A Comprehensive Review to Understand the Definitions, Advantages, Disadvantages and Applications of Machine Learning Algorithms

Md. Jamaner Rahaman
Department of Computer Science and Engineering
Leading University
Sylhet, Bangladesh

ABSTRACT

Machine learning (ML) means that first the machine learns with the help of algorithms then works automatically. In today's age people want to do almost everything automatically and efficiently. In that sense machine learning has made a revolutionary change because of its efficiency. An intelligent machine works faster than the human. The incidence of errors is conspicuously decreased by using ML. Depending on improving the necessity of ML algorithms in the present situation this paper tried to describe some ML algorithms especially supervised, unsupervised, semi-supervised and reinforcement learning including their definitions, advantages, disadvantages and area of work so that the people will understand which algorithm where to use. Particularly Support Vector Machine (SVM), Decision Trees, K-Nearest Neighbors (K-NN), Linear Regression, Logistic Regression for supervised learning. K-Means Clustering, Principal Component Analysis (PCA) for unsupervised learning. Basics of semi-supervised learning and reinforcement learning. Eventually from this paper people can easily get the idea of commonly used machine learning algorithms.

Keywords

Machine Learning, Support Vector Machine (SVM), Principal Component Analysis (PCA), Semi-Supervised Learning, Reinforcement Learning.

1. INTRODUCTION

In the near future, humans will no longer be able to think without machine learning. It is evident that machine learning (ML) is becoming more and more important every day due to its effectiveness. For example if anyone saw the videos on YouTube that time some more relatable videos suggested to us. This also has the impact of machine learning, because backend algorithms analyze the user browsing data and identify the user preference then shows the videos what users like. Machine learning basically knows the mechanism to manage data when it is taught [1]. Using ML the life of humans is easier than any of earlier eras.

Machine learning is to learn machines via algorithms and produce the output of what the machine has learnt. Sometimes a machine acts like a human, that means learning itself and doing it by itself. But sometimes it acts in the area of its expertise. For the bases of that machine learning (ML) algorithms have different types: supervised, unsupervised, semi-supervised and reinforcement learning. Evolutionary learning and deep learning are also part of machine learning [2]. Evolutionary learning is extensively used in production related work such as manufacturing, agriculture, power and energy etc. Deep learning also has different types of algorithms for example neural networks based on the nervous system and brain of a human [3]. One important thing has to be remembered: the performance of any machine learning technique depends on a good dataset that means quality and authenticity of data, and also selection of the suitable algorithm for the appropriate domain. So that it is necessary to know the behavior of all machine learning algorithms for applying the correct algorithm to the correct situations [4]. Usually the dataset is divided into two parts: training and testing then pre-processing, implementation of algorithm and result analysis.

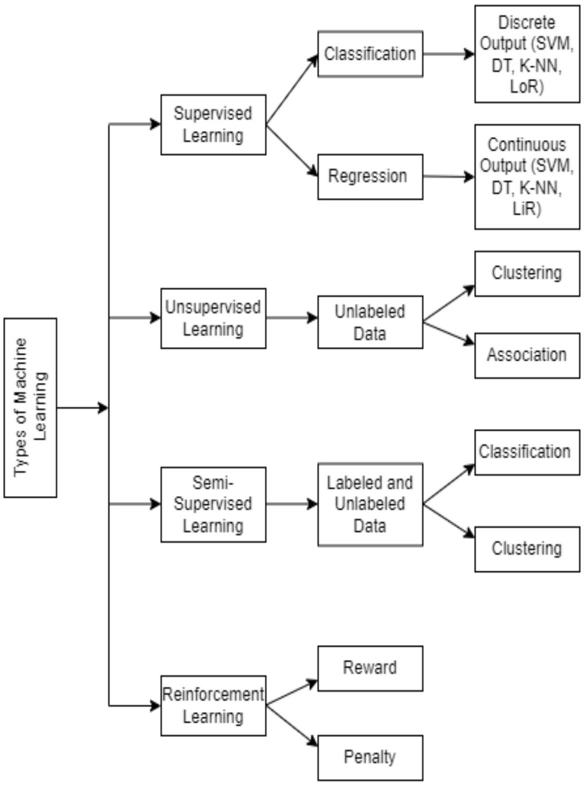

Fig 1: Types of Machine Learning

Early detection of disease is important to reduce human mortality. In recent era especially in healthcare system highly used machine learning to predict different types of diseases. In a traffic control system people also can use ML, first detect the number plate of a vehicle then processing the collective data with the help of ML, so that it can easily identify the fraud vehicle, and also decrease the road accident. Most importantly by using ML people can develop automatic traffic control systems. In the agriculture sector, using the ML to easily detect the vegetable disease, it is also a blessing for the farmer to protect their crop. At the same time robotics, natural language processing, pattern recognition, data analysis and many more things can be done correctly with the help of ML [5]. So that the human errors are reduced significantly. Without artificial intelligence, especially ML, the fourth industrial revolution can not even think. Everything has its pros and cons but people will have to ensure that machine learning is in good use so that the society will be benefited. Sometimes some of the people do not know the uses of ML algorithms that is why they are not interested to do work with ML. For those people the author attempts to investigate some ML algorithms to introduce the primary purposes of ML. The following few sections will describe several machine learning (ML) algorithms review.

2. SUPERVISED LEARNING

Supervised learning basically works on labeled data. Labeled data means that a dataset contains the input data at the same time of its output or target values also which a machine is supposed to predict. Supervised machine learning produces the output and matches with the target values that are already given to the dataset. The performance measurement of supervised machine learning depends on similarities with the predictive results and the target values of the dataset. Supervised learning algorithms also stop the learning when the expected result will come [6]. It has two types of categories such as classification algorithms and regression algorithms. In classification problems the output variable will be discrete, on the other hand in a regression problem the output variable will be continuous [7], [8]. Some commonly used supervised machine learning algorithms will be discussed below.

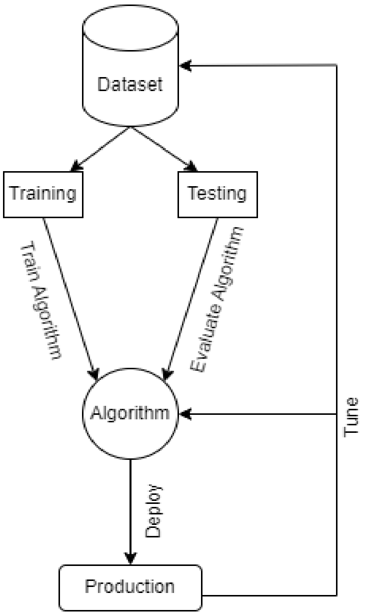

Fig 2: Workflow of Supervised Learning

2.1 Support Vector Machine (SVM)

One of the most popular and powerful supervised machine learning algorithms is SVM. It supports both classification and regression problems of supervised learning. The main goal of this technique is separating the different classes' objects by creating a hyperplane between classes. It is also called the decision boundary. Sometimes objects are not able to separate linearly so that the difficult mathematical situation will arrive and it is called a linearly non separable dataset that time a different way to solve this problem which is kernel [5]. Distance between any classes to the decision boundary is called margin. Maximizing the margin distance for both classes is the best way to avoid misclassification. It can be said that maximizing the margin distance minimizes the classification error [7]. The major advantage of SVM is that you can use this technique for big-dimensional data. Another one is if you can identify the correct hyperplane it produces better performance [9]. Disadvantages of SVM are struggle to identify accurate kernel function. In the case of noisy data, SVM does not work well. Credit card fraud detection, diagnosis of cancer, handwriting recognition, face detection and text categorization are the applications of SVM [5].

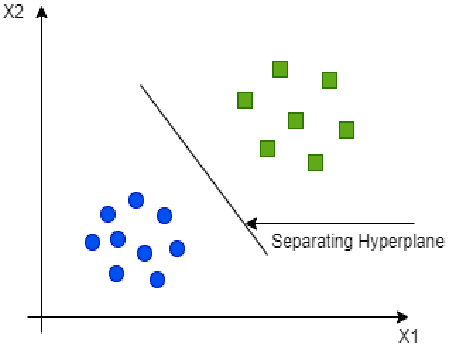

Fig 3: Support Vector Machine

2.2 Decision Trees (DT)

Among the supervised machine learning algorithms, the decision tree (DT) is one of the earliest and prominent learning algorithms. It allows classification and regression problems also. Decision variable is categorical in the classification tree. In regression tree decision variables are continuous [5]. DT has its route node, internal node and leaf node. The topmost node is called a route node that has no edges. It represents the whole dataset and is split into two or more identical sets. Internal nodes represent the feature attributes of the dataset. Data split in these points. The leaf node is not able to divide again. It represents the class labels or final outcome [7]. In DT continuously splitting the data from the route node into subsets based on the input features of the value and the process is repeatedly done for each child node until the leaf node will arrive. The nodes represent the choices and the edges represent the conditions [1]. The advantage of the decision tree is that it is used for both categorical and numerical data, easy to understand, less data preprocessing and faster [9]. Disadvantage of the decision tree is that it is unbalanced, giving local optimal solutions instead of globally optimal solutions [11]. Determining galaxy counts, control systems, financial analysis, detecting use of library books in future, predicting tumor problems are the applications of decision trees [5].

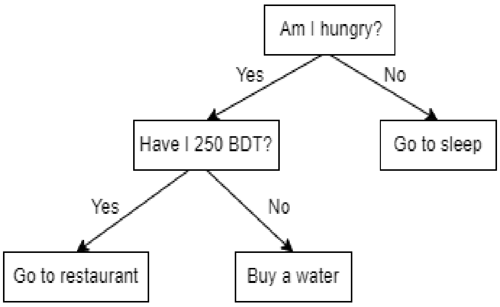

Fig 4: Decision Tree

2.3 K-Nearest Neighbors (KNN)

It is a simple but powerful supervised machine learning algorithm which is used for both classification and regression problems. Two major characteristics of KNN is that it is non-parametric which does not make assumptions about the distribution of the underlying data and lazy learning algorithm means that it does not train a model that is explicitly defined. The training data is saved and predictions are based on the k most similar neighbors (instances) in the dataset, it is the basic working principle of KNN. On the other hand it can be said that predictions are made directly by memorizing the training examples. The value of k is the most important part of KNN. Overfitting can occur with a small k, while a large k can smooth out predictions, but may also underfit [10]. Advantages of KNN, training is faster because there is no explicit training phase, implementable and easy to comprehend, it is applicable to both classification and regression tasks. Disadvantages of KNN, having irrelevant or redundant features in the dataset can cause performance degradation, large datasets require a high computational cost, kNN's performance can be affected by the k and distance metrics chosen. Applications of KNN include credit rating based on feature similarity, handwriting recognition, medical diagnosis of various diseases with similar symptoms, and recommendation systems [5], [11].

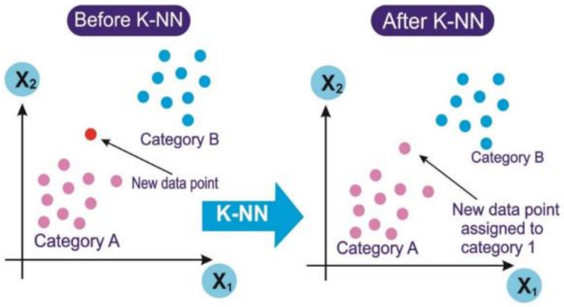

Fig 5: K-Nearest Neighbors [11].

2.4 Linear Regression

The fundamental and most simple supervised machine learning algorithm used in statistics is called linear regression. It follows the regression technique to model continuous variables [12]. Dependent variables (target variable) and independent variables (features) are also the main part of linear regression because the relationship between them is modeled by fitting a linear equation to observed data [13]. That means it can be said that linear regression worked with a labeled dataset where the output variable value depends on the input variable value [5]. Basically it works on prediction. Advantages of linear regression is that small to moderately large datasets are particularly well-suited for computational efficiency, easy to understand. Disadvantages of linear regression is that when the relationship is not truly linear and the feature is too many then overfit the data. Prediction of house price, students exam scores are the applications of linear regression.

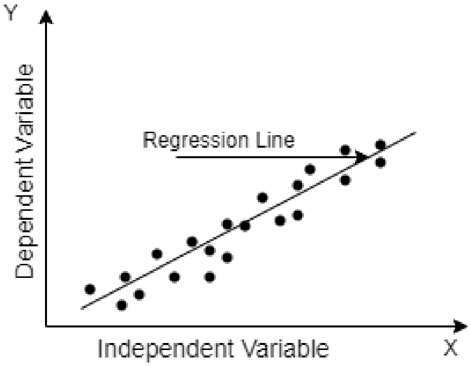

Fig 6: Linear Regression

2.5 Logistic Regression (LR)

It is a powerful and most commonly used algorithm for binary classification tasks. It is also called a classification algorithm which gives the categorical outcome. It is also frequently used in statistics and discrete data processing [14]. A threshold must be assigned to distinguish between two classes because it is a probability so that in logistic regression the predictive result or outcome will be 0 or 1 [15]. If the value between 0 and 0.5 (not exactly the 0.5) then the outcome is 0 and when the value is 0.5 or between 0.5 and 1 then the outcome is 1. Advantages of LR: simplicity, computational efficiency, training efficiency. Presents the probabilities associated with the class labels, which can be useful for making decisions. Disadvantages of LR: unable to solve nonlinear problems. Without identification of independent variables, working efficiency will be low. Spam detection, cancer diagnosis are the applications of logistic regression.

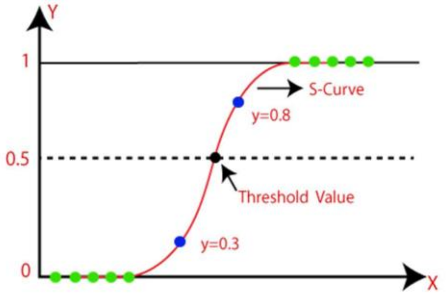

Fig 7: Logistic Regression [21].

3. UNSUPERVISED LEARNING

When the algorithms work with the unlabeled data then it is called unsupervised learning means that the target value is not defined, whereas in supervised learning people saw that the output value or predictive result need to be matched with the target value already defined in the dataset. Unsupervised machine learning algorithms trained by data and produce the output from the knowledge gained from the training that means it's up to the system to produce the output. In that case a large amount of data is required to train for unsupervised learning. The performance also depends on the large amount of data it has in unsupervised learning. Unsupervised learning used to find out the relationship in data where the outcome is unlabeled. Identify structures and patterns in data also the working principles of unsupervised learning. Types of unsupervised learning are clustering, association, anomaly detection, and autoencoder. Separating items into several groups is known as clustering. Discovering the relationship between variables in large datasets is called associations. Identifying the unusual patterns in the dataset is known as anomaly detection. Autoencoder used in neural networks for representation learning [16]. Below some popular unsupervised learning algorithms will be discussed.

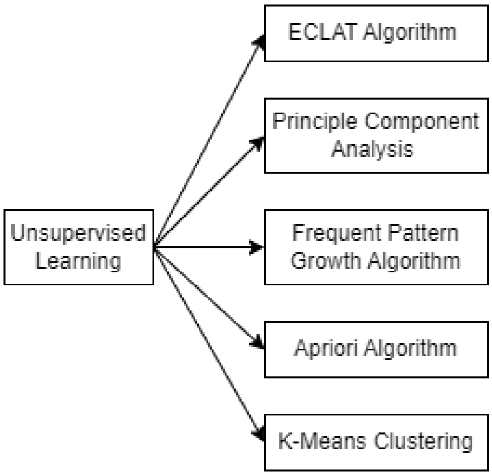

Fig 8: Types of Unsupervised Learning

3.1 K-Means Clustering

It is one of the most popular and simplest algorithms of unsupervised learning. K-Means clustering is basically used for clustering problems that means to differentiate the objects by their similarity into different groups or clusters [17]. By clustering, finding the similarities in the dataset and unlabeled data converted to a unique cluster. After clustering, it needs to find out the centroids or k-centers for each cluster. How many groups or clusters produced it shows the value of “k” [16]. Advantages of K-Means clustering is that it is easy to implement, in the sense of hierarchical clustering. K-Means clustering is faster and produces smaller groups. Disadvantages of K-Means clustering is that it is hard to find out the value of k, performance decreased because of different size of clusters [5]. Applications are compression of images, document classification, customer segmentation.

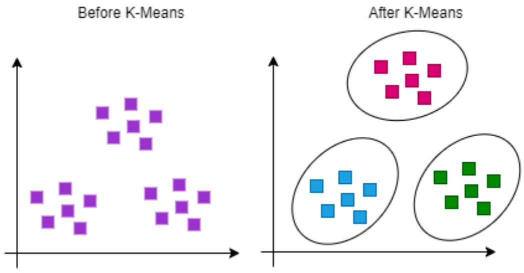

Fig 9: K-Means Clustering

3.2 Principal Component Analysis (PCA)

Whenever people talk about the reduction of dimensionality of a large dataset, PCA comes first. It reduces the dimensionality but the major information is still available in the dataset. It basically converts many variables into smaller groups so that it is easier to analyze the data and faster the performance. Advantages are simple calculation, terminate the noise, faster computation and improve the performance. Disadvantages are information loss by reducing dimensionality, difficult to interpret for its linear combinations of original features. Applications are computer vision, image compression etc.

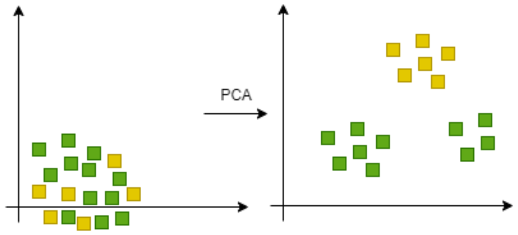

Fig 10: Principal Component Analysis

4. SEMI-SUPERVISED LEARNING (SSL)

Semi-supervised learning working with the both labeled and unlabeled dataset. The objective of semi-supervised learning is to classify the test data from the labeled and unlabeled data but first resolve the limitations of labeled and unlabeled data [18]. So it is clearly seen that semi-supervised learning is the combination of supervised and unsupervised learning where supervised learning works with the huge number of labeled data and unsupervised learning works with the unlabeled data so that it does not have any prior knowledge of target value. The main drawback of unsupervised learning is that it is unable to classify or cluster unknown data accurately. To overcome this problem semi-supervised learning proposed a model with few labeled patterns for training data and the other patterns used for test data [19]. Semi-supervised learning again divides into two types such as semi-supervised classification and semi-supervised clustering. Commonly used semi-supervised learning techniques are Self-Training, Co-Training, Graph-Based Methods, Generative Models. The advantages of semi-supervised learning is that it is cost-effective because of reducing the large amount of labeled data, and improved performance because of the combination of the labeled and unlabeled data. The disadvantages of semi-supervised learning is that it is complex to implement, and a big amount of unlabeled data needs to be processed in SSL which is hard. Applications of semi-supervised learning are text classification, image recognition, medical diagnosis, speech recognition etc.

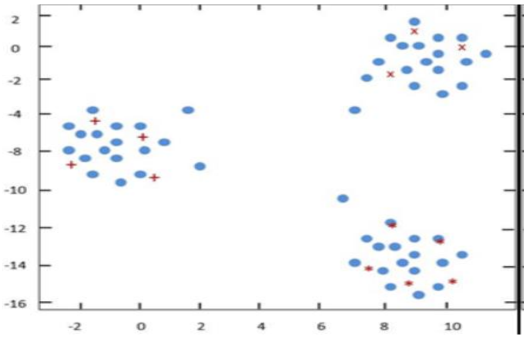

Fig 11: Semi-Labeled Dataset [18].

5. REINFORCEMENT LEARNING (RL)

Reinforcement Learning is a software agent which interacts with the surrounding environment to improve the performance for achieving the cumulative reward [20]. The objectives of RL are to maximize the rewards and develop a procedure for acting optimally within the environment for various situations. Learning from the consequences of action is the target of RL. There are several types of RL such as Q-Learning, SARSA, Deep Q-Network (DQN) etc. The advantage of RL is continuously learning from the environment and increasing the performance. It does not need any labeled data but it learns from the actions through rewards and penalties which is more practical. The disadvantages of RL are that it is computationally complex, so many times it is necessary to interact with the environment which is time consuming. Applications of RL are games, robotics, autonomous systems etc.

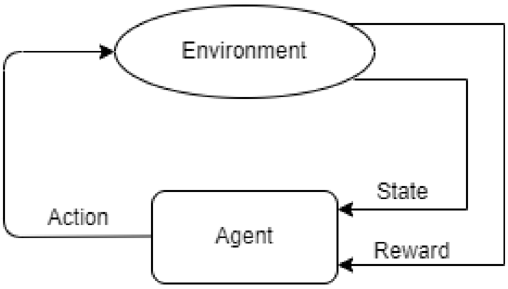

Fig 12: Reinforcement Learning

6. CONCLUSION

This study concludes with a comprehensive discussion of different types of machine learning algorithms. People will understand the uses of ML algorithms and apply them to their works. Dataset is one of the most important issues of any ML algorithm. From this research anyone can get the experience of which type of data to use in which algorithms. A large dataset will produce better output. Advantages, disadvantages and so on discussed in this paper so that the people easily identify the desired algorithms for their implementation. Everyone will know which algorithm fits for which dataset and also the category of the dataset. The performance of the ML algorithm can be easily checked by its limitations or disadvantages from this paper. Classification, regression and so many other techniques from the different learning are also classified and discussed in this paper. Overall this study will help the people to enhance their knowledge and skills for the machine learning algorithms point of view. Still some of the categories can be explored such as evolutionary learning and deep learning. In deep learning, FNN, CNN, RNN and so on are also the important names of ML algorithms. Hopefully in the future the author will work on that.

7. REFERENCES

[1] B. Mahesh, “Machine learning algorithms-a review,” Int. J. Sci. Res. IJSRInternet, vol. 9, no. 1, pp. 381–386, 2020.

[2] M. Fatima and M. Pasha, “Survey of machine learning algorithms for disease diagnostic,” J. Intell. Learn. Syst. Appl., vol. 9, no. 01, p. 1, 2017.

[3] A. Shrestha and A. Mahmood, “Review of deep learning algorithms and architectures,” IEEE Access, vol. 7, pp. 53040–53065, 2019.

[4] I. H. Sarker, “Machine learning: Algorithms, real-world applications and research directions,” SN Comput. Sci., vol. 2, no. 3, p. 160, 2021.

[5] S. Ray, “A quick review of machine learning algorithms,” presented at the 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon), IEEE, 2019, pp. 35–39.

[6] M. W. Berry, A. Mohamed, and B. W. Yap, Supervised and unsupervised learning for data science. Springer, 2019.

[7] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” BMC Med. Inform. Decis. Mak., vol. 19, no. 1, pp. 1–16, 2019.

[8] S. H. Shetty, S. Shetty, C. Singh, and A. Rao, “Supervised machine learning: algorithms and applications,” Fundam. Methods Mach. Deep Learn. Algorithms Tools Appl., pp. 1–16, 2022.

[9] P. C. Sen, M. Hajra, and M. Ghosh, “Supervised classification algorithms in machine learning: A survey and review,” presented at the Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018, Springer, 2020, pp. 99–111.

[10] A. E. Mohamed, “Comparative study of four supervised machine learning techniques for classification,” Int. J. Appl., vol. 7, no. 2, pp. 1–15, 2017.

[11] A. Ali and W. K. Mashwani, “A Supervised Machine Learning Algorithms: Applications, Challenges, and Recommendations,” Proc. Pak. Acad. Sci. Phys. Comput. Sci., vol. 60, no. 4, pp. 1–12, 2023.

[12] V. Nasteski, “An overview of the supervised machine learning methods,” Horiz. B, vol. 4, no. 51–62, p. 56, 2017.

[13] R. Choudhary and H. K. Gianey, “Comprehensive review on supervised machine learning algorithms,” presented at the 2017 International Conference on Machine Learning and Data Science (MLDS), IEEE, 2017, pp. 37–43.

[14] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi, “Supervised machine learning algorithms: classification and comparison,” Int. J. Comput. Trends Technol. IJCTT, vol. 48, no. 3, pp. 128–138, 2017.

[15] S. TIWARI, “Supervised Machine Learning: A Brief Introduction”.

[16] S. Naeem, A. Ali, S. Anam, and M. M. Ahmed, “An unsupervised machine learning algorithms: Comprehensive review,” Int. J. Comput. Digit. Syst., 2023.

[17] M. Suyala and S. Sharmab, “A Review on Analysis of K-Means Clustering Machine Learning Algorithm based on Unsupervised Learning,” 2024.

[18] A. E. Mehyadin and A. M. Abdulazeez, “Classification based on semi-supervised learning: A review,” Iraqi J. Comput. Inform., vol. 47, no. 1, pp. 1–11, 2021.

[19] Y. Reddy, P. Viswanath, and B. E. Reddy, “Semi-supervised learning: A brief review,” Int J Eng Technol, vol. 7, no. 1.8, p. 81, 2018.

[20] B. Kommey, O. J. Isaac, E. Tamakloe, and D. Opoku, “A Reinforcement Learning Review: Past Acts, Present Facts and Future Prospects,” IT J. Res. Dev., vol. 8, no. 2, pp. 120–142, 2023.

[21] R. Sharma, K. Sharma, and A. Khanna, “Study of supervised learning and unsupervised learning,” Int. J. Res. Appl. Sci. Eng. Technol., vol. 8, no. 6, pp. 588–593, 2020.