An Extensive Survey of Machine Learning Based Approaches on Automated Pathology Detection in Chest X-Rays

Ravidu Suien Rammuni Silva
University of Westminster
London, UK.
w1626709@my.westminster.ac.uk

Pumudu Fernando
Informatics Institute of Technology
Colombo, Sri Lanka.
pumudu.f@iit.ac.lk

***Abstract*—Radiography is one of the most common and eminent medical imaging technologies in the world to date. Chest radiography is a very powerful and successful way of diagnosing thoracic diseases of humans. With the latest advancements and development in computer hardware, computer vision and especially with the publicly available large-scale datasets, machine learning based approaches on automated pathology detection in chest radiography have become increasingly popular among researchers. Our study conducts an extensive survey on existing machine learning approaches, its datasets and techniques on pathology detection in Chest X-Rays. The paper presents popular and publicly available labelled Chest X-Rays datasets with its specifications and discusses about the labellers, labelling methodologies used by them in a comprehensive discussion. Then, popular effective Image Processing techniques for Chest X-Rays images are presented. Then the paper further discusses about the current machine learning architectures used and portraits the effectiveness of Deep Convolutional Neural Networks for the purpose. Finally, the paper concludes with a discussion with gaps in current literature, unexplored areas and possible future with them in Machine Learning based automated pathology detection on Chest X-Rays.**

I. Introduction

During the past few years, Medical Artificial Intelligence (Medical AI) has influenced many researchers in the community. Lung nodule detection [1], eye disease diagnosis [2], cancer detection [3] and localization are some examples for it. Currently, various modes and methodologies are being used as medical imaging techniques. Among them, Chest X-Rays (CXRs) are a very common and successful medical imaging method in the medical field. CXRs can be used to correctly identify, a number of diseases located in the area of heart and lung. Those vary from low-risk diseases to life-threatening diseases like Pneumothorax. Most of those diseases could be diagnosed and cured only if it was identified correctly and in the early stages. Identification of these diseases highly depends on the skill and the experience of the radiologist. This dependency developed the need for Computer Aided Diagnosis (CAD) systems. With the hardware and software technology we had in earlier days it wasn't possible even though the need for CAD systems were there. The first attempt for a CAD system was found in the 1960s [4]. With the advancements of the hardware and software technologies, a new development pathway was opened for Machine Learning and especially for Deep Learning. However, screening CXR through deep learning techniques still remains challenging mainly due to following reasons: 1) Complex and diverse visual patterns that indicate different thoracic abnormalities; 2) Generating near-perfect annotations for the large CXR datasets are very challenging and expensive. Despite the mentioned challenges, researchers consistently try to tackle them as the need for CAD systems especially for CXRs are still there in the world.

People in countries with poor health services suffer a lot having very low numbers of qualified radiologists. Evidence in [5] shows as of 2015 only two qualified radiologists were available in the country of Liberia for a population of around four million. These resource-poor areas in the world will also benefit from CAD systems.

A Study in [6] has shown that CAD systems in the detection of pulmonary nodules have increased the accuracy of the radiologists' accuracy from 89.4% to 94.0%. This shows that even the radiologists could use CXR pathology detection systems as a support system. Oxipit ChestEye [7] is a CAD suite for X-Ray diagnosis which can identify healthy CXRs with high accuracy. It also has recently received the CE certification to use as a medical device. Qure.ai [8] is another similar certified CAD suite. But these systems are still under research level and very fewer details on their performances have been published.

In this paper, we summarize, analyse and critically evaluate the existing literature. Hence, will be a good starting point for any researcher who wants to improve and build a quality CXR pathology detection system using machine learning. With that intention we divide our main topic into four sub-topics:

- Datasets
- Labellers
- Image Pre-processing methods
- Existing ML algorithms in CXR classification.

These topics were chosen and organized aligning to necessity and cruciality of their contribution to the ML-based systems for CXR classification. Then in the discussion, we drive a general and comparative discussion on the overall findings exhibited in this paper.

ISSN 2305-7254

II. In-depth analysis of the existing literature

A. Datasets

Neural Networks by its nature are commonly data hungry. When it comes to Convolutional Neural Networks, data plays an even bigger role. Not only the size or the number of training images matters but also the visual quality, relevance to the subject and the labelling quality matters too. These factors directly reflect on the models' quality, the accuracy and the performance (depending on the application) one tries to build. When the applied domain becomes Bioinformatics, those factors become more critical and important. Creating a quality medical image dataset requires a substantial amount of work and many of them cannot be found publicly to the researchers. In [9], they have used 2.3 million CXRs to train their model, using several datasets created from several sources in India. But none of them is available to the public. in "Table I" we summarize, analyse and compare the current datasets available for a researcher who would want to build a CAD system for CXR classification.

From "Table I" it is noticeable that although there are plenty of CXR datasets available, only very few of them are capable of training a complex neural network. From the shown list, only the last three datasets would be sufficient for building a CNN model as per the findings in [10], but it is possible to use techniques like transfer learning to make use of smaller datasets. Non-Medical image models can also be used for this purpose. [11] have illustrated how this technique benefits for building medical image classification models.

Another approach would be to merge and mix two matching data sets. But the attention should be given in splitting training, validation and test sets, not to include CXRs of the same person in any two of the splits. This merging method could reduce the effect of biasing of the Model that is going to be built to any unwanted random patterns that could exist, caused by the sourced patients or equipment used.

Reference [12] shows that having two views (frontal and lateral) does not make much of a difference than having only one view in Pneumonia diagnosis through CXRs and the skill of the radiologist is what contributes more. But it is an open question to what extent having two views of CXRs helps in diagnosing CXRs having multiple possible diseases.

Typically, most of the Digital X-Ray machines output a specialized image file format called Digital Imaging and Communications in Medicine (DICOM). Then it will be converted into more common types like PNG or JPEG as reflected in "Table I". But this will reduce bit-depth of the images from 16-bit to 12-bits [13] and maybe even lesser.

B. Labellers

A very important and crucial aspect of these datasets is the accuracy of their ground truth labels. One could argue that it is the most important aspect of a dataset. Even though a good neural network can tolerate a few corrupted or irrelevant input data/images, it's common sense that we can never expect a model to predict accurately by training it with inaccurate data.

Most labellers are created using *Natural Language Processing* and *Text Mining* techniques. Radiology reports can be found usually in a semi-structured way. Commonly radiologists document their findings in structured and titled sections. As per [14], those sections may include but not limited to *Title, Indication, Procedure, Findings, Impression and Footnotes.* The data specifically related to the labelling of the CXRs can be usually found in below titled sections of the radiology reports;

- Findings - A descriptive explanation of the main aspects of the X-Ray Image
- Impressions - A more abstract and short description concentrated predominantly on the most notable findings relating to the explanation in the Findings section in the report.
- [Other] - An explanation of the aspects not covered in both 1 and 2 sections. Could include potential uncertainties.

Reference [14] points out a survey that shows more than 50% of physicians just refers only to the 'Impressions' section of a radiologist's report. Some reports may only have one or two out of the above-mentioned sections. In rare cases in which none of the above is present, the final section of the report is considered. In [13] around 83%, 12% and 4% of them had Findings, Impressions and other sections respectively.

In "Table II" we review and summarize four widely used Medical Report labellers. Labellers [15], [16] are generalized on clinical narrative contents and biomedical publications like Patient Discharge summaries, Radiology reports and Electrocardiograms (ECGs). As pointed in [15], it is more biased for reports like Patient Discharge Summaries. But [15] is a Machine learning based tool where [16] is an ontology-based tool. Method [17] have used a hybrid method using both those types of labellers which they found to be generating more accurate results when applied to radiology reports. Both of those tools lack a very important feature which is the 'uncertainty' which could occur in radiology reports. This issue was addressed in [18] by improving [16]'s architecture. Both NegBio [18] and CheXpert [19] were tested on 650+ manually labelled reports in [13] and the test results don't show much of a difference, but in CheXpert [19] the authors' test on their own dataset shows significantly better test results than NegBio. In fact, NegBio has slightly higher Recall whereas CheXpert has higher Precision.

Machine Learning Vs. Rule-based Methodologies:

As shown in "Table II" the methodologies of these labelling tools are mainly of two types. Ontology-based/Rule-based or ML-based. Rule-based methods mostly rely on keywords and the rules related to those keywords. As shown in [20] these labellers mostly use regular expressions which limits its capabilities in grasping and identifying useful information from complex and extensive sentences. To overcome this issue, instead of regular expressions, attempts were made in [21] and [22] by incorporating dependency structures / parse trees. But none of these graph-based dependency methodologies has been made available publicly.

References [23], [24] takes a different approach using machine learning for the purpose. Typically, ML-based labelling approaches are data-hungry hence need significant amounts of manually annotated ground truth data for a well-performing model. Unavailability of such datasets with sufficient size, limits the performances of those approaches. Training on available small-sized and single-sourced datasets could cause the model to be overfitted and ill perform on unseen data. Conclusively, it can be noted, although ML-based approaches could potentially perform better with enough data, currently it is observed that rule-based systems generally perform better with the available amount of related data.

We also noticed that none of these data sets provides information about how critical the identified condition on the CXR is. This limitation makes the model unable to differentiate between deadly Pneumothoraxes versus treated low-risk Pneumothoraxes using its visual inequality in CXRs. Also, this could potentially mislead the training model, reducing the accuracies as shown in [25].

TABLE I. SUMMARY OF PUBLICLY AVAILABLE CXR DATASETS AND ITS SPECIFICATIONS

Institution	Labelled Diseases	Labelling Method(s) (Automated /Manual)	Source(s)	Specifications
Korean Institute of Tuberculosis (KIT)[26]	Tuberculosis	Manual	-	• 10,848 DICOM Images
The Indiana University[27]	Cardiomegaly, Atelectasis, Tortuous Aorta, Hypo-inflated Lung, Lung Opacity, Pleural Effusion, Lung, Hyperinflation, Cicatrix, Calcinosis	Mixed	Two hospital systems in Indiana Network for Patient Care	• 8,121 DICOM Images • Frontal + Lateral views
Japanese Society of Radiological Technology[28]	Lung Nodule	Manual	13 Medical centres - Japan 1 Institution - United States	• 247 DICOM Images • Image Size: 2048x2048 • 12 bits grayscale colour depth
U.S. National Library of Medicine[29] (Montgomery County Dataset)	Tuberculosis	Manual	-	• 138 PNG / DICOM Images. • Image Size: 4,020×4,892/ 4,892×4,020 pixels
U.S. National Library of Medicine[29] (Shenzhen chest Dataset)	Tuberculosis	Manual	Shenzhen No.3 People's Hospital	• 662 PNG Images. • Image Size: 3000x3000 pixels
National Institutes of Health, US[17] ***(NIH)***	Cardiomegaly, Atelectasis, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, Hernia	Automated	Hospitals affiliated to National Institutes of Health Clinical Centre	• 112,120 PNG/DICOM Image • Image Size: • 3000x2000 pixels
Stanford University[19]	Enlarged Cardio., Cardiomegaly, Lung Lesion, Lung Opacity, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, (Support Devices)	Automated	Stanford Hospital	• 224,316 JPEG/DICOM Images • Frontal + Lateral Views • Image Size: • 1024x1024*/ 320x320* pixels *Varies
Massachusetts Institute of Technology[13]	Enlarged Cardio., Cardiomegaly, Lung Lesion, Lung Opacity, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, (Support Devices)	Automated	Beth Israel Deaconess Medical centre	• 377,110 JPEG/DICOM Images from 227,827 imaging studies • Frontal + Lateral Views • 8-bit Colour Depth • Image Size: 1800x2000* pixels *Varies

C. Image Pre-processing methods

Data is not always there exactly the way we want. Hence data pre-processing is a very important step in building a quality ML model. With the rise of Convolutional Neural Networks, these pre-processing techniques became more important than they were for older machine learning techniques which included steps like 'feature extraction'. As CNNs does 'feature extraction' by itself the data we give as input must be in a clean manner and having the expected Region of Interest (ROI) on the image in contrast with the other area is advantageous. This will influence CNN to learn, 'what we need it to learn'. Reference [30] proposes a novel way to remove off-distribution samples from the dataset. This will prevent the model from unrelated or distorted samples which reduce the model performance in training.

Although there are numerous methods of Image Pre-processing, we consider two techniques which seem evidently help pathology detection in CXRs:

1) Image Enhancement: Experiments done in [31] shows how critical Image Enhancement techniques are to CNN Models. It compares and contrasts five chosen Image Enhancement methods; Contrast Limited Adaptive Histogram Equalization (CLAHE) [32], Successive Means Quantization Transform (SMQT) [33], Adaptive Gamma Correction [34], Laplace Operator and Wavelet Transform with the ability of them in improving the model quality. It shows that CLAHE

[32] and Laplace Operator methods make considerable improvements to the models' accuracies and Recall rates.

2) Image Segmentation: Image Segmentation is generally of three types; Rule-based, Model-based and Machine learning based methods. Rule-based methods segment images based on a set of rules defined on the shape, patterns/texture, colour-intensities, location of pixels and their influence on the other pixels in an image. Reference [35] is an example of a Model-based method. ML-based methods are based on pixels and their characteristics. Shallow learning approaches like Support Vector Machine (SVM) and Deep Neural Network (DNN) based approaches were widely used to segment images. Shallow learning-based approaches depend on handcrafted features, where Machine Learning based approaches automatically learn them replacing manual feature extraction steps, but this requires a sufficiently sized dataset.

TABLE II. AN OVERVIEW OF MEDICAL TEXT LABELLERS

Labeller	Methodology	Performance
General Medical Report Labellers		
DNorm-C[15]	Uses Machine Learning with a Named Entity Recognition (NER) [36] methodology based on linear-chain conditional random fields. Pairwise Learning was used as a Normalization Method.	Macro Avg. Precision 0.80 Recall 0.71 F-score 0.75 (Evaluated in [37])
MetaMap[16]	Ontology-based method for the detection of texts in Unified Medical Language System® (UMLS®) [38].	Macro Avg. Precision 0.84 Recall 0.88 F-score 0.86 (Evaluated in [17])
Specialized Radiology Report Labellers		
NegBio[18]	Ontology-based system. An extension of [16] and [39] enabling labelling of negation and uncertainty discovered in the radiology reports.	Macro Avg. Precision 0.71 Recall 0.74 F-score 0.71 (Evaluated in [13])
CheXpert[19]	Rule-based system. Extract information from the radiology reports by three stages; *Mention Extraction, Mention Classification and Mention Aggregation.*	Macro Avg. Precision 0.72 Recall 0.73 F-score 0.71 (Evaluated in [13])

Lung Segmentation; For CXRs, these image segmentation methods are used mainly for *Lung Segmentation.* Approach [40] is the first approach in the literature where the feature engineering and shallow learning was used for lung segmentation from CXRs. After that many approaches could be found like [41] which shows some performance improvements. Using fuzzy C-means (FCM) clustering, together with Gaussian kernels and space constraints, an Unsupervised machine learning approach was attempted in [42]. This was tested on [28] and shows promising accuracy results. A framework; *Structure Correcting Adversarial Network (SCAN)* was introduced in [43]. This approach improves the NN and attains human-level performances, accurately segmenting the heart and the lung fields in CXRs.

D. Existing ML architectures in CXR classification

ML algorithms are capable of learning both noticeable and hidden patterns and relationships that could exist among a given set of training data. This identification mainly happens in two ways; Supervised and Unsupervised. Some algorithms use a hybrid approach of these two; Semi-supervised learning. In supervised learning methodologies, humans influence the ML models, training with expecting outputs to corresponding inputs. Unsupervised learning methodologies look for the characteristics of the inputted data by itself. Apart from the image pre-processing techniques discussed in the previous section, Pathology Detection in CXRs is mainly done using supervised training. With the large-scale data availability as mentioned in section 1, almost all recent literature is found to be using Convolutional Neural Networks (CNNs). Although CNN classification is listed commonly under supervised learning, the internals of CNNs behave in an unsupervised manner extracting the relevant features by its own.

1) CNNs in CXR classification: When the data happens to be Image data, the process of identifying hidden characteristics of data becomes more complex. This complexity is mostly because of the visual pattern the data in an image represents. For example, it is possible to take an image and make it unrecognizable just by changing just the order of the pixels, while keeping the set of pixel data unchanged. Because of this reason, existing ML architecture did not perform well in the early stages of ML development. To overcome this in 1989, [44] came up with a brand new concept; Convolutional Neural Networks (CNNs). His approach did very well in identifying handwritten zip codes at that time. From that point, CNNs have come a long way to this date. It was found to be very useful even for identifying various pathologies in CXRs. Methods like [45] use feature extraction with Fully Connected Networks (FCNs) to detect abnormalities. Evidently, this method is very useful when we have a lesser amount of data. But when we have sufficient data and with techniques like transfer learning, CNNs would be a better way to go. Reference [46] shows how even models trained on non-medical images can be used as a transfer learning approach to build medical models effectively. Basically, abnormalities in CXRs are characterized under three types; Texture abnormalities, Focal abnormalities and Shape abnormalities [47]. CNNs do a good job in identifying these visual characteristics of CXRs, hence identifying the pathologies.

2) CNNs in existing Literature: Usage of CNNs for pathology detection in CXRs was first demonstrated by [17] using their own release of a dataset which we discussed in detail in section I of this paper. Initially, they classified eight common pathologies on CXRs and then they extended it to fourteen pathologies. As summarized in "Table III", it can be seen that DenseNet-121 [48] based models are common and are the best performing models. In [49] authors have researched the usage of LSTM [50] for the pathology classification. By that, they have explored the dependencies between the labels in a single X-Ray and its relation to the final prediction. But it is doubtful that whether the LSTM network is or is not affected by the ordering of the labels as LSTM is mainly based on text recognition and takes the sequential order of the text inputs into account in the predictions. However, their approach has increased the average AUC of the model by a considerable amount as shown in "Table III".

In CheXNet [51] the top Fully Connected Node of the original DenseNet architecture was removed and replaced with a GlobalAveragePoolingLayer and a Dense Layer with a sigmoid activation function. We noticed that although the authors were mainly concentrating on Pneumonia this architecture has shown success in all other pathologies as well. They again proved the applicability of the DenseNet for CXR classification and [52], [53] also have got some influence from it. With the findings of the research done in [53], we noticed that the potion of the dataset which is used to evaluate the model can dramatically change the evaluation results. This conclusion was due to the re-evaluation results of [51] achieved by [53] as shown in "Table III" on the official split of the [17] dataset using the methodology introduced by [51].

TABLE III. SUMMARY OF CURRENTLY AVAILABLE TOP-PERFORMING CNN-BASED MODELS; AUC SCORES ATTAINED BY EACH MODEL WAS DEPICTED FROM THE PUBLICATIONS. NOT AVAILABLE DATA ARE DENOTED BY "-". MACRO AVERAGES ARE COMPUTED ON AVAILABLE AUC SCORES. [19] AND [54] AVERAGES WERE NOT COMPARED AS VERY FEW DATA ARE AVAILABLE. MACRO AVERAGES ON THE OFFICIAL SPLIT ARE BOLDED. 'R-50' – RESNET-50, 'D-121'-DENSENET-121, 'G-CNN'-GUIDED CNN, 'ENS'-ENSEMBLED MODEL

	[17]	**[51]**	**[53]**	**[49]**	**[52]**	**[53]**	**[19]**	**[54]**	**[55]**	**[56]**	**[57]**
Dataset	CXR-14	CXR-14	CXR-14	CXR-14	CXR-14	CXR-14	CheXPert	CheXPert	MIMIC-CXR	CXR-14	CXR-14
Dataset Split	**Official**	Custom	**Official**	**Official**	Custom	**Official**	**Official**	**Official**	**Official**	Custom	**Official**
Backbone NN	R-50	D-121	D-121	D-121	D-121	D-121	Ens	Ens	Custom	G-CNN	G-CNN
Atelectasis	0.716	0.809	0.780	0.733	0.767	0.792	0.850	0.909	0.766	0.853	0.781
Cardiomegaly	0.807	0.925	0.882	0.858	0.883	0.881	0.900	0.910	0.840	0.939	0.885
Effusion	0.784	0.864	0.827	0.806	0.828	0.842	-	-	0.757	0.903	0.832
Infiltration	0.609	0.735	0.690	0.675	0.709	0.710	-	-	0.748	0.754	0.700
Mass	0.706	0.868	0.831	0.727	0.821	0.847	-	-	0.692	0.902	0.815
Nodule	0.671	0.780	0.781	0.778	0.758	0.811	-	-	0.568	0.828	0.765
Pneumonia	0.633	0.768	0.735	0.690	0.731	0.740	-	-	0.625	0.774	0.719
Pneumothorax	0.806	0.889	0.851	0.805	0.846	0.876	-	-	0.706	0.921	0.866
Consolidation	0.708	0.790	0.754	0.717	0.745	0.760	0.900	0.955	0.632	0.842	0.743
Edema	0.835	0.888	0.850	0.806	0.835	0.848	0.920	0.958	0.734	0.924	0.842
Emphysema	0.815	0.937	0.930	0.842	0.895	0.942	-	-	-	0.932	0.921
Fibrosis	0.769	0.805	0.822	0.757	0.818	0.833	-	-	0.761	0.864	0.835
Pleural Thickening	0.708	0.806	0.793	0.724	0.761	0.808	-	-	0.687	0.837	0.791
Hernia	0.767	0.916	0.932	0.824	0.896	0.934	-	-	0.815	0.921	0.911
Pleural Effusion	-	-	-	-	-	-	0.970	0.964	-	-	-
Macro-Avg.	**0.738**	0.841	**0.818**	**0.767**	0.807	**0.830**	**-**	**-**	0.717	0.871	**0.815**

In [52] they use PLCO dataset [58] in order to incorporate the spatial location of the disease on the image for the prediction making. We didn't include the final averaged AUC presented in the paper to the "Table III" as it uses a different split. It is unclear whether its performance is stable when multiple diseases and locations are present and when the locations are overlapping which could happen commonly. However, it shows that the spatial location information improves the AUC by achieving an AUC value of 0.87 on the PLCO dataset.

Having a lateral view X-Ray of the patient is helpful in identifying any disease indicates in an X-Ray more accurately. In that spirit [19] released the 'CheXpert' dataset which has multiple X-Rays in multiple views. As discussed in a previous section they claim to have better labeller as well. They have launched a competition [59] which only considers 5 pathologies out of 12 diseases labelled. Also, [54] is built focused on the competition which only showcases data related to the same 5 pathologies. Unfortunately, life-threatening and medically important diseases like Pneumothorax are not found in the selected evaluation pathology set.

3) Guided CNN networks: Although CNNs were made to learn on its own, in the literature we found that, if we influence the CNNs in a way that we want it to learn, it produces higher results. This is even useful for training CNNs with comparatively smaller, noisy datasets or for datasets with weak ground truth labels. Guided CNN networks have more potential also in accurately identifying the pathologies in CXRs. Below we summarize the findings of our thorough and critical investigation on the top publications done on three main publicly available datasets.

III. DISCUSSION

With current technologies and the improvements in Machine Learning, classification of pathologies in CXRs have achieved significant progress. Although the AUC scores which reflect on the quality of the built systems seem to be promising still there are some drawbacks in them which is why they are still not professionally used in medical systems. We already discussed some of those issues in this paper. Below we discuss few and more general drawbacks we found after a thorough investigation of the existing literature.

- Almost all the CXR datasets were constructed in one or few specific geographically closer sources. Due to this fact, the models built on those datasets may not perform in reality up to the exhibited evaluation results in the testing. The reason could be due to hidden and irrelevant patterns which could be due to the equipment, medical conditions of the patients, imaging procedures and

technologies. Training the algorithm in a more diverse set of CXR images would be a possible solution for this problem of sensitivity to Geographic Variation. We believe a well-planned systematic study, utilizing the available large-scale datasets from various sources, could build a system overcoming this issue.

- Another notable drawback of the existing models is that all of them just consider only on the CXR image or more images if available, to detect pathologies. But medical professionals investigate various external factors including but not limited to; gender, age, behavioural qualities, symptoms, medical history and various other clinical tests. Even though it is infeasible to build a model on every contributing feature mentioned before, we believe data like gender, age, symptoms and medical history could be incorporated for the prediction making and it will improve the quality of the predictions. But it is an open question whether these features actually contribute to the prediction making process and if so, to what extent and how those features affect the process.
- We noted that in almost all the approaches on CXR pathology detection, they have used existing architectures with few modifications. In some approaches, they have used model ensembling as a technique and it proved to be performing successfully. As we discussed this reflects on the abilities of some CNN architectures to perform notably well on just specific pathologies. This could be due to the architectural qualities of those CNNs. We believe proper manipulation of the predictions from each CNN type using sophisticated ensembling methodologies could gain some performance boost to the overall system.

TABLE IV. SUMMARY OF TOP RESEARCHES AND ITS ALGORITHMIC ANALYSIS DONE UNDER EACH LARGE DATASET. THIS TABLE CONTAINS A SUMMARY OF EACH REVIEWED TOP APPROACHES UNDER EACH DATASET WHICH THOSE WERE TRAINED ON.

	Approach Summary	**Gaps Exploration**
Dataset: **ChestX-ray14 [17]**		
[53]	• Incorporates *squeeze and excitation* blocks (SEnet) introduced in [60] into the CNN networks. • Uses the ability of SE blocks to further differentiate the areas with abnormalities from normal areas in CXRs. • Uses a multi-map transfer layer and max-min pooling methodology to deal with different pathologies separately.	• For some diseases like Hernia, SE blocks, multi-map layer and max-min pooling have reduced the AUC scores. • Heavy test time augmentation has been done generating five different image crops and a one horizontally flipped image from each test image. Less evidence has been provided on the effect on these augmentation techniques to the evaluation results
[57]	• Uses lung segmentation techniques to influence the CNN network more on the lung area for making the predictions	• Slight to no improvement in AUC scores even with heavy computations in the algorithm compared to [51] which only uses the original DenseNet-121 network.
[56]	• Proposes a three-branch Guided CNN; *Global Branch, Local Branch and Fusion Branch*. • Firstly, it localizes the abnormality in the global branch then uses a cropping procedure to crop that area and pass through that portion again in a CNN network in the local branch. Then the predictions from both Global and the Local branches were concatenated.	• The cropped image of the CXRs that indicates multiple diseases on multiple spanned spatial locations will be very similar to the uncropped image • If the NN fails to recognize true abnormal areas in the global branch then the local branch would never be able to make a useful prediction, which makes the local branch highly dependent on the global branch • Trained and tested on a custom split. Hence the evaluation results cannot be guaranteed. • This approach does not guarantee that CXRs from the same patient exist in different sets of dataset splits, which could falsely increase the evaluation accuracies.
Dataset: **CheXpert [51]**		
[54]	• Ranks at the top 2nd position in the CheXpert [59] competition • Used model ensembling and Label Smoothing Regularization (LSR) • Used a novel Conditional Training (CT) methodology which uses hierarchical dependencies among different pathologies to each other to improve the accuracy of the model predictions.	• The evaluation was only done on five pathologies out of twelve in the dataset. • Final predictions were taken from 6 trained CNNs which will make the model very large and will need high computational power to make predictions and even higher power to train or re-train the model. • Authors have acknowledged the generalization issues in Conditional Training methodology. This methodology will make the predictions significantly biased to the distribution that data comes from. Hence produce higher accuracy scores only on the trained distribution.
Dataset: **MIMIC-CXR [13]**		
[55]	• Specifically designed to take the two views of the X-Rays; frontal and lateral, in and work with them parallelly in making the final prediction	• AUC scores are significantly low compared to the related publications on other datasets. But no published work has been found working on MIMIC-CXR other than [55].

- As discussed in a previous section labelling of the CXRs also plays an important role. In some approaches as in [54], we found to be using techniques like LSR. It is understandable that labels could become a little noisy when large labelled datasets are being created. This is mainly because of the automation involved in labelling the samples. But proper procedures must be taken to eliminate these errors as much as possible.
- After taking a deep look at some large datasets' images [13], [17], [19] we found a considerable amount of noisy data with issues like undesired rotations, distortions, texts, signs and foreign objects. These errors could possibly harm the model performance. Methods like introduced in [30], could potentially help to overcome this issue.
- Looking at the existing literature we noticed that building a CAD system from the scratch involves different, sometimes technically opposite disciplines, as rule-based methodologies have incorporated to build up the ground truth labels, and DL is incorporated to build up the final model. Hence, it's a perfect blend of DL and rule-based methodologies. We didn't find any attempt made to build up a similar system based solely on rule-based methods or solely on DL based methods.
- After investigating the existing literature we believe that labellers mentioned in an earlier section could be further developed in a way that more useful information can be grasped from the radiology reports, which will allow labelling the CXR images not only on the presence of specific conditions but also how critical the identified condition is.
- We also noted that the literature lacks studies investigating how these ML-based systems could be practically used and their effectiveness in the practical clinical environment.
- As mentioned previously, no attempt has been made to incorporate patients' symptoms for the detection results. Hence, it would be a valuable contribution to the research community if such a study could be carried on.
- Also, looking at the overall picture of the existing literature, there is a notable requirement for a study which comprehensively analyse how the well-performing DL algorithms available to date performs and how those can be optimized, in a way it can produce even better results when trained on grayscale CXR images. We see [55] as a good initiative for this. We strongly believe this is very much possible with the large scale publicly available datasets, which could be used to repeatedly verify the performances of the algorithms.

IV. Conclusion

In this paper, the discussion was driven under four main sections. In the first section, we reviewed almost all the current publicly available datasets of CXRs to date. We also presented an in-depth analysis of those datasets. We identified how large-scale datasets could possibly support the research community in the development of quality DL models with reliable performances.

Even though it is possible to gather millions of CXRs even in a day, the challenge comes in labelling those gathered data. As manual labelling, which is also not perfect, is expensive and very inefficient, automated labelling plays an important role in creating complete annotated and usable datasets. It was found that even though ML-based labellers have more potential, rule-based labellers currently surpass their performance due to the lack of training data for ML-based models.

Then we noticed how various image processing techniques could help in utilizing small scale datasets into training machine learning models. However, with the availability of large-scale datasets, algorithms involving manual processes like feature extraction are being widely replaced by DL algorithms, specifically, well-performing CNN based algorithms.

Instead of attempting to build various new manual processing techniques, evidence in the recent literature shows attempts to build novel CNN architectures are more effective, allowing the algorithm to learn the 'features' itself and to learn how to neglect any noisy representations in the input data. Due to this reason, in the DL research community, enthusiasm to construct various new manual processing techniques like feature extraction has been gradually and increasingly getting replaced by the optimization and construction of new DL based algorithms.

Acknowledgements

We express our gratitude to Dr Harshana Bandara (MBBS, MD), Dr Nilmini Fernando (MBBS, DFM) and Dr Prasantha De Silva (MBBS, MSc, MD) for the valuable knowledge and insights shared with us for this paper.

References

[1] A. A. A. Setio *et al.*, 'Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge', *Medical Image Analysis*, vol. 42, pp. 1–13, Dec. 2017, doi: 10.1016/j.media.2017.06.015.

[2] V. Gulshan *et al.*, 'Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs', *JAMA*, vol. 316, no. 22, p. 2402, Dec. 2016, doi: 10.1001/jama.2016.17216.

[3] C. Chen, V. C. Kavuri, X. Wang, R. Li, H. Liu, and J. Huang, 'Multi-frequency diffuse optical tomography for cancer detection', in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, Brooklyn, NY, USA, Apr. 2015, pp. 67–70, doi: 10.1109/ISBI.2015.7163818.

[4] G. S. Lodwick, T. E. Keats, and J. P. Dorst, 'The Coding of Roentgen Images for Computer Analysis as Applied to Lung Cancer', *Radiology*, vol. 81, no. 2, pp. 185–200, Aug. 1963, doi: 10.1148/81.2.185.

[5] UMass Memorial Healthcare *et al.*, 'Diagnostic Radiology in Liberia: A Country Report', *JGR*, vol. 1, no. 2, Nov. 2015, doi: 10.7191/jgr.2015.1020.

[6] T. Kobayashi, 'Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs', p. 6, 1996.

[7] 'Oxipit', *Oxipit - The first AI chest X-Ray radiology suite for healthy patient reports*, 2020. https://oxipit.ai/.

[8] 'Qure.ai', *Qure.ai - Artificial Intelligence for Radiology*. https://qure.ai/.

[9] P. Putha *et al.*, 'Can Artificial Intelligence Reliably Report Chest X-Rays?: Radiologist Validation of an Algorithm trained on 2.3 Million X-Rays', *arXiv:1807.07455 [cs]*, Jun. 2019, Accessed: Aug. 18, 2020. [Online]. Available: http://arxiv.org/abs/1807.07455.

[10] C. Luo, X. Li, L. Wang, J. He, D. Li, and J. Zhou, 'How Does the Data set Affect CNN-based Image Classification Performance?', in *2018 5th International Conference on Systems and Informatics (ICSAI)*, Nanjing, Nov. 2018, pp. 361–366, doi: 10.1109/ICSAI.2018.8599448.

[11] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, 'Chest pathology detection using deep learning with non-medical training', in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, Brooklyn, NY, USA, Apr. 2015, pp. 294–297, doi: 10.1109/ISBI.2015.7163871.

[12] T. Lynch, 'Does the Lateral Chest Radiograph Help Pediatric Emergency Physicians Diagnose Pneumonia? A Randomized Clinical Trial', *Academic Emergency Medicine*, vol. 11, no. 6, pp. 625–629, Jun. 2004, doi: 10.1197/j.aem.2003.12.019.

[13] A. E. W. Johnson *et al.*, 'MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs', *arXiv:1901.07042 [cs, eess]*, Nov. 2019, Accessed: Jul. 24, 2020. [Online]. Available: http://arxiv.org/abs/1901.07042.

[14] J. R. Wilcox, 'The written radiology report', p. 4, 2006.

[15] R. Leaman, R. Khare, and Z. Lu, 'Challenges in clinical natural language processing for automated disorder normalization', *Journal of Biomedical Informatics*, vol. 57, pp. 28–37, Oct. 2015, doi: 10.1016/j.jbi.2015.07.010.

[16] A. R. Aronson and F.-M. Lang, 'An overview of MetaMap: historical perspective and recent advances', *J Am Med Inform Assoc*, vol. 17, no. 3, pp. 229–236, May 2010, doi: 10.1136/jamia.2009.002733.

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, 'ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases', p. 20, 2017.

[18] Y. Peng, 'NegBio: a high-performance tool for negation and uncertainty detection in radiology reports', p. 9.

[19] J. Irvin *et al.*, 'CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison', *arXiv:1901.07031 [cs, eess]*, Jan. 2019, Accessed: Jul. 24, 2020. [Online]. Available: http://arxiv.org/abs/1901.07031.

[20] H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman, 'ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports', *Journal of Biomedical Informatics*, vol. 42, no. 5, pp. 839–851, Oct. 2009, doi: 10.1016/j.jbi.2009.05.002.

[21] S. Sohn, S. Wu, and C. G. Chute, 'Dependency Parser-based Negation Detection in Clinical Narratives', p. 8.

[22] S. Mehrabi *et al.*, 'DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx', *Journal of Biomedical Informatics*, vol. 54, pp. 213–219, Apr. 2015, doi: 10.1016/j.jbi.2015.02.010.

[23] Y. Huang and H. J. Lowe, 'A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports', *Journal of the American Medical Informatics Association*, vol. 14, no. 3, pp. 304–311, May 2007, doi: 10.1197/jamia.M2284.

[24] C. Clark *et al.*, 'MITRE system for clinical assertion status classification', *J Am Med Inform Assoc*, vol. 18, no. 5, pp. 563–567, Sep. 2011, doi: 10.1136/amiajnl-2011-000164.

[25] A. G. Taylor, C. Mielke, and J. Mongan, 'Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study', *PLoS Med*, vol. 15, no. 11, p. e1002697, Nov. 2018, doi: 10.1371/journal.pmed.1002697.

[26] S. Ryoo and H. J. Kim, 'Activities of the Korean Institute of Tuberculosis', *Osong Public Health and Research Perspectives*, vol. 5, pp. S43–S49, Dec. 2014, doi: 10.1016/j.phrp.2014.10.007.

[27] D. Demner-Fushman *et al.*, 'Preparing a collection of radiology examinations for distribution and retrieval', *J Am Med Inform Assoc*, vol. 23, no. 2, pp. 304–310, Mar. 2016, doi: 10.1093/jamia/ocv080.

[28] J. Shiraishi *et al.*, 'Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule: Receiver Operating Characteristic Analysis of Radiologists' Detection of Pulmonary Nodules', *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, Jan. 2000, doi: 10.2214/ajr.174.1.1740071.

[29] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, 'Two public chest X-ray datasets for computer-aided screening of pulmonary diseases', p. 3.

[30] E. Çallı, K. Murphy, E. Sogancioglu, and B. van Ginneken, 'FRODO: Free rejection of out-of-distribution samples: application to chest x-ray analysis', *arXiv:1907.01253 [cs, eess, stat]*, Jul. 2019, Accessed: Aug. 08, 2020. [Online]. Available: http://arxiv.org/abs/1907.01253.

[31] X. Chen, 'Image enhancement effect on the performance of convolutional neural networks', p. 40.

[32] E. D. Pisano *et al.*, 'Contrast Limited Adaptive Histogram Equalization image processing to improve the detection of simulated spiculations in dense mammograms', *J Digit Imaging*, vol. 11, no. 4, pp. 193–200, Nov. 1998, doi: 10.1007/BF03178082.

[33] Mikael Nilsson, Mattias Dahl, and Ingvar Claesson, 'The Successive Mean Quantization Transform', in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Philadelphia, Pennsylvania, USA, 2005, vol. 4, pp. 429–432, doi: 10.1109/ICASSP.2005.1416037.

[34] S. Rahman, M. M. Rahman, M. Abdullah-Al-Wadud, G. D. Al-Quaderi, and M. Shoyaib, 'An adaptive gamma correction for image enhancement', *J Image Video Proc.*, vol. 2016, no. 1, p. 35, Dec. 2016, doi: 10.1186/s13640-016-0138-1.

[35] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, 'Active Shape Models-Their Training and Application', *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, Jan. 1995, doi: 10.1006/cviu.1995.1004.

[36] R. Grishman and B. Sundheim, 'Message Understanding Conference-6: A Brief History', p. 6, 1996.

[37] L. Kelly *et al.*, 'Overview of the ShARe/CLEF eHealth Evaluation Lab 2014', in *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, vol. 8685, E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, and E. Toms, Eds. Cham: Springer International Publishing, 2014, pp. 172–191.

[38] NIH, 'Unified Medical Language System (UMLS)', *NIH*. https://www.nlm.nih.gov/research/umls/index.html (accessed Oct. 16, 2020).

[39] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, 'A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries', *Journal of Biomedical Informatics*, vol. 34, no. 5, pp. 301–310, Oct. 2001, doi: 10.1006/jbin.2001.1029.

[40] M. F. McNitt-Gray, H. K. Huang, and J. W. Sayre, 'Feature selection in the pattern classification problem of digital chest radiograph segmentation', *IEEE Trans. Med. Imaging*, vol. 14, no. 3, pp. 537–547, Sep. 1995, doi: 10.1109/42.414619.

[41] B. van Ginneken, 'Automatic Segmentation of Lung Fields in Chest Radiographs', p. 8.

[42] Z. Shi, P. Zhou, L. He, T. Nakamura, Q. Yao, and H. Itoh, 'Lung Segmentation in Chest Radiographs by Means of Gaussian Kernel-Based FCM with Spatial Constraints', in *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, Tianjin, China, 2009, pp. 428–432, doi: 10.1109/FSKD.2009.811.

[43] W. Dai *et al.*, 'SCAN: Structure Correcting Adversarial Network for Organ Segmentation in Chest X-rays', *arXiv:1703.08770 [cs]*, Apr. 2017, Accessed: Aug. 02, 2020. [Online]. Available: http://arxiv.org/abs/1703.08770.

[44] LeCun Y, Denker JS, and Boser B, 'Backpropagation applied to handwritten zip code recognition', *Neural Comput 1*, pp. 541–551, 1989.

[45] H. R. H. Al-Absi, B. B. Samir, and S. Sulaiman, 'A Computer Aided Diagnosis System for Lung Cancer based on Statistical and Machine Learning Techniques', *JCP*, vol. 9, no. 2, pp. 425–431, Feb. 2014, doi: 10.4304/jcp.9.2.425-431.

[46] Y. Bar, I. Diamant, L. Wolf, and H. Greenspan, 'Deep learning with non-medical training used for chest pathology identification', Orlando, Florida, United States, Mar. 2015, p. 94140V, doi: 10.1117/12.2083124.

[47] B. van Ginneken, S. Katsuragawa, B. M. ter Haar Romeny, Kunio Doi, and M. A. Viergever, 'Automatic detection of abnormalities in chest radiographs using local texture analysis', *IEEE Trans. Med. Imaging*, vol. 21, no. 2, pp. 139–149, Feb. 2002, doi: 10.1109/42.993132.

[48] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, 'Densely Connected Convolutional Networks', *arXiv:1608.06993 [cs]*, Jan.

2018, Accessed: Aug. 06, 2020. [Online]. Available: http://arxiv.org/abs/1608.06993.

[49] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, 'Learning to diagnose from scratch by exploiting dependencies among labels', *arXiv:1710.10501 [cs]*, Feb. 2018, Accessed: Aug. 05, 2020. [Online]. Available: http://arxiv.org/abs/1710.10501.

[50] S. Hochreiter and J. Schmidhuber, 'Long Short-Term Memory', *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[51] P. Rajpurkar *et al.*, 'CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning', *arXiv:1711.05225 [cs, stat]*, Dec. 2017, Accessed: Aug. 05, 2020. [Online]. Available: http://arxiv.org/abs/1711.05225.

[52] S. Guendel *et al.*, 'Learning to recognize Abnormalities in Chest X-Rays with Location-Aware Dense Networks', *arXiv:1803.04565 [cs]*, Mar. 2018, Accessed: Aug. 05, 2020. [Online]. Available: http://arxiv.org/abs/1803.04565.

[53] C. Yan, J. Yao, R. Li, Z. Xu, and J. Huang, 'Weakly Supervised Deep Learning for Thoracic Disease Classification and Localization on Chest X-rays', *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 103–110, Aug. 2018, doi: 10.1145/3233547.3233573.

[54] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen, 'Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels', *arXiv:1911.06475 [cs, eess]*, Jun. 2020, Accessed: Aug. 03, 2020. [Online]. Available: http://arxiv.org/abs/1911.06475.

[55] J. Rubin, D. Sanghavi, C. Zhao, K. Lee, A. Qadir, and M. Xu-Wilson, 'Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks', *arXiv:1804.07839 [cs, stat]*, Apr. 2018, Accessed: Aug. 03, 2020. [Online]. Available: http://arxiv.org/abs/1804.07839.

[56] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, 'Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification', *arXiv:1801.09927 [cs]*, Jan. 2018, Accessed: Jul. 24, 2020. [Online]. Available: http://arxiv.org/abs/1801.09927.

[57] H. Liu, L. Wang, Y. Nan, F. Jin, Q. Wang, and J. Pu, 'SDFN: Segmentation-based deep fusion network for thoracic disease classification in chest X-ray images', *Computerized Medical Imaging and Graphics*, vol. 75, pp. 66–73, Jul. 2019, doi: 10.1016/j.compmedimag.2019.05.005.

[58] J. K. Gohagan, P. C. Prorok, R. B. Hayes, and B.-S. Kramer, 'The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: History, organization, and status', *Controlled Clinical Trials*, vol. 21, no. 6, pp. 251S-272S, Dec. 2000, doi: 10.1016/S0197-2456(00)00097-0.

[59] 'CheXpert: A Large Chest X-Ray Dataset And Competition', *CheXpert: A Large Chest X-Ray Dataset And Competition*. https://stanfordmlgroup.github.io/competitions/chexpert/.

[60] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, 'Squeeze-and-Excitation Networks', *arXiv:1709.01507 [cs]*, May 2019, Accessed: Aug. 08, 2020. [Online]. Available: http://arxiv.org/abs/1709.01507.