed](#)]
6. Frix, A.-N.; Cousin, F.; Refaee, T.; Bottari, F.; Vaidyanathan, A.; Desir, C.; Vos, W.; Walsh, S.; Occhipinti, M.; Lovinfosse, P.; et al. Radiomics in Lung Diseases Imaging: State-of-the-Art for Clinicians. *Pers. Med.* **2021**, *11*, 602. [[CrossRef](#)] [[PubMed](#)]

7. Castiglioni, I.; Rundo, L.; Codari, M.; Di Leo, G.; Salvatore, C.; Interlenghi, M.; Gallivanone, F.; Cozzi, A.; D'Amico, N.C.; Sardanelli, F. AI applications to medical images: From machine learning to deep learning. *Phys. Med.* **2021**, *83*, 9–24. [[CrossRef](#)]
8. Iqbal, M.J.; Javed, Z.; Sadia, H.; Qureshi, I.A.; Irshad, A.; Ahmed, R.; Malik, K.; Raza, S.; Abbas, A.; Pezzani, R.; et al. Clinical applications of artificial intelligence and machine learning in cancer diagnosis: Looking into the future. *Cancer Cell Int.* **2021**, *21*, 1–11. [[CrossRef](#)]
9. Liang, X.; Yu, J.; Liao, J.; Chen, Z. Convolutional Neural Network for Breast and Thyroid Nodules Diagnosis in Ultrasound Imaging. *BioMed Res. Int.* **2020**, *2020*, 1763803. [[CrossRef](#)]
10. Chang, Y.; Paul, A.K.; Kim, N.; Baek, J.H.; Choi, Y.J.; Ha, E.J.; Lee, K.D.; Lee, H.S.; Shin, D.; Kim, N. Computer-aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: A comparison with radiologist-based assessments. *Med. Phys.* **2016**, *43*, 554–567. [[CrossRef](#)]
11. Jin, Z.; Zhu, Y.; Zhang, S.; Xie, F.; Zhang, M.; Zhang, Y.; Tian, X.; Zhang, J.; Luo, Y.; Cao, J. Ultrasound Computer-Aided Diagnosis (CAD) Based on the Thyroid Imaging Reporting and Data System (TI-RADS) to Distinguish Benign from Malignant Thyroid Nodules and the Diagnostic Performance of Radiologists with Different Diagnostic Experience. *Med. Sci. Monit.* **2020**, *26*, e918452. [[CrossRef](#)]
12. Fujita, H. AI-based computer-aided diagnosis (AI-CAD): The latest review to read first. *Radiol. Phys. Technol.* **2020**, *13*, 6–19. [[CrossRef](#)] [[PubMed](#)]
13. Parmar, C.; Grossmann, P.; Bussink, J.; Lambin, P.; Aerts, H.J.W.L. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci. Rep.* **2015**, *5*, 13087. [[CrossRef](#)] [[PubMed](#)]
14. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* **2020**, *295*, 328–338. [[CrossRef](#)] [[PubMed](#)]
15. McCarthy, J.J.; Minsky, M.L.; Rochester, N. Artificial Intelligence. Research Laboratory of Electronics (RLE) at the Massachusetts Institute of Technology (MIT). 1959. Available online: <https://dspace.mit.edu/handle/1721.1/52263> (accessed on 3 March 2010).
16. McCarthy, J.; Minsky, M.L.; Rochester, N.; Shannon, C.E. A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Mag.* **2006**, *27*, 12.
17. Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; Wang, Y. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* **2017**, *2*, 230–243. [[CrossRef](#)] [[PubMed](#)]
18. Bera, K.; Schalper, K.A.; Rimm, D.L.; Velcheti, V.; Madabhushi, A. Artificial intelligence in digital pathology—New tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **2019**, *16*, 703–715. [[CrossRef](#)]
19. Wernick, M.N.; Yang, Y.; Brankov, J.G.; Yourganov, G.; Strother, S. Machine Learning in Medical Imaging. *IEEE Signal Process. Mag.* **2010**, *27*, 25–38. [[CrossRef](#)] [[PubMed](#)]
20. Erickson, B.J.; Korfiatis, P.; Akkus, Z.; Kline, T.L. Machine Learning for Medical Imaging. *RadioGraphics* **2017**, *37*, 505–515. [[CrossRef](#)]
21. Zhao, C.-K.; Ren, T.-T.; Yin, Y.-F.; Shi, H.; Wang, H.-X.; Zhou, B.-Y.; Wang, X.-R.; Li, X.; Zhang, Y.-F.; Liu, C.; et al. A Comparative Analysis of Two Machine Learning-Based Diagnostic Patterns with Thyroid Imaging Reporting and Data System for Thyroid Nodules: Diagnostic Performance and Unnecessary Biopsy Rate. *Thyroid* **2021**, *31*, 470–481. [[CrossRef](#)]
22. Park, V.; Han, K.; Seong, Y.K.; Park, M.H.; Kim, E.-K.; Moon, H.J.; Yoon, J.H.; Kwak, J.Y. Diagnosis of Thyroid Nodules: Performance of a Deep Learning Convolutional Neural Network Model vs. Radiologists. *Sci. Rep.* **2019**, *9*, 1–9. [[CrossRef](#)]
23. Cui, S.; Tseng, H.; Pakela, J.; Haken, R.K.T.; El Naqa, I. Introduction to machine and deep learning for medical physicists. *Med. Phys.* **2020**, *47*, e127–e147. [[CrossRef](#)]
24. Forghani, R.; Savadjiev, P.; Chatterjee, A.; Muthukrishnan, N.; Reinhold, C.; Forghani, B. Radiomics and Artificial Intelligence for Biomarker and Prediction Model Development in Oncology. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 995–1008. [[CrossRef](#)] [[PubMed](#)]
25. Guorong, W.; Dinggang, S.; Mert, R.S. *Machine Learning and Medical Imaging*; Academic Press: London, UK, 2016. [[CrossRef](#)]
26. El-Naqa, I.; Yang, Y.; Wernick, M.N.; Galatsanos, N.P.; Nishikawa, R. A support vector machine approach for detection of microcalcifications. *IEEE Trans. Med. Imaging* **2002**, *21*, 1552–1563. [[CrossRef](#)] [[PubMed](#)]
27. Hosny, A.; Parmar, C.; Quackenbush, J.; Schwartz, L.H.; Aerts, H. Artificial intelligence in radiology. *Nat. Rev. Cancer* **2018**, *18*, 500–510. [[CrossRef](#)] [[PubMed](#)]
28. Chartrand, G.; Cheng, P.M.; Vorontsov, E.; Drozdal, M.; Turcotte, S.; Pal, C.J.; Kadoury, S.; Tang, A. Deep Learning: A Primer for Radiologists. *RadioGraphics* **2017**, *37*, 2113–2131. [[CrossRef](#)]
29. Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D.S.W.; Karthikesalingam, A.; King, D.; Ashrafin, H.; Darzi, A. Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. *NPJ Digit. Med.* **2021**, *4*, 65. [[CrossRef](#)]
30. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [[CrossRef](#)] [[PubMed](#)]
31. Lundervold, A.S.; Lundervold, A. An overview of deep learning in medical imaging focusing on MRI. *Z Med. Phys.* **2019**, *29*, 102–127. [[CrossRef](#)]
32. Mazurowski, M.A.; Buda, M.; Saha, A.; Bashir, M.R. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *J. Magn. Reson. Imaging* **2019**, *49*, 939–954. [[CrossRef](#)]
33. Wang, L.; Yang, S.; Yang, S.; Zhao, C.; Tian, G.; Gao, Y.; Chen, Y.; Lu, Y. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. *World J. Surg. Oncol.* **2019**, *17*, 1–9. [[CrossRef](#)] [[PubMed](#)]

34. Liu, S.; Wang, Y.; Yang, X.; Lei, B.; Liu, L.; Li, S.X.; Ni, D.; Wang, T. Deep Learning in Medical Ultrasound Analysis: A Review. *Engineering* **2019**, *5*, 261–275. [[CrossRef](#)]
35. Erickson, B.J.; Korfiatis, P.; Kline, T.L.; Akkus, Z.; Philbrick, K.; Weston, A.D. Deep Learning in Radiology: Does One Size Fit All? *J. Am. Coll. Radiol.* **2018**, *15*, 521–526. [[CrossRef](#)]
36. Lambin, P.; Leijenaar, R.T.H.; Deist, T.M.; Peerlings, J.; de Jong, E.E.C.; van Timmeren, J.; Sanduleanu, S.; Larue, R.T.H.M.; Even, A.J.G.; Jochems, A.; et al. Radiomics: The bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **2017**, *14*, 749–762. [[CrossRef](#)]
37. Avanzo, M.; Stancanello, J.; El Naqa, I. Beyond imaging: The promise of radiomics. *Phys. Med.* **2017**, *38*, 122–139. [[CrossRef](#)] [[PubMed](#)]
38. Avanzo, M.; Wei, L.; Stancanello, J.; Vallières, M.; Rao, A.; Morin, O.; Mattonen, S.A.; El Naqa, I. Machine and deep learning methods for radiomics. *Med. Phys.* **2020**, *47*, e185–e202. [[CrossRef](#)] [[PubMed](#)]
39. Tseng, H.-H.; Wei, L.; Cui, S.; Luo, Y.; Haken, R.K.T.; El Naqa, I. Machine Learning and Imaging Informatics in Oncology. *Oncology* **2020**, *98*, 344–362. [[CrossRef](#)] [[PubMed](#)]
40. Zhou, H.; Jin, Y.; Dai, L.; Zhang, M.; Qiu, Y.; Wang, K.; Tian, J.; Zheng, J. Differential Diagnosis of Benign and Malignant Thyroid Nodules Using Deep Learning Radiomics of Thyroid Ultrasound Images. *Eur. J. Radiol.* **2020**, *127*, 108992. [[CrossRef](#)]
41. Yoo, Y.J.; Ha, E.J.; Cho, Y.J.; Kim, H.L.; Han, M.; Kang, S.Y. Computer-Aided Diagnosis of Thyroid Nodules via Ultrasonography: Initial Clinical Experience. *Korean J. Radiol.* **2018**, *19*, 665–672. [[CrossRef](#)] [[PubMed](#)]
42. Wang, X.; Agyekum, E.A.; Ren, Y.; Zhang, J.; Zhang, Q.; Sun, H.; Zhang, G.; Xu, F.; Bo, X.; Lv, W.; et al. A Radiomic Nomogram for the Ultrasound-Based Evaluation of Extrathyroidal Extension in Papillary Thyroid Carcinoma. *Front. Oncol.* **2021**, *11*, 625646. [[CrossRef](#)]
43. Gillies, R.J.; Schabath, M.B. Radiomics Improves Cancer Screening and Early Detection. *Cancer Epidemiol. Biomark. Prev.* **2020**, *29*, 2556–2567. [[CrossRef](#)]
44. Mayerhoefer, M.E.; Materka, A.; Langs, G.; Häggström, I.; Szczypiński, P.; Gibbs, P.; Cook, G. Introduction to Radiomics. *J. Nucl. Med.* **2020**, *61*, 488–495. [[CrossRef](#)]
45. Tunali, I.; Gillies, R.J.; Schabath, M.B. Application of Radiomics and Artificial Intelligence for Lung Cancer Precision Medicine. *Cold Spring Harb. Perspect. Med.* **2021**, *11*, a039537. [[CrossRef](#)] [[PubMed](#)]
46. Cao, Y.; Zhong, X.; Diao, W.; Mu, J.; Cheng, Y.; Jia, Z. Radiomics in Differentiated Thyroid Cancer and Nodules: Explorations; Application; and Limitations. *Cancers* **2021**, *13*, 2436. [[CrossRef](#)] [[PubMed](#)]
47. Araneo, R.; Bini, F.; Rinaldi, A.; Notargiacomo, A.; Pea, M.; Celozzi, S. Thermal-electric model for piezoelectric ZnO nanowires. *Nanotechnology* **2015**, *26*, 265402. [[CrossRef](#)] [[PubMed](#)]
48. Scorza, A.; Lupi, G.; Sciuto, S.A.; Bini, F.; Marinozzi, F. A novel approach to a phantom based method for maximum depth of penetration measurement in diagnostic ultrasound: A preliminary study. In Proceedings of the 2015 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Turin, Italy, 7–9 May 2015; pp. 369–374. [[CrossRef](#)]
49. Marinozzi, F.; Branca, F.P.; Bini, F.; Scorza, A. Calibration procedure for performance evaluation of clinical Pulsed Doppler Systems. *Measurement* **2012**, *45*, 1334–1342. [[CrossRef](#)]
50. Shen, Y.-T.; Chen, L.; Yue, W.-W.; Xu, H.-X. Artificial intelligence in ultrasound. *Eur. J. Radiol.* **2021**, *139*. [[CrossRef](#)]
51. Zhang, B.; Tian, J.; Pei, S.; Chen, Y.; He, X.; Dong, Y.; Zhang, L.; Mo, X.; Huang, W.; Cong, S.; et al. Machine Learning-Assisted System for Thyroid Nodule Diagnosis. *Thyroid* **2019**, *29*, 858–867. [[CrossRef](#)]
52. Wu, G.G.; Lv, W.Z.; Yin, R.; Xu, J.W.; Yan, Y.J.; Chen, R.X.; Wang, J.Y.; Zhang, B.; Cui, X.W.; Dietrich, C.F. Deep Learning Based on ACR TI-RADS Can Improve the Differential Diagnosis of Thyroid Nodules. *Front. Oncol.* **2021**, *11*, 575166. [[CrossRef](#)]
53. Koh, J.; Lee, E.; Han, K.; Kim, E.-K.; Son, E.J.; Sohn, Y.-M.; Seo, M.; Kwon, M.-R.; Yoon, J.H.; Lee, J.H.; et al. Diagnosis of thyroid nodules on ultrasonography by a deep convolutional neural network. *Sci. Rep.* **2020**, *10*, 1–9. [[CrossRef](#)]
54. Ko, S.Y.; Lee, J.H.; Yoon, J.H.; Na, H.; Hong, E.; Han, K.; Jung, I.; Kim, E.K.; Moon, H.J.; Park, V.Y.; et al. Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound. *Head Neck* **2019**, *41*, 885–891. [[CrossRef](#)]
55. Li, X.; Zhang, S.; Zhang, Q.; Wei, X.; Pan, Y.; Zhao, J.; Xin, X.; Qin, C.; Wang, X.; Li, J.; et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: A retrospective; multicohort; diagnostic study. *Lancet Oncol.* **2019**, *20*, 193–201. [[CrossRef](#)]
56. Ma, J.; Wu, F.; Zhu, J.; Xu, D.; Kong, D. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. *Ultrasonics* **2017**, *73*, 221–230. [[CrossRef](#)] [[PubMed](#)]
57. Buda, M.; Wildman-Tobriner, B.; Hoang, J.K.; Thayer, D.; Tessler, F.N.; Middleton, W.D.; Mazurowski, M.A. Management of Thyroid Nodules Seen on US Images: Deep Learning May Match Performance of Radiologists. *Radiology* **2019**, *292*, 695–701. [[CrossRef](#)] [[PubMed](#)]
58. Chi, J.; Walia, E.; Babyn, P.; Wang, J.; Groot, G.; Eramian, M. Thyroid Nodule Classification in Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network. *J. Digit. Imaging.* **2017**, *30*, 477–486. [[CrossRef](#)] [[PubMed](#)]
59. Kim, G.R.; Lee, E.; Kim, H.R.; Yoon, J.H.; Park, V.Y.; Kwak, J.Y. Convolutional Neural Network to Stratify the Malignancy Risk of Thyroid Nodules: Diagnostic Performance Compared with the American College of Radiology Thyroid Imaging Reporting and Data System Implemented by Experienced Radiologists. *AJNR Am. J. Neuroradiol.* **2021**, *42*, 1513–1519. [[CrossRef](#)] [[PubMed](#)]

60. Park, V.Y.; Lee, E.; Lee, H.S.; Kim, H.J.; Yoon, J.; Son, J.; Song, K.; Moon, H.J.; Yoon, J.H.; Kim, G.R.; et al. Combining radiomics with ultrasound-based risk stratification systems for thyroid nodules: An approach for improving performance. *Eur. Radiol.* **2021**, *31*, 2405–2413. [[CrossRef](#)] [[PubMed](#)]
61. Wei, R.; Wang, H.; Wang, L.; Hu, W.; Sun, X.; Dai, Z.; Zhu, J.; Li, H.; Ge, Y.; Song, B. Radiomics based on multiparametric MRI for extrathyroidal extension feature prediction in papillary thyroid cancer. *BMC Med. Imaging* **2021**, *21*, 20. [[CrossRef](#)] [[PubMed](#)]
62. Kwon, M.-R.; Shin, J.; Park, H.; Cho, H.; Hahn, S.; Park, K. Radiomics Study of Thyroid Ultrasound for Predicting BRAF Mutation in Papillary Thyroid Carcinoma: Preliminary Results. *Am. J. Neuroradiol.* **2020**, *41*, 700–705. [[CrossRef](#)]
63. Gu, J.; Zhu, J.; Qiu, Q.; Wang, Y.; Bai, T.; Yin, Y. Prediction of Immunohistochemistry of Suspected Thyroid Nodules by Use of Machine Learning-Based Radiomics. *AJR Am. J. Roentgenol.* **2019**, *213*, 1348–1357. [[CrossRef](#)]
64. Guo, R.; Guo, J.; Zhang, L.; Qu, X.; Dai, S.; Peng, R.; Chong, V.F.H.; Xian, J. CT-based radiomics features in the prediction of thyroid cartilage invasion from laryngeal and hypopharyngeal squamous cell carcinoma. *Cancer Imaging* **2020**, *20*, 81. [[CrossRef](#)]
65. Park, V.; Han, K.; Lee, E.; Kim, E.-K.; Moon, H.J.; Yoon, J.H.; Kwak, J.Y. Association Between Radiomics Signature and Disease-Free Survival in Conventional Papillary Thyroid Carcinoma. *Sci. Rep.* **2019**, *9*, 1–7. [[CrossRef](#)]
66. Wang, Y.; Yue, W.; Li, X.; Liu, S.; Guo, L.; Xu, H.; Zhang, H.; Yang, G. Comparison Study of Radiomics and Deep Learning-Based Methods for Thyroid Nodules Classification Using Ultrasound Images. *IEEE Access* **2020**, *8*, 52010–52017. [[CrossRef](#)]
67. Peng, S.; Liu, Y.; Lv, W.; Liu, L.; Zhou, Q.; Yang, H.; Ren, J.; Liu, G.; Wang, X.; Zhang, X.; et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: A multicentre diagnostic study. *Lancet Digit. Health* **2021**, *3*, e250–e259. [[CrossRef](#)]
68. Trimboli, P.; Bini, F.; Andrioli, M.; Giovanella, L.; Thorel, M.F.; Ceriani, L.; Valabrega, S.; Lenzi, A.; Drudi, F.M.; Marinozzi, F.; et al. Analysis of tissue surrounding thyroid nodules by ultrasound digital images. *Endocrine* **2015**, *48*, 434–438. [[CrossRef](#)] [[PubMed](#)]
69. Wu, E.; Wu, K.; Daneshjou, R.; Ouyang, D.; Ho, D.E.; Zou, J. How medical AI devices are evaluated: Limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **2021**, *27*, 582–584. [[CrossRef](#)] [[PubMed](#)]
70. Verburg, F.; Reiners, C. Sonographic diagnosis of thyroid cancer with support of AI. *Nat. Rev. Endocrinol.* **2019**, *15*, 319–321. [[CrossRef](#)]
71. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [[CrossRef](#)]
72. Bini, F.; Trimboli, P.; Marinozzi, F.; Giovanella, L. Treatment of benign thyroid nodules by high intensity focused ultrasound (HIFU) at different acoustic powers: A study on in-silico phantom. *Endocrine* **2018**, *59*, 506–509. [[CrossRef](#)]
73. Trimboli, P.; Bini, F.; Baek, J.H.; Marinozzi, F.; Giovanella, L. High intensity focused ultrasounds (HIFU) therapy for benign thyroid nodules without anesthesia or sedation. *Endocrine* **2018**, *61*, 210–215. [[CrossRef](#)]
74. Giovanella, L.; Piccardo, A.; Pezzoli, C.; Bini, F.; Ricci, R.; Ruberto, T.; Trimboli, P. Comparison of High Intensity Focused Ultrasound and radioiodine for treating toxic Thyroid nodules. *Clin. Endocrinol.* **2018**, *89*, 219–225. [[CrossRef](#)]
75. Benjamins, S.; Dhunoo, P.; Mesko, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. *NPJ Digit. Med.* **2020**, *3*, 118. [[CrossRef](#)]
76. Ben-Israel, D.; Jacobs, W.B.; Casha, S.; Lang, S.; Ryu, W.H.A.; de Lotbiniere-Bassett, M.; Cadotte, D.W. The impact of machine learning on patient care: A systematic review. *Artif. Intell. Med.* **2020**, *103*, 101785. [[CrossRef](#)]
77. Russ, G.; Trimboli, P.; Buffet, C. The New Era of TIRADSs to Stratify the Risk of Malignancy of Thyroid Nodules: Strengths, Weaknesses and Pitfalls. *Cancers* **2021**, *13*, 4316. [[CrossRef](#)] [[PubMed](#)]
78. Trimboli, P. Ultrasound: The Extension of Our Hands to Improve the Management of Thyroid Patients. *Cancers* **2021**, *13*, 567. [[CrossRef](#)] [[PubMed](#)]
79. Scappaticcio, L.; Piccardo, A.; Treglia, G.; Poller, D.N.; Trimboli, P. The dilemma of 18F-FDG PET/CT thyroid incidentaloma: What we should expect from FNA. A systematic review and meta-analysis. *Endocrine* **2021**, *73*, 540–549. [[CrossRef](#)] [[PubMed](#)]