doi:10.20944/preprints202103.0216.v1

Machine Learning: Algorithms, Real-World Applications and Research Directions

Iqbal H. Sarker[1,2*]

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 real-world applications, the knowledge of artificial 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 different machine learning techniques and their applicability in various real-world *applications* areas, such as cybersecurity, smart cities, healthcare, business, 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 not only the application developers but also the decision-makers and researchers in various real-world application areas, particularly from the technical point of view.

Keywords machine learning; deep learning; artificial intelligence; data science; data-driven decision making; predictive analytics; intelligent applications;

1 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, etc. The data can be structured, semi-structured, or unstructured, discussed briefly in Section 2, 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.

Artificial 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 specifically programmed and is generally referred 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 au-

[1] Swinburne University of Technology, Melbourne, VIC 3122, Australia.
[2] Department of Computer Science and Engineering, Chittagong University of Engineering & Technology, Chittagong-4349, Bangladesh.
[*] Correspondance: msarker@swin.edu.au (Iqbal H. Sarker)

doi:10.20944/preprints202103.0216.v1

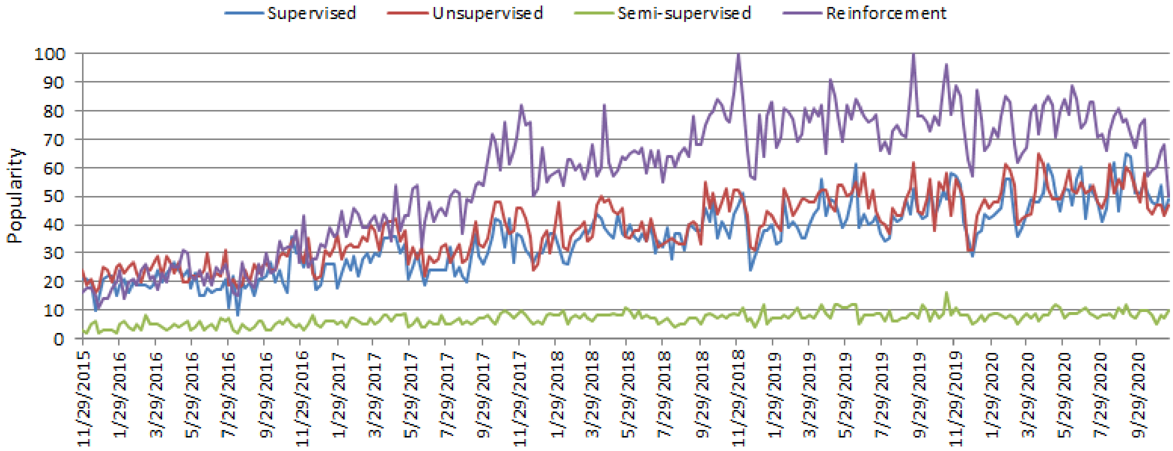


Fig. 1: The worldwide popularity score of various types of ML algorithms (supervised, unsupervised, semi-supervised, and reinforcement) in a range of 0 (min) to 100 (max) over time where x-axis represents the timestamp information and y-axis represents the corresponding score.

tomation. 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 briefly in Section 2. 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 five years. The *x-axis* of the figure indicates the specific 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 effectiveness and the efficiency 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, classification analysis, regression, data clustering, feature engineering and dimensionality reduction, association rule learning, or reinforcement learning techniques exist to effectively 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 different learning algorithms is different, even the outcome of different learning algorithms in a similar category may vary depending on the data characteristics [106]. Thus, it's 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 briefly in Section 4.

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 different machine learning techniques, and their applicability in various real-world 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 define the scope of our study by taking into account the nature and characteristics of various types of real-world 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.

- 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. Section 2 presents the types of data and machine learning algorithms in a broader sense and defines the scope of our study. We briefly discuss and explain different machine learning algorithms in Section 3. Various real-world application areas based on machine learning algorithms are discussed and summarized in Section 4. In section 5, we highlight several research issues and potential future directions, and finally, Section 6 concludes this paper.

2 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.

2.1 Types of Real-World Data

Usually, the availability of data is considered as the key to construct a machine learning model or data-driven real-world 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 briefly discuss these types of data.

- *Structured:* It has a well-defined 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-defined schemes, such as relational databases, structured data is 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 predefined format or organization for unstructured data, making it much more difficult 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 is 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 difference between "data" and "metadata" is that data is 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 significance for data users. A basic example of a document's metadata might be the author, file size, date generated by the document, keywords to define the document, etc.

In the area of machine learning and data science, researchers use various widely-used datasets for different purposes. These are, for example, cybersecurity datasets such as NSL-KDD [119], UNSW-NB15 [76], ISCX'12 [1], CIC-DDoS2019 [2], Bot-IoT [59], etc., smartphone datasets such as phone call logs [84] [101], SMS Log [29], mobile application usages logs [137] [117], mobile phone notification logs [73] etc., IoT data [57] [62] [16], agriculture and e-commerce data [138] [120], health data such as heart disease [92], diabetes mellitus [134] [83], COVID-19 [43] [74], etc., and many more in various application domains. The data can be in different types discussed above, which may vary from application to application in the real world. To analyze such data in a particular problem domain, and to extract the insights or useful knowledge from the data for building the real-world intelligent applications, different types of machine learning techniques can be used according to their learning capabilities, which is discussed in the following.

2.2 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 briefly 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

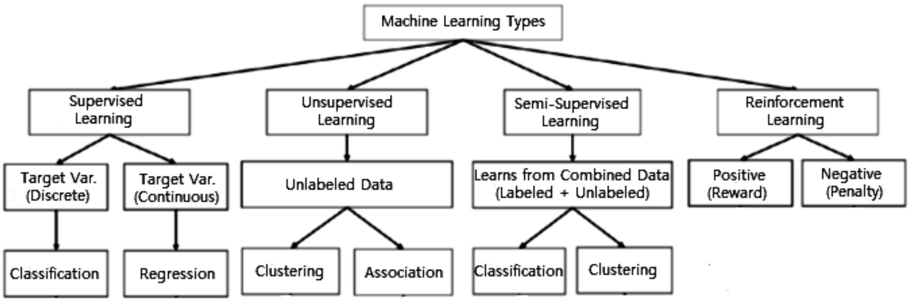


Fig. 2: Various types of machine learning techniques.

Table 1: Various types of machine learning techniques with examples.

Learning Type	Model building	Examples
Supervised	Algorithms or models learn from labeled data (Task-Driven Approach)	Classification, Regression.
Unsupervised	Algorithms or models learn from unlabeled data (Data-Driven Approach)	Clustering, Associations, Dimensionality Reduction.
Semi-supervised	Models are built using combined data (Labeled + Unlabeled)	Classification, Clustering.
Reinforcement	Models are based on reward or penalty (Environment-Driven Approach)	Classification, Control

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 identified to be accomplished from a certain set of inputs [105], i.e., a *task-driven approach*. The most common supervised tasks are "classification" that separates the data, and "regression" that fits 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 classification, 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, finding 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 [105] [41]. 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 is 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, text classification, etc.
- *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 efficiency [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 efficiency of sophisticated systems such as robotics, autonomous driving tasks, manufacturing, supply chain logistics, etc., however not preferable to use it for solving the basic or straightforward problems.

Thus, to build effective models in various application areas different types of machine learning techniques can play a significant 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.

3 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 learning-based 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.

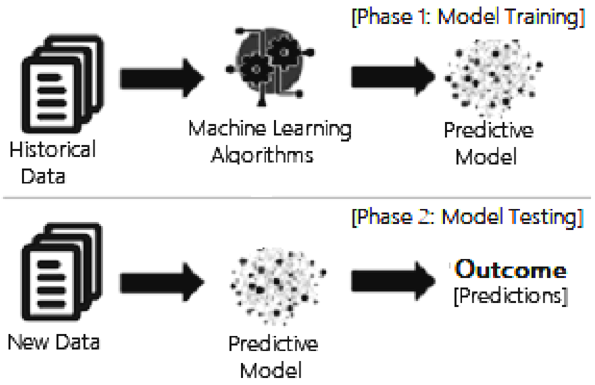


Fig. 3: A general structure of a machine learning based predictive model considering both the training and testing phase.

3.1 Classification Analysis

Classification 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 classification problem. In the following, we summarize the common classification problems.

- *Binary classification:* It refers to the classification tasks having two class labels such as "true and false" or "yes and no" [41]. In such binary classification 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 classification.
- *Multiclass classification:* Traditionally, this refers to those classification tasks having more than two class labels [41]. The multiclass classification does not have the principle of normal and abnormal outcomes, unlike binary classification tasks. Instead, within a range of specified classes, examples are classified as belonging to one. For example, it can be a multiclass classification task to classify various types of network attacks in the NSL-KDD [119] dataset, where the attack categories are classified 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 classification:* In machine learning, multi-label classification is an important consideration where an example is associated with several classes or labels. Thus, it is a generalization of multiclass classification, 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 classification. For instance, Google news can be presented under the categories of a "city name", "technology", or "latest news", etc. Multi-label classification includes advanced machine learning algorithms that support predicting various mutually non-exclusive classes or labels, unlike traditional classification tasks where class labels are mutually exclusive [82].

Many classification 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 classification, spam filtering, etc. To effectively classify the noisy instances in the data and to construct a robust prediction model, the NB classifier can be used [94]. The key benefit 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 affect due to its strong assumptions on features independence. Gaussian, Multinomial, Complement, Bernoulli, and Categorical are the common variants of NB classifier [82].

- *Linear Discriminant Analysis (LDA):* Linear Discriminant Analysis (LDA) is a linear decision boundary classifier created by fitting 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 classification 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 defined sigmoid function in Equation 1. It can overfit 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 overfitting 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 classification and regression problems, but it is more commonly used for classification.

$$g(z) = \frac{1}{1 + exp(-z)} \tag{1}$$

- *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 classifies new data points based on similarity measures (e.g., Euclidean distance function) [82]. Classification 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 classification as well as regression.
- *Support Vector Machine (SVM):* In machine learning, another common technique that can be used for classification, regression, or other tasks is a support vector machine (SVM) [56]. In high or infinite-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 classifier's generalization error. It is effective in high-dimensional spaces and can behave differently based on different mathematical functions known as the kernel. Linear, polynomial, radial basis function (RBF), sigmoid, etc., are the popular kernel functions used in SVM classifier [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) [88] is a well-known non-parametric supervised learning method. DT learning methods are used for both the classification 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 effective 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 classifies the instances. Instances are classified by checking the attribute defined 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].

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

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

- *Random Forest (RF):* A random forest classifier [19] is well known as an ensemble classification technique that is used in the field of machine learning and data science in various application areas. This method

doi:10.20944/preprints202103.0216.v1

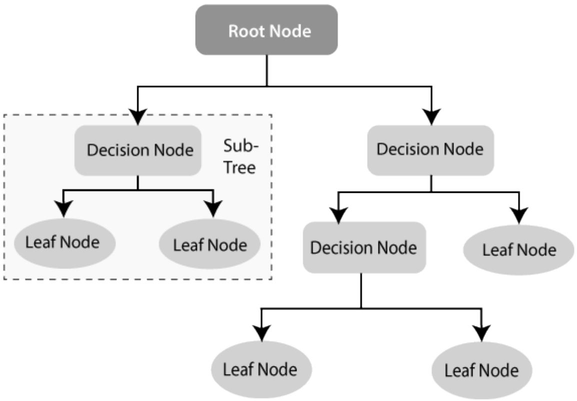


Fig. 4: An example of a decision tree structure.

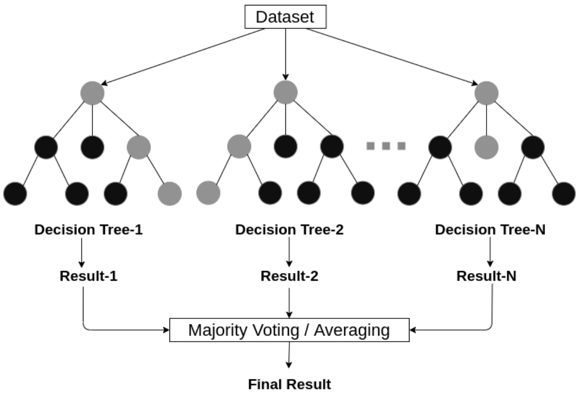


Fig. 5: An example of a random forest structure considering multiple decision trees.

uses "parallel ensembling" which fits several decision tree classifiers in parallel, as shown in Fig. 5, on different data set sub-samples and uses majority voting or averages for the outcome or final result. It thus minimizes the over-fitting 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 classification and regression problems and fits 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 classifiers by learning from their errors. This is developed by Yoav Freund et al. [35] and also known as "meta-learning". Unlike the random forest that uses parallel ensembling, Adaboost uses "sequential ensembling". It creates a powerful classifier by combining many poorly performing classifiers to obtain a good classifier of high accuracy. In that sense, AdaBoost is called an adaptive classifier by significantly improving the efficiency of the classifier, but in some instances, it can trigger overfits. AdaBoost is best used to boost the performance of decision trees, base estimator [82], on binary classification 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 final 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 & L2) [82], which reduces over-fitting, 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, α is the learning rate, and J_i is the training example cost of i^{th}, then Equ. 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 classification 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.

$$w_j \ := \ w_j - \alpha \frac{\partial J_i}{\partial w_j} \tag{4}$$

– *Rule-based Classification*: The term rule-based classification can be used to refer to any classification scheme that makes use of IF-THEN rules for class prediction. Several classification 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 classification 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 classification [127] [128]. The decision tree-based rules also provide significant accuracy in a prediction model for unseen test cases [106]. Since the rules are easily interpretable, these rule-based classifiers are often used to produce descriptive models that can describe a system including the entities and their relationships.

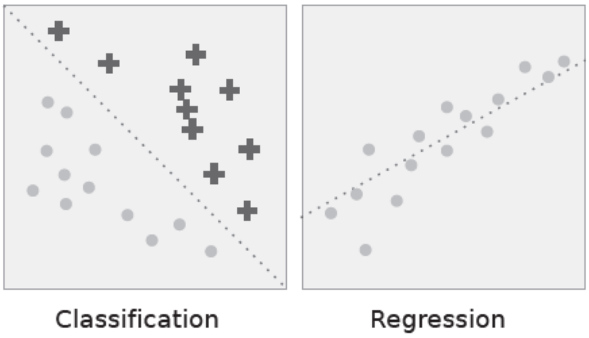


Fig. 6: Classification vs. Regression. In classification the dotted line represents a linear boundary that separates the two classes; in regression, the dotted line models the linear relationship between the two variables.

3.2 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 significant distinction between classification and regression is that classification predicts distinct class labels, while regression facilitates the prediction of a continuous quantity. Fig. 6 shows an example of how classification is different 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 fields, including financial 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 briefly in the following.

- *Simple and Multiple Linear Regression:* This is one of the most popular ML modeling techniques as well as the 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) by using the best fit straight line [41]. It is defined by the following equations -

$$y = a + bx + e \tag{5}$$

$$y = a + b_1x_1 + b_2x_2 + ... + b_nx_n + e \tag{6}$$

  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 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] defined in Equ. 6, whereas simple linear regression has only 1 independent variable, defined in Equ. 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 n^{th} in x [82]. The equation for polynomial regression is also derived from linear regression (polynomial regression of degree 1) equation, which is defined as below:

$$y = b_0 + b_1x + b_2x^2 + b_3x^3 + ... + b_nx^n + e \tag{7}$$

  Here, y is the predicted/target output, $b_0, b_1, ...b_n$ are the regression coefficients, x is an independent/ input variable. In simple words, we can say that if data is not distributed linearly, instead it is n^{th} 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-fitting 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 coefficients" ($L1$ penalty). As a result, LASSO appears to render coefficients to absolute zero. Thus, LASSO regression aims to find the subset of predictors that minimizes the prediction error for a quantitative response variable. On the other hand, Ridge regression uses $L2$ regularization [82], which is the "squared magnitude of coefficients" ($L2$ penalty). Thus, Ridge regression forces the weights to be small but never sets the coefficient value to zero, and does a non-sparse 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.

3.3 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 specific 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, behavioral analytics, etc., clustering can be used. In the following, we briefly 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, grid-based clustering is especially suitable. To obtain clusters, the principle is first 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.

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 identifies 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

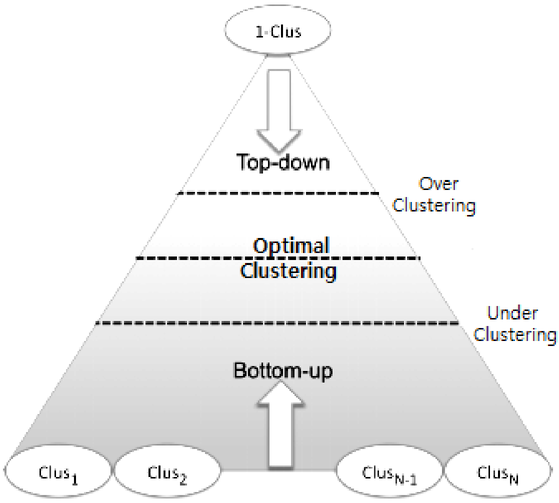


Fig. 7: A graphical interpretation of the widely-used hierarchical clustering (Bottom-up and Top-down) technique.

of cluster centers, the results can be inconsistent. Since extreme values can easily affect 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's a centroid-based algorithm that works by updating centroids candidates to be the mean of the points in a given region. To form the final set of centroids, these candidates are filtered 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 non-parametric 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 find 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 specification of the number of clusters in the data and can find arbitrarily shaped clusters. Although k-means is much faster than DBSCAN, it is efficient at finding 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 distribution-based clustering algorithm. A Gaussian mixture model is a probabilistic model in which all the data points are produced by a mixture of a finite number of Gaussian distributions with unknown parameters [82]. To find 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 first 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 flat clusters returned by k-means, which can help to make better decisions in the relevant application areas.

3.4 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 overfitting 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 briefly 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 overfitting 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 affects the effectiveness and efficiency of the target machine learning model. Chi-squared test, Analysis of variance (ANOVA) test, Pearson's correlation coefficient, 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 finding 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 coefficient between X and Y is defined as [41] -

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \qquad (8)$$

- *ANOVA:* Analysis of variance (ANOVA) is a statistical tool used to verify the mean values of two or more groups that differ significantly 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 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 χ^2 [82] statistic is an estimate of the difference between the effects of a series of events or variables observed and expected frequencies. The magnitude of the difference between the real and observed values, the degrees of freedom, and the sample size depends on χ^2. The chi-square χ^2 is commonly used for testing relationships between categorical variables. If O_i represents observed value and E_i represents expected value, then -

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad (9)$$

- *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 specified number of features. Features are ranked by the coefficients or feature significance 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 $L1$

doi:10.20944/preprints202103.0216.v1

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 coefficients 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 Classifier [82] is an example of a tree-based estimator that can be used to compute impurity-based function importances, which can then be used to discard irrelevant features.

– *Principal Component Analysis (PCA):* Principal component analysis (PCA) is a well-known unsupervised learning approach in the field 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 [81] [48]. Fig. 8 shows an example of the effect 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 effective machine learning model [98]. Technically, PCA identifies 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].

3.5 Association Rule Learning

Association rule learning is a rule-based machine learning approach to discover interesting relationships, “IF-THEN” 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, bioinformatics, etc. 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 ‘confidence’, which is introduced in [7].

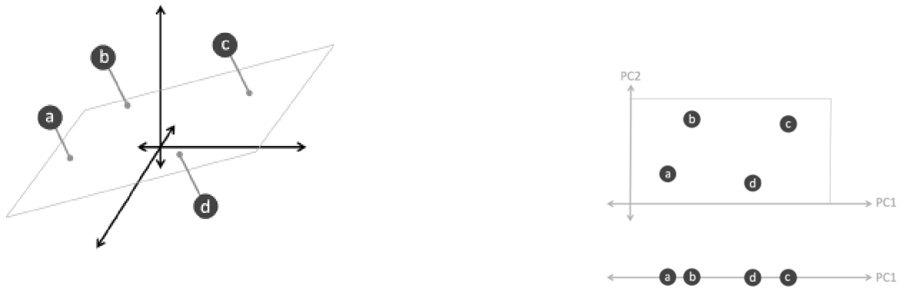
In the data mining literature, many association rule learning methods have been proposed, such as logic dependent [34], frequent pattern based [8] [49] [68], tree-based [42], etc. The most popular association rule learning algorithms are summarized below.

– *AIS and SETM:* AIS is the first 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 effort. 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 suffers from the same flaw as the AIS algorithm.

– *Apriori:* For generating association rules for a given dataset, Agrawal et al. [8] proposed the Apriori, Aprioiri-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 confidence. 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 bottom-up Lattice Traversal. ECLAT uses a depth-first search to find 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 efficient 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 difference 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 fit into memory for massive data sets,



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

Fig. 8: An example of a principal component analysis (PCA) and created principal components PC1 and PC2 in different dimension space.

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 non-redundant association rules. This algorithm first 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 decision making, particularly in a context-aware smart computing environment, where human or user preferences 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-specified constraints, such as minimum support and confidence value. The ABC-RuleMiner approach [104] discussed earlier could give significant results in terms of non-redundant rule generation and intelligent decision making for the relevant application areas in the real world.

3.6 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 defined 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 Model-free 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 model-free, is the key difference between model-free and model-based 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 doesn't 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 fields, such as game theory, control theory, operations analysis, information theory, simulation-based optimization, manufacturing, supply chain logistics, multi-agent systems, swarm intelligence, aircraft control, robot motion control, and many more.

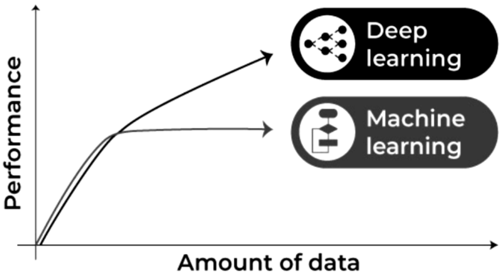


Fig. 9: Machine learning and deep learning performance in general with the amount of data.

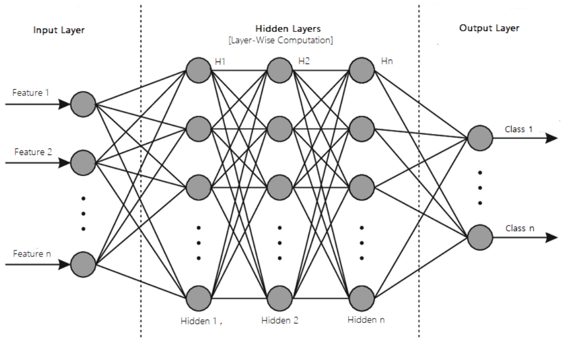


Fig. 10: A structure of an artificial neural network modeling with multiple processing layers.

3.7 Artificial Neural Network and Deep Learning

Deep learning is part of a wider family of artificial 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.

The most common deep learning algorithms are: Multi-layer Perceptron (MLP), Convolutional Neural Network (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 effective data-driven models for various purposes.

- *MLP:* The base architecture of deep learning, which is also known as the feed-forward artificial 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 classification, 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 field, such as AlexNet [60], Xception [24], In-

ception [118], Visual Geometry Group (VGG) [44], ResNet [45], etc.

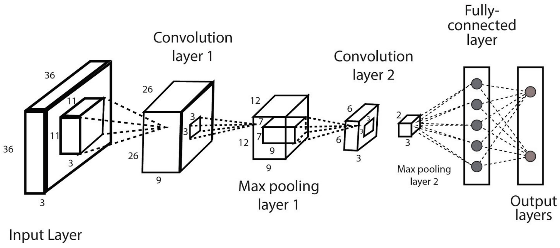


Fig. 11: An example of a convolutional neural network (CNN or ConvNet) including multiple convolution and pooling layers.

– *LSTM-RNN:* Long short-term memory (LSTM) is an artificial 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 differentiates it from other conventional networks. Thus, LSTM can be used when the data is 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, classification, regression, collaborative filtering, 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 artificial 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 classification analysis, regression, data clustering, feature selection and extraction, and dimensionality reduction, association rule learning, reinforcement learning, or deep learning techniques, can play a significant role for various purposes according to their capabilities. In Section 4, we discuss several application areas based on machine learning algorithms.

4 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 field 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, artificial 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 financial 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 traffic, find 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 classification 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 [129] [96]. Moreover, security policy rules generated by association rule learning techniques can play a significant role to build a rule-based security system [105]. Thus, we can say that various learning techniques discussed in Section 3, can enable cybersecurity professionals to be more proactive inefficiently 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 fields 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 traffic 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 traffic volume, resulting in serious issues such as delays, traffic 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 traffic is important, which is an indispensable part of a smart city. Accurate traffic prediction based on machine and deep learning modeling can help to minimize the issues [17] [31] [30]. For example, based on the travel history and trend of traveling through various routes, machine learning can assist transportation companies in predicting possible issues that may occur on specific routes and recommending their customers to take a different path. Ultimately, these learning-based data-driven models help improve traffic flow, increase the usage and efficiency of 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. [77] [33] [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 [63] [61]. 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 COVID-19 pandemic [111] [78] [10]. Overall, machine and deep learning techniques can help to fight 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 offers by analyzing browsing trends and click-through rates of specific 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 significant, 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-field 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 defined as the automated recognition of patterns and regularities in data, e.g., image analysis. Several machine learning techniques such as classification, 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. [53] [54] [5]. 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] [140] [137]. To build various personalized data-driven context-aware systems, such as smart interruption management, smart mobile recommendation, context-aware smart searching, decision making, etc. that intelligently assist end mobile phone users in a pervasive computing environment, machine learning techniques are applicable. For example, context-aware 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 classification methods can be used [106] [139]. Thus, various learning techniques discussed in Section 3 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 learning-based models can also apply to several other domains such as bioinformatics, cheminformatics, computer networks, DNA sequence classification, economics and banking, robotics, software engineering, and many more.

5 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, agriculture, etc. discussed in Section 4 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 Section 3 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, effectively modifying or enhance existing pre-processing methods, or proposing new data preparation techniques are required to effectively use the learning algorithms in the associated application domain.

To analyze the data and extract insights, there exist many machine learning algorithms, summarized in Section 3. Thus, selecting a proper learning algorithm that is suitable for the target application is challenging. The reason is that the outcome of different 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 effort, as well as the model's effectiveness and accuracy. In terms of model building, the techniques discussed in Section 3 can directly be used to solve many real-world issues in diverse domains, such as cybersecurity, smart cities, healthcare, etc. summarized in Section 4. 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 is bad to learn, such as non-representative, poor-quality, irrelevant features, or insufficient quantity for training, then the machine learning models may become useless or will produce lower accuracy. Therefore, effectively processing the data and handling the diverse learning algorithms are important, for a machine learning-based solution and eventually building intelligent applications.

6 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 briefly 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 identified create promising research opportunities in the field which must be addressed with effective 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.

Compliance with ethical standards

Conflict of interest The author declares no conflict of interest.

References

1. Canadian institute of cybersecurity, university of new brunswick, iscx dataset, url

http://www.unb.ca/cic/datasets/index.html/ (accessed on 20 october 2019).

2. Cic-ddos2019 [online]. available: https://www.unb.ca/cic/datasets/ddos-2019.html/ (accessed on 28 march 2020).
3. World health organization: Who. *http://www.who.int/*.
4. Google trends. In *https://trends.google.com/trends/*, 2019.
5. Nadia Adnan, Shahrina Md Nordin, Imran Rahman, and Amir Noor. The effects of knowledge transfer on farmers decision making toward sustainable agriculture practices. *World Journal of Science, Technology and Sustainable Development*, 2018.
6. Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 94–105, 1998.
7. Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.
8. Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proceedings of the International Joint Conference on Very Large Data Bases, Santiago Chile, pp.∼487–499.*, volume 1215, 1994.
9. David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.
10. Talha Burak Alakus and Ibrahim Turkoglu. Comparison of deep learning approaches to predict covid-19 infection. *Chaos, Solitons & Fractals*, 140:110120, 2020.
11. Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588, 1997.
12. Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.
13. Yuichiro Anzai. *Pattern recognition and machine learning*. Elsevier, 2012.
14. Sina F Ardabili, Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, and Peter M Atkinson. Covid-19 outbreak prediction with machine learning. *Algorithms*, 13(10):249, 2020.
15. Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49, 2012.
16. Fabrizio Balducci, Donato Impedovo, and Giuseppe Pirlo. Machine learning applications on agricultural datasets for smart farm enhancement. *Machines*, 6(3):38, 2018.
17. Azzedine Boukerche and Jiahao Wang. Machine learning-based traffic prediction models for intelligent transportation systems. *Computer Networks*, 181:107530, 2020.
18. Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
19. Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
20. Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees.* CRC press, 1984.
21. Longbing Cao. Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, 50(3):43, 2017.
22. Gail A Carpenter and Stephen Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, 37(1):54–115, 1987.
23. Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE, 2018.
24. François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.
25. Halil I Cobuloglu and İ Esra Büyüktahtakın. A stochastic multi-criteria decision analysis for sustainable biomass crop selection. *Expert Systems with Applications*, 42(15-16):6065–6074, 2015.
26. Amitabha Das, Wee-Keong Ng, and Yew-Kwong Woon. Rapid association rule mining. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 474–481. ACM, 2001.
27. Renato Cordeiro de Amorim. Constrained clustering with minkowski weighted k-means. In *2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 13–17. IEEE, 2012.
28. Anind K Dey. Understanding and using context. *Personal and ubiquitous computing*, 5(1):4–7, 2001.
29. Nathan Eagle and Alex Sandy Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.
30. Aniekan Essien, Ilias Petrounias, Pedro Sampaio, and Sandra Sampaio. Improving urban traffic speed prediction using data source fusion and deep learning. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–8. IEEE, 2019.
31. Aniekan Essien, Ilias Petrounias, Pedro Sampaio, and Sandra Sampaio. A deep-learning model for urban traffic flow prediction with traffic events mined from twitter. *World Wide Web*, pages 1–24, 2020.
32. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
33. Meherwar Fatima, Maruf Pasha, et al. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, 9(01):1, 2017.
34. Peter A Flach and Nicolas Lachiche. Confirmation-guided discovery of first-order rules with tertius. *Machine Learning*, 42(1-2):61–95, 2001.
35. Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156. Citeseer, 1996.
36. Hironobu Fujiyoshi, Tsubasa Hirakawa, and Takayoshi Yamashita. Deep learning-based image recognition for autonomous driving. *IATSS research*, 43(4):244–252, 2019.
37. Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 21(1):32–40, 1975.

38. Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.
39. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
40. Juan Guerrero-Ibáñez, Sherali Zeadally, and Juan Contreras-Castillo. Sensor technologies for intelligent transportation systems. *Sensors*, 18(4):1212, 2018.
41. Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, Amsterdam, Netherlands, 2011.
42. Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *ACM Sigmod Record*, volume 29, pages 1–12. ACM, 2000.
43. Stephanie A Harmon, Thomas H Sanford, Sheng Xu, Evrim B Turkbey, Holger Roth, Ziyue Xu, Dong Yang, Andriy Myronenko, Victoria Anderson, Amel Amalou, et al. Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. *Nature communications*, 11(1):1–7, 2020.
44. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.
45. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
46. Geoffrey E Hinton. A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade*, pages 599–619. Springer, 2012.
47. Robert C Holte. Very simple classification rules perform well on most commonly used datasets. *Machine learning*, 11(1):63–90, 1993.
48. Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.
49. Maurice Houtsma and Arun Swami. Set-oriented mining for association rules in relational databases. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 25–33. IEEE, 1995.
50. Mohammad Jamshidi, Ali Lalbakhsh, Jakub Talla, Zdeněk Peroutka, Farimah Hadjilooei, Pedram Lalbakhsh, Morteza Jamshidi, Luigi La Spada, Mirhamed Mirmozafari, Mojgan Dehghani, et al. Artificial intelligence and covid-19: deep learning approaches for diagnosis and treatment. *IEEE Access*, 8:109581–109595, 2020.
51. George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.
52. Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
53. Sachin S Kamble, Angappa Gunasekaran, and Shradha A Gawankar. Sustainable industry 4.0 framework: A systematic literature review identifying the current trends and future perspectives. *Process Safety and Environmental Protection*, 117:408–425, 2018.
54. Sachin S Kamble, Angappa Gunasekaran, and Shradha A Gawankar. Achieving sustainable performance in a data-driven agriculture supply chain: A review for research and applications. *International Journal of Production Economics*, 219:179–194, 2020.
55. Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.
56. S. Sathiya Keerthi, Shirish Krishnaj Shevade, Chiranjib Bhattacharyya, and Karuturi Radha Krishna Murthy. Improvements to platt's smo algorithm for svm classifier design. *Neural computation*, 13(3):637–649, 2001.
57. Vijay Khadse, Parikshit N Mahalle, and Swapnil V Biraris. An empirical comparison of supervised machine learning algorithms for internet of things data. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–6. IEEE, 2018.
58. Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.
59. Nickolaos Koroniotis, Nour Moustafa, Elena Sitnikova, and Benjamin Turnbull. Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset. *Future Generation Computer Systems*, 100:779–796, 2019.
60. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
61. Shashi Kushwaha, Shashi Bahl, Ashok Kumar Bagha, Kulwinder Singh Parmar, Mohd Javaid, Abid Haleem, and Ravi Pratap Singh. Significant applications of machine learning for covid-19 pandemic. *Journal of Industrial Integration and Management*, 5(4), 2020.
62. Prasanth Lade, Rumi Ghosh, and Soundar Srinivasan. Manufacturing analytics and industrial internet of things. *IEEE Intelligent Systems*, 32(3):74–79, 2017.
63. Samuel Lalmuanawma, Jamal Hussain, and Lalrinfela Chhakchhuak. Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: A review. *Chaos, Solitons & Fractals*, page 110059, 2020.
64. Saskia Le Cessie and Johannes C Van Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(1):191–201, 1992.
65. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
66. Huan Liu and Hiroshi Motoda. *Feature extraction, construction and selection: A data mining perspective*, volume 453. Springer Science & Business Media, 1998.
67. Gustavo López, Luis Quesada, and Luis A Guerrero. Alexa vs. siri vs. cortana vs. google assistant: a comparison of speech-based natural user interfaces. In *International Conference on Applied Human Factors and Ergonomics*, pages 241–250. Springer, 2017.
68. Bing Liu Wynne Hsu Yiming Ma. Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*, 1998.
69. James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

70. Mohammad Saeid Mahdavinejad, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi, and Amit P Sheth. Machine learning for internet of things data analysis: A survey. *Digital Communications and Networks*, 4(3):161–175, 2018.
71. André Marchand and Paul Marx. Automated product recommendations with preference-based explanations. *Journal of retailing*, 96(3):328–343, 2020.
72. Andrew McCallum. Information extraction: Distilling structured data from unstructured text. *Queue*, 3(9):48–57, 2005.
73. Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. Prefminer: mining user's preferences for intelligent mobile notification management. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, Germany, 12-16 September, pp.∼1223–1234. ACM, New York, USA.*, 2016.
74. Youssoufa Mohamadou, Aminou Halidou, and Pascalin Tiam Kapen. A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of covid-19. *Applied Intelligence*, 50(11):3913–3925, 2020.
75. Mohssen Mohammed, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashier. *Machine learning: algorithms and applications*. Crc Press, 2016.
76. Nour Moustafa and Jill Slay. Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In *2015 military communications and information systems conference (MilCIS)*, pages 1–6. IEEE, 2015.
77. Mehrbakhsh Nilashi, Othman bin Ibrahim, Hossein Ahmadi, and Leila Shahmoradi. An analytical method for diseases prediction using machine learning techniques. *Computers & Chemical Engineering*, 106:212–223, 2017.
78. Yujin Oh, Sangjoon Park, and Jong Chul Ye. Deep learning covid-19 features on cxr using limited training data sets. *IEEE Transactions on Medical Imaging*, 39(8):2688–2700, 2020.
79. Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
80. Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.
81. Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.
82. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
83. Sajida Perveen, Muhammad Shahbaz, Karim Keshavjee, and Aziz Guergachi. Metabolic syndrome and development of diabetes mellitus: Predictive modeling based on machine learning techniques. *IEEE Access*, 7:1365–1375, 2018.
84. Santi Phithakkitnukoon, Ram Dantu, Rob Claxton, and Nathan Eagle. Behavior-based adaptive call predictor. *ACM Transactions on Autonomous and Adaptive Systems*, 6(3):21:1–21:28), 2011.
85. Athanasios S Polydoros and Lazaros Nalpantidis. Survey of model-based reinforcement learning: Applications on robotics. *Journal of Intelligent & Robotic Systems*, 86(2):153–173, 2017.
86. Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
87. J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.
88. J. Ross Quinlan. C4.5: Programs for machine learning. *Machine Learning*, 1993.
89. Carl Rasmussen. The infinite gaussian mixture model. *Advances in neural information processing systems*, 12:554–560, 1999.
90. Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46, 2015.
91. Lior Rokach. A survey of clustering algorithms. In *Data Mining and Knowledge Discovery Handbook*, pages 269–298. Springer, 2010.
92. Saima Safdar, Saad Zafar, Nadeem Zafar, and Naurin Farooq Khan. Machine learning based decision support systems (dss) for heart disease diagnosis: a review. *Artificial Intelligence Review*, 50(4):597–623, 2018.
93. Iqbal H Sarker. Context-aware rule learning from smartphone data: survey, challenges and future directions. *Journal of Big Data*, 6(1):1–25, 2019.
94. Iqbal H Sarker. A machine learning based robust prediction model for real-life mobile phone data. *Internet of Things*, 5:180–193, 2019.
95. Iqbal H Sarker. Ai-driven cybersecurity: An overview, security intelligence modeling and research directions. *SN Computer Science*, 2021.
96. Iqbal H Sarker. Deep cybersecurity: A comprehensive overview from neural network and deep learning perspective. *SN Computer Science*, 2021.
97. Iqbal H Sarker, Yoosef B Abushark, Fawaz Alsolami, and Asif Irshad Khan. Intrudtree: A machine learning based cyber security intrusion detection model. *Symmetry*, 12(5):754, 2020.
98. Iqbal H Sarker, Yoosef B Abushark, and Asif Irshad Khan. Contextpca: Predicting context-aware smartphone apps usage based on machine learning techniques. *Symmetry*, 12(4):499, 2020.
99. Iqbal H Sarker, Hamed Alqahtani, Fawaz Alsolami, Asif Irshad Khan, Yoosef B Abushark, and Mohammad Khubeb Siddiqui. Context pre-modeling: an empirical analysis for classification based user-centric context-aware predictive modeling. *Journal of Big Data*, 7(1):1–23, 2020.
100. Iqbal H Sarker, Alan Colman, Jun Han, Asif Irshad Khan, Yoosef B Abushark, and Khaled Salah. Behavdt: A behavioral decision tree learning to build user-centric context-aware predictive model. *Mobile Networks and Applications*, pages 1–11, 2019.
101. Iqbal H Sarker, Alan Colman, Muhammad Ashad Kabir, and Jun Han. Phone call log as a context source to modeling individual user behavior. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (Ubicomp): Adjunct, Germany*, pages 630–634. ACM, 2016.
102. Iqbal H Sarker, Alan Colman, Muhammad Ashad Kabir, and Jun Han. Individualized time-series segmentation for mining mobile phone user behavior. *The Computer Journal, Oxford University, UK*, 61(3):349–368, 2018.
103. Iqbal H Sarker, Mohammed Moshiul Hoque, Md Kafil Uddin, and Tawfeeq Alsanoosy. Mobile data science and

intelligent apps: Concepts, ai-based modeling and research directions. *Mobile Networks and Applications*, pages 1–19, 2020.

104. Iqbal H Sarker and ASM Kayes. Abc-ruleminer: User behavioral rule-based machine learning method for context-aware intelligent services. *Journal of Network and Computer Applications*, page 102762, 2020.

105. Iqbal H Sarker, ASM Kayes, Shahriar Badsha, Hamed Alqahtani, Paul Watters, and Alex Ng. Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*, 7(1):1–29, 2020.

106. Iqbal H Sarker, ASM Kayes, and Paul Watters. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *Journal of Big Data*, 6(1):1–28, 2019.

107. Iqbal H Sarker and Khaled Salah. Appspred: predicting context-aware smartphone apps using random forest learning. *Internet of Things*, 8:100106, 2019.

108. Tobias Scheffer. Finding association rules that trade support optimally against confidence. *Intelligent Data Analysis*, 9(4):381–395, 2005.

109. Rohit Sharma, Sachin S Kamble, Angappa Gunasekaran, Vikas Kumar, and Anil Kumar. A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. *Computers & Operations Research*, 119:104926, 2020.

110. Shengli Sheng and Charles X Ling. Hybrid cost-sensitive decision tree, knowledge discovery in databases. In *PKDD 2005, Proceedings of 9th European Conference on Principles and Practice of Knowledge Discovery in Databases. Lecture Notes in Computer Science*, volume 3721, 2005.

111. Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Deep learning applications for covid-19. *Journal of Big Data*, 8(1):1–54, 2021.

112. Gökhan Silahtaroğlu and Nevin Yılmaztürk. Data analysis in health and big data: A machine learning medical diagnosis model based on patients' complaints. *Communications in Statistics-Theory and Methods*, pages 1–10, 2019.

113. David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

114. Beata Ślusarczyk. Industry 4.0: Are we ready? *Polish Journal of Management Studies*, 17, 2018.

115. Peter HA Sneath. The application of computers to taxonomy. *Journal of General Microbiology*, 17(1), 1957.

116. Thorvald Sorensen. method of establishing groups of equal amplitude in plant sociology based on similarity of species. *Biol. Skr.*, 5, 1948.

117. Vijay Srinivasan, Saeed Moghaddam, and Abhishek Mukherji. Mobileminer: Mining your frequent patterns on your phone. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing, Seattle, WA, USA, 13-17 September, pp.∼389–400. ACM, New York, USA.*, 2014.

118. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

119. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A Ghorbani. A detailed analysis of the kdd cup 99 data set. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pages 1–6. IEEE, 2009.

120. Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke. Challenges and research opportunities in ecommerce search and recommendations. In *ACM SIGIR Forum*, volume 54, pages 1–23. ACM New York, NY, USA, 2021.

121. Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. Constrained k-means clustering with background knowledge. In *Icml*, volume 1, pages 577–584, 2001.

122. Wei Wang, Jiong Yang, Richard Muntz, et al. Sting: A statistical information grid approach to spatial data mining. In *VLDB*, volume 97, pages 186–195, 1997.

123. Peng Wei, Yufeng Li, Zhen Zhang, Tao Hu, Ziyong Li, and Diyang Liu. An optimization method for intrusion detection classification model based on deep belief network. *IEEE Access*, 7:87593–87605, 2019.

124. Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):9, 2016.

125. Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

126. Ian H Witten, Eibe Frank, Leonard E Trigg, Mark A Hall, Geoffrey Holmes, and Sally Jo Cunningham. Weka: Practical machine learning tools and techniques with java implementations. 1999.

127. Chia-Chi Wu, Yen-Liang Chen, Yi-Hung Liu, and Xiang-Yu Yang. Decision tree induction with a constrained number of leaf nodes. *Applied Intelligence*, 45(3):673–685, 2016.

128. Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

129. Yang Xin, Lingshuang Kong, Zhi Liu, Yuling Chen, Yanmiao Li, Hongliang Zhu, Mingcheng Gao, Haixia Hou, and Chunhua Wang. Machine learning and deep learning methods for cybersecurity. *IEEE Access*, 6:35365–35381, 2018.

130. Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.

131. Mohammed Javeed Zaki. Scalable algorithms for association mining. *IEEE transactions on knowledge and data engineering*, 12(3):372–390, 2000.

132. Andrea Zanella, Nicola Bui, Angelo Castellani, Lorenzo Vangelista, and Michele Zorzi. Internet of things for smart cities. *IEEE Internet of Things journal*, 1(1):22–32, 2014.

133. Qiankun Zhao and Sourav S Bhowmick. Association rule mining: A survey. *Nanyang Technological University, Singapore*, 2003.

134. Tao Zheng, Wei Xie, Liling Xu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, and You Chen. A machine learning-based framework to identify type 2 diabetes through electronic health records. *International journal of medical informatics*, 97:120–127, 2017.

135. Yanxu Zheng, Sutharshan Rajasegarar, and Christopher Leckie. Parking availability prediction for sensor-enabled car parks in smart cities. In *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2015 IEEE Tenth International Conference on*, pages 1–6. IEEE, 2015.

136. Hengshu Zhu, Huanhuan Cao, Enhong Chen, Hui Xiong, and Jilei Tian. Exploiting enriched contextual information for mobile app classification. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1617–1621. ACM, 2012.
137. Hengshu Zhu, Enhong Chen, Hui Xiong, Kuifei Yu, Huanhuan Cao, and Jilei Tian. Mining mobile user preferences for personalized context-aware recommendation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):58, 2014.
138. He Zikang, Yang Yong, Yang Guofeng, and Zhang Xinyu. Sentiment analysis of agricultural product ecommerce review data based on deep learning. In *2020 International Conference on Internet of Things and Intelligent Applications (ITIA)*, pages 1–7. IEEE, 2020.
139. Sina Zulkernain, Praveen Madiraju, and Sheikh Iqbal Ahamed. A context aware interruption management system for mobile devices. In *Mobile Wireless Middleware, Operating Systems, and Applications*, pages 221–234. Springer, 2010.
140. Sina Zulkernain, Praveen Madiraju, Sheikh Iqbal Ahamed, and Karl Stamm. A mobile intelligent interruption management system. *J. UCS*, 16(15):2060–2080, 2010.