Received May 7, 2020, accepted June 14, 2020, date of publication June 24, 2020, date of current version August 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3004766

A Comparison of Transfer Learning Performance Versus Health Experts in Disease Diagnosis From Medical Imaging

HASSAAN MALIK[1], MUHAMMAD SHOAIB FAROOQ[1], (Member, IEEE), ADEL KHELIFI[2], ADNAN ABID[1], (Member, IEEE), JUNAID NASIR QURESHI[1,3], AND MUZAMMIL HUSSAIN[1]

[1]Department of Computer Science, University of Management and Technology, Lahore 54000, Pakistan
[2]Department of Computer Science and Information Technology, Abu Dhabi University, Abu Dhabi, United Arab Emirates
[3]Department of Computer Science, Bahria University Lahore Campus, Lahore 54000, Pakistan

Corresponding author: Muhammad Shoaib Farooq (shoaib.farooq@umt.edu.pk)

ABSTRACT Deep learning methods have huge success in task specific feature representation. Transfer learning algorithms are very much effective when large training data is scarce. It has been significantly used for diagnosis of diseases in medical imaging. This article presents a systematic literature review (SLR) by conducting a comparison of a variety of transfer learning approaches with healthcare experts in diagnosing diseases from medical imaging. This study has been compiled by reviewing research studies published in renowned venues between 2014 and 2019. Moreover, the data for the diagnosis performed by health care experts has also been acquired to perform a detailed comparative analysis for a wide range of diseases. The analysis has been performed on the basis of diseases, transfer learning approaches, type of medical imaging used. The comparative analysis is based on performance indices reported in studies which include diagnostic accuracy, true-positive (TP), false-positive (FP), true-negative (TN), false-negative (FN) sensitivity, specificity, and the area under the receiver operating characteristic curve (AUROC). A total of5,188articles were identified out of which 63 studies were included. Among them 21 research studies contain sufficient data to construct the evaluation tables that enable process of test accuracy of transfer learning having sensitivity ranged from 71% to 100% (mean 85.25%) and specificity ranged from 64% to 100% (mean 81.92%). Furthermore, health experts having sensitivity ranged from 33% to 100% (mean 85.27%) and specificity ranged from 82% to 100% (mean 91.63%).This SLR found that diagnostic accuracy of transfer learning is approximately equivalent to the diagnosis of health experts. The results also revealed that convolutional neural networks (CNN) have been extensively used for disease diagnosis from medical imaging. Finally, inappropriate exposure of diseases in transfer learning studies restricts reliable elucidation of the outcomes of diagnostic accuracy.

INDEX TERMS Transfer learning, heath experts, disease, medical imaging, SLR.

I. INTRODUCTION

In MEDLINE (a bibliographic database of biomedical and life sciences), the first paper was indexed with the MeSH term ''artificial intelligence'' (AI) dates back to July 1951, when a tortoise robot presented in the seminal paper ''Matter with mind; a neurological research robot'' was presented by Fletcher [1]. Currently in the field of AI, a large of number of scientific articles has been published, with numerous in the lay press [2]. The AI has transformed and improve the Quality of Life (QoL) through smart and intelligent applications such as face tagging, natural language translation (NLT) and speech recognition [3], [4]. A lot of advancement has been achieved by the research community in the domain of healthcare using AI; particularly in finding of patterns or identification of diseases from medical imaging, some researcher even thinks that in near future AI will replace and revolutionized the classical medical diagnostic approaches

The associate editor coordinating the review of this manuscript and approving it for publication was Cristian A Linte.

and change the role of doctors from diagnostician to ''information specialists''.

Diagnosis is the art of finding or identifying the nature of an illness through amalgamation and collections of data which allows them to accurately classify the disease and refers the medical treatment. Human diagnosticians achieve an adequate accuracy in classifying the disease by practicing on evident medical cases in the supervised diagnostic process [5].

Medical imaging has been considered the significant sources of diagnostic, but it is reliant on human elucidation. The need for, and accessibility of, diagnostic medical images is rapidly exceeding the capability and competency of the available healthcare specialists, especially in low and middle-income countries [6]. Automated process of disease diagnosis from medical imaging through the smart gadgets of AI, mainly in the field of transfer learning, might be able to solve this dilemma [7]. Reports of pre-trained algorithms exceeding humans in diagnostic assessment have generated considerable elation.

Transfer learning is an application of AI based on pre-trained learning that provides significant enhancement in accuracy and rate of diagnosis through medical imaging. There is a sturdy public engrossing and market demands that are driving the rapid production of such diagnostic products. The approaches of transfer learning provide architecture to exploit previously acquired knowledge to unravel new but relevant issues much more expeditiously and effectively using AI [8]. The transfer learning algorithms have the feature to fine-tune the model on the basis of previously trained data, allowing them to adjust to their provided input layers. This feature makes them the powerful tool for identifying and cataloging the pattern of diseases. Furthermore, the revealed features have not processed by medical technologists, but to a certain extent by the sequences they have trained from input data [9].

In contrast, deep learning algorithms has achieved astonishing and significant deviations to the medical engineering, with their findings in the field of image caption, computer vision and pattern recognition [10], [11]. Until now, the deep learning has faced three major issues in disease diagnostic procedures. First, access to a large amount of well-curated and labeled medical image databases. Second, highly specialized computing equipment's has vital role because the evaluation of deep learning algorithms depends on the parallel computing architectures, known as ''graphic processing units'' (GPU) [12]. Third, technical and numerical expertise is required to implement the deep learning algorithms. Transfer learning has ability to conquer such issues, where an algorithm has designed for a particular problem is repurposed and leveraged as initial point for learning on novel task. While transfer learning moderates a few of substantial computing assets request in developing a custom-made algorithm from inception, it nonetheless demands deep learning proficiency to deliver efficient and effective results.

In this SLR, we have sought the latest development of disease diagnostic performance by pre-trained algorithms for medical imaging compared with clinical experts, considering study behavior, exposure, and clinical significance to the globe. We have conducted this SLR to evaluate the diagnostic accuracy of transfer learning algorithms associated with health experts.

To the best of our knowledge, there is no such published SLR comparing the disease diagnostic performance between transfer learning models and clinical experts. Thus, we aimed to study the literature and provide a modern summary indicating the scope of pre-trained algorithms to disease diagnoses compared with human diagnostician. A taxonomy of disease diagnosis has been proposed using medical imaging to compare the accuracy of transfer learning algorithms and discussed their results by comparing with diagnostic accuracy of health experts. Furthermore, the proposed model for disease diagnosis has been presented. We hope this SLR would help healthcare experts' consciousness and comprehension of pre-trained learning-related medical practices.

This SLR has been categorized as follows: Section II discusses the methodology for conducting SLR by defining the research questions, search string strategy, inclusion-exclusion criteria of selected articles and quality assessment; Section III contains the results of research question; Section IV presents the discussions on the obtained results. Moreover, the proposed taxonomy and model for the diagnosis of diseases have been presented in this section; and finally, the SLR has been concluded in Section V.

II. RESEARCH METHODOLOGY

Petersen *et al*. [13] described that the purpose of SLR is to present an overview of a research publication area, classify its number of publications, and types of research studies, and outcomes available within it. In this systematic study the primary goal is to calculate the number of publications over time to explore different research trends. Furthermore, also identify the fora in which field of research has been published. The flow of systematic study process is shown in Fig. 1, which covers the search strategy for relevant research publications, classification scheme, and the mapping of publications.

A. RESEARCH QUESTIONS

The overall purpose of our systematic study research is to gain insight into the all possible solutions designed to address the comparison of pre-trained algorithms performance versus health experts in classifying diseases from different kinds of medical imaging. In order to achieve a comprehensive review on this research topic, the systematic study consists of four research questions (RQ) which are described with their corresponding motivations in Table 1. These RQs will allow us to classify the existing research trends, and to identify upcoming research studies aspect.

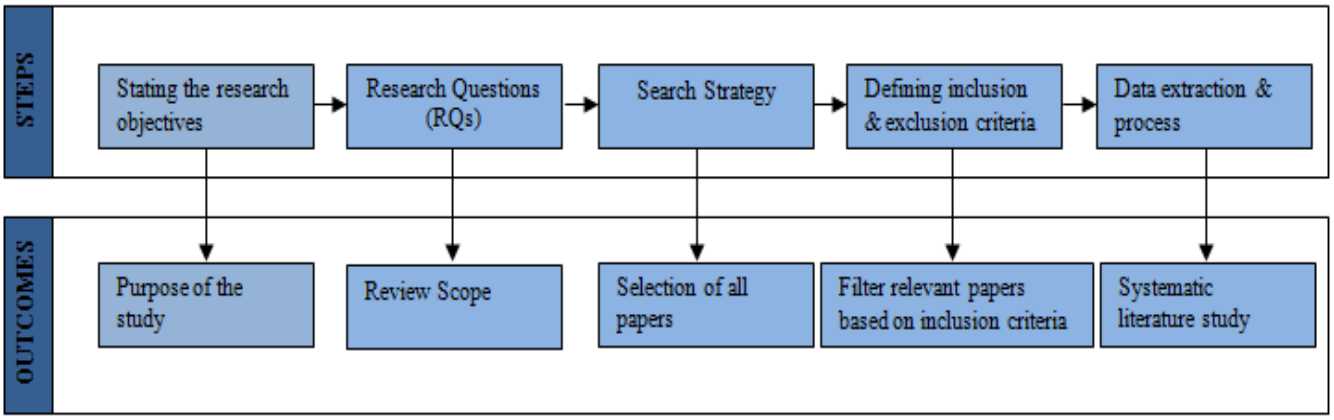

FIGURE 1. Systematic study process.

TABLE 1. Research questions.

No.	Research Question	Motivation
RQ1	What sort of benchmark datasets have been used in transfer learning based medical imaging research?	The purpose of this question is to identify the dataset used in medical imaging research for disease diagnosis.
RQ2	What types of classification methods / algorithms are used in transfer learning research in disease diagnosis?	The purpose of this question is to find the transfer learning methods and algorithms used for diagnosing diseases from medical imaging.
RQ3	What are the parameters / metrics on which the accuracy of classification methods / algorithms can be assessed in transfer learning for disease diagnosis from medical imaging?	The motivation behind this question is to identify the parameters / performance metrics of TL based algorithms like True Positive (TP) also known as Sensitivity or Recall, Misclassification Rate (Error Rate), False Positive (FP), True Negative (TN) also called Specificity, False Negative (FN), Precision, and ROC.
RQ4	What evidence / validation approach is there for disease diagnosis in health care sector?	The purpose of this question is to identify whether transfer learning studies of health care can be validated? Moreover, the purpose of this question is also to identify the experimental approach used for diagnosing different types of diseases.

B. SEARCH STRATEGY

In this systematic study, we searched for research studies that designed or validated a pre-trained model for the classification of any disease by using different medical or health imaging modalities. In additionally, authors compared the accuracy of diagnoses achieve by algorithms versus clinical or health experts. We searched 8 different databases such as Wiley library, IEEE Xplore, Ovid-MEDLINE, ACM Digital Library, Springer Link, Scopus, Science Direct, Taylor & Francis online and Conference Proceedings Citation Index for research studies published from January 2014, to December 2019, that validated pre-trained methods for the any kind of diseases diagnostics from medical imaging. We deliberately defined the cutoff of January 2014, to consider a conceded transform in the efficiency of algorithms with the advancement of transfer learning approaches. In 2014, Convolutional Neural Network (CNN) for object recognition designed and trained by Oxford's renowned Visual Geometry Group (VGG), enabled by modern concept of parallel computing architectures, made a significant breakthrough at the ''ImageNet Large- Scale Visual Recognition Challenge'' (ILSVRC) [14]. Manual searches were also done for related research studies, bibliographies and citations of selected articles were also undertaken to include any relevant papers that might have been missed during searches. We query the above mentioned well-reputed publication databases using a set of variant search keywords are shown in Fig.2.The primary search keywords were selected as key identifiers of study in the field of transfer learning. The secondary keywords were added to capture the research publications that identify the diagnostic accuracy of the health experts. Moreover, additional keywords were included to ensure detailed coverage. The retrieved results from the research publication portals consists of title of the papers, abstract, and publication outlet are stored in a personal knowledge base, which will be filtered according to the inclusion and exclusion criterion. The complete keywords of search string combinations for 8 different databases are mentioned in the Table 2.

C. STUDY SELECTION CRITERIA

The study selection process was done by identifying the most significant and relevant articles, which is also the initial objective of this systematic study. When the same research study obtained from more than one database, it was considered only once according to our search policy. Eligibility assessment of the selected papers was finalized by the authors who screened abstracts and titles of the search results

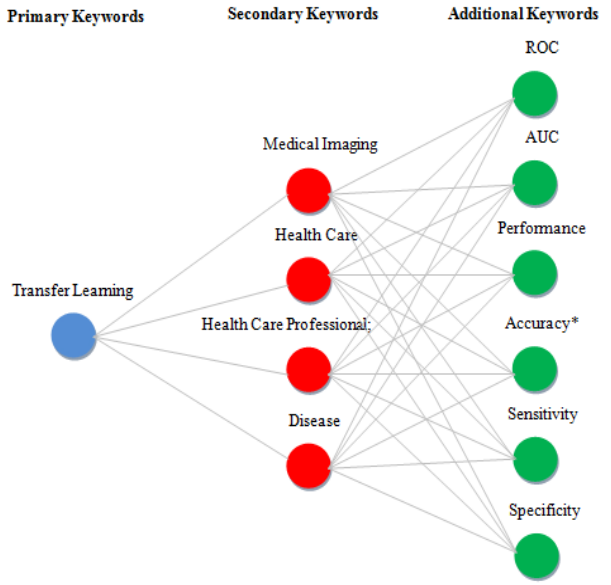

FIGURE 2. Search string keywords to identify the research studies.

TABLE 2. Search strategies for databases.

Database	Search Strategy
IEEE Xplore	((("Document Title": "transfer learning") AND ("Abstract":" roc" OR "auc" OR performance OR accuracy* OR sensitivity OR specificity) AND ("Abstract": "medical imaging" OR "health care" OR "health care professional" OR disease*)) Publication Year: 2014-2019
ACM Digital Library	(("transfer learning") AND ("roc" OR "auc" OR performance OR accuracy* OR sensitivity OR specificity) AND ("medical imaging" OR "health care" OR "health care professional" OR disease*)) Publication Year: 2014-2019
Ovid-Medline	(exp *transfer learning) or (roc or auc or performance or discriminate*) or ("health care" or "health care issues" or "disease*" or "health care professional" or "medical imaging")Publication Year: 2014-2019
Science Direct	Title, abstract, keywords: transfer learning roc auc performance accuracy* sensitivity specificity medical imaging health care health care professional disease* Publication Year: 2014-2019
Wiley Library	transfer learning roc auc performance accuracy sensitivity specificity medical imaging health care health care professional disease* Publication Year: 2014-2019
Springer Link	("transfer learning" AND ("roc" OR "auc" OR "performance" OR "accuracy* "OR "sensitivity" OR "specificity" OR "medical imaging" OR "health care" OR "health care professional" OR "disease*")) Publication Year: 2014-2019
Taylor & Francis Online	[All: "transfer learning"] AND [[All: "roc"] OR [All: "auc"] OR [All: "performance"] OR [All: "accuracy*"] OR [All: "sensitivity"] OR [All: "specificity"] OR [All: "medical imaging"] OR [All: "health care"] OR [All: "health care professional"] OR [All: "disease*"]] AND [Publication Date: (01/01/2014 TO 12/31/2019)]

independently. We did not apply any limitations on the target population, the disease outcome of interest. The inclusion criteria of the obtained studies were limited to the search string mentioned in Table 2. Moreover, the selected research studies that met at least one of the following exclusion criteria (EC) were excluded:

EC1. Papers which were not focused on binary classification of disease.

EC2. Diseases evaluated without medical imaging were excluded.

EC3. The studies investigating the image segmentation instead of image classification were excluded.

EC4. Studies based on non-human samples were excluded.

Fig. 3 shows the search process results. 63 studies were selected from 5,188 identified studies.

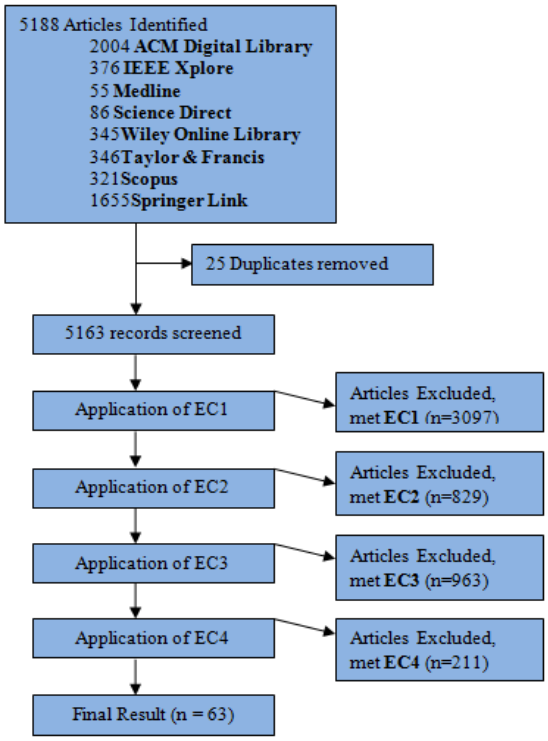

FIGURE 3. Studies selection process.

D. QUALITY ASSESSMENT

The quality assessment (QA) is usually carried out in systematic study but less in mapping studies. However, in order to evaluate the quality of our research, a questionnaire was designed to review the worth of the 63 included articles. The quality assessment (see Table 3) was calculated by the authors of this SLR.

Whether there exists a comparison of medical imaging data for health experts and transfer learning algorithms or otherwise? The possible answer to this question was "True (+1)" and "False (+0)."

Have the selected studies addressed an evident solution to the problems of disease diagnosis in medical imaging? The possible answer to this question was "True (+1)" and "False (+0)."

Are selected 63 research studies had been published in a recognized publication channel? This question was

TABLE 3. Quality Assessment of 63 selected studies.

Ref	Classification		Quality Assessment			
	Publication Channel	Year	a	b	c	score
[15]	Conference	2018	0	1	1.5	2.5
[16]	Journal	2019	0	1	2	3
[28]	Journal	2018	0	1	2	3
[60]	Journal	2018	0	1	2	3
[17]	Conference	2015	0	1	1.5	2.5
[18]	Journal	2017	0	1	2	3
[66]	Journal	2019	0	1	2	3
[19]	Journal	2018	1	1	2	4
[20]	Journal	2019	0	1	2	3
[21]	Journal	2018	1	1	2	4
[56]	Journal	2018	0	1	2	3
[57]	Journal	2018	0	1	2	3
[22]	Journal	2017	0	1	1.5	2.5
[23]	Journal	2019	0	1	2	3
[24]	Journal	2019	0	1	2	3
[25]	Journal	2018	0	1	2	3
[29]	Journal	2019	1	1	2	4
[61]	Journal	2019	1	1	2	4
[26]	Scientific Report	2017	1	1	2	4
[27]	Journal	2017	0	1	1.5	2.5
[30]	Letter	2017	0	1	2	3
[31]	Journal	2019	0	1	1	2
[32]	Journal	2019	1	1	2	4
[33]	Conference	2018	1	1	1.5	3.5
[34]	Journal	2018	0	1	2	3
[35]	Journal	2018	0	1	2	3
[77]	Journal	2018	0	1	2	3
[36]	Journal	2019	1	1	2	4
[37]	Journal	2019	1	1	2	4
[38]	Journal	2019	1	1	2	4
[39]	Journal	2018	0	1	0.5	1.5
[40]	Journal	2018	1	1	2	4
[41]	Journal	2017	1	1	2	4
[42]	Journal	2018	1	1	2	4
[43]	Journal	2019	1	1	1.5	3.5
[44]	Journal	2019	0	1	1.5	2.5
[45]	Journal	2018	0	1	1.5	2.5
[46]	Journal	2019	1	1	2	4
[47]	Journal	2018	1	1	2	4
[72]	Journal	2019	0	1	2	3
[48]	Journal	2019	0	1	2	3
[49]	Journal	2018	0	1	2	3
[50]	Journal	2019	0	1	2	3
[51]	Journal	2018	1	1	2	4
[70]	Journal	2019	0	1	2	3
[52]	Journal	2018	0	1	1.5	2.5
[53]	Journal	2018	1	1	2	4
[54]	Journal	2019	1	1	2	4
[55]	Journal	2019	0	1	1	2
[58]	Journal	2019	0	1	2	3
[59]	Journal	2019	0	1	2	3
[62]	Journal	2017	0	1	2	3
[63]	Journal	2019	0	1	2	3
[64]	Journal	2018	0	1	2	3
[65]	Journal	2018	1	1	2	4
[67]	Journal	2019	1	1	2	4
[68]	Journal	2019	0	1	2	3
[69]	Journal	2019	0	1	1.5	2.5
[71]	Journal	2017	0	1	2	3
[73]	Journal	2017	0	1	2	3
[74]	Journal	2019	0	1	1.5	2.5
[75]	Journal	2019	0	1	2	3
[76]	Journal	2019	0	1	2	3

answered by considering the Journal Citation Reports (JCR) 2019 with their quartile ranking such as Q1, Q2, Q3 and Q4. In addition, the computer science conference was ranked e.g. CORE
(A, B, and C).

The possible answers to this question were for conferences and workshops:

(+2) for CORE A,
(+1.5) for CORE B,
(+1) for CORE C,
(+0) Not present in CORE ranking.

The possible answers to this question were for journals, letters and scientific reports:

(+2) if it is quartile Q1,
(+1.5) if it is quartileQ2,
(+1) if it is quartileQ3,
(+0.5) if it is quartileQ4,
(+0) If it has no quartile ranking

The quality criterion (c) score is based on the fact that journal publications have more worth and value than conferences, workshops and seminars. Hence, the authors believe that publishing of research work in quartile ranked journals may be more complex and time consuming than in other publication channels.

III. RESULTS

This section answers the results of our four RQs described in Table 1. The QA score for each selected study listed in Table 3. Approximately 81% of the selected research studies obtain above average score, 14% of the articles hold average score, and 5% of the study hold the below average score. This QA score could also help the researchers and health experts to select the most significant and relevant papers for the diagnostics of diseases from medical imaging. We obtained the lists of all publication sources of included 63 studies, with their variant publication platforms, and the total numbers of articles per publication source are shown in Table 4. Two different types of publication platforms were observed such as letter and a scientific report. It has been calculated that 4.8 % of the selected research studies published in conferences, 1.6 % studies were presented as scientific report and research letter respectively. Moreover, 92.1 % of the included 63 research studies were published in journals. The overall distribution of all 63 selected studies (see Table 4) has been presented in Fig.4a. The journal papers have opted from Ovid-Medline has a ratio of 73%, IEEE has 4.8%, Wiley, Springer, Scopus, Taylor & Francis and Science direct, ACM digital library have a ratio 1.6% and 3.2% respectively. The journal wise distribution of the 63 included studies has been represented in Fig.4b.

A. SELECTION RESULTS

Our search identified 5,188 records, of which 5,163 were screened (Fig. 3). 63 studies were included in this systematic study[15]. These studies described lung cancer (8 studies), breast tumor (6 studies), diabetes (3 studies), knee injuries and Age-related macular degeneration (8 studies),

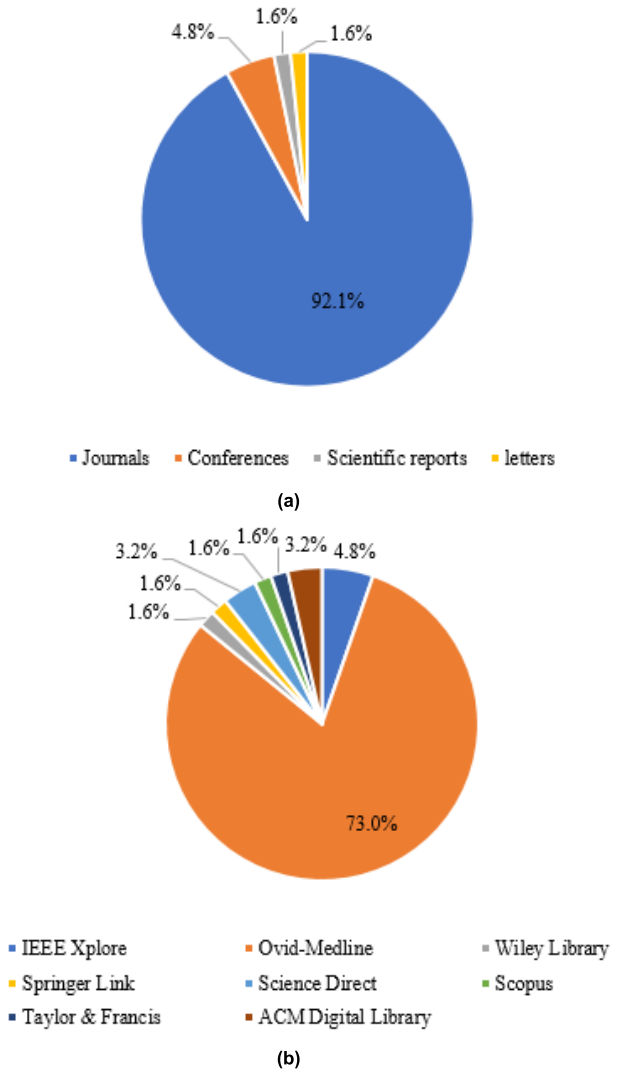

FIGURE 4. (a) Distribution of selected studies. (b) Journal-wise distribution of selected studies.

skin cancer (8 studies), retinopathy (4 studies), liver fibrosis (1 study), brain hemorrhage (2 studies), nasopharyngeal and thyroid cancer (6 studies), hip fractures (3 studies), Trauma and orthopaedics (1 study), Ophthalmology (1 study), oesophageal cancer (1 study), odontogenic tumors of the jaw(1 study), prostate cancer (1 study), femoral head osteonecrosis(1 study), alzheimer's disease (1 study), lymph node metastases (1 study), onychomycosis(1 study), spondylitis (1 study), Sjogren's syndrome (1 study), oesophageal cancer (1 study), helicobacter pylori gastritis (1 study) and Gastric cancer (1 study). Study characteristics are summarized in the Tables 5, 6, 7, and 8.

B. DATA EXTRACTION AND SYNTHESIS METHOD

The data extraction process was developed to provide the set of possible answer to the RQs listed in Table 1.

RQ1. What sort of benchmark datasets have been used in transfer learning based medical imaging research?

TABLE 4. Publication source.

Publication Source	Channel	Reference	No.	%
European Congress on Computational Methods in Applied Sciences and Engineering	Conference	[15]	1	1.6
Nature Medicine	Journal	[16][28][60]	3	4.7
2015 International Conference on Advances in Biomedical Engineering (ICABME)	Conference	[17]	1	1.6
Investigative Radiology	Journal	[18][66]	2	3.1
PLoS Medicine	Journal	[19]	1	1.6
European Journal of Cancer	Journal	[20]	1	1.6
JAMA Ophthalmology	Journal	[21][56][57]	3	4.7
Computers in Biology and Medicine	Journal	[22]	1	1.6
AAPM - American Association of Physicists in Medicine	Journal	[23]	1	1.6
Korean Journal of Radiology	Journal	[24]	1	1.6
Radiology	Journal	[25][29][61]	3	4.7
Nature	Scientific Report	[26]	1	1.6
IBM Journal of Research and Development	Journal	[27]	1	1.6
Springer Nature	Letter	[30]	1	1.6
Japanese Journal of Radiology	Journal	[31]	1	1.6
Biomedical Optics Express	Journal	[32]	1	1.6
2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)	Conference	[33]	1	1.6
Annals of Oncology	Journal	[34]	1	1.6
PLoS One	Journal	[35][77]	2	3.1
Theranostics	Journal	[36]	1	1.6
JAMA Network Open	Journal	[37]	1	1.6
Nature Biomedical Engineering	Journal	[38]	1	1.6
Cancer Communications	Journal	[39]	1	1.6
Proceeding of the National Academy of Sciences of the United State of America (PNAS)	Journal	[40]	1	1.6
Nature Biomedical Engineering	Journal	[41]	1	1.6

TABLE 4. *(Continued)***Publication source.**

Publication Source	Channel	Reference	No.	%
Translational Vision Science and Technology	Journal	[42]	1	1.6
International Ophthalmology	Journal	[43]	1	1.6
Clinical Endoscopy	Journal	[44]	1	1.6
Healthcare Informatics Research	Journal	[45]	1	1.6
Ophthalmology	Journal	[46][47][72]	3	4.7
IEEE Journal of Biomedical and Health Informatics	Journal	[48]	1	1.6
European Journal of Radiology	Journal	[49]	1	1.6
JAMA Dermatology	Journal	[50]	1	1.6
The Lancet	Journal	[51][70]	2	3.1
Quantitative Imaging in Medicine and Surgery	Journal	[52]	1	1.6
Oncologist	Journal	[53]	1	1.6
IEEE Access	Journal	[54]	1	1.6
Journal of Medical Imaging and Radiation Oncology	Journal	[55]	1	1.6
IEEE Transactions on Medical Imaging	Journal	[58]	1	1.6
American Roentgen Ray Society	Journal	[59]	1	1.6
JAMA - Journal of the American Medical Association	Journal	[62]	1	1.6
British Journal of Dermatology	Journal	[63]	1	1.6
Journal of Investigative Dermatology	Journal	[64]	1	1.6
Cell	Journal	[65]	1	1.6
Dentomaxillofacial Radiology	Journal	[67]	1	1.6
Head and Neck	Journal	[68]	1	1.6
Esophagus	Journal	[69]	1	1.6
ActaOrthopaedica	Journal	[71]	1	1.6
EBioMedicine	Journal	[73]	1	1.6
Skeletal Radiolog	Journal	[74]	1	1.6
Journal of Surgical Oncology	Journal	[75]	1	1.6
Endoscopy	Journal	[76]	1	1.6

Datasets are an integral part of the field of AI and have been cited in peer reviewed journals, conferences, letters and scientific reports. The medical image datasets for an AI application have adequate data volume, reusability and annotation. Each medical imaging datasets consists of metadata, data elements and identifiers. This combination is known as ''imaging examination''. Meta data element for medical imaging contains data made by an imaging modality; description of data depends on an order and annotations representing the content of a particular image. The Table 5 represents the datasets of variant diseases with their details. In the diabetes field [15], the Maastricht study of diabetes (T2DM) were used to diagnose diabetes from 8924 good quality images in the southern part of the Netherlands. From the clinical dataset, Optical coherence tomography (OCT) scan images were applied to diagnose the disease resulting in affected eyes by the macular fluid [47]. According to Bien *et al*., [19] the knee magnetic resonance imaging (MRI) reports were collected by Stanford Medical Center to detect the knee abnormalities, anterior cruciate ligament (ACL) tears, and meniscal tears. In identification of skin cancer, dermoscopic images from HAM 10000 dataset were collected from the International Skin Imaging Collaboration [20]. Adams *et al*., [55] used Anteroposterior hip radiograph dataset which contains 1160 images in the 8-bit PNG format. Burlina *et al*., [22] were applied to the 5664 color fundus images obtained from the NIH AREDS dataset to detect outer boundaries of the retina and resize them to 231 × 231 to conform the overFeat network. The National Eye Institute Age-Related Eye Disease Study (AREDS) dataset used by Burline *et al*., [56] which has total 67401 number of fundus photograph and it is identified as a gold-standard dataset. The open access series of breast ultrasonic dataset, which contains 882 images of unique breast masses, consists of 678 benign and 204 malignant lesions [23]. Cao *et al.*[58] were practiced prostrate mp-MRI dataset, which contains the data of 417 patients, preprocessed by intensity normalization and 3T scanners were used to take these images. Chee *et al.*[59] were used the hips dataset collected by Seoul National University Hospital (SNUH), which contains 673, 1346 MRI images of 16 years old or older patients and DICOM radiographic image archive were loaded by using python library 0.9.9v.

RQ2. What types of classification methods/algorithms are used in transfer learning research in disease diagnosis?

It has been observed that 55 studies were collected data by retrospectively and 8 studies used prospectively composed data as presented in (Table 6). Moreover, 17 studies have used datasets from open-access databases. When considering transfer learning and its approaches for medical diagnosis, there are two main processes. First process involves the classification including reduction of the potential outcomes (diagnosis) by comparing data to the particular outcomes. Second process involves the physiological data containing the medical data and images from different sources, which are used for diagnosing the disease. Moreover, transfer learning has been widely used for the purpose of dietary assessment [78]. The research has indicated that transfer learning is effectively implemented in different ways while considering the medical diagnosis. A brief review of

TABLE 5. Datasets description.

Ref	Year	Dataset	Description	Preprocessing	Instance	Format
[21]	2018	Retinal Photograph	Images were obtained with the help of eight academic institutions participating in the Imaging and Informatics in ROP.	Not mentioned	5511	PNG
[47]	2018	Clinical Dataset for macular fluid detection	OCT volume scans of eye affected by macular fluid were collected from the Vienna Reading Center database.	Not mentioned	1200	OCT scans image
[15]	2018	Maastricht Study type 2 diabetes (T2DM)	Study focusing on diabetes (type 2).	Not mentioned	10,000	Fundus Image obtained by AFC-230 camera by Nidek.
[55]	2019	Anteroposterior hip radiograph dataset	The images were collected from Royal Melbourne Hospital where patients had surgically confirmed NoF fractures.	Equivalent size and proportions.	1160	8-bit PNG images
[17]	2015	FDDSM database	Mammograms database haven been collected from "El farabi" center of radiology	Shock Filters	2,400	Images have been pixel size 85∗85 µm2 with 12 bits resolution
[18]	2017	2 Datasets: **1. Internal Cohort** Health Insurance Portability and Accountability Act complaint study for mammography 2012 **2. External Cohort** Breast Cancer Digital Repository (BCDR)	1. **Internal cohort** All patient undergoing mammography in 2012 at Institute of Diagnostic and Interventional Radiology, Switzerland. 2. **External Cohort** An external test cohort was obtained from publically available data set (BCDR).	Vidi Suite v2.0 used for image analysis	**Internal Cohort** (n = 3228) **External Cohort** (n = 35)	Mammography Imaging
[19]	2018	knee MRI examination	Knee MRI performed by Stanford University and manually analysis the reports in order to create a dataset.	Images were scaled to 256 × 256 pixels and converted into PNG file format.	1370	PNG
[20]	2019	HAM 10000	The dataset consists of dermoscopic images collected from the International Skin Imaging Collaboration (ISIC).	Not mentioned	2169	PNG
[22]	2017	NIH AREDS	The dataset contains colored fundus images.	Images were cropped by detecting the outer boundaries to the square shape and resize them into 231 * 231 pixels.	5664	PNG
[56]	2018	AREDS data set	The data set collected by National Eye Institute of4613individualsovera12-years study.	Data preprocessing contains blurring, sharpening horizontal flip ping and brightness adjustments.	67401	PNG
[23]	2019	Open Access Series of Breast Ultrasonic Data (OASBUD)	The dataset consists of images of unique breast masses.	Convert grayscale ultrasound images to (RGB)	882	Ultrasound images
[58]	2019	Prostate mp-MRI	This dataset obtained by 3T mp-MRI exams to "robotic-assisted laparoscopic prostatectomy"	Intensity Normalization and T2w & ADC registration	417	3T scanners
[59]	2019	Hips Datasets	This dataset consists of 16 years old or older patient with hip pain underwent MRI collected by "Seoul National University Hospital (SNUH)"	The images were resized to 224 * 224 pixels by using bilinear interpolation	673,1346	MR Images
[24]	2019	B-mode US Image datasets	The dataset of breast masses consists of 173 benign and 80 malignant patient images.	Not mentioned	253	ultrasonography
[25]	2018	dataset of liver CT images	Data set of patients with pathologic examination.	Not mentioned	7461	CT images
[26]	2017	"Multi centric Italian Lung Detection" (MILD)	The CT scans images were used from the MILD trials.	The low dose CT scans images were processed through 16-dectector row CT system	1,805	CT scans
[27]	2017	Dermoscopic images dataset	In 2016, the dataset released by "International Skin Imaging Collaboration" for "International Symposium on Biomedical Imaging" challenge.	Lesion Segmentation	900	PNG (Dermoscopic images)

TABLE 5. *(Continued)* **Datasets description.**

Ref	Year	Dataset	Description	Preprocessing	Instance	Format
[28]	2018	TCGA dataset	This images dataset collected from Cancer Genome Atlas.	The image which contains low amount of information and non-overlapped were removed.	1,634	Histopathology images
[61]	2019	F-FDG PET brain images	The dataset consists of brain images from the "Alzheimer's Disease NeuroImaging"	Grid method was used to preprocess the images.	2189	PET images
[29]	2019	Frontal chest radiographs	This dataset was based on radiographs images along with text reports.	The histogram equalization was applied for contrast enhancement and down sampling to a 224 x 224 input resolution	2,16,431	CXR Images
[62]	2017	Breast cancer dataset	Dataset was collected by two hospitals "Radboud University Medical Center" (RUMC) and "University of Medical Center Utrecht" (UMCU) in the Nether lands.	Not mentioned	399	RUMC images
[30]	2017	ISIC Dermoscopic Archive	This dataset was strictly composed for melanocytic lesions that are biopsy proven and annotated as malignant or benign.	Blurry images were removed from the test.	490	Clinical Images
[31]	2019	Breast Imaging Reporting and Data System (BI-RADS)	The collection of datasets based on ultrasound examination of breast masses and those masses were diagnosed as benign or malignant by pathology.	Ultrasound images were converted into jpeg and reduced the skin to chest wall.	480	Ultrasound DICOM images
[63]	2019	Digital clinical images of skin tumors and pigmented skin lesions	The dataset was obtained by the "University of Tsukuba Hospital" which consists of clinical images of skin tumors patients.	Pre-processing was not required	6009	Clinical Images
[32]	2019	1: RIM-ONE 2: DRISHTI-GS 3:ESPERANZA (private dataset)	The dataset **RIM-ONE** contains ophthalmic image and was created by the collaboration of three Spanish hospitals: "Hospital Universitario de Canarias", "Hospital Clínico San Carlos" and "Hospital Universitario Miguel Servet". The second dataset **DRISHTI-GS** consists of retinal fundus images and was collected at "Aravind Eye Hospital" in Madurai. **ESPERANZA** dataset based on fundus images with the view of 45 degrees and patient`s age ranging from 55 years to 86 years.	Preprocessing of contrast enhancement was not done.	2313	Fundus Images
[33]	2018	Not Reported	The dataset composed of brain CT scan images and collected by two local hospitals.	The images were resampled to isotropic resolution with having 250 x 250 mm fields-of-view in-plane.	329	CT scans images
[64]	2018	1: Asan dataset 2: MED-NODE dataset 3: Atlas Site images	The Asan dataset consists of clinical images of skin disease and was collected by "Department of Dermatology at Asan medical center". The second dataset MED-NODE was collected by "Department of Dermatology at University Medical Center Groningen". The third dataset known as atlas, was obtained from several dermatologic atlas sites	Resized image resolution for training and testing.	19,398	Clinical images.
[35]	2018	Asan dataset	The dataset was composed on patient demographics and clinical images acquired from 2003 to 2016.	Resized image resolution for training and testing.	598,854	Clinical Images.
[66]	2019	"Seoul National University Bundang Hospital" (SNUBH) dataset	The dataset was created by SNUBH where patients older than 16 years underwent Occipitomental view conventional radiography examinations.	DICOM files of the Waters' view radiographs were loaded and were cropped from the axis of the bilateral maxillary sinuses.	60,389	DICOM Water`s view radiograph

TABLE 5. *(Continued)* **Datasets description.**

Ref	Year	Dataset	Description	Preprocessing	Instance	Format
[69]	2019	Endocytoscopic system (ECS) images dataset	The dataset was composed of esophageal ECS examinations performed by "Saitama Medical Center".	Not mentioned	6235	ECS images
[38]	2019	Intracranial haemorrhage (ICH) dataset	The ICH dataset consists of unenhanced CT scans of brain.	Resized image resolution for training and testing.	904	CT scans
[39]	2018	Nasopharyngeal endoscopic images (NEI) dataset	The dataset contains images of clinicopathologic data and NEI of persons who underwent for routine screening at "Sun Yatsen University Cancer Center", China.	Excluded blurred images	33,507	Endoscopic images
[70]	2019	Thyroid imaging database	The dataset composed of ultrasound images collected from the thyroid imaging database at "Tianjin Cancer Hospital, Tianjin", China. It contains the information of adult patients aged 18 or older. The clinical diagnosis was done by radiologist of the above-mentioned hospital. The study was approved by the institutional board of the hospital.	The preprocessing was done by applying rotation, cropping, and adjustment of the saturation.	312,299	Ultrasound Images
[42]	2018	Not Mentioned	The OCT scans images were collected from the "Wuhan University Eye Center". All images were encrypted to protect the patients' health information. The dataset contains all kind of patients such as females, males, children's and adults. Moreover, the dataset also contains the information of same patients who underwent for their follow-up at different duration of time.	Excluded poor quality images	60,407	OCT Images

selected articles from transfer learning domain is presented in Table 6. The articles including different types of algorithms have been presented below in Fig. 5. The Fig. 5identifies the maximum number of articles using specific methods, architecture and the data source.

RQ3. What are the parameters/metrics on which the accuracy of classification methods/algorithms can be assessed in transfer learning for disease diagnosis in healthcare sector?

It has been noticed that 21articles used in this research study have provided sufficient data for the calculation of evaluation metrics. In transfer learning algorithms [18, 20, 23-25, 35, 39, 44, 59, 61, 64], the sensitivity ranged from 71.0 % to 100.0% (mean 85.25%) and specificity ranged from 64.0 % to 100·0% (mean 81.92%). Moreover, healthcare professional's [21, 26, 36, 37-38, 41, 51, 53, 66] sensitivity ranged from 33.0 % to 100.0% (mean 85.27%) and specificity ranged from 82.0% to 100·0% (mean 91.63%) are shown in Fig.6.

C. PERFORMANCE COMPARISON BETWEEN TRANSFER LEARNING AND HEALTH EXPERTS

All studies compared the diagnostic evaluation between transfer learning and diagnosticians were presented in Table 7. Performance parameters used for comparison included disease diagnostic accuracy, confusion matrix, sensitivity, specificity (see Fig.6) and the area under the receiver operating characteristic curve (AUROC) (see Fig.7). A total of four articles [16], [60], [34] and [46]examined the scenario where health experts were given additional clinical information alongside the image. Ardila *et al*., [16] tested single image versus the addition of historical images for both human diagnosticians and the transfer learning algorithms. Three studies also considered diagnostic performance in an algorithm-plus-clinician scenario [40], [41], [49]. Long *et al.*[41]., achieved a high accuracy compared with a panel of specialty doctors' predefined diagnostic decision and transcended the average levels of clinicians in most clinical situations except for treatment suggestion. Esteva *et al.*[30] also found that AI algorithms achieved comparable accuracy with or outperformed their human rivals. De Fauw *et al.*[60] reported results showed that AI's performance commensurate with retina specialists.

RQ4. What evidence/validation approach is there for disease diagnosis in health care?

Reference standards were wide ranging in line with variation of the target condition and the modality of imaging being used, with some studies adopting multiple methods (Table 8). 31 studies used histopathology; 22 studies used varying models of expert consensus; 2 studies used clinical follow-up; 1study used surgical confirmation; 3 studies used reading centre labels; 2 studies used imaging reports associated with open data sources and 2 studies used laboratory testing and 22 studies not reported any validation methods. Furthermore, 24 studies used random split sample validation techniques and 17 studies used resampling method.

TABLE 6. Algorithms, data source and image modality for the 63 included studies.

Ref	Image Modality	Algorithms Architecture	Name	Mean age (years)	Male (M)& Female (F)participants (%)	Data Source Number of images per training	Open Archive
[15]	Fundus Photography	ResNet	CNN	Not mentioned	M = 49, F=51	7931	No
[16]	CT	AlexNet	CNN	63 (Median)	M=53, F=47	512	No
[17]	Mammograms	ANN	ANN	Not mentioned	F=100	200	No
[18]	Mammograms	Vidi Suite v20	ANN	57	F=100	Study1: 95 Study2: 513	No
[19]	MRI	MRnet	CNN	38	M=49, F=41	1130	Yes
[20]	Dermatoscopy	ResNet-50	CNN	Not mentioned	Not mentioned	12378	Yes
[21]	Fundus Photography	CNN	CNN	Not mentioned	Not mentioned	4409	No
[22]	Fundus Photography	AlexNet	CNN	Not mentioned	Not mentioned	5664	Yes
[23]	Sonography	VGG-19	CNN	Not mentioned	Not mentioned	582	No
[24]	Ultrasonography	GoogleNet	CNN	47 (Median)	Not mentioned	790	No
[25]	CT	CNN	CNN	Training: 44 Testing: 48	F=28 (Training) F=43 (Testing)	7461	No
[26]	CT	ConvNet	CNN	Not mentioned	Not mentioned	490320	Yes
[27]	Microanatomy	Ensemble Method	Not mentioned	Not mentioned	Not mentioned	900	Yes
[28]	Microanatomy	Inception V3	CNN	Not mentioned	Not mentioned	825	Yes
[29]	Chest X-Rays	Resnet-18 DenseNet 121 AlexNet	CNN, Residual Network	Not mentioned	Not mentioned	180,000	No
[30]	Photographs	Inception-v3	CNN	Not mentioned	Not mentioned	129450	Yes
[31]	Breast Sonography	Inception-v2	CNN	Training: 55 Testing: 57	Not mentioned	947	Yes
[32]	Fundus Image	VGG-19	CNN	Not mentioned	Not mentioned	1560	Yes
[33]	CT	DenseNet	CNN	Not mentioned	Not mentioned	185	No
[34]	Dermatoscopy	Inception-v4	CNN	Not mentioned	Not mentioned	Not Reported	No
[35]	Photographs	ResNet-152	CNN	Asan1: 47 Asan2: 41	F = 55 (Asan1) F= 57 (Asan 2)	49567	Yes
[36]	OCT Photographs	VGG-16, ResNet-50, Inception-v3	CNN, Residual Network	51	82	28720	No
[37]	Chest-X Rays	CNN	CNN	Not mentioned	Not mentioned	87695	No
[38]	CT	Inception v3, VGG-16, ensemble v2, ResNet-50	CNN	Not mentioned	Not mentioned	704	No
[39]	Endoscopic	CNN	CNN	Training: 44 Testing: 46	F=30 (Training) F= 32 (Testing)	5557	No
[40]	X-Rays	DCNN	U-Net	Not mentioned	Not mentioned	100855	No
[41]	Ocular Photographs	DCNN	DCNN	Not mentioned	Not mentioned	886	No
[42]	OCT	ResNet	CNN	Not mentioned	Not mentioned	19815	No
[43]	Fundus Image	CNN	CNN	77	M=74, F=26	253	No
[44]	Endoscopic	VGG	CNN	69 (Median)	M=79, F=21	804	No
[45]	X-Rays	VGG-16	CNN	Not mentioned	Not mentioned	400	No
[46]	Fundus Image	Inception v3	CNN	Not mentioned	Not mentioned	140000	No
[47]	OCT	Deep Learning	CNN	Not mentioned	Not mentioned	840	No
[48]	Ultrasound	CNN	CNN	Training: Not mentioned Testing: 57	F = 90 (Testing)	6228	No
[49]	Ultrasound	Vidi Suite System	CNN	34	M=66, F=34	53	No
[50]	Dermoscopy	Inception v3	CNN	Not mentioned	Not mentioned	13724	Yes

TABLE 6. *(Continued)* **Algorithms, data source and image modality for the 63 included studies.**

Ref	Image Modality	Algorithms Architecture	Name	Mean age (years)	Male (M)& Female (F)participants (%)	Data Source Number of images per training	Open Archive
[51]	CT	ResNet	CNN	Not mentioned	Not mentioned	929	No
[52]	CT	3D CNN	CNN	61	M=39, F=61	1075	No
[53]	CT	CNN	CNN	60	M=56, F=44	2285	Yes
[54]	CT	CNN	DesNet	Not mentioned	M=44, F=56	523	No
[55]	X-Rays	AlexNet	CNN	Not mentioned	Not mentioned	512	No
[56]	Fundus	ResNet	CNN	Not mentioned	Not mentioned	59313	No
[57]	Fundus	ResNet-50	CNN	Not mentioned	Not mentioned	5913	No
[58]	MRI scans	Focal Net	CNN	Not mentioned	Not mentioned	Not Reported	No
[59]	MRI scans	ResNet	CNN	48	Not mentioned	1346	No
[60]	OCT	3D U-Net	CNN	Not mentioned	F=54 (Training) F= 55 (Testing)	14884	No
[61]	PET images	Inception-v3	CNN	76 (Male), 75 (Female)	M=53, F=47	1921	Yes
[62]	Histology	Google Net, ResNet, VGG-Net	CNN, Inception	Not mentioned	F=100	270	Yes
[63]	Photographs	GoogleNet	CNN	Not mentioned	Not mentioned	4867	No
[64]	Photographs	Ensemble: ResNet-152 + VGG-19	CNN	Asan1: 41 Asan2: 46	F= 55(Asan1) F=57 (Asan2)	19398	Yes
[65]	OCT	Inception v3	CNN	60	M=59, F=41	108312	Yes
[66]	MRI	Residual Network	CNN	59	M=60, F=40	8000	No
[67]	CT	AlexNet	CNN	Control: 66 Sjogren: 67	F=97 (Control) F=4 (Sjorgren) M= Not mentioned	400	No
[68]	Sonography	CNN	VGG	Training: 48 Testing: 50	F=82 (Training) F=85 (Testing)	594	No
[69]	Endoscopic	GoogleNet	CNN	Not mentioned	Not mentioned	240	No
[70]	Sonography	ResNet-50, Darnet-19	CNN	44 (Median)	F=75 (Training) F=77 (Testing)	42952	No
[71]	X-rays images of wrist, hand and ankle.	VGG-16	CNN	Not mentioned	Not mentioned	179521	No
[72]	Fundus Photographs	PNet, CNN, D-net	CNN	Not mentioned	Not mentioned	58402	Yes
[73]	Endoscopic	GoogleNet, Caffe	CNN	Training: 55 Testing: 50	F=55 (Training) F=57 (Testing)	32209	No
[74]	X-Rays	VGG-16	CNN	85	M=16, F=84	2978	No
[75]	Ultrasonography	ResNet V2 -50 Yolo v2	CNN	56	M=19, F=81	5007	No
[76]	Endoscopy	VGG-16, ResNet-50	CNN	Not mentioned	Not mentioned	24549	No
[77]	Dermatoscopy	VGG	CNN	Not mentioned	Not mentioned	362	No

IV. DISCUSSION

This is the first systematic study of diagnosis of disease accuracy of pre-trained algorithms versus clinical or medical experts using medical imaging. The studies have been selected in a careful manner with their diagnostic performance reporting and validation of the pre-trained algorithm was done. To summarize the findings of this SLR, taxonomy has been proposed (see Fig. 6) which consists of different diseases representing medical images and transfer learning algorithms used for diagnosis of each disease. Moreover, diagnostic accuracy obtained from the algorithms has also been compared with the health expert's examination.

A. TAXONOMY OF DISEASES DIAGNOSIS

The designed taxonomy (see Fig. 8) consists of multiple diseases which were diagnosed by using medical imaging. The diagnostics accuracy was measured by an applying variant kind of transfer learning algorithms. Furthermore, few of the diagnostic measurements were also compared with the health experts. The output was categorized into malignant and benign, the presence or not presence, referable or not referable, yes or no and normal or abnormal, because the results monitored in this SLR was based on binary classification. Ardila *et al.*,[16] applied3D-CNN on the current and prior low dose chest CT scans to predict the risk of

TABLE 7. Performance comparison between transfer learning and health experts.

Study	Year	Clinical Domain	TP	FP	FN	TN	Sensitivity	Specificity	AUROC
[18]	2017	Breast Cancer	25	11	10	24	0.716	0.696	0.79
[20]	2019	Skin Cancer	18	28	2	52	0.894	0.644	0.769
[23]	2019	Breast Cancer	45	18	8	92	0.851	0.834	0.893
[24]	2019	Breast Cancer	71	47	9	126	0.888	0.728	Not mentioned
[25]	2018	Lung Cancer	13	4	4	277	0.765	0.986	Not mentioned
[35]	2018	Skin Cancer	191	180	48	881	0.801	0.83	0.90
[39]	2018	Nasopharyngeal Cancer	29	279	22	25	0.895	0.708	Not mentioned
[44]	2019	Oesophageal cancer	458	47	51	358	0.898	0.883	Not mentioned
[59]	2019	Femoral head osteonecrosis	97	0	28	23	0.776	1	Not mentioned
[61]	2019	Alzheimer's disease	7	6	0	27	1	0.82	Not mentioned
[64]	2018	Onychomycosis	79	242	9	970	0.902	0.8	0.91
[70]	2019	Thyroid cancer	59	11	11	73	0.843	0.869	Not mentioned
[21]	2018	Retinopathy	51	0	3	46	0.94	0.94	Not mentioned
[26]	2017	Lung Cancer	314	38	68	219	0.822	0.852	Not mentioned
[36]	2019	Ophthalmology Age related	250	24	0	476	1	0.952	Not mentioned
[37]	2018	Lung Cancer	66	3	4	67	0.943	0.957	0.979
[38]	2019	Intracranial Hemorrhage	73	6	6	11	0.924	0.949	0.961
[41]	2017	Ophthalmology	14	0	1	42	0.9333	1	Not mentioned
[51]	2018	Respiratory Disease	10	13	20	107	0.333	0.892	Not mentioned
[53]	2019	Lung Cancer	24	3	1	22	0.96	0.88	Not mentioned
[66]	2019	Spondylitics	131	21	29	99	0.819	0.825	0.88

lung cancer. Skin classification was performed by Brinker *et al.*[20], on the dermoscopic images to detect the melanoma by using CNN algorithms. The results were also compared with dermatologist where CNN shows small variance in skin image classification task. For the identification of age-related macular degeneration (AMD), Schlegl *et al.* [47] employed a transfer learning algorithms for the detection of AMD using the OCT images of NIH AREDS dataset. The AMD severity was categorized into three classes. These classes include medically relevant 4-class, 3-class, and 2-class. The diagnostic accuracy of transfer learning algorithm was also compared with the clinical experts. The accuracies achieved by the transfer learning algorithms for these classes were 79.4%, 81.5%, and 93.4%, respectively. Moreover, the diagnostic accuracies reported by the health experts were 75.8%, 85.0% and 95.2% respectively. The result reveals that the performance of algorithms is approximately equivalent to the clinical experts. Abbasi *et al.*, [15] proposed a deep residual network (DRN) to predict the diabetic retinopathy from retinal images. The performance of DRN is significantly higher than health experts. In their study, Adams *et al.*, [55] described a AlexNet and GoogleNet model for detection of neck of femur (NoF) fractures by using anteroposterior hips x-rays and compare the results with experts. Remarkably, the pre-trained models were able to identify the NoF with similar levels as radiologic technologist. Cao *et al.*, [58] used pre-trained FocalNet technique for the detection joint prostate cancer and calculate Gleason score. For the assessment of their method, they used mp-MRI dataset of 417 patients and achieved sensitivity of 89.7%. Furthermore, with the comparison to radiologic technologists, FocalNet showed equivalent detection. Biopsy analysis becomes significant tasks for diagnosing the stage of cancer. Byra *et al.*, [23] analyzed transfer learning techniques to build a classifier based on sonography images of breast masses. After performing fine tuning, the pre-trained algorithms achieved AUC of 0.936 which was greater than the radiologists reading. Chee *et al.*, ý[59] used transfer learning algorithm to diagnose osteonecrosis of femoral head (OFH) based on MR images. The diagnostic accuracy of the algorithm was compared with the less, and experienced diagnostic radiographers. The sensitivity achieved by the algorithm for diagnosing OFH was non inferior to that of both radiologic technologists.

B. MODEL FOR DISEASE DIAGNOSIS

An effective and efficient model for diagnosis of diseases has been proposed in Fig. 9. The model consists of five main components, which are based on collection of datasets, splitting the data into training and testing sets, employing data augmentation techniques, fine-tuning of transfer learning algorithms, and measures diagnostics accuracy, and compare the results with health experts. The dataset used for the detection of diseases are of two types: open access or publicly available benchmark medical imaging datasets and private or non-public datasets.

In open access, the medical imaging datasets of skin classification was HAM1000, DermQuest, Mednode, DermIS and PH2, T2DM for diagnosing diabetes, FDDSM for detection of breast cancer, TCGA used for lungs cancer, F-FDG PET for Alzheimer's disease, BI-RADS consists of ultrasound images for identification of breast cancer, Asan, MEDNODE and Atlas site images database used for classification of skin disease, RIM-ONE and DRISHTI-GS contains fundus images for retinopathy have been available. While in private access, the available options were ESPERANZA

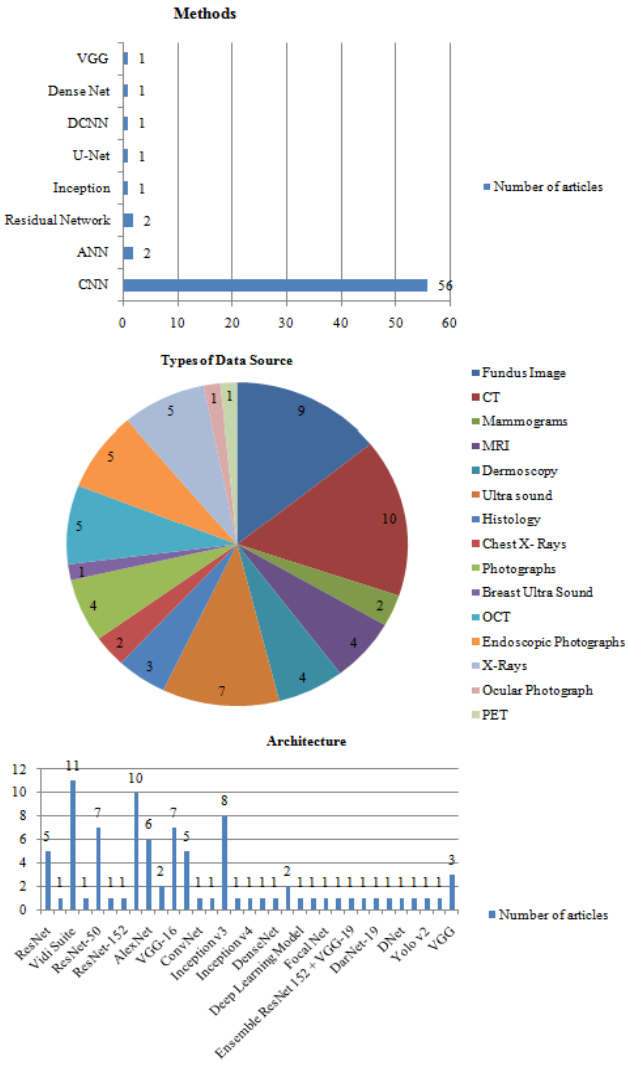

FIGURE 5. Synthesis of reviewed articles by type of transfer learning methods, data source and architecture.

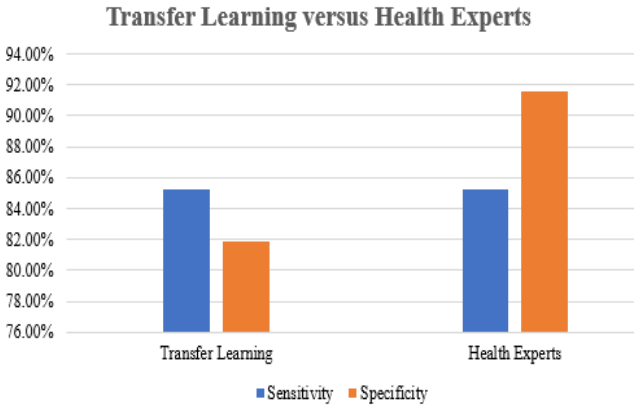

FIGURE 6. Comparison of Health Experts and transfer learning algorithms in terms of sensitivity and specificity.

dataset of fundus images for detection of glaucoma, Dermnet and IRMA skin used for identification of melanoma, and Anteroposterior hip radiograph database used for NoF. Data

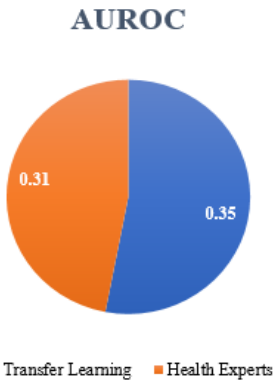

FIGURE 7. Average AUROC of transfer algorithms versus health experts.

augmentation techniques were applied to the datasets of medical imaging which were in low contrast. The different transformation operations were also used such as horizontal or vertical flip, shifts and zooms for achieving the best diagnostics accuracy. Furthermore, training and testing sets of images have also been defined in this stage. After that feature selection methods were used for selecting appropriate patterns. The variant types of transfer learning algorithms were employed to measure the diagnostics accuracy. All the results from different algorithms have been evaluated and compared with the health experts as shown in Fig.9.

C. PRINCIPLE FINDINGS

Our systematic study identified 63 articles on the field of transfer learning methods for diagnosis of disease. These research studies consisted from different medical fields, including lung cancer, breast tumor, diabetes, knee injuries and Age-related macular degeneration, skin cancer, retinopathy, liver fibrosis, brain hemorrhage, nasopharyngeal and thyroid cancer, hip fractures, Trauma and orthopaedics, ophthalmology, oesophageal cancer, odontogenic tumors of the jaw, prostate cancer, femoral head osteonecrosis, alzheimer's disease, lymph node metastases, onychomycosis, spondylitis, Sjogren's syndrome, oesophageal cancer, helicobacter pylori gastritis and Gastric cancer. Although numerous research studies have tried to discuss several medical topics, distinct pre-trained algorithms and training methods were employed across studies. Effectiveness and efficacy of validation methods contained varied results among different research studies. In selection criteria of this research study, articles about the comparisons of any disease diagnostic performance between pre-trained methods and clinical professionals were reviewed.

After careful selection of studies with transparent reporting of diagnostic performance and validation of the algorithm (see Table 7), it has been identified that the algorithms of transfer learning have almost equal specificity and sensitivity as compared to health experts. It can be said that the estimates of transfer learning algorithms have equivalent clinical accuracy (see Fig 6 & Fig. 7). However, there were several methodological deficiencies in most of the research studies

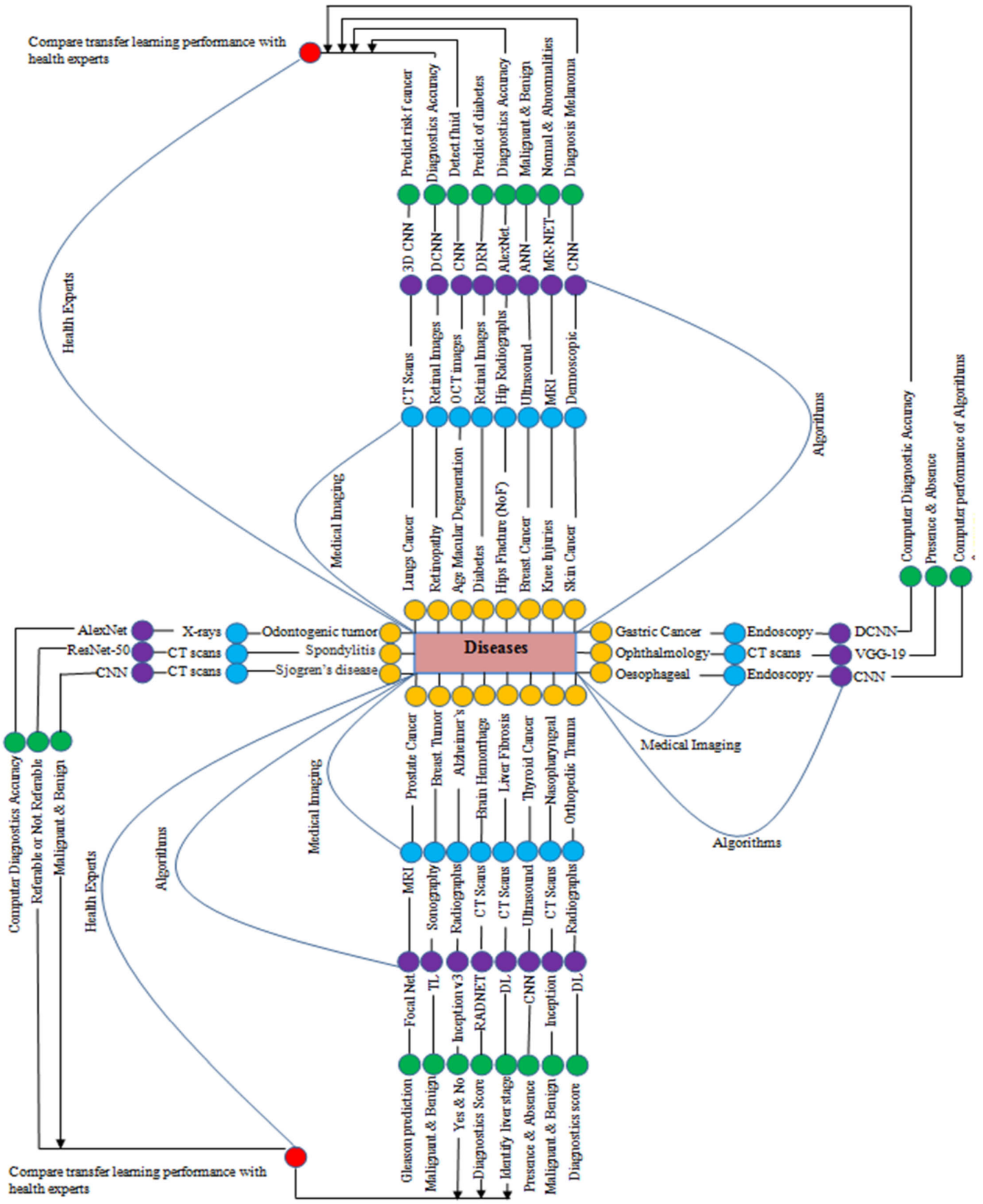

FIGURE 8. Taxonomy of multiple diseases with medical imaging and diagnostic accuracy of transfer learning algorithms versus health experts.

TABLE 8. Model validation for the 63 included studies.

Ref	Year	Target Disease	Reference Standards	Validation
[15]	2018	Diabetes	Laboratory testing	Cross Validations
[16]	2019	Lung cancer	Histology	Cross Validations
[17]	2015	Breast tumor	Histology	Cross Validations
[18]	2017	Breast tumor	Histology	Study 1: NA Study 2: temporal split-sample validation
[19]	2018	Knee injuries	Expert consent	Cross Validations
[20]	2019	Melanoma	Histology	Cross Validations
[21]	2018	Retinopathy	Expert consent	Resampling Process
[22]	2017	Age-related macular degeneration	Expert consent	Resampling Process
[23]	2019	Breast tumor	Histology	Resampling Process
[24]	2019	Breast tumor	Histology	Not mentioned
[25]	2018	Liver fibrosis	Histology	Resampling Process
[26]	2017	Lung cancer	Expert consent	Cross Validation
[27]	2017	Melanoma	Histology	Rotation Estimation
[28]	2018	Lung cancer	Histology	Not mentioned
[29]	2019	Lung conditions	Expert consent	Resampling Process
[30]	2017	Dermatological cancer	Histology	Resampling Process
[31]	2019	Breast tumor	Histology	Not mentioned
[32]	2019	Glaucoma	Expert consent	Resampling Process
[33]	2018	Brain haemorrhage	Expert consent	Not mentioned
[34]	2018	Melanoma	Histology	Not mentioned
[35]	2018	Skin disease	Histology	Rotation Estimation
[36]	2019	Age-related macular degeneration	Expert consent	Cross Validations
[37]	2019	Lung conditions	Expert consent	Rotation Estimation
[38]	2019	Intracranial haemorrhage	Expert consent	Cross Validations
[39]	2018	Nasopharyngeal malignancy	Histology	Rotation Estimation
[40]	2018	Trauma and orthopaedics	Expert consent	Not mentioned
[41]	2017	Ophthalmology	Expert consent	Resampling Process
[42]	2018	Macular pathology	Expert consent	Resampling Process
[43]	2019	Age-related macular degeneration	Expert consent	Not mentioned
[44]	2019	Oesophageal cancer	Histology	Not mentioned
[45]	2018	Odontogenic tumors of the jaw	Histology	Not mentioned
[46]	2019	Diabetic retinopathy	Expert consent	Not mentioned
[47]	2018	Macular diseases	Expert consent	Resampling Process
[48]	2019	Thyroid cancer	Histology	Resampling Process
[49]	2018	Breast tumor	Histology	Cross Validations
[50]	2019	Dermatological cancer	Histology	Not mentioned
[51]	2018	Lung fibrosis	Expert consent	Not mentioned
[52]	2018	Lung cancer	Histology	Rotation Estimation

TABLE 8. *(Continued)* Model validation for the 63 included studies.

Ref	Year	Target Disease	Reference Standards	Validation
[53]	2019	Lung cancer	Expert consent	Resampling Process
[54]	2019	Retinopathy	Expert consent	Cross Validations
[55]	2019	Hip fracture	Surgical confirmation	Cross Validations
[56]	2018	Age-related macular degeneration	Reading centre grader	Not mentioned
[57]	2018	Age-related macular degeneration	Reading centre grader	Not mentioned
[58]	2019	Prostate cancer	Histology	Resampling Process
[59]	2019	Femoral head osteonecrosis	Imaging reports	Not Reported
[60]	2018	Retinal disease	Follow-up	Rotation Estimation
[61]	2019	Alzheimer's disease	Follow-up	Not mentioned
[62]	2017	Lymph node metastases	Histology	Cross Validations
[63]	2019	Dermatological cancer	Histology	Resampling Process
[64]	2018	Onychomycosis	Histology	Cross Validations
[65]	2018	Retinal diseases	Expert consent	Cross Validations
[66]	2019	Spondylitis	Expert consent	Not mentioned
[67]	2019	Sjogren's syndrome	Expert consent	Not mentioned
[68]	2019	Thyroid cancer	Histology	Resampling Process
[69]	2019	Oesophageal cancer	Histology	Not mentioned
[70]	2019	Thyroid cancer	Histology	Not mentioned
[71]	2017	Fractures	Histology	Rotation Estimation
[72]	2019	Age-related macular degeneration	Reading centre grader	Not mentioned
[73]	2017	Helicobacter pylori gastritis	laboratory testing	Cross Validations
[74]	2019	Hip fractures	imaging reports	Cross Validations
[75]	2019	Malignant thyroid	Histology	Not mentioned
[76]	2019	Gastric cancer	Histology	Resampling Process
[77]	2018	Melanoma	Histology	Resampling Process

that need to be addressed. It is important to note that transfer learning diagnostic accuracy should be measured in isolation so that it may not affect the clinical practice. Several studies were deliberately ignored because they did not provide any comparisons with the health experts. It is also important to mention here that a few studies provided comparisons with health experts professional considering the same dataset.

Reviewed articles have revealed that any disease diagnostic performance of transfer learning algorithms have comparable with medical experts. The most effectively applied approach, CNNs yields substantial discriminative behavior on provision of the training datasets. Moreover, the methods of neural network mostly require large amount of data for training process, which feasible to apply on pre-trained models for rare diseases [65]. The combination of pre-trained models with other emerging technologies e.g. the distribution platforms of

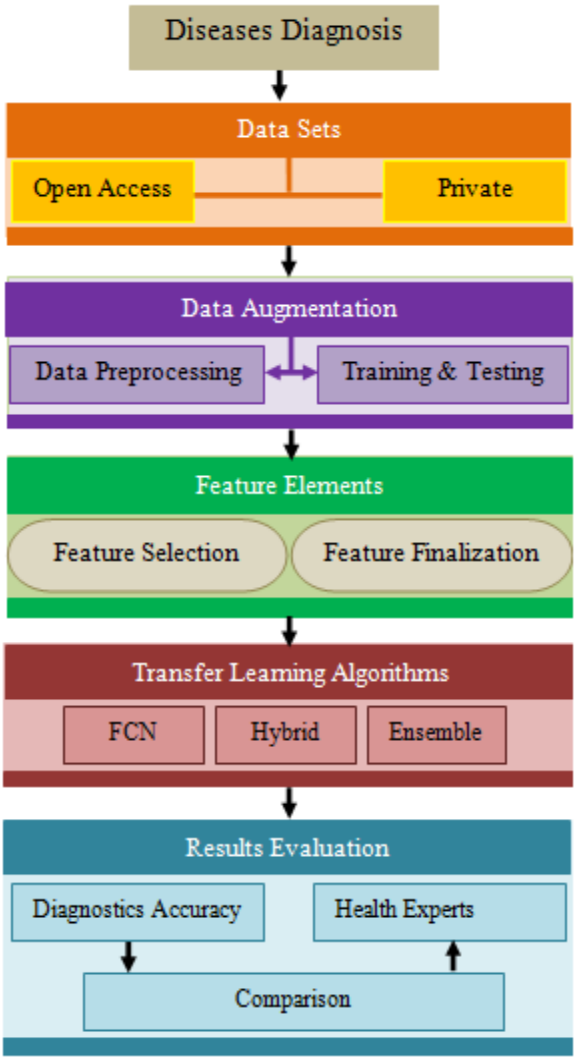

FIGURE 9. Model for diseases diagnosis.

cloud-based data would extend the performance use beyond clinical operations [41].

Most of the effectiveness of transfer learning approaches was observed in disease diagnosis from medical imaging across selected articles. Furthermore, computer-based methodologies expedite identification of clinical symptoms (e.g. benign and malignant) based on different features of medical imaging resulting in congruous outputs.

Pre-trained algorithms-based identification and classification of physical characteristics is covered during the training of machines, and this ability of machines is combined, and performance is assessed on appearance-based disease diagnostics such as skin diseases [20], [30]. Transfer learning-based medical imaging can reduce clinical tasks as well as the cognitive burden on medical professionals, which results in the increased efficiency in health care services.

Transfer learning performs in collaboration with clinical experts in order to analyze the medical images in an effective manner. Clinical image examination consists of disease identification problems whose output depends on the detection and interpretation of features such as patterns, colors and shapes. The quality of transfer learning networking enables continuous learning and training in order to achieve adequate accuracy [12]. As a result, significant achievement in disease diagnoses associated to the medical imaging evaluation. The distinguishing development of transfer learning, which human beings may not be able to acquire have contributed in improving the performance of clinical experts, as studied in the 63 articles reviewed in this paper.

The reviewed articles show that the achievement of pre-trained models was based on any disease diagnostics outcomes. However, the considerations of disease diagnostic outcomes require the achievement of meaningful suggestions. Recursive processes were used as diagnostics criteria for real-world situations, which are appraised by medical experts. Although the pre-trained learning abilities may lead towards other prospects and feasibility of diagnostic processes was inexorably determined by health professional in terms of clinical experience. Therefore, the concluding triumph of pre-trained learning models were controlled by healthcare professionals who are actual evaluators of diagnosis. The significance of the relationship between pre-trained algorithms and medical experts cannot be isolated.

Rapid growth of artificial intelligence technology provides promising outlook on diagnostic applications. However, the assessment initiated by medical specialists delivers an elementary role in continued bloom of artificial intelligence. In disease diagnosis application, transfer learning approaches cannot exist without interaction of human beings because concluding disease diagnosis must have real-world implications. The tireless learning capabilities of transfer learning algorithms were accompaniment cognitive fatigue in humans [12] and substantially improved the efficiency of medical experts. The extraordinary performance of transfer learning mechanisms is comparable with a health expert saves a lot of time in medical practice, which reduces the tension transition from clinician to the expert.

Despite being an auspicious moment for transfer learning approaches, there are certain problems that need to be addressed in impending stages. However, it is still not clear that whether transfer learning approaches can transform the assessment of medical experts in clinical setting. In addition, a hybrid system backed by both pre-trained algorithms and medical experts would present more efficient diagnostic practice.

These results can bring enhanced health process by diagnosis diseases from medical imaging. Data interpretation in this regard still remains an important task to the field of AI. New diagnostics reporting methods that address particular challenges of transfer learning mechanisms may improve future studies, ultimately enabling better assurance in results of future performance of this promising technology

V. CONCLUSION

The current developments in transfer learning have acquired comparable performance with health experts in different

fields. The streamlined efficiency and predictive performance regarding diagnosis of disease, especially in the tasks of medical imaging have exceeded the health experts with the capabilities and abilities. The continued development in technologies of pre-trained models have under pinned the clinical implications, which focus on principles of patient-centered health care. Moreover, these technologies should be considered for the purpose of artificial intelligence related research in technology-based medical research in the future. According to inclusion and exclusion criteria, our searched articles were limited to selected databases presented in Table 2. Some of journals and databases could not be explored because the scope of the article would have been enlarged. We could not target articles from other languages because of language barriers. Therefore, we have only targeted articles published in English language.

In future we expect more annotated data sources that will help developing specialized transfer learning algorithms which will result in even better accuracies for diseases diagnosis using medical imaging.

REFERENCES

[1] K. H. Fletcher, "Matter with a mind; a neurological research robot," *Research*, vol. 4, no. 7, pp. 305–307, 1951.
[2] Y. Shoham, R. Perrault, and E. Brynjolfsson, "The AI index 2018 annual report," AI Index Steering Committee, Human-Centered AI Initiative, Stanford Univ., Standford, CA, USA, Tech. Rep., 2018. [Online]. Available: http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf
[3] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, Nov. 2011.
[4] R. Hadsell, A. Erkan, P. Sermanet, M. Scoffier, U. Muller, and Y. LeCun, "Deep belief net learning in a long-range vision system for autonomous off-road driving," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nice, France, Sep. 2008, pp. 33–628.
[5] L. Faes, S. K. Wagner, D. J. Fu, X. Liu, E. Korot, J. R. Ledsam, T. Back, R. Chopra, N. Pontikos, C. Kern, G. Moraes, M. K. Schmid, D. Sim, K. Balaskas, L. M. Bachmann, A. K. Denniston, and P. A. Keane, "Automated deep learning design for medical image classification by health-care professionals with no coding experience: A feasibility study," *Lancet Digit. Health*, vol. 1, no. 5, pp. e232–e242, Sep. 2019.
[6] L. Zhang, H. Wang, Q. Li, M.-H. Zhao, and Q.-M. Zhan, "Big data and medical research in China," *Brit. Med. J.*, vol. 360, p. j5910, Feb. 2018.
[7] B. F. King, Jr., "Artificial intelligence and radiology: What will the future hold," *J. Amer. CollRadiol.*, vol. 15, pp. 501–503, Mar. 2018.
[8] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, "Transfer learning using computational intelligence: A survey," *Knowl.-Based Syst.*, vol. 80, pp. 14–23, May 2015.
[9] A. L. Fogel and J. C. Kvedar, "Artificial intelligence powers digital medicine," *NPJ Digit. Med.*, vol. 1, no. 1, p. 5, Dec. 2018.
[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.
[11] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Oct. 2012.
[12] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Med.*, vol. 25, no. 1, pp. 65–69, Jan. 2019.
[13] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *Proc. 12th Int. Conf. Eval. Assessment Softw. Eng. (EASE)*, Bari, Italy, Jun. 2008, pp. 71–80.
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[15] S. Abbasi-Sureshjani, B. Dashtbozorg, B. M. ter Haar Romeny, and F. Fleuret, "Exploratory study on direct prediction of diabetes using deep residual networks," in *VipIMAGE*. Basel, Switzerland: Springer, 2018, pp. 797–802.
[16] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, and S. Shetty, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Med.*, vol. 25, pp. 954–961, May 2019.
[17] N. G. B. Ayed, A. D. Masmoudi, D. Sellami, and R. Abid, "New developments in the diagnostic procedures to reduce prospective biopsies breast," in *Proc. Int. Conf. Adv. Biomed. Eng. (ICABME)*, Beirut, Lebanon, Sep. 2015, pp. 205–208.
[18] B. AS, M. Marcon, S. Ghafoor, M. C. Wurnig, T. Frauenfelder, and A. Boss, "Deep learning in mammography: Diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer," *Invest Radiol.*, vol. 52, no. 7, pp. 434–440, 2017.
[19] N. Bien *et al.*, "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet," *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002699.
[20] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, S. Fröhling, J. S. Utikal, and C. von Kalle, "A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task," *Eur. J. Cancer*, vol. 111, pp. 148–154, Apr. 2019.
[21] J. M. Brown, J. P. Campbell, A. Beers, K. Chang, S. Ostmo, R. V. P. Chan, J. Dy, D. Erdogmus, S. Ioannidis, J. Kalpathy-Cramer, and M. F. Chiang, "Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks," *JAMA Ophthalmol.*, vol. 136, no. 7, pp. 803–810, 2018.
[22] P. Burlina, K. D. Pacheco, N. Joshi, D. E. Freund, and N. M. Bressler, "Comparing humans and deep learning performance for grading AMD: A study in using universal deep features and transfer learning for automated AMD analysis," *Comput. Biol. Med.*, vol. 82, pp. 80–86, Mar. 2017.
[23] M. Byra, M. Galperin, H. Ojeda-Fournier, L. K. Olson, M. K. O'Boyle, C. E. Comstock, and M. P. Andre, "Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion," *Med. Phys.*, vol. 46, no. 2, pp. 746–755, 2019.
[24] J. S. Choi, B.-K. Han, E. S. Ko, J. M. Bae, E. Y. Ko, S. H. Song, M.-R. Kwon, J. H. Shin, and S. Y. Hahn, "Effect of a deep learning framework-based computer-aided diagnosis system on the diagnostic performance of radiologists in differentiating between malignant and benign masses on breast ultrasonography," *Korean J. Radiol.*, vol. 20, no. 5, p. 749, 2019.
[25] K. J. Choi, J. K. Jang, S. S. Lee, Y. S. Sung, W. H. Shim, H. S. Kim, J. Yun, J.-Y. Choi, Y. Lee, B.-K. Kang, J. H. Kim, S. Y. Kim, and E. S. Yu, "Development and validation of a deep learning system for staging liver fibrosis by using contrast agent–enhanced CT images in the liver," *Radiology*, vol. 289, no. 3, pp. 688–697, 2018.
[26] F. Ciompi, K. Chung, S. J. van Riel, A. A. A. Setio, P. K. Gerke, C. Jacobs, E. T. Scholten, C. Schaefer-Prokop, M. M. W. Wille, A. Marchianò, U. Pastorino, M. Prokop, and B. van Ginneken, "Towards automatic pulmonary nodule management in lung cancer screening with deep learning," *Sci. Rep.*, vol. 7, no. 1, p. 46479, Jun. 2017.
[27] N. C. F. Codella, Q.-B. Nguyen, S. Pankanti, D. A. Gutman, B. Helba, A. C. Halpern, and J. R. Smith, "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM J. Res. Develop.*, vol. 61, nos. 4–5, pp. 5:1–5:15, Jul. 2017.
[28] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature Med.*, vol. 24, pp. 1559–1567, Sep. 2018.
[29] J. A. Dunnmon, D. Yi, C. P. Langlotz, C. Ré, D. L. Rubin, and M. P. Lungren, "Assessment of convolutional neural networks for automated classification of chest radiographs," *Radiology*, vol. 290, no. 2, pp. 537–544, 2019.
[30] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, Jan. 2017.

[31] T. Fujioka, K. Kubota, M. Mori, Y. Kikuchi, L. Katsuta, M. Kasahara, G. Oda, T. Ishiba, T. Nakagawa, and U. Tateishi, "Distinction between benign and malignant breast masses at breast ultrasound using deep learning method with convolutional neural network," *Jpn. J. Radiol.*, vol. 37, pp. 466–472, Mar. 2019.

[32] J. J. Gómez-Valverde, A. Antón, G. Fatti, B. Liefers, A. Herranz, A. Santos, C. I. Sánchez, and M. J. Ledesma-Carbayo, "Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning," *Biomed. Opt. Express*, vol. 10, no. 2, pp. 892–913, 2019.

[33] M. Grewal, M. M. Srivastava, P. Kumar, and S. Varadarajan, "RADnet: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Washington, DC, USA, Apr. 2018, pp. 281–284.

[34] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, and L. Uhlmann, "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, 2018.

[35] S. S. Han, G. H. Park, W. Lim, M. S. Kim, J. I. Na, I. Park, and S. E. Chang, "Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network," *PLoS ONE*, vol. 13, no. 1, Jan. 2018, Art. no. e0191493.

[36] D.-K. Hwang, C.-C. Hsu, K.-J. Chang, D. Chao, C.-H. Sun, Y.-C. Jheng, A. A. Yarmishyn, J.-C. Wu, C.-Y. Tsai, M.-L. Wang, C.-H. Peng, K.-H. Chien, C.-L. Kao, T.-C. Lin, L.-C. Woung, S.-J. Chen, and S.-H. Chiou, "Artificial intelligence-based decision-making for age-related macular degeneration," *Theranostics*, vol. 9, no. 1, pp. 232–245, 2019.

[37] E. J. Hwang, S. Park, K.-N. Jin, J. I. Kim, S. Y. Choi, J. H. Lee, J. M. Goo, J. Aum, J.-J. Yim, J. G. Cohen, G. R. Ferretti, and C. M. Park, "Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs," *JAMA Netw. Open*, vol. 2, no. 3, 2019, Art. no. e191095.

[38] H. Lee, S. Yune, M. Mansouri, M. Kim, S. H. Tajmir, C. E. Guerrier, S. A. Ebert, S. R. Pomerantz, J. M. Romero, S. Kamalian, R. G. Gonzalez, M. H. Lev, and S. Do, "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets," *Nature Biomed. Eng.*, vol. 3, pp. 173–182, Dec. 2019.

[39] C. Li *et al.*, "Development and validation of an endoscopic images-based deep learning model for detection with nasopharyngeal malignancies," *Cancer Commun.*, vol. 38, no. 1, p. 59, Dec. 2018.

[40] R. Lindsey, A. Daluiski, S. Chopra, A. Lachapelle, M. Mozer, S. Sicular, D. Hanel, M. Gardner, A. Gupta, R. Hotchkiss, and H. Potter, "Deep neural network improves fracture detection by clinicians," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 45, pp. 11591–11596, 2018.

[41] E. Long, H. Lin, Z. Liu, X. Wu, L. Wang, J. Jiang, Y. An, Z. Lin, X. Li, J. Chen, J. Li, Q. Cao, D. Wang, X. Liu, W. Chen, and Y. Liu, "An artificial intelligence platform for the multihospital collaborative management of congenital cataracts," *Nature Biomed. Eng.*, vol. 1, no. 2, p. 0024, Feb. 2017.

[42] W. Lu, Y. Tong, Y. Yu, Y. Xing, C. Chen, and Y. Shen, "Deep learning-based automated classification of multi-categorical abnormalities from optical coherence tomography images," *Transl. Vis. Sci. Technol.*, vol. 7, no. 6, p. 41, Dec. 2018.

[43] S. Matsuba, H. Tabuchi, H. Ohsugi, H. Enno, N. Ishitobi, H. Masumoto, and Y. Kiuchi, "Accuracy of ultra-wide-field fundus ophthalmoscopy-assisted deep learning, a machine-learning technology, for detecting age-related macular degeneration," *Int. Ophthalmol.*, vol. 39, no. 6, pp. 1269–1275, 2019.

[44] K. Nakagawa, R. Ishihara, K. Aoyama, M. Ohmori, H. Nakahira, N. Matsuura, S. Shichijo, T. Nishida, T. Yamada, S. Yamaguchi, H. Ogiyama, S. Egawa, O. Kishida, and T. Tada, "Classification for invasion depth of esophageal squamous cell carcinoma using a deep neural network compared with experienced endoscopists," *Gastrointestinal Endoscopy*, vol. 90, no. 3, pp. 407–414, 2019.

[45] W. Poedjiastoeti and S. Suebnukarn, "Application of convolutional neural network in the diagnosis of jaw tumors," *Healthcare Informat. Res.*, vol. 24, no. 3, p. 236, 2018.

[46] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, S. Xu, S. Barb, A. Joseph, M. Shumski, J. Smith, A. B. Sood, G. S. Corrado, L. Peng, and D. R. Webster, "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy," *Ophthalmology*, vol. 126, no. 4, pp. 552–564, 2019.

[47] T. Schlegl, S. M. Waldstein, H. Bogunovic, F. Endstraßer, A. Sadeghipour, A.-M. Philip, D. Podkowinski, B. S. Gerendas, G. Langs, and U. Schmidt-Erfurth, "Fully automated detection and quantification of macular fluid in OCT using deep learning," *Ophthalmology*, vol. 125, no. 4, pp. 549–558, 2018.

[48] W. Song, S. Li, J. Liu, H. Qin, B. Zhang, S. Zhang, and A. Hao, "Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 3, pp. 1215–1224, May 2019.

[49] E. Stoffel, A. S. Becker, M. C. Wurnig, M. Marcon, S. Ghafoor, N. Berger, and A. Boss, "Distinction between phyllodes tumor and fibroadenoma in breast ultrasound using deep learning image analysis," *Eur. J. Radiol. Open*, vol. 5, pp. 165–170, Sep. 2018.

[50] P. Tschandl *et al.*, "Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks," *JAMA Dermatol.*, vol. 155, no. 1, pp. 58–65, 2019.

[51] S. L. F. Walsh, L. Calandriello, M. Silva, and N. Sverzellati, "Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: A case-cohort study," *Lancet Respiratory Med.* vol. 6, no. 11, pp. 837–845, 2018.

[52] S. Wang, R. Wang, S. Zhang, R. Li, Y. Fu, X. Sun, Y. Li, X. Sun, X. Jiang, X. Guo, X. Zhou, J. Chang, and W. Peng, "3D convolutional neural network for differentiating pre-invasive lesions from invasive adenocarcinomas appearing as ground-glass nodules with diameters $\leq$3 cm using HRCT," *Quant. Imag. Med. Surg.*, vol. 8, pp. 491–499, 2018.

[53] C. Zhang *et al.*, "Toward an expert level of lung cancer detection and classification using a deep convolutional neural network," *Oncologist*, vol. 24, no. 9, pp. 1159–1165, Sep. 2019, doi: 10.1634/theoncologist.2018-0908.

[54] Y. Zhang, L. Wang, Z. Wu, J. Zeng, Y. Chen, R. Tian, and J. Zhao, "Development of an automated screening system for retinopathy of prematurity using a deep neural network for wide-angle retinal images," *IEEE Access*, vol. 7, pp. 10232–10241, 2019.

[55] M. Adams, W. Chen, D. Holcdorf, M. W. McCusker, P. D. Howe, and F. Gaillard, "Computer vs human: Deep learning versus perceptual training for the detection of neck of femur fractures," *J. Med. Imag. Radiat. Oncol.*, vol. 63, no. 1, pp. 27–32, Feb. 2019.

[56] P. Burlina, N. Joshi, K. D. Pacheco, D. E. Freund, J. Kong, and N. M. Bressler, "Utility of deep learning methods for referability classification of age-related macular degeneration," *JAMA Ophthalmol.*, vol. 136, no. 11, pp. 1305–1307, 2018.

[57] P. M. Burlina, N. Joshi, K. D. Pacheco, D. E. Freund, J. Kong, and N. M. Bressler, "Use of deep learning for detailed severity characterization and estimation of 5-year risk among patients with age-related macular degeneration," *JAMA Ophthalmol.*, vol. 136, no. 12, pp. 1359–1366, 2018.

[58] R. Cao, A. M. Bajgiran, S. A. Mirak, S. Shakeri, X. Zhong, D. Enzmann, S. Raman, and K. Sung, "Joint prostate cancer detection and gleason score prediction in mp-MRI via FocalNet," *IEEE Trans. Med. Imag.*, vol. 38, no. 11, pp. 2496–2506, Nov. 2019, doi: 10.1109/TMI.2019.2901928.

[59] C. G. Chee, Y. Kim, Y. Kang, K. J. Lee, H.-D. Chae, J. Cho, C.-M. Nam, D. Choi, E. Lee, J. W. Lee, S. H. Hong, J. M. Ahn, and H. S. Kang, "Performance of a deep learning algorithm in detecting osteonecrosis of the femoral head on digital radiography: A comparison with assessments by radiologists," *Amer. J. Roentgenology*, vol. 213, no. 1, pp. 155–162, Jul. 2019, doi: 10.2214/AJR.18.20817.

[60] F. J. De *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Med.*, vol. 24, pp. 1342–1350, Aug. 2018.

[61] Y. Ding, J. H. Sohn, M. G. Kawczynski, H. Trivedi, R. Harnish, N. W. Jenkins, D. Lituiev, T. P. Copeland, M. S. Aboian, C. M. Aparici, S. Behr, R. R. Flavell, S.-Y. C. Huang, K. A. Zalocusky, L. Nardo, Y. Seo, R. A. Hawkins, M. H. Pampaloni, D. Hadley, and B. L. Franc, "A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain," *Radiology*, vol. 290, no. 2, pp. 456–464, 2019.

[62] B. E. Bejnordi , M. Veta, P. J. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *J. Amer. Med. Assoc.*, vol. 318, no. 22, pp. 2199–2210, 2017.

[63] Y. Fujisawa, Y. Otomo, Y. Ogata, Y. Nakamura, R. Fujita, Y. Ishitsuka, R. Watanabe, N. Okiyama, K. Ohara, and M. Fujimoto, "Deep learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumor diagnosis," *Brit. J. Dermatol.*, vol. 180, no. 61, pp. 373–381, 2019.

[64] S. S. Han, M. S. Kimm, W. Lim, G. H. Park, I. Park, and S. E. Chang, "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," *J. Invest Dermatol.*, vol. 138, no. 7, pp. 1529–1538, 2018.

[65] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. L. Ting, J. Zhu, J. Zhu, C. Li, S. Hewett, J. Dong, and K. Zhang, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018.

[66] Y. Kim, K. J. Lee, L. Sunwoo, D. Choi, C.-M. Nam, J. Cho, J. Kim, Y. J. Bae, R.-E. Yoo, B. S. Choi, C. Jung, and J. H. Kim, "Deep learning in diagnosis of maxillary sinusitis using conventional radiography," *Investigative Radiol.*, vol. 54, no. 1, pp. 7–15, Jan. 2019.

[67] Y. Kise, H. Ikeda, T. Fujii, M. Fukuda, Y. Ariji, H. Fujita, A. Katsumata, and E. Ariji, "Preliminary study on the application of deep learning system to diagnosis of Sjögren's syndrome on CT images," *Dentomaxillofacial Radiol.*, vol. 48, no. 6, Sep. 2019, Art. no. 20190019, doi: 10.1259/dmfr.20190019.

[68] J. H. Lee, J. H. Yoon, H. Na, E. Hong, K. Han, I. Jung, E.-K. Kim, H. J. Moon, V. Y. Park, E. Lee, and J. Y. Kwak, "Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound," *Head Neck*, vol. 41, no. 4, pp. 885–891, 2019.

[69] Y. Kumagai, K. Takubo, K. Kawada, K. Aoyama, Y. Endo, T. Ozawa, T. Hirasawa, T. Yoshio, S. Ishihara, M. Fujishiro, J.-I. Tamaru, E. Mochiki, H. Ishida, and T. Tada, "Diagnosis using deep-learning artificial intelligence based on the endocytoscopic observation of the esophagus," *Esophagus*, vol. 16, pp. 87–180, Dec. 2019.

[70] X. Li, S. Zhang, Q. Zhang, X. Wei, Y. Pan, J. Zhao, X. Xin, C. Qin, X. Wang, J. Li, F. Yang, Y. Zhao, M. Yang, Q. Wang, Z. Zheng, X. Zheng, X. Yang, C. T. Whitlow, and K. Chen, "Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: A retrospective, multicohort, diagnostic study," *Lancet Oncol.*, vol. 20, no. 2, pp. 193–201, Feb. 2019.

[71] J. Olczak, N. Fahlberg, A. Maki, A. S. Razavian, A. Jilert, A. Stark, O. Sköldenberg, and M. Gordon, "Artificial intelligence for analyzing orthopedic trauma radiographs," *Acta Orthopaedica*, vol. 88, no. 6, pp. 581–586, Nov. 2017.

[72] Y. Peng, S. Dharssi, Q. Chen, T. D. Keenan, E. Agrón, W. T. Wong, E. Y. Chew, Z. Lu, "DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs," *Ophthalmology*, vol. 126, no. 4, pp. 565–575, 2019.

[73] S. Shichijo, S. Nomura, K. Aoyama, Y. Nishikawa, M. Miura, T. Shinagawa, H. Takiyama, T. Tanimoto, S. Ishihara, K. Matsuo, and T. Tada, "Application of convolutional neural networks in the diagnosis of *Helicobacter pylori* infection based on endoscopic images," *EbioMedicine*, vol. 25, pp. 106–111, Nov. 2017.

[74] T. Urakawa, Y. Tanaka, S. Goto, H. Matsuzawa, K. Watanabe, and N. Endo, "Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network," *Skeletal Radiol.*, vol. 48, pp. 239–244, Jun. 2019.

[75] L. Wang, S. Yang, S. Yang, C. Zhao, G. Tian, Y. Gao, Y. Chen, and Y. Lu, "Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network," *World J. Surgical Oncol.*, vol. 17, no. 1, p. 12, Dec. 2019.

[76] L. Wu, W. Zhou, X. Wan, J. Zhang, L. Shen, S. Hu, Q. Ding, G. Mu, A. Yin, X. Huang, J. Liu, X. Jiang, Z. Wang, Y. Deng, M. Liu, R. Lin, T. Ling, P. Li, Q. Wu, P. Jin, and J. Chen, and H. Yu, "A deep neural network improves endoscopic detection of early gastric cancer without blind spots," *Endoscopy*, vol. 51, pp. 522–531, Jun. 2019.

[77] C. Yu, S. Yang, W. Kim, J. Jung, K.-Y. Chung, S. W. Lee, and B. Oh, "Acral melanoma detection using a convolutional neural network for dermoscopy images," *PLoS ONE*, vol. 13, no. 3, Mar. 2018, Art. no. e0193321.

[78] S. Mezgec and B. K. Seljak, "NutriNet: A deep learning food and drink image recognition system for dietary assessment," *Nutrients*, vol. 9, no. 7, p. 657, Jun. 2017.