A Review of Adrenal Lesions Diagnosis with Machine Learning

Bernardo Gonçalves

July 20, 2023

Abstract

Adrenal glands play a vital role in maintaining homeostasis under chronic stressors. They are susceptible to a range of malignant and benign lesions, including adrenal adenomas, which are the most common lesions. The diagnosis of these lesions is made via conventional imaging techniques, such as magnetic resonance imaging and computed tomography, is complex and depends heavily on the radiologist's knowledge and experience. Misdiagnosis due to the presence of pseudo lesions, imaging features overlap, or incorrect technique selection can lead to unnecessary costs and examinations, as most adrenal lesions do not require treatment. Machine learning methods have been proposed to improve adrenal lesion diagnosis. In this paper, we analyse all studies from 2017 until September 2022 that aim to diagnose or differentiate adrenal lesions using MRI or CT scans and machine learning methods. The studies that were not written in English, that were state-of-the-art reviews, that did not report which machine learning model was used, or that did not have the full text available were excluded from our analysis. For the sake of clarity, we divided our analysis into three categories accordingly to the goal of the studies: differentiation between adrenal adenomas and other lesions, differentiation between benign and malignant lesions, and studies that did not fit in any of the other groups. Despite the promising results that were reported, our analysis highlights the lack of deep learning studies, prospective studies, multicenter validations, and comparisons between the performance of the machine learning model and the radiologist's performance. Therefore, this paper serves as a comprehensive review of the current state-of-the-art in ML-based adrenal lesion diagnosis, while also identifying important research gaps that require further investigation.

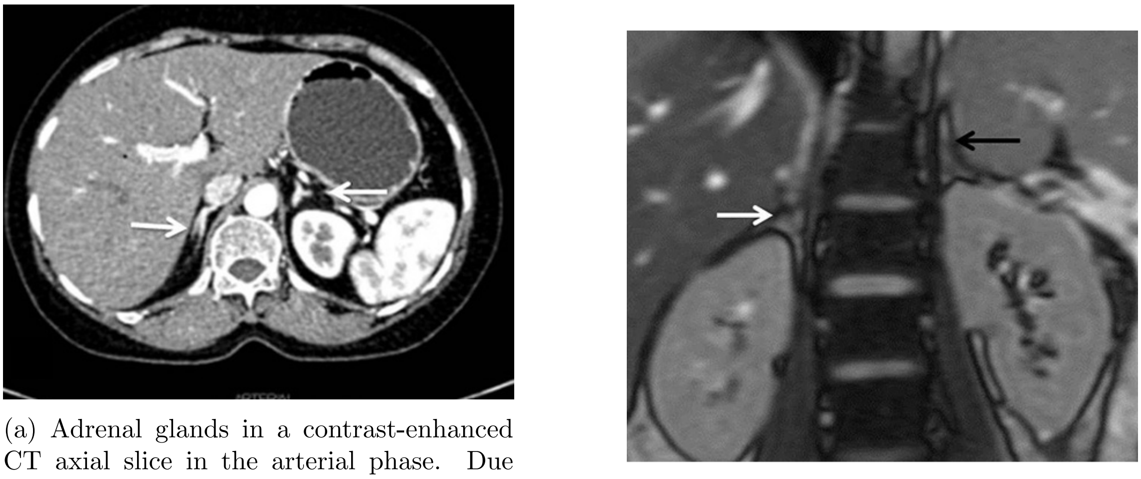
Keywords— Adrenal, Lesions, Machine Learning, Diagnosis, MRI, CT, Detection

1 Introduction

The adrenal glands, or suprarenal glands, are a component of the Hypothalamic-Pituitary-Adrenal (HPA) axis, which is responsible to maintain homeostasis in the presence of chronic stressors, activating a complex range of responses from the endocrine, nervous and immune systems, generally known as the stress response [1].

The adrenal glands can be affected by a wide variety of benign and malignant lesions. Approximately 9% of the global population is estimated to have adrenal lesions, which are mostly detected incidentally during abdominal imaging [2]. Adrenal adenomas represent 50 to 80% of all adrenal lesions [3]. Adenomas are often non-functional and remain asymptomatic, being discovered incidentally [4]. Adrenocortical carcinomas, despite being the most common primary adrenal lesion, are very rare, representing only 0.7–2.0 cases per million habitants per year [3]. Also, the adrenals are a frequent location of metastases [4], approximately 25% of patients with cancer have adrenal metastases on autopsy [3].

In general, non-functional lesions do not require any treatment, therefore it is crucial to differentiate between adenomas (typical non-functional lesions) and non-adenomas [4], to avoid unnecessary treatment. Commonly, adrenal adenomas have less than 1 cm in diameter and they can be lipid-rich or lipid-poor [5]. About 70-80 % of the adenomas are lipid-rich in contrast with the malignant lesions [4]. This results in a 20-30 % overlap between adenomas and malignant lesions in terms of intracytoplasmic lipid content [6].



(a) Adrenal glands in a contrast-enhanced CT axial slice in the arterial phase. Due to the high level of retroperitoneal fat, both glands are enhanced in this image slice. Reprinted from [5].

(b) Adrenal glands in an MR CSI coronal slice. Both glands have an intermediate signal intensity. Reprinted from [5].

Figure 1: Normal adrenal glands in CT and MR slices. The arrows indicate the localization of the glands.

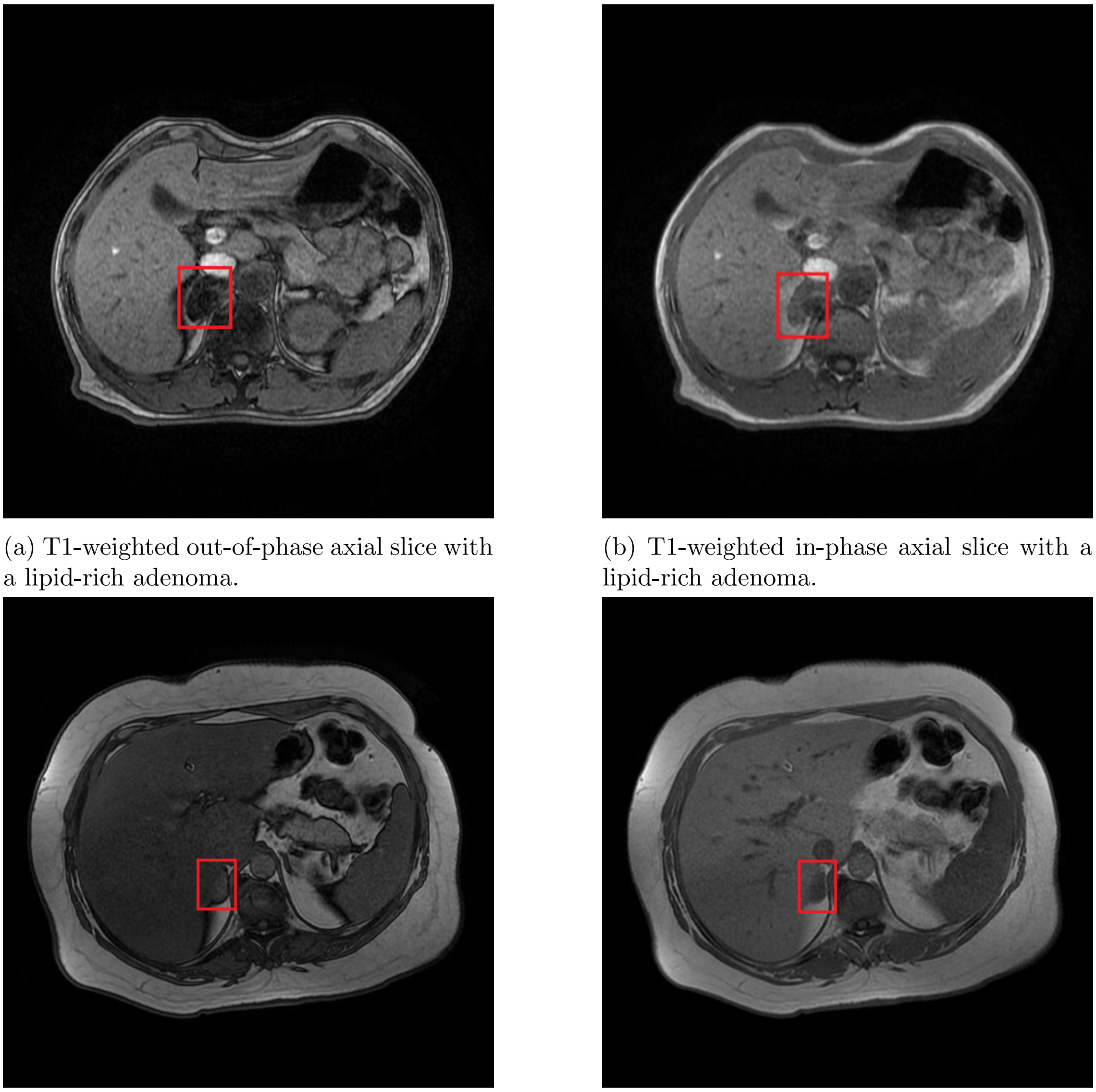
Structural medical imaging techniques are decisive in detecting and characterising adrenal lesions and complementary to functional imaging and endocrine evaluation in the assessment of functional lesions. Imaging techniques can also rule out invasive interventions. The most used imaging techniques to evaluate the adrenal glands are Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) [5].

Figure 1 shows the V- (left) and Y-shaped (right) normal glands. Figure 1a is an axial contrast-enhanced CT image in the arterial phase where both glands are enhanced due to the high retroperitoneal fat content. Figure 1b is a coronal MR Chemical Shift Image (CSI) out-of-phase showing normal adrenal glands as well.

The differential diagnosis between adenomas and non-adenomas is of extreme importance. This diagnosis can be hindered by the existence of lipid-poor adenomas that are more difficult to diagnose. Lipid-rich adenomas can be easily identified using unenhanced CT (less than 10 HU) [5] or CSI [4]. However, unenhanced CT is not indicated to detect lipid-poor adenomas [6]. In these cases, CSI presents itself as a better solution because of its improved sensitivity to low levels of lipid content and therefore it can detect 62-67% of the adenomas uncharacterized by unenhanced CT [6].

CSI is a fat-suppression technique that originates two sets of images: in-phase (IP) and out-of-phase (OP) images. In OP images the signal is the difference between the signals of water and fat molecules. In IP images the signal of both water and fat is added. Thus, there is a significant suppression of the signal from IP to OP images in lipid-rich lesions [7]. OP images are characterised by the so-called India ink artefact, which is a signal void in the margins of fatty and normal tissues [7], creating a darker boundary in lipid-rich lesions such as most adenomas. Figure 2 shows two adenomas, one lipid-rich (up) and the other lipid-poor (down) using CSI. The red boxes surround the adenomas. The signal difference in the region of the lesion between IP and OP images is much greater in lipid-rich adenomas than in lipid-poor adenomas, facilitating the diagnostic process of lipid-rich adenomas. Given these images, the diagnosis of adenoma can be made by visual evaluation or by quantitative indices such as the adrenal Signal Intensity Index (SII), Adrenal-to-Spleen Ratio (ASR), Adrenal-to-Liver Ratio (ALR) or the Adrenal-to-Muscle Ratio (AMR) [8]. A metanalysis with 1280 lesions (859 adenomas, 421 non-adenomas) documented a sensitivity of 94% and a specificity of 95% in detecting adrenal adenomas with the CSI technique using visual evaluation and/or quantitative methods [4]. Despite the elevated values, the authors assert that lipid-poor adenomas continue to present challenges, even when employing CSI.

When analysing lipid-poor adenomas, the addition of Dynamic Contrast-Enhanced (DCE) sequences is favourable, increasing the diagnostic performance [9]. In [10] 35 adrenal adenomas were analysed and it was concluded that the enhancement pattern of the adenomas is different from the one presented by malignant lesions. Adenomas present a homogeneous capillary blush in 18 seconds post-gadolinium images and a rapid washout in 45 seconds post-gadolinium images. In [4] it was concluded that the quantitative methods do not present a significant advantage to



(a) T1-weighted out-of-phase axial slice with a lipid-rich adenoma.

(b) T1-weighted in-phase axial slice with a lipid-rich adenoma.

(c) T1-weighted out-of-phase axial slice with a lipid-poor adenoma.

(d) T1-weighted in-phase axial slice with a lipid-poor adenoma.

Figure 2: Adrenal adenomas in axial MR CSI. The red rectangles surround the adenomas. Lipid-rich adenomas have a much greater intensity difference between in-phase and out-of-phase images.

the visual diagnosis. In addition, the authors do not recommend any additional imaging if the adenoma diagnosis is confirmed based on CSI. However, if the adenoma is not confirmed, DCE sequences would help distinguish between adenomas and malignant lesions [4].

The growing number of abdominal imaging studies has led to the increasing frequency of adrenal incidentalomas, which are usually benign, non-functional adenomas. Nevertheless, it is important to evaluate their functional status and malignancy, as soon as possible. The diagnosis of adrenal lesions is a complex process that involves both biochemical and radiological evaluation [11]. Adrenal radiologic evaluation via conventional imaging is a challenging process that depends largely on the experience and knowledge of the radiologist [12]. Several pitfalls can result in the misdiagnosis of adrenal lesions, such as the presence of pseudo lesions, overlap of imaging features of different lesions, or incorrect choice of the imaging technique [13]. New approaches to the diagnosis of adrenal lesions are crucial to avoid misdiagnosis, which can lead to increased treatment costs or unnecessary examination [12]. Additionally, it would be important to decrease the number of imaging exams that are necessary to complete the diagnosis. Machine learning methods have been proposed as a potential solution not only to improve the diagnostic accuracy and efficiency of adrenal lesions. Most of these machine learning methods analyse medical imaging features extracted using radiomic methods that will be briefly explored in **??**.

Radiomics is the extraction of quantitative features from medical images, such as Positron Emission Tomography (PET), MRI, or CT. Before the feature extraction is necessary to limit the amount of data that needs to be processed in order to extract features - ROI (region of interest) segmentation. This process can be manual, which is the gold standard, where a specialist selects the ROI. Manual segmentation is a time-consuming task that significantly depends on the skill of the operator. There are fully automatic methods for ROI segmentation however, they can fail in difficult cases, such as lesions with indistinct borders and are highly dependent on the quality of the image. For that reason, the usage of semi-automatic methods is preferable. These methods have minimal user interaction (seed identification or manual correction). The extracted quantitative features aim to describe the complexity of the individual region of interest. Ordinarily, these features are divided into 4 categories:

1. Shape-Based Features: numeric information respecting geometrics characteristics, like shape and size.

2. First-Order Statistics: distribution of voxel values without spatial information, generally histogram-based.

3. Second-Order Statistics: "texture" features, focus on the spatial relationships between voxels with similar grey levels.

4. High-Order Statistics: usage of filters to extract patterns from the images. From the resultant images, first and second-order features are extracted.

The most relevant radiomic features to the task in hand are selected using statistical approaches or machine learning [12]. Then, these features are used as the input of ML models to classify the region of interest. This type of workflow is widespread, appearing in 87% (20/23) of the analysed studies. Traditional machine learning models like K-means, Support Vector Machine (SVM), Logistic Regression (LogReg), or Random Forest (RanFor), are frequently used to classify radiomic features of regions of interest and have achieved high performance in different anatomical regions [12], [14]. Deep Learning models, such as Convolutional Neural Networks (CNNs), have been often applied to medical images from several anatomical regions with promising results [15]. Despite that, only 13% of the presented studies report the application of DL models to adrenal images. DL models are different from ML models mostly because they demand bigger annotated data sets and they do not rely on the feature extraction step (all features are automatically extracted and classified by the model).

2 Methods

The studies analysed in this section were selected according to the following PICO criteria:

P (patients) – patients with adrenal lesions.

I (interventions) – machine learning (including deep learning) modelling.

C (comparison) – standard of care imaging including Computed Tomography (CT) and Magnetic Imaging Resonance (MRI).

O (outcome) – lesion differentiation (benign/malign and subtyping) and lesions detection.

The studies were obtained by searching PubMed and Web of Science databases in June 2023. The following research string was used: (adrenal or suprarenal) AND (CT OR ”computed tomography” OR MRI OR ”magnetic resonance imaging” OR ”MRI scan” OR ”nuclear magnetic resonance” OR ”magnetic resonance” OR NMR) AND (”deep learning” OR ”convolutional networks” OR CNN OR ”neural networks” OR convolutional OR DNN OR SVM OR ”Support vector machine” OR ”decision tree” OR ”machine learning”). Studies that were: (a) reviews, (b) not written in English, (c) did not report a modelling method, (d) did not have the full text available were excluded from this research. The publication dates of the studies range from 2017 until June 2023.

The research resulted in 23 studies that were divided into 3 groups according with their object of study:

Group A: contains all the studies that focus on the differentiation between adrenal adenomas and other adrenal lesions.

Group B: contains all the studies that target the differentiation between benign and malignant adrenal lesions.

Group C: contains the remaining studies that did not fit any of the above categories.

Table 1 shows the distribution of the selected studies in terms of their group, image modality and model type. Overall, most of the studies adopt traditional machine learning models to classify imaging features (radiomics) from CT images to distinguish adenomas from other types of adrenal lesions, eg. Metastases, pheochromocytomas.

Group	**Image Modality**			**Model Type**			**Total**
	MRI	**CT**	**MRI+CT**	**ML**	**DL**	**ML+DL**	
A - Adenomas vs other lesions	4	7	1	11	1	0	12
B - Benign vs Malign	2	5	0	7	0	0	7
C - Other	1	3	0	2	1	1	4
Sum	7	15	1	20	2	1	23

Table 1: Studies distribution per group. CT: Computed Tomography; MRI: Magnetic Resonance Imaging; ML: Traditional Machine Learning; DL: Deep Learning.

In the next 3 sections, each group of papers will be analysed in detail, exploring their common and contrasting aspects. For each group, the analysis will be focused on the utilized data sets, models, and the obtained results.

3 Results

3.1 Group A - Adenomas vs other lesions

Group A comprises all studies that focus on the differentiation between adenomas and other lesions. This group corresponds to 52% of all analysed studies. Tables 2, 3 and 4 present an overview of the data sets, models, and results, respectively, for each study within group A.

Adenomas are the common lesion in all of these studies and they are compared with three different lesions: metastases [16]–[18]; pheochromocytomas [19]–[22], and carcinomas [23]–[25]. In [18]–[21] the data sets of adenomas consisted only in lipid-poor adenomas. Only one study had a non-binary data set with 3 classes: [26] data set has 3 classes: lipid-poor adenomas, lipid-rich adenomas and non-adenomas. The author of [27] do not specify another lesion type, performing differentiation between adenomas and non-adenomas. Most of the data sets (8 of 12 studies) contain CT images. From those, 6 use both images with and without contrast. The remaining have MRI datasets, all with CSI and T2W images. The sample size of the studies ranges from dozens to hundreds of lesions, which is closely related to the applied inclusion criteria and the initial sample size. For example, one study had a database of 336 patients however only 19 met the inclusion criteria, each with one lesion [24]. Most of the presented data sets are unbalanced with a much higher number of adenomas.

Reference	Image Modality	Sample Size (lesions)		
		Total	Adenomas	Other
[17]	U-CT	76	36	40
[19]	U/CE-CT	110	80	30
[20]	U-CT	265	181	84
[23]	CE-CT	54	25	29
[24]	U/CE-CT	19	9	10
[27]	U/CE-CT	115	83	32
[21]	U/CE-CT	280	188	92
[25]	U/CE-CT; T1W-OP/IP MRI	23	15	8
[22]	T1W-OP/IP; T2W MRI	60	40	20
[16]	T1W-OP/IP; T2W MRI	44	29	15
[18]	T1W-OP/IP; T2W MRI	63	23	40
[26]	T1W-OP/IP; T2W MRI	60	40	20

Table 2: Dataset Details for each article in Group A. CT: Computed Tomography; U: Unenhanced; CE: Contrast Enhanced; MRI: Magnetic Resonance Imaging; OP: Out-of-phase; IP: In-phase; T1W: T1-weighted; T2W: T2-weighted.

All studies, except one, performed lesion classification with ML models using radiomic features. The most frequent model is logistic regression (LogReg), followed by the support vector machine (SVM) and decision tree-based models. In every study, the region of interest (ROI) was manually selected by experts. The extraction of first and second-order statistics is a widespread practice, but only 3 studies extracted shape-based features, and none extracted higher-order statistics. The only study that has implemented a DL model has performed ROI (selected, cropped and labelled by experts) classification with a deep convolution neural network [27]. They reported the usage of augmentation techniques such as rotations and horizontal flips. There is only one study that implemented an unsupervised model (K-means) [24].

In [25] the data set consists of lipid-poor adenomas and carcinomas in both MRI and CT images. The objective of this work was to compare the different image modalities using the same machine learning approach. The authors have reported only the Area Under the Receiving Operating Characteristic Curve (AUC). The value presented in Table 4 refers to the best result, using CE-CT images and with MRI images the value decreases to 58 %.

Both works by Yi et al [19], [20] implemented a logistic regression model with the same radiomic features but using different CT images and achieved impressive results. However, [20] adds clinical features such as necrosis or calcification and lesion dimensions, to the radiomic features. Also, to improve feature selection the authors used the Least Absolute Shrinkage and Selection Operator (LASSO). These studies achieved the best results for this group. However, in terms of sensitivity, a higher value was obtained by [20]. Overall, there are 4 studies in which their metrics average more than 90%: [16], [19], [20], [27]. Three of them use CT images. [16] had a very small unbalanced dataset and used CSI images to differentiate renal cell carcinomas from adenomas.

3.2 Group B - Malign vs benign lesions

Group B consists of studies that aim at the differentiation between benign and malignant lesions. This group includes 30% of the studies. Tables 5, 6 and 7 display an overview of the data sets, models, and results, respectively, for each study within group B.

In terms of the image modalities in the data sets, this group is similar to group A. There are more studies that use CT images and all of them, except one, use CE-CT. Studies that analyse MRI data sets have both T1W chemical shift images and T2W images. The number of lesions analysed in each study varies from dozens to hundreds like in group A and most data sets are unbalanced. The study by Shoemaker et al had the largest and most balanced dataset of the analysed studies [28].

Unlike group A, most of the studies in group B implement semi-automatic region of interest segmentation. All studies except [29] combine different radiomic features. In [29] only second-order statistics are used as input of a Bayesian Spatial Gaussian Classifier. Group B does not include

Reference	**Type**	**Classification Model**	**ROI**	**Features**
[17]	ML	LogReg	Manual	1^{st}
[19]	ML	LogReg	Manual	1^{st}, 2^{nd}, higher
[20]	ML	LogReg	Manual	1^{st}, 2^{nd}, higher
[23]	ML	RanFor; LogReg	Manual	1^{st}, 2^{nd}, shape
[24]	ML	K-Means	Manual	1^{st}, 2^{nd}
[27]	DL	DCNN	Manual	-
[21]	ML	LinReg; SVM; RanFor	Manual	1^{st}; clinical
[25]	ML	LogReg	Manual	1^{st}, 2^{nd}, shape
[22]	ML	SVM	Manual	1^{st}
[16]	ML	LogReg	Manual	1^{st}
[18]	ML	LogReg	Manual	1^{st}, shape
[26]	ML	DecTre	Manual	1^{st}, 2^{nd}

Table 3: Modelling Details for each article in the Group A. ML: Traditional Machine Learning models; DL: Deep Learning models; LogReg: Logistic Regression; LASSO: Least Absolute Shrinkage and Selection Operator; DecTre: Decision Tree; RanFor: Random Forest; PCA: Principal Components Analysis; SVM: Support Vector Machine; 1^{st}, 2^{nd}, higher: first, second, higher order statistics, respectively; shape: shape-based features.

Reference	**Specificity - %**	**Sensitivity - %**	**Accuracy - %**	**AUC - %**
[17]	75.0	47.5	60.5	65.0
[19]	97.5	86.2	94.4	95.2
[20]	90.3	95.5	92.0	95.7
[23]	83.0	81.0	82.0	89.0
[24]	90.0	87.5	88.9	-
[27]	96.0	87.0	94.0	-
[21]	86.6	89.2	87.5	-
[25]	-	-	80.0	-
[22]	-	-	85.0	91.7
[16]	86.2	93.3	88.6	97.0
[18]	100	75.0	84.1	-
[26]	-	-	80.0	-

Table 4: Model metrics for each article in the Group A. AUC: Area Under the Receiving Operating Characteristic Curve.

Reference	Image Modality	Sample Size (lesions)		
		Total	Benign	Malign
[28]	U-CT	377	182	195
[30]	CE-CT	114	90	24
[29]	U/CE-CT	210	114	96
[33]	CE-CT	160	89	71
[34]	U/CE-CT	40	21	19
[31]	T1W-OP/IP; T2W MRI	122	112	10
[32]	T1W-OP/IP; T2W MRI	55	37	18

Table 5: Dataset Details for each article in Group B. CT: Computed Tomography; U: Unenhanced; CE: Contrast Enhanced; MRI: Magnetic Resonance Imaging; OP: Out-of-phase; IP: In-phase; T1W: T1-weighted; T2W: T2-weighted.

Reference	Type	Classification Model	ROI	Features
[28]	ML	LogReg	-	1^{st}, 2^{nd}
[30]	ML	NN	Semi-auto	1^{st}, 2^{nd}, higher, shape
[29]	ML	BayCla	Semi-auto	2^{nd}
[33]	ML	LogReg	Semi-auto	1^{st}, higher
[34]	ML	RanFor	Manual	1^{st}, 2^{nd}, higher, shape
[31]	ML	SVM	Manual; Semi-auto	2^{nd}, higher
[32]	ML	DecTre	Manual	1^{st}, 2^{nd}, higher, shape

Table 6: Modelling Details for each article in the Group B. ML: Traditional Machine Learning models; DL: Deep Learning models; LogReg: Logistic Regression; BayCla: Bayesian Classifier; NN: Neural Network; DecTre: Extra Trees Classifier. RanFor: Random Forest; SVM: Support Vector Machine; 1^{st}, 2^{nd}, higher: first, second, higher-order statistics, respectively; shape: shape-based features.

any study with a DL model, however, there are two studies that use neural networks [30], [31]. In [30] several optimisation algorithms for neural networks were experimented and the best results were achieved using the Bounded Particle Swarm Optimisation algorithm [1]. In [31] an SVM was implemented to perform binary classification, however, the authors also used a NN to perform type characterisation. The authors divided the dataset into 4 classes, each with one type of lesion, 3 benign (adenoma, cyst and lipoma) and 1 malign (metastasis). For both workflows, ROI selection was made using manual and semi-automatic segmentation, and the same radiomic features were used. The results presented in Table 7 refer to the binary classification using manual segmentation (the results using semiautomatic segmentation were worse), and they were the best results of this group. For the multiclass classification, the results were poor, despite the high values of specificity and accuracy, 96.2 % and 93.2 %, respectively, the sensibility is extremely low, 59.6 %, which can be explained by the high number of classes and the lack of balance in the data set. In this group, all the models are supervised learning models.

In this group there are only 2 studies where the average of the reported metrics is bigger than 90% - [31], [32], however, the last only reported accuracy and AUC. Both studies used the same type of images. In [32] applied an extra trees classifier to distinguish between benign and malign adrenal lesions without a drop in CS images. Extra Trees is a random ensemble of decision trees that is much faster than a RanFor and has differences in the input sampling and the selection of cut points [2].

[1] https://link.springer.com/chapter/10.1007/978-3-319-93025-1_2

[2] https://quantdare.com/what-is-the-difference-between-extra-trees-and-random-forest/

Reference	Specificity - %	Sensitivity - %	Accuracy - %	AUC - %
[28]	-	-	-	78.0
[30]	82.2	75.0	80.7	78.6
[29]	67.5	94.8	80.0	-
[33]	77.0	58.0	68.0	73.0
[34]	71.4	84.2	77.5	85.1
[31]	90.0	99.2	98.4	-
[32]	-	-	91.0	97.0

Table 7: Model metrics for each article in Group B. AUC: Area Under the Receiving Operating Characteristic Curve.

Reference	Image Modality	Sample size	Task
[35]	U-CT	38	Lesion Detection
[36]	CE-CT	229	Multiclass Classification
[37]	T2W-MRI	305	pheo vs non-pheo
[38]	U/CE-CT	83	Adenoma subtyping

Table 8: Dataset Details for each article in the Group C. CT: Computed Tomography; U: Unenhanced; CE: Contrast Enhanced; MRI: Magnetic Resonance Imaging; OP: Out-of-phase; IP: In-phase; T1W: T1-weighted; T2W: T2-weighted. Pheo: pheochromocytomas

3.3 Group C

Group C consists of the remaining studies that did not fit any of the prior defined groups. This group includes 18% of the studies. Tables 8, 9 and 10 display an overview of the data sets, models, and results, respectively, for each study inside group C.

In this group are 4 studies with distinct goals. [35] implements a fully convolutional neural network for lesion detection using a small data set of U-CT images. The network receives as input complete CT images and outputs the lesion probability map. To surpass the small data set issue, the authors applied traditional augmentation methods such as random crops and contrast variations and used a pre-trained network. The lesion probability map was then refined using a random walk-based algorithm.

A machine learning pipeline was created where the CNN embedding was used as input of an SVM to execute multiclass classification with a data set of CE-CT images [36]. The data set has 5 classes: carcinoma, non-functional adenoma, ganglioneuroma, myelolipoma and pheochromocytoma. A ResNet-101 pretrained in the ImageNet data set was used to create the feature embedding. To improve the toleration to intra-class variations the authors created a similarity feature learning module. Another relevant contribution was the usage of two DL networks each using a different CT image but with a weighted sharing strategy. The goal was to improve performance with a highly unbalanced data set.

[37] aimed at the differentiation between pheochromocytomas and non-pheochromocytomas with a T2W MRI dataset and a logistic regression model. The presented sample size includes an external and an internal data set. The external data set was used for external validation of the developed pipeline. [38] also used logistic regression to characterize adenomas with a CT dataset.

The results of the studies in this group cannot be compared due to the different objectives of each one. Nevertheless, the application of LogReg to analyse radiomic features is still a common practice that achieves impressive results even in multiclass classification.

4 Conclusions

When analysing the reported studies, one of the major difficulties was the lack of coherence between the metrics used to measure the performance of the developed models. The metrics selected in

Reference	Type	Model	ROI	Features
[35]	DL	FCN	Manual	-
[36]	DL + ML	CNN + SVM	Manual	CNN embedding
[37]	ML	LogReg	Semi-Auto	1^{st}, 2^{nd}, higher, shape
[38]	ML	LogReg	Manual	1^{st}, higher, shape

Table 9: Modelling Details for each article in Group C. ML: Traditional Machine Learning models; DL: Deep Learning models; LogReg: Logistic Regression; BSGC: Bayesian Spatial Gaussian Classifiers; CNN: Convolutional Neural Network; SVM: Support Vector Machine; 1^{st}, 2^{nd}, higher: first, second, higher-order statistics, respectively; shape: shape-based features.

Reference	Specificity - %	Sensitivity - %	Accuracy - %	AUC - %
[35]	-	76.29	-	-
[36]	95.9	83.7	85.2	-
[37]	75.0	85.7	84.0	90.6
[38]	92.8	91.5	92.2	90.2

Table 10: Model metrics for each article in the Group C. AUC: Area Under the Receiving Operating Characteristic Curve.

this review (specificity, sensitivity, accuracy and AUC) are common and relevant however, several studies did not report them, and some even reported only one metric, which is not enough to evaluate the performance of a predictive model. Indeed, the quality of reporting of predictive models is well-established as poor [12].

Almost all the reported studies stated that their datasets should be improved, because they were small or unbalanced or both, which proves the difficulty of finding a suitable medical dataset for machine learning proposes. Furthermore, most of the studies made single-centre retrospective studies and declared selection bias as a limitation. In fact, to better transfer these academic studies to clinical practice there are at least three shortcomings need to be tackled: lack of prospective studies, lack of external multicenter validation, and lack of comparison of diagnostic performance between the radiologist and the model [12].

There is a lack of studies using DL models to analyse adrenal imaging. Despite the impressive results of traditional ML models, such as the LogReg, DL models pose an advantage as they minimize the amount the feature engineering that is necessary to perform a good classification. In addition, DL models have achieved higher performance results than ML models in several medical imaging analysis tasks [39]. Another limitation related to traditional ML models is the necessity of segmentation methods to select the region of interest. From the analysed studies, it can be concluded that most of the studies used manual segmentation (which is time-consuming and prone to human error process), and those that compared manual and semi-auto segmentation stated that the models that classify features from ROI obtained with manual segmentation achieved the best results. DL models may play a crucial role in fully automatic adrenal lesion segmentation or detection as they have in other lesions and organs [39]. DL models can be trained for lesion detection or segmentation, where the lesion is classified and localised in the medical slice. To our knowledge only one paper performed adrenal lesion detection, which proves that this is still an emerging field.

The application of machine learning methods to the analysis of adrenal imaging has achieved outstanding results [16], [19], [20], [27], [31]. However, the usage of non-realistic datasets, the lack of validation, the extreme necessity of feature engineering and the lack of development of deep learning methods are still, shortcomings that need to be addressed to improve this field. To the best of our knowledge, this is the first review paper that compiles all studies that focus on the classification of adrenal lesions using traditional machine learning or deep learning models.

References

[1] OpenStaxCollege, *The adrenal glands – anatomy and physiology*, Last accessed 10 October 2022. [Online]. Available: http://pressbooks-dev.oer.hawaii.edu/anatomyandphysiology/chapter/the-adrenal-glands/.

[2] E. Dhamija, A. Panda, C. J. Das, and A. K. Gupta, *Adrenal imaging (part 2): Medullary and secondary adrenal lesions*, Jan. 2015. DOI: 10.4103/2230-8210.146859.

[3] B. Bracci, D. D. Santis, A. D. Gaudio, *et al.*, "Adrenal lesions: A review of imaging," *Diagnostics*, vol. 12, p. 2171, 9 Sep. 2022, ISSN: 2075-4418. DOI: 10.3390/diagnostics12092171. [Online]. Available: https://www.mdpi.com/2075-4418/12/9/2171.

[4] I. Platzek, D. Sieron, V. Plodeck, A. Borkowetz, M. Laniado, and R. T. Hoffmann, *Chemical shift imaging for evaluation of adrenal masses: A systematic review and meta-analysis*, Feb. 2019. DOI: 10.1007/s00330-018-5626-5. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30014203/.

[5] A. Panda, C. J. Das, E. Dhamija, R. Kumar, and A. K. Gupta, "Adrenal imaging (part 1): Imaging techniques and primary cortical lesions," *Indian Journal of Endocrinology and Metabolism*, vol. 19, pp. 8–15, 1 Jan. 2015, ISSN: 22309500. DOI: 10.4103/2230-8210.146858.

[6] G. M. Israel, M. Korobkin, C. Wang, E. N. Hecht, and G. A. Krinsky, "Comparison of unenhanced ct and chemical shift mri in evaluating lipid-rich adrenal adenomas," *American Journal of Roentgenology*, vol. 183, pp. 215–219, 1 2004, ISSN: 0361803X. DOI: 10.2214/ajr.183.1.1830215. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/15208141/.

[7] V. Jahanvi and A. Kelkar, "Chemical shift imaging: An indispensable tool in diagnosing musculoskeletal pathologies," *SA Journal of Radiology*, 2021, ISSN: 2078-6778. DOI: 10.4102/sajr. [Online]. Available: http://www.sajr.org.za.

[8] F. Fujiyoshi, M. Nakajo, Y. Fukukura, and S. Tsuchimochi, "Characterization of adrenal tumors by chemical shift fast low-angle shot mr imaging: Comparison of four methods of quantitative evaluation," *American Journal of Roentgenology*, vol. 180, pp. 1649–1657, 6 Jun. 2003, ISSN: 0361-803X. DOI: 10.2214/ajr.180.6.1801649. [Online]. Available: https://www.ajronline.org/doi/10.2214/ajr.180.6.1801649.

[9] M. Barat, A.-S. Cottereau, S. Gaujoux, *et al.*, "Adrenal mass characterization in the era of quantitative imaging: State of the art," *Cancers*, vol. 14, p. 569, 3 Jan. 2022, ISSN: 2072-6694. DOI: 10.3390/cancers14030569. [Online]. Available: https://www.mdpi.com/2072-6694/14/3/569.

[10] J. J. Chung, R. C. Semelka, and D. R. Martin, "Adrenal adenomas: Characteristic postgadolinium capillary blush on dynamic mr imaging.," *Journal of magnetic resonance imaging : JMRI*, vol. 13, pp. 242–8, 2 Feb. 2001, ISSN: 1053-1807. DOI: 10.1002/1522-2586(200102)13:2<242::aid-jmri1035>3.0.co;2-#. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11169830.

[11] P. Anagnostis, A. Karagiannis, K. Tziomalos, A. Kakafika, V. Athyros, and D. Mikhailidis, "Adrenal incidentaloma: A diagnostic challenge," *HORMONES*, vol. 8, pp. 163–184, 3 Jul. 2009, ISSN: 11093099. DOI: 10.14310/horm.2002.1233. [Online]. Available: http://www.hormones.gr/520/article/adrenal-incidentaloma:-a-diagnostic-challenge%E2%80%A6.html.

[12] H. Zhang, H. Lei, and J. Pang, "Diagnostic performance of radiomics in adrenal masses: A systematic review and meta-analysis," *Frontiers in Oncology*, vol. 12, Sep. 2022, ISSN: 2234-943X. DOI: 10.3389/fonc.2022.975183. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2022.975183/full.

[13] K. M. Elsayes, M. M. Elmohr, S. Javadi, *et al.*, "Mimics, pitfalls, and misdiagnoses of adrenal masses on ct and mri," *Abdominal Radiology*, vol. 45, pp. 982–1000, 4 Apr. 2020, ISSN: 2366-004X. DOI: 10.1007/s00261-019-02082-4. [Online]. Available: http://link.springer.com/10.1007/s00261-019-02082-4.

[14] M. W. Wagner, K. Namdar, A. Biswas, S. Monah, F. Khalvati, and B. B. Ertl-Wagner, "Radiomics, machine learning, and artificial intelligence—what the neuroradiologist needs to know," *Neuroradiology*, vol. 63, pp. 1957–1967, 12 Dec. 2021, ISSN: 0028-3940. DOI: 10.1007/s00234-021-02813-9. [Online]. Available: https://link.springer.com/10.1007/s00234-021-02813-9.

[15] A. Anaya-Isaza, L. Mera-Jiménez, and M. Zequera-Diaz, "An overview of deep learning in medical imaging," *Informatics in Medicine Unlocked*, vol. 26, p. 100 723, Jan. 2021, ISSN: 23529148. DOI: 10.1016/j.imu.2021.100723. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352914821002033.

[16] N. Schieda, S. Krishna, M. D. McInnes, *et al.*, "Utility of mri to differentiate clear cell renal cell carcinoma adrenal metastases from adrenal adenomas," *AJR. American journal of roentgenology*, vol. 209, W152–W159, 3 Sep. 2017, ISSN: 1546-3141. DOI: 10.2214/AJR.16.17649. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28742373/.

[17] W. Tu, R. Verma, S. Krishna, M. D. McInnes, T. A. Flood, and N. Schieda, "Can adrenal adenomas be differentiated from adrenal metastases at single-phase contrast-enhanced ct?" *American Journal of Roentgenology*, vol. 211, pp. 1044–1050, 5 Sep. 2018, ISSN: 15463141. DOI: 10.2214/AJR.17.19276. [Online]. Available: www.ajronline.org.

[18] W. Tu, J. Abreu-Gomez, A. Udare, A. Alrashed, and N. Schieda, "Utility of t2-weighted mri to differentiate adrenal metastases from lipid-poor adrenal adenomas," *Radiology: Imaging Cancer*, vol. 2, 6 Nov. 2020, ISSN: 2638616X. DOI: 10.1148/rycan.2020200011.

[19] X. Yi, X. Guan, C. Chen, *et al.*, "Adrenal incidentaloma: Machine learning-based quantitative texture analysis of unenhanced ct can effectively differentiate spheo from lipid-poor adrenal adenoma," *Journal of Cancer*, vol. 9, pp. 3577–3582, 19 2018, ISSN: 1837-9664. DOI: 10.7150/jca.26356. [Online]. Available: http://www.jcancer.org/v09p3577.htm.

[20] X. Yi, X. Guan, Y. Zhang, *et al.*, "Radiomics improves efficiency for differentiating subclinical pheochromocytoma from lipid-poor adenoma: A predictive, preventive and personalized medical approach in adrenal incidentalomas," *EPMA Journal*, vol. 9, pp. 421–429, 4 Dec. 2018, ISSN: 18785085. DOI: 10.1007/s13167-018-0149-3.

[21] H. Liu, X. Guan, B. Xu, *et al.*, "Computed tomography-based machine learning differentiates adrenal pheochromocytoma from lipid-poor adenoma," *Frontiers in Endocrinology*, vol. 13, Mar. 2022, ISSN: 1664-2392. DOI: 10.3389/fendo.2022.833413. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fendo.2022.833413/full.

[22] J. Liu, K. Xue, S. Li, Y. Zhang, and J. Cheng, "Combined diagnosis of whole-lesion histogram analysis of t1- and t2-weighted imaging for differentiating adrenal adenoma and pheochromocytoma: A support vector machine-based study," *Canadian Association of Radiologists Journal*, vol. 72, pp. 452–459, 3 Aug. 2021, ISSN: 0846-5371. DOI: 10.1177/0846537120911736. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0846537120911736.

[23] M. M. Elmohr, D. Fuentes, M. A. Habra, *et al.*, "Machine learning-based texture analysis for differentiation of large adrenal cortical tumours on ct," *CLINICAL RADIOLOGY*, vol. 74, 818.e1–818.e7, 10 Oct. 2019, ISSN: 0009-9260. DOI: 10.1016/j.crad.2019.06.021.

[24] F. Torresan, F. Crimì, F. Ceccato, *et al.*, "Radiomics: A new tool to differentiate adrenocortical adenoma from carcinoma," *BJS open*, vol. 5, 1 Jan. 2021, ISSN: 24749842. DOI: 10.1093/bjsopen/zraa061.

[25] L. M. Ho, E. Samei, M. A. Mazurowski, *et al.*, "Can texture analysis be used to distinguish benign from malignant adrenal nodules on unenhanced ct, contrast-enhanced ct, or in-phase and opposed-phase mri?" *American Journal of Roentgenology*, vol. 212, pp. 554–561, 3 Mar. 2019, ISSN: 0361-803X. DOI: 10.2214/AJR.18.20097. [Online]. Available: https://www.ajronline.org/doi/10.2214/AJR.18.20097.

[26] V. Romeo, S. Maurea, R. Cuocolo, *et al.*, "Characterization of adrenal lesions on unenhanced mri using texture analysis: A machine-learning approach," *Journal of Magnetic Resonance Imaging*, vol. 48, pp. 198–204, 1 Jul. 2018, ISSN: 10531807. DOI: 10.1002/jmri.25954. [Online]. Available: http://doi.wiley.com/10.1002/jmri.25954.

[27] M. Kusunoki, T. Nakayama, A. Nishie, *et al.*, "A deep learning-based approach for the diagnosis of adrenal adenoma: A new trial using ct," *The British Journal of Radiology*, vol. 95, 1135 Jul. 2022, ISSN: 0007-1285. DOI: 10.1259/bjr.20211066. [Online]. Available: https://www.birpublications.org/doi/10.1259/bjr.20211066.

[28] K. Shoemaker, B. P. Hobbs, K. Bharath, C. S. Ng, and V. Baladandayuthapani, "Tree-based methods for characterizing tumor density heterogeneity.," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 23, R. TE Altman, A. Dunker, L Hunter, M. Ritchie, T Murray, and Klein, Eds., pp. 216–227, 212669 Jan. 2018, ISSN: 2335-6936. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29218883http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5749399.

[29] X. Li, M. Guindani, C. S. Ng, and B. P. Hobbs, "Spatial bayesian modeling of glcm with application to malignant lesion characterization," *Journal of applied statistics*, vol. 46, pp. 230–246, 2 Jan. 2019, ISSN: 0266-4763. DOI: 10.1080/02664763.2018.1473348. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/31439980http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6706247.

[30] H. Koyuncu, R. Ceylan, S. Asoglu, H. Cebeci, and M. Koplay, "An extensive study for binary characterisation of adrenal tumours.," *Medical & Biological Engineering & Computing*, vol. 57, pp. 849–862, 4 Apr. 2019, ISSN: 01400118. DOI: 10.1007/s11517-018-1923-z. [Online]. Available: https://link.springer.com/article/10.1007/s11517-018-1923-z.

[31] M. Barstugan, R. Ceylan, S. S. S. Asoglu, H. Cebeci, and M. Koplay, "Adrenal tumor characterization on magnetic resonance images," *International Journal of Imaging Systems and Technology*, vol. 30, pp. 252–265, 1 Mar. 2020, ISSN: 0899-9457. DOI: 10.1002/ima.22358. [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=141526975&site=eds-livehttps://onlinelibrary.wiley.com/doi/abs/10.1002/ima.22358.

[32] A. Stanzione, R. Cuocolo, F. Verde, *et al.*, "Handcrafted mri radiomics and machine learning: Classification of indeterminate solid adrenal lesions," *Magnetic Resonance Imaging*, vol. 79, pp. 52–58, Jun. 2021, ISSN: 0730725X. DOI: 10.1016/j.mri.2021.03.009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/33727148https://linkinghub.elsevier.com/retrieve/pii/S0730725X21000394.

[33] M. B. Andersen, U. Bodtger, I. R. Andersen, K. S. Thorup, B. Ganeshan, and F. Rasmussen, "Metastases or benign adrenal lesions in patients with histopathological verification of lung cancer: Can ct texture analysis distinguish?" *European Journal of Radiology*, vol. 138, p. 109 664, May 2021, ISSN: 0720048X. DOI: 10.1016/j.ejrad.2021.109664. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0720048X21001443.

[34] A. W. Moawad, A. Ahmed, D. T. Fuentes, J. D. Hazle, M. A. Habra, and K. M. Elsayes, "Machine learning-based texture analysis for differentiation of radiologically indeterminate small adrenal tumors on adrenal protocol ct scans," *Abdominal Radiology*, vol. 46, pp. 4853–4863, 10 Oct. 2021, ISSN: 23660058. DOI: 10.1007/S00261-021-03136-2.

[35] L. Bi, J. Kim, T. Su, M. Fulham, D. Feng, and G. Ning, "Adrenal lesions detection on low-contrast ct images using fully convolutional networks with multi-scale integration," IEEE Computer Society, Jun. 2017, pp. 895–898, ISBN: 9781509011711. DOI: 10.1109/ISBI.2017.7950660.

[36] L. Bi, J. Kim, T. Su, M. Fulham, D. D. Feng, and G. Ning, "Deep multi-scale resemblance network for the sub-class differentiation of adrenal masses on computed tomography images," *Artificial Intelligence in Medicine*, vol. 132, p. 102 374, Oct. 2022, ISSN: 09333657. DOI: 10.1016/j.artmed.2022.102374. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0933365722001336.

[37] J. Kong, J. Zheng, J. Wu, *et al.*, "Development of a radiomics model to diagnose pheochromocytoma preoperatively: A multicenter study with prospective validation," *Journal of Translational Medicine*, vol. 20, 1 Dec. 2022, ISSN: 14795876. DOI: 10.1186/s12967-022-03233-w.

[38] Y. Zheng, X. Liu, Y. Zhong, F. Lv, and H. Yang, "A preliminary study for distinguish hormone-secreting functional adrenocortical adenoma subtypes using multiparametric ct radiomics-based machine learning model and nomogram," *Frontiers in Oncology*, vol. 10, Sep. 2020, ISSN: 2234943X. DOI: 10.3389/fonc.2020.570502.

[39] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, "A review on deep learning in medical image analysis," *International Journal of Multimedia Information Retrieval*, vol. 11, pp. 19–38, 1 Mar. 2022, ISSN: 2192-6611. DOI: 10.1007/s13735-021-00218-1. [Online]. Available: https://link.springer.com/10.1007/s13735-021-00218-1.