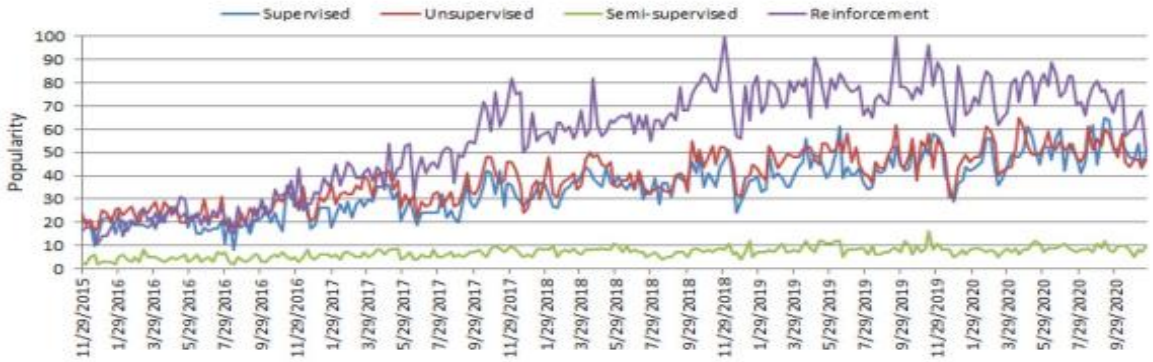
Machine Learning

Algorithms, Real-World Applications and Research Directions Iqbal H. Sarker1,2 Received: 27 January 2021 / Accepted: 12 March 2021 / Published online: 22 March 2021 © The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd 2021 Abstract In the current age of the Fourth Industrial Revolution (4IR or Industry 4.0), the digital world has a wealth of data, such as Internet of Things (IoT) data, cybersecurity data, mobile data, business data, social media data, health data, etc. To intelligently analyze these data and develop the corresponding smart and automated applications, the knowledge of artifcial intelligence (AI), particularly, machine learning (ML) is the key. Various types of machine learning algorithms such as supervised, unsupervised, semi-supervised, and reinforcement learning exist in the area. Besides, the deep learning, which is part of a broader family of machine learning methods, can intelligently analyze the data on a large scale. In this paper, we present a comprehensive view on these machine learning algorithms that can be applied to enhance the intelligence and the capabilities of an application. Thus, this study's key contribution is explaining the principles of diferent machine learning techniques and their applicability in various real-world application domains, such as cybersecurity systems, smart cities, healthcare, e-commerce, agriculture, and many more. We also highlight the challenges and potential research directions based on our study. Overall, this paper aims to serve as a reference point for both academia and industry professionals as well as for decision-makers in various real-world situations and application areas, particularly from the technical point of view. Keywords Machine learning · Deep learning · Artifcial intelligence · Data science · Data-driven decision-making ·

Introduction We live in the age of data, where everything around us is connected to a data source, and everything in our lives is digitally recorded [21, 103]. For instance, the current electronic world has a wealth of various kinds of data, such as the Internet of Things (IoT) data, cybersecurity data, smart city data, business data, smartphone data, social media data, health data, COVID-19 data, and many more. The data can be structured, semi-structured, or unstructured,

discussed briefy in Sect. "Types of Real-World Data and Machine Learning Techniques", which is increasing day-by-day. Extracting insights from these data can be used to build various intelligent applications in the relevant domains. For instance, to build a data-driven automated and intelligent cybersecurity system, the relevant cybersecurity data can be used [105]; to build personalized context-aware smart mobile applications, the relevant mobile data can be used [103], and so on. Thus, the data management tools and techniques having the capability of extracting insights or useful knowledge from the data in a timely and intelligent way is urgently needed, on which the real-world applications are based. Artifcial intelligence (AI), particularly, machine learning (ML) have grown rapidly in recent years in the context of data analysis and computing that typically allows the applications to function in an intelligent manner [95]. ML usually provides systems with the ability to learn and enhance from experience automatically without being specifcally programmed and generally refered to as the most popular latest technologies in the fourth industrial revolution (4IR or Industry 4.0) [103, 105]. "Industry 4.0" [114] is typically the ongoing automation of conventional manufacturing and industrial practices, including exploratory data processing, using new smart technologies such as machine learning automation. Thus, to intelligently analyze these data and to develop the corresponding real-world applications, machine learning algorithms is the key. The learning algorithms can be categorized into four major types, such as supervised, unsupervised, semi-supervised, and reinforcement learning in the area [75], discussed briefy in Sect. "Types of Real-World Data and MachineLearning Techniques". The popularity of these approaches to learning is increasing day-by-day, which is shown in Fig. 1, based on data collected from Google Trends [4] over the last fve years. The x-axis of the fgure indicates the specifc dates and the corresponding popularity score within the range of 0 (minimum) to 100 (maximum) has been shown in y-axis. According to Fig. 1, the popularity indication values for these learning types are low in 2015 and are increasing day by day. These statistics motivate us to study on machine learning in this paper, which can play an important role in the real-world through Industry 4.0 automation. In general, the efectiveness and the efciency of a machine learning solution depend on the nature and characteristics of data and the performance of the learning algorithms. In the area of machine learning algorithms, classifcation analysis, regression, data clustering, feature engineering and dimensionality reduction, association rule learning, or reinforcement learning techniques exist to efectively build data-

driven systems [41, 125]. Besides, deep learning originated from the artificial neural network that can be used to intelligently analyze data, which is known as part of a wider family of machine learning approaches [96]. Thus, selecting a proper learning algorithm that is suitable for the target application in a particular domain is challenging. The reason is that the purpose of diferent learning algorithms is diferent, even the outcome of diferent learning algorithms in a similar category may vary depending on the data characteristics [106]. Thus, it is important to understand the principles of various machine learning algorithms and their applicability to apply in various real-world application areas, such as IoT systems, cybersecurity services, business and recommendation systems, smart cities, healthcare and COVID-19, context-aware systems, sustainable agriculture, and many more that are explained briefy in Sect. "Applications of Machine Learning". Based on the importance and potentiality of "Machine Learning" to analyze the data mentioned above, in this paper, we provide a comprehensive view on various types of machine learning algorithms that can be applied to enhance the intelligence and the capabilities of an application. Thus, the key contribution of this study is explaining the principles and potentiality of diferent machine learning techniques, and their applicability in various realworld application areas mentioned earlier. The purpose of this paper is, therefore, to provide a basic guide for those academia and industry people who want to study, research, and develop data-driven automated and intelligent systems in the relevant areas based on machine learning techniques. The key contributions of this paper are listed as follows: – To defne the scope of our study by taking into account the nature and characteristics of various types of realworld data and the capabilities of various learning techniques. – To provide a comprehensive view on machine learning algorithms that can be applied to enhance the intelligence and capabilities of a data-driven application.



ig. 1 The worldwide popularity score of various types of ML algorithms (supervised, unsupervised, semi-supervised, and reinforcement) in a ange of 0 (min) to 100 (max) over time where x-axis represents the timestamp information and y-axis represents the corresponding score

– To discuss the applicability of machine learning-based solutions in various real-world application domains.

– To highlight and summarize the potential research directions within the scope of our study for intelligent data analysis and services. The rest of the paper is organized as follows. The next section presents the types of data and machine learning algorithms in a broader sense and defnes the scope of our study. We briefy discuss and explain diferent machine learning algorithms in the subsequent section followed by which various real-world application areas based on machine learning algorithms are discussed and summarized. In the penultimate section, we highlight several research issues and potential future directions, and the final section concludes

Types of Real-World Data and Machine Learning Techniques Machine learning algorithms typically consume and process data to learn the related patterns about individuals, business processes, transactions, events, and so on. In the following, we discuss various types of real-world data as well as categories of machine learning algorithms.

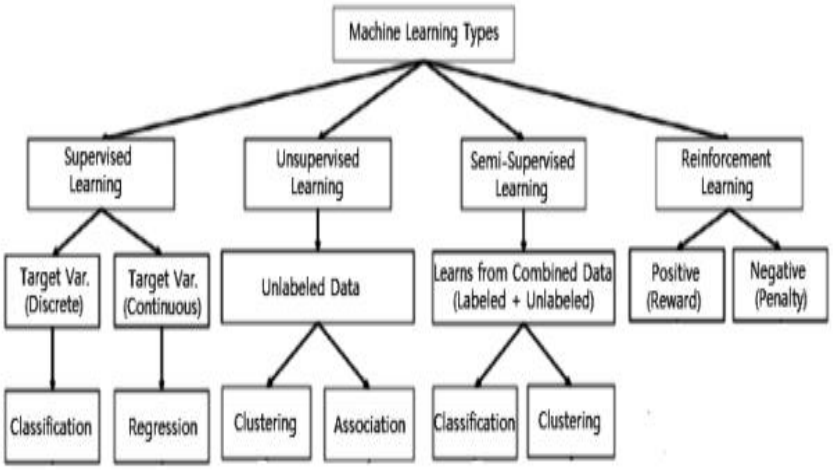
Types of Real-World Data Usually, the availability of data is considered as the key to construct a machine learning model or data-driven realworld systems [103, 105]. Data can be of various forms, such as structured, semi-structured, or unstructured [41, 72]. Besides, the "metadata" is another type that typically represents data about the data. In the following, we briefy discuss these types of data. – Structured: It has a well-defned structure, conforms to a data model following a standard order, which is highly organized and easily accessed, and

used by an entity or a computer program. In well-defned schemes, such as relational databases, structured data are typically stored, i.e., in a tabular format. For instance, names, dates, addresses, credit card numbers, stock information, geolocation, etc. are examples of structured data. – Unstructured: On the other hand, there is no pre-defned format or organization for unstructured data, making it much more difcult to capture, process, and analyze, mostly containing text and multimedia material. For example, sensor data, emails, blog entries, wikis, and word processing documents, PDF files, audio files, videos, images, presentations, web pages, and many

other types of business documents can be considered as unstructured data. – Semi-structured: Semi-structured data are not stored in a relational database like the structured data mentioned above, but it does have certain organizational properties that make it easier to analyze. HTML, XML, JSON documents, NoSQL databases, etc., are some examples of semi-structured data. – Metadata: It is not the normal form of data, but "data about data". The primary diference between "data" and "metadata" is that data are simply the material that can classify, measure, or even document something relative to an organization's data properties. On the other hand, metadata describes the relevant data information, giving it more signifcance for data users. A basic example of a document's metadata might be the author, fle size, date generated by the document, keywords to defne the document

Types of Machine Learning Techniques Machine Learning algorithms are mainly divided into four categories: Supervised learning, Unsupervised learning, Semi-supervised learning, and Reinforcement learning [75], as shown in Fig. 2. In the following, we briefy discuss each type of learning technique with the scope of their applicability to solve real-world problems. – Supervised: Supervised learning is typically the task of machine learning to learn a function that maps an input to an output based on sample input-output pairs [41]. It uses labeled training data and a collection of training examples to infer a function. Supervised learning is carried out when certain goals are identifed to be accomplished

Fig. 2 Various types of machine learning techniques



from a certain set of inputs [105], i.e., a taskdriven approach. The most common supervised tasks are "classifcation" that separates the data, and "regression" that fts the data. For instance, predicting the class label or sentiment of a piece of text, like a tweet or a product review, i.e., text classifcation, is an example of supervised learning. – Unsupervised: Unsupervised learning analyzes unlabeled datasets without the need for human interference, i.e., a data-driven process [41]. This is widely used for extracting generative features, identifying meaningful trends and structures, groupings in results, and exploratory purposes. The most common unsupervised learning tasks are clustering, density estimation, feature learning, dimensionality reduction, fnding association rules, anomaly detection, etc.

– Semi-supervised: Semi-supervised learning can be defined as a hybridization of the above-mentioned supervised and unsupervised methods, as it operates on both labeled and unlabeled data [41, 105]. Thus, it falls between learning "without supervision" and learning "with supervision". In the real world, labeled data could be rare in several contexts, and unlabeled data are numerous, where semi-supervised learning is useful [75]. The ultimate goal of a semi-supervised learning model is to provide a better outcome for prediction than that produced using the labeled data alone from the model. Some application areas where semi-supervised learning is used include machine translation, fraud detection, labeling data and text classifcation.

– Reinforcement: Reinforcement learning is a type of machine learning algorithm that enables software agents and machines to automatically evaluate the optimal behavior in a particular context or environment to improve its efciency [52], i.e., an environment-driven approach. This type of learning is based on reward or penalty, and its ultimate goal is to use insights obtained from environmental activists to take action to increase the reward or minimize the risk [75]. It is a powerful tool for training AI models that can help increase automation or optimize the operational efciency of sophisticated systems such as robotics, autonomous driving tasks, manufacturing and supply chain logistics, however, not preferable to use it for solving the basic or straightforward problems.

Thus, to build efective models in various application areas diferent types of machine learning techniques can play a signifcant role according to their learning capabilities, depending on the nature of the data discussed earlier, and the target outcome. In Table 1, we summarize various types of machine learning techniques with examples. In the following, we provide a comprehensive view of machine learning algorithms that can be applied to enhance the intelligence and capabilities of a data-driven application. Fig. 2 Various types of machine learning technique

Machine Learning

Tasks and Algorithms In this section, we discuss various machine learning algorithms that include classification analysis, regression analysis, data clustering, association rule learning, feature engineering for dimensionality reduction, as well as deep learning methods. A general structure of a machine learningbased predictive model has been shown in Fig. 3, where the model is trained from historical data in phase 1 and the outcome is generated in phase 2 for the new test data

Classifcation

- Analysis Classifcation is regarded as a supervised learning method in machine learning, referring to a problem of predictive modeling as well, where a class label is predicted for a given example [41]. Mathematically, it maps a function (f) from input variables (X) to output variables (Y) as target, label or categories. To predict the class of given data points, it can

be carried out on structured or unstructured data. For example, spam detection such as “spam” and “not spam” in email service providers can be a classifcation problem. In the following, we summarize the common classifcation problems.

- Binary classifcation: It refers to the classifcation tasks having two class labels such as “true and false” or “yes and no” [41]. In such binary classifcation tasks, one class could be the normal state, while the abnormal state could be another class. For instance, “cancer not detected” is the normal state of a task that involves a medical test, and “cancer detected” could be considered as the abnormal state. Similarly, “spam” and “not spam” in the above example of email service providers are considered as binary classifcation. Multiclass classifcation: Traditionally, this refers to those classifcation tasks having more than

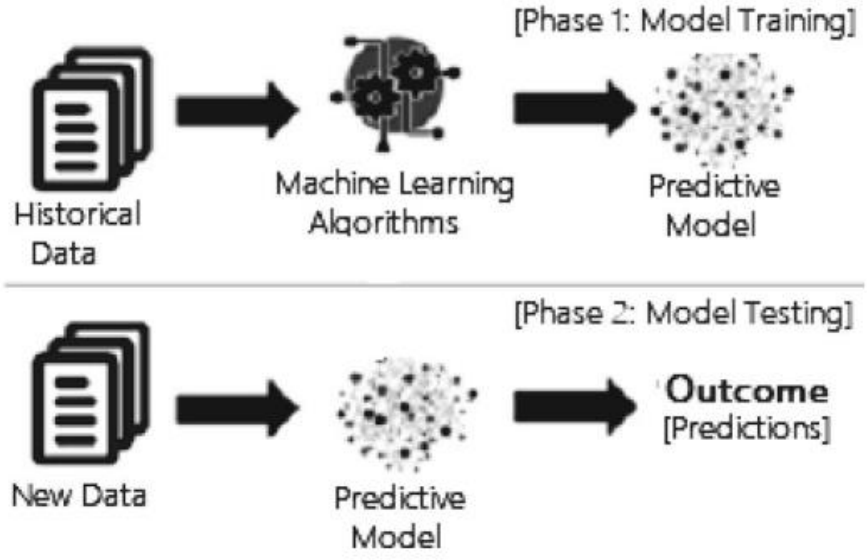


Fig. 3 A general structure of a machine learning based predictive model considering both the training and testing phase

labels [41]. The multiclass classifcation does not have the principle of normal and abnormal outcomes, unlike binary classifcation tasks. Instead, within a range of specifed classes, examples are classifed as belonging to one. For example, it can be a multiclass classifcation task to classify various types of network attacks in the NSL-KDD [119] dataset, where the attack categories are

classifed into four class labels, such as DoS (Denial of Service Attack), U2R (User to Root Attack), R2L (Root to Local Attack), and Probing Attack.

• Multi-label classifcation: In machine learning, multilabel classifcation is an important consideration where an example is associated with several classes or labels. Thus, it is a generalization of multiclass classifcation, where the classes involved in the problem are hierarchically structured, and each example may simultaneously belong to more than one class in each hierarchical level, e.g., multi-level text classifcation. For instance, Google news can be presented under the categories of a "city name", "technology", or "latest news", etc. Multilabel classifcation includes advanced machine learning algorithms that support predicting various mutually non-exclusive classes or labels, unlike traditional classifcation tasks where class labels are mutually exclusive

Many classifcation algorithms have been proposed in the machine learning and data science literature [41, 125]. In the following, we summarize the most common and popular methods that are used widely in various application areas.

– Naive Bayes (NB): The naive Bayes algorithm is based on the Bayes' theorem with the assumption of independence between each pair of features [51]. It works well and can be used for both binary and multi-class categories in many real-world situations, such as document or text classifcation, spam fltering, etc. To efectively classify the noisy instances in the data and to construct a robust prediction model, the NB classifer can be used [94]. The key beneft is that, compared to more sophisticated approaches, it needs a small amount of training data to estimate the necessary parameters and quickly [82]. However, its performance may afect due to its strong assumptions on features independence. Gaussian, Multinomial, Complement, Bernoulli, and Categorical are the common variants of NB classifer [82].

– Linear Discriminant Analysis (LDA): Linear Discriminant Analysis (LDA) is a linear decision boundary classifer created by ftting class conditional densities to data and applying Bayes' rule [51, 82]. This method is also known as a generalization of Fisher's linear discriminant, which projects a given dataset into a lower-dimensional space, i.e., a reduction of dimensionality that minimizes the complexity of the model or reduces the resulting model's computational costs. The standard LDA model usually suits each class with a Gaussian density, assuming that all classes share the same covariance matrix

[82]. LDA is closely related to ANOVA (analysis of variance) and regression analysis, which seek to express one dependent variable as a linear combination of other features or measurements.

– Logistic regression (LR): Another common probabilistic based statistical model used to solve classifcation issues in machine learning is Logistic Regression (LR) [64]. Logistic regression typically uses a logistic function to estimate the probabilities, which is also referred to as the mathematically defned sigmoid function in Eq. 1. It can overft high-dimensional datasets and works well when the dataset can be separated linearly. The regularization (L1 and L2) techniques [82] can be used to avoid over-ftting in such scenarios. The assumption of linearity between the dependent and independent variables is considered as a major drawback of Logistic Regression. It can be used for both classifcation and regression problems, but it is more commonly used for classifcation.

– K-nearest neighbors (KNN): K-Nearest Neighbors (KNN) [9] is an "instance-based learning" or non-generalizing learning, also known as a "lazy learning" algorithm. It does not focus on constructing a general internal model; instead, it stores all instances corresponding to training data in n-dimensional space. KNN uses data and classifes new data points based on similarity measures (e.g., Euclidean distance function) [82]. Classifcation is computed from a simple majority vote of the k nearest neighbors of each point. It is quite robust to noisy training data, and accuracy depends on the data quality. The biggest issue with KNN is to choose the optimal number of neighbors to be considered. KNN can be used both for classifcation as well as regression.

– Support vector machine (SVM): In machine learning, another common technique that can be used for classifcation, regression, or other tasks is a support vector machine (SVM) [56]. In high- or infnite-dimensional space, a support vector machine constructs a hyper-plane or set of hyper-planes. Intuitively, the hyper-plane, which has the greatest distance from the nearest training data points in any class, achieves a strong separation since, in general, the greater the margin, the lower the classifer's generalization error. It is efective in high-dimensional spaces and can behave diferently based on diferent mathematical functions known as the kernel. Linear, polynomial, radial basis function (RBF), sigmoid, etc.... are the popular kernel functions used in SVM

classifer [82]. However, when the data set contains more noise, such as overlapping target classes, SVM does not perform well.

– Decision tree (DT – Decision tree (DT): Decision tree (DT) [88] is a wellknown non-parametric supervised learning method. DT learning methods are used for both the classifcation and regression tasks [82]. ID3 [87], C4.5 [88], and CART [20] are well known for DT algorithms. Moreover, recently proposed BehavDT [100], and IntrudTree [97] by Sarker et al. are efective in the relevant application domains, such as user behavior analytics and cybersecurity analytics, respectively. By sorting down the tree from the root to some leaf nodes, as shown in Fig. 4, DT classifes the instances. Instances are classifed by checking the attribute defned by that node, starting at the root node of the tree, and then moving down the tree branch corresponding to the attribute value. For splitting, the most popular criteria are "gini" for the Gini impurity and "entropy" for the information gain that can be expressed mathematically as [82]. – Random forest (RF): A random forest classifer [19] is well known as an ensemble classifcation technique that is used in the feld of machine learning and data science in various application areas. This method uses

$$\text{Entropy} : H(x) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \quad (2)$$

$$\text{Gini}(E) = 1 - \sum_{i=1}^{c} p_i^2 \quad (3)$$

– Random forest (RF): A random forest classifer [19] is well known as an ensemble classifcation technique that is used in the feld of machine learning and data science in various application areas. This method uses

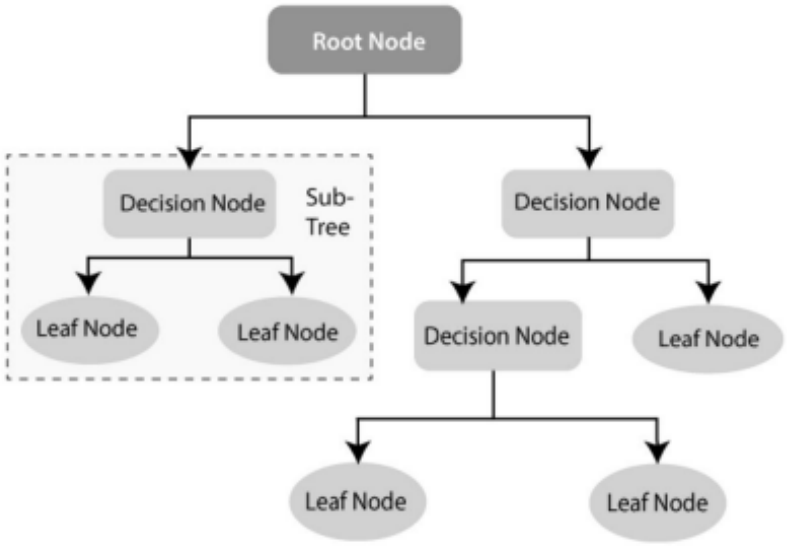


Fig. 4 An example of a decision tree structure

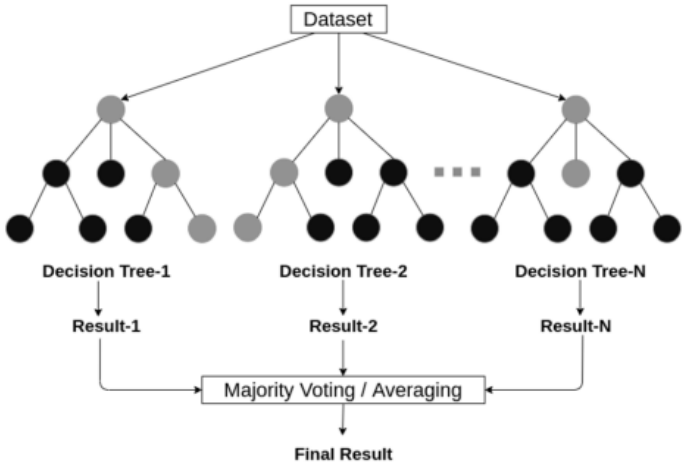


Fig. 5 An example of a random forest structure considering multiple decision trees

"parallel ensembling" which fts several decision tree classifers in parallel, as shown in Fig. 5, on diferent data set sub-samples and uses majority voting or averages for the outcome or fnal result. It thus minimizes the over-ftting problem and increases the prediction accuracy and control [82]. Therefore, the RF learning model with multiple decision trees is typically more accurate than a single decision tree based model [106]. To build a series of decision trees with controlled variation, it combines bootstrap aggregation (bagging) [18] and random

feature selection [11]. It is adaptable to both classifcation and regression problems and fts well for both categorical and continuous values.

– Adaptive Boosting (AdaBoost): Adaptive Boosting (AdaBoost) is an ensemble learning process that employs an iterative approach to improve poor classifers by learning from their errors. This is developed by Yoav Freund et al. [35] and also known as "metalearning". Unlike the random forest that uses parallel ensembling, Adaboost uses "sequential ensembling". It creates a powerful classifer by combining many poorly performing classifers to obtain a good classifer of high accuracy. In that sense, AdaBoost is called an adaptive classifer by signifcantly improving the efciency of the classifer, but in some instances, it can trigger overfts. AdaBoost is best used to boost the performance of decision trees, base estimator [82], on binary classifcation problems, however, is sensitive to noisy data and outliers.

– Extreme gradient boosting (XGBoost): Gradient Boosting, like Random Forests [19] above, is an ensemble learning algorithm that generates a fnal model based on a series of individual models, typically decision trees. The gradient is used to minimize the loss function, similar to how neural networks [41] use gradient descent to optimize weights. Extreme Gradient Boosting (XGBoost) is a form of gradient boosting that takes more detailed approximations into account when determining the best model [82]. It computes second-order gradients of the loss function to minimize loss and advanced regularization (L1 and L2) [82], which reduces over-ftting, and improves model generalization and performance. XGBoost is fast to interpret and can handle large-sized datasets well.

– Stochastic gradient descent (SGD): Stochastic gradient descent (SGD) [41] is an iterative method for optimizing an objective function with appropriate smoothness properties, where the word 'stochastic' refers to random probability. This reduces the computational burden, particularly in high-dimensional optimization problems, allowing for faster iterations in exchange for a lower convergence rate. A gradient is the slope of a function that calculates a variable's degree of change in response to another variable's changes. Mathematically, the Gradient Descent is a convex function whose output is a partial derivative of a set of its input parameters. Let, 훼 is the learning rate, and Ji is the training example cost of ith, then Eq. (4) represents the stochastic gradient descent weight update method at the j th iteration. In large-scale and

sparse machine learning, SGD has been successfully applied to problems often encountered in text classifcation and natural language processing [82]. However, SGD is sensitive to feature scaling and needs a range of hyperparameters, such as the regularization parameter and the number of iterations.

wj := wj – 훼 (4) 휕Ji 휕wj .

– Rule-based classifcation: The term rule-based classifcation can be used to refer to any classifcation scheme that makes use of IF-THEN rules for class prediction. Several classifcation algorithms such as Zero-R [125], One-R [47], decision trees [87, 88], DTNB [110], Ripple Down Rule learner (RIDOR) [125], Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [126] exist with the ability of rule generation. The decision tree is one of the most common rule-based classifcation algorithms among these techniques because it has several advantages, such as being easier to interpret; the ability to handle high-dimensional data; simplicity and speed; good accuracy; and the capability to produce rules for human clear and understandable classifcation [127] [128]. The decision tree-based rules also provide signifcant accuracy in a prediction model for unseen test cases [106]. Since the rules are easily interpretable, these rulebased classifers are often used to produce descriptive models that can describe a system including the entities and their relationship

Regression Analysis

Regression analysis includes several methods of machine learning that allow to predict a continuous (y) result variable based on the value of one or more (x) predictor variables [41]. The most signifcant distinction between classifcation and regression is that classifcation predicts distinct class labels, while regression facilitates the prediction of a continuous quantity. Figure 6 shows an example of how classifcation is diferent with regression models. Some overlaps are often found between the two types of machine learning algorithms. Regression models are now widely used in a variety of felds, including fnancial forecasting or prediction, cost estimation, trend analysis, marketing, time series estimation, drug response modeling, and many more. Some of the familiar types of regression algorithms are linear, polynomial, lasso and ridge regression, etc., which are explained briefy in the following.

• Simple and multiple linear regression: This is one of the most popular ML modeling techniques as well as a well-known regression technique. In this technique, the dependent variable is continuous, the independent variable(s) can be continuous or discrete, and the form of the regression line is linear. Linear regression creates a relationship between the dependent variable (Y) and one or more independent variables (X) (also known as regression line) using the best ft straight line [41]. It is defned by the following equations: where a is the intercept, b is the slope of the line, and e is the error term. This equation can be used to predict the

$$y = a + bx + e \quad (5)$$

$$y = a + b1x1 + b2x2 + \cdots + bnxn + e, \quad (6)$$

value of the target variable based on the given predictor variable(s). Multiple linear regression is an extension of simple linear regression that allows two or more predictor variables to model a response variable, y, as a linear function [41] defned in Eq. 6, whereas simple linear regression has only 1 independent variable, defned in Eq. 5.

• Polynomial regression: Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is not linear, but is the polynomial degree of nth in x [82]. The equation for polynomial regression is also derived from linear regression (polynomial regression of degree 1) equation, which is defned as below: Here, y is the predicted/target output, b0, b1, ...bn are the regression coefcients, x is an independent/ input variable. In simple words, we can say that if data are not distributed linearly, instead it is nth degree of polynomial then we use polynomial regression to get desired output.

• LASSO and ridge regression: LASSO and Ridge regression are well known as powerful techniques which are typically used for building learning models in presence of a large number of features, due to their capability to preventing over-ftting and reducing the complexity of the model. The LASSO (least absolute shrinkage and selection operator) regression model uses L1 regularization technique [82] that uses shrinkage, which penalizes "absolute value of magnitude of coefcients" (L1 penalty). As a result, LASSO appears to render coefcients to absolute zero. Thus, LASSO regression aims to fnd the subset of predictors that minimizes the predictiors error for a quantitative response

variable. On the other hand, ridge regression uses L2 regularization [82], which is the "squared magnitude of coefcients" (L2 penalty). Thus, ridge regression forces the weights to be small but never sets the coefcient value to zero, and does a nonsparse solution. Overall, LASSO regression is useful to obtain a subset of predictors by eliminating less important features, and ridge regression is useful when a data set has "multicollinearity" which refers to the predictors that are correlated with other predictors.

Cluster Analysis

Cluster analysis, also known as clustering, is an unsupervised machine learning technique for identifying and grouping related data points in large datasets without concern for the specifc outcome. It does grouping a collection of objects in such a way that objects in the same category, called a cluster, are in some sense more similar to each other than objects in other groups [41]. It is often used as a data analysis technique to discover interesting trends or patterns in data, e.g., groups of consumers based on their behavior. In a broad range of application areas, such as cybersecurity, e-commerce, mobile data processing, health analytics, user modeling and behavioral analytics, clustering can be used. In the following, we briefy discuss and summarize various types of clustering methods.

– Partitioning methods: Based on the features and similarities in the data, this clustering approach categorizes the data into multiple groups or clusters. The data scientists or analysts typically determine the number of clusters either dynamically or statically depending on the nature of the target applications, to produce for the methods of clustering. The most common clustering algorithms based on partitioning methods are K-means [69], K-Mediods [80], CLARA [55] etc.

– Density-based methods: To identify distinct groups or clusters, it uses the concept that a cluster in the data space is a contiguous region of high point density isolated from other such clusters by contiguous regions of low point density. Points that are not part of a cluster are considered as noise. The typical clustering algorithms based on density are DBSCAN [32], OPTICS [12] etc. The

density-based methods typically struggle with clusters of similar density and high dimensionality data.

– Hierarchical-based methods: Hierarchical clustering typically seeks to construct a hierarchy of clusters, i.e., the tree structure. Strategies for hierarchical clustering generally fall into two types: (i) Agglomerative—a "bottom-up" approach in which each observation begins in its cluster and pairs of clusters are combined as one, moves up the hierarchy, and (ii) Divisive—a "top-down" approach in which all observations begin in one cluster and splits are performed recursively, moves down the hierarchy, as shown in Fig 7. Our earlier proposed BOTS technique, Sarker et al. [102] is an example of a hierarchical, particularly, bottom-up clustering algorithm.

– Grid-based methods: To deal with massive datasets, gridbased clustering is especially suitable. To obtain clusters, the principle is frst to summarize the dataset with a grid representation and then to combine grid cells. STING [122], CLIQUE [6], etc. are the standard algorithms of grid-based clustering. – Model-based methods: There are mainly two types of model-based clustering algorithms: one that uses statistical learning, and the other based on a method of neural network learning [130]. For instance, GMM [89] is an example of a statistical learning method, and SOM [22] [96] is an example of a neural network learning method.

– Constraint-based methods: Constrained-based clustering is a semi-supervised approach to data clustering that uses constraints to incorporate domain knowledge. Application or user-oriented constraints are incorporated to perform the clustering. The typical algorithms of this kind of clustering are COP K-means [121], CMWK-Means [27], etc.

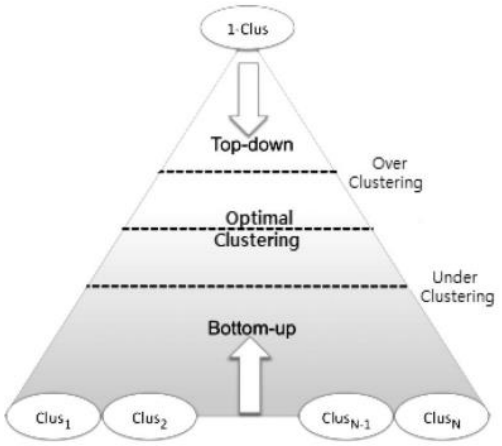


Fig. 7 A graphical interpretation of the widely-used hierarchical clustering (Bottom-up and top-down) technique

Many clustering algorithms have been proposed with the ability to grouping data in machine learning and data science literature [41, 125]. In the following, we summarize the popular methods that are used widely in various application areas.

– K-means clustering: K-means clustering [69] is a fast, robust, and simple algorithm that provides reliable results when data sets are well-separated from each other. The data points are allocated to a cluster in this algorithm in such a way that the amount of the squared distance between the data points and the centroid is as small as possible. In other words, the K-means algorithm identifes the k number of centroids and then assigns each data point to the nearest cluster while keeping the centroids as small as possible. Since it begins with a random selection of cluster centers, the results can be inconsistent. Since extreme values can easily afect a mean, the K-means clustering algorithm is sensitive to outliers. K-medoids clustering [91] is a variant of K-means that is more robust to noises and outliers.

– Mean-shift clustering: Mean-shift clustering [37] is a nonparametric clustering technique that does not require prior knowledge of the number of clusters or constraints on cluster shape. Mean-shift clustering aims to discover "blobs" in a smooth distribution or density of samples [82]. It is a centroid-based algorithm that works by updating centroid candidates to be the mean of the points in a given region. To form the fnal set of centroids, these candidates are

fltered in a post-processing stage to remove near-duplicates. Cluster analysis in computer vision and image processing are examples of application domains. Mean Shift has the disadvantage of being computationally expensive. Moreover, in cases of high dimension, where the number of clusters shifts abruptly, the mean-shift algorithm does not work well.

– DBSCAN: Density-based spatial clustering of applications with noise (DBSCAN) [32] is a base algorithm for density-based clustering which is widely used in data mining and machine learning. This is known as a nonparametric density-based clustering technique for separating high-density clusters from low-density clusters that are used in model building. DBSCAN's main idea is that a point belongs to a cluster if it is close to many points from that cluster. It can fnd clusters of various shapes and sizes in a vast volume of data that is noisy and contains outliers. DBSCAN, unlike k-means, does not require a priori specifcation of the number of clusters in the data and can fnd arbitrarily shaped clusters. Although k-means is much faster than DBSCAN, it is efcient at fnding high-density regions and outliers, i.e., is robust to outliers.

– GMM clustering: Gaussian mixture models (GMMs) are often used for data clustering, which is a distributionbased clustering algorithm. A Gaussian mixture model is a probabilistic model in which all the data points are produced by a mixture of a fnite number of Gaussian distributions with unknown parameters [82]. To fnd the Gaussian parameters for each cluster, an optimization algorithm called expectation-maximization (EM) [82] can be used. EM is an iterative method that uses a statistical model to estimate the parameters. In contrast to k-means, Gaussian mixture models account for uncertainty and return the likelihood that a data point belongs to one of the k clusters. GMM clustering is more robust than k-means and works well even with non-linear data distributions.

– Agglomerative hierarchical clustering: The most common method of hierarchical clustering used to group objects in clusters based on their similarity is agglomerative clustering. This technique uses a bottom-up approach, where each object is frst treated as a singleton cluster by the algorithm. Following that, pairs of clusters are merged one by one until all clusters have been merged into a single large cluster containing all objects. The result is a dendrogram, which is a tree-based representation of the elements. Single linkage [115], Complete linkage [116], BOTS [102] etc. are some examples of such techniques. The main

advantage of agglomerative hierarchical clustering over k-means is that the tree-structure hierarchy generated by agglomerative clustering is more informative than the unstructured collection of fat clusters returned by k-means, which can help to make better decisions in the relevant application areas.

Dimensionality Reduction and Feature Learning

In machine learning and data science, high-dimensional data processing is a challenging task for both researchers and application developers. Thus, dimensionality reduction which is an unsupervised learning technique, is important because it leads to better human interpretations, lower computational costs, and avoids overftting and redundancy by simplifying models. Both the process of feature selection and feature extraction can be used for dimensionality reduction. The primary distinction between the selection and extraction of features is that the "feature selection" keeps a subset of the original features [97], while "feature extraction" creates brand new ones [98]. In the following, we briefy discuss these techniques.

– Feature selection: The selection of features, also known as the selection of variables or attributes in the data, is the process of choosing a subset of unique features (variables, predictors) to use in building machine learning and data science model. It decreases a model's complexity by eliminating the irrelevant or less important features and allows for faster training of machine learning algorithms. A right and optimal subset of the selected features in a problem domain is capable to minimize the overftting problem through simplifying and generalizing the model as well as increases the model's accuracy [97]. Thus, "feature selection" [66, 99] is considered as one of the primary concepts in machine learning that greatly afects the efectiveness and efciency of the target machine learning model. Chi-squared test, Analysis of variance (ANOVA) test, Pearson's correlation coefcient, recursive feature elimination, are some popular techniques that can be used for feature selection.

– Feature extraction: In a machine learning-based model or system, feature extraction techniques usually provide a better understanding of the data, a way to improve prediction accuracy, and to reduce computational cost or training

time. The aim of "feature extraction" [66, 99] is to reduce the number of features in a dataset by generating new ones from the existing ones and then discarding the original features. The majority of the information found in the original set of features can then be summarized using this new reduced set of features. For instance, principal components analysis (PCA) is often used as a dimensionality-reduction technique to extract a lower- dimensional space creating new brand components from the existing features in a dataset [98].

Many algorithms have been proposed to reduce data dimensions in the machine learning and data science literature [41, 125]. In the following, we summarize the popular methods that are used widely in various application areas.

• Variance threshold: A simple basic approach to feature selection is the variance threshold [82]. This excludes all features of low variance, i.e., all features whose variance does not exceed the threshold. It eliminates all zero-variance characteristics by default, i.e., characteristics that have the same value in all samples. This feature selection algorithm looks only at the (X) features, not the (y) outputs needed, and can, therefore, be used for unsupervised learning.

• Pearson correlation: Pearson's correlation is another method to understand a feature's relation to the response variable and can be used for feature selection [99]. This method is also used for fnding the association between the features in a dataset. The resulting value is [–1, 1], where –1 means perfect negative correlation, +1 means perfect positive correlation, and 0 means that the two variables do not have a linear correlation. If two random variables represent X and Y, then the correlation coeffcient between X and Y is defned as [41]

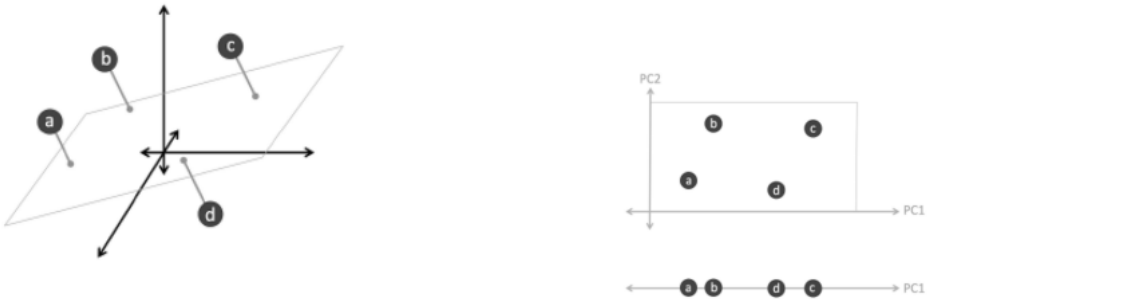
• ANOVA: Analysis of variance (ANOVA) is a statistical tool used to verify the mean values of two or more groups that difer signifcantly from each other. ANOVA assumes a linear relationship between the variables and the target and the variables' normal distribution. To statistically test the equality of means, the ANOVA method

(r(X, Y) = ∑n i=1(Xi – X¯)(Yi – Y¯) ∕∑n i=1(Xi – X¯)2 ∕∑n i=1(Yi – Y¯)2 . utilizes F tests. For feature selection, the results 'ANOVA F value' [82] of this test can be used where certain features independent of the goal variable can be omitted.

• Chi square: The chi-square 휒2 [82] statistic is an estimate of the diference between the efects of a series of events or variables observed and expected frequencies. The magnitude of the diference between the real and observed values, the degrees of freedom, and the sample size depends on 휒2. The chi-square 휒2 is commonly used for testing relationships between categorical variables. If Oi represents observed value and Ei represents expected value, then

• Recursive feature elimination (RFE): Recursive Feature Elimination (RFE) is a brute force approach to feature selection. RFE [82] fits the model and removes the weakest feature before it meets the specifed number of features. Features are ranked by the coefcients or feature signifcance of the model. RFE aims to remove dependencies and collinearity in the model by recursively removing a small number of features per iteration.

• Model-based selection: To reduce the dimensionality of the data, linear models penalized with the L1 regularization can be used. Least absolute shrinkage and selection operator (Lasso) regression is a type of linear regression that has the property of shrinking some of the coefcients to zero [82]. Therefore, that feature can be removed from the model. Thus, the penalized lasso regression method, often used in machine learning to select the subset of variables. Extra Trees Classifer [82] is an example of a tree-based estimator that can be used to compute impurity-based function importance, which can then be used to discard irrelevant features. • Principal component analysis (PCA): Principal component analysis (PCA) is a well-known unsupervised learn-ing



(a) An example of the original features in a 3D space. (b) Principal components in 2D and 1D space.

approach in the feld of machine learning and data science. PCA is a mathematical technique that transforms a set of correlated variables into a set of uncorrelated variables known as principal components [48, 81]. Figure 8 shows an example of the efect of PCA on various dimensions space, where Fig. 8a shows the original features in 3D space, and Fig. 8b shows the created principal components PC1 and PC2 onto a 2D plane, and 1D line with the principal component PC1 respectively. Thus, PCA can be used as a feature extraction technique that reduces the dimensionality of the datasets, and to build an efective machine learning model [98]. Technically, PCA identifes the completely transformed with the highest eigenvalues of a covariance matrix and then uses those to project the data into a new subspace of equal or fewer dimensions [82].

Association rule learning is a rule-based machine learning approach to discover interesting relationships, "IFTHEN" statements, in large datasets between variables [7]. One example is that "if a customer buys a computer or laptop (an item), s/he is likely to also buy anti-virus software (another item) at the same time". Association rules are employed today in many application areas, including IoT services, medical diagnosis, usage behavior analytics, web usage mining, smartphone applications, cybersecurity applications, and bioinformatics. In comparison to sequence mining, association rule learning does not usually take into account the order of things within or across transactions. A common way of measuring the usefulness of association rules is to use its parameter, the 'support' and 'confdence', which is introduced in [7]. In the data mining literature, many association rule learning methods have been proposed, such as logic dependent [34], frequent pattern based [8, 49, 68], and

tree-based [42]. The most popular association rule learning algorithms are summarized below

. – AIS and SETM: AIS is the frst algorithm proposed by Agrawal et al. [7] for association rule mining. The AIS algorithm's main downside is that too many candidate itemsets are generated, requiring more space and wasting a lot of efort. This algorithm calls for too many passes over the entire dataset to produce the rules. Another approach SETM [49] exhibits good performance and stable behavior with execution time; however, it sufers from the same faw as the AIS algorithm.

– Apriori: For generating association rules for a given dataset, Agrawal et al. [8] proposed the Apriori, Apriori-TID, and Apriori-Hybrid algorithms. These later algorithms outperform the AIS and SETM mentioned above due to the Apriori property of frequent itemset [8]. The term 'Apriori' usually refers to having prior knowledge of frequent itemset properties. Apriori uses a "bottom-up" approach, where it generates the candidate itemsets. To reduce the search space, Apriori uses the property "all subsets of a frequent itemset must be frequent; and if an itemset is infrequent, then all its supersets must also be infrequent". Another approach predictive Apriori [108] can also generate rules; however, it receives unexpected results as it combines both the support and confdence. The Apriori [8] is the widely applicable techniques in mining association rules.

– ECLAT: This technique was proposed by Zaki et al. [131] and stands for Equivalence Class Clustering and bottomup Lattice Traversal. ECLAT uses a depth-frst search to fnd frequent itemsets. In contrast to the Apriori [8] algorithm, which represents data in a horizontal pattern, it represents data vertically. Hence, the ECLAT algorithm is more efcient and scalable in the area of association rule learning. This algorithm is better suited for small and medium datasets whereas the Apriori algorithm is used for large datasets

. – FP-Growth: Another common association rule learning technique based on the frequent-pattern tree (FP-tree) proposed by Han et al. [42] is Frequent Pattern Growth, known as FP-Growth. The key diference with Apriori is that while generating rules, the Apriori algorithm [8] generates frequent candidate itemsets; on the other hand, the FP-growth algorithm [42] prevents candidate generation and thus produces a tree by the successful strategy of 'divide and conquer' approach. Due to its sophistication, however, FP-Tree is challenging to

use in an interactive mining environment [133]. Thus, the FP-Tree would not ft into memory for massive data sets, making it challenging to process big data as well. Another solution is RARM (Rapid Association Rule Mining) proposed by Das et al. [26] but faces a related FP-tree issue [133].

– ABC-RuleMiner: A rule-based machine learning method, recently proposed in our earlier paper, by Sarker et al. [104], to discover the interesting non-redundant rules to provide real-world intelligent services. This algorithm effectively identifies the redundancy in associations by taking into account the impact or precedence of the related contextual features and discovers a set of nonredundant association rules. This algorithm frst constructs an association generation tree (AGT), a top-down approach, and then extracts the association rules through traversing the tree. Thus, ABC-RuleMiner is more potent than traditional rule-based methods in terms of both non-redundant rule generation and intelligent decisionmaking, particularly in a context-aware smart computing environment, where human or user preference are involved. Among the association rule learning techniques discussed above, Apriori [8] is the most widely used algorithm for discovering association rules from a given dataset [133]. The main strength of the association learning technique is its comprehensiveness, as it generates all associations that satisfy the user-specifed constraints, such as minimum support and confdence value. The ABC-RuleMiner approach [104] discussed earlier could give signifcant results in terms of non-redundant rule generation and intelligent decisionmaking for the relevant application areas in the real world.

Reinforcement Learning

Reinforcement learning (RL) is a machine learning technique that allows an agent to learn by trial and error in an interactive environment using input from its actions and experiences. Unlike supervised learning, which is based on given sample data or examples, the RL method is based on interacting with the environment. The problem to be solved in reinforcement learning (RL) is defned as a Markov Decision Process (MDP) [86], i.e., all about sequentially making decisions. An RL problem typically includes four elements such as Agent, Environment, Rewards, and Policy.

RL can be split roughly into Model-based and Modelfree techniques. Model-based RL is the process of inferring optimal behavior from a model of the environment by performing actions and observing the results, which include the next state and the immediate reward [85]. AlphaZero, AlphaGo [113] are examples of the model-based approaches. On the other hand, a model-free approach does not use the distribution of the transition probability and the reward function associated with MDP. Q-learning, Deep Q Network, Monte Carlo Control, SARSA (State–Action–Reward–State–Action), etc. are some examples of model-free algorithms [52]. The policy network, which is required for model-based RL but not for modelfree, is the key diference between model-free and modelbased learning. In the following, we discuss the popular RL algorithms.

– Monte Carlo methods: Monte Carlo techniques, or Monte Carlo experiments, are a wide category of computational algorithms that rely on repeated random sampling to obtain numerical results [52]. The underlying concept is to use randomness to solve problems that are deterministic in principle. Optimization, numerical integration, and making drawings from the probability distribution are the three problem classes where Monte Carlo techniques are most commonly used.

– Q-learning: Q-learning is a model-free reinforcement learning algorithm for learning the quality of behaviors that tell an agent what action to take under what conditions [52]. It does not need a model of the environment (hence the term "model-free"), and it can deal with stochastic transitions and rewards without the need for adaptations. The 'Q' in Q-learning usually stands for quality, as the algorithm calculates the maximum expected rewards for a given behavior in a given state.

– Deep Q-learning: The basic working step in Deep Q-Learning [52] is that the initial state is fed into the neural network, which returns the Q-value of all possible actions as an output. Still, when we have a reasonably simple setting to overcome, Q-learning works well. However, when the number of states and actions becomes more complicated, deep learning can be used as a function approximator.

Reinforcement learning, along with supervised and unsupervised learning, is one of the basic machine learning paradigms. RL can be used to solve numerous

real-world problems in various felds, such as game theory, control theory, operations analysis, information theory, simulation-based optimization, manufacturing, supply chain logistics, multiagent systems, swarm intelligence, aircraft control, robot motion control, and many more.

Artifcial Neural Network and Deep Learning

Deep learning is part of a wider family of artifcial neural networks (ANN)-based machine learning approaches with representation learning. Deep learning provides a computational architecture by combining several processing layers, such as input, hidden, and output layers, to learn from data [41]. The main advantage of deep learning over traditional machine learning methods is its better performance in several cases, particularly learning from large datasets [105, 129]. Figure 9 shows a general performance of deep learning over machine learning considering the increasing amount of data. However, it may vary depending on the data characteristics and experimental set up. The most common deep learning algorithms are: Multilayer Perceptron (MLP), Convolutional Neural Network

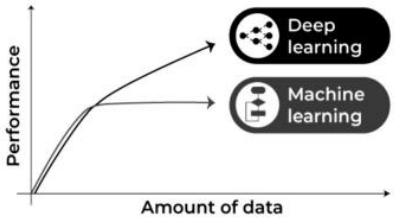


Fig. 9 Machine learning and deep learning performance in general with the amount of data

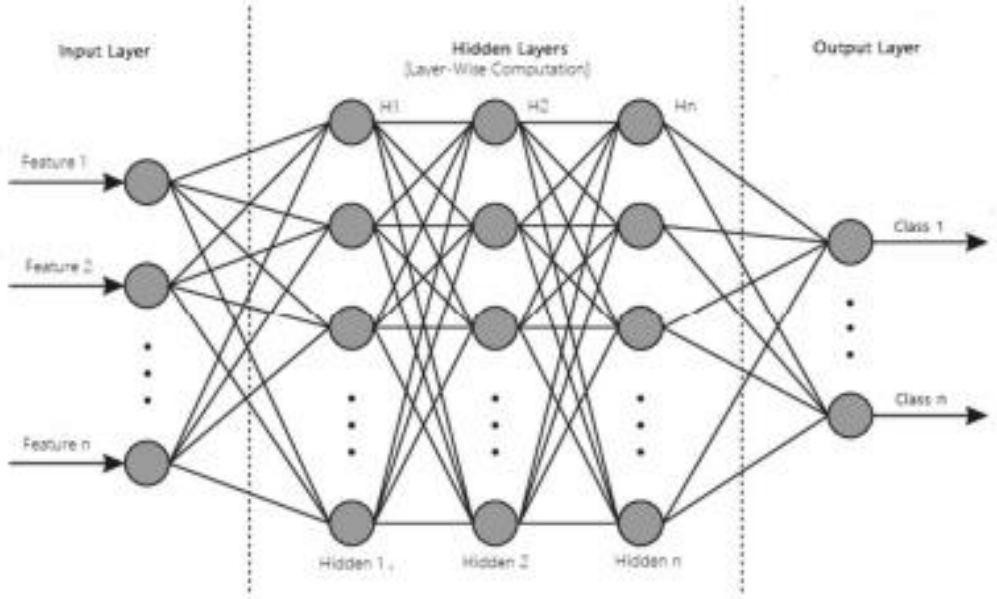
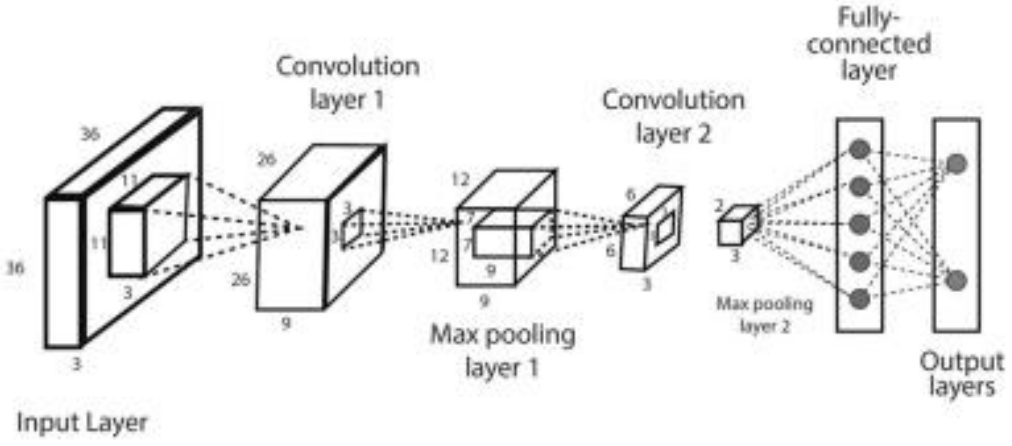


Fig. 10 A structure of an artificial neural network modeling with multiple processing layers



(CNN, or ConvNet), Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) [96]. In the following, we discuss various types of deep learning methods that can be used to build efective data-driven models for various purposes.

– MLP: The base architecture of deep learning, which is also known as the feed-forward artifcial neural network, is called a multilayer perceptron (MLP) [82]. A typical MLP is a fully connected network consisting of an input layer, one or more hidden layers, and an output layer, as shown in Fig. 10. Each node in one layer connects to each node in the following layer at a certain weight. MLP utilizes the "Backpropagation" technique [41], the most "fundamental building block" in a neural network, to adjust the weight values internally while building the model. MLP is sensitive to scaling features and allows a variety of hyperparameters to be tuned, such as the number of hidden layers, neurons, and iterations, which can result in a computationally costly model.

– CNN or ConvNet: The convolution neural network (CNN) [65] enhances the design of the standard ANN, consisting of convolutional layers, pooling layers, as well as fully connected layers, as shown in Fig. 11. As it takes the advantage of the two-dimensional (2D) structure of the input data, it is typically broadly used in several areas such as image and video recognition, image processing and classifcation, medical image analysis, natural language processing, etc. While CNN has a greater computational burden, without any manual intervention, it has the advantage of automatically detecting the important features, and hence CNN is considered to be more powerful than conventional ANN. A number of advanced deep learning models based on CNN can be used in the feld, such as

AlexNet [60], Xception [24], Inception [118], Visual Geometry Group (VGG) [44], ResNet [45], etc.

– LSTM-RNN: Long short-term memory (LSTM) is an artifcial recurrent neural network (RNN) architecture used in the area of deep learning [38]. LSTM has feedback links, unlike normal feed-forward neural networks. LSTM networks are well-suited for analyzing and learning sequential data, such as classifying, processing, and predicting data based on time series data, which diferentiates it from other conventional networks. Thus, LSTM can be used when the data are in a sequential format, such as time, sentence, etc., and commonly applied in the area of time-series analysis, natural language processing, speech recognition, etc.

In addition to these most common deep learning methods discussed above, several other deep learning approaches [96] exist in the area for various purposes. For instance, the self-organizing map (SOM) [58] uses unsupervised learning to represent the high-dimensional data by a 2D grid map, thus achieving dimensionality reduction. The autoencoder (AE) [15] is another learning technique that is widely used for dimensionality reduction as well and feature extraction in unsupervised learning tasks. Restricted Boltzmann machines (RBM) [46] can be used for dimensionality reduction, classifcation, regression, collaborative fltering, feature learning, and topic modeling. A deep belief network (DBN) is typically composed of simple, unsupervised networks such as restricted Boltzmann machines (RBMs) or autoencoders, and a backpropagation neural network (BPNN) [123]. A generative adversarial network (GAN) [39] is a form of the network for deep learning that can generate data with characteristics close to the actual data input. Transfer learning is currently very common because it can train deep neural networks with comparatively low data, which is typically the re-use of a new problem with a pre-trained model [124]. A brief discussion of these artifcial neural networks (ANN) and deep learning (DL) models are summarized in our earlier paper Sarker et al. [96]. Overall, based on the learning techniques discussed above, we can conclude that various types of machine learning techniques, such as classifcation analysis, regression, data clustering, feature selection and extraction, and dimensionality reduction, association rule learning, reinforcement learning, or deep learning techniques, can play a signifcant role for various purposes according to their capabilities. In the following section, we discuss several application areas based on machine learning algorithms

. Applications of Machine Learning

In the current age of the Fourth Industrial Revolution (4IR), machine learning becomes popular in various application areas, because of its learning capabilities from the past and making intelligent decisions. In the following, we summarize and discuss ten popular application areas of machine learning technology.

– Predictive analytics and intelligent decision-making: A major application feld of machine learning is intelligent decision-making by data-driven predictive analytics [21, 70]. The basis of predictive analytics is capturing and exploiting relationships between explanatory variables and predicted variables from previous events to predict the unknown outcome [41]. For instance, identifying suspects or criminals after a crime has been committed, or detecting credit card fraud as it happens. Another application, where machine learning algorithms can assist retailers in better understanding consumer preferences and behavior, better manage inventory, avoiding out-of-stock situations, and optimizing logistics and warehousing in e-commerce. Various machine learning algorithms such as decision trees, support vector machines, artifcial neural networks, etc. [106, 125] are commonly used in the area. Since accurate predictions provide insight into the unknown, they can improve the decisions of industries, businesses, and almost any organization, including government agencies, e-commerce, telecommunications, banking and fnancial services, healthcare, sales and marketing, transportation, social networking, and many others.

– Cybersecurity and threat intelligence: Cybersecurity is one of the most essential areas of Industry 4.0. [114], which is typically the practice of protecting networks, systems, hardware, and data from digital attacks [114]. Machine learning has become a crucial cybersecurity technology that constantly learns by analyzing data to identify patterns, better detect malware in encrypted trafc, fnd insider threats, predict where bad neighborhoods are online, keep people safe while browsing, or secure data in the cloud by uncovering suspicious activity. For instance, clustering techniques can be used to identify cyber-anomalies, policy violations, etc. To detect various types of cyber-attacks or intrusions machine learning classifcation models by taking into account the impact of security features are useful [97]. Various deep learning-based security models can also be used on the large scale of security datasets [96, 129]. Moreover, security policy rules generated by association rule learning

techniques can play a signifcant role to build a rule-based security system [105]. Thus, we can say that various learning techniques discussed in Sect. Machine Learning Tasks and Algorithms, can enable cybersecurity professionals to be more proactive inefciently preventing threats and cyber-attacks.

– Internet of things (IoT) and smart cities: Internet of Things (IoT) is another essential area of Industry 4.0. [114], which turns everyday objects into smart objects by allowing them to transmit data and automate tasks without the need for human interaction. IoT is, therefore, considered to be the big frontier that can enhance almost all activities in our lives, such as smart governance, smart home, education, communication, transportation, retail, agriculture, health care, business, and many more [70]. Smart city is one of IoT's core felds of application, using technologies to enhance city services and residents' living experiences [132, 135]. As machine learning utilizes experience to recognize trends and create models that help predict future behavior and events, it has become a crucial technology for IoT applications [103]. For example, to predict trafc in smart cities, parking availability prediction, estimate the total usage of energy of the citizens for a particular period, make context-aware and timely decisions for the people, etc. are some tasks that can be solved using machine learning techniques according to the current needs of the people.

– Traffic prediction and transportation: Transportation systems have become a crucial component of every country's economic development. Nonetheless, several cities around the world are experiencing an excessive rise in trafc volume, resulting in serious issues such as delays, trafc congestion, higher fuel prices, increased CO2 pollution, accidents, emergencies, and a decline in modern society's quality of life [40]. Thus, an intelligent transportation system through predicting future trafc is important, which is an indispensable part of a smart city. Accurate trafc prediction based on machine and deep learning modeling can help to minimize the issues [17, 30, 31]. For example, based on the travel history and trend of traveling through various routes, machine learning can assist transportation companies in predicting sustainable modes of transportation, and limit real-world disruption by modeling and visualizing future changes.

– Healthcare and COVID-19 pandemic: Machine learning can help to solve diagnostic and prognostic problems in a variety of medical domains, such as disease prediction, medical knowledge extraction, detecting regularities in data,

patient management, etc. [33, 77, 112]. Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus, according to the World Health Organization (WHO) [3]. Recently, the learning techniques have become popular in the battle against COVID-19 [61, 63]. For the COVID-19 pandemic, the learning techniques are used to classify patients at high risk, their mortality rate, and other anomalies [61]. It can also be used to better understand the virus's origin, COVID-19 outbreak prediction, as well as for disease diagnosis and treatment [14, 50]. With the help of machine learning, researchers can forecast where and when, the COVID-19 is likely to spread, and notify those regions to match the required arrangements. Deep learning also provides exciting solutions to the problems of medical image processing and is seen as a crucial technique for potential applications, particularly for COVID19 pandemic [10, 78, 111]. Overall, machine and deep learning techniques can help to fght the COVID-19 virus and the pandemic as well as intelligent clinical decisions making in the domain of healthcare.

– E-commerce and product recommendations: Product recommendation is one of the most well known and widely used applications of machine learning, and it is one of the most prominent features of almost any e-commerce website today. Machine learning technology can assist businesses in analyzing their consumers' purchasing histories and making customized product suggestions for their next purchase based on their behavior and preferences. E-commerce companies, for example, can easily position product suggestions and ofers by analyzing browsing trends and click-through rates of specifc items. Using predictive modeling based on machine learning techniques, many online retailers, such as Amazon [71], can better manage inventory, prevent out-of-stock situations, and optimize logistics and warehousing. The future of sales and marketing is the ability to capture, evaluate, and use consumer data to provide a customized shopping experience. Furthermore, machine learning techniques enable companies to create packages and content that are tailored to the needs of their customers, allowing them to maintain existing customers while attracting new ones.

– NLP and sentiment analysis: Natural language processing (NLP) involves the reading and understanding of spoken or written language through the medium of a computer [79, 103]. Thus, NLP helps computers, for instance, to read a text, hear speech, interpret it, analyze sentiment, and decide which aspects are signifcant, where machine learning techniques can be used. Virtual personal

assistant, chatbot, speech recognition, document description, language or machine translation, etc. are some examples of NLP-related tasks. Sentiment Analysis [90] (also referred to as opinion mining or emotion AI) is an NLP sub-feld that seeks to identify and extract public mood and views within a given text through blogs, reviews, social media, forums, news, etc. For instance, businesses and brands use sentiment analysis to understand the social sentiment of their brand, product, or service through social media platforms or the web as a whole. Overall, sentiment analysis is considered as a machine learning task that analyzes texts for polarity, such as "positive", "negative", or "neutral" along with more intense emotions like very happy, happy, sad, very sad, angry, have interest, or not interested etc.

– Image, speech and pattern recognition: Image recognition [36] is a well-known and widespread example of machine learning in the real world, which can identify an object as a digital image. For instance, to label an x-ray as cancerous or not, character recognition, or face detection in an image, tagging suggestions on social media, e.g., Facebook, are common examples of image recognition. Speech recognition [23] is also very popular that typically uses sound and linguistic models, e.g., Google Assistant, Cortana, Siri, Alexa, etc. [67], where machine learning methods are used. Pattern recognition [13] is defned as the automated recognition of patterns and regularities in data, e.g., image analysis. Several machine learning techniques such as classifcation, feature selection, clustering, or sequence labeling methods are used in the area.

– Sustainable agriculture: Agriculture is essential to the survival of all human activities [109]. Sustainable agriculture practices help to improve agricultural productivity while also reducing negative impacts on the environment [5, 25, 109]. The sustainable agriculture supply chains are knowledge-intensive and based on information, skills, technologies, etc., where knowledge transfer encourages farmers to enhance their decisions to adopt sustainable agriculture practices utilizing the increasing amount of data captured by emerging technologies, e.g., the Internet of Things (IoT), mobile technologies and devices, etc. [5, 53, 54]. Machine learning can be applied in various phases of sustainable agriculture, such as in the pre-production phase - for the prediction of crop yield, soil properties, irrigation requirements, etc.; in the production phase—for weather prediction, disease detection, weed detection, soil nutrient management, livestock management, etc.; in processing phase—for demand

estimation, production planning, etc. and in the distribution phase - the inventory management, consumer analysis, etc.

– User behavior analytics and context-aware smartphone applications: Context-awareness is a system's ability to capture knowledge about its surroundings at any moment and modify behaviors accordingly [28, 93]. Context-aware computing uses software and hardware to automatically collect and interpret data for direct responses. The mobile app development environment has been changed greatly with the power of AI, particularly, machine learning techniques through their learning capabilities from contextual data [103, 136]. Thus, the developers of mobile apps can rely on machine learning to create smart apps that can understand human behavior, support, and entertain users [107, 137, 140]. To build various personalized data-driven context-aware systems, such as smart interruption management, smart mobile recommendation, context-aware smart searching, decision-making that intelligently assist end mobile phone users in a pervasive computing environment, machine learning techniques are applicable. For example, contextaware association rules can be used to build an intelligent phone call application [104]. Clustering approaches are useful in capturing users' diverse behavioral activities by taking into account data in time series [102]. To predict the future events in various contexts, the classifcation methods can be used [106, 139]. Thus, various learning techniques discussed in Sect. "Machine Learning Tasks and Algorithms" can help to build context-aware adaptive and smart applications according to the preferences of the mobile phone users.

In addition to these application areas, machine learningbased models can also apply to several other domains such as bioinformatics, cheminformatics, computer networks, DNA sequence classifcation, economics and banking, robotics, advanced engineering, and many more.

Challenges and Research Directions

Our study on machine learning algorithms for intelligent data analysis and applications opens several research issues in the area. Thus, in this section, we summarize and discuss the challenges faced and the potential research opportunities and future directions.

In general, the effectiveness and the efficiency of a machine learning-based solution depend on the nature and characteristics of the data, and the

performance of the learning algorithms. To collect the data in the relevant domain, such as cybersecurity, IoT, healthcare and agriculture discussed in Sect. "Applications of Machine Learning" is not straightforward, although the current cyberspace enables the production of a huge amount of data with very high frequency. Thus, collecting useful data for the target machine learning-based applications, e.g., smart city applications, and their management is important to further analysis. Therefore, a more in-depth investigation of data collection methods is needed while working on the real-world data. Moreover, the historical data may contain many ambiguous values, missing values, outliers, and meaningless data. The machine learning algorithms, discussed in Sect "Machine Learning Tasks and Algorithms" highly impact on data quality, and availability for training, and consequently on the resultant model. Thus, to accurately clean and pre-process the diverse data collected from diverse sources is a challenging task. Therefore, efectively modifying or enhance existing pre-processing methods, or proposing new data preparation techniques are required to efectively use the learning algorithms in the associated application domain. To analyze the data and extract insights, there exist many machine learning algorithms, summarized in Sect. "Machine Learning Tasks and Algorithms". Thus, selecting a proper learning algorithm that is suitable for the target application is challenging. The reason is that the outcome of diferent learning algorithms may vary depending on the data characteristics [106]. Selecting a wrong learning algorithm would result in producing unexpected outcomes that may lead to loss of efort, as well as the model's efectiveness and accuracy. In terms of model building, the techniques discussed in Sect. "Machine Learning Tasks and Algorithms" can directly be used to solve many real-world issues in diverse domains, such as cybersecurity, smart cities and healthcare summarized in Sect. "Applications of Machine Learning". However, the hybrid learning model, e.g., the ensemble of methods, modifying or enhancement of the existing learning techniques, or designing new learning methods, could be a potential future work in the area. Thus, the ultimate success of a machine learning-based solution and corresponding applications mainly depends on both the data and the learning algorithms. If the data are bad to learn, such as non-representative, poor-quality, irrelevant features, or insufcient quantity for training, then the machine learning models may become useless or will produce lower accuracy. Therefore, efectively processing the data and handling the diverse learning algorithms are important, for a machine learning-based solution.

Conclusion

In this paper, we have conducted a comprehensive overview of machine learning algorithms for intelligent data analysis and applications. According to our goal, we have briefy discussed how various types of machine learning methods can be used for making solutions to various real-world issues. A successful machine learning model depends on both the data and the performance of the learning algorithms. The sophisticated learning algorithms then need to be trained through the collected real-world data and knowledge related to the target application before the system can assist with intelligent decision-making. We also discussed several popular application areas based on machine learning techniques to highlight their applicability in various real-world issues. Finally, we have summarized and discussed the challenges faced and the potential research opportunities and future directions in the area. Therefore, the challenges that are identifed create promising research opportunities in the feld which must be addressed with efective solutions in various application areas. Overall, we believe that our study on machine learning-based solutions opens up a promising direction and can be used as a reference guide for potential research and applications for both academia and industry professionals as well as for decision-makers, from a technical point of view

Declaration

Conflict of internet The author declares no conflict of interest