PriMera Scientific Surgical
Research and Practice
Volume 2 Issue 1 July 2023
DOI: 10.56831/PSSRP-02-041
ISSN: 2836-0028

PriMera Scientific Publications

An XAI Model for Malignancy Detection of the Pulmonary Nodules: Building Trust by Reducing AI Risk

Type: Research Article
Received: June 21,2023
Published: June 27,2023

Citation:
Mahua Pal. "An XAI Model for Malignancy Detection of the Pulmonary Nodules: Building Trust by Reducing AI Risk". PriMera Scientific Surgical Research and Practice 2.1 (2023): 20-29.

Mahua Pal*

Department of Sciences and Commerce, J. D. Birla Institute, Kolkata, India

***Corresponding Author:** Mahua Pal, Department of Sciences and Commerce, J. D. Birla Institute, Kolkata, India.

Abstract

This is the era of Artificial Intelligence, where AI models are growing fast for making critical decisions for several predictive models. Here AI is giving solution to the medical practices for diagnosis as well as prognosis with its increasing intelligence to simplify the ambiguity and complexity in data to carry out clinical decisions. Several research studies have truly demanded the need of AI-based systems and how to enhancing their capabilities to help medical practitioners. However, instead of giving highest effort for making most accurate AI based models, still now assessing the magnitude and impact of human trust on AI technology demands substantial attention. In the last decade many AI based CAD models were developed which hardly could persuade the experienced medical practitioners to accept the machine-specified decisions. In this research work, it was attempted to interpret and explain a supervised AI model built using XGBoost on Lung cancer detection by using XAI (Explainable AI) tools - two post-hoc methods (LIME and SHAP) and one ante-hoc method, to provide satisfactory explanations to medical practitioners, thereby minimize the AI risk factor in implementation of the model and reinforce the trust to the medical experts and patients in accepting such model. In this paper, the results of all three XAI tools were illustrated using heatmaps to select important input bi- omarkers that contributed more in detection of the benign or malignancy state of the pulmonary nodules. Finally the supervised AI model was rebuilt using only those important input features and it was found out that the metrics like specificity, precision, the AUC of the newer model under the ROC curve were giving better result in prediction of lung cancer nodule state. It is a study to explore how XAI tools highlight the contributions of input features in an AI model and how that AI model's performance can be fine-tuned based on the outputs of XAI mechanism.

Keywords: AI; LIME; Pulmonary nodules; SHAP; XAI; XGBoost

Introduction

Modern intelligent systems such as image based medical diagnostic AI systems are becoming the most accurate techniques for predicting cancer. But these AI models due to its unexplainable black box problem lack transparency and create a barrier for clinical implementation due to trust issue. Resolving this issue, recent development in the area of XAI shows the way to overcome the problem of deep learning architectures by providing after the fact explanations. XAI output is not only used for validating the prediction, but also recommends some correction in the prediction model and explores some new potential biomarkers. These explanations can be local, global post-hoc, ante-hoc, visual and textual. Many research works have been conducted using CNN, DNN models in the medical fields such as dermatology, ophthalmology, radiology, etc. Few AI models [6] were also developed using different ML algorithms in diagnosing lung-cancer where pulmonary nodules are the abnormal growths in one or both lungs that show up on CT scans. Lung cancer is the foremost contributor to cancer-related mortality, resulting in 1.3 million cancer deaths per year worldwide [3]. Due to the lack of the adequate number of qualified medical doctors and other healthcare professionals in India, there is a major concern and associated burden toward the cancer morbidity and mortality. Here AI based diagnosis CAD models could be acting like an Augmented Doctor [1], an assistant system to the clinical experts and medical practitioners where medical practitioners control the process. Henceforth, this CAD model is a Human-In-The-Loop (HITL) model [7] where interpretations could be made afterwards (post-hoc) or interpretations could be derived in their architecture (ante-hoc). Usually an ante-hoc method based AI model suffers from lower accuracy than that of post-hoc methods, as its modeling capacity may be limited due to the architectural restriction.

An XAI model could be utilized for the preliminary investigation of the states of pulmonary nodules by identifying the reasons as explanations to doctors and patients. Developing an XAI model using SHAP and LIME tools in detection of lung cancer from the CT scan reports of patients was a new attempt that could assist the medical practitioners in their first step of the lung cancer medical treatment process. With this perspective, this XAI model was developed using a binary AI classification algorithm - XGBoost and two XAI tools were applied on the top of this classifier for global and local post-hoc interpretation. The ante-hoc explanation feature of the tree explainer 'XGBoost' was also taken into account for the explanations of the malignancy of the pulmonary nodules. This intelligent automation could reduce the stress of preliminary examination and detection process taken by the healthcare system to identify the malignant cases and thus medical experts' team can concentrate on the emergency cases.

Many a time the experimental results of AI Model looks uncertain for unrelated data features of an image to detect pulmonary nodules besides providing the correct prediction and sometime there may be serious contrariety to give proper explanation in how different deep learning models gives the same prediction. Our proposed model focuses that XAI can provide trust in AI systems through it's an elevated predictability and understandability. We have arranged this research paper in the next few sections in the following order with the background motivation of this work along with some literature reviews in Section 2, proposed model with methodology in Section 3, implementations and result discussions in Section 4 and finally the conclusion with future work in Section 4, respectively.

Background

The importance and different methods of explainabi lity for this kind of AI intelligent model for detection of the disease such as lung cancer, based on medical images thoroughly discussed in this section. Visual heat maps are the most popular methods for explaining and interpreting image-based classification. Local explanations are useful for explaining a single data sample at a time as an assistive diagnosis system, whereas global-level explanations are useful for understanding the model performance as a whole to identify decision biases. Class Activation Mapping (CAM) [13], GRAD- CAM [14], gradient-based methods like Saliency, Integrated Gradient, DeepLift [16] or methods based on mathematical decomposition like Layerwise-Relevance Propagation (LRP), Agglomerative Contextual Decomposition (ACD) and SHapley Additive exPlanations (SHAP) [2, 15] methods access the model parameters to understand the attribution of the classification results to the individual pixels and the model architecture. Model-agnostic perturbation-based methods can be used for model independent explanation without knowledge of their internals. LIME [18], Occlusion, RISE, and Extremal Perturbation are the examples which differ in the occlusion strategy (procedure and perturbation) [8].

Ahmed et al. [9] applied XAI techniques in stack ensemble ML framework which was built using generalized linear model (GLM), random forest (RF), Gradient boosting machine (GBM), extreme Gradient boosting machine (XGBoost), and Deep Neural Network (DNN) and visualized the risk factors of lung and bronchus cancer (LBC) mortality from the stack ensemble model's output in global and local scales after considering the input features such as air-pollution, socio-economic status and etc. Siddhartha et al [10] proposed an XAI model to predict the postoperative life expectancy in the lung cancer patients after surgery by using SHAP and LIME techniques on the top of ML model built using Random Forest algorithm. They had examined the data of those patients who were already qualified for surgery after the detection with lung cancer. Bartczak et al. [11] proposed another XAI model to predict the postoperative life expectancy in the lung cancer patients after surgery by using SHAP and Ceteris Paribus methods on the top of ML model built using hyperparameters cross validation and logistic regression. Venugopal et al. [12] had developed a model with 20-layer deep residual CNN and applied occlusion technique on a nodule and thus clinical attribution heatmaps on the nodules provided aid to the radiologists to identify features aiding classification. So far, very insignificant amount of research works have been conducted to enforce trust in the healthcare system to accept AI CAD model required for preliminary automated investigations of lung cancer malignancy and the predictions with minimal AI risk. These models won't replace doctors, but facilitate them to provide faster and better service. Henceforth, there need a lots of research work, discovery of new explainable algorithms and fine tunings which motivated us to work in this field. The XAI Model, in this paper explained the reasons while predicting the state of cancerous cells in lungs through heat maps by checking the contributions of the input biomarkers obtained from CT scan report of a patient and this XAI model was fine tuned by rebuilding it using the important input biomarkers and prediction was verified with the help of the outputs of SHAP and LIME methods.

Proposed Model with Methodology

In this research work we propose an XAI based CAD model that uses SHAP and LIME techniques on the top of an XGBoost classifier to classify benign and malignant state of lung cancer. The main goal of this work was to employ AI based techniques to detect infectious pulmonary nodules and to use XAI to understand and compare the strategy of each model to increase trust. Accordingly XAI techniques were applied to detect lung cancer using the knowledgebase of radiological biomarkers from LIDC-IDRI's diagnostic data and lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions [17]. The Flow diagram of proposed model is given Figure 1.

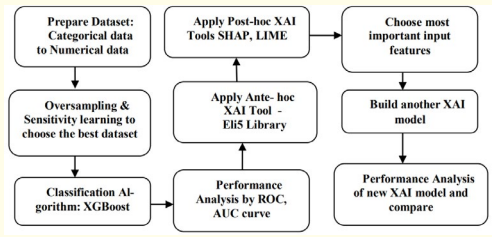

Figure 1

The XAI techniques were applied to interpret the predictions made by the XGBoost classification model. Finally, based on the highlighted features visualized in the heatmaps, the important biomarkers were selected that had higher influence to predict the malignancy of pulmonary nodules and unnecessary biomarkers were excluded to build another model that could speed up the execution time of the model.

The Lung Image Database Consortium image collection (LIDC-IDRI) provides a web-accessible international resource for development, training, and evaluation of intelligent CAD models for lung cancer detection and diagnosis. In this study, 1307 data which was released in October 2011 was investigated. This dataset contained images from a clinical thoracic CT scan and an associated XML file that recorded the results of a two-phase image annotation process performed by four experienced thoracic radiologists. In each CT scan, the marked lesions belong to one of three categories.

nodule > or =3 mm	***nodule <3 mm***	***non-nodule > or =3 mm***
May be Malignant	Could not be annotated; Not Malignant	Not Malignant

Features listed in the dataset for conducting this work with categorical values were:

Calcification, Internal _Structure, Lobulation, Margin, Sphericity, Spiculation, Subtlety, Texture, Number_of _Nodules, Nodule_Size_ More_ than_3mm, Malignancy. We classified Malignancy into two states – Malignant (High-Risk) and Benign (Low-Risk). The annotated grading 1 and 2 under Malignancy, featured in the pylidc documentation [4] were considered as Benign class and the grading 3, 4 and 5 were considered as Malignant class. After referring to the pylidc documentation, we arranged the dataset with categorical values under each input feature attributes. The categorical data were converted to numerical. Since the dataset had 493 Low-risk / Benign data and 814 High-risk / Malignant data, the problem related to the imbalance classes was handled by comparing the original dataset performance with the over- sampled dataset performance. The oversampling of the dataset was done using the ADASYN, ROS and SMOTE algorithm separately, and finally the ADASYN method was chosen to balance the dataset. The 80% data were used for training the model and 20% data were used for testing. XGBoost classifier supports binary classification. After oversampling and sensitivity analysis, these ten different features or biomarkers were fed into the supervised binary classification AI model built using XGBoost tree- explainer classifier. The SHAP and LIME XAI tools were applied on the top of this classification model to highlight the important features with their corresponding risk factors from their visual heatmaps. Thus an XAI model was produced. The important input features which contributed more to the prediction of malignancy have been taken into account and that AI model has been reconstructed (the classifier algorithm remained same, i.e., XGBoost classifier). Thus the first model got fine tuned as the performance of this reconstructed model was found better than that of the previous one. SHAP and LIME tools were again applied to validate the model prediction by observing the output of SHAP and LIME methods and these interpretations were found consistent with biological intuition which enforce trust to the system.

Implementation & Results Discussions

The performance of the XGBoost AI model for detecting lung cancer at its initial stage was illustrated in this section. XGBoost [5] is a popular scalable machine learning system for tree boosting. DNN is a state of the art model which is popularly used in medical diagnosis and needs a humongous amount of data to show their relevance, so we preferred the XGBoost model for this problem with 1307 data. The coding of the flow of tasks (Fig 1) was done using Python language in Google Colaboratory platform which provides high-end supports. This dataset had an imbalance-dataset problem which was resolved by using an oversampling algorithm. SMOTE (Synthetic Minority Oversampling Technique), ADASYN (Adaptive Synthetic) are the methods that generate synthetic data. The XGBoost classification model's performances with those oversampled data were separately evaluated based on the different metrics such as AUC value, F1 Score, precision and specificity etc. We used 5 fold cross validation technique here. Area under the ROC Curve (AUC) which was 0.952 signified aggregate measure of performance across all possible classification thresholds was out- standing. Therefore, ADASYN [19] was finally chosen as the oversampling technique for this model based on the AUC value and other metrics (Table 1) and finally the total number of data was 1466 in the dataset. Based on the performance value of important metrics (Table 1) after dividing the dataset into 80-20, 70-30, 60-40 for training and testing, 80-20 train-test division of dataset was implemented in this model.

Method	*Train-Test%*	*Metrics*	*AUC*	*F1 Score*
ADASYN	80-20	Accuracy = 0.929 Precision = 0.908 Recall (TPR) = 0.977 Specificity (TNR) = 0.861	0.952	0.941237
	70- 30	Accuracy = 0.928 Precision = 0.913 Recall (TPR) = 0.940 Specificity (TNR) = 0.915	0.934	0.926303
	60-40	Accuracy = 0.939 Precision = 0.931 Recall (TPR) = 0.946 Specificity (TNR) = 0.933	0.940	0.93844
SMOTE	80-20	Accuracy = 0.930 Precision = 0.883 Recall (TPR) = 0.994 Specificity (TNR) = 0.841	0.950	0.935218
	70- 30	Accuracy = 0.942 Precision = 0.896 Recall (TPR) = 0.994 Specificity (TNR) = 0.849	0.944	0.942459
	60-40	Accuracy = 0.915 Precision = 0.904 Recall (TPR) = 0.890 Specificity (TNR) = 0.903	0.947	0.934048

Table 1: Comparing SMOTE vs. ADASYN methods.

Machine Learning Model

The performance of the XGBoost classification model (Table 1) showed that the AUC of the model under the ROC curve was 0.952. The accuracy of the model is 0.929 and specificity is 0.861. (Fig 2).

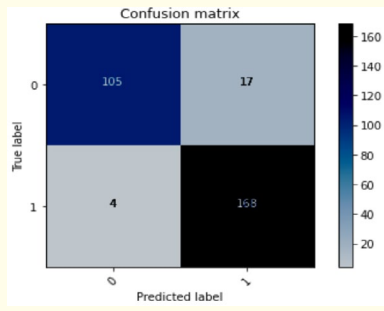

Figure 2

Accuracy = 0.929.
Precision = 0.908.
Recall (TPR) = 0.977.
Specificity (TNR) = 0.861.
Fallout (FPR) = 1.393e-01.
gmean = 0.9168644872558049.

Explainable AI

Ante-hoc Methods: The critical risk factors for the detection of the state of the pulmonary nodules using the XGBoost model were explained by its inherent feature explanation methods. XGBoost has its default library, eli5, which was used to find out the impact of the individual feature over the prediction by the model (Table 2). It was observed that Calcification was the input feature contributing maximum in the prediction of the state of the nodules.

Feature	***Weight***
Calcification	0.512
Internal Structure	0.2331
Lobulation	0.0848
Margin	0.0584
Sphericity	0.0307
Spiculation	0.0291
Subtlety	0.0277
Texture	0.0242
Number of Nodules	0
Nodule_Size More_ than_3mm	0

Table 2: Feature vs. Weight Contribution list.

Post-hoc Methods

SHAP: The SHAP tool was used to highlight the risk factors for explanations and interpretations of the predictions. Once the result is explained and interpreted, the trusts of the end-users will automatically build up. Fig. 3 illustrated the SHAP Summary plot which validated the result obtained by the inherent feature explanation methods of XGBoost classifier (Table 2).

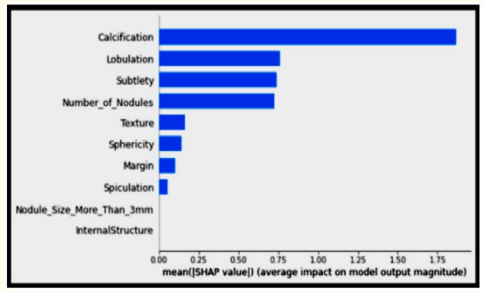

Figure 3

The feature contribution summary plot (Fig4) illustrated top feature contributions and also illustrated data point distribution to provide visual indicators of how feature values affect predictions. Calcification, Number_of_Nodules, Texture of the nodules, and Subtlety had a high positive impact while detecting the state of the pulmonary nodules as malignant. Here, the red colour signified the positive influence on declaring the nodule was malignant, whereas the blue colour implied the negative influence.

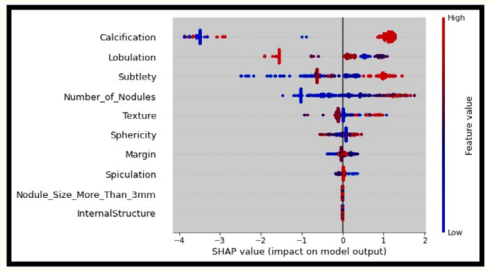

Figure 4

The dependency plots (Fig 5) illustrated the relationship of the calcification of the nodules with the texture of the nodules. Similarly, few other dependency plots were created. The higher the SHAP value of an input feature or biomarker, the higher the influence of that biomarker in the detection of cancerous cells, thus the nodules with 'popcorn' calcification pattern and irregular shape with solidified texture were more prone to be of malignant nature (cancerous cells). In this way dependency plots could explain any dependency between the features which can lead to some unseen properties that could be revealed by these plots.

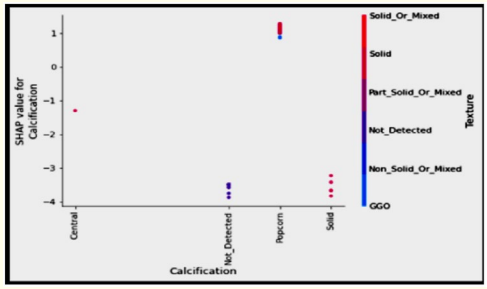

Figure 5

The partial dependency plot visualized the relationship of a biomarker with its SHAP value to understand its overall impact in the final prediction of the system. The higher the SHAP values implicated, the higher the risk factors.

The Force Plot was used for local prediction and it was formed based on any single record by calculating the SHAP values of the biomarkers of that individual patient (one specific single record). Here the force plot (Fig 6) visualized -2.43 was the model output for the sample with index 1 using the following coding in Python:

shap.force_plot(shap_explainer.expected_value,test_shap_values[1,:], X_test_disp.iloc[1,:]) All of the features' (Calcification, Sphericity, Texture, etc) values led to the prediction value of -2.43, which was then converted to a value of 0 class (Benign). SHAP plotted the top most influential features for the sample under study. Features in red colour influenced positively, i.e. dragged the prediction value closer to class 1 (Malignant), features in blue colour influenced the opposite. The Force Plot is very helpful to detect individual patients' status from their CT scan report.

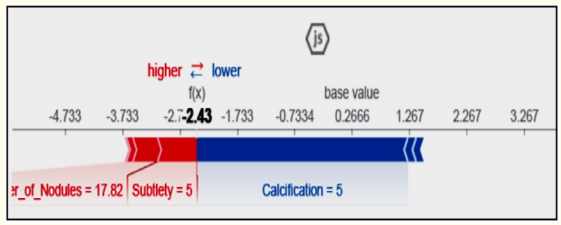

Figure 6

LIME: LIME makes surrogate models to understand the feature importance. The plots are used for local prediction and it gives faster output than SHAP tool. LIME produced model-agnostic explanations for a particular instance or the vicinity of a particular instance and here (Fig 7) the values of Calcification, Lobulation, Nodule_Size_More_Than_3mm, Internal Structure classified the sample as 'Benign'. Although the number_of_nodules influenced in other way, but it was visible that the sizes of all the nodules were less than 3mm, henceforth this case came under benign class under this XAI Model. Henceforth, our research work classified those physiological attributes as a high-risk factors for lung cancer, for which the medical practitioners required a post-hoc clarification and provided explanations consistent with biological intuition.

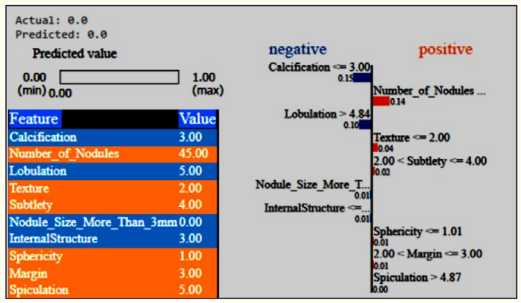

Figure 7

Similarly, another AI model using XGBoost classifier was created with few biomarkers and the performance of the

new model was evaluated based on the result of important metrics. Here, the second AI model was built by selecting seven important features only, such as 'Calcification', 'Lobulation', 'Margin', 'Sphericity', 'Subtlety', 'Texture', 'Number_of_Nodules, 'Malignancy'. The values of the important metrics of the second AI Model are shown in Fig 8.

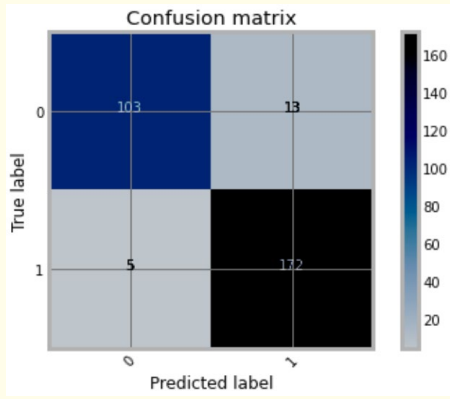

Figure 8

Accuracy = 0.939.
Precision = 0.930.
Recall (TPR) = 0.972.
Specificity (TNR) = 0.888.
Fallout (FPR) = 1.121e-01.

It was understood that after selecting seven features out of eleven features, the performance of the model was better than that of the first model. The results of SHAP for the new model were displayed in Fig 9. The influence of biomarkers remained the same to predict the malignancy of pulmonary nodules. Henceforth, any new test data can also be fed into this XAI model and the model can predict along with highlighting those biomarkers as reasons for prediction which endow with the confidence to the medical practitioners, doctors and patients to trust the model.

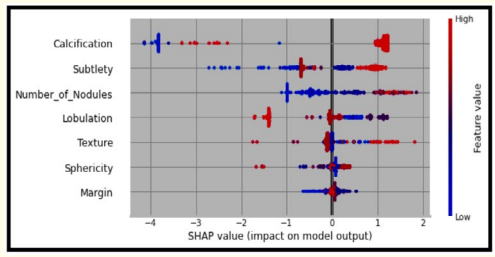

Figure 9

Conclusion with Future Work

In this research work, the prediction of XGBoost AI binary classification model was successfully interpreted by XAI tools and the interpretations achieved by using the XAI tools were applied to rebuild and reform the classification model by selecting only the most important input features and that made the execution speed of the new model faster than the previous one. Through experimental results we observed that, many times AI approaches looks indecisive for extraneous oversampling data features of an image to detect pulmonary nodes (although it may correctly predicted), and also sometime there may be serious contrariety to give proper explanation in how different deep learning models gives the same prediction. Our proposed model focuses that XAI can provide trust in AI

systems especially that are used in high stake decision process and can minimize AI implementation risk factors. Similar XAI models could be built to diagnose other diseases as well. Here the XAI model was formed using XGBoost's inherent architecture, and two post-hoc XAI tools SHAP and LIME. We could explore other new XAI methods and different statistical approaches on hybrid AI models for assured predictions and those trustable prediction models could be extended to multiclass detection models from a binary classification model to grade cancer stages.

References

1. AMA passes first policy recommendations on augmented intelligence.
2. An introduction to explainable AI with Shapley values.
3. The Global Burden of Disease 2004 Update, WHO.
4. Hancock, M., Pylidc , MIT.
5. Rathe A. Random Forest vs XGBoost vs Deep Neural Network, Kaggle.
6. Zhu P and Ogino M. "Guideline-based additive explanation for computer-aided diagnosis of lung nodules". In Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support (2019): 39-47.
7. Singh A, Sengupta S and Lakshminarayanan V. "Explainable deep learning models in medical image analysis". Journal of Imaging 6.6 (2020): 52.
8. Lucieri A., et al. "Achievements and Challenges in Explaining Deep Learning based Computer-Aided Diagnosis Systems". arXiv preprint (2020).
9. Ahmed ZU., et al. "Explainable artificial intelligence (XAI) for exploring spatial variability of lung and bronchus cancer (LBC) mortality rates in the contiguous USA". Scientific reports 11.1 (2021): 1-15.
10. Siddhartha M, Maity P and Nath R. "Explanatory artificial intelligence (XAI) in the prediction of post-operative life expectancy in lung cancer patients". Int J Sci Res 8 (2020).
11. Bartczak M and Partyka M. Chapter 8 Story Lungs: eXplainable predictions for post operational risks.
12. Venugopal VK., et al. "Unboxing AI-radiological insights into a deep neural network for lung nodule characterization". Academic radiology 27.1 (2020): 88-95.
13. Zhou B., et al. "Learning deep features for discriminative localization". In Proceedings of the IEEE conference on computer vision and pattern recognition (2016): 2921-2929.
14. Selvaraju RR., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In Proceedings of the IEEE international conference on computer vision (2017): 618-626.
15. Lundberg SM and Lee SI. "A unified approach to interpreting model predictions". In Proceedings of the 31st international conference on neural information processing systems (2017): 4768-4777.
16. Shrikumar A, Greenside P and Kundaje A. "Learning important features through propagating activation differences". In International Conference on Machine Learning (2017): 3145-3153.
17. The Cancer Imaging Archive (TCIA) Public Access, LIDC-IDRI.
18. Tulio Ribeiro M, Singh S and Guestrin C. "Why Should I Trust You?": Explain- ing the Predictions of Any Classifier. arXiv e-prints (2016): arXiv-1602.
19. He H., et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (2008): 1322-1328.