

Domain-specific Explainable AI Solutions for Healthcare

Ravisankar PRIYADHARSINI¹, Guru Subbiah DEVI LAKSHMI²,
Jayavarthanaraj Soundharavalli MALAVIKA³

¹ Department of Artificial Intelligence and Data Science, Rajalakshmi Engineering College,
Thandalam, Chennai, 602105, India
priyadharsini.r@rajalakshmi.edu.in (*Corresponding author)

² Department of Information Technology, Easwari Engineering College,
Ramapuram, Chennai, 600089, India
devilakshmi.gs@eec.srmmp.edu.in

³ Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering,
Kalavakkam, Chennai, 603110, India
malavika572000@gmail.com

Abstract: Explainable AI (XAI) has emerged as a pivotal solution for addressing different concerns such as lack of transparency, interpretability, trust, accountability, and clinical decision-making, particularly in the healthcare sector. XAI strives to bridge the gap between AI's high-performance predictive capabilities and the need for understandable, trustworthy, and accountable AI systems. The dataset selected for this study includes chest X-ray images of patients affected by Covid-19 along with the corresponding patient metadata. The primary goal of constructing the proposed system is twofold: firstly, to determine whether an individual is affected by Covid-19 or pneumonia and secondly, to provide accountability for the AI models responsible for such classification decisions. The Deep Learning architectures, namely CNN+LSTM, EfficientNetB0, and ResNet18 are employed for classification tasks and the CNN+LSTM model outperforms the other two models, achieving a training accuracy of 99.42% and a testing accuracy of 83.66%. The preprocessed metadata and image-derived features were provided⁵ as inputs to the XGBoost and Random Forest classifiers. XGBoost and Random Forest achieved a validation accuracy of 98.69%, while their training accuracy reached 99.26% and 99.67%, respectively. The explainability of the deep learning models is visually illustrated by using the XAI technique Grad-CAM. Similarly, the machine learning models are interpreted by using two XAI techniques: SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). This comprehensive approach combines accurate prediction with an improved transparency, enabling the stakeholders to better understand the underlying decision-making process through multimodal data fusion.

Keywords: EfficientNetB0, Explainable AI, CNN+LSTM, Grad-CAM, LIME, Random Forest Classifier, ResNet18, SHAP, XGBoost Classifier.

1. Introduction

Artificial Intelligence (AI) is becoming increasingly important in real-time scenarios due to its ability to analyze large datasets, automate complex tasks, and make decisions based on data history. Explainable AI (XAI) aims to address these issues by providing transparent and accountable AI systems. XAI techniques can be categorized into model-based, post-hoc, model-specific, or model-agnostic techniques. XAI methods like Grad-CAM, LIME, and SHAP are employed to explain AI decisions, enhance trust, and ensure regulatory compliance. (Holzinger et al., 2022)

XAI applications in finance include explaining credit scores, loan approval decisions, fraud detection, transparency in algorithmic trading systems, risk assessment models, and automated responses in customer service chatbots. (Adams & Hagra, 2020; Ariza-Garzon et al., 2020) XAI can identify biases in educational AI systems, preventing unfair treatment and opportunities. XAI also boosts parental engagement by providing transparent insights into student

progress and explains the logic behind tutoring selections. (Rachha & Seyam, 2023; Khosravi et al., 2022). XAI is also used in clinical trial design and radiology to explain image-based diagnostic decisions. (Gupta et al., 2021; van der Velden et al., 2022; Onari et al., (2022). Similarly, XAI has an impact in various fields like manufacturing and quality control, marketing and advertising, human resources, cyber security and more.

Explainable AI is a promising solution for healthcare, addressing the "black box" problem associated with complex AI models. It helps identify responsible factors when AI makes errors, fostering trust and reducing the likelihood of erroneous decisions. However, implementing Explainable AI faces challenges such as striking a balance between model complexity and interpretability, integrating AI into the existing healthcare infrastructure, and ensuring healthcare professionals can easily understand and trust the provided explanations. This work aims to develop an effective and robust Explainable AI model tailored specifically for the healthcare

domain, enhancing the interpretability and trustworthiness of AI-generated outcomes. This paper also addresses unique challenges related to healthcare data, such as dealing with sensitive patient information and diverse, multidimensional medical data. The goal is to empower healthcare stakeholders to make informed decisions based on AI-driven recommendations, advancing healthcare practices and patient well-being.

The proposed framework is assessed using performance metrics such as the accuracy, precision, recall, F1-score, and confusion matrix analysis. A comparative evaluation is conducted between Deep Learning models, including CNN+LSTM, EfficientNetB0, and ResNet18, and Machine Learning approaches such as Random Forest and XGBoost. Furthermore, the study examines the contribution of multimodal fusion and evaluates model interpretability through the Grad-CAM, SHAP, and LIME explanations.

The main contributions of the proposed model are as follows:

- Developing a domain-specific Explainable AI (XAI) framework for the identification of COVID-19 and pneumonia cases using chest X-ray images and associated patient metadata.
- Designing and evaluating deep learning architectures (CNN+LSTM, EfficientNetB0, and ResNet18) for an accurate disease identification.
- Extracting and integrating image-based features with preprocessed metadata through a multimodal fusion approach.
- Implementing and comparing machine learning classifiers (XGBoost and Random Forest) on the fused feature set to enhance the predictive performance.
- Providing model interpretability and accountability using XAI techniques, including Grad-CAM for deep learning models and SHAP and LIME for machine learning models.
- Achieving a balance between a high classification accuracy and transparency, thereby supporting trustworthy decision-making in healthcare applications.

The remainder of this research paper is organized as follows. Section 2 presents the related work

and discusses the existing studies relevant to the proposed approach. Section 3 describes the proposed system and its overall framework. Further on, Section 4 presents the image classification process using deep learning models, while Section 5 describes the classification of metadata using machine learning models. Section 6 discusses the development of the Explainable AI (XAI) models employed for interpreting the predictions. Section 7 presents the experimental results and performance evaluation. Finally, Section 8 concludes the paper by summarizing the key findings and outlining possible directions for future research.

2. Related Work

The need for understanding the decisions made by deep neural networks (DNNs) is emphasized by proposing a framework for analysing explainable deep learning methods based on three key dimensions. The framework also covers the methods for evaluating the explanations and designing them for users. This comprehensive analysis serves as a valuable resource for researchers in the field of explainable AI (Ras et al., 2022). There are 17 XAI methods, including LIME, Anchors, GraphLIME, and SHAP. The evolving nature of XAI, future methods, and the importance of measuring explanation quality are emphasized and the potential for incorporating human experience and conceptual knowledge is highlighted (Holzinger et al., 2022). The significance of trustworthy Artificial Intelligence (AI) and Explainable AI (XAI) is examined by focusing on their current state, definitions, methods, and concerns. A framework is proposed for deploying XAI systems, which emphasizes the collaboration across fields, and suggests ways to make intelligent systems trustworthy and ethical (Ali et al., 2023). A model is proposed for explainable AI, a multidisciplinary research area, to understand the stakeholders' perspectives and develop interdisciplinary collaboration to understand the interplay between explainability approaches, stakeholders' perspectives, and context (Langer et al., 2021).

The potential of combining various types of health data improves smart care by the process of turning data into knowledge, including health records, medical images, DNA facts, sensor

details, and behavioural information. Multimodal data is handled by using features, rule systems, machine learning, deep learning, and natural language processing tools. The need for good data, system integration, patient privacy, and moral considerations is a major concern (Shaik et al., 2023). A deep learning system for COVID-19 detection is built using AI. It fine-tunes three pre-trained models, namely VGG16, ResNet50, and InceptionV3, on an enriched dataset. The results show a high accuracy, with InceptionV3 leading at 98.10% and VGG16 at 98.30%. Merging the models is said to improve the findings (Guefrechi et al., 2021). A robust classification model is built for chest X-ray analysis during the COVID-19 pandemic. The model categorizes chest X-rays into normal, pneumonia, and tuberculosis, with additional severity classifications for COVID-19 cases. The model reaches various accuracy rates, with VGG-16 achieving 95.9% for pneumonia, DenseNet-161 achieving 98.9% for pneumonia, and ResNet-18 achieving up to 76% for severe COVID-19 X-rays. The importance of deep learning for medical diagnostics, particularly in mass screening for COVID-19 is emphasized (Shelke et al., 2021).

XAI aims to bridge this gap by explaining internal decisions to healthcare professionals. Its successful integration requires human involvement and collaboration among stakeholders (Yang, 2022). Heart rate variability (HRV) in individuals from Delhi, India, who recovered from COVID-19 is analysed using seven machine learning models. The XGBoost model outperformed the others, indicating a need for regular follow-ups and early interventions. An interpretable AI model was introduced for home monitoring, promoting the use of wearable technologies. This was a pioneering effort in analyzing cardiac status for individuals who recovered from COVID (Gupta et al., 2021). The importance of interpretable machine learning (ML) models in healthcare, especially during the COVID-19 pandemic is highlighted by four models - logistic regression, decision tree, pyFUME, and RIPPER – which were used for classifying COVID-19 patients based on blood samples. RIPPER showed the best accuracy at 97.78%. The interpretability was made possible through SHAP explanations and PyFUME's fuzzy logic capabilities (Onari et al., 2022). Explainable AI's potential in medical

image analysis is discussed by highlighting its shift from handcrafted features to learned features through neural networks and potential biases (van der Velden et al., 2022).

Gradient and perturbation-based XAI techniques for explaining deep neural network decisions on multi-modal medical images are crucial for an effective clinical decision-making. Gradient-based methods, like Guided BackProp and DeepLift, leverage the gradient signal to estimate the importance of features in predicting model outcomes. On the other hand, perturbation-based methods, such as occlusion, LIME, and kernel SHAP, utilize input-output sampling pairs to estimate feature importance (Jin et al., 2023). A framework for automatic text-based diagnosis in clinical settings is proposed, focusing on balancing accuracy and interpretability. It integrates Bayesian Network Ensembles with Entity-Aware Convolutional Neural Networks (Tan, M. & Le, Q. V. 2019) improving AI-powered healthcare. The framework uses real Electronic Medical Record (EMR) documents for a better accuracy and reasonable diagnosis explanations. The results show the framework's effectiveness in maintaining accuracy and enhancing interpretability (Chen et al., 2020). The significance of customizing AI application explanations to cater to diverse stakeholders' needs is achieved by simplifying ANN inference using CART, and integrating XAI tools for transparency, comprehensibility, and interpretability (Wang, Toly Chen & Chiu, 2023).

The Case-Based Reasoning (CBR) method, which combines quantitative and qualitative techniques, has been applied to breast cancer management. It demonstrated a classification accuracy comparable to that of k-Nearest Neighbor algorithms but with an enhanced explainability, allowing medical experts to understand case similarities through shared patient characteristics (Lamy et al., 2019). The existing XAI models in healthcare face significant challenges, including the black-box nature of deep learning models, which limits their interpretability and weakens trust in critical diagnostic scenarios. Additionally, many XAI techniques struggle to integrate multimodal healthcare data, such as medical images and metadata, while providing clinically relevant, actionable explanations tailored to diverse stakeholders, highlighting the need for a more

comprehensive and transparent approach. The recent XAI approaches utilized in the medical imaging field are compared in Table 1.

3. Proposed System

This research work aims to present a system for providing explainability to the AI models that was developed especially for the healthcare domain. A Covid19 chest-X-ray dataset is used for developing this system. This dataset also includes the corresponding metadata for all the images. The main objective for developing this model is to classify whether a person has COVID-19 or pneumonia and to provide accountability for the AI models that make these classification decisions by giving explanations for each stakeholder such as clinicians, patients, and data scientists to understand the underlying decision-making process.

The original dataset consisted of 930 chest X-ray images accompanied by a metadata file containing 951 records and 29 attributes. During the data preparation phase, inconsistencies such as missing records and incomplete metadata entries were identified and addressed. After data cleaning, the images were organized into two diagnostic categories: COVID-19 and pneumonia, comprising 559 and 233 images, respectively.

To enhance the effectiveness of model training and improve generalization, extensive preprocessing

was performed on all images. This included image resizing, normalization, and noise reduction. Subsequently, data augmentation techniques such as rotation, horizontal flipping, and scaling were applied to generate additional training samples and mitigate class imbalance. As a result, the dataset was expanded to a total of 2,700 chest X-ray images, including 1,200 COVID-19 cases and 1,500 pneumonia cases.

The final processed dataset was then partitioned using a stratified train-test split to maintain the original class distribution. Accordingly, 80% of the images (2,160 samples) were allocated for training, which included 960 COVID-19 and 1,200 pneumonia images, while the remaining 20% (540 samples) were reserved for testing, including 240 COVID-19 and 300 pneumonia images. This balanced distribution ensured a reliable model training and an unbiased performance evaluation.

As shown in Figure 1, the system uses the predefined Covid19 chest-X-ray dataset [21] that comprises medical images, accompanied by the corresponding metadata, utilized for Image Augmentation. This process serves the dual purpose of introducing diversity into the training data and ensuring dataset balance. The resultant dataset, comprising both original and augmented images, forms the basis for constructing the DL model. Three DL models are built, namely ResNet18, EfficientNetB0 and CNN+LSTM. Subsequently, a XAI model - GradCAM is developed for the

Table 1. Comparison of Recent XAI Approaches in COVID-19 Medical Imaging

Citation	Methodology	Techniques Used	Key Contribution	Limitations
Rajpoot et al. (2024)	Ensemble deep learning for COVID-19 detection using CT and X-ray images	CNN ensemble + XAI methods	Improves diagnostic accuracy and integrates explainability for clinical trust	Limited generalization; no multimodal fusion
Soares et al. (2024)	Explainable deep learning framework for COVID-19 detection using CT scans	Deep neural networks + interpretability methods	Focuses on explainable diagnosis using CT imaging	Restricted to CT modality; lacks metadata integration
Chatterjee et al. (2024)	Comparative study of XAI techniques for COVID-19 classification	ResNet, DenseNet + Grad-CAM, saliency maps	Evaluates multiple interpretability techniques	Lacks a hybrid ML-DL framework and multimodal data fusion
Sun et al. (2023)	Evaluation framework for XAI methods using perturbation analysis	CNN + perturbation-based XAI validation	Provides reliability analysis of explanation methods	Focuses on evaluation rather than classification performance
Zhang et al. (2024)	Hybrid capsule network with attention mechanism for COVID-19 detection	Capsule Network + CNN-attention	Enhances feature representation and classification accuracy	Limited explainability; no use of multiple XAI techniques

previously constructed DL model. Simultaneously, the metadata, which consists of augmented image metadata, undergoes preprocessing. This metadata, in conjunction with features extracted from the images, is employed to build a ML model. The ML models built are XGBoost and Random Forest. A corresponding XAI model is then established for the previously constructed ML models, by using SHAP and LIME.

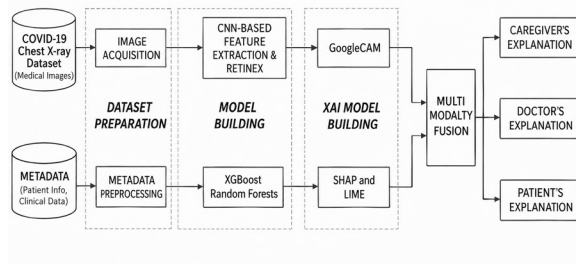


Figure 1. System Design

In a combined effort to enhance interpretability, the stakeholders are not only presented with the prediction results but they are also provided with insights into the decision-making process. The workflow is defined on two paths, one related to where the image dataset is pre-processed i.e. augmented, and fed as input to the DL models for image classification and the other to where the metadata of those images is pre-processed and fed as input to the ML models for metadata classification. Then the workflow proceeds with XAI model building for both the DL and ML models to provide explanations for various stakeholders.

To ensure a reliable evaluation and prevent data leakage, strict patient-level separation was maintained throughout the dataset preparation process. Images and metadata belonging to the same patient were included exclusively in either the training set or the testing set, but never in both. This safeguard prevented the models from learning patient-specific patterns that could artificially inflate the performance metrics. Furthermore, preprocessing, feature extraction, and multimodal fusion procedures were performed independently after dataset partitioning to avoid an unintended information transfer between datasets. The proposed framework was evaluated only on unseen patient data samples to ensure robustness, fairness, and the generalizability of the classification results.

For the training set, a 5-fold cross-validation strategy was employed to ensure robustness and reduce overfitting. Model performance was evaluated using standard metrics including

accuracy, precision, recall, and the F1-score, averaged across the validation folds.

The processing pipeline follows a sequential order:

1. **Preprocessing:** Raw chest X-ray images are resized to a uniform resolution (224×224), normalized, and denoised to ensure consistency.
2. **Data Augmentation:** Augmentation techniques such as rotation, horizontal flipping, and scaling are applied exclusively to the training data to increase diversity and address class imbalance.
3. **Feature Extraction:** Preprocessed and augmented images are passed through deep learning models (CNN+LSTM, EfficientNetB0, and ResNet18) to extract high-level feature representations.
4. **Multimodal Fusion:** Extracted image features are combined with preprocessed metadata (e.g. age, gender, clinical attributes).
5. **Classification:** The fused feature set is used as input for machine learning classifiers (XGBoost and Random Forest) for the final prediction.

To improve the reliability of the proposed framework, the experiments were conducted through multiple runs using fixed random seed initialization. Cross-validation was additionally applied to minimize the bias caused by a single train-test split. The final results were obtained by considering the average performance across repeated experiments. A comparative statistical analysis was carried out among the Deep Learning and Machine Learning models using metrics such as accuracy, precision, recall, and the F1-score. These measures helped ensure the consistency, reproducibility, and generalizability of the developed system.

The novelty of this approach lies in the integration of deep learning (DL) and machine learning (ML) models with explainability techniques to address the critical challenges in healthcare. By combining Grad-CAM for DL models with SHAP and LIME for ML models, this system ensures a robust performance for both image and metadata classification and also provides a comprehensive framework for explainability purposes. This dual-path approach uniquely bridges the gap between data-driven predictions and stakeholder trust. Such an integration is particularly transformative in healthcare, where transparency and accountability are essential.

4. Image Classification Using DL Models

The image dataset along with the corresponding metadata was provided to the sequential augmentor for augmentation in order to balance it. The transformations were implemented using a sequential function from the image library. Both the COVID-19 and pneumonia-related images along with the respective metadata were given as input to the first level of augmentation where they underwent horizontal flipping and rotation transformations. The result obtained for the first-level augmentation combined with the original dataset is given as input for the second-level augmentation where it undergoes vertical flipping, shear, and scaling transformations.

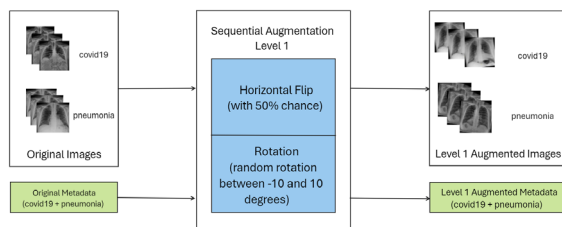


Figure 2. Level 1 Augmentation

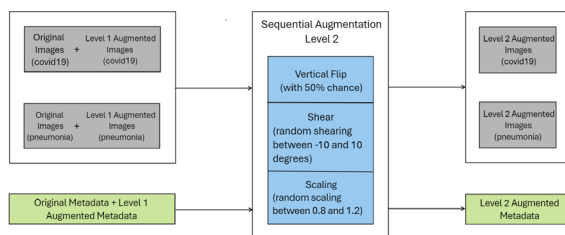


Figure 3. Level 2 Augmentation

The augmented images along with the original images are used for training the DL models – ResNet18, EfficientNetB0 and CNN+LSTM as shown in Figure 4.

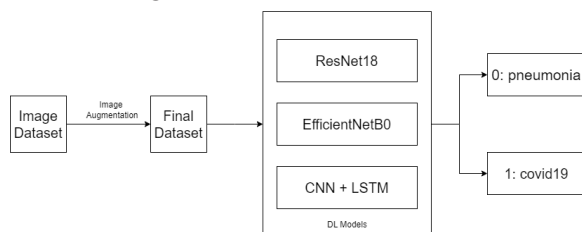


Figure 4. DL Models for Image Classification

The ResNet-18 architecture, a variant of the ResNet family, is used for training deep networks. Images from two classes, ‘COVID-19’ and ‘pneumonia’, are loaded and resized, and the datasets are split into training and validation sets.

Data loaders are created for batch processing. The network’s initial convolutional layers extract hierarchical characteristics, with residual blocks with skip connections reducing the vanishing gradient issue. The residual output is sent to a fully connected layer, representing the two classes needed for the binary classification task. The model uses cross-entropy loss as the objective function and the Adam optimizer with a learning rate of 0.001. The training loop iterates over 10 epochs, the model’s performance being evaluated on a separate validation dataset. The model’s training accuracy is 96.48% and its validation accuracy is 83.66%.

The EfficientNet-B0 model architecture is a highly efficient convolutional neural network (CNN) for image classification tasks. The final feature vector is forwarded to the fully connected layer for classification purposes, where the model is trained to minimize cross-entropy loss which is the objective function and the Adam optimizer is employed with a learning rate of 0.001. The batch size is 32 and the training loop iterates over 10 epochs, while the validation loop evaluates the model’s performance on a separate validation dataset. The training accuracy is 97.96% and the validation accuracy is 81.49%.

The hybrid neural network model, is designed for processing sequential data, particularly images. This hybrid approach enables the system to effectively capture spatial patterns in medical images and sequential dependencies with regard to features, making it particularly advantageous for classification tasks. To preserve the spatial structure of feature maps, the output of the convolutional layers is flattened to a 1D vector and reshaped to fit the input shape expected by the LSTM layers. The LSTM layers are stacked hierarchically, with a decreasing number of memory cells from 128 to 16, enabling the model to learn increasingly abstract and compressed representations of the sequential data. The model is capped with a dense output layer employing the Softmax activation function, facilitating multi-class classification. The model uses categorical-cross-entropy loss as the objective function and the Adam optimizer with a learning rate of 0.0001. The batch size is 16 and the training loop iterates over 50 epochs, the model’s performance being evaluated on a separate validation dataset. Among the three deep learning models considered in this study -ResNet18, EfficientNetB0, and

CNN+LSTM -the CNN+LSTM model obtained a training accuracy of 99.42% and a validation accuracy of 83.66%. The validation results indicate that CNN+LSTM achieved a comparatively better performance than ResNet18 and EfficientNetB0 on the evaluated dataset.

5. Metadata Classification Using ML Models

Finding and using the appropriate methods to handle any null or missing values in the metadata is part of the metadata preparation procedure. The `isnull()` function was used to calculate and locate the missing values. Imputation techniques were used to address the few missing values in the metadata. The mean or mode imputation technique is used to impute features such as age, sex, and RT PCR positive, depending on whether the data is categorical or numerical. ML imputation is used to infer features such as survival, intubated, intubation present, Date, location, DOI, URL, license, and offset were among the attributes that were eliminated because they were insignificant for this work. Among the characteristics related to extra oxygen, extubation, temperature, pO2 saturation, leukocyte, neutrophil, and lymphocyte counts, the neutrophil count, and lymphocyte count had mostly null value, hence they were also dropped. Encoding was performed for the categorical variables that were present in the metadata.

Feature extraction is the process of capturing and representing relevant information or patterns in an image using a set of features. Deep learning techniques, particularly CNNs, have become powerful tools for feature extraction, as they automatically learn hierarchical features through convolutional layers. The CNN+LSTM model is loaded and the last three layers are removed, allowing the model to extract features. Principal Component Analysis (PCA) can be applied to reduce the dimensionality of image features while retaining the most significant information. It is often used in conjunction with other feature extraction methods to simplify the representation of data, making it more computationally efficient and revealing important underlying patterns. The extracted features are then combined with the preprocessed metadata and are used for training the ML models – the XGBoost Classifier and Random Forest Classifier as shown in Figure 5.

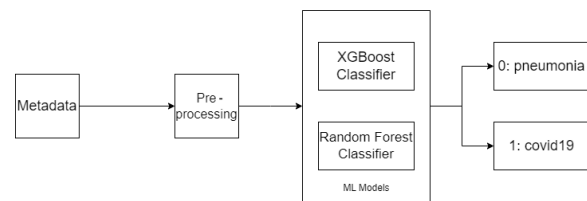


Figure 5. ML Model for Metadata Classification

The XGB Classifier refers to a specific implementation of XGBoost for classification tasks. During its creation, hyperparameters such as the number of trees (n estimators), learning rate (learning rate), and the maximum depth of the trees (max depth) can be specified. The hyperparameters used here are as follows: n estimators = 100, learning rate = 0.001, and max depth = 5. During training, XGBoost sequentially builds a series of decision trees, with each tree correcting the errors made by the ensemble of existing trees. Its training accuracy was 99.26% and its testing accuracy was 98.69%.

Random Forest employs a technique called bootstrap aggregating, or bagging, to create multiple decision trees. Each tree is trained on a bootstrapped sample of the training data, where random samples are drawn and then replaced. The hyperparameters used are n estimators = 100 and max depth = 5. The n estimators parameter specifies the number of decision trees that will be created by using Random Forest. A higher number of trees can potentially lead to a more robust and accurate model. However, this also leads to diminishing returns, and a balance needs to be struck between computational efficiency and model performance. The max depth parameter controls the maximum depth of each decision tree in the Random Forest. A tree's depth represents the number of nodes from the root to the farthest leaf. Limiting the depth helps prevent overfitting by restricting the complexity of individual trees, promoting a better generalization for unseen data. The training accuracy was 99.67% and the testing accuracy was 98.69%.

6. XAI Model Building

Explainable AI (XAI) model building involves developing machine learning models that not only make accurate predictions but also provide transparent and interpretable explanations for their decisions. Unlike traditional black-box models,

which may be difficult to understand and trust, XAI models aim to provide insights into the factors influencing their predictions, making them more understandable and trustworthy for humans. The XAI models used are Grad-Cam for providing a visual explanation of the output of the DL model - CNN+LSTM and Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) for providing the textual explanation of the output of the ML model - Random Forest.

As shown in Figure 6, Grad-CAM is a technique for visualizing and interpreting convolutional neural networks (CNNs) by highlighting the regions of an input image that contribute the most to the model's prediction. It achieves this by computing the gradient of the target class score with respect to the feature maps of the last convolutional layer. By combining these gradients with the feature maps, Grad-CAM generates a heatmap that highlights the regions of the image that are the most relevant to the predicted class. The implementation begins with the function `get_img_array()` that loads and preprocesses an image from a given path. This function uses the `keras.utils.load_img()` and `keras.utils.img_to_array()` methods to load the image and convert it into a NumPy array, respectively. Additionally, the image is resized to a specified target size using the `target_size` parameter. The resulting image array is then expanded along the batch dimension to create a 4D array suitable for the model input. The `make_gradcam_heatmap` function takes several inputs, including the image array, the CNN model, the name of the last convolutional layer, and an optional argument `pred_index` specifying the predicted class index. It operates by leveraging TensorFlow's GradientTape mechanism to compute gradients and generate the Grad-CAM heatmap. The function begins by constructing a gradient model using `keras.models.Model`.

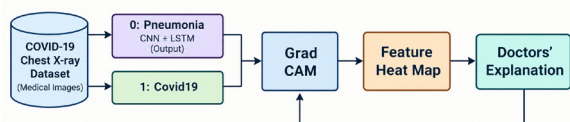


Figure 6. Grad-CAM for CNN + LSTM

As shown in Figure 7, SHAP is a powerful tool for explaining the output of machine learning models. It provides insights into the contributions

of individual features towards model predictions, thereby enhancing interpretability and trust in the model's decisions. SHAP values are based on the Shapley values from the cooperative game theory, which quantify the marginal contribution of each feature to the difference between the actual prediction and the average prediction. The `plot_type = bar` parameter specifies the type of summary plot to be generated, wherein features are represented as bars sorted by their average SHAP values. In addition to the summary plot, individual instance interpretations are visualized using force plots generated by the `shap.force_plot()` function. The new version of the Random Forest Classifier with the same hyperparameters was built using only the important features. Here, the important features are RT PCR positive, `pca_features`, `in_icu`, `age` and `intubated`. The training and testing accuracy for the new version of the Random Forest Classifier are the same as for the previous one. The testing accuracy is 98.69%.

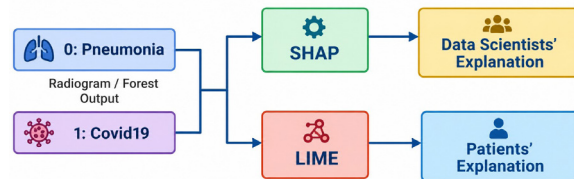


Figure 7. SHAP and LIME for the Random Forest Classifier

LIME is a model-agnostic framework designed to provide local interpretability for machine learning models. It explains individual predictions by approximating the model's behaviour around a specific instance through the use of interpretable surrogate models. The selected features included RT-PCR positivity, PCA-derived features, ICU admission status, patient age, chest X-ray view, and intubation status. The training and testing accuracy for the new version of the Random Forest Classifier are the same as for the previous one. The testing accuracy is 98.69%.

7. Experimental Results

The dataset used in this paper consists of publicly available chest X-rays and CT scans of people who have been diagnosed with COVID-19 or are suspected of having it, as well as people who have had pneumonia caused by other viruses and bacteria, such as MERS, SARS, and ARDS. The Covid19 chest-X-ray dataset [21] includes data

that was indirectly gathered from hospitals and doctors as well as information that was publicly available (the University of Montreal Ethics Committee has given its approval for this). The source of this dataset was a GitHub repository.

To enhance the robustness of the proposed system, the metadata features were carefully examined to reduce the effect of the variables that could strongly influence the diagnosis outcome. Only relevant and meaningful features were considered during model development to ensure a balanced learning. In addition, ablation studies were conducted by selectively removing metadata and image feature groups to analyze their contribution to the overall classification performance. The results helped in understanding the importance of each feature set within the multimodal framework. This analysis further strengthened the robustness and interpretability of the proposed Explainable AI system.

To further improve the reliability of the Explainable AI framework, additional validation measures were incorporated for assessing the explanations. Faithfulness and stability analyses were considered to examine whether the explanations remained aligned with the model predictions under varying conditions and repeated evaluations. Furthermore, the generated Grad-CAM, SHAP, and LIME explanations were qualitatively examined for their clinical relevance and interpretability in the context of chest X-ray analysis. This assessment helped ensure that the highlighted regions and influential features were medically meaningful and supportive of transparent healthcare decision-making. consistency and trustworthiness of the generated images.

The Grad-CAM heatmap highlights the image regions that the deep learning model focused on most when making its decision about the X-rays. A heat map is generated for the given image. It is a graphical representation where hotter colours indicate areas with a higher importance according to the model. The heatmap is layered on top of the original X-ray image. The areas with the most intense colours on the heatmap, indicate which parts of the X-ray were most critical for the model's decision. This would provide a visual explanation to the doctors for trusting the AI model as shown in Figure 8.

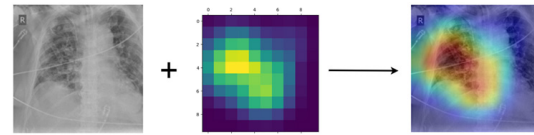


Figure 8. Doctor's Explanation

Figure 9 shows the explanation for Data Scientists which includes the details of the model responsible for prediction along with the related hyperparameters used. It also includes the SHAP explanation which consists of the expected value of the predicted class along with the SHAP values of all features. It conveys whether a feature has a positive or negative contribution towards the prediction. A visualization of the SHAP results is also shown.

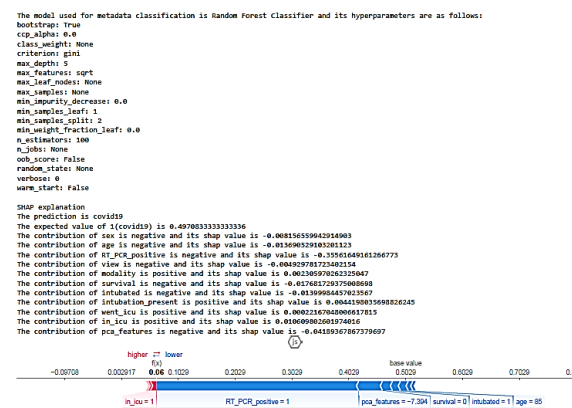


Figure 9. Data Scientist's Explanation

Figure 10 shows the explanation for patients which includes the probability of the prediction and important features responsible for the prediction. This differs from patient to patient data and hence LIME is used for the explanation. The features which are in medical terms that cannot be understood by patients are explained in simpler words. The priority or weightage of the importance of the features are visualized through the bar graph.

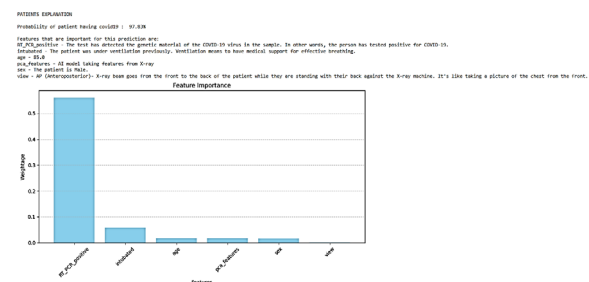


Figure 10. Patient's Explanation

7.1 Performance Analysis

The DL models - ResNet18, EfficientNetB0 and CNN+LSTM - are successfully built, trained and validated on the respective datasets. Hyperparameters in deep learning models are parameters that are not learned from the data but are set prior to the training process. They control the overall behaviour of the model and its training process. Tuning hyperparameters is a crucial step in optimizing the performance of a deep learning model. The hyperparameters of all the three models are included in Table 2.

Table 2. Hyperparameters for the DL Models

Hyper parameters	ResNet18 & EfficientNetB0	CNN+LSTM
Batch size	32	16
Epochs	10	50
Loss Function	Binary Cross-Entropy	Categorical Cross-Entropy
Optimizer	Adam	Adam
Learning Rate	0.001	0.0001
Training Size	80%	80%
Testing Size	20%	20%

The performance metric used is accuracy. Accuracy is a commonly used metric for evaluating the performance of a classification model. It is a measure of the overall correctness of the model in predicting the classes of the given set of samples. Accuracy is calculated as the ratio of the correctly predicted images to the total number of images. Comparatively, the CNN+LSTM works better than other models, achieving a training accuracy of 99.42% and a validation accuracy of 83.66%, while the training accuracy for the ResNet18 and EfficientNetB0 models is 96.48% and 97.96%, respectively and the validation accuracy for ResNet18 and EfficientNetB0 is 83.66% and 81.49%, respectively. Table 3 shows the comparative analysis of the results for the DL models.

The ML models - XGBoost and Random Forest Classifiers - are successfully built, trained, and validated on the respective datasets. Hyperparameters for machine learning models are settings or configurations that are not learned

from the data during training but are set prior to training. These parameters control the learning process and directly impact the performance and behaviour of the model. The hyperparameters of both models are shown in Table 4.

Table 4. Hyperparameters for the ML Models

Hyper parameters	Used values
n_estimators	100
Learning Rate	0.001 (XGBoost)
max_depth	3 (XGBoost) and 5 (Random Forest)
Training Size	80%
Testing Size	20%

Comparatively, the Random Forest Classifier works better than the XGBoost Classifier achieving a training accuracy of 99.67% and a testing accuracy of 98.69%, while the training and testing accuracy for XGBoost is 99.26% and 98.69%, respectively. Table 5 shows the comparative analysis of the results for the ML models.

The SHAP and LIME models can be evaluated by building another version of the ML model with only the important features. To conclude that the explanations given by the SHAP and LIME models are correct, this version of the model either performs the same as the previous one or it should perform better. By applying SHAP for the Random Forest Classifier model, the important features are derived. The new version of the Random Forest Classifier with the same hyperparameters is built using only the important features. In this case, the important features are RT PCR positive, pca features, in_icu, age and intubated. The training and testing accuracy for the new version of the Random Forest Classifier is the same as for the previous one. The testing accuracy is 98.69%. By applying LIME for the Random Forest Classifier model, the important features are derived. The new version of the Random Forest Classifier with the same hyperparameters is built using only the important features. In this case the important features are survival, intubated, intubation present. The training and testing accuracy for the new version of the Random Forest Classifier is the same as for the previous one. The testing accuracy is 98.69%.

Table 3. Evaluation of the DL Models

DL Model	Training Accuracy	Validation Accuracy	Precision	Recall	F1-Score	Specificity
ResNet18	96.48%	83.66%	84.21%	82.75%	83.47%	85.10%
EfficientNetB0	97.96%	81.49%	82.30%	80.95%	81.62%	83.40%
CNN+LSTM	99.42%	83.66%	85.12%	83.90%	84.50%	86.25%

Table 5. Evaluation of the ML Models

ML Model	Training Accuracy	Validation Accuracy
XGBoost Classifier	99.26%	98.69%
Random Forest Classifier	99.67%	98.69%

The proposed hybrid model shows a strong performance, where CNN+LSTM achieves a high training accuracy, while Random Forest performs better as generalization is concerned when multimodal features are used. This indicates that combining deep learning with machine learning helps in improving robustness and reducing overfitting. The use of both chest X-ray images and patient metadata makes the system more reliable compared with the case when only a single type of data was used, which is important in real clinical scenarios. In addition, explainability methods such as Grad-CAM, SHAP, and LIME help in understanding how the model makes decisions, thereby increasing transparency and trust among clinicians. The good balance between recall and specificity also shows that the model can correctly identify both the normal cases and the cases involving either Covid-19 or pneumonia.

8. Conclusion

This study presents a domain-specific explainable AI framework for healthcare that focuses on identifying COVID-19 and pneumonia cases using chest X-ray images combined with patient metadata. This system integrates deep learning models (ResNet18, EfficientNetB0, and CNN+LSTM) for feature extraction and classification, and CNN+LSTM achieved the highest training accuracy (99.42%) with a validation accuracy of 83.66%. The extracted image features, when fused with metadata, were further utilized by machine learning models, with the Random Forest classifier demonstrating a strong performance (a 99.67% training accuracy

and a 98.69% testing accuracy). To enhance interpretability, explainability techniques including Grad-CAM, SHAP, and LIME were applied, providing insights tailored to different stakeholders such as clinicians, patients, and data scientists. Additionally, a reduced-feature model based on XAI-identified important features achieved a comparable performance, indicating the relevance of the selected features. Overall, the results suggest that the proposed framework can achieve an effective classification while improving transparency; however, the conclusions are drawn in alignment with the observed validation results, and further evaluation on larger and more diverse datasets would strengthen its applicability in real-world clinical settings.

However, despite achieving promising classification and explainability results, the proposed framework has certain limitations. The dataset used in this study may contain variations in image quality, class distribution, and data collection conditions, which can introduce bias during model learning. In addition, the reliability of the labels and metadata depends on the quality of the original dataset annotations. Although the developed models performed well on the available data, their effectiveness across different hospitals, imaging devices, and patient populations requires further investigation. Another limitation lies in the absence of external validation using independent clinical datasets, which is important for evaluating real-world generalizability. Future work will focus on incorporating larger and more diverse datasets to improve the robustness and clinical applicability of the proposed Explainable AI framework.

REFERENCES

- Adams, J. & Hagraas, H. (2020) A Type-2 Fuzzy Logic Approach to Explainable AI for regulatory compliance, fair customer outcomes and market stability in the Global Financial Sector. In: *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 19-24 July 2020, Glasgow, UK. <https://doi.org/10.1109/FUZZ48607.2020.9177542>.
- Ali, S., Abuhmed T., El-Sappagh, S. et al. (2023) Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*. 99, Art. No. 101805. <https://doi.org/10.1016/j.inffus.2023.101805>.
- Ariza-Garzon, M. J., Arroyo, J., Caparrini, A. et al. (2020) Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending. *IEEE Access*. 8, 64873-64890. <https://doi.org/10.1109/ACCESS.2020.2984412>.

- Chatterjee, S., Dey, N. & Ashour, A. S. (2024) Comparative analysis of explainable AI techniques for COVID-19 detection using deep neural networks. *Journal of Imaging*. 10(2), 45.
- Chen, J., Dai, X., Yuan, Q. et al. (2020) Towards Interpretable Clinical Diagnosis with Bayesian Network Ensembles Stacked on Entity-Aware CNNs. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 2020, Online*. Stroudsburg, USA, Association for Computational Linguistics. pp. 3143–3153.
- Guefrechi, S., Jabra, M. B., Ammar, A et al. (2021) Deep learning based detection of COVID-19 from chest X-ray images. *Multimedia Tools and Applications*. 80(21-23), 31803-31820. <https://doi.org/10.1007/s11042-021-11192-5>.
- Gupta, A., Jain, J., Poundrik, S. et al. (2021) Interpretable AI Model-Based Predictions of ECG changes in COVID-recovered patients. In: *2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART), 8-10 December 2021, Paris/ Creteil, France*. New York, USA, IEEE. pp. 1-5. <https://doi.org/10.1109/BioSMART54244.2021.9677747>.
- Holzinger, A., Saranti, A., Molnar, C. et al. (2022) Explainable AI Methods - A Brief Overview. In: Holzinger, A., Goebel, R., Fong, R. et al. (eds.) *xxAI - Beyond Explainable AI. xxAI 2020* (part of the series *Lecture Notes in Computer Science*, vol 13200). Cham, Switzerland, Springer, pp. 13-38.
- Jin W., Li X., Fatehi M. et al. (2023) Generating post-hoc explanation from deep neural networks for multi-modal medical image analysis tasks. *MethodsX*. 10, Art. No. 102009. <https://doi.org/10.1016/j.mex.2023.102009>.
- Khosravi, H., Shum, S. B., Chen, G. et al. (2022) Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*. 3, Art. No. 100074. <https://doi.org/10.1016/j.caeai.2022.100074>.
- Lamy, J.-B., Sekar, B., Guezennec, G. et al. (2019) Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine*. 94, 42-53. <https://doi.org/10.1016/j.artmed.2019.01.001>.
- Langer, M., Oster, D., Speith, T. et al. (2021) What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*. 296, Art. No. 103473. <https://doi.org/10.1016/j.artint.2021.103473>
- Onari, M. A., Nobile, M. S., Grau, I. et al. (2022) Comparing Interpretable AI Approaches for the Clinical Environment: an Application to COVID-19. In: *2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 15-17 August 2022, Ottawa, ON, Canada*. New York, USA, IEEE. <https://doi.org/10.1109/CIBCB55180.2022.9863020>.
- Patel, P. (2020) Chest X-ray (Covid-19 & Pneumonia). <https://www.kaggle.com/datasets/prashant268/chest-xray-covid19-pneumonia> [Accessed: 1st August 2024].
- Rachha, A. & Seyam, M. (2023) Explainable AI In Education: Current Trends, Challenges, And Opportunities. In: *SoutheastCon 2023, 1-16 April 2023, Orlando, FL, USA*. New York, USA, IEEE. pp. 232-239.
- Rajpoot, N., Jain, S. & Gupta, A. (2024) Explainable ensemble deep learning model for COVID-19 detection using chest imaging. *IEEE Access*. 12, 14523–14538.
- Ras, G., Xie, N., van Gerven, M. et al. (2022) Explainable Deep Learning: A Field Guide for the Uninitiated. *Journal of Artificial Intelligence Research*. 73, 329-396. <https://doi.org/10.1613/jair.1.1320>.
- Shaik, T., Tao, X., Li, L. et al. (2023) A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. *Information Fusion*. 102, Art. No. 102040. <https://doi.org/10.1016/j.inffus.2023.102040>.
- Shelke, A., Inamdar, M., Shah, V. et al. (2021) Chest X-ray Classification Using Deep Learning for Automated COVID-19 Screening. *SN Computer Science*. 2(4), 300. <https://doi.org/10.1007/s42979-021-00695-5>.
- Soares, E., Angelov, P., Biaso, S., et al. (2024) An explainable deep learning framework for COVID-19 diagnosis using CT images. *Neural Computing and Applications*. 36, 11245–11260.
- Sun, L., Mo, Z., Yan, F. et al. (2023) Evaluation of explainable artificial intelligence methods in medical imaging: A perturbation-based approach. *Scientific Reports*. 13, Art. No. 46493.
- van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A. et al. (2022) Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*. 79, Art. No. 102470. <https://doi.org/10.1016/j.media.2022.102470>.
- Wang, Y.-C., Toly Chen, T.-C. & Chiu, M.-C. (2023) A systematic approach to enhance the explainability of artificial intelligence in healthcare with application to diagnosis of diabetes. *Healthcare Analytics*. 3, Art. No. 100183. <https://doi.org/10.1016/j.health.2023.100183>.
- Yang, C.C. (2022) Explainable Artificial Intelligence for Predictive Modeling in Healthcare. *Journal of Healthcare Informatics Research*. 6, 228–239. <https://doi.org/10.1007/s41666-022-00114-1>.
- Zhang, Y., Wu, H. & Li, X. (2024) Attention-based capsule network for COVID-19 detection from chest X-ray images. *SN Applied Sciences*. 6, Art. No. 579.



This is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.