International Journal of Scientific Research and Engineering Development-– Volume 7 Issue 5, Sep-Oct 2024
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

A Systematic Literature Survey on Machine Learning

Ranjith R*, Dr. Srinivasan Arumugam**
*(PG Scholar, CSE Department, Madha Engineering College, Chennai – 69.
Email: ranjithrr2001@gmail.com)
** (Professor, CSE Department, Madha Engineering College, Chennai – 69.
Email: a.srinivasan2025@gmail.com)

----------**********----------

Abstract:

This paperwork does a thorough literature review of the various types of machine-learning algorithms, classifying them as supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. This review gives a thorough overview of basic algorithms in each category, covering their theoretical foundations, practical applications, and current advances. This paper tries to provide comprehensive knowledge of the machine learning landscape by highlighting the advantages and disadvantages of various approaches through an analysis of significant research contributions and trends. This literature survey will be a useful resource for researchers, practitioners, and anybody interested in the current state and future directions of machine learning.

Keywords — **Machine Learning, Algorithms, Applications.**

----------**********----------

I. INTRODUCTION

The rapid development of machine learning attracts everyone's attention and becomes a key concept in every field of research and enhancements across various domains are made by developing systems to learn from data and improve themselves without any external interventions [1]. The field of artificial intelligence known as "machine learning" is primarily concerned with creating different algorithms that can learn from data. Unlike traditional programming, where instruction need to be fed explicitly, machine learning models can identify patterns and can make decisions based on the inference. Machine learning algorithms are broadly classified into supervised, unsupervised, semi-supervised, and reinforcement learning [14].

II. CLASSIFICATION OF DATA

In this section, we'll discuss what data is and how it is classified. Data is a unique piece of information, usually combined in different ways to produce different meanings. It is collected, analysed, measured, and reported, and can often be represented using graphs, images, and other analysis tools. Raw data can be a form of number or character collection before it's been cleaned and corrected by data engineers or analysts. It should be corrected so that we can remove outliers. This process of cleaning the data is known as 'Data Pre-processing', and the data filtered from this process is called 'Processed Data' or 'Cleaned Data'. Various forms of data can exist, such as structured, semi-structured, or unstructured.

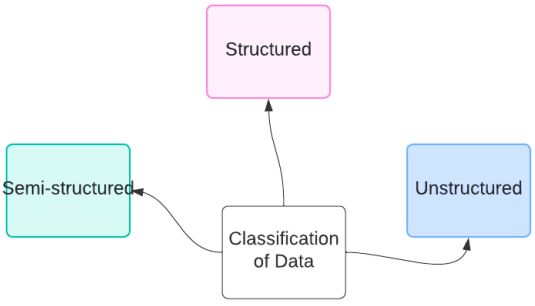

Figure 1 Structure of Data

A. Structured Data

We can see this kind of data in our day-to-day office lives, such as structured data in a form of spreadsheets, relational databases, etc. Structured data has a well-defined structure and follows some standard order that is highly organized and accessed efficiently, such as rows and columns in a table. Using such data, ML models can be fitted almost perfectly, and their prediction scores will also be high.

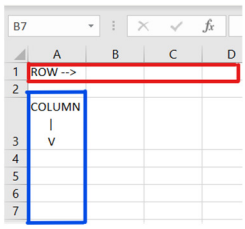

Figure 2 Spreadsheet - Excel

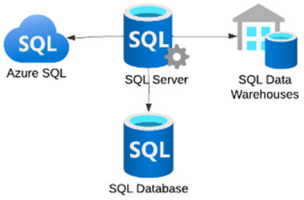

Figure 3 Relational Database

B. Unstructured Data

Unstructured data, which does not at all follow any order, can't be fit into some organized way that we do with structured data. We can otherwise refer to this kind of data as raw data. It often represents data in a format that can't be searched for or organized into traditional databases, such as images, texts, audio, and videos.

Figure 4 Images

Figure 5 Texts

Figure 6 Audios

Figure 7 Videos

C. Semi-structured Data

Typical examples of semi-structured data include HTML tags, XML tags, NoSQL databases that follow a JSON format, etc. This is also a kind of data that doesn't fit in an organized way like structured data does. Though it follows some hierarchical order to look like it is organized.

Figure 8 HTML

Figure 9 XML

Figure 10 JSON

III. CLASSIFICATION OF MACHINE LEARNING TECHNIQUE

How the source (data) has some classification, likewise we do have categories in ML algorithms which were classified based on its learning methodology, the type of dataset, the size, the presence of outliers, and many factors are there [12]. In this section, we are going to explore those algorithms and it's learning nature.

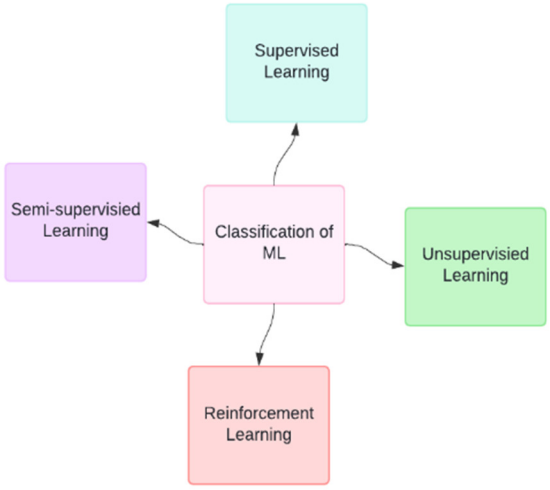

Figure 11 ML Classification Chart

A. Supervised Learning Technique

The name 'supervise' means someone or something will be watched by someone. If you think in an ML way, we can say a machine learns a thing under some supervision. The exact definition for Supervised ML would be the ML model trains on the labelled dataset, and each training examples are paired with its output label, then the goal is for the model to learn a way to efficiently map the input features to their output labels so that it can be able to predict the label for new or unseen data. So many algorithms are classified under this category

for training and fitting a model for the given dataset, we are going explore a few such algorithms here.

1) Linear Regression: This algorithm is considered to be very simple and easily understandable to everyone. The main picture here is that this algorithm tends to find a relationship between a dependent (Y) and an independent (X) variable. We all know the line equation, i.e., Y = mX + c, where m is a slope and c is an intercept. Say, for example, if you are predicting a house price Y based on the given value X, the plotted regression line can show the price increase for every sq. ft. of house size. The ultimate goal is to find the value of m and c so that the line fits most of the data points as close as possible.

2) Support Vector Machine: This algorithm is mostly used for classification but also used for regression. It attempts to find the best boundary called a hyperplane that separates the data points into classes. Say, for example, you have two data groups plotted in a graph and draw a hyperplane that maximum separates the group with the maximum distance between the data points and the plane. These closest data points are called support vectors; hence, this algorithm has the name.

3) Random Forest: This ML algorithm falls under the category of ensemble learning and can be used in classification and regression tasks. This approach primarily consists of creating numerous decision trees and combining their predictions to increase accuracy and reduce overfitting. This algorithm is considered to be powerful for its feature of reducing overfitting. It can produce more accurate predictions by leveraging the diversity of multiple decision trees.

4) Decision Tree: This algorithm is pretty straightforward, as it creates a tree-like model to make decisions. Based on the features or attributes present in the given data, it splits the data to develop branches, with each leaf node representing a decision. This algorithm is the base for the Random Forest algorithm. Say, for example, that predicting the voting of a person by giving their age and salary as input will make a decision that if the age is greater than [18], the decision leads to an outcome that voting is possible.

***5) K – Nearest Neighbour (KNN)*:** This algorithm is widely used for the problems of classification, as it attempts to classify the data points based on their proximity. The notation 'K' denotes the number of neighbours that the algorithm needs to take to compare the Euclidean distances between them, and hence a new data point can be classified as one of the categories to which the Euclidean distance between the K neighbours is less.

Naive Bayes, Gradient Boosting Machines (GBM), Artificial Neural Networks (ANN), Ridge Regression, Lasso Regression, Elastic Net, AdaBoost, Perceptron, Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA), Stochastic Gradient Descent (SGD), Polynomial Regression are some of other supervised algorithms.

B. Unsupervised Learning Technique

It is a type of ML where the model is trained on data where it doesn't have labelled outputs like in the supervised type. This type of model finds the underlying hidden patterns, structures, or relationship in the data instead of learning from pre-defined data. This type of technique is most commonly used for exploratory data analysis (EDA), to understand the structure of the data.

1) K – Means Clustering: This algorithm is used for clustering data into different groups based on their similarity. The main goal is to group the data points into K clusters, each of which has a centroid – the cluster with the closest mean and each individual data point belongs to one of those clusters. Centroids are nothing but a randomly picked data point (by some methods). Now you can assign each and every data points to one cluster by comparing its Euclidean distance, thus forming K clusters. These above steps were repeated till some halting conditions were met. Once then, final clusters are formed. This algorithm is widely used for customer segmentation, document categorization, and image compression.

2) Hierarchical Clustering: This algorithm is also used for clustering the similar data points. Unlike the K-means seen above, where the number of clusters K is defined at the start, here this clustering builds a tree-like structure called a dendrogram that represents a data point in a hierarchy, allowing us to decide the number of clusters at any time by cutting the tree at the closest level. Also, there are two types of this clustering, such as Agglomerative clustering (bottom-up) and Divisive (top-down).

3) Principal Component Analysis (PCA): It is also known as the dimensionality reduction technique, which is used to simplify the datasets with many variables (features) while keeping as much needed information (variance) as possible. This algorithm transforms the dataset into a new set of variables called principal components, which are linear combinations of the original data. This algorithm helps in removing data redundancy and overfitting problems.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Models (GMM), Independent Component Analysis (ICA), t-Distributed Stochastic Neighbour Embedding (t-SNE), Autoencoders, Self-Organizing Maps (SOM), Apriori Algorithm (for association rule learning), FP-Growth Algorithm (Frequent Pattern), Affinity Propagation, Mean Shift Clustering, Latent

Dirichlet Allocation (LDA) (for topic modelling), Isomap, Spectral Clustering, Non-negative Matrix Factorization (NMF), Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH clustering) are some of the other unsupervised algorithms.

C. Semi-Supervised Learning Technique

In this type of ML, the model is initially trained on a set of labelled data to understand the core relationship and then again refines itself by fitting with the set of unlabelled data to find the underlying pattern and hidden relationships. This iterative process between labelled and unlabelled data makes the model give better accuracy on abundant unlabelled data prediction.

1) Self-Training: The main idea of this algorithm is to train on the labelled dataset iteratively and attempt to predict the labels of the unlabelled dataset. Say, for example, that to predict whether an email is 'spam' or 'not spam', train the labelled dataset with the help of a decision tree algorithm, then use the classifier to predict the unlabelled and unseen data. This algorithm can be used if you are dealing with a less-labelled dataset and trying to predict a huge amount of unlabelled data.

2) Generative Models: This algorithm aims to model a distribution of the data itself, and these models try to discover and learn the underlying pattern and data generation process; hence, this can be used as a semi-supervised learning algorithm by combining the labelled and unlabelled data. This generative model learns the joint probability distribution, P(X|Y), where X is the feature and Y is the label. The ultimate goal is to learn P(Y|X), the probability of a label given an observation. Thus, this model estimates the P(X) data distribution with the use of unlabelled data and estimates the label distribution P(Y|X) with the use of labelled data.

3) Label Propagation: It is a graph-based algorithm, where the data points are represented as nodes, and edges between nodes represent the similarity between points. Labelled nodes (data points) propagate their labels to the neighbouring unlabelled node over the graph. This process iterates until the labels are stabilized, with each unlabelled node being assigned a label based on its neighbours. This algorithm suits very well when the data has a natural grouping ability, like social networking or a recommendation system.

Co-Training, Gaussian Mixture Models (GMM), Transductive Support Vector Machines (TSVM), Semi-Supervised k-Means Clustering, Label Spreading, Graph-Based Semi-Supervised Learning, Laplacian SVM, Graph Convolutional Networks (GCN), Multi-View Learning, Consistency Regularization, Mean Teacher Model, Pseudo-Labelling, Entropy Minimization, Tri-Training, Generative Adversarial Networks (GANs) for Semi-Supervised Learning are some other semi-supervised algorithms.

D. Reinforcement Learning Technique

This learning type is considered to be the most powerful learning technique in ML for problem-solving that involves sequential decision-making in a real-time environment. It can learn from the direct feedback in the form of penalties or rewards and interaction, which are most crucial here. Unlike supervised and unsupervised learning, which focus on learning from static datasets.

1) Q: Learning: This algorithm makes the model learn a Q-value (or action value), which guides the agent to the cumulative reward for taking a certain action in a specific state [8]. This algorithm is used to obtain the optimal action-selection policy for an agent interacting with an environment. This algorithm doesn't require a model to learn about the environment; instead, it directly learns from the agent's experiences by updating the Q-value. The ultimate aim is to find and tell the agent to take what action in each state that'll maximize the overall reward over time.

2) Monte-Carlo Tree Search (MCTS): This algorithm makes decisions by simulating future scenarios and selecting actions based on these simulations. This algorithm attempts to maximize the rewards through trial and error in an environment. This is an iterative algorithm that searches through possible future outcomes in a tree-like structure. The minimum idea is to simulate random games or actions from a given state and then use the results of this simulation to make better decisions on its next run.

IV. GLIMPSE OF ARTIFICIAL INTELLIGENCE AND DEEP LEARNING

Artificial intelligence is one of the foremost fast-developing fields of computer science, specially focused on creating machines that can be capable of performing human tasks and even impossible things by humans [2]. The tasks can be of problem-solving, learning, natural language understanding, perception, and decision-making. Machine learning and Deep learning are considered to be the subsets of AI. It encompasses technologies and approaches, from rule-based systems to expert systems.

Deep learning is a subset of ML, which is again a subset of AI. This was invented by scientists to inspire how our human brain cell works. DL was

also known as an Artificial Neural Network (ANN). How our human brain contains n-no layers of neurons, like we are here creating artificial neurons that can process data in some hierarchical manner, allowing a model to learn even more complex patterns that a normal ML algorithm won't fit.

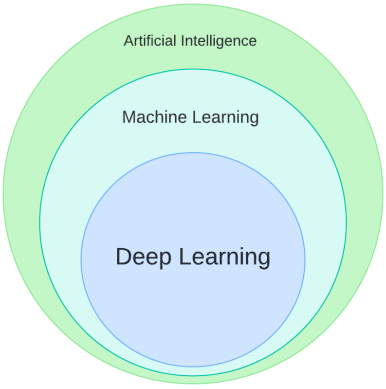

Figure 12 Subset Chart of AI, ML, DL

V. REAL-TIME APPLICATIONS

If we talk about applications, it never ends. The truth is ML is spreading over every industrial domain and gives a better application than a human can give. In which, we'll discuss about few such domains and their ML applications [15], [16].

A. Healthcare

1) Medical Imaging: ML algorithms can be able to analyse medical images [3] like X-rays, CT scans, MRI scans, etc., thereby assisting doctors in real-time by diagnosing conditions like cancer, fracture, heart disease, TB cells, etc.

2) Personalized Medicare: ML helps individual patients' conditions by learning their past medical history data, genetic data and providing them personalized analytics and even support by recommendations on future things [14].

3) Predictive Analysis: ML also helps in predicting things such as the likelihood of a patient's readmission and progression of the disease, supporting doctors promptly.

A. Accounts or Finance

1) Fraud detection: The ML model can be able to analyse real-time data transactions to detect unusual activities and reports to the customer and services, thereby preventing fraud [10].

2) Algorithm Trading: Today, ML models can be able to analyse daily stocks and find trends and patterns [9]. Thereby predicting the market value of a particular stock in a particular period. Investors and stock market analysts find this to be highly beneficial. Even with this, people are starting small businesses too. So far, this application ML has reached and is still developing [5], [11].

B. Cybersecurity

1) Threat Detection: ML models can be able to monitor the network traffic [6] and system logs in real-time to detect and respond to cyber-attacks such as phishing, malware, etc.

2) Intrusion Detection System: ML models can also identify unusual activities [7] and anomalies happening in the network by mitigating security risks before they break into systems.

VI. CHALLENGES AND ENHANCEMENTS

The field of ML, though reaches almost every industrial domain and performs tasks that can't even a human can imagine, yet it is still considered a developing field. Here are some challenges that are currently faced by the ML.

A. Quality and Quantity of Data

High-quality labelled data is very essential for training effective ML models. However, obtaining such data is very expensive and time-consuming. In some datasets, certain categories are underrepresented, leading to a biased model with very poor accuracy. Noisy, incomplete, or inaccurate data can largely impact the ML model's accuracy.

B. Overfitting and Underfitting

ML Model prediction score depends on dataset quality and training quality too. Training makes the model perform well or poorly. In such case, the word overfitting means a model performs well for the training data but not for the testing/unseen/new data. Similarly, underfitting means the model fails to train well and yields a poor performance

Other than these, even many more factors that are affecting the ML performance.

C. Enhancements

*1) **Explainable AI (XAI):*** It is a method and process of making ML models more interpretable and transparent, especially in critical areas like healthcare, finance, etc., with Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) to trust the results given by the models.

*2) **Federated Learning:*** It is a technique used to train ML models across multiple devices or servers without the data being exchanged or centralized instead of keeping a local data sample. This approach enhances privacy and security when it comes to dealing with sensitive data. It often includes Edge AI, where the models are built using edge computing and deployed on edge devices such as smartphones, IoT devices, etc.

*3) **Green AI:*** With the developing concern about the environmental impact of large-scale ML models, there is a hustle toward developing more energy-efficient algorithms and models [13].

VII. CONCLUSION

This paper has gotten us through the fundamental concepts of ML precisely enough by covering its diverse algorithms, and practical applications. We've dealt with the field of its impact in various industrial domains and challenges in today's trend. We also discussed the future enhancements on how ML is going to solve complex tasks that humans can't even imagine in action. Hence, the insights gained from this survey will be a valuable guide for researchers and ML practitioners aiming to develop real-time innovative solutions, thereby contributing to the wellness of society.

REFERENCES

[1] L. Capogrosso, F. Cunico, D. S. Cheng, F. Fummi and M. Cristani, et al. "A Machine Learning-Oriented Survey on Tiny Machine Learning," in IEEE Access, vol. 12, pp. 23406-23426, 2024, doi: 10.1109/ACCESS.2024.3365349.

[2] Jannani, N. Sael and F. Benabbou, et al. "Artificial Intelligence for Quality of Life Study: A Systematic Literature Review," in IEEE Access, vol. 12, pp. 62059-62088, 2024, doi: 10.1109/ACCESS.2024.3395369.

[3] Firas H. Almukhtar, Shahab Wahhab Kareem, Farah Sami Khoshaba, et al. Design and development of an effective classifier for medical images based on machine learning and image segmentation, Egyptian Informatics Journal, Volume 25, 2024, 100454, ISSN 1110-8665, https://doi.org/10.1016/j.eij.2024.100454.

[4] Jensen, Benjamin, et al. Algorithmic Stability: How AI Could Shape the Future of Deterrence. Center for Strategic and International Studies (CSIS), 2024. JSTOR, http://www.jstor.org/stable/resrep60683. Accessed 29 Sept. 2024.

[5] Koshiyama Adriano, Kazim Emre, Treleaven Philip, Rai Pete, et al. Towards algorithm auditing: managing legal, ethical and technological risks of AI, ML and associated algorithms R. Soc. Open Sci.11230859, 2024, http://doi.org/10.1098/rsos.230859.

[6] Methaq A. Shyaa, Noor Farizah Ibrahim, et al. Evolving cybersecurity frontiers: A comprehensive survey on concept drift and feature dynamics aware machine and deep learning in intrusion detection systems, Engineering Applications of Artificial Intelligence, Volume 137, Part A, 2024, 109143, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2024.109143.

[7] Muath Asmar, Alia Tuqan, et al. Integrating machine learning for sustaining cybersecurity in digital banks, Heliyon, Volume 10, Issue 17, 2024, e37571, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2024.e37571.

[8] Qian Xie, et al. Application of Q-learning algorithm based on heterogeneous sensor networks in higher education teaching management, Measurement: Sensors, Volume 33, 2024, 101188, ISSN 2665-9174, https://doi.org/10.1016/j.measen.2024.101188.

[9] Rihab Najem, Meryem Fakhouri Amr, Ayoub Bahnasse, Mohamed Talea, et al. Advancements in Artificial Intelligence and Machine Learning for Stock Market Prediction: A Comprehensive Analysis of Techniques and Case Studies, Procedia Computer Science, Volume 231, 2024, Pages 198-204, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2023.12.193.

[10] Surendranadha Reddy Byrapu Reddy, Praneeth Kanagala, Prabu Ravichandran, et al. Effective fraud detection in e-commerce: Leveraging machine learning and big data analytics, Measurement: Sensors, Volume 33, 2024, 101138, ISSN 2665-9174, https://doi.org/10.1016/j.measen.2024.101138.

[11] Latrisha N. Mintarya, Jeta N.M. Halim, Callista Angie, Said Achmad, Aditya Kurniawan, Machine learning approaches in stock market prediction: A systematic literature review, Procedia Computer Science, Volume 216, 2023, Pages 96-102, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2022.12.115.

[12] Oleiwi, Safa & Jawad, Zahraa & Alzuaidi, Yasser. (2023). LITERATURE REVIEW ON THE VARIETY KINDS OF AI AND ML TECHNOLOGIES CLASSIFICATIONS AND LIMITATIONS. 17. 80-95.

[13] Verdecchia, R., Sallou, J., & Cruz, L. (2023). A systematic review of Green AI. WIREs Data Mining and Knowledge Discovery, 13(4), e1507. https://doi.org/10.1002/widm.1507.

[14] Brnabic, A., Hess, L.M. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. BMC Med Inform Decis Mak 21, 54 (2021). https://doi.org/10.1186/s12911-021-01403-2.

[15] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x.

[16] Barenkamp, M., Rebstadt, J. & Thomas, O, et al. Applications of AI in classical software engineering. AI Perspect 2, 1 (2020). https://doi.org/10.1186/s42467-020-00005-4.

