Knowledge and Information Systems (2023) 65:3881–3921
https://doi.org/10.1007/s10115-023-01894-7

REVIEW

Computer-aided diagnosis systems: a comparative study of classical machine learning versus deep learning-based approaches

Ramzi Guetari[1] · Helmi Ayari[1] · Houneida Sakly[2]

Received: 23 September 2022 / Revised: 23 April 2023 / Accepted: 25 April 2023 /
Published online: 24 May 2023

Abstract
The diagnostic phase of the treatment process is essential for patient guidance and follow-up. The accuracy and effectiveness of this phase can determine the life or death of a patient. For the same symptoms, different doctors may come up with different diagnoses whose treatments may, instead of curing a patient, be fatal. Machine learning (ML) brings new solutions to healthcare professionals to save time and optimize the appropriate diagnosis. ML is a data analysis method that automates the creation of analytical models and promotes predictive data. There are several ML models and algorithms that rely on features extracted from, for example, a patient's medical images to indicate whether a tumor is benign or malignant. The models differ in the way they operate and the method used to extract the discriminative features of the tumor. In this article, we review different ML models for tumor classification and COVID-19 infection to evaluate the different works. The computer-aided diagnosis (CAD) systems, which we referred to as classical, are based on accurate feature identification, usually performed manually or with other ML techniques that are not involved in classification. The deep learning-based CAD systems automatically perform the identification and extraction of discriminative features. The results show that the two types of DAC have quite close performances but the use of one or the other type depends on the datasets. Indeed, manual feature extraction is necessary when the size of the dataset is small; otherwise, deep learning is used.

Keywords Machine learning · Deep learning · Computer-aided diagnosis system (CAD) · Feature extraction · Convolutional neural network · Tumor classification

✉ Ramzi Guetari
ramzi.guetari@ept.ucar.tn

Helmi Ayari
helmi.ayari@ept.rnu.tn

Houneida Sakly
houneida.sakly@esiee.fr

[1] SERCOM Laboratory, Polytechnic School of Tunisia, University of Carthage, PO Box 743, La Marsa 2078, Tunisia

[2] RIADI Laboratory, National School of Computer Sciences, University of Manouba, Manouba 2010, Tunisia

1 Introduction

Medical imaging is becoming one of the most important techniques for the detection and treatment of pathology, because of the numerous advantages it presents as well as its progressions and ranks that are steadily increasing. The various medical imaging modalities allow the acquisition of information on tissues and/or certain functions while minimizing the invasive manipulations for the patient and also guide the practitioner in the treatment procedure.

There are two essential steps in interpreting the diagnostic information from medical imaging: the first one is the identification of patterns, called "forms" in the lexicon of image processing, while the second one is the association of these models with a suspected diagnosis. The success of these two steps is strongly related to the clinician's skills and is always subject to his judgment. It is well known that two practitioners, having the same level of expertise, often give more or less different interpretations of the same imaging data. In addition, the tasks of identifying structures are tedious for a human operator, especially when it comes to 3D acquisitions.

In addition to the subjectivity of each practitioner's diagnosis, the human perception, especially the vision system, has its own limitations and may be unable to recognize certain relevant information. For example, in traditional medicine, the processing of information provided by medical magnetic resonance imaging (MRI) datasets usually relies on the principles of "expert's opinion" and the specialist's experience in diagnosis and interpretation. For the evaluation of most neuroradiology problems, this method, e.g., research, description, and evaluation of focal lesions, is completely sufficient in daily clinical usage. This becomes more difficult in case of diseases only detectable by subtle volumetric changes or by signal disturbances. In this case, human perception is often unable to detect information relevant to the correct diagnosis.

In order to make the diagnostic task as objective as possible, different systems for analyzing and processing various types of medical images have been developed in recent years. These systems allow especially to characterize human tissues, suspicious regions, pathologies, etc., and to produce an automatic or semi-automatic diagnosis [1]. These are computer-aided diagnosis (CAD) systems. CAD systems use the most advanced image processing techniques combined with AI and ML to establish the most reliable diagnosis possible. They have the advantage of going beyond the limits of human memory and being able to detect pathological changes that cannot be detected by physicians. They are intended to improve the quality of patient care [2].

There are two main types of CAD systems developed by the scientific community: (i) systems that operate based on features extracted by human experts and (ii) systems that learn by themselves the features they rely on to discriminate pathologies. There are also hybrid systems that rely on both manually extracted features from human experts and automatically learned features, but these are not the focus of this work.

This paper reviews recent works in the field of CAD considering the two main identified categories (Classical CAD and DL). The classification of thyroid and pulmonary nodules as well as the COVID-19 diseases was interested to study and evaluate these CAD systems. The paper highlighted the advantages and disadvantages of each of the two types of CAD, to identify when it is appropriate to use the first or the second type, and to conclude with perspectives for the improvement of computer-aided diagnosis systems.

The paper is organized as follows: Section 1 is the introduction; Sect. 2 presents the strategy followed to collect the research materials that made this work possible. Section 3 covers

Table 1 Selected digital libraries

Digital libraries	URL
SpringerLink	http://link.springer.com
IEEE Xplore	http://ieeexplore.ieee.org
Google Scholar	http://scholar.google.com
ACM Digital Library	http://dl.acm.org

the fundamentals of supervised machine learning models. Section 4 provides an overview of the medical decision support system. Image feature extraction is covered in Sect. 5. Section 6 presents CADs based on handcrafted features and the corresponding ML models. Section 7 discusses deep learning-based CADs with as well as the most popular models used in designing CADs. Section 8 is a discussion about the models as well as different works, and finally, Sect. 9 concludes the whole work.

2 Search strategy

The bibliographic references used in this work were collected from four digital libraries: SpringerLink, IEEE Xplore, Google Scholar and ACM Digital Library. The summary is presented in Table 1. The choice of these digital libraries was guided by the ease of access and the availability of a significant number of documents related to our research topic, which made the task of selecting articles quite complex. Our choices were guided by certain questions that we asked ourselves as well as by the date of publication. We have privileged the most recent ones.

A total of 122 references were finally selected to conduct this research. The large majority of the publications we have selected are papers published in journals (75%) and conferences (22%). We have considered only a few articles (3%) from workshops, books or other sources.

Our investigations were mainly guided by the following research questions:

- RQ1: What are the research works carried out on CAD systems in the medical field?
- RQ2: What are the machine learning models used in CAD systems?
- RQ3: What are the metrics used for the evaluation of CAD systems?
- RQ4: On which features are CAD systems based to establish a diagnosis?
- RQ5: How are these characteristics identified?
- RQ6: What technical problems do CAD systems face?

3 Supervised machine learning

Today, when one talks about AI, most people think of ML. In reality, ML is one of the areas of AI that focuses on the ability to learn how to perform a task based on observation of an environment. It is the development, analysis, and implementation of methods that allow a machine to evolve through a learning process and thus perform tasks that are difficult or impossible to accomplish by more traditional algorithmic means. ML algorithms allow computers to train on data inputs and use statistical analysis to produce values that fall within a specific range. For this reason, ML facilitates the use of computers in building models from sample data to automate decision-making processes based on the input data [3]. Applied

to classification problems, it is about learning to distinguish the different elements of this environment through the observation of examples.

Supervised learning consists of having a learning set, whose objects are labeled. For example, a label could be a clearly identified class to which an object belongs. A test set composed of objects whose membership classes are unknown is also needed in supervised learning. The goal of learning from labeled examples is to construct a function that best approximates an unknown function that generates random, independent, and identically distributed data, of which one has only a few examples. A learning system based on examples is composed of three main modules:

- A generator that generates random data called input vectors. These vectors are independent and identically distributed according to an unknown probability distribution $P(x)$.
- A supervisor that associates for each input vector x an output y (the class) following an equally unknown probability distribution $P(x, y)$.
- A learning machine that implements a family of functions that must produce for each input vector x an output $\hat{y}$ as close as possible to the output y of the supervisor.

3.1 Supervised machine learning models

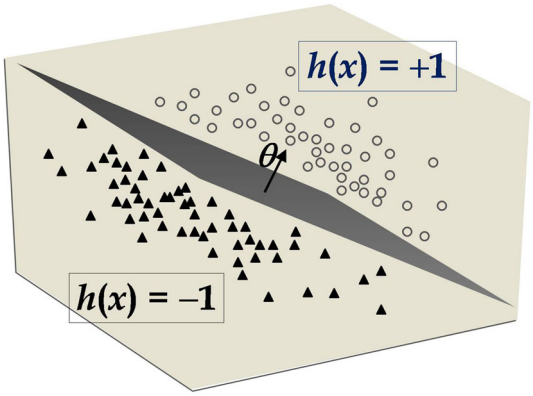
As applications of ML, classifiers can be built to distinguish an edible mushroom from a poisonous one, to predict whether it will rain tomorrow based on current weather conditions, or to produce a medical diagnosis (determining whether a patient has a disease based on a set of symptoms). The considerable challenge in ML is to develop an algorithm that can efficiently solve classification problems of different types. The quality of a learning algorithm thus depends on its ability to build a classifier that generalizes the observed phenomenon. ML approaches are continuously developed but some stand out for their efficiency and the applications in which they are implemented. Among these algorithms, the K-nearest neighbor method (KNN) [4, 5], artificial neural networks (ANN) [6], decision trees and random forests [7], and genetic algorithms [8, 9] can be mentioned.

Generally speaking, a ML-based system searches for regularities in an available dataset to extract the appropriate knowledge, without a predefined model. This method is now being developed thanks to the increase in computer power and the accumulation of huge quantities of data [10]. The regularities that a ML model looks for are translated into a features map. These features, depending on the learning model, can be defined by experts or automatically identified by the model itself.

Recently, the emergence of techniques based on ANNs has revitalized AI in general and the design of CAD tools in particular. ANNs are highly connected networks of elementary processors (Neurons) organized in layers and operating in parallel. Each neural network has an input layer, an output layer and a number of intermediate layers called hidden layers. Each artificial neuron receives a variable number of inputs from the upstream neurons and produces a unique output. Each of these inputs is associated with a weight w representative of the strength of the connection. The output of an artificial neuron then branches to feed a variable number of downstream neurons.

An artificial neuron computes the value a which is the weighted sum of the inputs according to the following expression: $a = \sum w_i \cdot x_i + b_i$, where w_i is the weight and b_i is a bias. The connections between the neurons that make up the network describe the topology of the model; it can be any, but most often it is possible to distinguish a certain regularity. The collective behavior of a set of neurons allows the emergence of higher-order functions with

Fig. 1 Linear separation



respect to the elementary function of a single neuron. The setting of the values of the weights which link the layers in the network is what constitutes the training of the system. Supervised learning involves associating the inputs x_i of a neural network with well-defined outputs y_i during the learning phase. The idea is therefore to provide much-known input–output pairs (x_i, y_i) of data and to vary the weights according to these samples in order to reduce the error between the output of the network and the desired output. In other words, the goal is to progressively improve the neural network to predict the correct y given x.

3.2 The learning process

Supervised learning requires a set of data $S_n = \{(x^{(i)}, y^{(i)}), i = 1, \ldots, n\}$ called the training set. The value n indicates the cardinal of the set S, $x^{(i)}$ is the ith object to be classified and $y^{(i)}$ is the label associated with $x^{(i)}$. S_n is the information provided to the learning algorithm to train it. A classifier h is an application defined by: $h : \Omega \longrightarrow I$ where h is the classifier, Ω is the vector space where the feature vectors representing the objects to be classified are defined and I is the set of labels. I can for example be the set $\{-1, +1\}$, where -1 could mean "malignant tumor" and $+1$ "benign tumor".

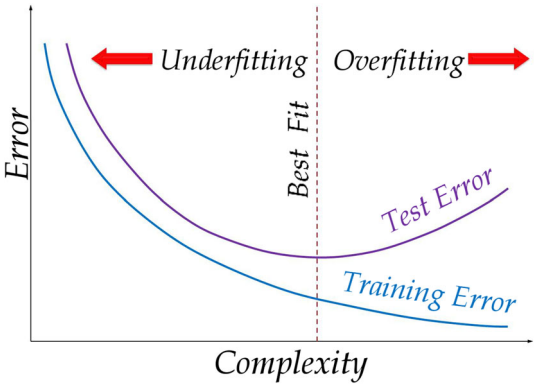
The task of the learning algorithm given an object $x^{(i)}$ is to find the corresponding label $y^{(i)}$. Considering the tumor classification task, the classifier divides the space into two halves: one labeled benign tumor, and the other malignant tumor. The classifier determines the hyperplane (known as the decision boundary, Fig. 1) defined by the parameters θ and θ_0 such that $\theta \cdot x + \theta_0 = 0$, $\theta \in \Omega$ and $\theta_0 \in \mathbb{R}$.

Then the evaluation of the quality of the classifier is needed with respect to the training set S_n by computing the training error E_n as defined in (1):

$$E_n(h) = \frac{1}{n} \sum_{i=1}^{n} \left[h\left(x^{(i)}\right) \neq y^{(i)} \right] \tag{1}$$

$$h\left(x^{(i)}\right) \neq y^{(i)} = \begin{cases} 1, \text{ if error} \\ 0, \text{ otherwise.} \end{cases} \tag{2}$$

Fig. 2 Classifier optimization



Let's rewrite the formula of the learning error with respect to θ and θ_0:

$$E_n(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^{n} \left[y^{(i)} \cdot \left(\theta \cdot x^{(i)} + \theta_0 \right) \leq 0 \right] \tag{3}$$

In the learning phase, it is essential to minimize the learning error or even reduce it to zero. It is quite possible that there are several classifiers that reduce the learning error to zero. Which one to choose in this case? The classifier that also minimizes the test error should be chosen. This error is defined in the same way as the learning error but applies to a dataset outside the training data and whose labels are unknown. When the classifier choice problem is transformed into an optimization problem, two aspects should be taken into account: the "*Loss*" and the "*Generalization*". If the learning and test errors are both large then the model has not learned to classify objects very well and this phenomenon is called "*underfitting*". If the learning error is minimal but the test error is large then the classifier has learned the training set with its noise, it is therefore not possible to generalize it; this is called "*overfitting*". Optimizing the "*Loss*" consists mainly in minimizing the learning error and the "*Regularization*" consists in making the model general so that the test error is also minimal as shown in Fig. 2. The following section will discuss the feed-forward neural network and how to be trained.

3.3 Training feed-forward neural networks

Feed-forward neural networks with multiple hidden layers are complex models that attempt to capture the representation of examples to the output unit in a way that facilitates the actual prediction task. It is this part of learning the representation that makes the learning problem difficult. It turns out that a simple (stochastic) gradient descent algorithm is capable of finding a good solution to the resolution parameters, provided that the model is given some overcapacity. It is therefore necessary to find an algorithm able to evaluate this gradient, i.e., the derivative of the loss with respect to the parameters. The solution to this problem is an algorithm known as the backpropagation algorithm [11–13].

Parameterized models are functions that depend on input data and parameters (weights and biases) that, when combined with the inputs, produce a result. Since the input data cannot be modified, the quality of the result produced by the model depends on the parameters that can be trained. At the initialization of a neural network, the weights and biases of the model

receive random values, which makes it globally useless and inappropriate for the resolution of a classification problem for example. It is therefore necessary to train the model so that it can adjust these parameters in an efficient way allowing it to provide an adequate solution to the problem at hand. The output of a neural network is a value $\hat{y} = f(x, w)$, x is the input, w is the set of trainable parameters, and $\hat{y}$ is the result.

In supervised learning, this output is subject to a cost function $C(y, \hat{y})$, which compares the real output y with the output of the model $\hat{y}$. ML mainly consists in optimizing the cost function, usually by minimizing them by gradient-based methods. A gradient-based method is a method/algorithm that finds the minima of a function, assuming that one can easily compute the gradient of this function which is assumed to be continuous and differentiable almost everywhere. The update of the parameters is done according to the expression (4) for the weights and (5) for the biases (where w_{ij}^{l} is the weight applied between neuron i of layer l and neuron j of layer $l+1$, b_i^l is the bias and η is the learning rate.).

$$w_{ij}^{l} = w_{ij}^{l} - \eta \cdot \frac{\partial C\left(y, \hat{y}\right)}{\partial w_{ij}^{l}} \tag{4}$$

$$b_{i}^{l} = b_{i}^{l} - \eta \cdot \frac{\partial C\left(y, \hat{y}\right)}{\partial b_{i}^{l}} \tag{5}$$

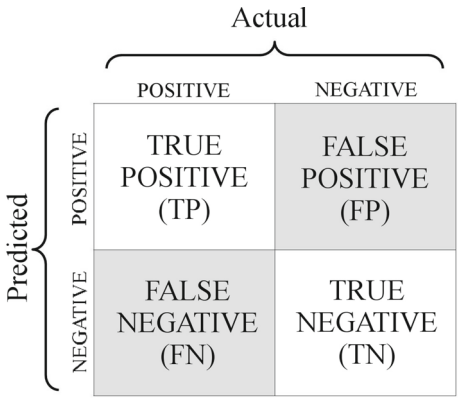
If the neural network is not trained, there is a very high probability that y and $\hat{y}$ are very different and the error is given by the cost function $C(y, \hat{y})$. Since the expected value y for the neurons of the last layer is known, it is quite easy to know to what extent the weights associated with each neuron of the penultimate layer contributed to this error. And if the error of the penultimate layer is known, the error of the previous layer could be computed, and so on, until the first one. The weights and biases are thus adjusted from the last layer of the network to the first by calculating the appropriate partial derivative and applying the adjustments formulated in (4) and (5). This process of adjusting the weights and biases of the last to the first layer is called backpropagation. The feed-forward / backpropagation process is repeated a number of times to minimize the cost function and allow the network to learn. The following sections will discuss the most used evaluation metrics for machine learning models.

3.4 Performance evaluation of classification algorithms models

Many choices of ML models are available to us. The nature of the problem to be solved helps to guide the appropriate choice. For example, a classification algorithm cannot be applied to a regression problem. Defining a ML model is important, but it must be relevant. In supervised learning, the objective is to create models that generalize the learned situation, i.e., when faced with new data, the machine is able to develop predictive models. It is, therefore, necessary to know how to evaluate any learning algorithm on its dataset, avoiding as much as possible the bias of over-fitting. A rigorous evaluation of the performance of an algorithm is an essential step for its deployment.

In a classification problem, and this is the case in CAD systems, the accuracy is mostly used to evaluate the performance of the model, but this is not enough to really judge the model. That said, a model may appear to perform well for one performance indicator and quite the opposite for another. It is usually necessary to combine two or more indicators to judge the performance of a model. In the following, some of the most commonly used performance indicators to evaluate a ML model are presented.

Fig. 3 The Confusion Matrix



3.4.1 Confusion matrix

A confusion matrix (Fig. 3) is a tool to measure the performance of a ML model by checking how often its predictions are correct compared to reality in classification problems. Correct and incorrect predictions are highlighted and divided by class. The results are then compared with the real values. True positives (*TP*) are cases where the model has predicted a positive case and the actual class of the object is also positive. Similarly, true negatives (*TN*) are cases where the model has predicted a negative case and the real class of the object is also negative. False positive (*FP*) case is referred to when the real class of the object is negative and the predicted class is positive. The false negative (*FN*) case is a case where the real class of the object is positive and the predicted class is negative.

3.4.2 Accuracy

The accuracy is the ratio of the number of correct predictions to the total number of data in the dataset as expressed by the formula (6).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

Accuracy seems to be a very good indicator of performance and it is if the data is balanced in the dataset. Otherwise, it becomes almost useless to measure the performance of a ML model and a very bad model can give good accuracy.

3.4.3 Precision

The precision is the proportion of true positive results in the set of cases having been predicted positive for a certain class; it is expressed by the formula (7).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

3.4.4 Recall (or sensitivity) and specificity

Recall is the fraction of correctly predicted TP cases for a class relative to all cases actually belonging to that class. This is expressed by the formula (8)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

The specificity is estimated by the proportion of tn predicted by the model compared to all the cases listed as negative in the dataset; it corresponds to the expression (9).

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{9}$$

3.4.5 F1-score

Precision and recall are complementary measures to the accuracy allowing to better qualify the performance of a ML model and particularly in cases where the classes are not distributed in a balanced way. The idea here is to combine the two measures into one to evaluate a ML model which gives us a metric called F1-score or Sørensen–Dice coefficient or Dice Similarity Coefficient (DSC). It is the harmonic mean between precision and recall expressed by the formula (10).

$$F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

3.4.6 Receiver operating characteristic (ROC)

The ROC curve is a graphical representation of the relationship between the sensitivity and the specificity. The ordinate represents the sensitivity and the abscissa corresponds to the quantity $(1 - Specificity)$. It illustrates the performance of a binary classification system when its discrimination threshold varies. Although it is possible to construct the ROC curve manually, the use of specialized calculation software is recommended because it requires a precise procedure which will not be detailed in this paper.

3.4.7 Area under the ROC curve (AUC)

The area under the curve (abbreviated as AUC) is the measure of the area under the plot of a mathematical function. Formally, this value corresponds to the integral of the function studied. It provides an aggregate measure of performance for all possible classification thresholds. The value of the AUC varies from 0 to 1. A model whose predictions are 100% false has an AUC of 0.0; a model whose predictions are 100% correct has an AUC of 1.0. The area under the ROC curve can be interpreted as the probability that, among two randomly selected subjects, e.g., one with a disease and one without, the value of the marker is higher for the diseased than for the non-diseased. Therefore, an AUC of 0.5 (50%) indicates that the marker is non-informative. An increase in AUC indicates improved discriminatory abilities, with a maximum of 1.0 (100%).

4 Medical decision support systems

As medical knowledge has continued to grow, physicians can no longer master all the medical knowledge needed to recognize diseases or to determine the best therapeutic treatment. They often use external sources of information, traditionally colleagues, books, and scientific literature, to find the information they need in order to establish their diagnoses and to define a therapy procedure. Nevertheless, in spite of the large amount of easily accessible resource materials available online, the help from colleagues, etc., finding a solution to a given patient's problem is a difficult task.

Very early on, the qualities of the computer (memory, speed, computing power) emerged as potential solutions to this problem, and computer-aided medical decision support systems (CAMDS) were developed. CAMDS are software systems that "provide clinicians, staff, patients, and other individuals with knowledge and person-specific information, intelligently filtered and presented at appropriate times, to enhance health and health care" [14]. These systems are designed to assist the practitioner in his reasoning with a view to identifying a diagnosis and choosing the appropriate therapy. They are not intended to replace the decision-maker by offering "ready-made" solutions.

CAD systems are a category of CAMDS that help radiologists to make an accurate diagnosis with the benefit of information generated by computerized image analysis. The concept of CAD systems has emerged over the last twenty years to provide some answers to problems related to the limits of human perception capabilities and thus help radiologists to effectively evaluate medical images to detect lesions and pathological changes at an early stage [15]. The purpose of a CADs is to classify images according to a set of defined classes. CADs are automated systems that provide radiologists with a second opinion that is devoid of human subjectivity and assist them in the diagnosis by providing them with a mapping of suspicious parts in the medical images. CAD systems are an interdisciplinary technology, they use a variety of disciplines and concepts such as AI, computer vision, image processing, etc. Their principle is the extraction of the characteristics of a medical image and the design of a prediction and detection model from a learning base. This generally empirical model allows quantifying the probability that an area of a test image is pathological.

The analysis and processing of medical images is a sensitive area where the cost of the error can be fatal. The goal of a CAD is to increase the sensitivity of detection and reduce the rate of false positives and false negatives. A CAD system requires a thorough analysis of the digitized medical images in order to extract the most significant characteristics that lead to the automatic or semi-automatic diagnosis. The objective of a CAD is to **classify** the analyzed images according to well-defined criteria. For example, it is a question of classifying suspicious tumor zones of human tissue into benign or malignant tumors. Feature extraction is a stage of image processing that often goes hand in hand with classification. Indeed, to establish a rule of classification (supervised or not), one generally bases oneself on a set of numerical criteria describing the object or the phenomenon observed. In practice and depending on the context, two types of characteristics can be extracted:

- Generic descriptors that do not necessarily have a physical or biological interpretation (SIFT [16], SURF [17], oriented gradient histograms [18], shape background [19], bags of words [20], etc.),
- Descriptors having a physical meaning. Typically, in the biomedical field, these may be morphological characteristics describing the observed objects (cell size, vessel thickness, etc.).

The extraction and selection of features from a medical image is therefore an essential task that requires sophisticated image processing methods. The performance of a CAD system depends on this selection since these features are used for image classification and thus for diagnosis [21].

5 Image feature extraction

In computer science, an image is a 2D digital signal and can be interpreted as a 2D matrix of brightness or color values. Such an interpretation makes it possible to represent an image in different ways. Indeed, the images can be seen as a set of pixels, characterized by their variable light intensity, representing a concrete or abstract scene. They can be modeled as a function f whose values $f(x, y)$ vary according to the spatial coordinates. An image is also a 2D signal, or a set of textures or patterns arranged in a specific manner to describe the scene.

The methods of analysis of images have a common basis: the global or local characterization of the images that would be dealt with. This characterization involves the analysis and processing of the image that focuses on low-level attributes or features (also called descriptors). These descriptors may be at the pixels level [22, 23] (histograms, co-occurrence matrices), shapes (geometric attributes), or regions (patterns, textures). A low-level feature is a set of values that are extracted directly from the image. The extraction of the low-level characteristics represents the first abstraction with respect to the raw image, the goal is that this characteristic must be a distinctive criterion with respect to the visual entities that one seeks to characterize. These low-level descriptors can be used for the entire image as they can be used for part of the image.

Extracting visual descriptors on the entire image (global descriptors) reduces the amount of required processing and the cost of searching for the most similar images. Algorithms allowing the extraction of global descriptors are simple, fast, and not memory-consuming. These descriptors also are more robust to noise that can affect the signal. However, the global approach does not allow an efficient search of objects (in a broad sense) in the image. Moreover, the attributes of small objects are embedded in a single global descriptor of the image. Conversely, the descriptors extracted from a part of the image (local descriptors) are efficient but require a lot of processing. Local features, also known as structural features, as opposed to global features, have locality property and focus specifically on the various areas of the image. Local descriptors can be (i) regions of the image obtained either by segmentation of the entire image or by search for regions of interest, or (ii) points of interest. This consists in treating the shape or the morphology of the image's characteristics. This analysis is based on nonlinear image processing techniques. The use of appropriate morphological operations makes it possible to simplify the images while preserving the essential shape of the geometric structures (loop cavitations, connected objects, etc.) and eliminating the noise.

The principle of the morphological analysis is to detect in the image elements such as contour, edge density, edge direction, number of horizontal or vertical lines, co-termination and parallel lines [24], spatial/layout relations [25], number of endpoints, number of cross points, horizontal curves at top or bottom, etc. The purpose of contour analysis is to look for areas of an image with remarkable local properties, such as abrupt and local changes in intensity, texture changes, specific points, and so on. These techniques are often used in preprocessing to solve more difficult problems (object detection, interpretation, registration, tracking of objects in a video, etc.). These techniques have, however, some problems such as

the noise in the image that can affect the accuracy of the detection and contour coding that closely depends on the starting point of the tracking algorithm.

Local descriptors allow for keeping localized information in the image, thus avoiding certain details being lost in the rest of the image. Although this type of representation brings undeniable advantages such as the coding of certain knowledge on the structure or the supply of certain knowledge on the type of components constituting this object, it also has very notable disadvantages. Indeed, they are sensitive to different variations and the quantity of observations produced is very large, which implies the processing of a large volume of data.

Other descriptors may be extracted from an image using other types of analysis. The statistical analysis of an image consists of representing objects by statistical measures. The statistical analysis techniques range from the simplest, such as the distribution of pixels in the various regions of the image and the histograms, to the more complex ones like principal component analysis. Metric characteristics are physical measurements of the image such as the height, width, and ratio of these two measurements. Adaptive features are extracted directly from the image and require a learning step. Indeed, the system uses a representation close to the original image and builds the characteristics extractor, and optimizes it. These descriptors are not very relevant in the field of CAD.

The purpose of the feature extraction process is not only to reduce the amount of input information but also to obtain the most discriminating information for the recognition step. This is an essential phase of a recognition system. However, some feature extraction techniques are accompanied by an irreversible loss of information. As a result, a compromise between the quality and the quantity of information is essential. In the ideal case, the descriptors must be robust to the different variations, especially geometric transformations (rotation, translation, etc.), the variation of the viewing angles, the variation of scale, etc.

The choice of extracted features is often guided by the desire for invariance or robustness with respect to transformations of the image. A descriptor is qualified as good if it can describe the content of the image with great variance of scale and interoperability, and if it can distinguish any type of objects. A survey of existing methods of extracting feature descriptors is presented in [26]. These descriptors are subdivided into four classes: shape descriptors, color descriptors, texture descriptors, and motion descriptors.

When analyzing medical images, which is the subject of this paper, the decision about the nature of a tumor depends on a number of visual features that can be supplemented by additional examinations. A CAD system requires the association of symptoms (physical characteristics of thyroid nodules for example) with features extracted from the images. This association is performed either manually by specialists or automatically by the CAD system itself. In both cases, the learning step is the pillar on which the diagnostic assistance will be based. The success of this step depends on the images that will be used for training (dataset), the selected discrimination characteristics, the accuracy of the image annotation and, in the case of manual annotation, the objectivity of this process.

6 CADs based on handcrafted features

The purpose of a CAD based on handcrafted features is to classify images according to a set of defined classes. Object classification consists of establishing a procedure that associates a class (among other classes) with a given data. The supervised classification consists of having a learning set, whose objects and classes to which they belong are clearly identified, and a test set, composed of objects whose membership classes are unknown.

Image classification is an important task in the field of computer vision, object recognition, and ML. The use of low-level image descriptors and the bag-of-words model are at the heart of today's image classification systems. Most image classification environments have three steps: (i) the extraction of the low-level descriptors in images, (ii) the creation of a vocabulary of visual-words, and (iii) the learning of the model of images' classes, during which the procedure (classifier) to associate an object to a class is defined. Then follows the phase of classification (or prediction) during which the classification rules established (or learned) during the previous step are used to deduce the class of an object that is, a priori, unknown.

In the first step of extracting the low-level descriptors [27]. The most common choices in recent methods are SIFT [16], SURF [17], or DSIFT [28]. The second step is the creation of the visual vocabulary, the usual choice for this step is the use of a K-means algorithm and the creation of a bag-of-words. The third step is the training of the classifier, many systems often choose support vector machines with linear or nonlinear kernels. Most of these systems are then evaluated on small datasets that fit well in central memory, such as Caltech-101 [29], Caltech-256 [30], or PASCAL VOC [31].

As an example, the thyroid ultrasound image has different features that can be extracted visually. These characteristics are interpreted by radiologists and the results of the diagnosis depend on the experience of the latter. Several studies have attempted to deal with the characteristics of thyroid nodules and have identified some features as suggestive ones. Based on the latter, a Thyroid Imaging Reporting and Data System (TIRADS) [32] has been developed to classify thyroid nodules based on their probabilities of malignancy. The creators of TIRADS develop a grouping of ultrasound signs into ten original aspects related to categories TIRADS 1 to 5 of increasing probability of malignancy. The TIRADS system is based on 4 signs strongly suggestive of malignancy, namely (i) nodule thicker than wide, (ii) nodule with irregular borders, (iii) microcalcifications and (iv) very hypoechoic nodule. The TIRADS classification makes it possible to select the nodules for which an ultrasound-guided cytopunction is indicated (mandatory examination before any surgical indication). This system is, however, not always reliable, since accuracy is often based on the personal experience of radiologists.

6.1 Significant visual attributes

Classification, whether supervised or not, is based on rules established in relation to a set of numerical criteria of an image describing the object or phenomenon. These criteria represent visual characteristics and their extraction consists of mathematical transformations calculated on the pixels of a digital image. The classification rules are similarity rules applied to the criteria representing an image or a part of an image. The following presents some of the most important features used in medical image classification.

6.1.1 Texture characteristics

The texture is a local spatial variation in pixel intensities and orientation [33]. It is a very important characteristic used for object detection in the region of interest (ROI) in the image, and its goal is to get distinctive features that describe the inner intensity distribution of different classes in images by texture analysis. A number of structural, spectral, and statistical model-based techniques are used to compute texture feature descriptors such as gray level

histogram, run-length matrix, gray level cooccurrence matrix (GLCM), and contourlet transform coefficient. GLCM is a widely used statistical texture analysis method that analyzes regional textures using second-order texture features. It can reflect the position distribution characteristics between pixels with the same gray level, which is the basis for calculating the texture feature, as well as the comprehensive information of the gray level of the image in the adjacent direction [34], the adjacent interval, the amplitude of the change, and so on. It also describes the frequency with which different gray-scale value combinations appear in an image and measures the connection between neighbor pixels at four angles (0°, 45°, 90°, 135°) [35]. Given an image I of size $N \times N$ in grayscale, the GLCM matrix is computed by Eq. (11) where the shift operator $(\Delta x, \Delta y)$ indicates a position (ignoring edge effects) that can be applied to any pixel in the image.

$$P(i, j) = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} \begin{cases} 1, \text{if } I(x, y) = i \text{ and } I(x + \Delta x, y + \Delta y) = j \\ 0, \text{otherwise.} \end{cases} \tag{11}$$

A whole series of new descriptors have been derived from the GLCM to highlight the most important parameters of these cooccurrence matrices.

Correlation (12): a measure of the similarity of the image's different gray levels in the row and column directions, where μ and σ are the mean and standard deviation, respectively.

$$\text{Correlation} = \sum_{i,j=0}^{N-1} P(i, j) \frac{(i - \mu)(j - \mu)}{\sigma^2} \tag{12}$$

Energy (13): the energy is a measure of the image's localized change and represents the rate of change in the color/brightness/magnitude of pixels over time and space.

$$\text{Energy} = \sum_{i,j=0}^{N-1} P(i, j)^2 \tag{13}$$

Entropy (14): it is defined in images as the corresponding states of intensity level to which individual pixels can adapt.

$$\text{Entropy} = \sum_{i,j=0}^{N-1} -\log P(i, j) P(i, j) \tag{14}$$

Contrast (15): it is the amount of color or grayscale differentiation that exists between various image features in both analog and digital images.

$$\text{Contrast} = \sum_{i,j=0}^{N-1} P(i, j)(i - j)^2 \tag{15}$$

Homogeneity (16): used to determine the distribution of gray and/or color values within an image.

$$\text{Homogeneity} = \sum_{i,j=0}^{N-1} P(i, j) \frac{P(i, j)}{1 + (i - j)^2} \tag{16}$$

Auto-correlation (17): the degree of similarity between a given time series and a lagged version of itself over successive time intervals.

$$\text{Auto-correlation} = \sum_{i,j=1}^{1} P(i,j) \tag{17}$$

Cluster-shade (18): helps assess perceptual concepts of uniformity by measuring the asymmetry of the GLCM matrix

$$\text{Sha} = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\left(i + j - \mu_x - \mu_y\right)^3 \times P(i,j) \tag{18}$$

Dissimilarity (19): represents the distance between pairs of pixels in the ROI.

$$\text{Dissimilarity} = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}|i-j| \times P(i,j) \tag{19}$$

6.1.2 Shape characteristics

The shape is a significant visual aspect in image processing. It is defined by a set of characteristics that allow one to identify a specific form of a representative class. Some entity extraction approaches, such as the Zernike moment [36, 37], domain of shape change, and one-dimensional function, are employed to define a shape S (Area A defined by Eq. 20, Perimeter P specified by Eq. 21, Circularity C expressed by Eq. 22, etc.) [38]. The link between the area of a shape and the area of a circle with the same perimeter is represented by Circularity [39]. These features describe the irregularities used by radiologists to make a diagnosis.

$$A(S) = \sum\sum I(x,y)\Delta A \tag{20}$$

where ΔA symbolizes the size of each pixel and $I(x,y) = 1$ if the pixel belongs to the region, $(x,y) \in S$, and 0 else.

$$P(S) = \int \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} \tag{21}$$

where x_i and y_i symbolizes the coordinates of pixel i.

$$C(S) = \frac{4\pi \times A(S)}{P(S)^2} \tag{22}$$

6.1.3 Intensity characteristics

This feature allows us to characterize the intensity of different regions of an image. One of the approaches widely used is the histogram approach [38] that allows the analysis by intensity values on the whole or parts of an image. Intensity is described by parameters such as the average of the pixel values (23), the variance (24) (and the standard deviation (25)), and the skewness (26).

$$\mu = \sum_{i,j=1}^{n} P(i,j) \tag{23}$$

$$\sigma^2 = \sum_{i,j=1}^{n} P(i,j)(i-\mu)^2 \tag{24}$$

$$\sigma = \sqrt{\sum_{i,j=1}^{n} P(i,j)(i-\mu)^2} \tag{25}$$

$$\text{Skewness} = \sum_{i,j=1}^{n} (i-\mu)^3 P(i,j) \tag{26}$$

6.2 Feature detection methods

The richness, diversity, and complexity of the data involved in image segmentation make this feature extraction step necessary. Their theoretical scope covers a wide spectrum, from data analysis methods to the effective minimization of Bayes' error. There are many methods for the detection of various types of features and some of the most popular ones are represented in the following.

6.2.1 The local binary pattern (LBP)

LBP (27) is an alternative and unifying approach to the traditionally divergent statistical and structural models of texture analysis [40]. Due to its computational simplicity, efficiency, and discriminating power, it has been widely used in various applications. It proceeds by labeling the pixels by thresholding the neighborhood of each pixel to obtain a result considered as a binary value [34, 41]. It is robust to monotonic changes in the grayscale that can be caused by illumination variations.

$$\text{LBP}_{P,R} = \sum_{n=0}^{p-1} s(g_p - g_c)2^p \tag{27}$$

where g_c is the gray level intensity value of the center pixel and g_p is the value of the center pixel's equivalent surrounding pixel. The radius of the circular neighborhood is R, and P is the number of neighbors. The indicator function, $s(x)$, is defined as in (28).

$$s(x) = \begin{cases} 1, \text{if } x \geq 0 \\ 0, \text{if } x < 0 \end{cases} \tag{28}$$

6.2.2 Scale invariant feature transform (SIFT)

SIFT is a method, developed in 1999 and very popular in the field of computer vision, to extract features (or points of interest or key points) [42] from an image and compute their descriptors. A descriptor is a vector that describes the neighborhood of the feature to which it is associated [43]. It is used to find the pairs of features that are most similar in two images. To facilitate this matching step, the descriptor must have many invariance properties (rotation, scale, illumination). Thus, the descriptors of two features that are identical except for a geometric or photometric change must be as close as possible. The matching step is then to compare the descriptors. SIFT has 4 majors steps:

- Key point Detection

- Key point Localization
- Orientation Assignment
- Key point Descriptor / Feature matching

The algorithm starts the calculation of the difference of Gaussians (DoG) to detect the local extrema (AKA key points) in the spatial scale. Then, using a threshold value, more precise key point locations are identified. The third step assigns the orientation to describe the key points with image rotation invariance. Finally, in the last step, a set of key point descriptors is generated, each of which generates orientation histograms [43].

6.2.3 Speeded-up robust feature (SURF)

SURF [44] is a local feature detection and descriptor algorithm based on properties similar to those of SIFT, while reducing complexity. The SURF algorithm is performed in two steps [45, 46]: a first step in which the points of interest are identified and a second step allowing the extraction of the descriptors. The detection of the points of interest is performed by approximating the convolution of the Gaussian second-order partial derivatives using box-type filters. This first step also allows the construction and localization of the scale space. In the second step of descriptor extraction, a descriptor vector corresponding to a point of interest is generated. The pixels included in a circular neighborhood around the point of interest are considered and the Haar wavelet responses (HWRs) [47] in the horizontal and vertical directions are calculated. This allows us to obtain the main orientation in an orientation window of size $\pi/3$, which is moved along the given sliding step. A square region R centered on the point of interest is then oriented along the main orientation to allow the descriptor vector to be invariant to image rotation. HWRs are again computed for each pixel in the 4×4 sub-regions that make up the square region R. A four-dimensional vector is then computed for each sub-region from the HWRs of the pixels it contains.

6.2.4 Oriented fast and rotated BRIEF (ORB)

ORB [48] uses features from accelerated segment test (FAST) key point detector and binary BRIEF descriptor. The concept of ORB is to extract less and the best features from an image [49]. Compared with SIFT and SURF, the computational cost of ORB is less and its magnitude is faster than SURF. It detects a large number of key points by using the FAST key point detector and then to find relevant features that have less sensibility to noise and generate better results from the extracted key points it uses a Harris corner detector [50]. Moreover, for the quick key focuses on the direction, ORB uses OFAST and for the BRIEF descriptor with orientation (rotation) angle it uses RBRIEF, and lastly applies orientation to transform the used sample in BRIEF descriptor [49, 51].

6.3 Features' selection

Feature selection is the process of minimizing the data by reducing the redundant, correlated, and irrelevant features of the dataset. It helps in dimensionality reduction by selecting the optimal feature subset that contains the necessary information, which could enhance the learning algorithm speed and improve the performance and the accuracy of the classification [52]. The importance of feature selection is that it conserves the original meanings of original features, which provides superior readability and interpretation for the classification model

[53]. Hereafter, two of the most representative methods for feature selection, namely PCA and LDA, are represented, knowing that other methods have been used for this purpose such as the RF used in [38].

6.3.1 Principal component analysis (PCA)

PCA [54] is a linear dimension reduction technique that employs an orthogonal transformation to transform a set of correlated variables to a set of uncorrelated variables if the dataset is regularly distributed jointly [55], and is widely applied in many different fields. The main concept of PCA is to convert *a*-dimension data into a new *b*-dimension data called principal component where (b≤a). Principal components are new orthogonal *b*-dimensional features that are built from the original *a*-dimensional features. In further depth, all of the freshly constructed PCs are linear functions with rapid reduction of variance that are uncorrelated with one another. The first few PCs maintain the majority of the variation from the original data, whereas the last few PCs with little fluctuation can be useful in outlier detection [56]. PCA looks for a linear combination of variables from which to extract the most variation. After removing this variance, PCA looks for a second linear combination that iteratively explains the greatest proportion of the remaining variation. This method is known as the principal axis method, and it produces orthogonal (uncorrelated) factors [57]. PCA has 4 steps [55, 58]:

- Step 1: Data Normalization.
- Step 2: Calculation of the covariance matrix of the features.
- Step 3: Calculation of the eigenvector and eigenvalue of the covariance matrix.
- Step 4: Projection of the raw data into *k*-dimensional subspace.

6.3.2 Linear discriminant analysis (LDA)

LDA [59] is a dimensionality reduction technique and statistical-based pattern ML classifier widely used in different fields and used to solve many classifications and data visualization problems. The concept of LDA is to by maximize the average between-class scatter and decrease the average within-class scatter. This method uses label information to learn appropriate transformation directions by using the label information to earn a discriminant projection [60]. LDA is used to find a linear transformation that categorizes several classes. It is a simple and robust algorithm [61], and when the datasets are linearly separable, LDA performs exceptionally well in classification tasks. In reality, LDA typically works effectively, especially for low-dimensional issues [62].

6.4 Classification models

This section describes some of the most common machine learning techniques. There are numerous classification tasks that can be performed. Because each algorithm is used to solve a specific problem, each task frequently necessitates a different algorithm. Classification models include support vector machine, random forest, Naive Bayes, and K nearest neighbor.

6.4.1 Support vector machine (SVM)

SVM [63] is a ML model that is commonly used for regression and classification problems. It works by mapping a subset of data into a high-dimensional space, where the data is then split

into two classes by a linear separator (Fig. 1) known as a hyperplane [64]. Note that there can be an infinite number of hyperplanes to choose from. By lowering the distance between the hyperplane focuses and maximizing the margin between the classes, SVM can select the hyperplane that does the best job of classifying the data [65]. The margin is the distance between each class's boundary and its nearest point [66]. These points closest to the boundary are referred to as support vectors [67, 68]. In the linear case [69], The two classes are linearly separable, which means that it is possible to find at least one hyperplane defined by a vector with a bias that is capable of separating the classes with zero error. There are many cases where even the use of the optimal hyperplane will not be effective in linearly separating the data sets. For that reason, SVM can create a nonlinear hypersurface of decisions to classify nonlinearly separable data and transform the input vector to a vector with higher dimension feature space.

6.4.2 Random forest (RF)

RF [70] is a decision tree-based method that performs effectively in a classification task. It is a mixture of tree predictors, with each tree relying on vector values chosen at random and independently. By fitting each tree on a bootstrap sample rather than the original sample, it reduces over-fitting and improves accuracy by averaging prediction across multiple different trees [71]. RF trains the image dataset by infusing randomness into the training of the trees, followed by combining the outputs of each trained tree into a single classifier. The test image is then sent down through each tree until it reaches the leaf node in the second phase. The average of the posteriors across all trees is then used to classify the input image [43]. The fundamental difference between the two is that a random forest can produce more reliable ensemble forecasts than a decision tree [72].

6.4.3 Naive Bayes (NB)

NB [73] is a probabilistic classifier based on the Bayes theorem that describes the probability of an event based on prior knowledge of the conditions that could be associated with the event. This model is a relatively simple and sophisticated classification strategy that is suitable for building and analyzing very large datasets [74]. NB is an extremely capable and scalable method, and its model is simple to construct and can handle massive amounts of data [68]. It uses three steps to predict the class label with the highest posterior probability: The dataset is turned into a frequency table in the first stage. After determining the probabilities, the second step is to create a likelihood table. In the final stage, the posterior probability for each class is calculated using the NB equation shown in expression (29)

$$P\,(a|y) = \frac{P\,(y|a) \cdot P\,(a)}{P\,(y)} \tag{29}$$

where $P\,(a|y)$ indicates the posterior probability of a class, $P(a)$ represents the class prior probability, $P\,(y|a)$ shows the likelihood which is the probability of predictor given class, and $P(y)$ indicates the predictor's prior probability.

6.4.4 K nearest neighbors (KNN)

The KNN [75] model is a non-parametric supervised ML algorithm widely used for pattern classification and regression problems. KNN is a simple and easy classifier to implement, it

does not require a huge memory for storage and has an impressive performance in multi-model classes [43]. It acts as a classifier using two factors, namely the similarity measure or distance function and the chosen k value, with the performance depending on the aforementioned factors. For any new data point, KNN first calculates the distance between all data points and gathers those that are close to it. The algorithm then organizes those closest data points based on their distance from the target data point by employing various distance functions. Furthermore, the next step is to collect a specific number of data points with the shortest distance among all of them and categorize them based on their distance.

6.5 Related works

This literature review is particularly interested in work concerning the diagnosis of thyroid nodules. Several methods were presented to classify thyroid nodules as benign or malignant; these methods extract diverse attributes to generate the feature vector and employ various ML techniques (classifiers). Table 2 includes a summary of the related work mentioned in this paper.

Shankar et al. [69] proposed a model for classifying thyroid data that makes use of optimal feature selection and a kernel-based classifier process. They used Multi-Kernel SVM (MK-SVM) to create classification models. The novelty and goal of this proposed model as feature selection is to improve the performance of the classifying process using improved grey wolf optimizer [76]. The datasets were obtained from the UCI repository. The proposed thyroid classification achieves 97.49% accuracy, 99.05% sensitivity,94.5% specificity, and F1-score of 98.26%.

In [77], the authors used internal characteristics such as content and echogenicity and external characteristics such as margin, shape, and orientation to determine the malignancy of the thyroid nodule along with the CAD system. They first applied an adaptive median filter with by bilateral filter to reduce the noise of the images. Then, active contour and morphology operations were applied for the segmentation of the nodules. After the extraction of geometric and texture features, they used the multi-layer perception to classify the internal characteristics and SVM to classify external characteristics. CaThyDB is the used dataset, and the classification of this technique achieved 100% sensitivity, 95.45% specificity, 97.78% accuracy, and 98.87% F1-score which means that the proposed CAD system can be used by radiologists in classifying thyroid nodules and the perspective of this article is to develop a fully automatic system.

Prochazka et al. [78] designed a CAD that uses only direction-independent features by two-threshold binary decomposition. Thus, during image acquisition, the CAD is not influenced by the orientation and inclination angle of the ultrasound transducer. To calculate direction-independent features, authors applied several algorithms such as histogram analysis and segmentation-based fractal texture analysis algorithm. Axial US images of 40 patients were obtained from their clinic's database system. They also extracted various features such as histogram parameters, fractal dimension, and mean brightness value in different grayscale bands from 40 thyroid nodules (20 malignant and 20 benign). SVM and RF classifiers were used to classify data. Along with the use of leave-one-out cross-validation method, SVM achieved higher accuracy with 94.64%, while random forests achieved 92.42%. Concluding that the proposed system with SVM classifier has better accuracy and is reliable for radiologists to use to diagnose the thyroid nodules.

In [79], two distinct ML techniques, SVM and RF were evaluated for thyroid disorder diagnosis. The experiment was carried out using the Thyroid dataset from the UCI Machine

Table 2 Related literature review

Method	Classifiers	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
Shankar et al. [69]	MK-SVM	97.49	99.05	94.50	98.26
Colakoglu et al. [80]	RF	86.80	85.20	87.90	85.99
Nugroho et al. [77]	SVM	97.78	100.00	95.45	98.87
Ataide et al. [82]	RF	99.33	99.39	99.25	99.35
Aboudi et al. [38]	SVM	93.29	98.45	90.94	95.80
	RF	96.13	93.27	97.04	94.67
Prochazka et al. [78]	SVM	94.64	NA	NA	NA
	RF	92.42	NA	NA	NA
Olatunji et al. [68]	RF	90.91	93.00	100.00	96.00
	SVM	84.09	86.00	92.00	89.00
	NB	81.82	93.00	93.00	93.00
Alyas et al. [81]	RF	94.80	94.80	91.00	94.80
	NB	NA	93.00	78.00	NA
	KNN	NA	59.00	91.00	NA
Shivastuti et al. [79]	SVM	93.00	93.00	NA	91.00
	RF	92.00	92.00	NA	88.00
Rehman et al. [65]	NB	100.00	100.00	100.00	100.00
	LR	100.00	100.00	100.00	100.00
	KNN	97.84	96.00	98.00	97.00
	SVM	86.02	76.00	80.00	84.00
	DT	75.34	67.00	80.00	62.00
Chandel et al. [86]	KNN	93.44	NA	NA	NA
	NB	22.56	NA	NA	NA
Prochazka et al. [83]	RF	95.00	95.00	95.00	95.00
	SVM	91.60	95.00	90.00	93.26

Learning Repository. The accuracy, precision, recall, and F-score of these two ML techniques were compared using four performance metrics. Based on the experimental results shown in Table 2, SVM is found to be superior to random forest for thyroid disorder diagnosis.

Aboudi et al. [38] presented a new CAD system developed to classify thyroid nodules trained with 447 US images of thyroid nodules. Statistical feature extraction methods were used on these images to extract features. The most relevant and non-redundant features were chosen using a feature selection method based on RF and the multi-objective particle swarm optimization algorithm. The nodules were then classified using SVM and RF. The classification performance metrics were evaluated using 10-fold cross-validation. Using the contour-based ROI, their proposed CAD achieved a maximum accuracy of 94.28% for SVM and 96.13% for RF.

In [80], Colakoglu et al. evaluate the diagnostic utility of ML-based quantitative texture analysis for distinguishing benign and malignant thyroid nodules. The RF classifier was used to evaluate 306 quantitative textural aspects of 235 thyroid nodules from a total of 198 individuals. Reproducibility tests and a wrapper approach were used to select features and reduce dimensions. The proposed method was compared with histopathologic or cytopathologic findings as a reference method for diagnostic accuracy, sensitivity, specificity, and AUC. For the results, 284 of 306 initial texture features demonstrated high reproducibility (intraclass correlation $\geq$ 0.80). The RF classifier correctly detected 87 of 102 malignant thyroid nodules and 117 of 133 benign thyroid nodules, for a diagnostic sensitivity of 85.2%, specificity of 87.9%, F1-score of 85.99%, and accuracy of 86.8%. The AUC of the model was 0.92. ML classification can accurately distinguish between benign and malignant thyroid nodules using quantitative textural analysis of thyroid nodules.

In [81], various ML algorithms such as decision tree, RF, KNN, and ANN are applied to a dataset to create a comparative analysis in order to better predict disease based on parameters established from the dataset. The dataset was obtained from the UCI data repository on thyroid disease. It contains 7200 multivariate records. In addition, the dataset has been modified to ensure accurate classification prediction. For a more accurate comparison of the dataset, the classification was performed on both the sampled and unsampled datasets. The authors obtained the highest accuracy for the RF algorithm after manipulating the dataset, as shown in Table 2.

Ataide et al. [82] used the geometric and morphological (G–M) feature extraction techniques and RF model to classify the thyroid nodules from ultrasound images and to reduce the subjectivity in the current diagnostic process. The classification of the thyroid nodules is based on TIRAD guidelines. 27 G–M features were extracted from images and only 11 were considered significant and were selected according to TIRADS guidelines to be evaluated by ML techniques. The open-source Digital Database of Thyroid Ultrasound Images (DDTI) dataset, from the Instituto de Diagnostico Medico, was used for this work. The result of G–M features combined with RF classifier was compared with state-of-the-art methods showing the efficiency of the proposed method by achieving 99.33% accuracy, 99.39% sensitivity, 99.25% specificity, and 99.35% F1-score which are the highest values.

In [68], authors developed ML-based system that can serve as early warning systems by detecting TC at an early stage (pre-symptomatic stage). Furthermore, they aimed to achieve the highest possible accuracy while using the fewest features possible. While there have been previous attempts to use ML to predict thyroid cancer, this is the first attempt to use a Saudi Arabian dataset as well as to target diagnosis in the pre-symptomatic stage (preemptive diagnosis). RF, ANN, SVM, and NB techniques were used in this work, and each was chosen for its unique capabilities. The highest accuracy rate obtained was 90.91%, with 93.00%

sensitivity, 100.00% specificity, and 96.00% F1-score as shown in Table 2 with the RF technique.

In [83], 60 US images of thyroid nodules (20 malignant, 40 benign) were split into small patches of 17 × 17 pixels and used to extract various direction-independent features using two-threshold binary decomposition, a method that decomposes an image into a group of binary images. The features were then used in RF and SVM classifiers to differentiate between benign and malignant nodules. The 10-fold cross-validation approach was used to evaluate the classification. The performance of individual patches was then averaged to classify the complete nodules. The performance of this approach was evaluated by various performance indicators (see Table 2) including the area under the ROC curve (0.971 for RF and 0.965 for SVM) and the F1-score (95% for RF and 93.25% for SVM).

Ur Rehman et al. [65] used a dataset obtained from DHQ Teaching Hospital, Dera Ghazi Khan, Pakistan with which they trained and validated various classifiers including KNN, NB, SVM, decision tree, and logistic regression. In addition to the features selected to qualify the malignant or benign nature of thyroid nodules, they considered three additional features namely pulse rate, body mass index, and blood pressure. The experiment consisted of three iterations; the first iteration did not use feature selection, whereas the second and third iterations used an L1-, L2-based feature selection technique [84, 85]. The experiment was evaluated and analyzed, and many factors such as accuracy, precision, and AUC were considered. The results presented in Table 2 indicate that the classifiers that used L1-based feature selection achieved higher overall accuracy, sensitivity, specificity, and F1 score with the same value compared with those that did not use feature selection and the L2-based feature selection technique.

In [86], the authors experimented with various classical ML models to classify thyroid diseases based on parameters such as TSH, T4U, and goiter. In particular, they applied the KNN and NB algorithms and the Rapid miner tool was used for the experimental study. The accuracy of the KNN algorithm was 93.44%, while that of Naive Bayes was 22.56% which tells us that KNN is an efficient algorithm for this classification, while Naive Bayes seems to be inadequate.

7 CADs based on deep learning

In an article published in the journal "Annals of Oncology" [87] researchers showed that a CAD system based on Google's Inception v4 Convolutional Neural Network (CNN) architecture outperformed experienced dermatologists in detecting skin cancer from a series of digitized images. The system was trained on more than 100,000 dermoscopic images of skin lesions and moles labeled as benign or malignant. The system's performance was then compared to that of 58 dermatologists, including 30 experts. The physicians, on average, correctly identified 87% of melanomas and 71% of non-cancerous lesions. When they obtained larger images, and more detailed information (age and gender, position of the skin lesion), these rates grew to 89% and 76%. The machine detected 95% of melanomas from the first set of images.

Similarly, in a paper published in the journal "Archives of pathology and laboratory medicine" [88], a team of researchers from the Naval Medical Center and DeepMind (Google AI) showed the significant superiority of LYNA (Lymph Node Assistant) CAD performance over that of physicians in diagnosing breast cancer, while doctors on average barely achieve

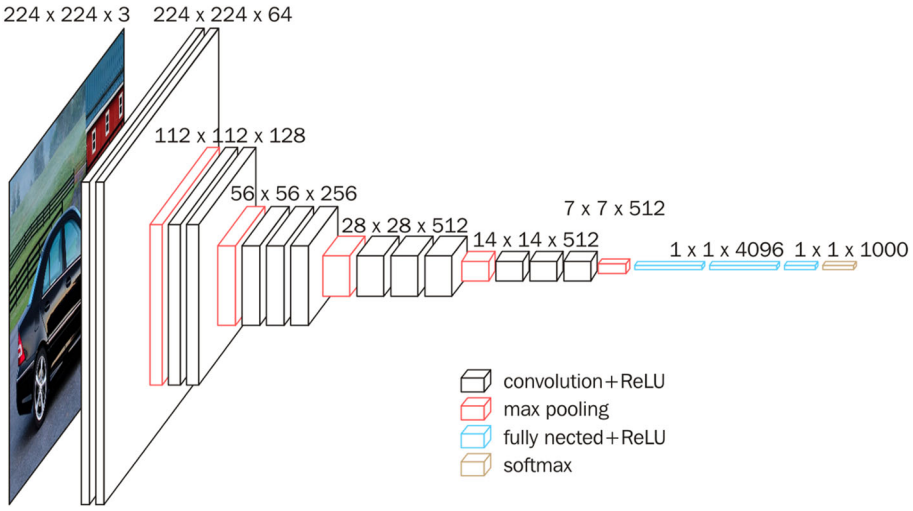


Fig. 4 VGG-16 CNN architecture [90]

an AUC of 81%, LYNA achieved a slide-level AUC of 99% and a tumor-level sensitivity of 91%.

These are only two examples among many others of the capabilities of deep learning in diagnosis assistance. These CAD systems are based on a deep learning model called convolutional neural networks (ConvNet or CNN) are going to be discussed in the following section.

7.1 Convolutional neural networks

Before talking about convolutional neural networks (CNN) it is important to define what deep learning (DL) is. Deep ANNs are distinguished by the existence of many intermediate layers, hence the notion of depth. This feature allows them to solve much more complex ML problems than shallow neural networks. Deep learning refers to ML algorithms that are based on deep ANNs.

CNNs are a particular type of deep ANN. Although capable of solving a multitude of problems, CNNs are best suited for image classification. They came to prominence in 2012 at the annual *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) when the AlexNet model broke all records.

CNNs work in a similar way to traditional supervised learning methods except that they do not need a "manual" feature extraction algorithm, such as SIFT for example. They take care of all the feature extraction and description work during the learning phase. The primary goal of this learning phase is to minimize the classification error in order to optimize both the classifier parameters and the features. Moreover, the specific architecture of CNNs allows to extract features of different complexity, from the simplest to the most sophisticated [89]. The automatic extraction and prioritization of features, which are adapted to the given problem, is one of the strengths of CNNs. The CNN architecture is formed by a stack of distinct layers (Fig. 4; [90]) that transform the input from a given layer into an output to the next one. CNNs have two components: the Hidden layers allowing extracting the relevant features

and the classification part. In hidden layers, the network performs a series of convolutions and pooling operations during which the features are detected. In the classification part, the fully connected layers serve as a classifier on top of these extracted features [3]. In object classification, they assign a probability for an object on the image is what the algorithm predicts it is. The hidden layers of a CNN are a succession of alternations of convolutional layers and grouping layers.

7.1.1 Convolution layer

The CNNs take their name from this layer because it is its fundamental component. It scans the image to locate the presence of a set of features by performing a convolution filtering. The principle of this filtering is to "slide" on the image a window representing the wanted feature (also called filter) and calculate the convolution product between the feature and each portion of the scanned image. This convolution filtering operation is applied to each filter for each image [91]. The result for each pair $(image, filter)$ is a so-called feature map that indicates the presence of features in the image. The filters are not pre-defined but learned by the network during the training phase. This is the strength of the CNNs which are able to determine by themselves the discriminating features of an image, by adapting to the problem at hand.

7.1.2 Pooling layer

A pooling (sub-sampling) layer is usually placed after a convolution layer. It receives as input each of the feature maps produced by the layer that precedes it and applies the pooling operation that consists of reducing the size of the images while preserving their important characteristics. CNNs apply max pooling which consists in splitting the image into regular cells, often square and of small size (2×2), and to keep only the maximum value of each of them. The pooling layer allows reducing the number of parameters which also reduces the computation time in the network. The pooling layer makes the network less sensitive to the position and orientation of the features.

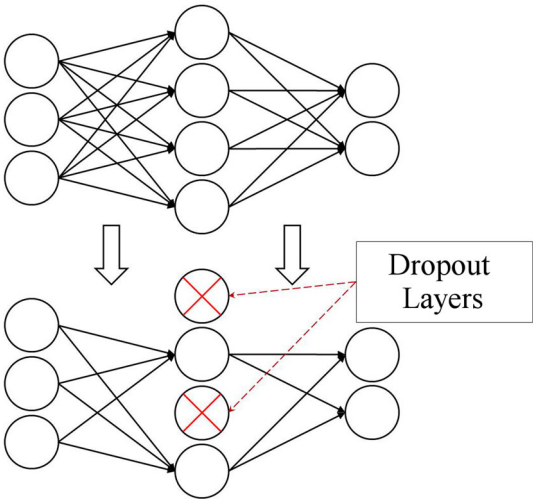
7.1.3 Rectification layer

A Rectified Linear Unit (ReLU for short) is a nonlinear activation function defined by: $f(x) = max(0, x)$. Since this function is linear for positive inputs only, it is called a piecewise linear function or hinge function. Its role is to remove any negative value from the previous filtering operations by replacing it with zero. It establishes a linear relationship between each positive input and the dependent variable.

7.1.4 Fully connected layer

The last layer of a CNN is always fully connected and is used for classification. It receives a vector from the different feature maps computed by the convolution and pooling layers as input and constructs a vector as output whose size is the number of classes in the image classification problem [92]. Each element of the vector indicates the probability for the input image to belong to a class. These probabilities are computed by an activation function, *softmax* most often, which returns a value between 0.0 and 1.0, knowing that the sum of all returned values is equal to 1.0.

Fig. 5 Dropout



The fully connected layer allows us to reconstruct the image from the various previous filterings. It determines the link between the position of the features in the image and a class. Indeed, the input table being the result of the previous layer corresponds to an activation map for a given feature: the high values indicate the location (more or less precise according to the pooling) of this feature in the image. If the location of a feature at a certain place in the image is characteristic of a certain class, then the corresponding value is given a high weight in the table.

7.1.5 Regularization and optimization techniques

Other layers are very often inserted in the CNNs whose goal is to reduce the learning time but especially to avoid over-fitting. During training, a regularization function is used to avoid over-fitting.

Dropout

Over-fitting can be avoided by using a variety of techniques throughout the training phases. Among these techniques, "dropout" is the best known. The *dropout* [93] consists in randomly deleting units and their connections during the learning process (Fig. 5), which allows to generate slightly different models with different neurons configurations at each iteration. The idea is to force the model to operate each neuron individually, as its neighbors can be randomly deactivated at any time. The "early stop approach" is used in some circumstances to protect the system, while it is being trained and verified and to halt the training before the completion of all epochs to prevent the system from becoming over-fitted.

Batch normalization

The normalization is a pre-processing of numerical data to bring them to a common scale without distorting their shape. *Batch normalization* has several beneficial effects that can be mentioned (i) the acceleration of the learning process, (ii) the resolution of the internal covariate shift problem: the input of each layer is distributed around the same mean and standard deviation, and finally (iii) the smoothing of the loss function which, in turn, by optimizing the model parameters, improves the learning speed of the model.

Table 3 Results for the 3 architectures

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
CNN	84.15	83.96	84.32
DNN	82.37	80.66	83.90
SAE	82.59	83.96	81.35

RMSprop
RMSprop [94] is a method of adaptive learning. It was created as a solution to the problem of the dramatically declining learning rate. In order to dispose of history fast after discovering a convex bowl, it uses an exponentially declining average. In reality, it resembles Adadelta and is described in (30) where $E|g^2|_t$ is the averaging factor: $g_t = \nabla_\theta J(\theta)$.

$$\begin{cases} E|g^2|_t = \gamma E|g^2|_{t-1} + (1-\gamma)\, g_t^2 \\ \theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E|g^2|_t + \epsilon}} g_t \end{cases} \tag{30}$$

C-Adam
The Adam optimization algorithm [95] is an extension of stochastic gradient descent that has recently been more widely adopted for deep learning applications in computer vision and natural language processing. It is an optimization algorithm for updating iterative network weights based on training data. The method computes individual adaptive learning rates for different parameters from estimates of the first and second moments of the gradients. Empirical results show that Adam works well in practice and compares favorably to other stochastic optimization methods.

7.2 Related works

CAD systems based on deep learning have been introduced in the fields of medical imaging, overcoming the limitations of the feature extraction step. Indeed, rather than manually extracting features, deep neural networks allow direct use of the input image and avoid the programmer having to assume the task of discovering the best representation of the data. They automatically learn which characteristics are most relevant to a given problem. Table 6 summarizes the related work mentioned in this paper.

7.2.1 Thyroid nodule classification

In [96], Li et al. propose an approach based on the faster R-CNN neural network [97, 98] to which they have made modifications to make it adapted to the detection of ultrasound images. These modifications, according to [96], improved the ability of Faster R-CNN to detect cancerous regions (papillary carcinoma) in thyroid ultrasounds. The basic Faster R-CNN architecture receives as input a 3 channel image of size 224×224. The first convolution layer using 96 filters 7×7 with a stride of 2 in x and y activated by a ReLU layer. This is followed by a max pooling layer using the 2 stripe with 3×3 regions, processed by normalized contrast which gives 96 feature maps 55×55. The following layers 2, 3, 4, 5 perform the same operation as already described. The 6 and 7 layers are fully connected serving to extract the features from the 5 layer which is input as a vector. The output layer is a softmax layer for the classification. The modifications consist of concatenating the 3rd and 5th convolution

layers and adding a spatially constrained layer before the output layer. The concatenation of shallow and deep layers of the CNN allows to detect fuzzy or small areas of tumor and the spatial constraint layer allows extracting the characteristics of the surrounding region in which the cancerous regions reside. The approach was tested on a dataset of ultrasound images taken on 53 men and 247 women aged 10 to 85 years, of which 250 cases were diagnosed with papillary thyroid cancer and 50 cases were diagnosed with a normal thyroid. The dataset contains a total of 4670 images each with one to three cancerous regions. 200 images from this dataset were used for the training of the neural network and 100 for its validation. The results obtained are as follows: accuracy: 91.6%, precision: 93.3%, recall: 93.5%, specificity: 81.3%, F1-score: 93.4%, and AUC: 0.938.

Ma et al. [99] proposed a hybrid method for diagnosing thyroid nodules in ultrasound images, consisting of the use of a cascade of two CNNs architectures pre-trained with the ImageNet dataset and a splitting method, to detect thyroid nodules from these images. The first CNN architecture consists of 15 convolutional layers and two max pooling layers with a window size of 3×3 that follow the first and the third convolution layers, respectively. A padding size of two pixels is used in these two pooling layers. In addition, the parametric rectified linear unit function (PReLu) [100] is used as the activation function. A distribution over the two class labels is generated after the output of the last convolutional layer with a softmax layer. The second CNN architecture consists of four convolution layers, four pooling layers, and two fully connected layers with 64 and 1 outputs respectively. The first two convolutional layers are both followed by the max pooling layers with a window size of 3, a padding of 1, and a stride of 2. The third convolutional layer is followed by a max pooling layer with a stride of 2 and a window size of 2. Finally, a max pooling layer with a stride of 8 and a window size of 8 follows the fourth convolutional layer. The activation function is a rectified linear unit (ReLu). After the output of the second fully connected layer, a softmax layer is used to generate a distribution over the two class labels by minimizing the cross-entropy loss between the predicted labels and the ground truth labels. The two networks are fine tuning and trained separately to obtain feature maps. These maps are merged, and used as input to a Softmax classifier. A 10-fold cross-validation with the same dataset was used to refine the architecture. The performance of the system was evaluated by the area under the ROC curve, the FROC (free-response receiver operating characteristic) analysis [101, 102] and the JAFROC (jackknife alternative free-response receiver operating characteristic) analysis [103, 104]. Experimental results validated on 15,000 images show that this cascade architecture is very efficient for the detection of thyroid nodules and results in an AUC value of 98.51%. The FROC and JAFROC analyses show that the performance of such an architecture is significantly improved compared to "traditional" methods.

Liu et al. [105] presented a hybrid approach combining traditional hand crafted features and deep learning attributes to discriminate the nature of thyroid nodules (benign or malignant). The VGG-F architecture [106], consisting of 5 convolutional pooling groups and 2 fully connected layers, and pre-trained on the ImageNet dataset was used as the basis for this architecture. The results from the lower layers represent the low-level features, including edge, direction, and intensity features, which are combined with those extracted manually (HOG, LBP, Gray Level Co-Occurrence Matrix), Scale Invariant Feature Transform (SIFT), and VLAD (Vector of Locally Aggregated Descriptors). The advantage of this combination of features is that, apart from high-level feature descriptors generalized by CNNs, the manually extracted ones, which are manufactured low-level feature descriptors, retain their relevance advantages. 1037 clinically verified ultrasound images of thyroid nodules provided by the Cancer Hospital of Chinese Academy of Medical Sciences were used for evaluation. This dataset consisted of 651 images representing benign tumors and 386 images corresponding to

malignant tumors. The type and location of all nodules were annotated by the physicians. The results showed the effectiveness of this architecture with an accuracy of 93.1%, a sensitivity of 90.8%, a specificity of 94.5%, and an AUC of 0.977.

Chen et al. [107] presented a non-invasive and automatic approach for the characterization of thyroid nodules in ultrasound images based on the adaptation of the parameters of a deep neural network (DNN) model. Several experiments have been conducted in order to fix the optimal parameters of the DNN. The confusion matrix was used to compare the difference between the model's predictions and actual values. The prediction accuracy of the model has reached 93%. The Receiver Operating Characteristic (ROC) curve is very close to the boundary of the rectangle. The AUC (Area Under The Curve) is of 0.935, which shows that the proposed model is performing well. Experimental results show the superiority of the proposed DNN algorithm, since it has the highest accuracy rate of 93% and 95% respectively on the actual medical dataset and the UCI standard dataset, compared to some algorithms traditional learning such as random forest and other learning algorithms.

7.2.2 Lung nodule classification

In [108] the authors use a model based on residual blocks to classify lung nodules. The latter are of various shapes and sizes which require focusing on global as well as local features. Authors use non-local neural networks [109, 110], also known as self-attention layers which reduce the number of parameters needed. For local feature identification, they use residual blocks [111] with a kernel size of 3×3. The model was trained on the LIDC-IDRI dataset [112] (containing 1,018 computed tomography (CT) scans annotated by radiologists) and validated by the "10-fold cross validation" technique. With an AUC=95.62%, the proposed method significantly outperformed the other reference methods.

Tran et al. [113] propose a CNN architecture named LdcNet, using the focal loss function proposed by Lin et al. [114]. This architecture has been trained on the LIDC/IDRI dataset [112] while applying data augmentation to consolidate the initial dataset with new images. Non-uniform transformations, such as stretching or tilting, were not applied due to the importance of nodule shape for classification. The LdcNet architecture consists of 3 blocks of 3 successive convolutional layers each, each block is followed by a max pooling layer. Two fully connected layers complete the architecture and allow the classification of nodules from the features extracted from the previous convolutional blocks. This architecture achieved a sensitivity of 96.0%, an accuracy of 97.2% and a specificity of 97.3%.

Song et al. [115] used 3 different techniques for lung nodule classification: a CNN, a DNN, and a Stacked Autoencoder (SAE). The LIDC-IDRI dataset [112] was used for training and validation of each of the 3 models.

- The CNN consists of two convolution layers composed of 32 filters of size 5×5. Each convolution layer is followed by a max pooling layer of a kernel of size 2×2. Two fully connected layers followed by a Softmax classifier complete the CNN architecture.
- The DNN consists of an input layer that receives the grayscale images whose size is 28×28. The input layer is followed by four fully connected layers of respective sizes 784×1, 512×1, 256×1 and 64×1. A dropout layer with a parameter of 0.6 has been inserted between the third and fourth fully connected layers. The activation function of the fourth layer is ReLU. A Softmax classifier is applied at the end of the DNN.
- The SAE is a multi-layer sparse auto-encoder. An auto-encoder is an ANN that is often used in learning discriminative features from a dataset. It consists of an encoder and a decoder. The encoder consists of a set of layers of neurons, which process the data in order

Table 4 Results obtained by Vaid et al. [116]

Finding	Precision (%)	Recall (%)	F1-score (%)
Normal	98.6	96.0	97.3
COVID-19	91.7	97.1	94.3

to obtain a new representation of it while the layers of neurons of the decoder analyze the encoded data and try to rebuild the original data. In [115], several auto-encoders and Softmax classifiers are combined to build an SAE network with several hidden layers and a final Softmax classifier. Since the SAE is used for classification in this work, the hidden layer generated by the auto-encoder is directly used for classification, thus negating the decoding part of the auto-encoder.

The results obtained by this work are summarized in Table 3 . The CNN architecture performed the best of the 3 models. These results do not improve on previous work for fairly obvious reasons. The CNN is not deep enough and the authors could have used for example 3×3 filters to capture more detail. The poorer performance of the DNN and the SAE compared to the CNN can be explained by the very nature of these two architectures which are formed by fully connected layers, although obtained in different ways. The performance of convolution layers in the extraction of shape and texture features far exceeds the possibilities of fully connected layers.

7.2.3 Identification of COVID-19 disease infection

The COVID-19 virus does not affect everyone in the same way. It has been verified, in many cases, that some individuals can contract the virus and therefore be contagious without developing symptoms. The tests for COVID-19 are not 100% reliable. On the other hand, the symptoms of COVID-19 correspond to the symptoms of other much less dangerous respiratory infections. Additional or more reliable methods than the RT-PCR (Reverse Transcription-Polymerase Chain Reaction) test whose reliability is 95% or the antigenic test (much less reliable than PCR) are needed. Many research works have focused on the impact of deep learning in the detection of COVID-19 from medical images, in particular, Chest X-rays and Computed Tomography (CT) scans.

Vaid et al. [116] adopted a transfer learning approach based on the VGG-19 architecture for COVID-19 case classification from patient anterior–posterior chest radiographs. They used a publicly available dataset for learning and evaluating their model.[1] This dataset is composed of frontal chest radiographs from 181 patients from different geographical locations (Italy, China, Australia, etc.) and labeled by expert radiologists. The small number of images explains the use of transfer learning. The VGG-19 model was pre-trained on lmageNet and modified by adding a multilayer perceptron trained on the chest radiograph dataset for the classification of COVID-19 positive and negative cases. The results of this model are presented in Table 4.

In [117], the authors presented a CAD allowing the detection of infections caused by COVID-19. The process starts with an image pre-processing phase to remove regions unnecessary for diagnosis and in particular the diaphragm which is composed of pixels with high intensity that negatively affect the distinction and quantification of lung disease patterns. The original image is then processed on one the one hand by a histogram equalization algorithm

[1] https://github.com/ieee8023/covid-chestxray-dataset.

Table 5 Classification report of the method proposed in [117]

	Precision (%)	Recall (%)	F1-score (%)
Normal	96	91	93
COVID-19	73	98	84
Other pneumonia	96	96	96

and on the other hand by the application of a bilateral low pass filter, each of the two processing results in a new image. These images are combined with the original image to form a pseudo-color image. It is this image that is used for classification by a VGG16 model pre-trained on ImageNet (transfer learning) to classify the chest radiographs into three classes: (i) infected with COVID-19, (ii) Pneumonia other than COVID-19, and (iii) normal cases (no pneumonia). The dataset used is an assembly of chest radiography images from several different public medical repositories comprising 8474 posteroanterior 2D radiological images of the chest. 415 images in the dataset represent confirmed cases of COVID-19 disease, 5179 represent non-COVID-19 pneumonia, and 2880 represent normal cases (no pneumonia). In this study, the class weight technique, which consists in adjust weights inversely proportional to class frequencies in the input data, combined with data augmentation, is applied during training to reduce the potential consequences of imbalanced data. The results of this work are summarized in Table 5.

In [118] the authors analyzed the effectiveness of a VGG16-based deep learning model for the identification of pneumonia and COVID-19 from chest radiographs. They used a public dataset of radiological images of healthy patients, patients with pneumonia, and patients with COVID-19.[2] They applied the hold-out technique to split the dataset into a learning subset composed of 80% of the dataset content selected at random and a test/validation subset composed of the remaining 20% of images. The model obtained an Accuracy, a Precision, a Sensitivity and an F1-Score of 86% each, a Specificity of 93%.

Misra et al. [119] used residual learning based on three multi-channel pre-trained ResNet architectures for COVID-19 diagnosis from chest radiographs. The three models were retrained to classify radiographs on a one-against-all basis of (a) normal or diseased, (b) pneumonia or non-pneumonia, and (c) COVID-19 or non-COVID-19. Slightly more than 6000 chest radiographs from three different public datasets were used for this work. These data were from the RSNA Pneumonia Detection Challenge,[3] the COVID-19 image dataset,[4] and the COVID-19 radiographs [120]. The final dataset consisted of 1579 radiographs of healthy individuals, 4245 individuals with pneumonia and 184 individuals with COVID-19. The three ResNet architectures were assembled and refined based on the final dataset to recognize the 3 classes of healthy, pneumonia other than COVID-19, and COVID-19. The results demonstrated the effectiveness of the approach with 94% accuracy and 100% recall on the dataset.

In [121], in order to automatically diagnose COVID-19, Ouyang et al. developed a dual-sampling attention network for thoracic computed tomography (CT) classification. Based on two 3D ResNet34 networks, the authors used different sampling strategies by integrating a gradient-based attention mechanism. Gradient-based attention methods are methods usually conducted offline, allowing to reveal important regions influencing the network prediction. In

[2] https://public.roboflow.ai/classification/covid-19-and-pneumoniascans.

[3] https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/RSNA-Pneumonia-Detection-Challenge-2018.

[4] https://www.kaggle.com/datasets/andrewmvd/convid19-x-rays.

Table 6 Related literature review of deep learning

Method	Classifiers	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	AUC (%)	F1-score (%)
Thyroid nodule classification							
Li et al. [96]	R-CNN	91.60	93.50	81.30	93.30	93.80	93.40
Ma et al. [99]	CNN	NA	NA	NA	NA	98.51	NA
Liu et al. [105]	VGG-F	93.10	90.80	94.50	NA	97.70	NA
Chen et al. [107]	DNN	95.00	NA	NA	NA	93.50	NA
Lung nodule classification							
Al-Shab et al. [108]	ANN	NA	NA	NA	NA	95.62	NA
Tran et al. [113]	LdcNet	97.20	96.00	97.30	NA	NA	NA
Song et al. [115]	CNN	84.15	83.96	84.32	NA	NA	NA
	DNN	82.37	80.66	83.90	NA	NA	NA
	SAE	82.59	83.96	81.35	NA	NA	NA
Identification of COVID-19 disease infection							
Vaid et al. [116]	VGG19	NA	96.00	NA	98.60	NA	97.30
	VGG19	NA	97.10	NA	91.70	NA	94.30
Heidari et al. [117]	VGG16	NA	91.00	NA	96.00	NA	93.00
	VGG16	NA	98.00	NA	73.00	NA	84.00
	VGG16	NA	96.00	NA	96.00	NA	96.00
Civit-Masot et al. [118]	VGG16	86.00	86.00	93.00	86.00	NA	86.00
Misra et al. [119]	ResNet architectures	94.00	100	NA	NA	NA	NA
Ouyang et al. [121]	3D ResNet34 networks	87.50	86.90	90.10	NA	94.40	82.00

their work, the authors extended gradient-based attention by developing an online trainable component that uses segmented pneumonia infection regions to ensure that the network can make decisions based on these infection regions. The dual sampling strategy adopted in this work is used to mitigate the imbalance in the size of the infected regions between COVID-19 and Community Acquired Pneumonia. This imbalance is in part a result of the rapid progression of COVID-19 after symptom appearance. The model was trained, tested, and validated on one of the largest multi-center CT dataset for COVID-19 from 8 hospitals. 2186 CT scans of 1588 patients were used for the training and validation phase (5-fold cross-validation). The tests were performed on another dataset composed of 2796 CT scans of 2057 patients. This model obtained an accuracy of 87.5%, a sensitivity of 86.9%, a specificity of 90.1%, an F1 score of 82.0%, and an AUC of 0.944 (Table 6).

8 Discussion

The first finding from this study is the effectiveness of both classical and deep learning approaches in tumor classification. The question that then arises is "Why continue to explore both approaches?" Some ask why not focus on deep learning since it relieves us of the tedious work of feature identification and selection. To answer this kind of question, a step back should be taken and analyze the effectiveness of each of the two approaches and especially when is this effectiveness put to the test.

The effectiveness of either approach depends on the efficient and accurate selection of features to determine whether a tumor is benign or malignant and to differentiate acute pulmonary edema from COVID-19 pneumonia. Deep learning has demonstrated very good performances in terms of identification and selection of features, but on the other hand, it needs a huge volume of data on which it will be trained. The availability of data in sufficient quantities, to train a CNN for example, is not always guaranteed. And even if these data are available, if they are not well chosen, they can lead to inefficient systems because of the phenomena of under-fitting and over-fitting.

Although the performance of both approaches can be affected by an imbalance in the training data, deep learning is more sensitive to it because the features are extracted automatically. If the dataset representing one of the classes is not sufficient, there is a high chance that the values of the features learned and representing this class are wrong. In this case, a performance indicator, such as accuracy, on its own is not very meaningful. Indeed, if for example in a thyroid nodule dataset, there are 90% benign nodules and 10% malignant nodules, if the classifier is wrong on all the malignant nodules, it will still have an accuracy of 90%, which is not negligible.

If for small datasets deep learning is rather inefficient, this is where classical learning can be distinguished. As long as the data sets are small, the learning phase is not time-consuming. The features are chosen by experts on scientific and experimental grounds. Again, it all depends on the feature selection process which, if done manually, is quite subjective and prone to interpretation.

To address the issues that arise in each of the two models, solutions have been proposed. For the handcrafted selection of features, in addition to relying on the expertise of the various actors involved in the CAD design process, in particular physicians and image processing experts, rigorous methods based on mathematical and statistical foundations have been developed to help better identify the relevant features for classification (refer to paragraphs 6.1, 6.2 and 6.3).

Proven solutions also exist to solve the problem of unbalanced classes. On the one hand, there is data augmentation which allows, by applying certain modifications on the elements of the initial dataset, to virtually increase its cardinal. It is thus possible to multiply the number of objects of a dataset by 5 or even by 6. This solution, which has shown its efficiency in a good number of applications, does not really solve the problem but allows us to have a sufficient number of data to learn well the characteristics allowing the classifications. That said, some transformations may not be adequate for certain types of objects. The paragraph 7.2.2 pointed out that the shape of a lung nodule is crucial for a good diagnosis and, therefore, non-uniform transformations, such as stretching or tilting should not be applied. Approaches really dedicated to unbalanced data have been implemented to solve the problem of class imbalance. Several approaches have been used acting at the data level or at the algorithm level as well as some hybrid methods. Deep generative models [122] seem to be among the most promising. These are unsupervised learning methods whose goal is to learn to describe the underlying distribution of unlabeled training data and, from there, also learn to generate brand new data from this distribution, thus establishing a balance between classes by enriching those that lack data.

9 Conclusion

To conclude this comparative study, it is necessary to remind that in the healthcare sector, AI covers all areas, from biological and imaging diagnostics to technotherapies (technological therapies: prostheses, connected implants), and even biotherapies (genetic and cell therapy). AI enables the exploration of medical data accumulated in recent years and has brought medicine into a new and major phase of innovation comparable to the one that led to the development of antibiotics after the Second World War. The two major contributions that are driving AI in the healthcare field are (i) the scientific contribution in terms of R&D, which promotes discoveries on diagnosis, prognosis, and prediction of the evolution of pathologies, and (ii) the economic and socio-industrial contributions. This leads to the automation of numerous tasks, the transformation of care paths, and the organization of care by measuring the impact of treatments.

In the context of medical diagnosis of a pathology, whether preventive or curative, and which requires a significant effort from medical staff, AI is emerging as a real alternative that is transforming the professions and practices of the healthcare sector. Thanks to a great capacity to analyze data compared to humans, it has facilitated the establishment of a diagnosis for diseases that are complex to detect: eye diseases, cancers, melanomas, etc. Among the latest developments in this field, the analysis of medical imagery for the macroscopic recognition of skin tumors or breast cancer, etc. has been added to these capabilities to further refine the proposed diagnosis.

The question that arises and that this research attempted to solve through this literature review is: “In diagnosing, can AI do as well as doctors?” Experience shows that ML now exceeds the ability of doctors to detect tumors in medical images. This paper has examined the possibilities of the so-called classical ML algorithms (SVM, RF, KNN, etc.) compared to deep learning and the performances are quite comparable when the whole process is established with rigor. It has been presented in the discussion part when one or the other of the two approaches was necessary and, roughly speaking, deep learning is more efficient when the learning data is abundant while it presents weaknesses if the datasets are small. In this case, classical methods are the best choice, provided that the characteristics involved in

the diagnosis are well-identified. Nevertheless, the choice of one or the other approach is not only a question of the size of the datasets. Indeed, the size of a dataset can more or less be rectified by proven techniques such as, for example, data augmentation or data generation using auto-encoders. Other problems that can affect the quality of an ML model and its performance are the dataset structure. It has been shown that some poorly compiled datasets can lead to the over-fitting. The quality of a dataset is not always guaranteed. and software solutions have been developed to minimize the risk of over-fitting. The first step is to detect the over-fitting itself by the cross validation technique but this can only be done after the learning phase. Then, other techniques, such as dropout or ensemble learning can be used to limit the risk.

What remains to be done? Current work in this area consists of improving the scores of different algorithms by searching for new models or combining existing models as well as assessing mathematical means of evaluating loss and generalization functions. Attempts using reinforcement learning combined with usual classifiers (neural networks, SVM, etc.) are being investigated. We are almost at the beginning of this revolution in the medical field, other models and other innovations will surely come in the near future.

References

1. Alam TM, Shaukat K, Khelifi A, Aljuaid H, Shafqat M, Ahmed U, Nafees SA, Luo S (2022) A fuzzy inference-based decision support system for disease diagnosis. Comput J. https://doi.org/10.1093/comjnl/bxac068
2. Devnath L, Summons P, Luo S, Wang D, Shaukat K, Hameed IA, Aljuaid H (2022) Computer-aided diagnosis of coal workers' pneumoconiosis in chest x-ray radiographs using machine learning: a systematic literature review. Int J Environ Res Public Health 19(11):6439. https://doi.org/10.3390/ijerph19116439
3. Dharmale SG, Gomase SA, Pande S (2022) Comparative analysis on machine learning methodologies for the effective usage of medical wsns. In: Proceedings of data analytics and management. Springer, pp 441–457. https://doi.org/10.1007/978-981-16-6285-0_36
4. Zhang S, Li X, Zong M, Zhu X, Wang R (2018) Efficient KNN classification with different numbers of nearest neighbors. IEEE Trans Neural Netw Learn Syst 29(5):1774–1785. https://doi.org/10.1109/TNNLS.2017.2673241
5. Guo G, Wang H, Bell D, Bi Y, Greer K (2003) KNN model-based approach in classification, vol 2888, pp 986–996. https://doi.org/10.1007/978-3-540-39964-3_62
6. Jain AK, M, J, Mohiuddin KM (1996) Artificial neural networks: a tutorial. Computer 29(3):31–44. https://doi.org/10.1109/2.485891
7. Bernard S, Heutte L, Adam S (2009) On the selection of decision trees in random forests. In: 2009 International joint conference on neural networks, pp 302–307. https://doi.org/10.1109/IJCNN.2009.5178693
8. Man KF, Tang KS, Kwong S (1996) Genetic algorithms: concepts and applications. IEEE Trans Ind Electron 43(5):519–534. https://doi.org/10.1109/41.538609
9. Holland JH (1992) Genetic algorithms. Sci Am 267(1):66–73
10. Rajurkar P, Mohod S, Pande S (2021) The study of various methodologies in the development of recommendation system. In: 2021 9th International conference on reliability, infocom technologies and optimization (trends and future directions) (ICRITO), pp 1–5. https://doi.org/10.1109/ICRITO51393.2021.9596125
11. Hecht–Nielsen R (1989) Theory of the backpropagation neural network. In: International 1989 joint conference on neural networks, pp 593–6051. https://doi.org/10.1109/IJCNN.1989.118638
12. Leung H, Haykin S (1991) The complex backpropagation algorithm. IEEE Trans Signal Process 39(9):2101–2104. https://doi.org/10.1109/78.134446
13. Brunel A, Mazza D, Pagani M (2019) Backpropagation in the simply typed lambda-calculus with linear negation. Proc ACM Program Lang. https://doi.org/10.1145/3371132
14. Berner ESL, Osheroff JA, Tamblyn R (2009) Clinical decision support systems: state of the art

15. Devnath L, Luo S, Summons P, Wang D, Shaukat K, Hameed IA, Alrayes FS (2022) Deep ensemble learning for the automatic detection of pneumoconiosis in coal worker's chest X-ray radiography. J Clin Med 11(18):5342. https://doi.org/10.3390/jcm11185342
16. Lowe DG (2004) Distinctive image features from scale invariant keypoints. Int J Comput Vis 60(2):91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94
17. Bay H, Ess A, Tuytelaars T, Gool LV (2008) Speeded-up robust features (SURF). Comput Vis Image Underst 110(3):346–359. https://doi.org/10.1016/j.cviu.2007.09.014
18. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection, vol 1, pp 886–893. https://doi.org/10.1109/CVPR.2005.177
19. Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell 24(4):509–522. https://doi.org/10.1109/34.993558
20. Sivic J, Zisserman A (2003) Video Google: a text retrieval approach to object matching in videos. In: Ninth IEEE international conference on computer vision, Nice, France, pp 1470–1477
21. Srinivas C, KS NP, Zakariah M, Alothaibi YA, Shaukat K, Partibane B, Awal H (2022) Deep transfer learning approaches in performance analysis of brain tumor classification using mri images. J Healthc Eng. https://doi.org/10.1155/2022/3264367
22. Woods R, Cherry S, Mazziotta J (1992) Rapid automated algorithm for aligning and reslicing PET images. J Comput Assist Tomogr 16:620–633
23. Roche A, Malandain G, Pennec X, Ayache N (1998) The correlation ratio as a new similarity measure for multimodal image registration, vol 1496, pp 1115–1124. https://doi.org/10.1007/BFb0056301
24. Iqbal Q, Aggarwall JK (1999) Applying perceptual grouping to content–based image retrieval: building images
25. Flickner M, Sawhney H, Niblack W, Ashley J, Huang Q, Dom B, Gorkani M, Hafner J, Lee D, Petkovic D, Steele D, Yanker P (1995) Query by image and video content: the QBIC system. Computer 28(9):23–32. https://doi.org/10.1109/2.410146
26. Bernal J, Vilarino F, Sánchez J (2010) Feature detectors and feature descriptors: where we are now
27. Kashif M, Deserno TM, Haak D, Jonas S (2016) Feature description with sift, surf, brief, brisk, or freak? A general question answered for bone age assessment. Comput Biol Med 68(1):67–75. https://doi.org/10.1016/j.compbiomed.2015.11.006
28. Bosch A, Zisserman A, Munoz X (2007) Image classification using random forests and ferns. In: 2007 IEEE 11th international conference on computer vision, pp 1–8. https://doi.org/10.1109/ICCV.2007.4409066
29. Li F, Fergus R, Perona P (2004) Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: 2004 Conference on computer vision and pattern recognition workshop, pp 178–178. https://doi.org/10.1109/CVPR.2004.383
30. Griffin G, Holub A, Perona P (2007) Caltech-256 object category dataset. CalTech Report
31. Everingham M, Van Gool L, Williams C, Winn J, Zisserman A (2010) The pascal visual object classes (VOC) challenge. Int J Comput Vis 88:303–338. https://doi.org/10.1007/s11263-009-0275-4
32. Horvath E, Majlis S, Rossi R, Franco C, Niedmann JP, Castro A, Dominguez M (2009) An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management. J Clin Endocrinol Metab 94(5):1748–1751. https://doi.org/10.1210/jc.2008-1724
33. Dong Y, Wang T, Yang C, Zheng L, Song B, Wang L, Jin M (2019) Locally directional and extremal pattern for texture classification. IEEE Access 7:87931–87942. https://doi.org/10.1109/ACCESS.2019.2924985
34. Chunmei X, Mei H, Yan Z, Haiying W (2020) Diagnostic method of liver cirrhosis based on mr image texture feature extraction and classification algorithm. J Med Syst 44:1–8. https://doi.org/10.1007/s10916-019-1508-x
35. Guo F, Li W, Tang J, Zou B, Fan Z (2020) Automated glaucoma screening method based on image segmentation and feature extraction. Med Biol Eng Comput 58(10):2567–2586. https://doi.org/10.1007/s11517-020-02237-2
36. Arvacheh EM, Tizhoosh HR (2005) Pattern analysis using Zernike moments. In: 2005 IEEE instrumentation and measurement technology conference proceedings, vol 2, pp 1574–1578. https://doi.org/10.1109/IMTC.2005.1604417
37. Liao SX, Pawlak M (1998) On the accuracy of Zernike moments for image analysis. IEEE Trans Pattern Anal Mach Intell 20(12):1358–1364. https://doi.org/10.1109/34.735809
38. Aboudi N, Guetari R, Khlifa N (2020) Multi-objectives optimisation of features selection for the classification of thyroid nodules in ultrasound images. IET Image Process 14(9):1901–1908. https://doi.org/10.1049/iet-ipr.2019.1540
39. Ryszard SC et al (2007) Image feature extraction techniques and their applications for cbir and biometrics systems. Int J Biol Biomed Eng 1(1):6–16

40. Humeau-Heurtier A (2019) Texture feature extraction methods: a survey. IEEE Access 7:8975–9000. https://doi.org/10.1109/ACCESS.2018.2890743
41. Alpaslan N, Hanbay K (2020) Multi-resolution intrinsic texture geometry-based local binary pattern for texture classification. IEEE Access 8:54415–54430. https://doi.org/10.1109/ACCESS.2020.2981720
42. Meena KB, Tyagi V (2020) A hybrid copy-move image forgery detection technique based on Fourier–Mellin and scale invariant feature transforms. Multimed Tools Appl 79(11):8197–8212. https://doi.org/10.1007/s11042-019-08343-0
43. Bansal M, Kumar M, Kumar M (2021) 2d object recognition: a comparative analysis of SIFT, SURF and ORB feature descriptors. Multimed Tools Appl 80(12):18839–18857. https://doi.org/10.1007/s11042-021-10646-0
44. Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features. Lect Notes Comput Sci 3951:404–417. https://doi.org/10.1007/11744023_32
45. Alwan HB, Ku-Mahamud KR (2020) Cancellable face template algorithm based on speeded-up robust features and winner-takes-all. Multimed Tools Appl 79(39):28675–28693. https://doi.org/10.1007/s11042-020-09319-1
46. He Q, He B, Zhang Y, Fang H (2019) Multimedia based fast face recognition algorithm of speed up robust features. Multimed Tools Appl 78(17):24035–24045. https://doi.org/10.1007/s11042-019-7209-0
47. Struzik ZR, Siebes A (1999) The Haar wavelet transform in the time series similarity paradigm. In: Proceedings of the third European conference on principles of data mining and knowledge discovery. Springer, Berlin, pp 12–22
48. Rublee E, Rabaud V, Konolige K, Bradski G (2011) Orb: An efficient alternative to sift or surf. In: 2011 International conference on computer vision, pp 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544
49. Chhabra P, Garg NK, Kumar M (2020) Content-based image retrieval system using ORB and SIFT features. Neural Comput Appl 32(7):2725–2733. https://doi.org/10.1007/s00521-018-3677-9
50. Harris C, Stephens M (1988) A combined corner and edge detector. In: Proceedings of the 4th Alvey vision conference, pp 147–151
51. Gupta S, Kumar M, Garg A (2019) Improved object recognition results using SIFT and ORB feature detector. Multimed Tools Appl 78(23):34157–34171. https://doi.org/10.1007/s11042-019-08232-6
52. Pande S, Khamparia A, Gupta D (2021) Feature selection and comparison of classification algorithms for wireless sensor networks. J Ambient Intell Humaniz Comput 66:1–13. https://doi.org/10.1007/s12652-021-03411-6
53. Hancer E, Xue B, Zhang M (2020) A survey on feature selection approaches for clustering. Artif Intell Rev 53(6):4519–4545. https://doi.org/10.1007/s10462-019-09800-w
54. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemom Intell Lab Syst 2(1–3):37–52. https://doi.org/10.1016/0169-7439(87)8008-9
55. Reddy GT, Reddy MPK, Lakshmanna K, Kaluri R, Rajput DS, Srivastava G, Baker T (2020) Analysis of dimensionality reduction techniques on big data. IEEE Access 8:54776–54788. https://doi.org/10.1109/ACCESS.2020.2980942
56. Geetharamani R, Sivagami G (2021) Iterative principal component analysis method for improvised classification of breast cancer disease using blood sample analysis. Med Biol Eng Comput 59(10):1973–1989. https://doi.org/10.1007/s11517-021-02405-y
57. Ricciardi C, Valente AS, Edmund K, Cantoni V, Green R, Fiorillo A, Picone I, Santini S, Cesarelli M (2020) Linear discriminant analysis and principal component analysis to predict coronary artery disease. Health Inform J 26(3):2181–2192. https://doi.org/10.1109/ACCESS.2020.2980942
58. Zhao H, Zheng J, Xu J, Deng W (2019) Fault diagnosis method based on principal component analysis and broad learning system. IEEE Access 7:99263–99272. https://doi.org/10.1109/ACCESS.2019.2929094
59. Ye J, Janardan R, Li Q (2004) Two-dimensional linear discriminant analysis. Adv Neural Inf Process Syst 17:66
60. Hou Q, Wang Y, Jing L, Chen H (2019) Linear discriminant analysis based on kernel-based possibilistic c-means for hyperspectral images. IEEE Geosci Remote Sens Lett 16(8):1259–1263. https://doi.org/10.1109/LGRS.2019.2894470
61. Dornaika F, Khoder A (2020) Linear embedding by joint robust discriminant analysis and inter-class sparsity. Neural Netw 127:141–159. https://doi.org/10.1016/j.neunet.2020.04.018
62. Hand DJ (2006) Classifier technology and the illusion of progress. Stat Sci 21(1):1–14. https://doi.org/10.1214/088342306000000060
63. Noble WS (2006) What is a support vector machine? Nat Biotechnol 24(12):1565–1567. https://doi.org/10.1038/nbt1206-1565

64. Danjuma KJ, Wajiga GM, Garba EJ, Ahmadu AS, Longe OB (2022) Accuracy assessment of machine learning algorithm (s) in thyroid dysfunction diagnosis. In: 2022 IEEE Nigeria 4th international conference on disruptive technologies for sustainable development (NIGERCON), pp 1–5. https://doi.org/10.1109/NIGERCON54645.2022.9803113
65. Abbad Ur Rehman H, Lin C, Mushtaq Z, Su S (2021) Performance analysis of machine learning algorithms for thyroid disease. Arab J Sci Eng 46(10):9437–9449. https://doi.org/10.1007/s13369-020-05206-x
66. Jackins V, Vimal S, Kaliappan M, Lee MY (2021) Ai-based smart prediction of clinical disease using random forest classifier and Naive Bayes. J Supercomput 77(5):5198–5219. https://doi.org/10.1007/s11227-020-03481-x
67. Sampath P, Packiriswamy G, Pradeep Kumar N, Shanmuganathan V, Song O, Tariq U, Nawaz R (2020) Iot based health-related topic recognition from emerging online health community (med help) using machine learning technique. Electronics 9(9):1469. https://doi.org/10.3390/electronics9091469
68. Olatunji SO, Alotaibi S, Almutairi E, Alrabae Z, Almajid Y, Altabee R, Altassan M, Ahmed Basheer MI, Farooqui M, Alhiyafi J (2021) Early diagnosis of thyroid cancer diseases using computational intelligence techniques: a case study of a Saudi Arabian dataset. Comput Biol Med 131:104267. https://doi.org/10.1016/j.compbiomed.2021.104267
69. Shankar K, Lakshmanaprabu SK, Gupta D, Maseleno A, De Albuquerque HCV (2020) Optimal feature-based multi-kernel svm approach for thyroid disease classification. J Supercomput 76(2):1128–1143. https://doi.org/10.1007/s11227-018-2469-4
70. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
71. Xie J, Wang C (2011) Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. Expert Syst Appl 38(5):5809–5815. https://doi.org/10.1016/j.eswa.2010.10.050
72. Islam SS, Haque MS, Miah MSU, Sarwar TB, Nugraha R (2022) Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study. PeerJ Comput Sci 8:898. https://doi.org/10.7717/peerj-cs.898
73. Rish I et al (2001) An empirical study of the Naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence, vol 3, pp 41–46
74. Divya K, Sirohi A, Pande S, Malik R (2021) An iomt assisted heart disease diagnostic system using machine learning techniques. In: Cognitive internet of medical things for smart healthcare: services and applications, pp 145–161. https://doi.org/10.1007/978-3-030-55833-8_9
75. Peterson LE (2009) K-nearest neighbor. Scholarpedia 4(2):1883
76. Mirjalili S, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. Adv Eng Softw 69:46–61. https://doi.org/10.1016/j.advengsoft.2013.12.007
77. Nugroho HA, Frannita EL, Ardiyanto I, Choridah L et al (2021) Computer aided diagnosis for thyroid cancer system based on internal and external characteristics. J King Saud Univ Comput Inf Sci 33(3):329–339. https://doi.org/10.1016/j.jksuci.2019.01.007
78. Prochazka A, Gulati S, Holinka S, Smutek D (2019) Classification of thyroid nodules in ultrasound images using direction-independent features extracted by two-threshold binary decomposition. Technol Cancer Rese Treat 18:1533033819830748. https://doi.org/10.1177/1533033819830748
79. Shivastuti KH, Manhas J, Sharma V (2021) Performance evaluation of svm and random forest for the diagnosis of thyroid disorder. Int J Res Appl Sci Eng Technol 9:945–947. https://doi.org/10.1007/978-981-15-6202-0_39
80. Colakoglu B, Alis D, Yergin M (2019) Diagnostic value of machine learning-based quantitative texture analysis in differentiating benign and malignant thyroid nodules. J Oncol. https://doi.org/10.1155/2019/6328329
81. Alyas T, Hamid M, Alissa K, Faiz T, Tabassum N, Ahmad A (2022) Empirical method for thyroid disease classification using a machine learning approach. BioMed Res Int. https://doi.org/10.1155/2022/9809932
82. Ataide EJ, Ponugoti N, Illanes A, Schenke S, Kreissl M, Friebe M (2020) Thyroid nodule classification for physician decision support using machine learning-evaluated geometric and morphological features. Sensors 20(21):6110. https://doi.org/10.3390/s20216110
83. Prochazka A, Gulati S, Holinka S, Smutek D (2019) Classification of thyroid nodules in ultrasound images using direction-independent features extracted by two-threshold binary decomposition. Technol Cancer Res Treat 18:1533033819830748. https://doi.org/10.1016/j.compmedimag.2018.10.001
84. Nie F, Huang H, Cai X, Ding C (2010) Efficient and robust feature selection via joint l2,1-norms minimization. In: Proceedings of the 23rd international conference on neural information processing systems, vol 2, pp 1813–1821. Curran Associates Inc., Red Hook

85. Jin J, Xiao R, Daly I, Miao Y, Wang X, Cichocki A (2021) Internal feature selection method of csp based on l1-norm and Dempster–Shafer theory. IEEE Trans Neural Netw Learn Syst 32(11):4814–4825. https://doi.org/10.1109/TNNLS.2020.3015505
86. Khushboo C, Kunwar V, Sabitha S, Choudhury T, Mukherjee S (2016) A comparative study on thyroid disease detection using k-nearest neighbor and Naive Bayes classification techniques. CSI Trans ICT 4(2):313–319. https://doi.org/10.1007/s40012-016-0100-5
87. Haenssle HAEA (2018) Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 29(8):1836–1842. https://doi.org/10.1093/annonc/mdy166
88. Liu Y, Kohlberger T, Norouzi M, Dahl GE, Smith JL, Mohtashamian A, Olson N, Peng LH, Hipp JD, Stumpe MC (2018) Artificial intelligence-based breast cancer nodal metastasis detection: insights into the black box for pathologists. Arch Pathol Lab Med 143(7):859–868. https://doi.org/10.5858/arpa.2018-0147-OA
89. Alam TM, Shaukat K, Khan WA, Hameed IA, Almuqren LA, Raza MA, Aslam M, Luo S (2022) An efficient deep learning-based skin cancer classifier for an imbalanced dataset. Diagnostics 12(9):2115. https://doi.org/10.3390/diagnostics12092115
90. Ferguson M, Ak R, Lee Y-TT, Law KH (2017) Automatic localization of casting defects with convolutional neural networks. In: 2017 IEEE international conference on big data (big data), pp 1726–1735. https://doi.org/10.1109/BigData.2017.8258115
91. Kumar P, Singh P, Pande S, Khamparia A (2022) Plant leaf disease identification and prescription suggestion using deep learning. In: Proceedings of data analytics and management. Springer, pp 547–560. https://doi.org/10.1007/978-981-16-6285-0_43
92. Yadav N, Pande S, Khamparia A, Gupta D (2022) Intrusion detection system on iot with 5g network using deep learning. Wirel Commun Mob Comput 2022:66. https://doi.org/10.1155/2022/9304689
93. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958
94. Dauphin Y, De Vries H, Bengio Y (2015) Equilibrated adaptive learning rates for non-convex optimization. Adv Neural Inf Process Syst 28:66. arXiv:1502.04390
95. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
96. Li H, Weng J, Shi Y, Gu W, Mao Y, Wang Y, Liu W, Zhang J (2018) An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images. Sci Rep. https://doi.org/10.1038/s41598-018-25005-7
97. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
98. Li H, Huang Y, Zhang Z (2017) An improved faster R-CNN for same object retrieval. IEEE Access 5:13665–13676. https://doi.org/10.1109/ACCESS.2017.2729943
99. Ma J, Wu F, Jiang T, Zhu J, Kong D (2017) Cascade convolutional neural networks for automatic detection of thyroid nodules in ultrasound images. Med Phys 44(5):1678–1691. https://doi.org/10.1002/mp.12134
100. Shen W, Zhou M, Yang F, Yang C, Tian J (2015) Multi–scale convolutional neural networks for lung nodule classification. In: Information processing in medical imaging. Springer, Cham, pp 588–599. https://doi.org/10.1007/978-3-319-19992-4_46
101. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH (1977) A free response approach to the measurement and characterization of radiographic observer performance. In: Application of optical instrumentation in medicine VI, vol 0127. SPIE, pp 124–135. https://doi.org/10.1117/12.955926
102. Bandos AI, Rockette HE, Song T, Gur D (2009) Area under the free-response ROC curve (FROC) and a related summary index. Biometrics 65(1):247–256. https://doi.org/10.1111/j.1541-0420.2008.01049.x
103. Chakraborty DP, Berbaum KS (2004) Observer studies involving detection and localization: modeling, analysis, and validation. Med Phys 31(8):2313–2330. https://doi.org/10.1118/1.1769352
104. Chakraborty DP (2008) Validation and statistical power comparison of methods for analyzing free-response observer performance studies. Acad Radiol 15(12):1554–1566. https://doi.org/10.1016/j.acra.2008.07.018
105. Liu T, Xie S, Yu J, Niu L, Sun W (2017) Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 919–923. https://doi.org/10.1109/ICASSP.2017.7952290
106. Bar Y, Diamant I, Wolf L, Lieberman S, Konen E, Greenspan H (2015) Chest pathology detection using deep learning with non-medical training. In: 2015 IEEE 12th international symposium on biomedical imaging (ISBI), pp 294–297. https://doi.org/10.1109/ISBI.2015.7163871

107. Chen D, Niu J, Pan Q, Li Y, Wang M (2017) A deep-learning based ultrasound text classifier for predicting benign and malignant thyroid nodules. In: 2017 International conference on green informatics (ICGI), pp 199–204. https://doi.org/10.1109/ICGI.2017.39
108. Al-Shabi M, Lan BL, Chan WY, Ng KH, Tan M (2019) Lung nodule classification using deep local-global networks. Int J Comput Assist Radiol Surg 14:1815–1819. https://doi.org/10.1118/1.3633941
109. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 7794–7803. https://doi.org/10.1109/CVPR.2018.00813
110. Zhang H, Goodfellow I, Metaxas D, Odena A (2019) Self-attention generative adversarial networks. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, vol 97. PMLR, pp 7354–7363
111. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778. https://doi.org/10.1109/CVPR.2016.90
112. Armato SGEA (2011) The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 38(2):915–931. https://doi.org/10.1118/1.3528204
113. Tran GS, Nghiem TP, Nguyen VT, Luong CM, Burie JC (2019) Improving accuracy of lung nodule classification using deep learning with focal loss. J Healthc Eng. https://doi.org/10.1155/2019/5156416
114. Lin T, Goyal P, Girshick R, He K, Dollar P (2020) Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 42(02):318–327. https://doi.org/10.1109/TPAMI.2018.2858826
115. Song QZ, Zhao L, Luo XK, Dou XC (2017) Using deep learning for classification of lung nodules on computed tomography images. J Healthc Eng 2017:1–7. https://doi.org/10.1155/2017/8314740
116. Vaid S, Kalantar R, Bhandari M (2020) Deep learning Covid-19 detection bias: accuracy through artificial intelligence. Int Orthopaed 44:1539–1542. https://doi.org/10.1007/s00264-020-04609-7
117. Heidari M, Mirniaharikandehei S, Z, KA, Danala G, Qiu Y, Zheng B (2020) Improving the performance of cnn to predict the likelihood of Covid-19 using chest X-ray images with preprocessing algorithms. Int J Med Inform 144:104284. https://doi.org/10.1016/j.ijmedinf.2020.104284
118. Civit-Masot J, Luna-Perejón F, Domínguez Morales M, Civit A (2020) Deep learning system for Covid-19 diagnosis aid using x-ray pulmonary images. Appl Sci 10(13):4640. https://doi.org/10.3390/app10134640
119. Misra S, Jeon S, Lee S, Managuli R, Jang IS, Kim C (2020) Multi-channel transfer learning of chest X-ray images for screening of Covid-19. Electronics 9:66. https://doi.org/10.3390/electronics9091388
120. Cohen JP, Morrison P, Dao L (2020) Covid-19 image data collection: prospective predictions are the future. arXiv arXiv:2003.11597
121. Ouyang X, Huo J, Xia L, Shan F, Liu J, Mo Z, Yan F, Ding Z, Yang Q, Song B, Shi F, Yuan H, Wei Y, Cao X, Gao Y, Wu D, Wang Q, Shen D (2020) Dual-sampling attention network for diagnosis of Covid-19 from community acquired pneumonia. IEEE Trans Med Imaging 39:2595–2605. https://doi.org/10.1109/TMI.2020.2995508
122. Sampath V, Maurtua I, Aguilar Martín JJ, Gutierrez A (2021) A survey on generative adversarial networks for imbalance problems in computer vision tasks. J Big Data. https://doi.org/10.1088/1742-6596/1693/1/012160

Ramzi Guetari is an Associate Professor in Computer Science at the Polytechnic School of Tunisia. He obtained his PhD in Computer Science from the University of Savoie in France and worked as a Research Engineer at the French National Institute for Research in Computer Science and Control (INRIA) where he was seconded to the World Wide Web Consortium (W3C). He participated in the development and promotion of Web technologies and notably participated in the development of Amaya, the W3C test bed. He was at the origin of the internationalization of Amaya. For more than 10 years, Ramzi has been working in the industrial field and has led large-scale missions for international organizations as well as for important other major international companies. Ramzi's research work has focused on distributed information systems and fundamental computing. Over the past ten years, he has worked on artificial intelligence and machine learning and has supervised a substantial number of Master and Doctoral theses.

Helmi Ayari obtained a Master Degree in Intelligent Imaging and Artificial Vision Systems from the Tunisian Higher Institute of Informatics (ISI) in 2021. His Master thesis was supervised by Dr. Ramzi Guetari. He worked on Computer-Aided Diagnosis Systems and in particular on feature extraction and dimentionality reduction. He is currently pursuing a PhD at the Polytechnic School of Tunisia. His main research area is financial and credit risk assessment methods using machine learning.

Houneida Sakly is a PhD and Engineer in Medical Informatics. She is a member of the research program "Deep Learning Analysis of Radiologic Imaging" in Stanford university. Certified in Healthcare Innovation with MIT-Harvard Medical school. Her main field of research is the Data Science applied to the Healthcare. She is a member of the Integrated Science Association (ISA) at the Universal Scientific Education and Research Network (USERN) in Tunisia. Currently, she is serving as a lead editor for various book and special issue in the field of Digital Transformation and Data Science in Healthcare. Recently, she has won the Best Researcher Award in the International Conference on Cardiology and Cardiovascular Medicine- San Francisco, United States.