AI Integration in Medical Imaging: Advanced Analysis of Chest X-ray

Danushka Bandara
Intelligent Technologies Research Group, ACE, UEL
London, UK
u2347071@uel.ac.uk

Thamo Sutharssan
Department of Engineernig ACE, UEL
London, UK
0000-0003-1836-0517

Mhd Saeed Sharif
Intelligent Technologies Research Group, ACE, UEL
London, UK
s.sharif@uel.ac.uk

***Abstract*—In this research, we introduce two types of Artificial Intelligence (AI) models for classifying chest X-rays, binary and categorical. These models were trained and validated utilizing Convolutional Neural Network (CNNs) and transfer learning techniques. The binary classification model performed well in classifying normal and abnormal X-rays. The categorical classification model showed good abilities to recognize pathological states such as cardiomegaly and infiltration. However, it faced challenges when radiographic patterns overlapped. We used a dataset of 2,463 chest X-ray images with various pathological conditions and improved CNN architectures with two validation approaches to ensure robustness and reliability. This study contributes to the growing literature on AI in medical imaging, showing enhanced clinical outcomes with robust performance and predictive capabilities.**

Keywords— AI, Medical Imaging, chest X-ray, CNN, VGG16 model architecture, binary classification model, categorical classification model, cardiothoracic pathological conditions

I. Introduction

Medical imaging is a critical component of modern healthcare. The chest X-ray provides significant insights into the diagnosis and treatment of a variety of respiratory and cardiothoracic pathologies [1]. Interpreting chest X-rays is still challenging due to their complicated image details and requires a significant amount of knowledge. Inconsistencies and errors during interpretation could be triggered by a variety of human factors such as fatigue, mental stress, and different levels of skills and experience [2]. These factors emphasize the need for supplementary diagnostic tools for radiologists and physicians to ensure accurate diagnoses.

Prior studies have verified that AI has the potential to strengthen the efficiency and precision of medical image processing [3]. CNN has demonstrated significant potential in analyzing and classifying medical images with high precision [4]. Some studies have shown that AI systems assist radiologists by providing automated, accurate and consistent analysis of medical images, reducing physicians' workload and speeding up the diagnostic process [5].

The primary problem of this research is to address the variability and potential errors in interpreting chest X-rays caused by human factors. Misdiagnoses and inconsistent readings can lead to inappropriate treatments and bad outcomes for patients. We hypothesize that AI models can improve the accuracy and consistency of identifying and classifying chest X-rays. By automating diagnostic processes, AI models can assist radiologists and physicians in better clinical decision-making. In this study, we aim to develop AI models utilizing advanced CNN architectures and transfer learning techniques, focusing on VGG-16 [6]. The specific selection of VGG-16 comes from its proven effectiveness in feature extraction for image classification tasks. Even though VGG-16 is an older design, in some image classification tasks it performed well compared to more complicated designs like Resnet or EfficientNet. It is easy to construct and understand because of its 2D Convolutional (Conv2D) layer structure, which includes max-pooling layers and fixed filter sizes. Compared to the residual connection of ResNet or the compound scaling strategies of EfficientNet, this simplicity leads to fewer operation layers and lowers computational complexity. Additionally, VGG-16 works well with smaller batch sizes and less memory. It is also easier to fine-tune for transfer learning applications, especially when dealing with smaller datasets and limited hardware resources. While ResNet and EfficientNet can achieve higher accuracy through more sophisticated operations, VGG-16 offers a practical balance between accuracy and computational efficiency [7]. This literature shows that VGG-16 is a better choice for small datasets and lower computational power. The uniform structure and deep layers of VGG-16 facilitate the extraction of fine-grained features, which makes it an excellent choice for identifying subtle differences in medical images such as chest X-rays [7].

The aim of this research is to develop and validate two AI models: one for binary classification and another for categorical classification of chest X-rays. Our techniques employ cutting-edge deep learning methodologies to achieve high classification accuracy in both past preliminary assessments and more extensive diagnostic tasks. This binary classification model will greatly reduce the load on radiologists by classifying normal and abnormal X-rays. The categorical classification model aims to discriminate between multiple conditions with great precision in pneumonia, atelectasis, pneumothorax, mass, infiltration, nodule, cardiomegaly, effusion, and normal.

This work serves as a contribution to the AI field on optimizing diagnostic workflows. It is the first full clinical evaluation of such models to set a benchmark for AI-driven medical diagnostics in the future. The above AI models are being evaluated and validated under rigorous conditions. After those evaluation processes, the practical efficacy of these AI models become apparent, which improves patient management and health outcomes.

II. Methodology

This research uses chest X-ray images collected from Sri Lankan hospitals and publicly available datasets including the National Institute of Health (NIH) datasets [8], MIMIC dataset [9] and the chest X-ray 8 dataset [8]. As shown in Fig. 2, 2463 chest X-ray images were collected, representing 8 cardiothoracic pathological conditions and a diverse range of demographic backgrounds. This variety of datasets is critical for constructing an AI model that can generalize across multiple demographic patient groups ensuring greater efficiency while avoiding demographic biases [10]. As shown in Fig. 1, the dataset includes eight types of cardiothoracic

pathological conditions such as pneumothorax, pneumonia, atelectasis, infiltrate, mass, nodule, cardiomegaly, and effusion, along with normal cases. The pathological conditions were classified and labelled according to radiologists' reports and existing labels from the public datasets. It is crucial to accurately depict these pathological conditions in the dataset to train AI models effectively [11].

The methodology started with key steps, the first of which was data preparation. The entire dataset was resized to 150 x150 pixels size [12]. To normalize the input dataset, we employed several processing steps. These steps involved rescaling to normalize values of pixels, as well as using augmentation strategies such as zoom, rotation, shear, width and height shifts, and horizontal flips using the ImageDataGenerator library from Keras. These preprocessing techniques prepared the dataset for model training and improved the model robustness by mimicking various scanning conditions [13].

As shown in Fig. 3, the dataset was segmented into validation, training, and testing with percentages of 15%, 75%, 10% respectively [14]. This distribution ensures the training set is sufficient for effectively training the models, while providing enough data for the validation and testing to assess the model's performance and reduce overfitting [15]. Dataset splitting is a common practice in machine learning for healthcare applications, as it ensures an adequate dataset for validation and testing. It helps models evaluate performance and make predictions on unseen images [16].

In this study, two AI models were created: binary and categorical classification models. The binary classification model was created using a custom CNN architecture. The model architecture includes multiple Conv2D layers increasing filter size of 32, 64 and 128. Each Conv2D layer is followed by a MaxPooling2D layer, which decreases spatial dimensions. After the convolutional and pooling operation, a Flattened layer is used to turn 3D feature maps into 1D feature vectors, which can then be processed by fully connected layers. To prevent overfitting during training, a dropout layer is applied after flattening the data. The final phase of the architecture involves fully connected Dense layers. Nonlinearity is incorporated into the model using a hidden Dense layer with Rectified Linear Unit (ReLU). The output layer creates a probability score between 0 and 1 using sigmoid activation [17]. Input chest X-ray images can potentially be classified into normal or abnormal utilizing this binary classification model architecture.

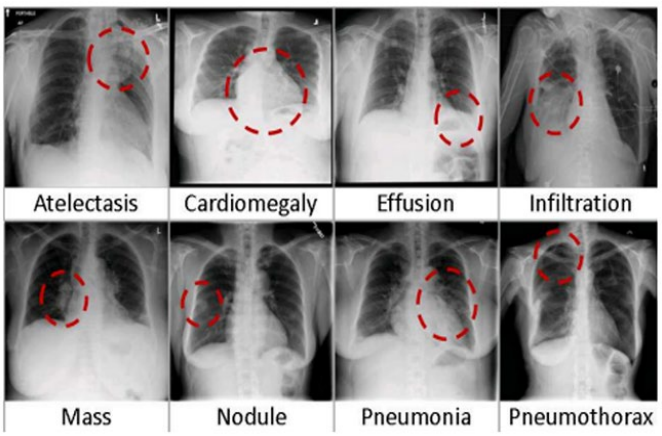

Fig. 1. Main pathological conditions of the dataset [7].

According to the block diagram of Fig. 4, a categorical classification model was developed encompassing VGG-16 pre-trained architecture to classify chest X-rays into 8 pathological conditions along with a normal condition. The VGG-16 model was employed as a feature extractor, which was pretrained on the ImageNet dataset. To capture hierarchical image features, it omitted fully connected layers and instead relied on deep convolutional layers.

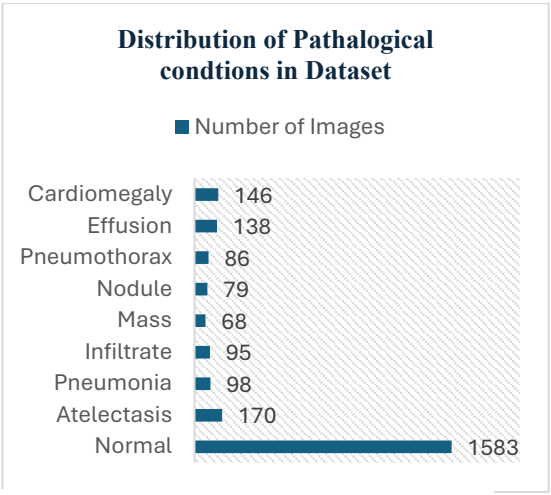

Fig.2. The distribution of Thoracic Condition in Dataset.

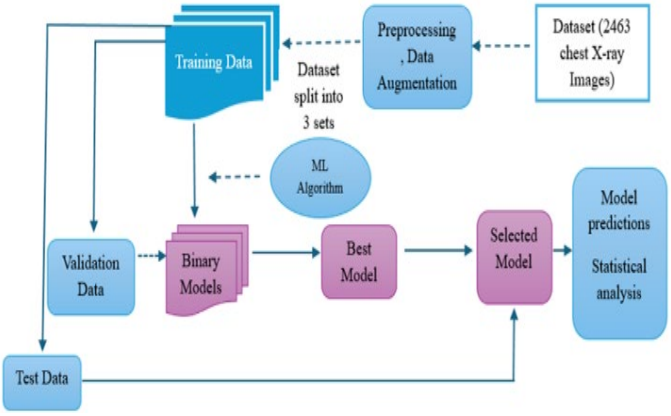

Fig. 3. Block diagram of binary classification model.

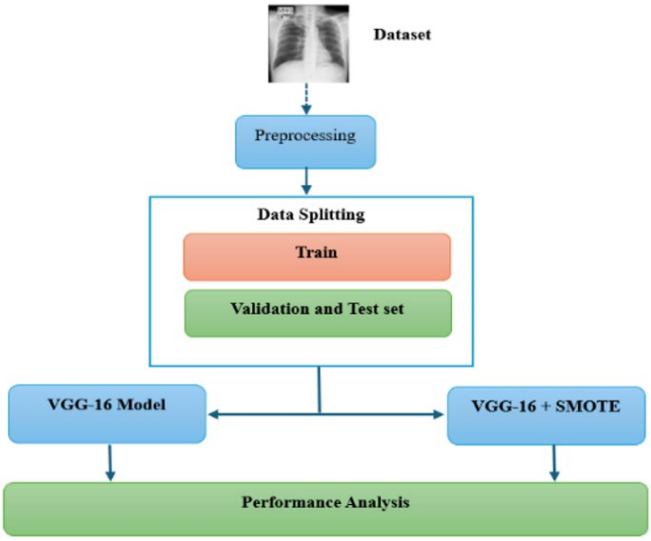

Fig.4. Block diagram of categorical classification model.

These Convolutional layers are grouped into five blocks, increasing filter sizes (32, 64, 128, 256 and 512). Blocks 1 and 2 contain two Conv2D layers with 64 and 128 kernels, while blocks 3,4, and 5 each contain three Conv2D layers with 256, 512 filters. Custom fully connected layers were added after feature extraction. This includes a Flatten layer, which is followed by a Dense layer consisting of 512 units and ReLU activation [18]. A 0.5 dropout rate was added to the Dropout layer to prevent overfitting. To classify chest X-rays into their respective pathological conditions, a final dense layer has been equipped with Softmax activation. In order to overcome issues of the sparse gradient on noisy problems, the model architecture was built using Adam optimizer and trained with categorical cross entropy loss.

The categorical classification model also trained using 224 x 224 pixels to further assess model's accuracy. The higher resolution enables the model to capture the finer details in radiographic images especially for subtle conditions that are difficult to distinguish at lower resolutions. Furthermore, SMOTE (Synthetic Minority Oversampling Technique) was utilized to balance the dataset by creating synthetic samples for under presented classes [19] including Pneumothorax, Mass, and Nodule. This strategy helped solve class imbalance, which is typical problem in medical imaging datasets and frequently limits the model's capability to generalize well.

Both AI models were trained using TensorFlow and Keras. The binary categorization model was trained with a batch size of 20 over 30 epochs. Similarly, the categorical classification model used the same framework and was trained with a 32-batch size over 10 epochs. An artificial data augmentation technique was utilized to enhance the dataset to meet real-world varied image conditions [20].

These metrics were utilized to evaluate model performances, including accuracy, precision, specificity, recall, F1 Score, ROC (Receiver Operating Characteristic), and AUC (Area Under Curve). Cross-validation techniques such as K-fold cross validation, were utilized to assess the model performance across multiple subsets of the dataset [20]. An independent test dataset was utilized to evaluate model performance on unseen chest X-ray images. The research was conducted ensuring ethical compliance throughout, and patient data privacy was protected in the process according to General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA) standards [22].

III. RESULTS AND FINDINGS

This section presents findings and results from analyzing two AI models that were developed to classify chest X-rays: binary and categorical. To evaluate their performances, we monitored real-time accuracy and loss throughout each epoch of training and validation phases. In addition, evaluated the models on previously undisclosed dataset and conducted a detailed statistical analysis of models' performances.

The main goal was to assess each model's capability to accurately identify an X-ray into a predetermined category. The binary classification model was supposed to classify between normal and abnormal categories, while the categorical model was to categorize a single chest x-ray image under one of eight pathological conditions or normal. To ensure consistency in performance evaluation both models were evaluated using the same methodology. We assessed each model's diagnostic capabilities and limitations using several evaluation statistics, which were mentioned in the methodology section. The results are discussed extensively in next subsections, which also discuss the impact of these findings for medical imaging technologies and include data visualization to demonstrate model effectiveness.

A. Binary Classification Model

This subsection presents findings of the binary categorization model, developed to differentiate chest X-rays into normal and abnormal. The assessment focuses on metrics for performance during the training, validation and testing phases, as well as the analysis of metrics for unseen chest X-ray images. The methods used were exactly those described in the methodology for training CNNs on the chest X-ray dataset. The binary model was trained under 30 training epochs, which showed significant improvement in accuracy as well as loss reduction on both phases of training and validation. As shown in Fig. 5 and Fig. 6, the model accuracy of training phase was 74.54% with a corresponding loss 0.5332. During the training phase, Accuracy gradually improved to 98.48% and loss was reduced to 0.0346 by the 30th epoch. Meanwhile, validation accuracy increased significantly, starting at 61.25% and reaching 99.31%, showing the model robust performances and predictive capabilities. Upon model prediction on the test set, the binary categorization model reached 99.42% accuracy and 1.85% loss. These results show the model's capability to accurately classify chest X-rays into normal and abnormal.

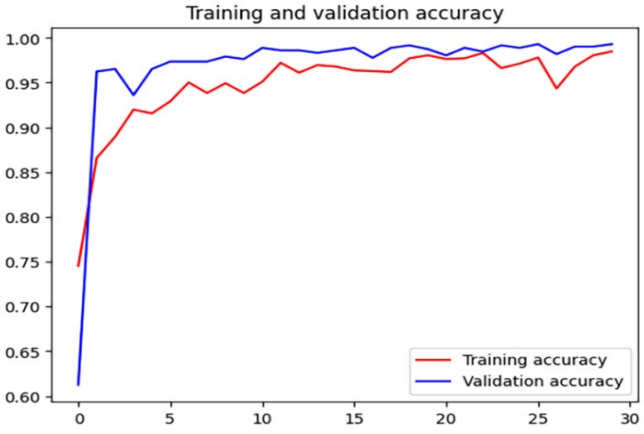

Fig.5. The validation and training accuracy of binary classification model. (X: Epochs, Y: Accuracy)

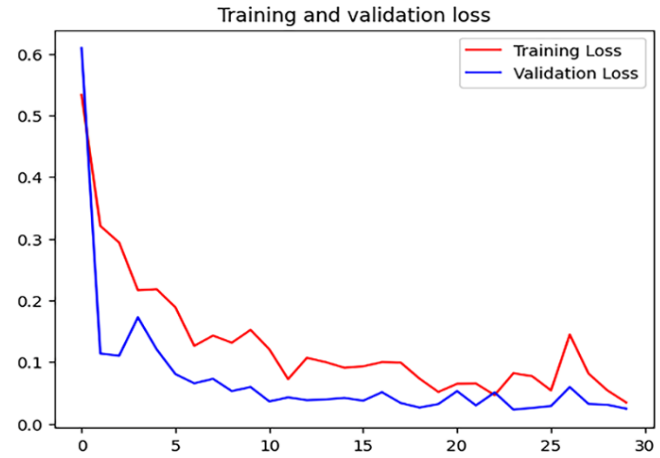

Fig. 6. Graphical representation of training and validation loss of binary classification model. (X: Epochs, Y: Loss)

In this research, a sample of 10 unseen chest x-ray image sample was used to evaluate model prediction and performance. This test was performed to evaluate model generalizability in real-world circumstances. The models' predictions (y_pred) were compared with true labels (y_true). The performance was statistically assessed using the confusion matrix as shown in Fig. 7, which gave statistical insights into the model's accuracy by counting TN (True Negatives), FP (False Positives), TP (True Positives), and FN (False Negatives). In addition, the model evaluation metrics described in the methodology section were derived from the confusion matrix. These metrics provided an in-depth understanding of the model's capability in classifying unseen data.

According to the matrix in Fig. 7, the following metrics were calculated. As shown in TABLE 1, binary classification models' practical applicability for unseen data achieved an accuracy rate of 90%. Fig. 10 illustrates, ROC curve which has an AUC of 0.96, indicating excellent prediction between normal (without any pathological condition) and abnormal (with pathological condition) cases. Furthermore, Cohen's kappa score of 0.80, indicating the amount of agreement among prediction and true classification, validates the model's accuracy for diagnosis.

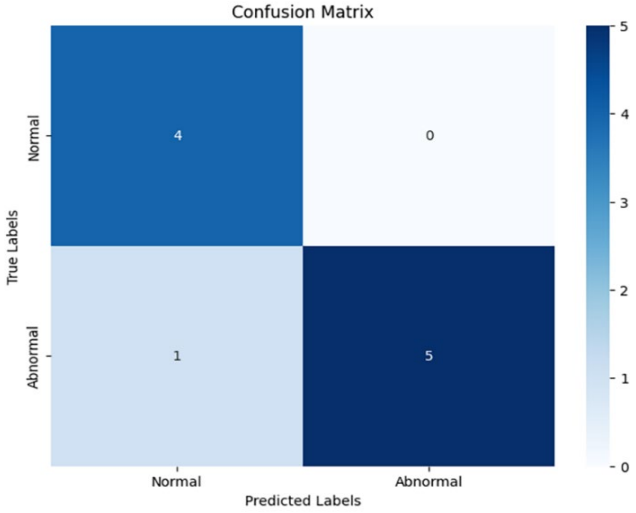

Fig. 7. Confusion matrix of the binary classification model.

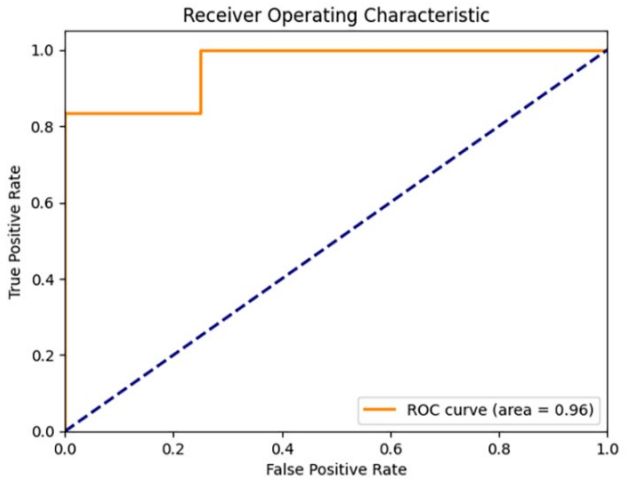

Fig. 8. ROC curve: binary classification.

TABLE I. CALCULATION OF EVALUATION METRICS FOR A BINARY CLASSIFICATION MODEL.

Metrics	Calculation
Accuracy	$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} = \frac{5+4}{(5+4+0+1)} = 90\%$
Precision	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{5}{5+0} = 1 = 100\%$
Recall	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{5}{5+1} = 83.33\%$
Specificity	$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{4}{4+0} = 100\%$
F1 Score	$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precesion} + \text{Recall}} = 2 \times \frac{1 \times 0.833}{1 + 0.833} \approx 90.9\%$

B.Categorical Classification Model

This section highlights the findings and results of the model developed to predict different clinical conditions and identify normal images from chest X-rays. Using VGG-16 pre-trained model architecture, the categorical classification model was trained and validated on a different group of chest X-rays, classifying them into nine categories.

As shown in Fig. 9 and Fig. 10, the model was trained over twenty epochs, beginning with starting accuracy of 61.36% and an initial loss of 2.773, indicating the challenge of differentiating between various categories. After the training phase, the model had an accuracy of 71.8% and loss of 0.747 for validation. This progressive increase in accuracy and progressive decrease in loss throughout epochs indicates its ability to train and respond to subtle differences in the dataset. The model obtained 69.3% accuracy on the test dataset, which included 518 images from 8 pathological conditions and normal images. As shown in Table 3, model prediction was evaluated using a comprehensive statistical approach to determine the F1 score, precision, recall, and sensitivity, in addition a confusion matrix and ROC curve analysis for each pathological condition classes.

As shown in Fig. 11, the confusion matrix provides inconsistent results, with accurate predictions in certain pathologies and certain misclassification in other pathological conditions. As an example, 'Atelectasis' was incorrectly classified as normal while 'pneumothorax' was misidentified as 'Effusion', indicating a probable overlap in radiological features or inadequate training data for the model for these pathological conditions. Fig. 12 illustrates that the ROC curves for each class showed different outcomes. 'Infiltrate' and 'Cardiomegaly' achieved exceptional AUC values of 1.00, showing excellent classification. However, classification such as 'pneumothorax' and 'effusion' showed lower AUC values (0.50 and 0.44 respectively) reflecting model's limitation in these classes.

The categorical classification model showed significantly better results after applying SMOTE to correct class imbalance and increasing the image resolution to 224 x 224. After 20 epochs, the training accuracy had improved to 73.25% with a validation accuracy of 71.18%. Training loss and validation loss were 0.7489 and 0.7420, respectively. This modification enhanced the model's capability to learn more from the same dataset, increasing its capability to classify different pathological conditions. In addition, the model improved its test accuracy to 70.84%.

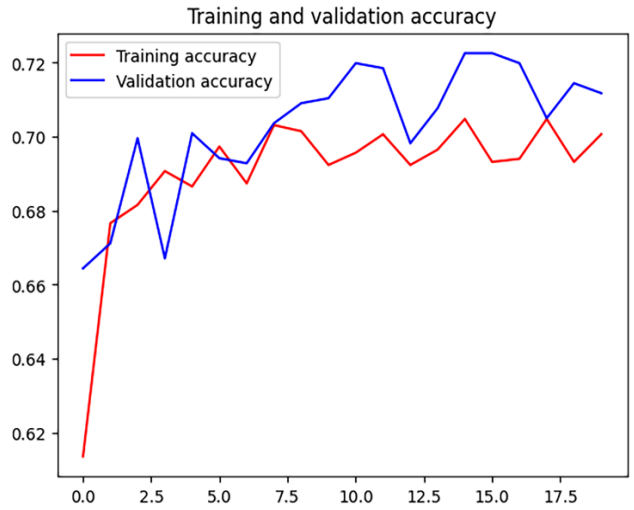

Fig 9. The validation and training accuracy of categorical classification model. (X: Epochs, Y: Accuracy).

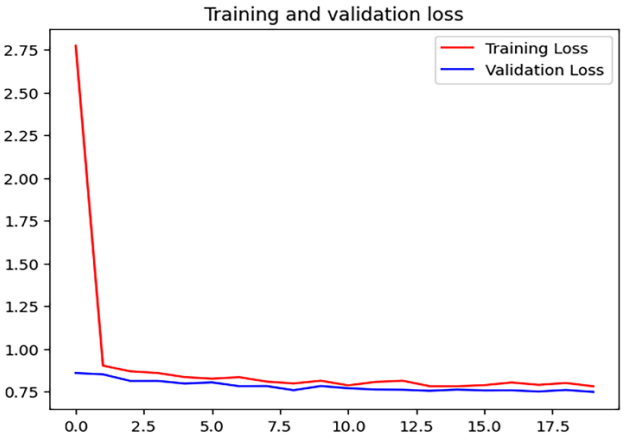

Fig 10. The validation and training loss of categorical classification model. (X: Epochs, Y-a: Loss).

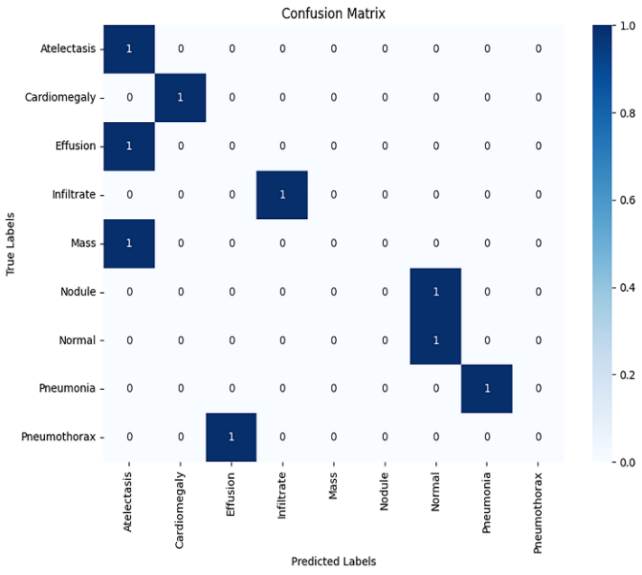

Fig. 11. Confusion matrix of the categorical classification model.

TABLE II. CALCULATION OF EVALUATION METRICS FOR A CATEGORICAL CLASSIFICATION MODEL

Metric	Value
Accuracy	55.56%
Precision	43%
Sensitivity	56%
F1 Score	46%

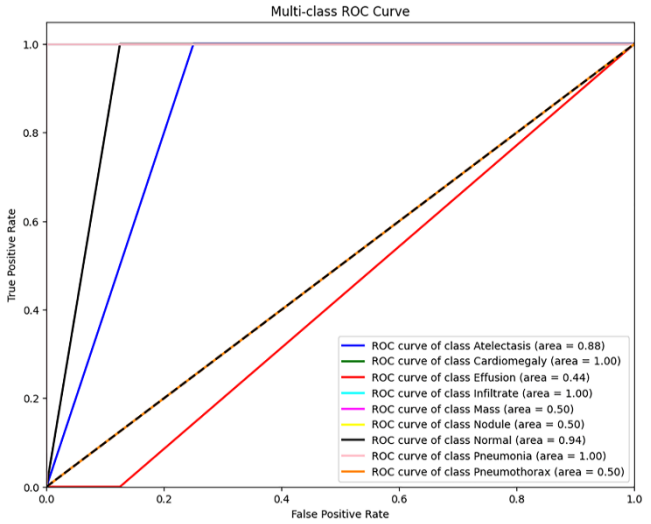

Fig. 12. ROC curve of categorical classification model.

IV. Discussion

The aim of this research was to apply cutting-edge CNN architecture to enhance the categorization of chest X-rays into distinct pathological conditions to enhance diagnostic accuracy and precision in medical imaging. This study demonstrates how to employ both binary and categorical classification model to maximize detection of normal and abnormal data while accurately predicting specific pulmonary conditions.

The binary classification model achieved a high level of performance classifying normal and abnormal chest X-rays, archiving an accuracy of 98.48% in the training phase and 99.31% in the validation phase after 30 epochs. It achieved a test accuracy of 99.42%. When evaluated with unseen data, key metrics were calculated as follows: precision of 100%, recall of 83.33% and F1 score of approximately 90%, thus showing strong diagnostic capabilities. The VGG-16 architecture-based categorical classification model, evaluation metrics were calculated at 61.36% yielding a validation accuracy of 71.18%, and test accuracy of 69.31%. Even though it was effective in certain classes such as ‘cardiomegaly’ and ‘infiltrate’, there was variability in others providing some improvement in differentiation between closely related pathologies.

The findings highlight some of the possibilities and limitations of existing AI techniques for medical image analysis. The binary classification model with high accuracy and minimal loss suggests that it is suitable for clinical use. The categorical classification model, while useful in some areas, needs extensive refinement particularly in the situation of overlapping conditions. The complexities of multi-class categorization and potential dataset imbalances contribute to these challenges.

Key strengths of this research include diverse data collection, advanced CNN architecture and sound validation techniques followed to enhance reliability and model robustness. However associated challenges such as dataset representativeness, overfitting and computational demand remain. To address the data imbalance challenge, techniques like SMOTE were used which considerably improved classification performance for underrepresented classes like pneumothorax, mass and nodule. This illustrates the need to correct class imbalance in medical imaging datasets to minimize biases in model predictions. Even though the dataset size of 2463 images are one of major limitation, since it could not have been enough for high performance multi class classification. The limited dataset makes overfitting even if we used techniques like data augmentation, SMOTE, dropout and early stopping. To increase model generalizability future research should collect larger and more balanced dataset.

While the findings indicate that AI models have potential to transform clinical practice improving the speed and accuracy of diagnostics, thus reducing radiologists' workload and allowing for early detection of pathological conditions. For effective clinical integration, it is essential to provide smooth integration with current IT systems such as PACs (Picture Archiving & Communication System) as well as maintain trust among users through transparent and interpretable AI models. Visual interpretation of model's decision-making process could potentially be provided by employing methods like GRAD-CAM (Gradient Weighted Class Activation Mapping). It allows physicians to get a better understand how AI model comes to its findings. This transparency is critical for health care providers to establish confidence in AI- assisted diagnosis.

Furthermore, addressing ethical issues such as bias, fairness and transparency in the AI model is essential for ensuring inclusive healthcare outcomes for all patients. Despite the dataset of this research is diverse, it may not be representing every demographic group. This gap might raise improper biases, and it leads to reducing the validity of clinical applications. Following that, the primary focus should be on developing datasets that are more varied, inclusive and properly represent all communities. The incorporation of fairness criteria will also help to examine how the AI models generalize across diverse subgroups, ensuring that the AI system's predictions are accurate and useful for all patients, regardless of their background.

Future research should be conducted on increasing the diversifying training dataset, advancing machine learning algorithms, and improving validation processes. Other opportunities for future advancements are the extending AI application to other diagnostic medical images such as US (Ultrasound Scanner) images, CT (Computed Tomography images and MRI (Magnetic Resonance Imaging) images. By addressing the limitations of current study design and exploring new methods, future research on AI integration in healthcare settings will be enhanced significantly. It leads to better patient outcomes and more effective health care delivery.

V. CONCLUSION

The development models for the processing of chest X-ray images provided a significant step in medical diagnostic integration. The binary classification model classified chest X-ray images highly accurately and robustly distinguished normal from abnormal X-rays. While the categorical classification model predicted pre-specified pulmonary conditions such as 'cardiomegaly' and 'infiltration'. The CNNs demonstrated high diagnostic accuracy and adaptability with new datasets, suggesting possible wide application to a variety of clinical contexts. Hence AI serves as a supplementary tool for radiologists, reducing workload, speeding up diagnosis process, while providing more accurate results for patient.

When integration AI into clinical practice, hurdles include system compatibility, ethical consideration and data protection, ethical consideration and cultural adaptation within the medical profession. These are the areas where future research must focus. To increase the acceptance of AI applications across all imaging modalities, we must improve the variety of training datasets and better model interpretability and transparency.

With advancement in machine learning together with computational hardware, the AI models will become more complex and take on much more complex tasks, transforming medical imaging techniques. The broader adoption of AI in clinical settings has a promising future in relation to the delivery of efficient, personalized and patient centered healthcare. Addressing the ethical, technological and practical challenges that exist is essential for successful integration.

Furthermore, this study adds certain insights into the role of AI in medical image interpretation and sets some grounds for further innovation and studies. The combination of ethical and practical considerations will improve diagnoses and treatment in medicine, resulting in improved health outcomes and more efficient health care delivery.

VI. RECOMMENDATIONS

Several strategic enhancements are recommended to improve the future research and application of AI algorithms in diagnostic medical imaging, especially for chest X-rays. An expanded dataset should include a large variety of images that represent different pathological conditions. This is necessary for comprehensive learning and preventing bias, which may impact diagnostic accuracy [23]. Enhancing the quality of image annotation by localizing the affected area will help the model recognize associate features and improve diagnostic precision.

Incorporating full radiological reports for each image will provide invaluable contextual information, boosting training data and helping AI models to better align with the human diagnostic process [24]. Utilizing GRAD-CAM would assist in understanding how AI models read images by highlighting specific regions that affect prediction and helping model evaluation [25]. Improving partnership with medical experts guarantees that AI tools are clinically relevant and effectively integrated into medical work flows. High-end computational resources are critical for managing the complexity of AI structures of AI and the computation intensity of training procedures, enabling robust model training and exploration of advanced AI architectures.

Lastly, there must be a considerable focus on ethical implications and regulatory standards. Developing clear guidelines regarding the moral implementation of AI in medical settings. Ethical use of AI in healthcare, as well as the strict protection of patient data. Implementing these recommendations will enhance scope, accuracy, and clinical

significance of AI in medical imaging, resulting in better patient care and health care efficiency.

REFERENCES

[1] H. Zhang and Y. Qie, "Applying deep learning to medical imaging: A Review," Applied Sciences, vol. 13, no. 18, p. 10521, Jan. 2023, doi: 10.3390/app131810521.

[2] M. S. Alyahya, H. H. Hijazi, M. N. Alolayyan, F. J. Ajayneh, Y. S. Khader, and N. A. Al-Sheyab, "The association between cognitive medical errors and their contributing organizational and individual factors," Risk Management and Healthcare Policy, vol. Volume 14, no. 14, pp. 415–430, Feb. 2021, doi: 10.2147/rmhp.s293110.

[3] M. Khalifa and M. Albadawy, "AI in diagnostic imaging: revolutionizing accuracy and efficiency," Computer Methods and Programs in Biomedicine Update, vol. 5, pp. 100146–100146, 2024, doi: 10.1016/j.cmpbup.2024.100146.

[4] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical image analysis using deep learning algorithms," *Frontiers in Public Health*, vol. 11, no. 1273253, Nov. 2023, doi: 10.3389/fpubh.2023.1273253.

[5] L. Pinto-Coelho, "How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications," Bioengineering, vol. 10, no. 12, p. 1435, Dec. 2023, doi: doi.org/10.3390/bioengineering10121435.

[6] C. Hsieh, "Human-centered multimodal deep learning models for chest X-Ray diagnosis." Accessed: Sep. 28, 2024. [Online]. Available: https://www.ijcai.org/proceedings/2023/0817.pdf

[7] S. Mascarenhas and M. Agarwal, "A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification," IEEE Xplore, Nov. 01, 2021. https://ieeexplore.ieee.org/document/9687944

[8] A. E. W. Johnson et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," Scientific Data, vol. 6, no. 1, p. 317, Dec. 2019, doi: 10.1038/s41597-019-0322-0.

[9] 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," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, doi: 10.1109/cvpr.2017.369.

[10] D. Ueda et al., "Fairness of artificial intelligence in healthcare: Review and recommendations," Japanese Journal of Radiology, vol. 42, no. 1, pp. 3–15, Aug. 2023. doi:10.1007/s11604-023-01474-3

[11] Y. Yang, K. Sun, Y. Gao, K. Wang, and G. Yu, "Preparing data for artificial intelligence in pathology with clinical-grade performance," Diagnostics, vol. 13, no. 19, p. 3115, Jan. 2023, doi: 10.3390/diagnostics13193115.

[12] Kieu, S.T. et al. (2021) 'Covid-19 detection using integration of deep learning classifiers and contrast-enhanced canny edge detected X-ray images', IT Professional, 23(4), pp. 51–56. doi:10.1109/mitp.2021.3052205.

[13] A. Ait Nasser and M. A. Akhloufi, "A review of recent advances in deep learning models for chest disease detection using radiography," Diagnostics, vol. 13, no. 1, p. 159, Jan. 2023, doi: doi.org/10.3390/diagnostics13010159.

[14] Yadaw, A.S. et al. (2020) 'Clinical features of COVID-19 mortality: Development and validation of a clinical prediction model', The Lancet Digital Health, 2(10). doi:10.1016/s2589-7500(20)30217-x.

[15] H. Liz, J. Huertas-Tato, M. Sánchez-Montañés, J. Del Ser, and D. Camacho, "Deep learning for understanding multilabel imbalanced Chest X-ray datasets," Future Generation Computer Systems, Mar. 2023, doi: 10.1016/j.future.2023.03.005.

[16] B. Koçak, R. Cuocolo, D. P. dos Santos, A. Stanzione, and L. Ugga, "Must-have qualities of clinical research on artificial intelligence and machine learning," Balkan Medical Journal, vol. 40, no. 1, pp. 3–12, Jan. 2023, doi: https: 10.4274/balkanmedj.galenos.2022.2022-11-51.

[17] M. Soric, D. Pongrac, and I. Inza, "Using convolutional neural network for chest X-ray image classification," 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Sep. 2020. doi:10.23919/mipro48935.2020.9245376

[18] A. H. Md. Linkon, Md. M. Labib, T. Hasan, M. Hossain, and M.-E. -Jannat, "Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: An extensive study," Informatics in Medicine Unlocked, vol. 24, p. 100582, 2021, doi: doi.org/10.1016/j.imu.2021.100582.

[19] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance," Information Sciences, vol. 505, pp. 32–64, Dec. 2019, doi: 10.1016/j.ins.2019.07.070.

[20] M. Yaqub et al., "State-of-the-art cnn optimizer for brain tumor segmentation in magnetic resonance images," Brain Sciences, vol. 10, no. 7, p. 427, Jul. 2020, doi: 10.3390/brainsci10070427.

[21] Airola, A. *et al.* (2011) 'An experimental comparison of cross-validation techniques for estimating the area under the ROC curve', computational statistics & data analysis, 55(4), pp. 1828–1844. doi:10.1016/j.csda.2010.11.018.

[22] Said, A., Yahyaoui, A. and Abdellatif, T. (2024) 'HIPAA and GDPR compliance in IOT healthcare systems', Advances in Model and Data Engineering in the Digitalization Era, pp. 198–209. doi:10.1007/978-3-031-55729-3_16.

[23] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, "Addressing bias in big data and AI for health care: A call for open science," *Patterns*, vol. 2, no. 10, p. 100347, Oct. 2021, doi: 10.1016/j.patter.2021.100347.

[24] Agarwal, N. et al. (2023) Combining human expertise with artificial intelligence: Experimental evidence from radiology [Preprint]. doi:10.3386/w31422.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," International Journal of Computer Vision, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.