IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204191 |

Machine Learning Methods for Intelligent Data Analysis and Real-World Applications

D. Diana Juliel, K. Aarthi, M. Jamunarani, B. Padmapriya

Assistant Professor, Kalaignar Karunanidhi Institute of Technology, Coimbatore, India

P.G. Student, Kalaignar Karunanidhi Institute of Technology, Coimbatore. India

ABSTRACT: The digital world is life with data in this Fourth Industrial Revolution era, including data from the Internet of Things (IoT), cybersecurity, mobile, business, social media, health, and other sources. Understanding artificial intelligence (AI), and specifically machine learning (ML), is essential to intelligently analysing these data and creating the associated intelligent and automated applications. There are many different kinds of machine learning algorithms in the field, including supervised, unsupervised, semi-supervised, and reinforcement learning. Furthermore, deep learning, a subset of a larger class of machine learning techniques, is capable of large-scale, intelligent data analysis. We provide a thorough analysis of various machine learning techniques in this paper, which may be used to improve an application's intelligence and functionality. Therefore, the fundamental contribution of this study is to provide an explanation of the principles underlying various machine learning approaches and how they may be applied in a variety of real-world application areas, including e-commerce, cybersecurity systems, smart cities, healthcare, and agriculture, among many others. Based on our investigation, we also emphasize the difficulties and future directions for research. The overall goal of this article is to provide a technical point of reference for experts in the industry and academia, as well as for decision-makers in a variety of real-world scenarios and application domains.

KEYWORDS: Machinelearning·Deeplearning·Artificialintelligence·Datascience·Data-drivendecision-making· Predictive analytics · Intelligent applications

I. 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 . 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 citydata,businessdata,smartphonedata,socialmediadata, healthdata,COVID-19data,andmanymore.Thedata can be structured, semi-structured, or unstructured, discussed briefly in Sect. “Types of Real-World Data and MachineLearning Techniques”, which is increasing day-by-day. Extracting insights from these data can be used to build variousintelligentapplicationsintherelevantdomains. Generally speaking, the type and characteristics of the data as well as the functionality of the learning algorithms determine how successful and efficient a machine learning solution is. To efficiently create data-driven systems, machine learning algorithms can be used in classification analysis, regression, data clustering, feature engineering and dimensionality reduction, association rule learning, or reinforcement learning.For instance, to build a data-driven automated and intelligent cybersecurity system, the relevant cybersecurity data can be used to build personalized context-aware smart mobile applications, the relevant mobile data can be used , 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. AI technology is widely used throughout industry, government, and science. Some high-profile applications include advanced web search engines (e.g., Google Search); recommendation systems (used by YouTube, Amazon, and Netflix); creative tools (e.g., ChatGPT and AI art); and superhuman play and analysis in strategy games (e.g., chess and Go). ML usually provides systems with the ability to learn and enhance from experience automatically without being specifically pro- gammed and is generally referred to as the most popular latest technologies in the fourth industrial revolution (4IR orIndustry4.0).“Industry4.0”istypically theongoingautomationofconventionalmanufacturingand industrial practices, including exploratory data processing, using new smart technologies such as machine learning automation. Thus, to intelligently analyze these data and to developthecorrespondingreal-worldapplications, machine learning algorithms isthekey.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204191 |

Thelearningalgorithmscan be categorized into four major types, such as supervised, unsupervised, semi-supervised, and reinforcement learning in the area discussed briefly in Sect. “Types ofReal-World Data and MachineLearning Techniques”. The popularity of these approaches to learning is increasing day-by-day, which is shown in Fig. 1, based on data collected from Google Trends over the last five y ears. 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, classificationanalysis,regression,dataclustering, feature engineering and dimensionality reduction, association rule learning, or reinforcement learning techniques exist to effectively build data-driven systems. Besides, deep learning originatedfromtheartificial neural network that can be used to intelligently analyze data, which is known as part of a wider family of machine learning approaches. Thus, selecting a proper learningalgorithmthatissuitableforthetargetapplication 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.Thus,itisimportanttounderstandtheprinciples 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 context-aware systems, sustainable agriculture, and many more that are explained briefly in Sect. “Applications of Machine Learning”.

Based on the importance and potentiality of “Machine Learning” to analyze the data mentioned above, in this paper, we provide a comprehensive view on various types of machine learning algorithms that can be applied to enhance the intelligence and the capabilities of an application. Thus, the key contribution of this study is explaining the principles and potentiality of different machine learning techniques, and their applicability in various real-worldapplicationareasmentionedearlier.Thepurpose of this paper is, therefore, to provide a basic guide forthose 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.

Thekeycontributionsofthispaperarelistedas 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.

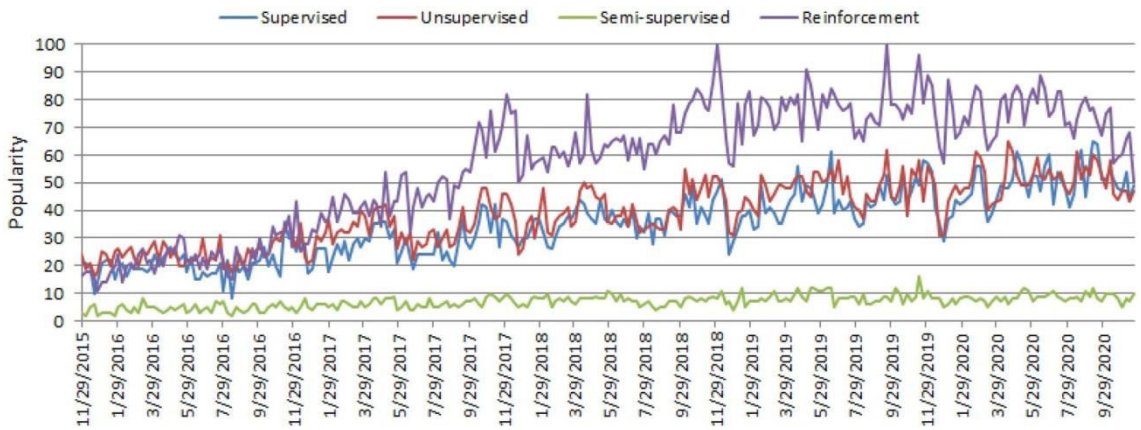


Fig. 1 The worldwide popularity score of various types of ML algorithms (supervised, unsupervised, semi-supervised, and reinforcement) in a range of 0 (min) to 100 (max) over time where x-axis represents the timestamp information and y-axis represents the corresponding score discuss the applicability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

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204191 |

in various real-world application domains.Tohighlightandsummarizethe potential research directions within the scope of our study for intelligent data analysis and services.

II. 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.

III. 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 . Data can be of various forms, such as structured, semi-structured, or unstructured . 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 asrelational databases, structured data are typically stored, i.e., in a tabular format. For instance, names, dates, addresses, credit card numbers, stock information, geolocation, etc. are examples of structured data.Unstructured: On the other hand, there is no pre-defined 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 are not stored in a relational database like the structured data mentioned above, but it does have certain organizational proper- ties 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 are simply the material that can classify, measure, or even document something relative to an organization's data properties. On the other hand, metadata describes the relevant data information, giving it more 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 docu ment, 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, UNSW-NB15, ISCX'12, CIC-DDoS2019, Bot-IoT , etc., smartphone datasets such as phone call logs , SMS Log , mobile application usages logs, mobile phone notification logs etc., IoT data, agriculture and e-commerce data, health data such as heart disease , diabetes mellitus ,COVID-1etc., 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.

Machine Learning algorithms are mainly divided into four categories: Supervised learning, Unsupervised learning, Semi-supervised learning, and Reinforcement learning, as shown in Fig. 2. In the following, we briefly discuss each type of learning technique with the cope of their applicability to solve real-world problems.

Supervised: Supervised learning is typically the task of machine learning to learn a function that maps an input to an output based on sample input-output pairs. 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, 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 super- vised learning. Unsupervised: Unsupervised learning analyses unlabelled datasets without the need for human interference, i.e., a data-driven process. 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,

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204191 |

dimensionality reduction, finding association rules, anomaly detection, etc.

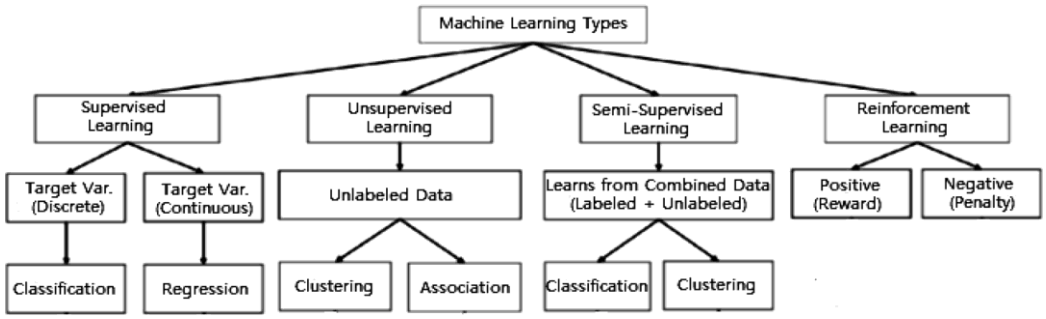


Fig. 2 Various types of machine learning techniques

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. Thus, it falls between learning "without supervision" and learning "with supervision". In the real world, labeled data could be rare in several contexts, and unlabeled data are numerous, where semi-supervised learning is useful]. 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,labelling data and text classification. 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, 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. 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 and supply chain logistics, however, not preferable to use it for solving the basic or straight forward 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.

earning 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, associa- tions, 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

Table 1 Various types of machine learning techniques with examples

IV. 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.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204191 |

V. 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 fol- lowing, 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.

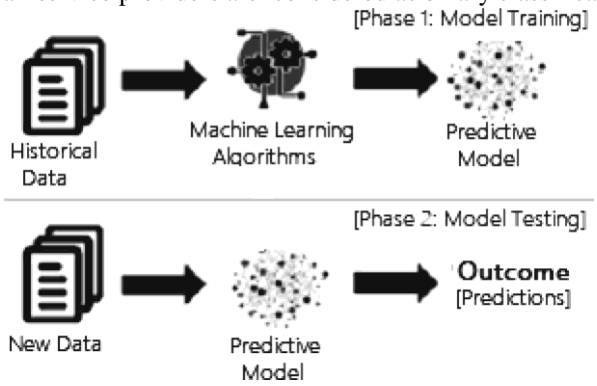


Fig. 3 A general structure of a machine learning based predictive model considering both the training and testing phase

- Multiclass classification: Traditionally, this refers to those classification tasks having more than two classlabels. 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 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.

Many classification algorithms have been proposed in the machine learning and data science literature. 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 . 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. 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 . 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.

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. This method is also known as a

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204191 |

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 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) . Logistic regression typically uses a logistic function to estimate the probabilities, which is also referred to as the mathematically defined sigmoid function in Eq. 1. It can overfit high-dimensional datasets and works well when the dataset can be separated linearly. 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.

Decision tree (DT): Decision tree (DT) is a well- known non-parametric supervised learning method. DT learning methods are used for both the classification and regression tasks and CART are well known for DT algorithms. over, recently proposed BehavDT, and Intrud Tree 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.

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

$$\text{Entropy} : H(x) = -\sum_{i=1} p(xi) \log 2\, p(xi) \qquad (2)$$

K-nearest neighbors (KNN): K-Nearest Neighbors

$$\text{Gini}(E) = 1 - p^2$$

(KNN) is an "instance-based learning" or non-gen- eralizing 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). 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) . 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., c

Random forest (RF): A random forest classifier 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 uses

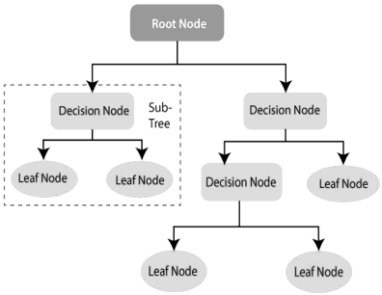


Fig. 4 An example of a decision tree structure

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204191 |

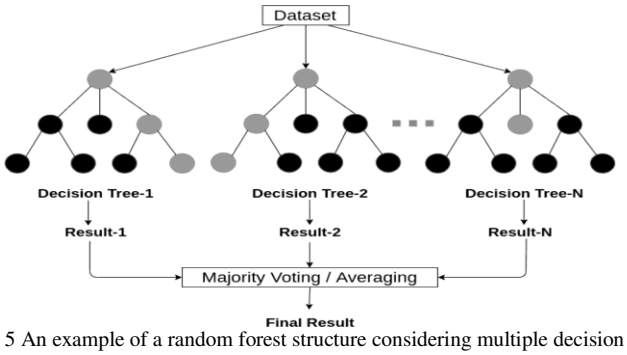


Fig. 5 An example of a random forest structure considering multiple decision trees

"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 aver- ages for the outcome or final result. It thus minimizes the over-fitting problem and increases the prediction accuracy and control . Therefore, the RF learning model with multiple decision trees is typically more accurate than a single decision tree based model. To build a series of decision trees with controlled variation, it combines bootstrap aggregation (bagging) and random feature selection. It is adaptable to both classification and regression problems and fits well for both categorical and continuous values.(XGBoost) is a form of gradient boosting that takes more detailed approximations into account when determining the best model . It computes second-order gradients of the loss function to minimize loss and advanced regularization (L1 and L2), 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) 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, par- ticularly 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, is the learning rate, and J_i is the training example cost of ith, then Eq. (4) represents the stochastic gradient descentweight 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. However, SGD is sensitive to feature scaling and needs a range of hyperparameters, such as the regularization parameter and the number of iterations.

$$w := w - a^{6Ji}.$$

Adaptive Boosting (AdaBoost): Adaptive Boosting (AdaBoost) is an ensemble learning process that employs an iterative approach to improve poor classi- fiers by learning from their errors. This is developed by Yoav Freund et al. 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, on binary classification problems, however, is sensitive to noisy data and outliers.Extreme gradient boosting (XGBoost): Gradient Boost- ing, like Random Forests 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 use gradient descent to optimize.

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, One-R, decision trees, DTNB, Ripple Down Rule learner (RIDOR), Repeated Incremental Pruning to Produce Error Reduction (RIPPER) 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. The decision tree-based rules also provide significant accuracy in a prediction model for unseen test cases. 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.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204191 |

VI. 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. Figure 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. 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 defined in Eq. 6, whereas simple linear regression has only 1 independent variable, defined in Eq. 5.

- Polynomial regression: Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is not linear, but is the polynomial degree of nth in x. The equation for polynomial regression is also derived from linear regression (polynomial regression of degree 1) equation, which is defined as below:
-

Some of regression algorithms are linear, polynomial,

$$y = b0 + b1x + b2x2 + b3x3 + \cdots + bnxn + e. \quad (7)$$

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 a well-known regression technique. In this technique, the dependent variable is continuous, the independent variable(s) can be continuous or discrete, and the form of the regression line is linear. Linear regression creates a relationship between the dependent variable (Y) and one or more independent variables (X) (also known as regression line) using the best fit straight line . It is defined by the following equations:Here, y is the predicted/target output, b0, b1, ...bn are the regression coefficients, x is an independent/ input variable. In simple words, we can say that if data are not dis- tributed linearly, instead it is nth degree of polynomial then we use polynomial regression to get desired output.LASSO and ridge regression: LASSO and Ridge regression are well known as powerful techniques which are typically used for building learning models in presence of a large number of features, due to their capability to preventing over-fitting and reducing the complexity of the model. The LASSO (least absolute shrinkage and selection operator) regression model uses L1 regularization technique that uses shrinkage, which penalizes

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

"absolute value of magnitude of coefficients" (L1 pen- alty). As a result, LASSO appears to render coefficients

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

to absolute zero. Thus, LASSO regression aims to find the subset of predictors that minimizes the prediction 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

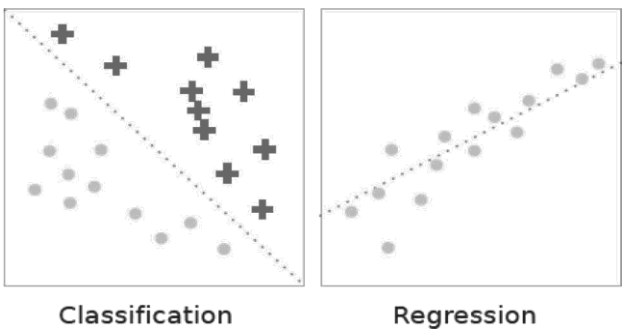


Fig. 6 Classification vs. regression. In classification the dotted line represents a linear boundary that separates the two classes; in regression, the dotted line models the linear relationship between the two variables for a quantitative response variable.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

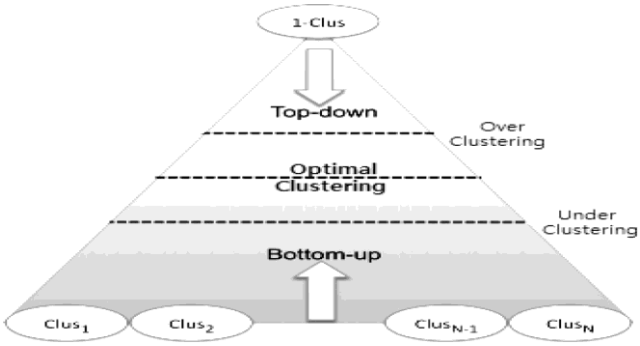
| DOI: 10.15680/IJIRCCE.2024.1204191 |

VII. 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 thanobjects in other groups. It is often used as a data analysis technique to discover interesting trends or patterns in data, e.g., groups of consumers based on their behavior. In a broad range of application areas, such as cybersecurity, e-commerce, mobile data processing, health analytics, user modeling and behavioral analytics, clustering can be used. In the following, we briefly discuss and summarize various types of clustering methods.

Partitioning methods: Based on the features andsimilarities in the data, this clustering approach categorizes the data into multiple groups or clusters. The datascientists 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, K-Mediods, CLARA 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 , OPTICS 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 hierarchicalclustering 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. 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 CLIQUE , 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 statisti- cal learning, and the other based on a method of neural network learning]. For instance, GMM is an example of a statistical learning method, and SOM 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



Among the association rule learning techniques discussed above, Apriori is the most widely used algorithm for discovering association rules from a given dataset. 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 sup- port and confidence value. The ABC-Rule Miner approach 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.

VIII. 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

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204191 |

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 MarkovDecision Process (MDP), 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 infer- ring optimal behavior from a model of the environment by performing actions and observing the results, which include the next state and the immediate reward. Alp- haZero, AlphaGo 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. 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. 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. It does not need a model of the environment(hence the term "model-free"), and it can deal with stochastic transitions and rewards without the need for adaptations. The 'Q' in Q-learning usually stands for quality, as the algorithm calculates the maximum expected rewards for a given behavior in a given state.Deep Q-learning: The basic working step in Deep Q-Learning 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 reason- ably 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 approximation.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.

IX. 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. The main advantage of deep learning over traditional machine learning methods is its better performance in several cases, particularly learning from large datasets. Figure 9 shows a general performance of deep learning over machine learning considering the increasing amount of data. However, it may vary depending on the data characteristics and experimental set up.The most common deep learning algorithms are: Multi- layer Perceptron (MLP), Convolutional Neural Network

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

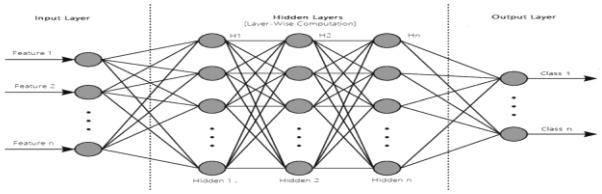


Fig. 10 A structure of an artificial neural network modeling with multiple processing layers

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204191 |

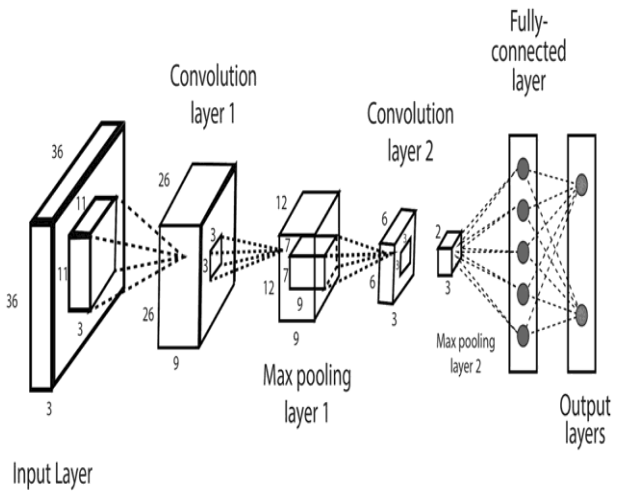


Fig. 11 An example of a convolutional neural network (CNN or Con- vNet) including multiple convolution and pooling layers

(CNN, or ConvNet), Long Short-Term Memory Recurrent Neural Network (LSTM-RNN). 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) . 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, 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) enhances the design of the standard ANN, consisting of convolutional layers, pooling layers, as well as fully connected layers, as shown in Fig. 11.

X. 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. The basis of predictive analytics is capturing and exploiting relationships between explanatory variables and predicted variables from previous events to predict the unknown outcome. 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. 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 isone of the most essential areas of Industry 4.0., which is typically the practice of protecting networks, systems, hardware, and data from digital attacks. 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 neigh- borhoods 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.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204191 |

XI. 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. Thus, the ultimate success of a machine learning-based solution and corresponding applications mainly depends on both the data and the learning algorithms. If the data are bad to learn, such as non-representative, poor-quality, irrelevant features, or 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.

XII. CONCLUSION

A thorough review of machine learning methods for intelligent data analysis and applications is presented in this work. In line with our objective, we have briefly discussed the ways in which diverse machine learning techniques can be applied to solve a range of real-world problems. Both the data and the effectiveness of the learning algorithms are necessary for a machine learning model to be successful. Before assisting with intelligent decision-making, the system must first train its advanced learning algorithms using real-world data and knowledge relevant to the intended application. In order to demonstrate how machine learning techniques can be applied to a variety of real-world problems, we also covered a number of well-liked application areas. We have now reviewed and discussed the difficulties encountered as well as the possibilities.

REFERENCES

1. Canadian institute of cybersecurity, university of new brun- swick, iscx dataset, 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. Worldhealthorganization:WHO.http://www.who.int/.
4. Googletrends.Inhttps://trends.google.com/trends/,2019.
5. Adnan N, Nordin Shahrina Md, Rahman I, Noor A. The effects of knowledge transfer on farmers decision making toward sus- tainableagriculturepractices.WorldJSciTechnolSustainDev. 2018.
6. Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications.In:Proceedingsofthe1998ACMSIGMODinter- national conference on Management of data. 1998; 94–105
7. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. In: ACM SIGMOD Record. ACM. 1993;22: 207–216
8. AgrawalR,GehrkeJ,GunopulosD,RaghavanP.Fastalgorithms forminingassociationrules.In:ProceedingsoftheInternational Joint Conference on Very Large Data Bases, Santiago Chile. 1994; 1215: 487–499.
9. Aha DW, Kibler D, Albert M. Instance-based learning algo- rithms. Mach Learn. 1991;6(1):37–66.
10. AlakusTB,TurkogluI.Comparisonofdeeplearningapproaches to predict covid-19 infection. Chaos Solit Fract. 2020;140:
11. AmitY,GemanD.Shapequantizationandrecognitionwithran- domized trees. Neural Comput. 1997;9(7):1545–88.
12. AnkerstM,BreunigMM,KriegelH-P,SanderJ.Optics:ordering pointstoidentifytheclusteringstructure.ACMSigmodRecord. 1999;28(2):49–60.
13. Anzai Y. Pattern recognition and machine learning. Elsevier; 2012.
14. ArdabiliSF,MosaviA,GhamisiP,FerdinandF,Varkonyi-Koczy AR,ReuterU,RabczukT,AtkinsonPM.Covid-19outbreakpre- diction with machine learning. Algorithms. 2020;13(10):249.
15. Baldi P. Autoencoders, unsupervised learning, and deep architec- tures. In: Proceedings of ICML workshop on unsupervised and transfer learning, 2012; 37–49 .
16. alducciF,ImpedovoD,PirloG.Machinelearningapplications onagriculturaldatasetsforsmartfarmenhancement.Machines. 2018;6(3):38.
17. Boukerche A, Wang J. Machine learning-based traffic predic- tion models for intelligent transportation systems. Comput Netw. 2020;181
18. Breiman L. Bagging predictors. Mach Learn. 1996;24(2):123–40.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204191 |

19. BreimanL.Randomforests.MachLearn.2001;45(1):5–32.
20. BreimanL,FriedmanJ,StoneCJ,OlshenRA.Classificationand regression trees. CRC Press; 1984.
21. CaoL.Datascience:acomprehensiveoverview.ACMComput Surv (CSUR). 2017;50(3):43.
22. Carpenter GA, Grossberg S. A massively parallel architecture for a self-organizing neural pattern recognition machine. Comput Vis Graph Image Process. 1987;37(1):54–115.
23. ChiuC-C,SainathTN,WuY,PrabhavalkarR,NguyenP,Chen Z,KannanA,WeissRJ,RaoK,GoninaE,etal.State-of-the-art speechrecognitionwithsequence-to-sequencemodels.In:2018 IEEEInternationalConferenceonAcoustics,SpeechandSignal Processing (ICASSP), 2018 pages 4774–4778. IEEE .
24. CholletF.Xception:deeplearningwithdepthwiseseparablecon- volutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1251–1258, 2017.
25. CobulogluH,BüyüktahtakınIE.Astochasticmulti-criteriadeci-sionanalysisforsustainablebiomasscropselection.ExpertSyst Appl. 2015;42(15–16):6065–74.
26. DasA,NgW-K,WoonY-K.Rapidassociationrulemining.In: ProceedingsofthetenthinternationalconferenceonInformation and knowledge management, pages 474–481. ACM, 2001.
27. deAmorimRC.Constrainedclusteringwithminkowskiweighted k-means.In:2012IEEE13thInternationalSymposiumonCom- putational Intelligence and Informatics (CINTI), pages 13–17. IEEE, 2012.
28. Dey AK. Understanding and using context. PersonUbiquit Com- put. 2001;5(1):4–7.
29. Eagle N, Pentland AS. Reality mining: sensing complex social systems. Person Ubiquit Comput. 2006;10(4):255–68.
30. EssienA,PetrouniasI,SampaioP,SampaioS.Improvingurban trafficspeedpredictionusingdatasourcefusionanddeeplearn- ing. In: 2019 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE. 2019: 1–8. .
31. EssienA,PetrouniasI,SampaioP,SampaioS.Adeep-learning modelforurbantrafficflowpredictionwithtrafficeventsmined from twitter. In: World Wide Web, 2020: 1–24 .
32. EsterM,KriegelH-P,SanderJ,XiaoweiX,etal.Adensity-based algorithmfordiscoveringclustersinlargespatialdatabaseswith noise. Kdd. 1996;96:226–31.
33. FatimaM,PashaM,etal.Surveyofmachinelearningalgorithms for disease diagnostic. J Intell Learn Syst Appl. 2017;9(01):1.
34. Flach PA, Lachiche N. Confirmation-guided discovery of first- order rules with tertius. Mach Learn. 2001;42(1–2):61–95.
35. FreundY, Schapire RE, et al. Experiments with a new boosting algorithm. In: Icml, Citeseer. 1996; 96: 148–156

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details