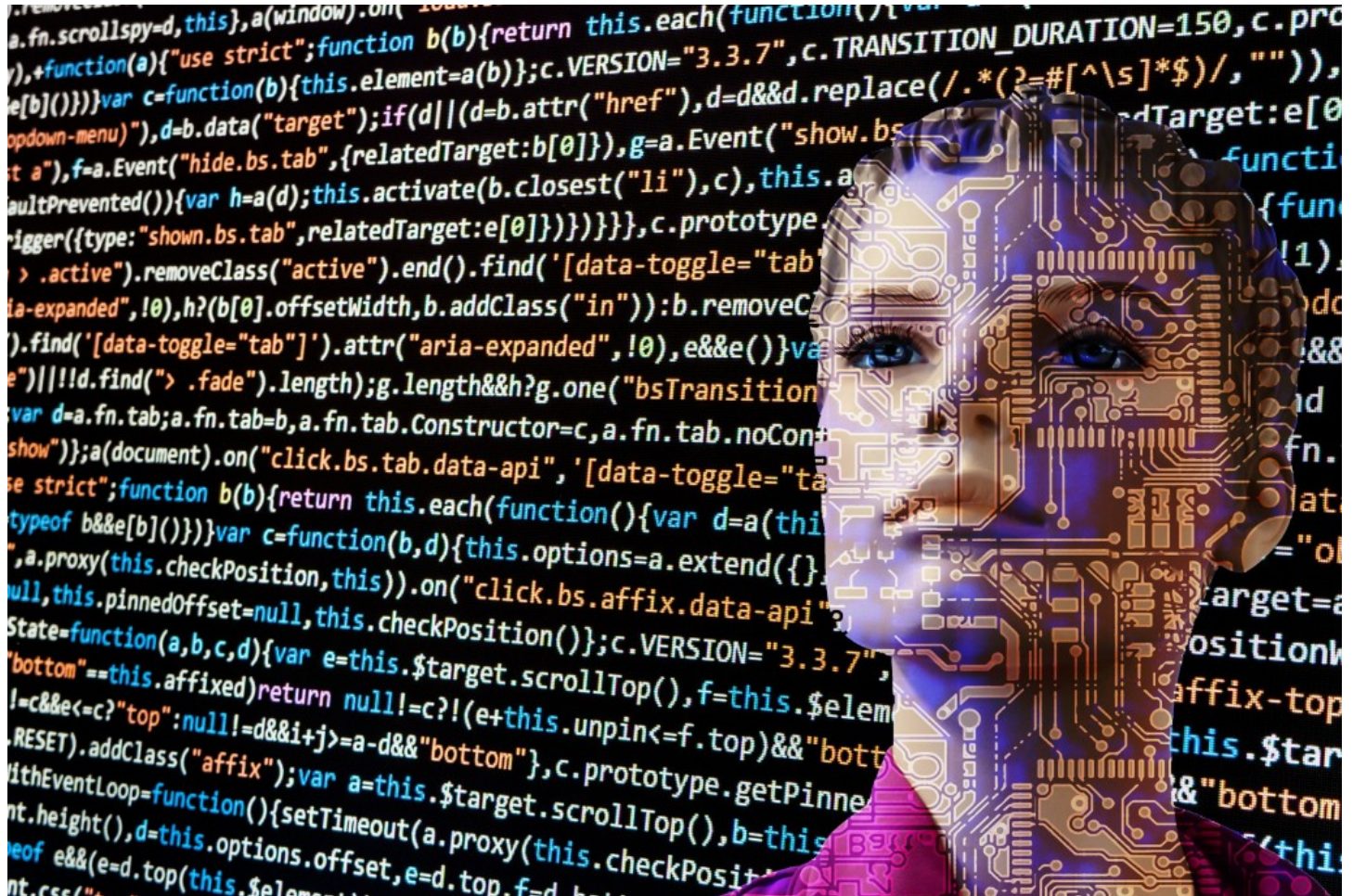


Bhusan Chettri's Tutorial of Interpretable Machine Learning in Safe-Critical Applications.

This tutorial from Bhusan Chettri provides an overview of Interpretable Machine Learning (IML) a.k.a Explainable AI (xAI) taking into account safety-critical application domains such as medicine, finance and security.



London, United Kingdom Sep 26, 2022 ([Issuewire.com](https://www.issuewire.com)) - This tutorial is the second installment of the interpretable AI tutorial series by Dr Bhusan Chettri a PhD graduate in AI and Voice Technology from Queen Mary University of London. The tutorial talks about the need for explanations from AI and Machine Learning (ML) models by providing two examples in order to provide a good context about the IML topic. Finally, it describes some of the important concepts a.k.a criterias that any ML/AI model in safe-critical applications must satisfy for their successful adoption in real- world setting. But, before getting deeper into this edition, it is worth revisiting briefly the first part of this tutorial series of [AI, Machine Learning, Deep Learning and Their Interpretability](#)

Part1 mainly focussed on providing an overview about various aspects related to AI, Machine Learning, Data, Big-Data and Interpretability. It is a well known fact that *data* is the driving fuel behind the success of every machine learning and AI applications. The first part described how vast amounts of data are generated (and recorded) every single minute from different mediums such as online transactions, use of different sensors, video surveillance applications and social media such as Twitter, Instagram, Facebook etc. Today's fast growing digital age that leads to generation of such massive data,

commonly referred as *Big Data*, has been one of the key factors towards the apparent success of current AI systems across different sectors.

The tutorial also provided a brief overview of AI, Machine Learning, Deep Learning and highlighted their relationship: deep learning is a form of machine learning which involves use of artificial neural network with more than one hidden layers for solving a problem by learning patterns from training data; machine learning involves solving a given problem by discovering patterns within the training data but it does not involve use of neural networks (PS: machine learning using neural networks is simply referred as deep learning); AI is a general terminology that encompasses both machine learning and deep learning. For example, a simple chess program which involves a sequence of hard-coded if-else rules defined by a programmer can be regarded as an AI which does not involve use of data i.e there is no data-driven learning paradigm. To put it in simple terms, deep learning is a subset of machine learning and machine learning is a subset of AI.

The tutorial also briefly talked about the *back-propagation* algorithm which is the engine of neural networks and deep learning models. Finally, it provided a basic overview of IML stressing their need and importance towards understanding how a model makes a judgment about a particular outcome. It also briefly discussed a *Post-hoc* IML framework (that takes a pre-trained model to understand their behavior) showcasing an ideal scenario with a human in a loop for making the final decision of whether to accept or reject the model [prediction or a particular outcome. Please read Tutorial of AI, Machine Learning, Deep Learning and Their Interpretability for more details.](#)

Why do we need interpretability? The need for explainable AI

One important question that arises is why do we need interpretability or explanations from ML models at all? Well, Bhusan Chettri explains that the need arises when an AI or ML model needs to satisfy some criteria (for example trustworthiness) for its successful adoption in real-world setting or business applications. Before elaborating further, Bhusan Chettri considers the following two examples to build up the context on this topic. More specifically, the examples below aim to reflect how interpretability helps discover problems caused by biases in dataset that influence model decisions which can be catastrophic in safety-critical applications.

Example 1 – Consider an automatic system trained to detect between a dog and a wolf, a *wolf monitoring system*, which is to be deployed in some neighborhood. Whenever the system detects a wolf it triggers an alarm to raise awareness to the people in that neighborhood for taking safety precautions. However, when it detects a dog (for example a husky) the system just continues its normal functioning i.e keeps scanning the images around the neighborhood in real time. Now, what happens if the training data used for training such automatic systems contained biases? Dataset biases can occur from a variety of sources for example during data collection, compilation and post processing. Often the underlying data collection process involved is automatic with little or no human involvement in post verification. One reason for this is due to enormous dataset sizes. Thus errors are prone to occur at various stages of data collection, for instance a wolf image mistakenly labeled as dog and vice-versa in the training set from crowdsourcing workers hired for data annotation. Or, let's say images in a training dataset for a particular class (say Dog or Wolf class) contained some cues when images were recorded?

On a similar note, as illustrated in [this research article](#) from Washington University, the researchers found that the classifier (dog vs wolf detector) had learned to detect wolves based on other contents within the image. In simple words, the model had learned to use presence/absence of snow (white patches) as a potential cue for detecting wolves. This suggests that most of the wolf pictures in the

training set contained white patches of snow or ice and the model had used this as a cue which is actually not relevant to the problem. To put it another way, as there is a very tiny difference between a wolf and a husky dog, ML models may have found it quite hard to exploit that tiny cue for making the right decisions. Hence ML models exploited a back-door as a way to provide right answers but for wrong reasons. This means if a dog image contained some snow in the background, then the model could classify it as a wolf with a high degree of confidence. Similarly, if a wolf picture is not having any snow in the background, then this model may misclassify it as a dog with a high probability. Now, imagine the consequence for such a *false acceptance* in a real world scenario. The system would not trigger alarm and the public may find wolves roaming freely around the streets – and the damage it may cause would be costly. This model, therefore, is not trustworthy. It is heavily biased to irrelevant cues in the training data. It is worth noting that it would be quite challenging for a normal human being to think that context in the image (snow in this case) could have been used as a potential cue for class discrimination. However, it is very possible for machines. Infact, machines can do things that are way beyond human expectations, imaginations and understanding. Thus interpretation and reasoning behind why and how a ML model comes to a particular outcome or decision is very important to ensure trustworthiness and its reliability.

Example2 – Medicine: Consider an automatic Tuberculosis (TB) detector. The machine learning model has been trained to automatically detect if a person has TB or not. It has been trained using both a healthy person's data and a person's with TB. Let's imagine that a large sample of data has been collected for this. Raw data collected from each person contains various information collected in a form of a questionnaire related to TB such as duration of cough, body temperature, presence of blood samples in the cough, etc. The model would simply learn to make judgment about whether a person has a TB by exploiting correlation between input features and the answers: TB or no TB. This model might show good performance on a held- out test set also. But can such a model be deployed in hospitals in real time for making automatic decisions without any human involvement? The answer is No. It is quite risky to just solely make a judgment if a person has a TB or not on this model outcome itself. Such safety-critical domain requires a certain level of trust and assurance that these models do behave as they should when deployed in real-time applications.

Thus these models need to satisfy certain criterias for their safe adoption in real-time use cases. For this human-in-the-loop is often used and appears to be an important part in assessing their trustworthiness before making final decisions.

Thus, for this particular problem, the IML framework can be used in the following way. First, the automatic detector generates the prediction for the new input (i.e the new patient details) by extracting relevant features just the same way as done during model training. An output value closer to 1 would indicate the presence of TB while a value closer to 0 indicates the patient is not having TB with high likelihood. A value of 1 indicates that the model is highly confident that the patient has TB while a value of 0 indicates otherwise with high certainty. Now to produce an explanation for this prediction, the obtained prediction/outcome along with the features extracted for the input instance are usually passed to the IML framework i.e the explainer in a post-hoc setting (where the IML framework operates on pre-trained models). The explainer then returns an explanation for the prediction of this input instance which could be a list of words highlighting important features (among all the features) that the model used for making the particular decision. For example, it may assign large weights to input features such as *blood in cough, continuous coughing for more than 2 weeks*. With such explanations which directly correlate to TB, it becomes easier for a doctor (human-in-the-loop) to make the final decision and produce a report with high confidence whether the patient has TB or not. Without such explanations, it is difficult to make decisions based solely on models prediction.

Next, the article will discuss some of the important criterias that any AI or ML model needs to satisfy for its successful adoption in real-time usage. The article will further describe how IML can help in meeting these criterions.

Trust:

Imagine the degree of trust and faith a patient has on a Doctor; the trust an individual has on the Pilot flying the plane being boarded to travel to a new destination. In a similar context, for application domain such as medicine, finance or security it is very important that humans (or users) have the similar degree of faith and trust (just the way humans do in real-world scenarios like the case of a doctor or a pilot as highlighted earlier) on the AI or ML models that shall be deployed for use in real time. Now to achieve this, one simple yet crucial step in order to promote trust to ML users is by making use of IML methods to ensure that data used to train such models are free from inherent biases or confounders that influence model decisions. This in turn can often be achieved by performing data exploratory analysis using initially trained models and analyzing data for test instances of confidently classified data samples. With such approaches one can confirm whether models are exploiting the right features from the input data towards making decisions. Check out [Analyzing The Predictions Of a CNN-Based Replay Spoofing Detection System](#) article for more on this.

Robustness: It is very important that AI & ML models are robust against small perturbation on input features. Their behavior must not change when bad actors present manipulated input instances to the model to manipulate model decisions for their personal gain. Research has demonstrated that today's high-performing AIs are often easily fooled when crafted input instances imperceptible to human eyes or ears are presented to them. The systems produce completely different output class labels (in case of an image classifier) or might produce a different audio (in case of Speech synthesis systems). Such carefully crafted instances are called *adversarial input* and the field is referred to as *adversarial machine learning*, a very hot topic in the field of machine learning but beyond the scope of this article. Although many different techniques to combat adversarial inputs have been proposed by ML researchers, none of them offers a full- proof solution. Therefore, lack of robustness in AI/ML models remains a pressing concern when they are deployed in safe-critical applications such as medicine, finance or autonomous vehicles. One way to tackle this issue and make ML models more robust could be to use the IML framework for analyzing such adversarial examples to derive better understanding about how and why models show different behavior on such input instances. This in turn may help build robust ML models.

Reliability:

The behavior of ML models must remain similar when tested across multiple instances. Often ML models fail to show good generalization or perform unreliably when they get deployed in a real world setting. In other words, their behavior is often different in real-world usage in comparison to the one found in laboratory settings (where they are trained and tested). Thus reliability of ML models is also one key factor in safe-critical applications to ensure that models behave in a similar way as they behaved in laboratory conditions. One approach towards building reliable ML models could be to ensure that test dataset mimic real-world settings where those models are likely to be deployed.

Legal requirements:

As per GDPR – General Data Protection Regulation – a European Union regulation, every user of automatic systems possess a right to explanation of the systems decisions. By incorporating an IML framework in building such an automatic system, it becomes easier to explain the reasons behind an

outcome or predictions in a meaningful and human-friendly way.

Fairness:

ML models must make decisions by exploiting the right cues from the training data. However, it is often very challenging to ensure that training dataset are clean and bias free. Biases may occur during data compilation, post processing or even during data collection through crowdsourcing platforms (for example Amazon Mechanical Turk). Therefore, it becomes difficult to often avoid such ML models to be unfair when the underlying data used for model training is biased already. Therefore, such ML models often behave wrongly or show discriminatory behaviour all because of biased training data being used to learn model parameters. One approach towards avoiding this problem could be to not use certain features from the input data that show such biased behavior. However, it is again difficult to detect such implicit correlations within the training data before training a ML model. Thus, the IML framework can be utilized in such situations to detect if the ML models under question are unfair by analyzing whether discriminative features (right cues/features) have been utilized by models in making predictions.

Summary

In this tutorial, Bhusan Chettri provided an insight on xAI and IML taking into consideration safe-critical application domains such as medicine, finance and security where deployment of ML or AI requires satisfaction of certain criterias (such as fairness, trustworthiness, reliability etc). To that end, Dr Bhusan Chettri who earned his PhD in Machine Learning and AI for Voice Technology from QMUL, London described why there is a need for interpretability on today's state-of-the-art ML models that offer impressive results as governed by a single evaluation metric (for example classification accuracy). Bhusan Chettri elaborate this in detail by taking two simple use cases of AI systems: wild-life monitoring (a case of dog vs wolf detector) and automatic tuberculosis detector. He further detailed how biases in training data can affect models from being adopted in real-world scenarios and that understanding training data and performing initial data exploratory analysis is equally crucial so as to ensure models behave reliably in the end during deployment. Stay tuned for more on the topics of explainable AI. The next edition of this series shall discuss different taxonomies of interpretable machine learning. Furthermore, various methods of opening black-boxes: towards explaining behavior of ML models shall be described. Stay Tuned.

Media Contact

Bhusan Chettri

bhusanchettri5@gmail.com

Source : Bhusan Chettri

[See on IssueWire](#)

