**VIVEKANAND EDUCATION SOCIETY’S INSTITUTE OF TECHNOLOGY**

Department of **Computer Engineering**



**Literature Survey** on

Paper:***Explainable Unsupervised Machine Learning for Cyber-Physical Systems***

Under the subject: Machine Learning (**ML**)

Year: **B.E**. Semester : **VII**

Submitted by

Sanjana Asrani, D17B, 01

Under the guidance of

Subject Teacher

**Dr, Sharmila Sengupta**

(2023-2024)

**Abstract**

**Cyber-Physical Systems (CPS)** are integrated systems that merge the physical and digital worlds, connecting and coordinating the behavior of physical components and processes with digital control and communication systems. In CPS, physical elements, such as sensors, actuators, and processes, are tightly interconnected with computational and communication components, enabling them to interact, monitor, and adapt to one another in real-time. CPS are characterized by their ability to sense, compute, and act in a coordinated manner to achieve specific goals.Ex. :Self-Driving Cars (Autonomous Vehicles): Self-driving cars use a combination of sensors (e.g., lidar, cameras, radar), GPS, and onboard computers to navigate the physical world. They make real-time decisions to control steering, acceleration, and braking based on their surroundings and traffic conditions. For this domain, **UnSL** is needed as CPS generates a high amount of unlabeled data at a rapid pace.

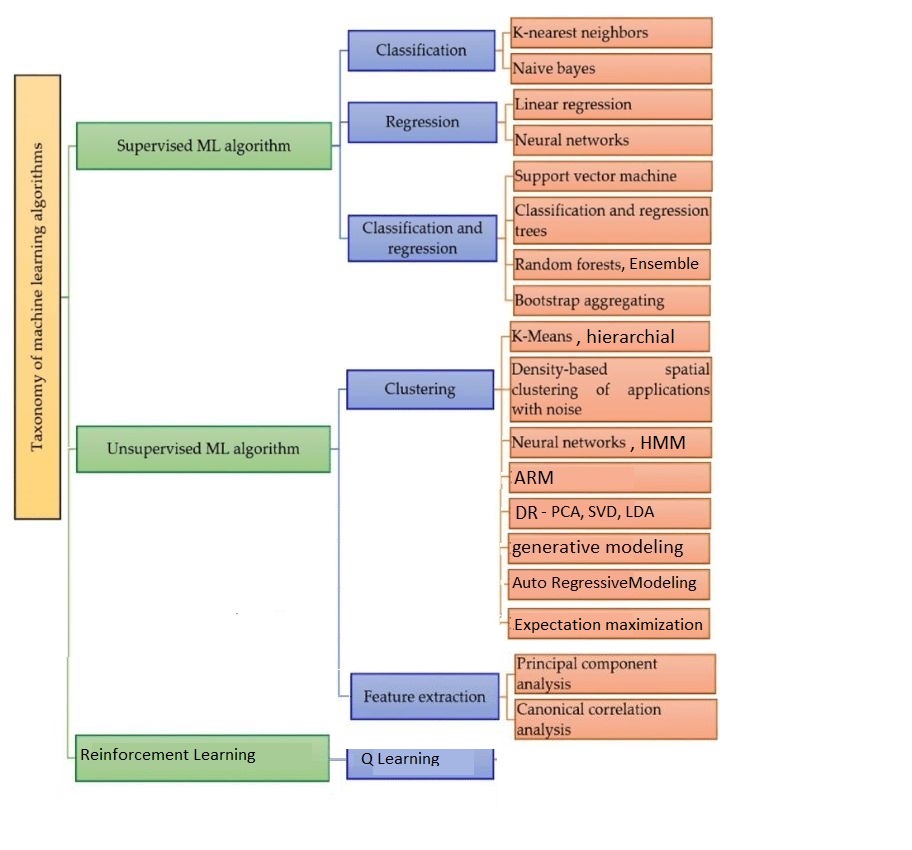
**Explainable AI (XAI)/ Interpretable AI** techniques make machine learning models provide insights into their decision-making process by revealing factors, features, or data that influenced their predictions or classifications. This transparency helps users understand why a model arrived at a specific outcome, which can aid in trust, interpretation, and the ability to improve model performance or address biases. Companies like Amazon, Google, NVIDIA Fiddler lab, IBM, and national institutes are working towards bridging the gap between hardcore data scientists who are building the models and the business teams using these models to make decisions.

**Introduction**

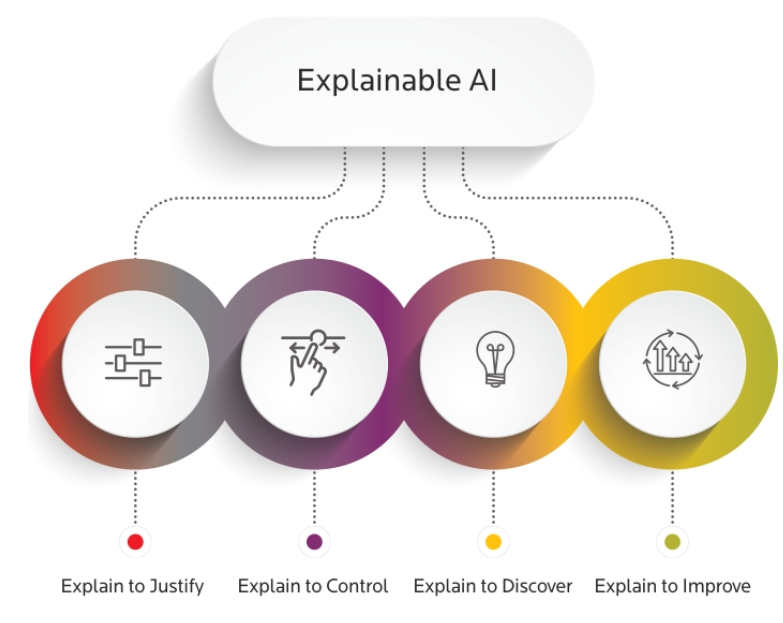
In the era of AI, achieving both effectiveness and interpretability is vital. This literature survey focuses on the emerging domain of unsupervised explainable AI, specifically within Self-Organizing Maps (SOMs). It explores innovative techniques enhancing SOM interpretability and transparency, enabling a deeper understanding of complex data structures. By examining current research and future prospects, this survey navigates the evolving landscape of unsupervised, explainable SOMs. Topics covered include SOM complexity, model transparency, and the interpretability of data patterns.

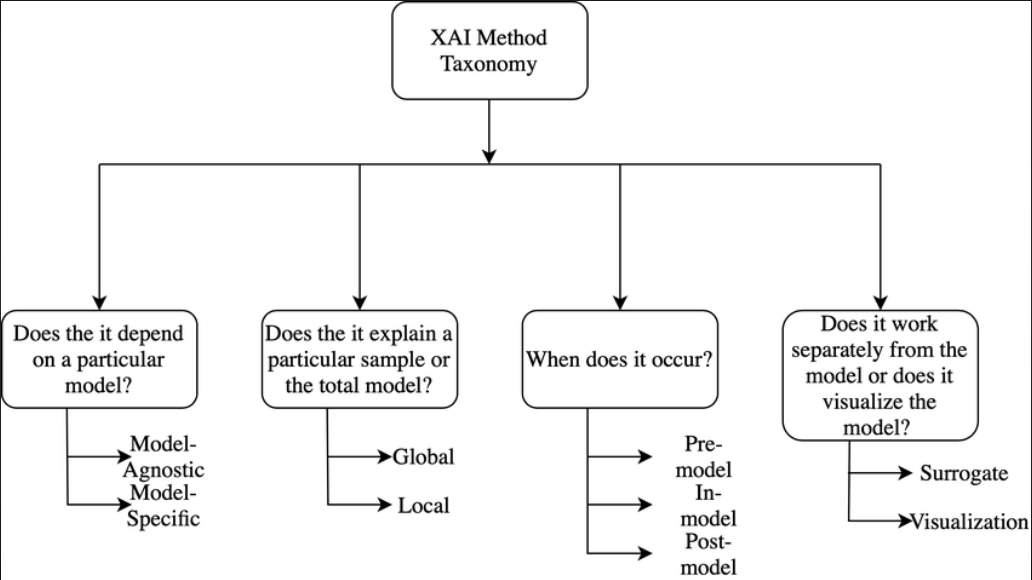
**Literature Survey**

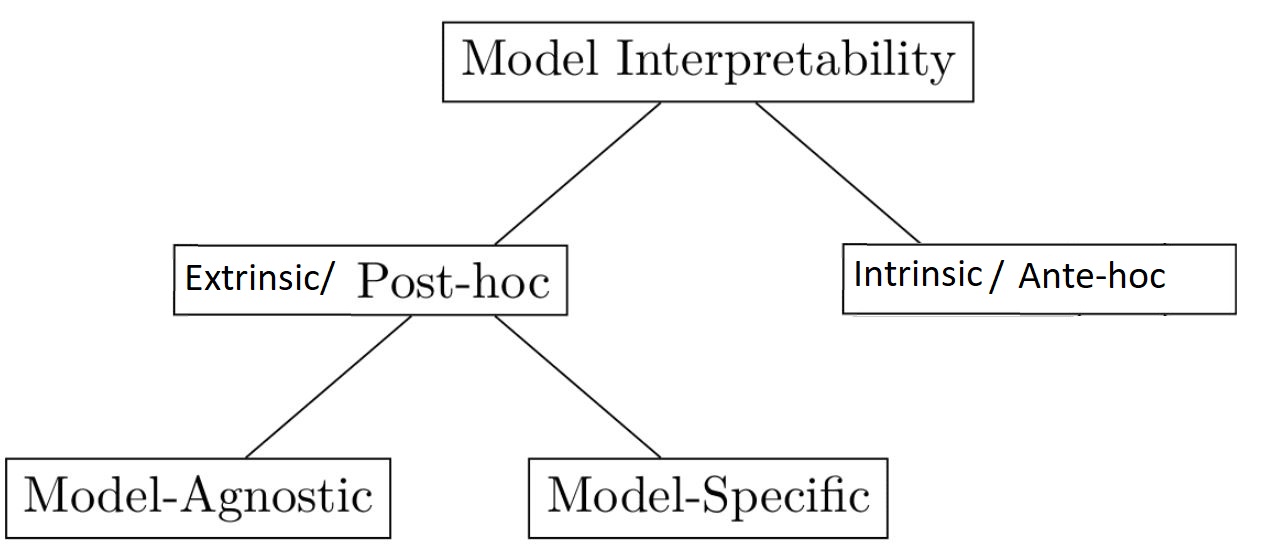
Taxonomy of all ML Algorithms:



**What the explanations should consist of? What to explain for?**

****

****

****

**Result of the interpretation method:**

– Feature summary statistic:

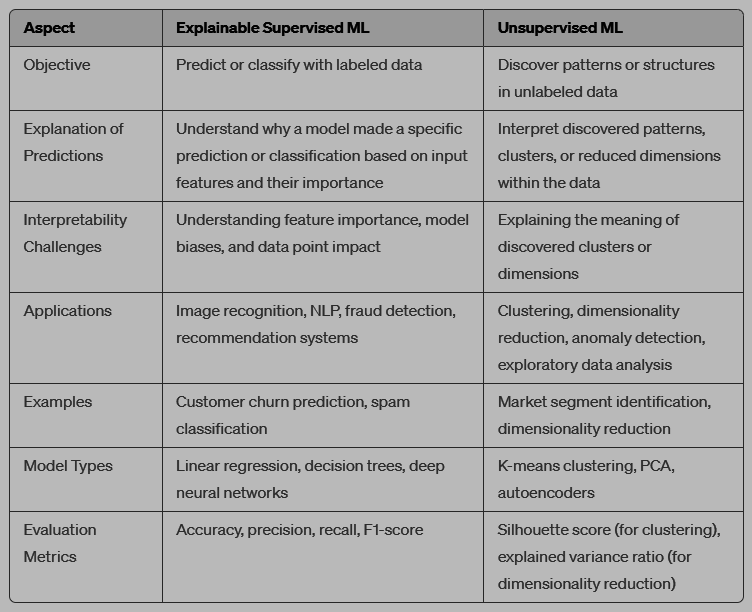
– Feature summary visualization

– Model internals

– Data point

– Intrinsically interpretable model

**Explainable Supervised vs. Unsupervised ML:**

****

**Why XUNSL:**

Unsupervised methods can be used for exploratory data analysis, identifying hidden structures, In situations where complex, multidimensional data requires and reducing dimensionality before applying SML for predictive tasks. Simplification and insight extraction, UML can be more useful in simplifying the problem for further analysis.

**2 approaches:**

1. Generate new algos

2. Modify existing algos: use existing XSL to interpret XUnSL (EXPLAIN-IT), process:

2.1. Clustering of input data, generating class labels.

2.2. Classifier trained on above

2.3. Explanation of classification using model Agnostic methods like **LIME** which stands for Local Interpretable Model-agnostic Explanations.

**Model agnosticism** refers to the property of LIME using which it can give explanations for any given supervised learning model by treating it as a ‘black box’ separately. This means that LIME can handle almost any model that exists out there in the wild!

**Local explanations** mean that LIME gives explanations that are locally faithful within the surroundings or vicinity of the observation/sample being explained.

**SOM** stands for Self-Organizing Map. It is a type of artificial neural network and a machine learning algorithm used for data visualization and dimensionality reduction. SOMs are also known as Kohonen maps, after their inventor, Teuvo Kohonen.

1. Unsupervised Learning: SOMs are a form of unsupervised learning, which means they don't require labeled data. Instead, they identify patterns and relationships in unlabeled data.

2. Topology-Preserving: SOMs are designed to preserve the topological properties of the input data. They map high-dimensional data to a lower-dimensional grid, where similar data points are placed close to each other on the grid.

3. Neuron Grid: A SOM consists of a grid of neurons or nodes. Each neuron represents a weight vector that corresponds to a specific location in the input space.

4. Training Process: During training, the SOM adapts its neuron weights to the input data. Neurons that are closer to a particular input vector adjust their weights to become more similar to that input.

5. Competitive Learning: SOMs use competitive learning, where the neuron whose weight vector is closest to the input is considered the winner. This neuron's weights are adjusted to better match the input.

6. Data Clustering and Visualization: SOMs are often used for data clustering and visualization. They can reveal underlying patterns and group similar data points on the map, making it easier to understand the data's structure.

7. Dimensionality Reduction: By projecting high-dimensional data onto a lower-dimensional map, SOMs can be used for dimensionality reduction. This simplifies data while preserving essential information.

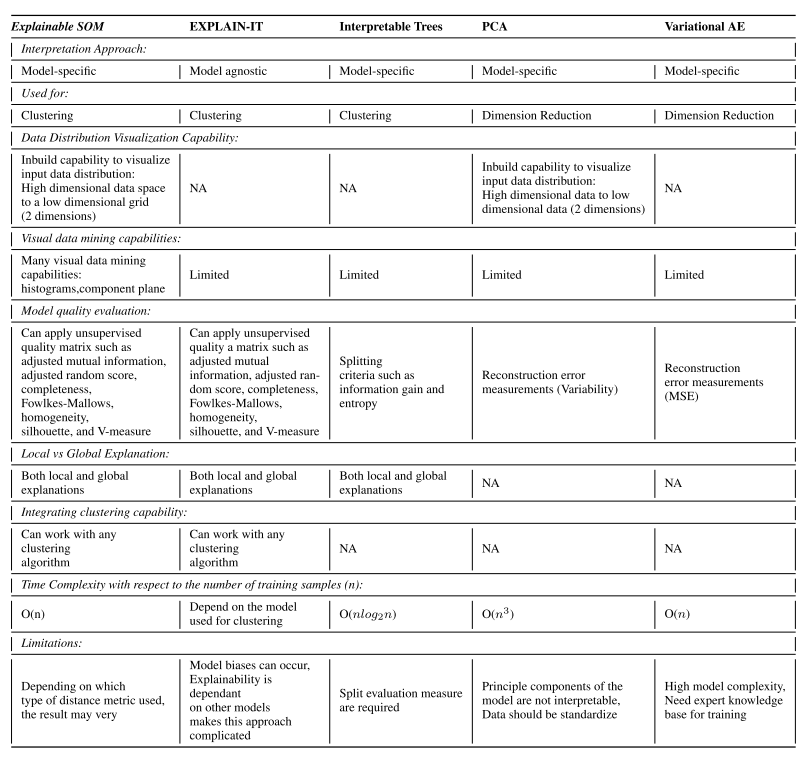
8. Applications: SOMs have been applied in various fields, including data mining, image processing, speech recognition, and exploratory data analysis. They are useful for tasks like customer segmentation, anomaly detection, and feature extraction.

9. Grid Topology: The grid topology of a SOM can be one-dimensional, two-dimensional, or even higher-dimensional, depending on the nature of the data and the desired output.

In summary, Self-Organizing Maps (SOMs) are a type of unsupervised neural network that can organize and represent complex data in a way that preserves the topological relationships between data points. They are particularly useful for data clustering, visualization, and dimensionality reduction.

**In this work:**

An Explainable Model Specific Algorithm focused on CPS for SOM to generate local and global interpretations was developed. Comparison with existing XUnSL algorithms was made.



A Variational Autoencoder (VAE) is a type of artificial neural network used in unsupervised machine learning and generative modeling. It is an extension of the traditional autoencoder architecture, designed to learn probabilistic representations of data. VAEs are particularly useful for generating new data points that resemble the training data, making them popular in applications like image generation and data compression.

**Using explainable SOM for CPS:**

1. Safety and security
2. Process optimization
3. Sales strategies
4. Generalizability
5. Real time operators

**Conclusion**

In this comprehensive literature survey, we have delved into the evolving landscape of unsupervised explainable artificial intelligence, with a particular emphasis on the application of Self-Organizing Maps (SOMs) in CPS.

**References**

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9536751>