 http://dx.doi.org/10.14500/aro.12038

In-depth Analysis on Machine Learning Approaches: Techniques, Applications, and Trends

Abdulhady A. Abdullah[1], Nergz S. Mohammed[2], Maryam Khanzadi[3], Safar M. Asaad[4,5], Zrar Kh. Abdul[6] and Halgurd S. Maghdid[4†]

[1]Abdulhady Abas Abdullah Artificial Intelligence and Innovation Centre, University of Kurdistan Hewler, Erbil, Iraq

[2]Department of Computer Science, Faculty of Science, Soran University, Soran, Kurdistan Region – F.R. Iraq

[3]Department of Health Information Technology Engineering, University of Tehran, Tehran, Iran

[4]Department of Software Engineering, Faculty of Engineering, Koya University, Danielle Mitterrand Boulevard, Koya, KOY45, Kurdistan Region – F.R. Iraq

[5]Department of Computer Engineering, College of Engineering, Knowledge University, Erbil 44001, Kurdistan Region – F.R. Iraq

[6]Department of Computer, College of Science, Charmo University, Sulaymaniyah, Kurdistan Region – F.R. Iraq

***Abstract*—Machine learning (ML) approaches cover several aspects of daily life tasks, including knowledge representation, data analysis, regression, classification, recognition, clustering, planning, reasoning, text recommendation, and perception. The ML approaches enable applications to learn and adapt with or without being directly programmed from previous data or experience. The ML techniques, coupled with current technologies, provide a range of solutions, starts from vision-based applications to text-generation applications. To this end, this article presents a comprehensive overview of the approaches of ML, including supervised, unsupervised, semi-supervised, reinforcement, and self-learning. This review critically examines the roles performed by these aforementioned approaches in terms of their weaknesses and strengths. Furthermore, within this study, a new comparative analysis is conducted by reviewing existing studies and evaluating ML techniques using metrics including data requirement, accuracy, complexity, interpretability, scalability, applications, and challenges. Thereafter, the implemented ML techniques are classified, and their key findings are examined. The comprehensive review demonstrates that neither standalone nor hybrid ML techniques can completely satisfy all of the evaluated metrics, the necessity of customized solutions based on the requirements of particular applications.**

ARO-The Scientific Journal of Koya University
Vol. XIII, No. 1 (2025), Article ID: ARO.12038. 13 pages
DOI: 10.14500/aro.12038
Received: 06 February 2025; Accepted: 07 May 2025
Regular review paper; Published: 22 May 2025
†Corresponding author's e-mail: halgurd.maghdid@koyauniversityorg

***Index Terms*—Comparative metrics, Learning challenges, Machine learning algorithms, Machine learning structures.**

I. Introduction

A technology that allows us to produce intelligent systems capable of imitating human intelligence is called artificial intelligence (AI). Machine learning (ML) is a branch of AI which enables machines to understand without being directly programmed from previous data or skills (Christine, et al., 2020). Why should a machine be learned, even though we can program it? Well, there are two main reasons; first, the builders cannot predict all possible scenarios. Second, the builders happen to not know how to program a solution themselves (Weihao, Di and Theo, 2020). Fig. 1 below demonstrates the classes of ML.

Supervised learning (SUL) is the ML technique in which machines are trained in using training records, and machines calculate the output based on that data (Jwan, Abas and Tarik, 2024). SUL able to further separate into two kinds of problems: Classification and regression techniques. The classification procedures are used when the output variable is categorical, which includes two classes, such as yes-no, true-false, and female-male. The second type of SUL is regression real-value performance variable estimation, including special cases of forecasting future values out of recent or past values in a time series (Hooman, et al., 2019).

Unsupervised learning (USL) is another type of ML in which its models are trained without any guidance using an unlabeled dataset and can operate on that data. The concept of USL algorithms can be described within the idea of

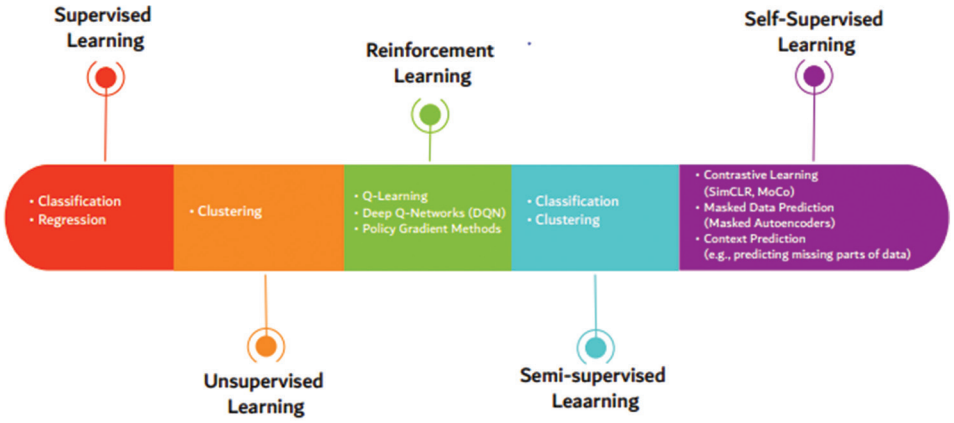

Fig. 1. The machine learning approaches with their tasks.

clustering, when clustering is a process of grouping a set of items into unique clusters such that the items with high similarities remain in a group and have less or no similarity with the items of any other group (Alboukadel, 2017).

The third kind of ML is semi-supervised learning (SSL), in the middle of the training datasets of supervised and USL, and the SSL problem starts with a sequence of both labeled and unlabeled data (Changde, Changying and Huiguang, 2021). From the state-of-the-art, the ML algorithms' accuracy differs in terms of the characteristics and size of the data sets between the training and testing sets. There is no one suitable ML algorithm to resolve all the problems.

Currently, most of the real-life solutions, when they are running via implementing the ML algorithms, are enhanced by using new paradigms, which are reinforcement learning (RIF) and self-learning (SEL) algorithms. The RIF is working on the expense of measuring the rewards and penalties of the actions according to the characteristics of application's environment. While the SEL algorithms mostly develop autonomously, to provide continuous enhancement and promise the intended targets via continuous learning.

However, the aforementioned ML paradigms need further investigation as well as they need to be analyzed according to the application solutions. To this end, this article makes a major contribution by providing a comprehensive review of ML approaches and offering a detailed comparative analysis that systematically contrasts SUL, USL, SSL, RIF, and SEL paradigms across critical operational metrics, thereby aiding researchers and practitioners in selecting the most appropriate learning technique based on application needs. Furthermore, a major contribution of this study is the detailed comparative analysis that systematically evaluates these approaches based on data requirements, complexity, accuracy, interpretability, scalability, applications, and challenges. By synthesizing recent research findings from 2016 to 2024 and identifying the strengths and limitations of each paradigm, the article also provides critical insights for researchers and practitioners. The provided comparison shows that no single method is universally optimal, highlighting the importance of selecting techniques tailored to specific application needs. In another vain, the review discusses emerging trends such as transfer learning, scalable SSL, and privacy-preserving techniques, offering valuable directions for future research.

The structure of this article is as follows: In section II, the theoretical background is explained. Section III investigates some literature reviews regarding supervised, unsupervised, semi-supervised, reinforcement, and SEL methods are described. A new comparative analysis of the current study of using ML algorithms via data requirements, complexity, accuracy, interpretability, scalability, application, and challenges is presented in section IV. Section V and VI provide the discussions and conclusions of this review, respectively.

II. Background

Researchers have proposed a huge number of methods in this field; hence, this section focused on ML classes. In general, ML has three kinds of learning: supervised, unsupervised, and semi-supervised. SUL that includes a classification algorithm and regression algorithm. USL contains a clustering algorithm. SSL is between SUL and USL. All these techniques and methods are explained in detail in the next subsections.

It's the fact that there is a big connection between AI and ML fields, including deep learning (DL) (Abdullah, et al., 2024). As illustrated in Fig. 2, ML is a subdomain of computer science used to analyze data, which automates the structuring of analytical models. ML algorithms aim to learn from the existing data without being explicitly programmed. New data sets adapt independently, which means the learned machine characteristic comes from the iterative feature of the models applied to the data set. This independent adoption is the main aspect of learning. Most often, this implies using a set of historical outcomes to estimate future outcomes (Sabr, 2025).

A. Supervised ML

SUL contains several methods that are applying something that must be learned from previous unused information, utilizing target illustrations to anticipate future effects. Beginning from the investigation of a well-known preparation

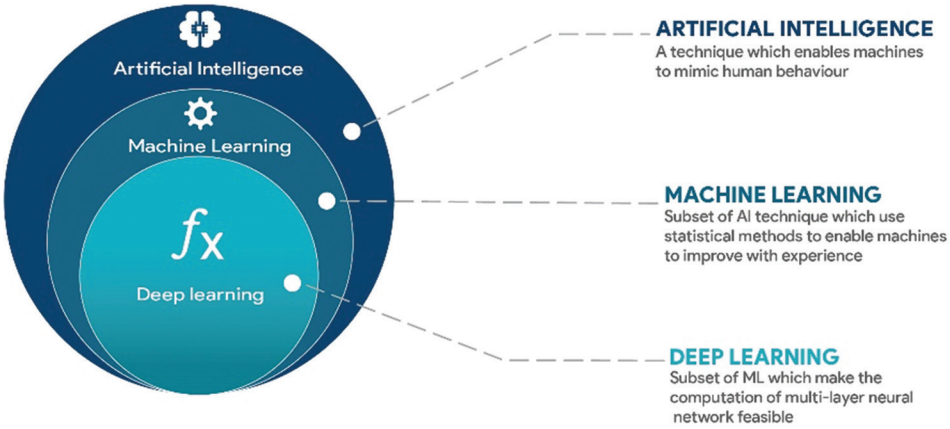

Fig. 2. The relation between artificial intelligence, machine learning and deep learning (Adamu, 2019).

dataset, the learning procedure yields an induced applies to obtain forecasts regarding the yield prices. The structure can supply labels for each further input after adequate preparation. The learning system can moreover match its product against the proper, planning yield, and discover mistakes in arranging to adjust the show appropriately (Gao, et al., 2014). The supervised ML consists of regression and classification techniques. Both of those methods are used for forecasting machines. The main differences among those two techniques are label value in regression is numerical. However, the classification procedure is categorical (Hemant and Rishabh, 2017). To well understand, the procedure of supervision is exposed in Fig. 3.

Classification

A classification technique, traditionally, is a purpose that evaluates the feature structures so that the label divorces one class into positive standards and the other into negative standards, the classification method is used for prediction in ML and works with the labeled data. However, the output variable for classification algorithms is categorical or (discrete), such as recognition of a type of car in a photo, what the weather will be like today or a message from a friend. More detail of these processes is implemented in Fig. 4. (Yongjun and Siyu, 2020). There are several applications running via utilizing classification tasks, for instance, bank customer loan pays willingness prediction, gait recognition, user positioning, email spam classification, web news classification, leaf diseases classification, and cancer tumor cell identification (Upasana, 2019). The classification has several techniques and methods, such as KNN, linear discriminant analysis, regression trees, learning vector quantization, support vector machines, naive Bayes, bagging and random forest, boosting, and stochastic gradient descent (Deepti and Dilip, 2018).

Regression

The Regression task is a numerical way that examines and recognizes the connection between two or more variables of attention. The regression is also to make a relation between a single dependent variable and one or a set of independent variables, as shown in Fig. 5. (Carlos, et al., 2019). Furthermore, the main idea beyond executing regression analysis is to know which features are important that can be failed to observe and how they are manipulating each other (Sung-Jin, et al., 2016). In another vein, feature selection is an important step in the data cleaning for the regression functions. This is because the regression identifies or chooses the most related variables that are contributing to the prediction process.

Linear regression and Logistic regression are the most well-known techniques of the regression task. The linear regression is to make the linear relationship between the dependent and non-dependent variables. While, the logistic regression is used to estimate the probability of happing an event based on set of given independent variables (Carlos, et al., 2019).

The regression is an essential step in most ML algorithms. There are several real-life applications that relaying on utilizing regression analysis starts from medicine report cases to the house price estimates or financial forecasting.

B. Unsupervised ML

USL is a type of ML where a show must explore for formerly hidden patterns in a dataset through no labels and with a minimum of human observation. In USL, a dataset is provided devoid of labels, and typical studies useful properties of the group of the dataset. We do not speak the classical what it needs to be studied but agree to it to invent patterns and attraction decisions from the unlabeled information (Sarfaraz, et al., 2019). The procedures in USL are harder than in SUL since we have little or no information about the data. USL responsibilities normally include combining related instances, dimensionality decrease, and density approximation. One more term for USL is “knowledge discovery” (Kushal, 2020). The most generally used USL procedures are k-means, hierarchical cluster analysis, and expectation maximization (Hui, Ping and Duo, 2019). Public USL methods contain clustering, and dimensionality decrease. Fig. 6 explains the process of USL.

Clustering

Cluster analysis is the most public USL algorithm. Sense that you don’t recognize how many clusters are in the data beforehand, when running the typical (Abdullah, et al.,

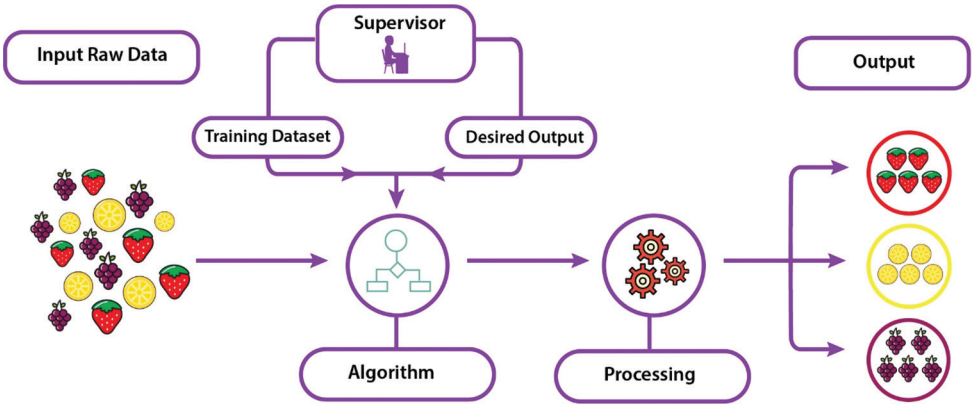

Fig. 3. Supervised learning paradigm.

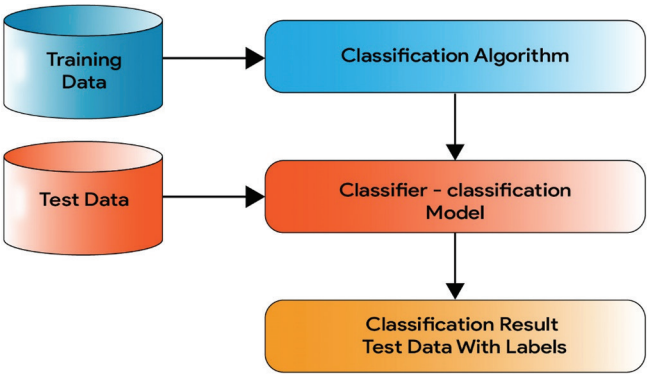

Fig. 4. Classification process.

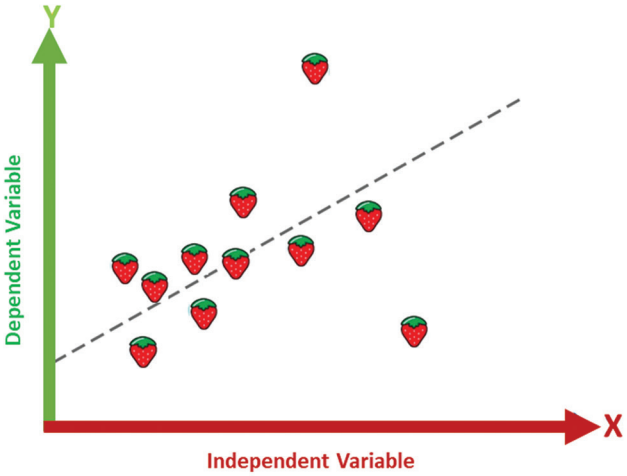

Fig. 5. A Regression line between the independent and dependent variables.

2024). Different from many other numerical techniques, the final output labels are not known previously. This kind of algorithm can help in solving many obstacles. It makes available information about where patterns and associations in data happen, but not what those might be or what they mean. The objective of cluster analysis is to discover related groups of topics, where "similarity" among separate pairs of subjects means the overall amount above the entire set of individualities (Guo, et al., 2021). Some of the cluster methods are partition clustering, hierarchical clustering, and fuzzy clustering (Chunrong, et al., 2019). Fig. 7 below explains the clustering procedure.

C. Semi-Supervised ML

Semi-supervised is one of the methods of AI among the training datasets of supervised and USL. The SSL problem starts with a sequence of both labeled and unlabeled data, semi-supervised learning aims to categorize some of the unlabeled data using the labeled evidence set (Haitao and Zhenhua, 2018). The idea behind SSL is to learn from structured and unstructured information to increase the predictive power of the models. Regularly, personalities unrelated to the domain imagine robots invading advancements and taming people. But AI is extremely diverse, or at least, much more than that possibility (Rui, Feiping and Xuelong, 2017). SUL held the initial kind of learning is investigated in the field of AI. Considering its inception, infinite techniques differing in the complexity of the humble logistic regression to the massive neural network should be examined to enhance accuracy and sinister power. SSL practices the classification method to classify data assets and the clustering procedure to arrange it into different sections. Fig. 8 explains the semi-supervised process.

D. Reinforcement ML

The RIL approach is one of the current ML methods when an agent learns from the vicinity by executing actions to make a decision. The actions are also improved by receiving the feedback in the form of rewards or penalties, as shown in Fig. 9. The most important components of an RIF paradigm are: Agent, environment, state, action, reward, policy, and value functions. Furthermore, the agent observes the recent updates stated of the environments and it provides the action based on the available policy. Thus, the agent is to maximize the cumulative reward over time, while the environment transitions from the old state to a new state and then calculate the reward. Thereafter, the agent revises its policy based on the reward to improve future decisions (Zhang et al., 2021).

The Q-leaning, Deep Q-learning (DQN), Policy Gradient (PG), and Actor-Critic are the most common or core algorithms of the RIF approach (Zhou, Huang and Fränti, 2022). The Q-learning is a model-free algorithm which

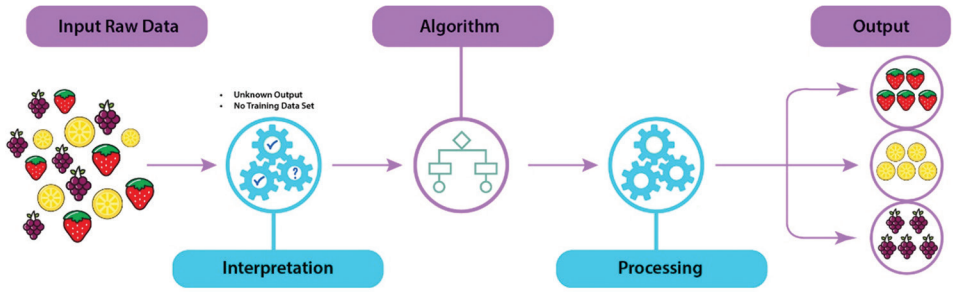

Fig. 6. Unsupervised learning.

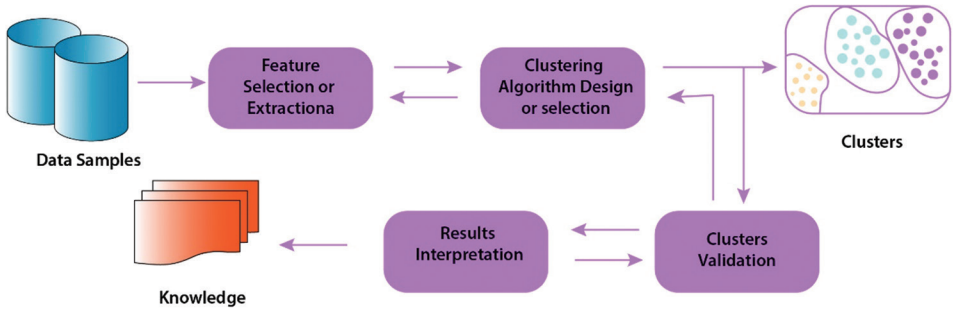

Fig. 7. Clustering process.

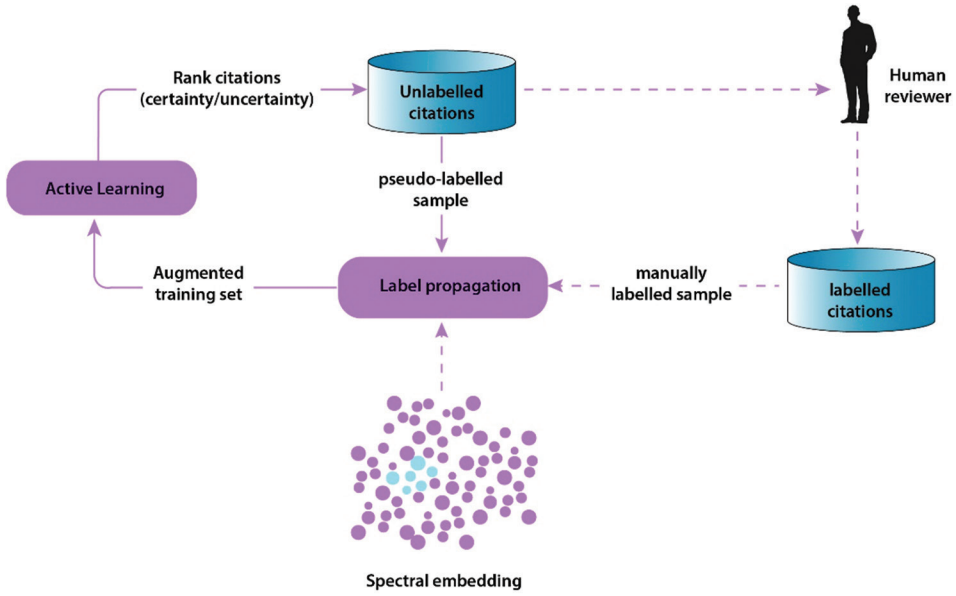

Fig. 8. Semi-supervised learning process.

learns a Q-value for each state-action pair, while the DQN algorithm combines Q-learning with deep neural networks. Furthermore, the PG methods equally enhance the policy and function values. The actor-critic methods use two linked models, where an actor is to select actions and a critic is to evaluate the models. To this end, these methods show that the RL is unique, since the agent learning from trial and error which focuses on long-term rewards rather than only current outcomes (Zhang et al., 2021).

With the era of ML approaches, Robotics, gaming, energy systems, agriculture digitization, driverless cars, healthcare, and finance are the most well-known applications for today's life. However, the navigating safely, developing personalized treatment strategies for chronic diseases, performing complex manipulation tasks, portfolio management, and optimizing energy utilization are still remained as challenges (Perera and Kamalaruban, 2021).

E. SEL Approach

A SEL approach in the context of ML refers to the new paradigm that can enhance their performance and outcomes based on data that they collect through their experiences

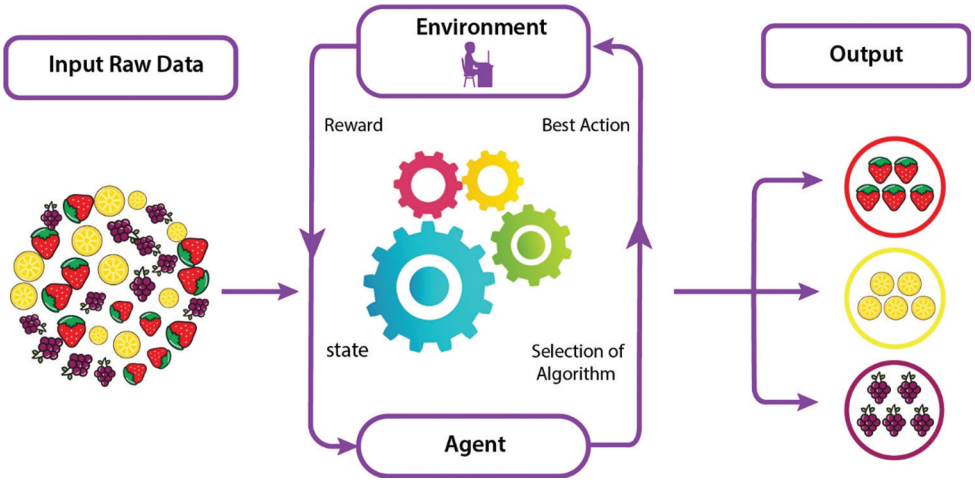

Fig. 9. Reinforcement learning paradigm.

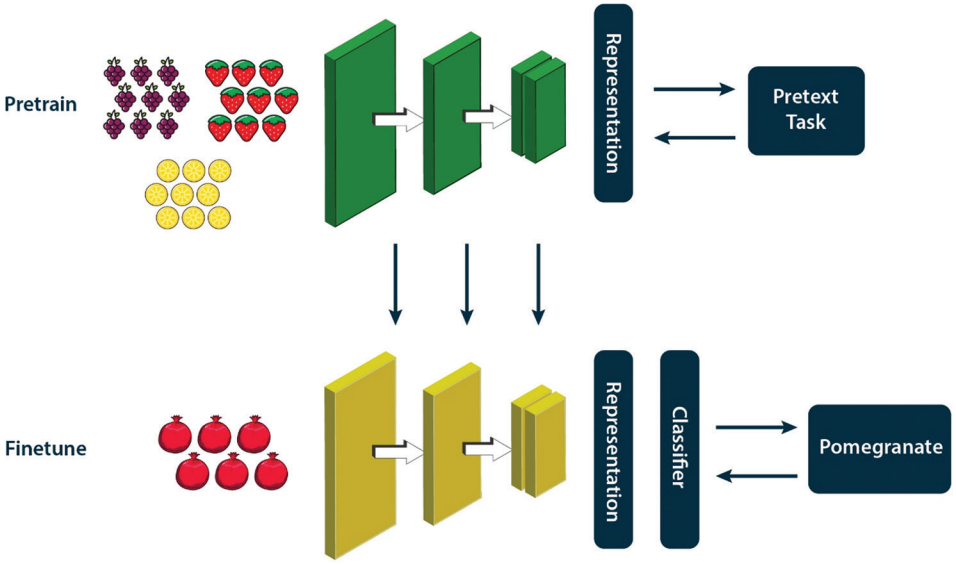

Fig. 10. Fine-tuning process in self-learning paradigm.

without being explicitly programmed for those tasks. This approach is often associated with USL, where the algorithm tries to identify patterns and relationships in data without prior labels or instructions, as shown in Fig. 10. Furthermore, the USL, feature discovery, and adaptivity are the most important aspects of the SEL approach (Wu and He, 2023). For example, within USL, the system trained from data without having any labeled or corrections. While with feature discovery, the ability to autonomously discover the representations needed for feature detection or classification from raw data. However, with the adaptivity, the systems adjust their based on upcoming data, repeatedly updating their weights or model of the world in general (Iqbal et al., 2023).

The current solutions via SEL methods are anomaly detection, recommendation systems, predictive maintenance in manufacturing, dynamic pricing models, autonomous robotics, and text-generation like chat-GPTs. However, such solutions need further research due to several challenges starts from identifying unusual patterns, biased issues for personalized recommendations, predicting are likely to fail or require maintenance, adjusting prices in real-time, to the operating in complex and unpredictable environments.

In another vain, SEL approaches are particularly valuable in environments where it is impossible or impractical to manually label all the possible scenarios or outcomes that a system might need to handle. They allow systems to adapt over time and improve their accuracy, making them highly effective in complex, real-world applications where new data continuously emerges.

III. Current ML Algorithms

Due to the importance of learning techniques, which are used widely in many areas, a huge amount of research has been done in this area. In this section, several works regarding supervised, unsupervised, and semi-supervised methods briefly are discussed.

The authors of the paper (Xin, et al., 2020) discussed the significant challenge of the social stream SSL classification

technique called classification over drifting and evolving stream (CODES). One goal is to set the training group consisting of both structured and unstructured examples. The extreme learning machine (ELM) based on the SSL strategy can preserve the ELM input drawing area without output knowledge of emergent data in the social stream. CODES can achieve effective learning achievement above drifting and growing social streams while improving practical importance among the real-world social stream utilizations.

Pilot studies conducted by (Yuan and Marc, 2019), for this purpose, the authors contrast three separate visual-inertial odometry (VIO) methods based on learning: Supervised, semi-supervised, and unsupervised. VIO, that used pictures and inertial estimations to evaluate the movement, is supposed to be unique of the main tools for virtual fact and argument fact. The goal is to reach more accurate, robust, and efficient localization. The outcomes demonstrate the semi-supervised model better supervised methods as well as unsupervised ones.

An attempt to create a model that can be automatically built to recognize a species of iris have been done by (Ajay, et al., 2018), the authors used KNN classification SUL algorithms. Take advantage of the iris dataset, which contains 150 data samples in three groups, each containing 50 samples. The libraries used are Pandas, NumPy, Scikit-learn, and Matplotlib. The result shows the forecast for Class 0 (Setosa) and Calss 2 (Virginica) is one hundred percentage right, however, the forecast for Class 1 (Versicolor) is 4% inaccurate.

A model suggested for intrusion detection has been made by (Manjula and Balachandra, 2016). The authors used classification procedures, specifically, logistic regression, Gaussian naive Bayes, support vector machine, and random forest, those techniques are tested by a dataset, namely the NSL-KDD dataset. Code is done using a python programming language. Consequences display that the Random Forest Classifier outperforms supplementary approaches in assessing whether the records of the traffic are usual or a raid. It has an accuracy of 99%.

A technique was proposed to detect fake users by (BalaAnand, et al., 2019), the authors used the Graph-based SSL algorithm (EGSLA). The data are taken from Twitter. More precisely, the data set contains 2,915,147 tweets were recovered from 21,473 users through the duration from 12–2017 to 2–2018. The EGSLA technique is examined through the existing game hypothesis, support vector machine, KNN, and decision tree methods. The outcomes were visible that the suggested EGSLA procedure succeeds 90.3% precision in recognizing forgery (fake) users.

A procedure for identifying the distance among a score and a group pattern has been made by (Kristina and Miin-Shen, 2020), tried to establish an USL framework by using the k-means procedure, thus as that is independent of initializations externally variable determination and can likewise discover an optimum number of groups at the same time. The authors compared this technique with several other algorithms. The consequences indeed indicate the best feature of the recommended U-kmeans clustering procedure.

The textual documents have been proposed by (Aiman and Rosnafisah, 2017), the authors for this purpose an approach that used the KNN algorithm for the classification to construct an ML system in R software. The data were taken from two websites: (egov.kz and government.kz). Finally, the authors found out that the highest percentage of accuracy when the value of k ranges from one to fifty. The accuracy dropped sharply above the 50's.

An approach was proposed to produce a price of the cars have been done by (Nitis, et al., 2018), for this intention utilized the regression techniques that are multiple linear, gradient boosted trees, and random forest. Records used during that analysis were taken from the German e-commerce website then data training compiled via with Python language. The dataset includes 304,133 records and 11 columns. The outcomes were then compared through mean absolute error (MAE) as a measure. Gradient boosted trees give the best attainment with MAE = 0.28, the second of the best is a random forest with MAE = 0.35, followed by multiple linear regression with MAE = 0.55 errors. Therefore, researchers assumed that gradient boosted trees can be advisable to build the price assessment form.

A process was recommended to categorize internet traffic detection done by (Mrudul, et al., 2019), the authors used KNN and naive Bayes classification techniques. The authors concentrate on six statistical variables of the fifty variables achievable in the UNSW NB dataset. The result illustrates that the KNN procedure gives a precision 85%, while the Naive Bayes process reached 54% of precision.

An effort to understand the shortcomings of the KNN technique was made by (Gongde, et al., 2003), the authors propose a new kind of KNN technique for classification. To validate the technique, tests were carried out on some publicly available datasets obtained from the UCI ML repository. The results indicate that the model based on KNN compares well with C5.0 algorithm and the KNN method. The KNN model significantly decreases the amount of data tuples in the final classification model with an average rate of 90.41% reduction.

In another work by (Fangming, Oayou and Xinying, 2010), the authors suggest a proposal to study a Mahalanobis distance with a minor quantity of well-known knowledge, by using a graph-based semi-supervised output broadcast technique to improve the classification knowledge that is given through the customer, and later uses a process of increased biased relevant component analysis to study a Mahalanobis distance purpose. Later, those procedures, the authors used Mahalanobis instead of Euclidean distance as a metric function to find the distance between points of the KNN classifier. For this purpose, take advantage of the UCI datasets. The result shows that technique able to be importantly increases the precision of the KNN classification methods.

Attempts to build a model that can be automatically constructed to recognize a data uncertainty have been done by (Nergz and Beitollahi, 2022), the authors used numerous experiential methods and ML procedures and the combination of radial basis function network with the particle-swarm

TABLE I
Recent Advances in Machine Learning Techniques (2010–2024)

Year	Authors	Technique (s) used	Key findings
2016	Manjula and Balachandra	LOR, Gaussian naive Bayes, SVM, RF	The random forest achieved highest accuracy for intrusion detection.
2017	Aiman and Rosnafisah	SUL (KNN)	High accuracy in classifying textual documents.
2018	Ajay et al.	SUL (KNN)	High accuracy in classifying iris species.
2018	Nitis et al.	MLR, gradient boosted trees, RF	Gradient boosted trees achieved the best performance for price prediction.
2019	Yuan and Marc	SUL, USL, SSL	Semi-supervised outperformed both supervised and unsupervised.
2019	BalaAnand et al.	SSL (EGSLA)	90.3% accuracy in detecting fake users on Twitter.
2019	Mrudul et al.	SUL (Naive Bayes, KNN)	KNN achieved higher precision than Naive Bayes.
2020	Xin et al.	SUL (CODES)	Improved learning performance on drifting social streams.
2020	Kristina and Miin-Shen	USL (K-means)	Improved clustering performance independent of initializations.
2022	Xin et al.	SSL	Comprehensive review of SSL techniques and their applications.
2022	Cao et al.	SSL	Introduces the open-world assumption in SSL, handling out-of-distribution data.
2022	Bromley et al.	SUL	Proposes MaskSup for improved semantic segmentation.
2022	Kaiming et al.	SSL (adversarial training)	Enhances model robustness by denoising features in adversarial settings.
2023	Smith et al.	SSL	Reduces confirmation bias in pseudo-labels, improves robustness.
2023	Jones et al.	SSL	Combines semantic and instance similarity for better performance.
2023	Lee et al.	SUL	Joint optimization for segmentation, depth estimation, and edge detection.
2023	Kim et al.	SSL	Enhances SSL with adversarial training for robust model performance.
2023	Patel et al.	USL	Proposes a new method for clustering high-dimensional data efficiently.
2024	Wang et al.	SUL	Enhances supervised models with transfer learning techniques.
2024	Zhao et al.	SSL	Addresses imbalance in datasets with a novel SSL approach.
2024	Gupta et al.	USL	Efficient real-time anomaly detection in streaming data.
2024	Li et al.	SSL	Applies graph-based SSL for better social network insights.
2024	Thompson et al.	SUL	New techniques for enhancing text classification accuracy.
2024	Rodriguez et al.	SSL	Combines multiple SSL methods for improved performance.
2024	Hernandez et al.	USL	Proposes an unsupervised approach for improving image quality.
2024	Nguyen et al.	SSL	Integrates multiple data modalities for better SSL performance.
2024	Tan et al.	SUL	Enhances supervised learning with active learning strategies.
2024	Chen et al.	USL	Uses reinforcement signals to guide unsupervised learning processes.
2024	Park et al.	SSL	Proposes a scalable SSL approach for large datasets.
2024	Roberts et al.	SUL	Improves recommendation systems with personalized models.
2024	Kaiming et al.	SSL (masked modeling)	Proposes an advanced masked modeling approach for SSL in vision tasks.
2022	Zhou, Huang and Fränti	RIL	Future of motion planning algorithms in robots
2021	Zhang et al.	RIL	Data privacy and adaptive learning capability, and their prospects in real-time monitoring, out-of-clinic diagnosis are challenged
2021	Perera and Kamalaruban	RIL	Reinforcement learning has a notable potential which has not been utilized
2023	Wu and He	SEL	Solve the problem of weak explanation of model
2023	Iqbal et al.	SEL	Careful tuning and experimentation are essential to determine the optimal combination of manual features

CODES: Classification over drifting and evolving stream, SUL: Supervised learning, USL: Unsupervised learning, SSL: Semi-supervised learning, RIF: Reinforcement learning, SEL: Self-learning

optimization algorithm to categorize data in the presence of uncertainty. The simulation specifically produced the following outcomes: (F- Measure, recall, precision, and accuracy) are (96%, 95%, 97%, and 97%). The following is an obvious indication that the suggested model outperforms conventional ML techniques in identifying ambiguous data. The suggested approach outperforms all ML-based approaches by an average of 4.4%. As well as the authors in another worked on classify uncertain data (Darbaz, Al-Barznji and Mohammed, 2024) used a combination ML based methods with DL techniques The outcomes of the suggested hybrid model depend on well-known evolution metrics F- Measure = 95%, recall = 94%, precision = 96%, as well as accuracy = 97%.

Table I summarizes and integrates recent research studies from 2016 to 2024 alongside older works, providing a comprehensive and up-to-date overview of significant advances in supervised, unsupervised, semi-supervised, and self-SUL techniques. Each entry highlights the innovative methods and key findings, reflecting the evolving landscape of ML research.

IV. Comparative Analysis

This section presents a comparative analysis of supervised, unsupervised, semi-supervised, and self-SUL techniques based on the recent advancements outlined in the literature review. The comparison considers various aspects such as data requirements, complexity, accuracy, interpretability, scalability, applications, and challenges.

A. Data Requirements

SUL

Requires large amounts of labeled data. Effective for tasks where annotated data is abundant, such as image classification and speech recognition.

USL

Works with unlabeled data, suitable for exploratory data analysis where the goal is to identify hidden patterns without predefined labels, like clustering and anomaly detection.

SSL

Utilizes both labeled and unlabeled data, making it ideal for situations where labeled data is scarce or expensive to obtain. It strikes a balance by leveraging the abundance of unlabeled data to improve learning accuracy.

RIF

RL differs significantly from other learning paradigms because it typically requires neither labeled data (as in SUL) nor solely unlabeled data (as in USL). Instead, RL operates through an agent interacting with an environment to learn policies based on rewards. This interaction generates the data (in the form of state, action, and reward tuples) that RL algorithms use to learn.

SEL

Generates supervisory signals from the data itself, reducing the need for labeled data. It is effective in domains like natural language processing and computer vision, where creating large labeled datasets is challenging.

B. Complexity

SUL

Generally moderate to high, depending on the algorithm used. Complex models like deep neural networks require significant computational resources and expertise.

USL

Typically low to moderate complexity. Algorithms like K-means clustering are relatively straightforward but can become complex with high-dimensional data.

SSL

High complexity due to the integration of both labeled and unlabeled data. Methods like adversarial training and graph-based approaches add to the complexity.

RIF

The complexity of RL can be quite high, primarily because the environment itself can be highly dynamic and the learning process is based on sequential decision-making.

SEL

Can range from moderate to high complexity. Techniques like masked modeling and contrastive learning involve sophisticated architectures and training strategies.

C. Accuracy

SUL

Generally high, especially when ample labeled data is available. Models can be fine-tuned to achieve state-of-the-art performance in specific tasks.

USL

Accuracy varies widely, often dependent on the nature of the data and the specific algorithm. Model evaluation can be challenging due to the lack of ground truth.

SSL

Often achieves higher accuracy than USL and can approach the accuracy of SUL, especially with well-designed algorithms that leverage the unlabeled data effectively.

RIF

Accuracy in RL is typically framed in terms of the optimality of the learned policy rather than traditional accuracy metrics. The goal is to maximize cumulative rewards, which may not always align with achieving high accuracy in predictions, like in SUL.

SEL

Shows promising accuracy, particularly in tasks where large-scale unlabeled data is available. The learned representations can be fine-tuned for specific downstream tasks, achieving competitive performance.

D. Interpretability

SUL

High for simple models (e.g., decision trees, linear regression) but lower for complex models like deep neural networks.

USL

Varies, often challenging due to the lack of labeled data. Techniques like clustering provide some interpretability by grouping similar data points.

SSL

Depends on the combination of techniques used. Graph-based methods offer some interpretability, but overall, the complexity can reduce interpretability.

RIF

Similar to complex models in supervised and self-SUL, RL models can struggle with interpretability, especially in high-dimensional spaces or when using deep neural networks as function approximators (Deep RIF).

SEL

Generally lower interpretability due to the complex nature of the tasks and models. However, certain techniques, like contrastive learning, provide some insights into the representations learned.

E. Scalability

SUL

Can be challenged with large datasets due to the need for extensive labeled data and computational resources.

USL

Generally scalable, as many algorithms can handle large datasets efficiently.

SSL

Scalability can be challenging due to the complexity of integrating labeled and unlabeled data. Techniques like scalable SSL aim to address this issue.

RIF

Scalability in RL can be a challenge due to the need for extensive interaction with the environment, which can be

computationally expensive and slow, particularly in real-world scenarios.

SEL

Often highly scalable, as it can leverage vast amounts of unlabeled data. Techniques like masked modeling are designed to handle large-scale data efficiently.

F. Applications

SUL

Widely used in tasks like image and speech recognition, medical diagnosis, financial forecasting, and more.

USL

Applied in customer segmentation, topic modeling, anomaly detection, and exploration of data analysis.

SSL

Useful in natural language processing, bioinformatics, web content classification, and scenarios with limited labeled data.

RIF

RL is extensively applied in areas such as robotics (for complex control tasks), gaming (e.g., AI playing video games or board games like Go), autonomous vehicles (for dynamic decision-making tasks), and optimization problems in operations research.

SEL

Effective in natural language processing, computer vision, and other domains where generating labeled data is expensive or impractical.

G. Challenges

SUL

Overfitting, scalability, and the high cost of data labeling.

USL

Model evaluation, interpretability, and convergence issues.

SSL

Data integration, model complexity, and computational costs.

RIF

Some of the primary challenges in RL include the dependency on quality and diversity of the reward signal.

SEL

Complexity of model training, interpretability, and ensuring the quality of self-generated labels.

This comparative analysis highlights the unique strengths and challenges of each learning paradigm shown in Table II, providing insights into their applicability to various tasks and domains. Understanding these nuances helps researchers and practitioners select the most appropriate methods for their specific needs, driving innovation and performance improvements in ML applications.

V. Discussions

The comparative analysis of SUL, USL, SSL, RIF, and SEL learning paradigms reveals several insights into their respective strengths, limitations, and applications. SUL remains a cornerstone of ML, particularly in domains where large amounts of labeled data are available. Its high accuracy and effectiveness in tasks such as image and speech recognition, medical diagnosis, and financial forecasting make it a preferred choice for many applications. However, reliance on extensive labeled datasets poses a significant challenge, both in terms of data collection costs and scalability. Additionally, while simpler models like decision trees offer high interpretability, more complex models such as deep neural networks can suffer from overfitting and reduced interpretability.

USL excels in exploratory data analysis, where the goal is to uncover hidden patterns and structures without predefined

TABLE II
Comparison of Learning Paradigms: SUL, USL, SSL, RIF, and SEL Approaches

Aspect	SUL	USL	SSL	RIF	SEL
Data requirements	Requires large amounts of labeled data	Works with unlabeled data	Utilizes both labeled and unlabeled data	Interact with to generate data through the agent's actions.	Generates supervisory signals from the data itself
Complexity	Moderate to high	Low to moderate	High	High	Moderate to high
Accuracy	Generally high	Varies widely	Often higher than unsupervised, lower than supervised	Can achieve high where clear metrics	Varies; can improve over time as more data is processed
Interpretability	High for simple models, low for complex models	Varies, often challenging	It depends on the combination of techniques used	Generally lower due to complex tasks and models	Moderate; as models evolve, tracking changes and understanding decisions
Scalability	Can be challenging with large datasets	Generally scalable	Can be challenging due to data integration	Varies depending on the complexity of the environment	Often highly scalable
Applications	Image and speech recognition, medical diagnosis, financial forecasting	Customer segmentation, topic modeling, anomaly detection	NLP, bioinformatics, web content classification	Gaming, autonomous vehicles, and robotics.	NLP taks, computer vision, tasks with vast unlabeled data
Challenges	Overfitting, scalability, high cost of data labeling	Model evaluation, interpretability, convergence issues	Data integration, model complexity, computational costs	Complexity of model training, high computational cost, defining appropriate rewards and penalties	Complexity of model training, interpretability

SUL: Supervised learning, USL: Unsupervised learning, SSL: Semi-supervised learning, RIF: Reinforcement learning, SEL: Self-learning

labels. Techniques like clustering, dimensionality reduction, and anomaly detection are valuable in applications ranging from customer segmentation to topic modeling and fraud detection. The main advantage of an SUL is its ability to work with unlabeled data, making it scalable and versatile. However, the accuracy of unsupervised models can vary widely, and the lack of ground truth makes model evaluation and interpretability challenging.

SSL learning strikes a balance between SUL and USL approached by leveraging both known and unknown items. This approach is particularly useful in scenarios where the known item is scarce or expensive to obtain, such as in medical imaging or NLP tasks. The SSL algorithms often achieve higher accuracy than purely unsupervised methods and can approach the performance of supervised models when designed effectively. However, the complexity of integrating labeled and unlabeled data, along with the computational costs, poses significant challenges (Papers with Code) (Papers with Code).

In another vain, the RIF involves an agent interacting with an environment to generate data through actions, which makes it highly complex and dependent on designing effective reward systems. This learning type achieves high accuracy in environments with clear success metrics, such as games or simulations. However, it faces challenges in interpretability due to the complexity of tasks and models, and scalability can vary greatly with environmental and state space complexities. It is primarily applied in areas such as gaming, autonomous vehicles, robotics, and real-time decision-making. The major challenges include the complexity of training models, the computational demands, and the necessity of designing a reward system that effectively balances short- and long-term goals.

While the SEL utilizes existing knowledge to generate new insights, requiring a foundational dataset to begin learning and often improving accuracy over time as it processes more data and adapts to new inputs. The complexity of SEL systems can be moderate to high, depending on the mechanisms used for knowledge extraction and adaptation. While these systems are generally scalable within the learning algorithm's capabilities, they face interpretability challenges as models evolve and adapt. SEL is applied in self-correcting algorithms, dynamic decision-making systems, and continuous learning environments within AI systems. Key challenges include maintaining continuous learning without data drift, managing computational resources effectively, and adapting autonomously to evolving datasets without human oversight.

Several emerging trends are shaping the future of ML across these paradigms. Transfer learning, for instance, enhances supervised models by leveraging pre-trained models on related tasks, thus reducing the need for large labeled datasets. Adversarial training, particularly in semi-supervised and self-SUL, improves model robustness and generalization (Papers with Code). Additionally, scalable methods for large datasets and techniques that address data imbalance are becoming increasingly important as the volume and variety of data continue to grow (Papers with Code).

Ongoing research is focused on improving the explainability and fairness of ML models, ensuring that they are not only accurate but also transparent and unbiased. Privacy-preserving learning techniques, such as federated learning, are also gaining traction, enabling the use of sensitive data without compromising privacy. The integration of multiple data modalities in SSL and the development of more efficient training algorithms for self-SUL are promising areas for future exploration.

The review of learning paradigms underscores the unique advantages and challenges of each approach. By understanding these nuances, researchers and practitioners can better select and tailor methods to their specific needs, driving innovation and performance improvements in ML applications. The continuous evolution of these paradigms, driven by advancements in techniques and increasing data availability, holds significant potential for future breakthroughs in AI.

VI. Conclusion

In this study, a review of the ML approaches, including SUL, USL, SSL, RIF, and SEL, has been investigated within current applications. From the reviewed and discussed studies, the ML methods and techniques, accuracy differs depending on the properties (attributes) and the volume of the data sets amongst the training sets and testing sets. In the diverse world of ML, each method provides unique insights and solutions for specific problems. The SUL is highly structured, requiring large amounts of labeled data to train models that achieve high accuracy. This method excels in applications such as image and speech recognition, where precise outputs based on clear examples are essential. However, its reliance on extensive labeled datasets can lead to challenges such as overfitting and the high cost of data annotation. In contrast, USL thrives on unlabeled data, uncovering hidden patterns and relationships without predefined labels. This approach is particularly beneficial in scenarios like customer segmentation and anomaly detection, where the data may lack explicit instructions but still contains valuable insights. Challenges in USL include difficulties with model evaluation and interpretability of results, which can vary significantly based on the complexity of the data.

The SSL bridges the gap between supervised and USL by utilizing both labeled and unlabeled data. This method improves learning accuracy without the exhaustive need for labeled data, making it suitable for applications like natural language processing and bioinformatics, where obtaining comprehensive labeled datasets can be costly or impractical. The RIF and SEL reveal a spectrum from highly structured learning environments to adaptive, interactive systems. RIF learns from its environment through trial and error, offering solutions for real-time decision-making in complex scenarios. Meanwhile, SEL adapts and evolves autonomously, promising ongoing improvement across various applications.

Together, these learning methodologies highlight the versatility of ML. They offer a glimpse into a future where

machines not only calculate and predict but also discover, adapt, and continually learn from their interactions, much like living beings. The unique strengths and challenges of each method emphasize the importance of selecting the right approach based on the specific requirements and constraints of the task.

References

Abdullah, A.A., Abdulla, S.H., Toufiq, D.M., Maghdid, H.S., Rashid, T.A., Farho, P.F., Sabr, S.S., Taher, A.H., Hamad, D.S., Veisi, H., and Asaad, A.T., 2024. *NER-RoBERTa: Fine-Tuning RoBERTa for Named Entity Recognition (NER) within Low-Resource Languages*. [arXiv Preprint].

Abdullah, A.A., Ahmed, A.M., Rashid, T., Veisi, H., Rassul, Y.H., Hassan, B., Fattah, P., Ali, S.A., and Shamsaldin, A.S., 2024. *Advanced Clustering Techniques for Speech Signal Enhancement: A Review and Metanalysis of Fuzzy c-Means, k-Means, and Kernel Fuzzy c-Means Methods*. [arXiv Preprint].

Adamu, J.A., 2019. Advanced stochastic optimization algorithm for deep learning artificial neural networks in banking and finance industries. *Risk and Financial Management*, 1(1), p. 8.

Aiman, M., and Rosnafisah, B.S., 2017. *Using KNN Algorithm for Classification of Textual Documents*. Malaysia, IEEE.

Ajay, S.S., Thirunavukkarasu, K., Prakhar, R., and Sachin, G., 2018. *Classification of IRIS Dataset Using Classification Based KNN Algorithm in Supervised Learning*. IEEE, United States.

Alboukadel, K., 2017. *Practical Guide to Cluster Analysis in R Unsupervised Machine Learning*. 1st ed. STHDA, Bad Vilbel.

BalaAnand, M., Karthikeyan, N., Karthik, S., Varatharajan, R., Manogaran, G., and Sivaparthipan, C.B., 2019. An enhanced graph-based semi-supervised learning algorithm to detect fake users on Twitter. *The Journal of Supercomputing*, 75, pp. 6085-6105.

Carlos, F., Flora, F., Miguel, G., Azevedo, O., Sousa, N., and Erlhagen, W., 2019. Gait Classification of Patients with Fabry's Disease Based on Normalized Gait Features Obtained using Multiple Regression Models. In: *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.

Changde, D., Changying, D., and Huiguang, H., 2021. Multimodal deep generative adversarial models for scalable doubly semi-supervised learning. *Information Fusion*, 68, pp. 118-130.

Christine, W.C., Tontiwachwuthikul, P., Zeng, F., and Liang, Z.Z., 2020. Recent progress and new developments of applications of artificialintelligence (AI), knowledge-based systems (KBS), and machine learning (ML) in the petroleum industry. *Petroleum*, 6(4), pp. 319-320.

Chunrong, W., Jia, L., Isokawa, T., Jun, Y., and Yunni, X., 2019. Efficient Clustering Method Based on Density Peaks with Symmetric Neighborhood Relationship. In: *The International Exchange Program of National Institute of Information and Communications (NICT)*.

Darbaz, M.H., Al-Barznji, K., and Mohammed, N.S., 2024. Accurate uncertainty dataset classification using hybrid deep learning models. *International Journal of Advanced Processing Systems*, 10, pp. 15-30.

Deepti, S., and Dilip, S.S., 2018. *Prediction of Diabetes using Classification Algorithms*. IEEE and Elsevier, India.

Fangming, G., Oayou, L., and Xinying, W., 2010. *Semi-Supervised Weighted Distance Metric Learning for kNN Classification*. IEEE Xplore, Changchun.

Gao, H., Shiji, S., Jatinder, N.D.G., and Cheng, W., 2014. Semi-supervised and unsupervised extreme learning machines. *IEEE Transactions on Cybernetics*, 44(12), pp. 2405-2417.

Gongde, G., Wang, H., Bell, D., Bi, Y., and Greer, K., 2003. *KNN Model-Based Approach in Classification*. Northern Ireland, UK, Springer.

Guo, P., Lijuan, W., Jun, S., and Fang, D., 2021. A hybrid unsupervised clustering-based anomaly detection method. *Tsinghua Science and Technology*, 26(2), pp. 146-153.

Haitao, G., and Zhenhua, L., 2018. *Safe Semi-Supervised Learning from Risky Labeled and Unlabeled Samples*. IEEE Xplore, Japan.

Hemant, K.G., and Rishabh, C., 2017. *Comprehensive Review on Supervised Machine*. IEEE, United States.

Hooman, H.R., Tran, N.K., Betts, E.V., Howell, L.P., and Green, R., 2019. Artificial intelligence and machine learning in pathology: The present landscape of supervised methods. *Academic Pathology*, 6, p. 2374289519873088.

Hui, Y., Ping, Y., and Duo, L., 2019. *Study on Deep Unsupervised Learning Optimization Algorithm Based on Cloud Computing*. IEEE, United States.

Iqbal, S., Qureshi, A.N., Aurangzeb, K., Alhussein, M., Haider, S.I., and Rida, I., 2023. AMIAC: Adaptive medical image analyzes and classification, a robust self-learning framework. Neural Computing and Applications, 1(1), pp. 1-29.

Jwan, K., Abas, A.A., and Tarik, R., 2024. *Exploring Public Service and Prosocial Motivation Using Machine Learning*. [Authorea Preprints].

Kristina, P.S., and Miin-Shen, Y., 2020. *Unsupervised K-Means Clustering Algorithm*. Vol. 8. IEEE Xplore, Japan.

Kushal, R.D., 2020. *Analysing the Role of Supervised and Unsupervised Machine Learning in IoT*. IEEE, United States.

Manjula, C.B., and Balachandra, M., 2016. *Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection*. India, ScienceDirect.

Mrudul, D., Ritu, S., Saniya, S., and Krutika, M., 2019. *Internet Traffic Detection using Naïve Bayes and K-Nearest Neighbors (KNN) Algorithm*. Indea, IEEE.

Nergz, S.M., and Beitollahi, H., 2022. Accurate classification in uncertainty dataset using particle swarm optimization-trained radial basis function. *NeuroQuantology*, 20(6), pp. 166-179.

Nitis, M., Prajak, C., Thongchai, K., Suwat, R., Sabir, B., and Pitchayakit, B., 2018. *Prediction of Prices for Used Car by Using Regression Models*. IEEE Xplore, Bangkok, Thailand.

Perera, A.T.D., and Kamalaruban, P., 2021. Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews*, 17(1), p. 110618.

Rui, Z., Feiping, N., and Xuelong, L., 2017. *Semi-Supervised Classification Via Both Label and Side Information*. IEEE, United States.

Sarfaraz, H., Kandel, P., Bolan, C.W., Wallace, M.B., and Bagci, U., 2019. Lung and pancreatic tumor characterization in the deep learning era: Novel supervised and unsupervised learning approaches. *IEEE Transactions on Medical Imaging*, 38(8), pp. 1777-1787.

Sabr, S.S., 2025. A Comprehensive Part-of-Speech Tagging to Standardize Central-Kurdish Language: A Research Guide for Kurdish Natural Language Processing Tasks. s.l.: arXiv.

Sung-Jin, K., Chan-Ho, K., Sang-Yong, J., and Yong-Jae, K., 2016. *Shape Optimization of a Hybrid Magnetic Torque Converter Using the Multiple Linear Regression Analysis*. IEEE, United States.

Upasana, 2019. *Al Zone*. Available from: https://dzone.com/articles/introduction-to-classification-algorithms [Last accessed on 2025 Oct 8].

Weihao, H., Di, S., and Theo, B., 2020. Guest editorial: Applications of artificial intelligence in modern power systems: Challenges and solutions. *Journal of Modern Power Systems and Clean Energy*, 8(6), pp. 1-2.

Wu, B., and He, S., 2023. Self-learning and explainable deep learning network toward the security of artificial intelligence of things. *The Journal of Supercomputing*, 79(4), pp. 4436-4467.

Xin, B., Chao, Z., Donghang, L., Yongjiao, S., and Yuliang, M., 2020. Efficient incremental semi-supervised classification over drifting and evolving social

streams. *IEEE Access*, 8, p. 1.

Yongjun, Z., and Siyu, Y., 2020. *Semi-Supervised Active Learning Image Classification method Based on Tri-Training Algorithm*. IEEE, United States.

Yuan, T., and Marc, C., 2019. A *Case Study on Visual-Inertial Odometry using Supervised, Semi-Supervised and Unsupervised Learning Methods*. IEEE, United States.

Zhang, K., Wang, J., Liu, T., Luo, Y., Loh, X.J., and Chen, X., 2021. Machine learning-reinforced noninvasive biosensors for healthcare. *Advanced Healthcare Materials*, 10(17), p. 2100734.

Zhou, C., Huang, B., and Fränti, P., 2022. A review of motion planning algorithms for intelligent robots. *Journal of Intelligent Manufacturing*, 33(2), pp. 387-424.