4.4.1 Evaluation Metrics for Explaining AI Decisions

As discussed in Sect. 4.3, the methods of explaination are broadly classified into three types namely Model-based Explainations, Attribute-based Explainations, and Example-based Explainations. In this section, the quantitative metrics are being discussed for measuring the qualities of these explaination methods. Apart from the evaluation of the AI explaination methods, proper selection of the evaluation metrics plays a major role in evaluating the system accurately. Different types of metrics that evaluate the extent of the explainability by different methods are as follows:

- **Subjective metrics**: It is designed for questioning the users based on tasks and the explainations provided, these questions are asked when the task gets executed or afterward for obtaining the subjective response from the user on the explainations. Some of the examples of these types of metrics are the confidence, and trust of users that have an enormous grasp over the focal points for evaluating the explainable system. Hoffman et al. (2018) proposed a metric that is used for the subjective evaluation of an AI system. It considers factors like user trust, understanding, and satisfaction. Zhou et al. (2016) have looked over the factors like the uncertainty that affect the trust of users in informed machine Learning decision makings. They established that explaination generated because of influence in the training data points remarkably affects the user's trust in case of informed decision making.
- **Objective metrics**: It mentions the objective information of a task to a user before or after the task is being performed. Such kinds of examples are human metrics, which include behavior and physiological measures of humans when informed decision making takes place, another such metric is task-related metrics, which include time length for completing the task and performance of the task. Schmidt and Biessmann (2019) showed that fast and accurate decisions mean instinctively understanding the explainations provided. It resulted in deriving a trust metric that is based on the explainability metrics.
- **Computational metrics**: These metrics are known as mathematical indicators for determining the quality of explainations generated by an XAI system. The measurement of these kinds of explainations is generally being carried out by using necessarily developed equations. Thus, these metrics may be used without any kind of human intervention as guidelines for preparing the explaination techniques.
- **Cognitive metrics**: The explainations provided to the end-users are being measured by using cognitive metrics. The assessment of human subjects is a blunt indication of explainations, as we know the initial goal of XAI is to convey the reasons behind machine judgments to people.

Accuracy is one of the most commonly used metric (Rosenfeld 2021), it is very easy for understanding although noticing only these metrics would give an incomplete suggestion regarding the performance of a model. Multiple established metrics are there which would provide a thorough insight into the performance of the model. The metrics used for quantifying the explainations are generally very specific to the different types of machine learning problems and models. Some of the widely used

Table 4.4 Metrics for quantitatively explaining AI decisions

Types of explaination	Metrics	Explaination properties
Model-based explainations	Model size (Guidotti et al. 2018)	Simplicity
	Interaction strength (Markus et al. 2021)	Simplicity
	Level of agreement (Lakkaraju et al. 2017)	Clarity
Attribution-based explainations	Effective complexity (Nguyen and Martínez 2020)	Broadness and simplicity
	Recall of important features (Ribeiro et al. 2016)	Soundness
	Selectivity and continuity (Montavon et al. 2018)	Soundness and clarity
	Mutual information (Nguyen and Martínez 2020)	Broadness and soundness
	Sensitivity (Montavon et al. 2018)	Soundness
Example-based explainations	Diversity (Nguyen and Martínez 2020)	Simplicity
	Non-representatives (Nguyen and Martínez 2020)	Simplicity and completeness

metrics are Loss, Confusion Matrix, Accuracy, Mean Absolute Error, Root Mean Square Error, Accuracy, etc.

Table 4.4 shows the metrics for quantitatively explaining the AI decisions for explaining the classification.

4.5 Use-Case: Explaining Deep Learning Models Using Grad-CAM

As a use-case of XAI, we implemented some of the visual methods of XAI for explaining the learned models for classifications of fresh tea leaves (Banerjee et al. 2022). We have applied a well-known technique for introducing visual explainations of predictions from Neural Network-based models, which makes them more interpretable. We have used transfer learning for classifying different types of tea leaves. The aim here is to reveal the underlying features that are responsible to explain the relationship between the predictions of a tea-leaf classification model by using popular visualization techniques. In this work, the Grad-CAM method has been used that use the gradients of the targeted concept. We performed our experiment on our image dataset created by imaging and labeling the freshly harvested tea leaves