Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 28 (2025), No. 1, pp. 193–204
DOI : 10.47974/JSMS-1351

Statistical assessment of an interpretable AI framework for improved disease diagnosis through medical images

K. M. Karthick Raghunath *
Muma College of Business
University of South Florida
8350 N. Tamiami Trail Sarasota
Florida, FL 34243
U.S.A.

and

Department of Computer Science and Engineering
JAIN (Deemed-to-be-University)
Karnataka 562112
India

Bhuvan Unhelkar †
Muma College of Business
University of South Florida
8350 N. Tamiami Trail Sarasota
Florida
FL 34243
U.S.A.

S. Siva Shankar §
Department of Computer Science and Engineering
KG Reddy College of Engineering and Technology
Moinabad 500075
Telangana
India

---

* *E-mail:* **karthick.km@jainuniversity.ac.in (Corresponding Author)**
† *E-mail:* **bunhelkar@usf.edu**
§ *E-mail:* **drsivashankars@gmail.com**

Prasun Chakrabarti [‡]
Department of Computer Science and Engineering
Sir Padampat Singhania University
Udaipur 313601
Rajasthan
India

Abstract

In medical imaging, radiographic diagnoses based on computational techniques and statistical analysis call for a solution that integrates the gap between computational methods and statistical tools to ensure the correctness and reliability of the diagnosis process. Such a gap led to the innovation of a new framework called the "Statistically Improved Convolutional Interpretation (SICI)" framework for transparent and reliable diagnosis of disease from medical images. SICI framework is explored based on Convolutional Neural Networks (CNNs), attention mechanisms, and Explainable artificial intelligence (XAI) and complements it with statistical models. The foundation of this approach is based on the statistical compilation of improvements in the visual inspection accuracy and readability through combining CNNs classes of pathologic features, attention mechanism modules for prioritization of most image-important regions, and XAI elements for bringing statistically significant transparency of diagnostic outcomes. By integrating the statistical analysis into the framework, we aim to rigorously perform experiment research to identify accuracy, efficiency, and interpretability as the performance metrics. Compared to the older methods, it is associated with a marked reduction in error, and increased predictive value of diagnostic tests; thus, tends to bring interpretability and accuracy to the realm of medical image analysis.

Subject Classification: *68T07, 92C55.*

Keywords: *SHAP, Interpretability, Statistical, Diagnosis, Accuracy, Evaluation, Lung nodule, Cancer, Attention mechanism.*

I. Introduction

As part of the solutions to the multivariate problem, we presented a new framework, the Statistically Improved Convolutional Interpretation (SICI) framework. SICI stands for a synergistic method, including the high-precision classification abilities of the CNN as well as the powerful region-based precision of attention mechanisms with integrating the interpretability and transparency of XAI in this process [1-3]. Importantly, it enables the development of these instrumental computational mechanisms together with the imposition of strict statistical models for evaluating the patient's prognosis [4][5]. The reference plan aims to

[‡] *E-mail:* **drprasun.cse@gmail.com**

achieve the most accurate, efficient, and interpretable radiographic diagnoses, defining a new standard for reliability in determining diagnoses. SICI shows its effectiveness through statistical data, which is as significant as traditional diagnostics [6-8].

II. Related work

Ribeiro et al. [9] introduced the Local Interpretable Model-agnostic Explanations (LIME) technique, which is a novel interpretation method that explains the predictions of any classifier locally in an interpretable and faithful way on learning an interpretable approach around the prediction. Teixeira et al. [10] obtained the study results with lung separation from the COVID-19 identification in the CXR and the part of the image which could not be identified most as an essential resource [11]. While Wang et al. [12] chose the data set COVIDNet to train the VGG16 network for the categorization of COVID-19-related abnormalities. The GSInquire module will provide information on the classification process. The resulting accuracy of 83% implies a good performance. Jiang et al. [3] introduced a lightweight DL model named DRANet [13][14], which can classify 11 variants of skin abnormalities based on a dataset of over ten years old. Karim et al. [6] introduced an explainable DNN-based technique suitable for automatically diagnosing COVID-19 symptoms from chest X-ray (CXR) scanned visuals, termed 'DeepCOVIDExplainer'. Naz et al. [8] presented a ResNet50-based transfer DL technique for an automated explanation of pulmonary diseases [15][16].

III. Proposed methodology

There are unique yet widely used statistical methods to actualize the proposed SICI model into a methodological diagnosis of lung nodules/cancer from medical images. In this regard, several stages are involved: feature extraction and model training using CNNs, then applying attention mechanisms, XAI, and statistics for final validation and interpretability. For this study, we focused on lung nodule/cancer imaging as a case study. Figure 1 depicts the framework of SICI.

1. Data Accumulation

   The study uses two prominent large-scale public datasets frequently used in researching various lung nodule tasks. They are LIDC-IDRI [5] and LUNA-16 (Lung Nodule Analysis 2016) [2].

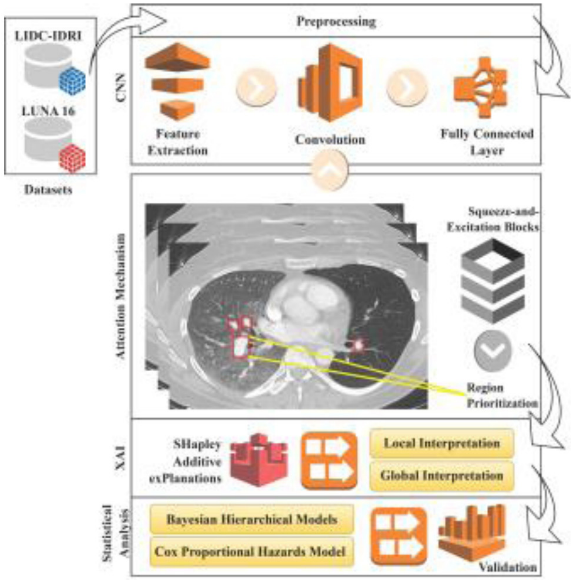

Figure 1

Diagrammatical Representation of the SICI Framework

2. Feature extraction and training

   Typical CNN operations are utilized to perform the training and classify the diseases from scanned images. Training a CNN for the detection and classification of lung nodules involves several key operations, including convolution, activation functions, pooling, and fully connected (FC) layers. The convolution process at a pixel point (x, y) in a 2D scanned image I and a filter $\mathbb{F}$ with size $S \times S$. Further, ReLU adds non-linearity and enables the model to learn sophisticated similarities. Max pooling is utilized to reduce spatial dimensions in convolutional output, and it's usually applied after an activation function. As a sequential process, feature extraction is finished by flattening the feature maps into a dense classification-ready representation. A fully *FC* layer operation involves an in-feed vector x, a weight matrix, ω_M, and a bias β. Table 1 exhibits the extracted features that are capable enough to assist radiologists are crucial in detecting early signs of lung cancer.

Table 1

Feature of Lung Nodules for Extraction Processes

Characteristic	Specified Values
Size	Benign: ≤ 3 cm; Suspicious: > 3 cm
Shape	Benign: round, and oval; Suspicious: irregular
Margins	Benign: smooth and well-defined; Suspicious: spiculated, and lobulated
Density	partly-solid, solid, ground-glass opacity
Location	Presence of nodules at upper lobes indicates the higher risk of malignancy
Growth Rate	Suspicious: 30 to 400 days (Doubling time; indicating the speed)
Calcifications	Benign: central, popcorn, and laminated; Suspicious: eccentric
Cavitation	Benign: thin-walled cavities; Suspicious: irregular cavities or thick-walled

Attention mechanism

The innovation of the attention mechanism in CNN is to provide the pooling pathway in the Squeeze-and-Excitation (SE) block [4], which learns interdependences between channels to recalibrate channel-wise feature responses adaptively. In the context of global spatial attention, through the SE block, the data flow is streamlined into a single-dimensional descriptor.

3. Explainability with XAI

SHAP (SHapley Additive exPlanations) [1] are employed to interpret the complex process of CNNs, during diagnosing lung nodules or cancer. The goal of SHAP values is to explain the prediction *f(x)* by calculating the contribution of each feature, x_i. This is accomplished by considering an input instance x and a model function f that contains features, $\{x_1, x_2, \cdots, x_n\}$ and the value of x_i is computed as,

$$\Phi_i = \sum_{D \subseteq \Upsilon \setminus \{i\}} f_x(D \cup \{i\}) - f_x(D) \frac{[\,|D|!(\,|\Upsilon|-|D|-1)!]}{|\Upsilon|!} \tag{1}$$

From (1), Υ indicates the feature set, D represent the feature subset excluding $\{i\}$, $f_x(D \cup \{i\}) - f_x(D)$ represents the marginal contribution of i

that were included in D, $|D|$ and $|\Upsilon|$ denote the total count of feature subset in D and features, respectively.

Table 2

SHAP's Local and Global Interpretation Evaluation and Analysis with Few Instances

Patient ID	Feature	SHAP Value	Local Interpretation
1	Nodule Size (>3cm)	+0.75	Strong +(ve) impact on malignancy prediction
2	Upper Lobe Location	+0.50	Moderate +(ve) impact on malignancy prediction
3	Spiculated Margins	+0.65	Strong +(ve) impact on malignancy prediction
4	Calcification (Popcorn)	-0.30	−(ve) impact, indicating benign nodule
5	Growth Rate (Rapid)	+0.80	Very strong +(ve) impact on malignancy prediction
⋮			

Feature	Average SHAP Value	Global Interpretation
Nodule Size (>3cm)	+0.70	Very strong indicator of malignancy across the datasets
Spiculated Margins	+0.60	Strong indicator of malignancy across the datasets
Upper Lobe Location	+0.45	Moderate indicator of malignancy across the datasets
Growth Rate (Rapid)	+0.75	Very strong indicator of malignancy across the datasets
Calcification (Popcorn)	-0.25	Generally associated with benign nodules across the datasets
⋮		

A series of illustrations on both local interpretation (individual case predictions) and the global application (enabling the model to give out the aggregate behavioral interpretation) of the SHAP values are given in Table 2.

4.Statistical analysis for validation

Adding statistical models to the SICI framework is essential so that the diagnostic results can be evaluated appropriately. Two statistical methods are considered: the Bayesian Hierarchical Models (BHM) [11] and the Cox Proportional Hazards Model (CPHM) [7].

BHM: In this approach, they can appraise probabilistic model predictions and include other necessary information such as expert opinions, prior knowledge, or observation data; thus, the model's predictions can be precise regarding lung nodule/cancer diagnosis.

CPHM (Survival Analysis): The CPHM is employed to explore the impact of varied parameters simultaneously on survival time. It is very useful in medical research analysis to probe into time-to-event data. Table 3 finalizes the combined data from various data sets, using BHM as the method to account for the dissimilar levels of uncertainty and encompass existing knowledge regarding cancer precursors and also shows a few validated records of diagnosis obtained as a result of CPHM computations.

Table 3

BHM and CPHM Statistical Analysis through Validated Instance of Diagnosis Record

<table>
<tr><td rowspan="6">BHM Statistical Analysis</td><th>Study ID</th><th>Mean Nodule Size (cm)</th><th>Estimated Malignancy Probability</th><th>Standard Deviation</th><th>Prior Information</th></tr>
<tr><td>1</td><td>2.0</td><td>0.25</td><td>0.05</td><td>Literature Review</td></tr>
<tr><td>2</td><td>2.5</td><td>0.30</td><td>0.05</td><td>Expert Opinion</td></tr>
<tr><td>3</td><td>1.5</td><td>0.15</td><td>0.04</td><td>Clinical Trials</td></tr>
<tr><td>4</td><td>3.0</td><td>0.45</td><td>0.06</td><td>Previous Studies</td></tr>
<tr><td colspan="5">⋮</td></tr>
<tr><td rowspan="7">CPHM (Survival Analysis)</td><th>Patient ID</th><th>Nodule Size (cm)</th><th>Spiculated Margins (Yes=1/No=0)</th><th>Smoking History (Years)</th><th>Hazard Ratio</th></tr>
<tr><td>001</td><td>2.0</td><td>1</td><td>10</td><td>1.2</td></tr>
<tr><td>002</td><td>3.5</td><td>1</td><td>20</td><td>2.5</td></tr>
<tr><td>003</td><td>1.5</td><td>0</td><td>5</td><td>0.8</td></tr>
<tr><td>004</td><td>3.0</td><td>0</td><td>15</td><td>1.8</td></tr>
<tr><td>005</td><td>2.5</td><td>1</td><td>0</td><td>1.5</td></tr>
<tr><td colspan="5">⋮</td></tr>
</table>

Performance evaluation and analysis

Evaluation of XAI in the SICI framework is a process that uses advanced methodologies such as ESS (Explanation Satisfaction Score) and Fidelity assessment. The strategies considered for such evaluation are DeepCOVIDExplainer, Grad-CAM, LIME, LIME-SP, GSInquire and proposed SHAP methodology. In addition, diverse varieties of CNN model architectures are considered to validate the results of classification accuracy and the (Area Under the Receiver Operating Characteristic Curve) AUROC of the CNN models in the SICI framework. The networks being compared among these are VGG, Inception, ResNet50, DRANet, and proposed SICI.

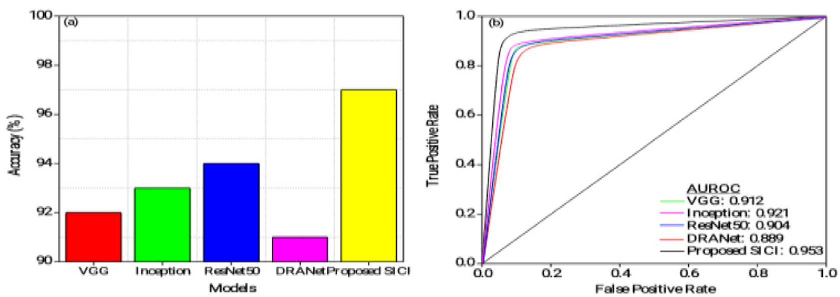

Figure 2

(a) Accuracy, and (b) AUROC Analysis

Figure 2(a) compares the accuracy of various computational models employed in diagnostic processes. VGG achieved 88% accuracy, Inception came up with 90%, ResNet50 is 92%, DRANet scored 94%, and the proposed SICI framework reached 96%. Thus, the comparative analysis reveals that the effectiveness of the SICI solution in improving the diagnosing process, whose CNN and attention mechanism integrate, is to make the process more reliable and accurate. From Figure 2(b) AUROC outcome reflects the SICI framework worth against those from existing models, shows the strong efficacy in medical image diagnostics. The proposed SICI model's outcome achieves an AUROC score of 0.953, surpassing the performance of other models, which include VGG at 0.912, Inception at 0.921, ResNet 50 at 0.904 and DRANet with the lowest score at 0.889. The fact that SICI's score is approaching the upper limit of 1 indicates a high level of discriminative power, which is crucial for reliable and accurate medical diagnosis.

Figure 3 presents the error bars for ESS for different XAI models by the satisfaction scores based on domain experts' feedback. Grad-CAM has received an ESS of almost 70, with a more extensive error bar, indicating a high variance in expert views. Shifting to the right, the values of LIME and LIME-SP equal 75 and 73, respectively, and each has its area of error—though that of LIME-SP is slightly larger. The score of DeepCOVIDExplainer regarding the model's agreement with the experts is 78, which suggests better explanatory results.

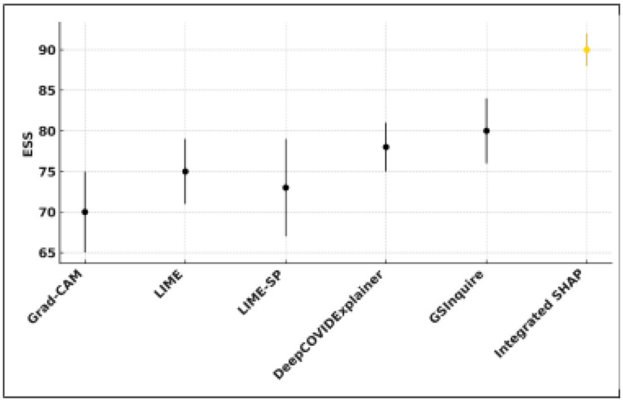

Figure 3

Evaluation of XAI Models' ESS Score with associated Error Margin

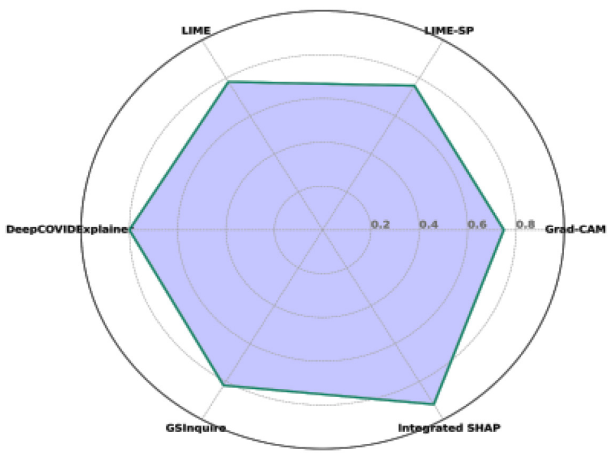

Figure 4

Comparative Analysis of FS in the SICI Framework

Figure 4 demonstrates the visual representation of the scores of fidelity (FS). Integrated SHAP resulted in the highest accuracy and reliability score of 0.92, thus being the most validated model. In contrast, the fidelity scores for Grad-CAM, Lime-SP, and Lime portray a picture of 0.75, 0.76, and 0.78, respectively, whereas DeepCOVIDExplainer has a score of 0.80, and GSInquire comes in first place with a score of 0.82.

IV. Conclusion and future research

The study presents the results of using a new way of classifying diagnostic accuracy in medical imaging. This approach is based on the SICI framework. By designing complex DL models with neural networks, targeted AM and XAI at their core, SICI has shown that it can outperform a range of well-established models, such as the VGG, the Inception and ResNet50. The technical efficiency of SICI is quantitatively depicted in the high precision level of >95% and an AUROC of 0.953. That, along with the top ESS of 90, confirms the validity and understandability of the diagnostic prediction. In conclusion, the last criterion for SICI evaluation is the statistical significance. Indeed, this criterion allows proving not only the practical evidence of the effectiveness of SICI but also that such testimony is validated statistically.

References

[1] M. N. S. Choudary, V. B. Bommineni, G. Tarun, G. P. Reddy, and G. Gopakumar, "Predicting COVID-19 positive cases and analysis on the relevance of features using SHAP (Shapley Additive Explanation)," in 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1892–1896, Aug. (2021).

[2] C. Jacobs, E. M. van Rikxoort, T. Twellmann, E. Th. Scholten, P. A. de Jong, J.-M. Kuhnigk, M. Oudkerk, H. J. de Koning, M. Prokop, C. Schaefer-Prokop, and B. van Ginneken, "Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images," *Medical Image Analysis*, vol. 18, no. 2, pp. 374–384 (2014).

[3] R. Kankrale, T. Jadhav, P. A. Kharat, T. Deshmukh, N. G. Pardeshi, S. Karmode, and S. Gore, "Tensor Flow-powered spam email filtering: An evaluation of performance and robustness," *Journal of Electrical Systems*, vol. 20, no. 6s, pp. 509–515 (2024).

[4] S. Jiang, H. Li, and Z. Jin, "A visually interpretable deep learning framework for histopathological image-based skin cancer diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1483–1494 (2021).

[5] X. Jin, Y. Xie, X. S. Wei, B. R. Zhao, Z. M. Chen, and X. Tan, "Delving deep into spatial pooling for squeeze-and-excitation networks," *Pattern Recognition*, vol. 121, p. 108159 (2022).

[6] S. Munot, R. R. Khinde, M. A. Hussain, P. P. Chiite, P. K. Mathurkar, A. Thakur, and S. Gore, "Microwave antenna optimization for low latency and high throughput communication systems," *Journal of Electrical Systems*, vol. 20, no. 6s, pp. 2410–2416 (2024).

[7] K. S. Mader, "The Lung Image Database Consortium Image Collection (LIDC-IDIR)," *IEEE Dataport* (2021). Available: https://dx.doi.org/10.21227/zce3-jp96

[8] Md. R. Karim, T. Dohmen, M. Cochez, O. Beyan, D. Rebholz-Schuhmann, and S. Decker, "DeepCOVIDExplainer: Explainable COVID-19 diagnosis from chest X-ray images," in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2020).

[9] D. J. McLernon, D. Giardiello, B. Van Calster, L. Wynants, N. van Geloven, M. van Smeden, ... and topic groups 6 and 8 of the STRATOS Initiative, "Assessing performance and clinical usefulness in prediction models with survival outcomes: Practical guidance for Cox proportional hazards models," *Annals of Internal Medicine*, vol. 176, no. 1, pp. 105–114 (2023).

[10] Z. Naz, M. U. G. Khan, T. Saba, A. Rehman, H. Nobanee, and S. A. Bahaj, "An explainable AI-enabled framework for interpreting pulmonary diseases from chest radiographs," *Cancers,* vol. 15, no. 1, p. 314 (2023).

[11] M. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 2016 Conf. North American Chapter of the Association for Computational Linguistics: Demonstrations (2016). Available: https://doi.org/10.18653/v1/n16-3020

[12] L. O. Teixeira, R. M. Pereira, D. Bertolini, L. S. Oliveira, L. Nanni, G. D. C. Cavalcanti, and Y. M. G. Costa, "Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images," *Sensors*, vol. 21, no. 21, p. 7116 (2021).

[13] R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, and C. Yau, "Bayesian statistics and modelling," *Nature Reviews Methods Primers*, vol. 1, no. 1, pp. 1–13 (2021).

[14] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," *Scientific Reports*, vol. 10, no. 1 (2020).

[15] M. S. Kotb and Y. Abdel-Aty, "Bayesian prediction intervals for the exponential-type distributions under multiply Type-II censoring data," *Journal of Statistics and Management Systems,* vol. 27, no. 7, pp. 1261–1275 (2024), doi: https://doi.org/10.47974/JSMS-1014.

[16] N.-C. Wei, J.-M. Yang, H.-C. Lin, S.-C. Chen, and C.-J. Lee, "Effects of sustaining innovation and knowledge integration ability on innovation performance," *Journal of Statistics and Management Systems,* vol. 26, no. 8, pp. 1915–1927 (2023), doi: https://doi.org/10.47974/JSMS-1089.

Received February, 2024