Decoding AI: Transparent Models for Understandable Decision-Making

Satyanarayan Kanungo

Independent Researcher, USA

***Abstract*:** In this paper, Explainable AI methods have been presented for transparent decision-making in medical field analysis scenarios. It has used three different explainable methods and applied those in the data set of medical images with an aim to enhance the decision-making comprehensiveness given by the CNN (“*Convolutional Neural Network”*). It has used two ML(Machine Learning) methods such as LIME (“Local Interpretable Model-Agnostic Explanation”) and SHAP (“*SHapley Additive exPlanations”*) with an alternative approach of explanation, the CIU (“*Contextual Importance and Utility”*) method. Furthermore, it has assessed explanations by evaluation of the human and conducted user studies built on explanations by SHAP, CIU and LIME. A set of tests have been carried out in the setting of a web-related survey and stated the understanding and explanation of the explanations. It has also quantitatively analysed three groups of users where (n=20,20,20) with three diverse explanation forms. It has also identified notable differences in the decision-making of humans between various settings of explanation support.

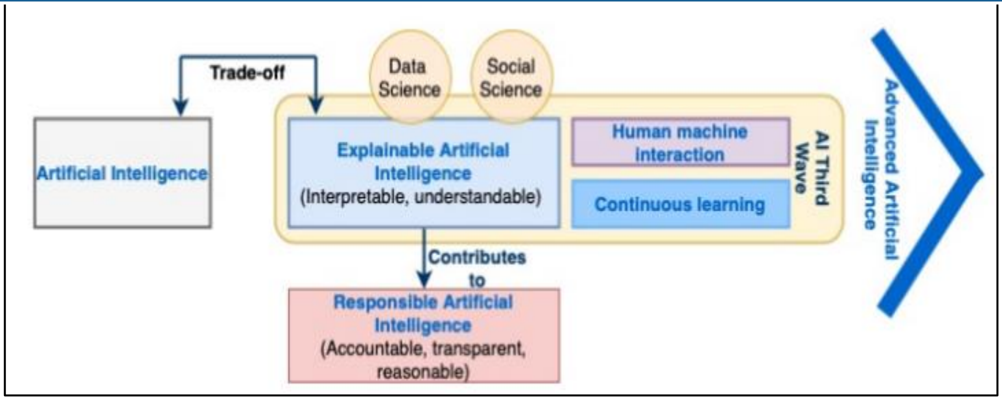
***Keywords*:** image recognition, artificial intelligence, expandable AI, medical image

Introduction

In conventional diagnostics, this paper captured some images of possible lesions checked in a medical setting by one of the doctors. This approach has been time-consuming and dependent on the doctors’s attention, examined for thousand times. AI and deep learning have been used to extract the data from several images and interest in the field has been growing in finance, forensics, education, scientific research and medical diagnostics. The use of CNN network has been used to produce higher accuracy results with comparison to standard approaches. Through the utilisation of ML techniques, the method of lesion detection has been automated with accuracy. The use of XAI (Explainable Artificial Intelligence) can be helpful to foster trust and transparency [1]. It can be helpful to clarify all the decisions initiated by black-box model to establish it intuitive for humans. The below figure has diagnosed additional explanations and the way it can enhance the truthfulness of these methods. This paper has used two ML(Machine Learning) methods such as LIME (“Local Interpretable Model-Agnostic Explanation”) and SHAP (“*SHapley Additive exPlanations”*) with an alternative approach of explanation, the CIU (“*Contextual Importance and Utility”*) method. Furthermore, it has assessed explanations by evaluation of the human and conducted user studies built on explanations by three models.

Background

Methods of Explainable AI have been developed to achieve high transparency to produce AI system explanations. It has investigated various approaches to interpreting autonomous systems and making it understandable to all humans [2]. It has helped to evaluate the limitations and strengths of the ML model and facilitate understandability. It has used one approach such as an explanation of post hoc to extract useful data on the process of black-box model. The main goal of the XAI model has been creating explainable models to maintain strong learning efficiency. The picture below has discussed the basic concept of the XAI as it contributes to responsible AI.


Trade-off
Data Science
Social Science
Artificial Intelligence
Explainable Artificial Intelligence
(Interpretable, understandable)
Human machine interaction
Continuous learning
AI Third Wave
Advanced Artificial Intelligence
Contributes to
Responsible Artificial Intelligence
(Accountable, transparent, reasonable)

XAI methods

LIME

The LIME (“*Local Interpretable Model-Agnostic Explanation”*) model has been developed to help its users by establishing explanations for the decisions of the black-box model in some instances [3]. The explanation of LIME has been based on the behaviour of the classifier models and that can be decision tree or linear regression, identified in the equation.

In this model, x has been used as instance being and its explanation has been considered the fidelity term maximisation. f has represented the model of black-box and it has explained by an explainer and represented by the sign g. It has tried to match all information in the vicinity prediction that must be explained.

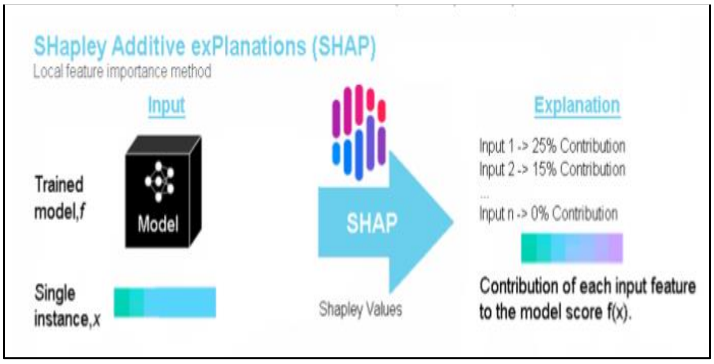
$$explanation(x) = argmin_{g \in G} \rfloor (f, g, \pi_x) + \Omega(g)$$

SHAP

This paper has used the SHAP model to generate the AI explanations and examined the Deep Explainer and Gradient Explainer of SHAP. The algorithm of Karnel SHAP has been chosen to estimate the SHAP values and it provided the best outcomes despite being slow [4]. The SHAP model has been used to explain predictions and it use the coalitions concept to computer the features of Shapley Value for predicting (x) by the model of Black-box (f). It has calculated marginal contribution and it has been stated in the below picture.

$$\phi_j(x) = \frac{1}{M} \sum_{m=1}^{M} \phi_j^m$$

$$\phi_j^m = \hat{f}(x_{+j}^m) - \hat{f}(x_{-j}^m)$$


SHapley Additive exPlanations (SHAP)
Local feature importance method
Input
Trained model,f
Model
Single instance,x
SHAP
Shapley Values
Explanation
Input 1 -> 25% Contribution
Input 2 -> 15% Contribution
...
Input n -> 0% Contribution
Contribution of each input feature to the model score f(x).

CIU

This paper has explained the method of CIU built on the utility of the feature which has been useful for future prediction [5]. This model has used two essential evaluation methods and one of them is the model of Contextual Importance and the other one is the contextual utility. The model of CIU is different from SHAP and LIME as it does not use the model of an intermediate surrogate to make linearity assumptions. The Contextual Utility and the contextual Importance have been used to generate the interpretation and explanation built on the features of the set of data. The below picture has described the mathematical calculations of CU and CI,

$$CI_j(\vec{C},\{i\}) = \frac{cmax_j(\vec{C},\{i\}) - cmin_j(\vec{C},\{i\})}{absmax_j - absmin_j}$$

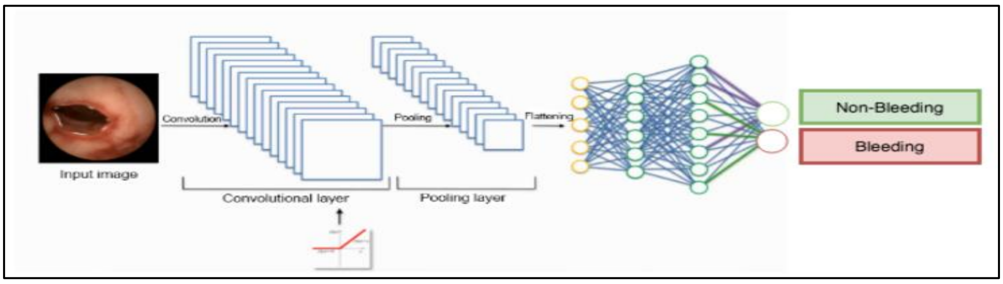
$$CU_j(\vec{C},\{i\}) = \frac{out_j(\vec{C}) - cmin_j(\vec{C},\{i\})}{cmax_j(\vec{C},\{i\}) - cmin_j(\vec{C},\{i\})}$$

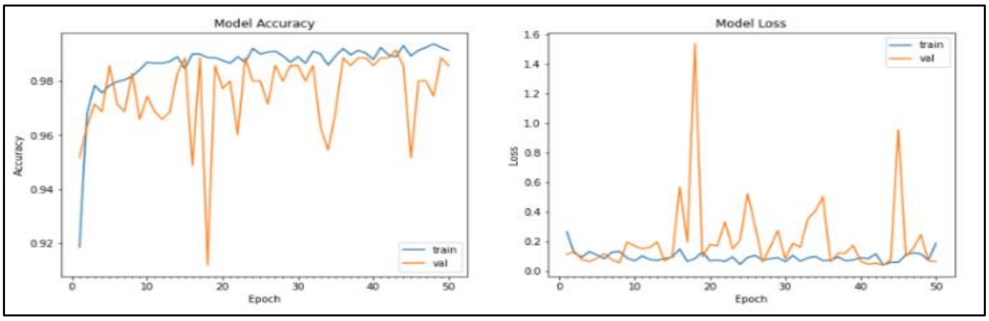
Here the CI has provided the inputs (i) for a specific output (j) in the C. context.*absmax*is considered the maximum value for the possible output and it absmin has been considered the minimum value for the possible output.

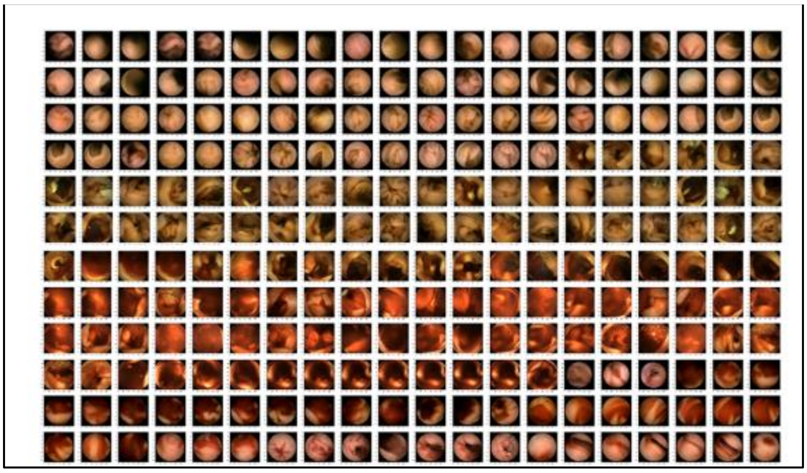
Methodology

“***The Image Data Set***” has been considered in this paper by the use of Video Capsule Endoscopy (VCE) to visualise the entire gastro tract of the patients. The main aim of this process is to identify red lesions segments in a small bowel, a key organ where bleeding occurs from unknown reaseons. More then 3200 images have been available in the data set of this Endoscopy process. This paper has aimed at 3200 images and among them, only 10% has been used for testing and the rest for training [6].

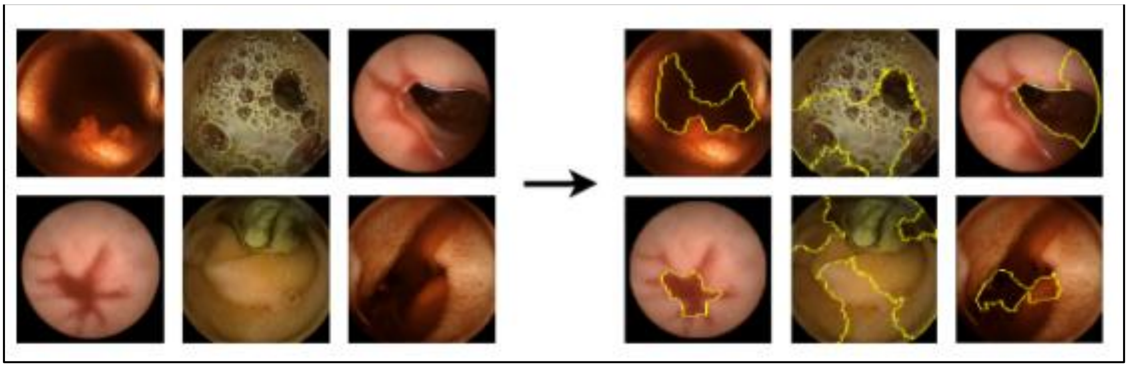
It has also discussed the ***implementation process*** of the ***Black-Box Model*** and it has split the labels and data into the validation sets and training which are assigned randomly. These images have represented the situation in the medical application, shown below figure and it has included both non-bleeding and bleeding examples. All the 3200 pictures have been resized to pixels of 150 * 150 for accurate and faster communication [7]. It has achieved an accuracy of more than 98% in the validation. It has trained the CNN model built on labels and these were created utilising the annotated images of respiratory as the reference point.







This paper has included setting without the explanations to use the explainable AI methods. It has decided to utilise three different methods such as LIME model, SHAP model and CIU model [8]. The implementation of SHAP and LIME methods that has been done on Triton computing and on the other hand, the explanations of CIU were generated by the application of RStudio. In the below picture, it has describe the explanation of LIME model and marks the field contributing to non-bleeding scenarios [9].



Design of user study

This paper has used the explianble AI algorithm to generalise it and invited users in the place of medical profesionals. The the users have completed a line of tests in the web-related survey and showcased the experience of the users. Itahs gathered users from the academic environment and it has collected the information from 60 usrers (n=60), 20 usrs in each group [10]. It has performed test from 20 users without any explanation with the support of LIME explanation, it has also performed test from 20 years without any explalantion with the support of SHAP explanation and rest tests have been conducted with 20 users with the support of CIU model [11].

Models	Total	Gender		Degree			XAI Understanding	
		Female	**Male**	**PhD Degree**	**Master's Degree**	**Bachelor 's Degree**	**Yes**	**No**
LIME	20	6	14	3	12	5	12	8
SHAP	20	7	13	6	12	2	8	12
CIU	20	7	13	5	9	6	9	11

Table 1: Demographics of study participants

Analysis and result

This paper has compared the time required the to initiate the explanations and the performances at the time of generating explanations with those three explainable AI methods. The model of LIME needed 11seconds per image and more than 5 minutes and it also needs 20 seconds for around 28 images with the sample_numb equals to 2500 and features_number equals to 10. It has needed around one hour and 45 minutes to generate all the exoalnations on the validation pictures (354). The application of SHAP needed more approximately 10 seconds per pictures andapproximately 4 min and 35 seconds for around 28 pictures with the sample-numbers equals to 3000 [12]. Likewise, the CIU model needs 8.5 seconds for in image, approximately 4 minutes for 28 pictures, and near by 1 hour and 18 minutes for those 354 images.

		LIME	SHAP	CIU
Time Comparison	1 Picture	11.4 s	9.8 s	8.5 s
	354 pictures	1 h 45 min	1 h 30 min	1 h 18 min
	28 pictures	5 min 20 s	4 min 30s	4 min

Table 2: Time Comparison

Quantitative analysis

This paper has used statistics to assess the data and exploratory statistics after composing three different user groups; LIME, SHAP and CIU explanations [13]. It has first examined the information from the setting of three user and investigated the dberseness between medians and means of the human decision-making.

	Measures	Lime User Study	
		With Explanation	Without Explanation
Correct	Median	14.5	15.00
	Mean	13.15	12.90
Incorrect	Mean	1.80	2.05
	Median	1.50	1.00

Table 3: LIME users mean and median value

	Measures	SHAP User Study	
		With Explanation	Without Explanation
Correct	Median	15.00	14.00
	Mean	12.40	13.05
Incorrect	Mean	2.60	1.85

	Median	1.00	2.00

Table 4: SHAP users mean and median value

	Measures	**CIU User Study**	
		Without Explanation	**With Explanation**
Correct	**Mean**	14.30	15.90
	Median	13.00	16.00
Incorrect	**Mean**	1.70	1.00
	Median	2.00	0.00

Table 5: CIU users mean and median value

	Measure	**LIME**	**SHAP**	**CIU**
Correct Explanations and incorrect Explanations	**Mean**	8.55	8.65	10.35
	Median	9.50	9.50	11.00

Table 6: Correct Explanations and incorrect Explanations

Table 3-5 has shown the mean data and median data of the incorrect and correct decisions for every explanation type as the setting of non-explanation for all the three user examinations. It has faced notable differences in the mean and media valus of user decision-making procedure [14] ***Table 6*** has shown differences in the relation means in the understanding of users explanations built on incorrect explanations derived fro the correct ones.

Discussion

This paper has assessed notable diversity in the decision-making of he human between three user groups worked with diverse methods of explanation support [15]. The results in the paper has suggested that the model of CIU has been more essential to the users to establish the correct decisions thn other two SHAP and LIME models. It can be concluded from the result section that the explanations of the CIU generated model were more clear to the selected users and it provided strong support in the process of decision-making [16]. It can also be stated from the results that the users have supported the CIU explanation as it takes small time to finish the user study in comparison with SHAP and LIME explanation. The results in the paper also shown insight about the questions concerning the utilisation of the XAI methods. In the two studies out of three examinations, it can be sated the users performed exceptionally with the support of the explanation [17]. The method of SHAP has provided low correct answers in comparison with others relatively.

Limitation and future work

Limitation

This paper has many limitations though it has provided the evaluation of AI methods to support the decision-making process of humans in the medical domain.

- The focus of the present study has been limited to the angle medical set of data. The present utilisation of explanation support has mainly focused on complex cases in the medical field for various diagnoses [18].
- It needs to expand the present scope and apply the explanatory data in real-life circumstances, it can use the information in the scenarios of the real world and it can facilitate the practical application.

Future work

- Improvement of the methods can generalise the provided explanations by using diverse medical set of data and it can provide greater support for the medical experts.
- The participant number in the paper has been limited to 60 and it should test the methods with more number of users to produce more accurate statistical test outcomes [19].

Conclusion

From the above discussion it can be concluded that it has explained the use of three diverse explainable method and how it can be obscured to diverse data set of medical domain and provide significant support in decision-making in the medical field. It has discussed SHAP, LIME and CIU explainable AI methods to suggest the notable diversification in decision-making of humans, with the CIU method shown as the best support for decision-making [20]. It has used application-related evaluation and a clear idea of how to apply it in the processing of medical images. It has also discussed the limitations of the paper as well as its future recommendations.

References

[1] Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S. and Turini, F., 2019. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, *34*(6), pp.14-23.

[2] Nimri, R., Battelino, T., Laffel, L.M., Slover, R.H., Schatz, D., Weinzimer, S.A., Dovc, K., Danne, T. and Phillip, M., 2020. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. *Nature medicine*, *26*(9), pp.1380-1384.

[3] Kim, B., Park, J. and Suh, J., 2020. Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, *134*, p.113302.

[4] Jaigirdar, F.T., Rudolph, C., Oliver, G., Watts, D. and Bain, C., 2020, December. What information is required for explainable AI?: A provenance-based research agenda and future challenges. In *2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC)* (pp. 177-183). IEEE.

[5] Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S., 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), p.18.

[6] Li, X.H., Cao, C.C., Shi, Y., Bai, W., Gao, H., Qiu, L., Wang, C., Gao, Y., Zhang, S., Xue, X. and Chen, L., 2020. A survey of data-driven and knowledge-aware explainable ai. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), pp.29-49.

[7] Roscher, R., Bohn, B., Duarte, M.F. and Garcke, J., 2020. Explainable machine learning for scientific insights and discoveries. *Ieee Access*, *8*, pp.42200-42216.

[8] Kumar, T.S., 2020. Data mining based marketing decision support system using hybrid machine learning algorithm. *Journal of Artificial Intelligence*, *2*(03), pp.185-193.

[9] Spinner, T., Schlegel, U., Schäfer, H. and El-Assady, M., 2019. explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics*, *26*(1), pp.1064-1074.

[10] Tjoa, E. and Guan, C., 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, *32*(11), pp.4793-4813.

[11] Samek, W. and Müller, K.R., 2019. Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, pp.5-22.

[12] Fernandes, M., Vieira, S.M., Leite, F., Palos, C., Finkelstein, S. and Sousa, J.M., 2020. Clinical decision support systems for triage in the emergency department using intelligent systems: a review. *Artificial Intelligence in Medicine*, *102*, p.101762.

[13] Malhi, A., Knapic, S. and Främling, K., 2020. Explainable agents for less bias in human-agent decision making. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers 2* (pp. 129-146). Springer International Publishing.

[14] Zhang, K., Xu, P. and Zhang, J., 2020, October. Explainable AI in deep reinforcement learning models: A shap method applied in power system emergency control. In *2020 IEEE 4th conference on energy internet and energy system integration (EI2)* (pp. 711-716). IEEE.

[15] Kim, B., Park, J. and Suh, J., 2020. Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, *134*, p.113302.

[16] Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S. and Turini, F., 2019, July. Meaningful explanations of black box AI decision systems. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 9780-9784).

[17] Tyler, N.S., Mosquera-Lopez, C.M., Wilson, L.M., Dodier, R.H., Branigan, D.L., Gabo, V.B., Guillot, F.H., Hilts, W.W., El Youssef, J., Castle, J.R. and Jacobs, P.G., 2020. An artificial intelligence decision support system for the management of type 1 diabetes. *Nature metabolism*, *2*(7), pp.612-619.

[18] Fitriyani, N.L., Syafrudin, M., Alfian, G. and Rhee, J., 2020. HDPM: an effective heart disease prediction model for a clinical decision support system. *IEEE Access*, *8*, pp.133034-133050.

[19] Satish, Karuturi S R V, and M Swamy Das. "Review of Cloud Computing and Data Security." IJAEMA (The International Journal of Analytical and Experimental Modal Analysis) 10, no. 3 (2018): 1- 8.

[20]. Satish, Karuturi S R V, and M Swamy Das. "Multi-Tier Authentication Scheme to Enhance Security in Cloud
Computing." IJRAR (International Journal of Research and Analytical Reviews) 6, no. 2 (2019): 1-8.

[20] Xu, Feiyu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. "Explainable AI: A brief survey on history, research areas, approaches and challenges." In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pp. 563-574. Springer International Publishing, 2019.

[21] Srivastav and S. Mandal, "Radars for Autonomous Driving: A Review of Deep Learning Methods and Challenges," in IEEE Access, vol. 11, pp. 97147-97168, 2023, doi: 10.1109/ACCESS.2023.3312382.

[22] Das, S., Agarwal, N., Venugopal, D., Sheldon, F.T. and Shiva, S., 2020, December. Taxonomy and survey of interpretable machine learning method.In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 670-677). IEEE.