from *Banuri experimental tea farm*, Palampur, India. Our data consists of 965 tea leaf images where the combination of 1-Leaf 1-bud consists of 261 data samples, 2-Leaf 1-bud has 174 data samples, 3-Leaf 1-bud have 279 data samples, 4-Leaf 1-bud has 199 data samples, and 5-Leaf 1-bud has 52 data samples. Because of the class imbalance in our dataset, we have implemented data augmentation, which reduced the class imbalance. We used the final layer of convolution for producing a localization map that highlights the most important regions of the image while determining the class of the tea-leaf. While performing our experiments, we found that the Grad-CAM method takes out features from the pre-trained VGG16 model. We implemented transfer learning by using well-known pre-trained backbone models like VGG16 and InceptionV3. For our tea leaf dataset, we discovered that the performance evaluation metrics of our built model on 1L1B(1-Leaf 1-bud) have overall best performance, with precision, recall, and F1-scores of 0.90, 0.90, 0.90. Our 2L1B data (2-Leaf 1-bud) has precision, recall, and F1scores of 0.73, 0.62, and 0.67, respectively. Whereas 3L1B (3-Leaf 1-bud) has a precision of 0.80, recall of 0.82, and F1-score of 1. Moreover, 4L1B (4-Leaf 1-bud) has a comparatively low precision of 0.69, but has a good recall of 0.92, and has an F1score of 0.79. 5L1B (5-Leaf 1-bud) has a precision of 1 and has a comparatively poor recall of 0.18, and an F1score of 0.30 which shows the model has certain limitations in the identification of the actual distinguishing features responsible for classifications for tea leaves. Our trained classification model, which employs VGG16, predicts an average of 0.83, 0.69, 0.74 precision, recall, and F1-score, with an accuracy of 80%. And by using InceptionV3 our model could able to have achieved an accuracy of 76.41% respectively. We obtained explainations based on the classification in our dataset of the tea leaf images by providing the pre-trained models as input to the Grad-CAM pipeline for producing class-specific heat maps. We used Grad-CAM as an explaination method for explaining our prediction of the model. The model generates explainations for different leaf image categories. Grad-CAM exclusively highlights the important portions of the image which are of utmost importance for predicting the class of the tea leaf image. Figure 4.7 shows the generated heatmaps for explaining the most important image regions which are influential in predicting the result using pre-trained model VGG16 and InceptionV3 models.

The heatmap produced by Grad-CAM for the pre-trained VGG16 model focuses particularly on the Leaf and Bud portion of the input image, whereas the heatmap generated by Grad-CAM with pre-trained InceptionV3 model has an inappropriate focus on the stem region of the leaf image. Grad-CAM uses the final feature map of the model for creating heatmaps that highlight the image pixels responsible for the prediction of the image class.

4.6 Challenges and Future Directions

Explainable AI is an active area of research for handling the black-box nature of deep learning models. The most recent literature survey, reveals that though there

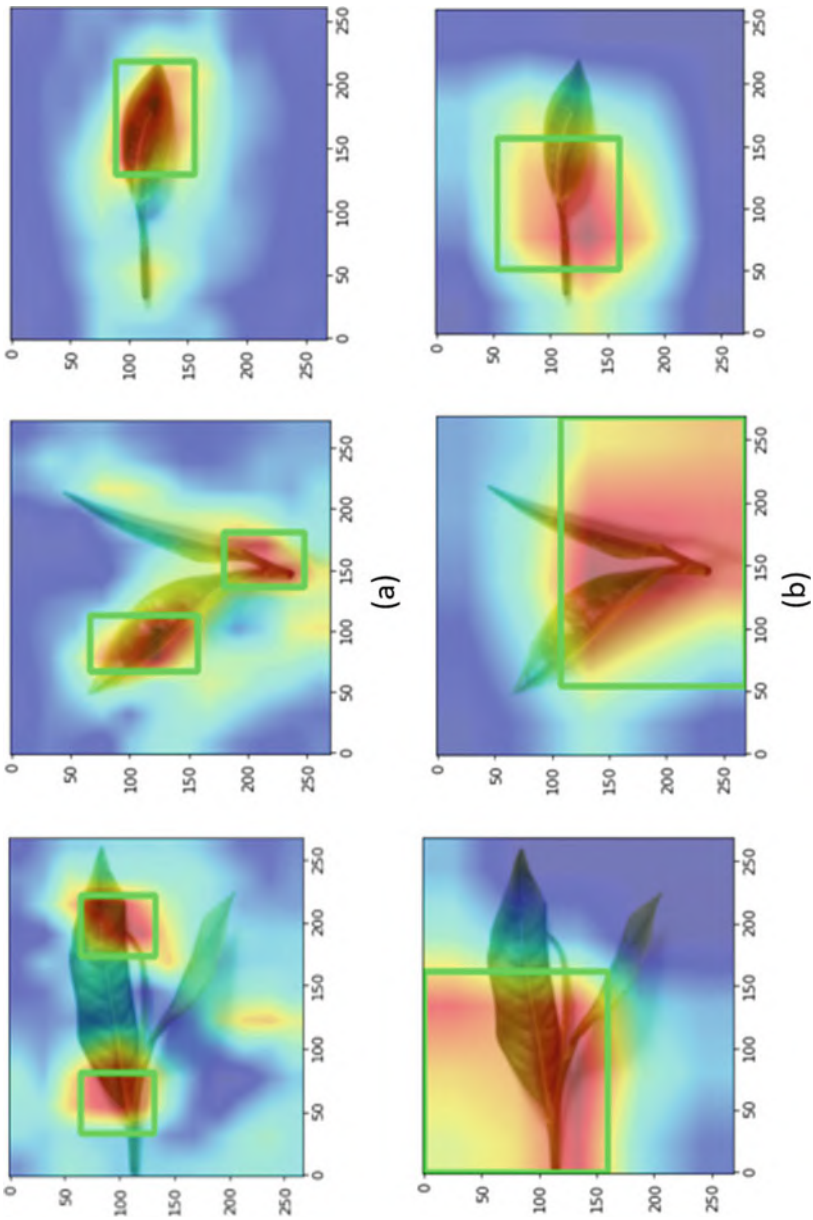

Fig. 4.7 Explaining the prediction by using Grad-CAM and pre-trained **a** VGG16, and **b** InceptionV3 models (Banerjee et al. 2022)

are several works published in this domain of research there are still many important problems to be resolved by the research communities. Though there are a lot of AI black-box methods are there which need explainations but due to high-dimensional issues, the image classification and object detection problem in computer vision is challenging. The black-box models that yield highly accurate results normally need to be explained using visual explaination for better perception of the underlying working of the model behind the same. However, the main challenge for visual explainations of a classifier's output arises from the complex nature of the AI model and underlying data. Since the visual explaination method relies on the algorithm for the generation of human perceptible feedback which becomes subjective in nature. Moreover, it is very challenging to properly explain complex classification models accurately. The classifiers which provide good performances are gradually becoming more complex due to the involvement of the numerous parameters and the operations performed by them which in turn makes them complex and tough to explain. Because of the different architectures of the AI methods, the problem to design an effective framework for explaining the underlying decision process increases. Other challenges include the use of multiple types of data for training the models. For explaining an AI model, the most common strategy is to trace back to the input data. Different types of data require different types of explainations. Though the AI model based on image data is relatively easier for generating visual interpretations of the decision process, the same using the textual, speech, or nominal data is difficult to get explained similarly. Even in image data-based AI models, there are no objective metrics available that measures the quality of explainations. This paves the way for further research in this area. In this section, we endeavored to identify a few main research questions in the field of XAI. These questions include some necessary aspects like how do we evaluate the Extent of Explaination for an AI model? While referring to the AI perspective, metrics are defined as any quantification of the extent of explaination which helps in evaluating its quality and suitability.

Evaluation metrics explain the model's performance. It is used in measuring the quality of the explaination. Though there are multiple metrics used for tasks of classification, ranking, clustering, regression, modeling topics, etc. there are very limited metrics are there for measuring the extent of explaination in different data types and AI tasks. This in turn forms another research question whether a method of explaination can be devised that is invariably applicable to any data type or AI model?

There are different methods of explainable AI that target explaining the decision-making of the AI system and thus can identify the problems associated with the underlying model. But it needs to be investigated how explainable AI models can help in improving the prediction accuracy or inhibiting the failure points in real-life problems. XAI consists of a set of frameworks and tools which helps in interpreting and understanding the predicted output of the AI models. It is also imperative to study the applicability of the explaination method on a local instance of data or globally on an entire dataset. There are techniques where an AI model can be explained locally and globally based on the given input data but rarely any specific XAI method is available which can effectively evaluate the quality of explainations both in the local

and global dataset. Also, there is a necessity for more quantitative evaluation metrics which would provide a comparison between different types of explainability methods and would quantify the acceptance of these kinds of methods for increasing trust in them. Moreover, these kinds of metrics could be used for standardizing the XAI solutions.

4.7 Conclusion

Explainable AI (XAI) is an emerging technological paradigm of which most enterprises are conscious. The methods and processes of XAI provide several advantages. Explainability in pre-modeling is a feasible but under-focused approach for avoiding transparency problems. Pre-modeling explainability methods mainly focus on the explainability of data rather than the model itself. Whereas Post-modeling/Post-hoc explainability is a collection of different types of methods with a common goal of gaining a better understanding of the working of the trained model. Based on Post-hoc explaination the methods are classified into model-agnostic and model-specific techniques. Moreover, the metrics used for explaining the AI decisions are quantitatively evaluated as Subjective metrics, Objective metrics, Computational metrics, and Cognitive metrics for evaluating the AI system accurately. Although there is a necessity for more quantitative evaluation metrics which would provide a comparison between different types of explainability methods and would quantify the acceptance of these methods for enhancing trust in them. For increasing transparency in the developed model, it is necessary to produce an intuitive explaination. In this chapter, we have presented a comprehensive overview of the methods and metrics for explaining decisions made by AI models. We also covered the taxonomy of XAI in ample detail and discussed different strategies used for providing explainations behind the working of the data-based learned models. We have presented a selected overview of works to assist researchers and practitioners in understanding insights, accessible resources, and unresolved difficulties in using XAI methodologies. A use-case of implementing a popular XAI visualization method is also been demonstrated in this chapter. Though the research work in the area of XAI is in full swing but still has many gray areas to be addressed by the global AI communities. The research directions section in this chapter is an endeavor to summarize the identified research gaps and unanswered research questions for prospective XAI researchers.

References

Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)

Antwarg, L., Miller, R.M., Shapira, B., Rokach, L.: Explaining anomalies detected by autoencoders using shap. arXiv preprint arXiv:1903.02407 (2019)