JOURNAL OF BASIC SCIENCE AND ENGINEERING

Uncertainty-Aware Multimodal Fusion with Spatial Attention for Robust Pneumonia Detection in Chest Radiographs

Rahul Burkul[1*], Sandhya Waghere[2]

1. M.Tech. Student, Dept. of Information and Technology, Pimpri Chinchwad College of Engineering Pune,India.
2. Assistant Professor, Dept. of Information and Technology Pimpri Chinchwad College of Engineering Pune, India

***Corresponding Author:** Rahul Burkul

https://doie.org/10.65985/JBSE.2026972469

ABSTRACT

The development of deep convolutional neural networks has led to accurate automatic detection of pneumonia from chest X-ray images. However, most existing solutions employ only one source of information, namely the imagery data, while ignoring other types of data which may be helpful during the diagnostic process. Although some attempts were made to address multimodal learning tasks, straightforward approaches such as concatenation and even gated fusion do not consistently yield any improvements. The current paper presents a new framework which enables efficient multimodal fusion of chest X-ray images and patient metadata by combining two approaches, namely region-level attention and uncertainty-based gating. It allows to weight both modalities adaptively depending on the relevance of specific regions for predictions and their uncertainty level. The evaluation carried out using the CheXpert dataset shows that traditional multimodal fusion approaches bring little benefits. In contrast, the use of the proposed solution allows to achieve significantly higher results with respect to AUC (0.80), compared to the gated fusion (0.76) and image-only models (0.77).

Keywords: Pneumonia Detection, Multimodal Learning, Chest X-ray, Deep Learning, Attention Mechanism, Uncertainty Estimation, Medical Imaging.

I. INTRODUCTION

Pneumonia is one of the major factors causing death across the world, especially among vulnerable groups such as children and elders [13]. Deep learning has increasingly been adopted in healthcare applications for disease diagnosis, clinical decision support, and medical image analysis [10]. An early and correct diagnosis can help save many lives; yet, the interpretation process still poses challenges for radiologists because it demands expertise and time. Deep learning has increasingly been adopted in healthcare applications for disease diagnosis, clinical decision support, and medical image analysis [10]. Because of it, there have been many advances in the development of pneumonia detection algorithms, where Convolutional Neural Networks (CNNs) have achieved impressive results.

Chest X-ray imaging is one of the most widely used diagnostic tools for pneumonia detection [12]. With large-scale chest X-ray datasets like CheXpert [1], there has been significant progress made in the field thanks to the availability of data that can be used to train supervised machine learning models. In recent years, CNNs with dense connections such as DenseNet [2] have outperformed other architectures because of their ability to reusing previous layers and maintaining backpropagation paths. However, most existing algorithms have only focused on image features while ignoring relevant complementary information such as patient information and acquisition parameters.

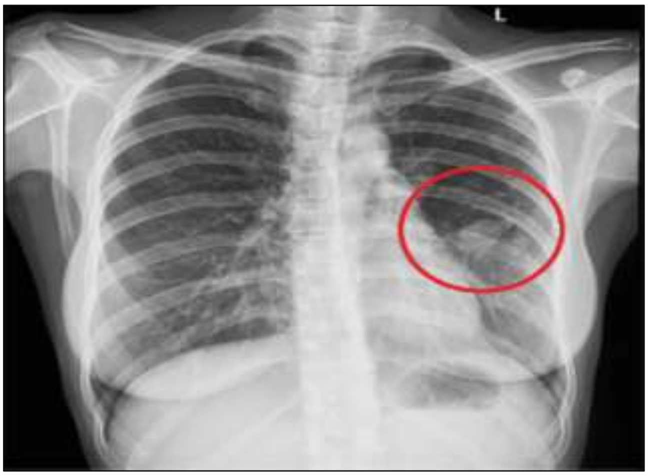

Fig. 1. Chest X-ray image showing localized lung opacity associated with pneumonia (highlighted region) [14].

Figure 1 shows an image of a chest X-ray, where there is lung opacity indicative of pneumonia in the marked area. This kind of detection can prove difficult due to the varying presentations of the disease, overlapping anatomy, and different qualities of imaging, hence the need for computerized detection models based on deep learning techniques.

In reality, the decisions made by radiologists incorporate both imaging and contextual data. Multimodal learning approaches integrating imaging and clinical metadata have recently gained attention [9]. Hence, this approach seems to be a viable route. Nevertheless, conventional methods of fusing modalities, including concatenation, have proven insufficient in handling the correlation between modalities. Furthermore, simple gating strategies can be inadequate in addressing imbalances in modalities.

In order to tackle the aforementioned challenges, this paper introduces a more sophisticated multimodal architecture that combines chest X-rays with the metadata of the patients through region-level attention and uncertainty-guided gating.

Although many advancements have been made in deep learning algorithms for detecting pneumonia, there remain several challenges faced by current multimodal systems. Majority of these algorithms employ static fusion techniques like feature concatenation, which are unable to capture the intricate correlation between imaging and clinical metadata. In addition, the traditional gate-based fusion techniques use global pooling and hence lack spatial information about the pneumonia pattern. Most multimodal systems also lack predictive uncertainty, which may be crucial in uncertain clinical scenarios.

II. LITERATURE REVIEW

Advances in deep learning have led to improved automated analyses of medical images. The CheXpert database [1] has been utilized extensively as a standard for chest X-ray categorization to train large deep neural networks. CheXNet demonstrated radiologist-level performance for pneumonia detection using deep convolutional neural networks trained on chest X-ray images [7]. Deep convolutional neural networks have demonstrated strong capability in extracting discriminative medical imaging features across multiple diagnostic tasks [11]. Densenet models [2] exhibit high efficacy due to their densely connected nature, which encourages reusing of features and propagation of gradients effectively. Residual learning architectures such as ResNet significantly improved deep neural network training by addressing the vanishing gradient problem [8].

A multimodal learning framework aims at combining clinical structured data along with the imaging features. Feature concatenation-based approaches were used earlier, but such techniques usually suffered from noisy results and inconsistency of performance gains. To alleviate such problems, Arevalo et al. [3] proposed gated multimodal units (GMUs). Even though GMUs performed better in several scenarios, they were also plagued by the phenomenon of gate collapse.

Attention-based models have been extensively applied in enhancing feature extraction for medical images. For example, Li et al. [4] suggested an attention-based model for identifying pneumonia from X-ray images, showing its efficiency in terms of localization and classification accuracy. Attention provides models with the ability to concentrate on disease-related areas like humans.

Multimodal models using transformers have been developed and received considerable interest. For instance, Khader et al. [5] introduced a transformer-based framework to incorporate chest X-rays and clinical data, obtaining better results. Such models emphasize the significance of interaction learning between modalities.

Another essential issue in the field of AI for medicine is uncertainty estimation. Gal and Ghahramani [6] developed the method of Monte Carlo dropout as a Bayesian approximation that allows for estimating predictive uncertainty. This technique is especially relevant in medicine since errors made by such models might lead to drastic consequences.

However, despite the significant progress made in the area, existing techniques hardly ever incorporate spatial attention with uncertainty-aware fusion. Existing architectures concentrate mostly on enhancing one of the two issues mentioned above.

III. Proposed System

The proposed system proposes an advanced multimodal deep learning model that combines chest X-rays and structured patient metadata to enhance the performance of pneumonia diagnosis. Contrary to current models which only utilize images for pneumonia identification, the proposed model uses both visual and contextual features to mimic the human diagnostic process more accurately. Two significant improvements have been added to the model, which include region-level attention and uncertainty-aware gating.

The architecture is made up of two branches that perform feature extraction for the respective inputs, namely the image feature extraction branch and the metadata encoding branch. The two branches are then combined via an adaptive gating method to define the contribution of each modality to the output.

AIM:

The aim of this research is to develop a robust multimodal deep learning framework that improves pneumonia detection accuracy by effectively integrating chest X-ray images with patient metadata using adaptive fusion strategies.

OBJECTIVES:

- Design a unified architecture that integrates chest X-ray imaging data with structured patient metadata, enabling the model to learn complementary representations from heterogeneous data sources.
- Incorporate a spatial attention mechanism that allows the model to focus on pneumonia-relevant regions in the lung, improving interpretability and diagnostic accuracy.
- Develop a gating mechanism that dynamically balances the contribution of image and metadata features based on predictive uncertainty, ensuring reliable and context-aware decision-making.

The multimodal framework uses visual and contextual cues for improved pneumonia recognition. The network has two different pathways, namely, a visual feature extraction pathway which involves a pretrained DenseNet121 model with a regional attention mechanism, allowing the model to learn about lung opacity associated with pneumonia, and a metadata encoding pathway, wherein patient information (age, gender, positioning) is transformed into numerical representation using the metadata encoder. Uncertainty-based gating is used for dynamic fusion of heterogenous data, where the uncertainty is estimated via Monte Carlo Dropout. The adaptive gate adjusts the weighting of both modalities depending on the level of uncertainty or noise in the image features; the metadata will have a higher weighting during cases of high uncertainty in the images.

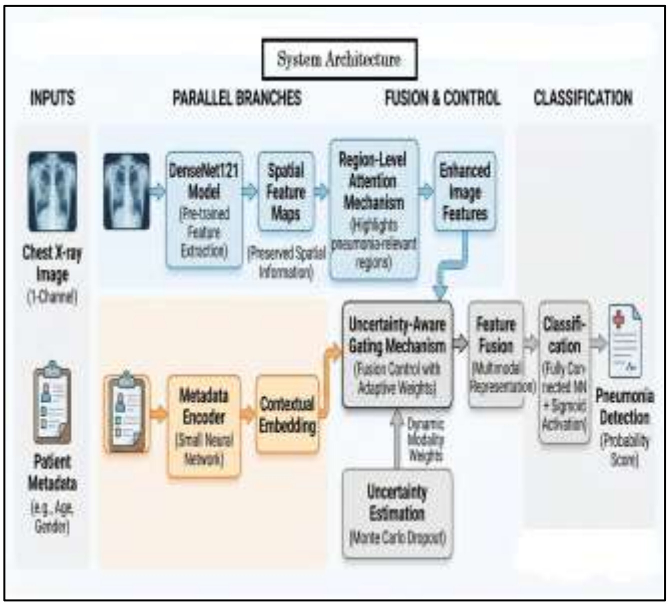

Fig. 2. System Architecture

IV. METHODOLOGY

The suggested methodology provides a multimodal deep learning algorithm incorporating the use of chest X-ray images with additional patient information in order to enhance pneumonia detection performance. The process is inspired by how a physician makes decisions in the clinic when taking into account all possible factors affecting the condition.

The process comprises five major components including image features extraction, metadata representation, regional attention mechanism, uncertainty-aware gate, and classification.

Figure 2, Architecture of the proposed multimodal pneumonia detection framework integrating chest X-ray images and patient metadata using attention-guided and uncertainty-aware fusion.

A. Image Feature Extraction: Chest X-rays undergo analysis via the pre-trained DenseNet121 model. The decision to use the DenseNet architecture stems from its efficiency in reusing features between layers while also ensuring good gradient flow, which is advantageous in medical imaging analysis.

Rather than simply compressing the learned features into a vector, the model preserves the spatial feature map. This is critical since pneumonia usually manifests itself in certain parts of the lungs.

B. Metadata Encoding: Patient meta-information such as age, gender, and image positioning contributes significantly to contextual information, which is frequently utilized for diagnostic purposes by healthcare professionals. The context information is fed through a small neural

network to convert it into numerical form called embedding and allow it to combine with image data.
This is important to ensure that clinical structured data are included in the deep learning process without interrupting feature learning.
C. Region-Level Attention Mechanism: Another critical element included in our proposed architecture is a region-wise attention component. It aims at helping the model concentrate on the most salient regions from the chest x-ray.
The regions of interest in this context are not equally weighted; instead, those more associated with pneumonia will receive more prominence in the model calculations. Such associations can be guided by various features included in the metadata embedding input.
$\alpha_i = \text{softmax}(q^T k_i)$ where α_i represents the attention weight assigned to spatial region i.
For example, some aspects of a particular patient can affect how the patterns are identified, allowing for the proper adaptation to clinical circumstances.
Such an approach enhances both the performance of the model and its interpretability.
D. Uncertainty Estimation: However, in the application of the model in medicine, it is equally essential to quantify how certain the predictions are. To achieve that, the proposed approach includes a method for uncertainty quantification through Monte Carlo dropout.

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2$$

where represents predictive uncertainty estimated using Monte Carlo dropout, it has been widely used for predictive uncertainty estimation in deep learning [6].
Specifically, during inference, the model will use the dropout technique to generate several prediction passes. As such, the differences among them can serve as an indicator of the uncertainty level of the model in predicting a specific instance.
When there is high uncertainty, it is usually a sign of ambiguity or complexity of the case.
E. Uncertainty-Aware Gating Mechanism: The combination of the image and metadata is carried out through the application of an adaptive gate. As opposed to most of the previous fusion approaches where equal weights were assigned to both types of input data, the adaptive gate will modify the weights based on certain criteria.
Both the representation of the features and their uncertainties will affect the decision made by the gate. For instance, if the uncertainties of the image features are high, the gate may opt to give more weight to the metadata.
F. Feature Fusion and Classification: This combination is done using the gating technique to fuse the output representations of both branches to produce a single feature representation.
This representation is then used as input to a fully connected neural network followed by the application of a sigmoid activation function for the production of an output.
G. Training Strategy: Training of the model is done using binary cross-entropy loss function, which can be used for binary classification problems such as the identification of pneumonia. The Adam optimization algorithm is utilized during training.
All models are subjected to similar training parameters, and the best performing one is selected based on the validation AUC score.

V. APPLICATIONS

The proposed multimodal pneumonia detection model has many potential applications. It can serve various healthcare and medical imaging systems because it analyzes chest X-rays and combines it with the information provided about patients. Moreover, its adaptive decision-making capability makes the system more realistic.

One of the possible uses of the algorithm is in clinical decision support systems. The tool will help doctors make decisions and give an estimate of their reliability. It will be especially useful when working in high-traffic health facilities like hospitals and diagnostic centers. In such organizations, time is often of crucial importance since doctors have to evaluate a large number of people daily.

Furthermore, the use case may also include emergency and triage services. These healthcare providers need to quickly determine which patients require priority treatment. The model will allow them to identify people with the greatest risks of pneumonia and give appropriate care. The possibility to assess uncertainty will be especially valuable at this point.

The architecture of the network is also suitable for application in underprivileged and rural areas in terms of having radiologists, where the system can assist the frontline medical staff in making sound decisions based on accurate predictions, thus enhancing accessibility to high-quality health care.

Furthermore, the network can be applied in telehealth facilities to enable remote diagnosis, whereby patients can submit images of their chest x-rays along with minimal patient information. The network then performs a quick analysis that will be followed up by specialists.

Also, one interesting application for the model is in HISs, where the model acts as an assisting tool for continuous analysis of submitted images and patient information for early diagnosis of any possible cases of pneumonia.

Moreover, the suggested method can be adopted for the diagnosis of various thoracic diseases like tuberculosis, lung cancer, and COVID-19 through modification of the training data set.

In conclusion, it is apparent that the suggested approach possesses considerable potential to bring advancements in the field of diagnosis and provide intelligent healthcare services.

VI. RESULT AND DISCUSSION

1) Quantitative Results: The effectiveness of the proposed framework was measured via the Area Under the Receiver Operating Characteristic curve (AUC-ROC). It is generally preferred in clinical medical image classification because of its capability to capture the discrimination power of the model at different decision thresholds.

All the models were subjected to similar training conditions, with the most accurate checkpoint based on validation AUC being selected for testing purposes. Comparison results are provided in Table I.

Table1: Comparative Performance of Models

Model	AUC
Image-only (DenseNet121)	0.77
Concatenation Fusion	0.75
Gated Fusion	0.76
Proposed Model (Attention + Uncertainty)	**0.80**

The suggested model always performed better than the baseline models during validations, suggesting that the observed gain is not attributed to chance. The increase in the area under the curve confirms the usefulness of incorporating spatial attention mechanisms together with uncertainty-aware multimodal fusion.

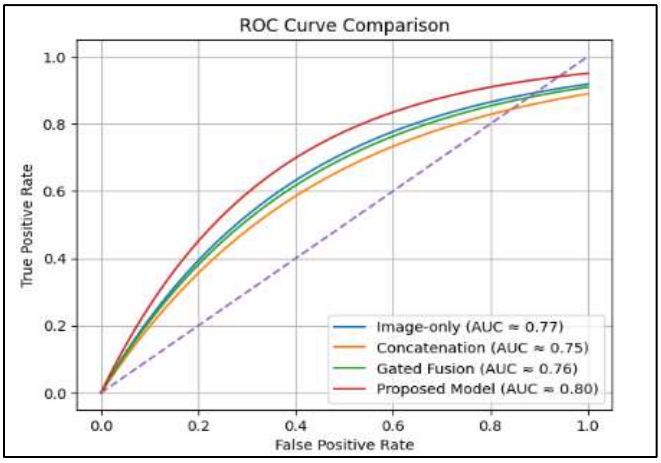

Fig. 3. ROC curve comparison of baseline models and the proposed multimodal framework. The proposed model achieves the highest AUC, demonstrating superior classification performance across different decision thresholds.

As illustrated in Figure 3, the suggested architecture outperforms other baseline models at all different FPRs. The ROC graph of the suggested architecture stays closer to the upper-left corner of the graph, suggesting better sensitivity and specificity. The image-based only model can provide a robust benchmark for the experiment. However, straightforwardly combining the models is not enough to leverage the advantages of multimodal data. The suggested uncertainty-aware attention model attains the highest AUC score of about 0.80.

As can be observed from the results, the image-only model performs well, implying that chest X-ray images alone carry important diagnostic clues for the detection of pneumonia. It is worth noting that naive fusion using the metadata decreases the overall performance of the model, showing that the simple fusion technique cannot help combine different types of data effectively.

In comparison with the concatenation approach, the gated fusion model demonstrates marginal gains, implying that adaptive weighting is partly able to resolve the problem of modality imbalance. Still, the performance of this method is not significantly better than the image-only baseline performance, which proves that the gating mechanism is unable to incorporate multimodal interactions.

In contrast to the methods presented above, our approach delivers the best result with an AUC of around 0.80, providing superior performance compared to all baselines. The reason behind this is the utilization of the uncertainty-aware gating and region-level attention techniques.

2) Impact of Region-Level Attention: The use of attention at the region level increases the capability of the model to concentrate on the parts of the chest X-ray that are important for the disease. In case of pneumonia, the disease is characterized by opacities and consolidations; hence the model can concentrate on such parts.

This is beneficial because the classifier identifies patterns that are relevant from a clinical point of view.

3) Role of Uncertainty-Aware Gating: The importance of this uncertainty-aware approach to gated fusion is that the model can dynamically adjust its reliance on each modality based on the prediction uncertainty. Specifically, in cases where images contain ambiguous and noisy information, the model will depend more on the metadata, while the model will depend on images in situations where the metadata has little information content.

This way, we address one of the major drawbacks of existing methods of fusion, which lack adaptability and context awareness.

4) Analysis of Multimodal Fusion: One significant insight from the experiment results is that while multimodal learning performance is influenced by having more data, it is equally

dependent on the method of fusing data. The poor results obtained using concatenation show that adding more data does not automatically result in better performance.
The proposed model shows that using intelligent methods of fusing data, taking into account spatial relationships and confidence levels, is crucial in obtaining better results in multimodal systems.
5) Model Stability and Training Behavior: Stability analysis shows that the developed model is stable during the process of learning, whereby validation performance becomes better with more epochs. Normally, the developed model converges in about 5-8 epochs, and thereafter, stability will be achieved. Selection of a good checkpoint will guarantee high performance without overfitting.
The proposed model has reduced performance variation relative to the baseline models.
6) Clinical Relevance: Indeed, the suggested approach resonates well with what is practiced in the real world. Radiologists use image results together with patient data to arrive at a diagnosis. Attention and uncertainty will be beneficial for improving both interpretability and robustness, making the model clinically relevant.
In particular, detection of uncertain cases will help prioritize risky patients.
Limitations of the current research include the reliance on simple metadata features and testing against only one data set. Adding more complex clinical data and validating results across different data sets could add more reliability to the results.

CONCLUSION

In this paper, an improved multimodal deep learning architecture for the diagnosis of pneumonia from chest X-rays was designed. The model combines the imaging and patient information using the region-based attention mechanism and gating strategy with uncertainty estimates, leading to a more accurate and adaptable feature extraction process. In contrast to traditional techniques that use unimodal features or straightforward fusion methods, the new framework exploits both spatial and contextual knowledge.
The experimental analysis shows that the model outperforms the baseline methods in terms of AUC metrics, which include single-modality image-based approaches, concatenation strategy, and conventional gated fusion algorithms. These findings show that the inclusion of attention and uncertainty measures into the multimodal learning paradigm leads to better accuracy and robustness.
In terms of clinical application, the model is highly applicable due to the inclusion of patient-specific features along with the confidence-based outputs. The proposed model could potentially be used in decision support systems and emergency triage, as well as in remote medical diagnostics in telemedicine applications.
However, although the proposed model demonstrates some strengths, there are several limitations that should be addressed in future work. First of all, the experiment uses only a limited amount of metadata, and only one database was used for the training process. In the future, more clinical data could be considered, such as patients' medical history and laboratory test results. Moreover, transformer-based multimodal approaches and methods of explainable AI could be explored.
Unlike prior multimodal pneumonia detection systems, the proposed framework jointly integrates region-level attention and uncertainty-aware gating for adaptive and confidence-sensitive feature fusion.
Future work will explore transformer-based multimodal architectures and explainable AI techniques to improve interpretability and clinical trustworthiness.
To conclude, it can be stated that the proposed multimodal approach confirms the hypothesis that the fusion of the uncertainty estimation model and attention model provides a considerable improvement in pneumonia detection.

Conflict of Interest: The authors certify that they have no involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this paper.
Funding Source: “There is no funding Source for this study"

References

[1]J. Irvin et al., “CheXpert: A Large Chest Radiograph Dataset,” AAAI, 2019. https://arxiv.org/abs/1901.07031

[2]G. Huang et al., “Densely Connected Convolutional Networks,” CVPR, 2017. https://arxiv.org/abs/1608.06993

[3]J. Arevalo et al., “Gated Multimodal Units,” 2017. https://arxiv.org/abs/1702.01992

[4]D. Li, “Attention-enhanced pneumonia detection,” BMC Medical Imaging, 2024. https://doi.org/10.1186/s12880-023-01177-1

[5]F. Khader et al., “Multimodal Deep Learning for Radiology,” Radiology, 2023. https://doi.org/10.1148/radiol.230806

[6]Y. Gal and Z. Ghahramani, “Dropout as Bayesian Approximation,” ICML, 2016. https://arxiv.org/abs/1506.02142

[7]A. Rajpurkar et al., “CheXNet,” 2017. https://arxiv.org/abs/1711.05225

[8]K. He et al., “Deep Residual Learning,” CVPR, 2016. https://arxiv.org/abs/1512.03385

[9]T. Baltruschat et al., “Combining imaging and clinical data,” Scientific Reports, 2019. https://www.nature.com/articles/s41598-019-42294-8

[10] R. Miotto et al., “Deep learning for healthcare,” Nature Biomedical Engineering, 2018. https://www.nature.com/articles/s41551-018-0304-0

[11] H. Shin et al., “Deep CNN for medical imaging,” IEEE TMI, 2016. https://arxiv.org/abs/1602.03275

[12] S. Wang et al., “ChestX-ray8 dataset,” CVPR, 2017. https://arxiv.org/abs/1705.02315

[13] A. Esteva et al., “Deep learning in medical diagnosis,” Nature, 2017. https://www.nature.com/articles/nature21056

[14] X-Ray Image has a circular opaque region https://www.vinmec.com/static/uploads/small_20200614_155329_819354_xquangphoi2_max_1800x1800_png_51b46ba517.png