Machine Learning Technique and Applications – An Classification Analysis

[1]J Xin Ge and [2]Yuan Xue
[1]School of Chemistry and Chemical Engineering, Nanjing University, Jiangsu, China, 210093.
[1]ktxin@nju.edu.cn, [2]yuanxuenu93@gmail.com

ArticleInfo
Journal of Machine and Computing (http://anapub.co.ke/journals/jmc/jmc.html)
Doi : https://doi.org/10.53759/7669/jmc202101022
Received 30 March 2021; Revised form 30 June 2021; Accepted 20 August 2021; Available online 05 October 2021.

Abstract – The digitally-enhanced environment is susceptible to massive data, such as information security data, internet technology data, cellular internet, patient records, media data, corporate data, and so on, in the current era of Industry 4.0. Understanding of Machine Learning (ML) is essential for intelligently evaluating these sets of data and developing related "intelligent" and "automated" solutions. Different forms of ML algorithms e.g. reinforcement learning, semi-supervised, unsupervised and supervised learning exist in this segment. In addition, deep learning, which is a wider segment of ML techniques, can smartly evaluate datasets on a massive scale. In this research, a comprehensive analysis of ML techniques and classification analysis algorithms that are applicable to develop capabilities and intelligence of applications are analyzed. Therefore, this research's contribution is illustrating the key principles of various ML techniques and their application in different real-life application realms e.g. e-commerce, healthcare, agriculture, smart cities, cyber-security systems etc. Lastly, this paper presents a discussion of the challenges and future research based on this research.

Keywords – Machine Learning (ML), Internet of Things (IoT), Classification Analysis, Digital Technology.

I. INTRODUCTION

Our lives are digitally recorded, and everything around us is linked to a data source. In today's modern world, data from the Internet of Things (IoT), digital security, urban planning, business, mobile phone data, media data, medical records, COVID-19 data, and several forms of datasets are accessible. The data may be organized, semi-structured, or unstructured, and it is increasing at an exponential pace. This dataset's information may be used to build a number of smart applications in the relevant areas. For example, pertinent security data [1] may be utilized to create a data-driven sophisticated computerized security system, and appropriate cellular internet could be utilized for formulate personalized context-aware intelligent apps for mobile devices, and so on As a consequence, real-world applications require data management tools and methods that can quickly and intelligently extract insights or useful information from data.

In recent decades, artificial intelligence (AI) has made significant progress in the fields of information processing and computation, allowing programs to operate intelligently. ML is generally considered as one of the most popular technical breakthroughs of the fourth industrial revolution, since it allows systems to enhance and develop without having to be explicitly coded (Industry 4.0). The term "Industry 4.0 denotes to the progressive automations of traditional operations and manufacturing, integrating the processing of datasets that is investigative, via the application of new technology solutions such as machine learning robots [2].

As a result, machine learning algorithms are critical for successfully evaluating this data and developing suitable practical applications. Unsupervised, supervised, reinforcement learning, and semi-supervised learning are the four main kinds of learning methods. These methods of learning are getting more popular. Experts say that in 2015, the popularity indicator values for different learning styles were low, but that they are now increasing on a regular basis. These results inspire us to write this essay on deep learning, which has the potential to play a significant role in authentic robots as part of Industry 4.0. The kind and qualities of data, as well as the effectiveness of machine learning, influence the usefulness and profitability of computational solutions.

To successfully create data-driven technologies, machine learning methods such as clustering algorithm, modelling, data segmentation, features engineering and dimensional reduction, sequential pattern recognition, and learning algorithm exist. Furthermore, supervised learning is derived from the convolutional neural network, which is part of a larger family of ML algorithms that may be used to cognitively evaluate data. As a result, choosing an appropriate classification model for a given domain's goal application is difficult. This is because various learning techniques serve different goals, and the outputs of various learning algorithms in the same category might change based on data features.

As a result, it's critical to comprehend the concepts of diverse machine learning techniques and their relevance to a variety of real-world applications, such as IoT network, information security solutions, business or recommendations structures, intelligent buildings, universal health care and COVID-19, context-aware applications, sustainable farming, etc. Considering the relevance and capability of ML to evaluate datasets illustrated above, we issue a detailed analysis of the types of ML techniques, which are applicable in enhancing the capacity and knowledge of an operation in this article.

As a result, the report's major contribution is to explain the concepts and possibilities of various ML approaches, as well as their relevance in the aforementioned actual industrial applications. As a result, the goal of this article is to give a starting point for academics and industry professionals who would like to study, investigate, and build data-driven advanced automation technologies using ML algorithms in the diverse domains. The Section II below presents the types of actual data and ML methodologies. Section III focuses of the machine learning task and algorithm. Section IV presents the challenges and recommends directions for future research. Lastly, Section V concludes the paper.

II. REAL-LIFE DATA MACHINE LEARNING (ML) TECHNIQUES

Data is generally consumed and processed by ML algorithms in order to understand patterns and correlations about persons, corporate processes, interactions, occurrences, etc. We will go through several types of real data and different types of ML algorithms in the sections below.

Real-Life Data

The presence of data is typically thought to be the most important factor in developing a ML technique or data-driven real-world structures. Semi-structured, structured and unstructured data are examples of different types of data. Furthermore, "metadata" is a sort of data that generally represents information about data. We'll go through these sorts of data in a little more detail below.

- *Structured:* This has a well-defined architecture, adheres to a database schema that follows a set of rules, is well-organized and accessible, and is used by an institution or system software. Structured data is generally kept in a statistical manner in well-defined systems, such as SQL model. Structured data includes things like names, date, locations, credit card details, inventory data, localization, and so on.
- *Unstructured:* Unstructured data has no pre-defined structure or classification, rendering it considerably harder to gather, handle, and evaluate, since it mostly consists of textual and media. Unstructured data includes things like sensor information, mails, blog posts, forums, and word processors, as well as Pdfs, voice recordings, videos, pictures, slideshows, websites, and a variety of other documentation.
- *Semi-structured:* Semi-structured data, unlike structured data, is not kept in a database system, it has some key characteristics that make it possible to examine. Semi-structured data includes NoSQL databases, JSON documents, XML and HTML among others.
- *Metadata:* It is "metadata," as opposed to "regular statistics." The major variation between metadata and data is that data is typically information that may be utilized to classify, measure, or even explain anything about an association's data qualities. Metadata defines and prioritizes relevant information for data consumers. The publishers, image format, dates created by publication, tags to describe the content, and so on are only a few examples of metadata in a manuscript.

Scientists in the fields of ML and data science employ a variety of publically available datasets for diverse reasons. NSL-KDD, Bot-IoT, and other cybersecurity analytics, as well as smartphones sets of data including telephone call records, SMS records, and mobile software use logs, are examples [3]. In various software sectors, IoT data, agricultural and e-commerce data, patient records such as cardiovascular disease, type 2 diabetes, COVID-19, and others may be found. The input can be in any of the forms listed above, and they can differ from one program to the next in the actual world. As stated below, several types of ML algorithms may be employed to assess this form of data in certain sectors of problems and retrieve knowledge or appropriate data from it in order to build reasonably intelligent systems.

Machine Learning Techniques

Reinforcement, Semi-supervised, unsupervised and supervised learning are the four primary categories of ML techniques, as illustrated in Fig. 1. We will go through each sort of learning approach and the extent to which it may be used to tackle real-world situations in the sections that follow.

- *Supervised:* ML is typically utilized in training patterns, which translate to entries on outputs with respect to experimental inputs and outputs pairs, which is known as supervised methods. In order to elucidate a signal, it employs training data set and a set of training instances. Whenever a task-driven method is used, supervised learning happens when specific planning is done to be achieved from a specific set of input data. "Classification," which separates data, and "structural equation modeling," which matches data, are the two most typical supervised operations. Text categorization, for illustration, is an example of supervised learning. It involves forecasting the classifier or emotion of a piece of text, such as a tweet or a customer reviews.
- *Unsupervised:* Unsupervised learning is a data-driven method that examines unmarked datasets but without user intercession. This is commonly used for generating feature extraction, detecting relevant configurations, classifications in findings, and experimental reasons. Clustering, computation, pattern recognition, data preprocessing, discovering frequent patterns, outlier detection, and other unsupervised cognitive domain are among the most popular.
- *Semi-supervised:* Because it works with the both labeled and unlabeled data, semi-supervised learning is a combination of the unsupervised and supervised approaches discussed above. As a result, it lies between "unsupervised" and "supervised" learning. In the actual world, classification may be scarce in a variety of situations, but unlabelled abounds, making semi-supervised learning a viable option. The primary objective of a semi-supervised learning approach is to general more enhanced predictive results compared to the ones retrieved from the inputs of models. Semi-supervised learning is utilized in a variety of applications, including language processing, intrusion detection, dataset labelling, and text categorization.

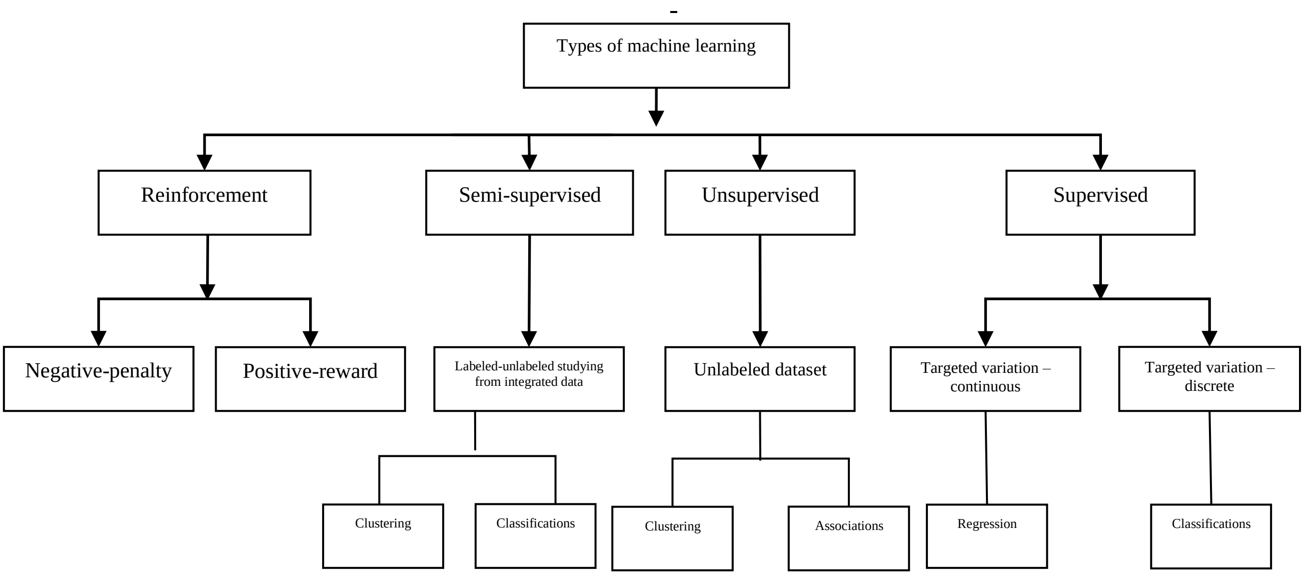

Fig 1. Types of machine learning methods

- *Reinforcement:* This is a type of learning, which allows elements of an application and systems to assess the optimum actions in certain environments and situation to effectively optimize outputs, i.e. an approach that is based on the ecosystem. The major purpose of this learning type that is centred on incentives and penalties is to utilize ecological enthusiasts' expertise to take immediate action to increase incentive or decrease risk. It is a strong tool for developing Artificial intelligence systems that may assist enhance mechanization or enhance the operating effectiveness of complicated techniques such as robots, autonomous vehicle activities, production, and distribution network operations, but it is not recommended for tackling simple or fundamental issues.

III. MACHINE LEARNING TASK AND ALGORITHM

The ML algorithm, also known as classification analysis, will be discussed in this part. In Fig. 2, the overall framework of an ML-based classification algorithm is depicted, with the model being learned based on the application of statistics in the first phase with results being created for fresh training dataset in phase 2. A basic framework for a ML-based prediction model that takes into consideration both the testing and training stages.

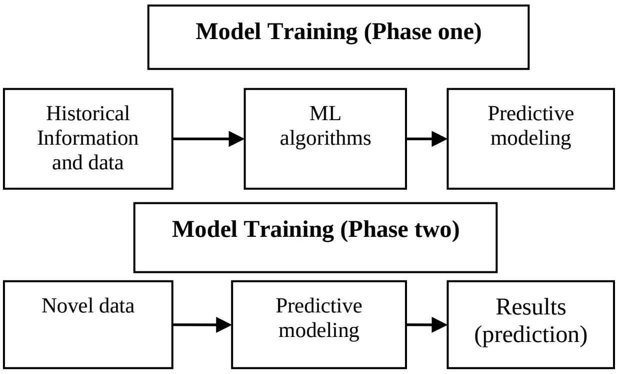
Model Training (Phase one)
Historical Information and data
ML algorithms
Predictive modeling
Model Training (Phase two)
Novel data
Predictive modeling
Results (prediction)

Fig 2. A framework of an ML-based classification algorithm

Classification Analysis

In computer vision, classifying is considered a supervised training approach, and it also refers to a computational modeling issue in which a training sample is anticipated for a specific case. It translates a function (f) from input parameters (X) to output results (Y) as a destination, labeling, or classifications in mathematics. It may be used on unstructured and structured data to estimate the classification of supplied data sets. Fraud detecting (such as "bait" and "not bait," for instance) can be a categorization issue with hosting companies. We'll go over the most frequent categorization issues in the next section.

Rating activities using multiclass identifiers, such as "correct and incorrect" or "yes and no," are known as classification models. In classification model challenges like these, one category may represent the normal condition, whereas another might be the unexpected situation. For example, the usual condition of a task involving a diagnostic procedure is "tumor not found," while the pathological form is "disease discovered." Likewise, in the preceding example of companies and organizations, "bait" and "not bait" are categorical classifications.

Segmentation problems with more than binary classification symbols are known as classification models. In contrast to binary classification problems, multi - label classifying has no concept of normal and pathological results. Rather, samples are categorized as matching to one of several categories. For instance, in the NSL-KDD database, classifying different forms of networking assaults into four training dataset such as denial-of-service, root-local-attack and user-to-root-attack, including assessment of strike, which could be a transfer for the tasks involved in the learning process.

Inter segmentation is an essential factor in computer vision when one sample is linked with many categories or brands. Therefore, it is an extension of classification tasks, in which the initial problem courses are hierarchical, constructed and each sample can correspond to several categories at each hierarchy at the same time, e.g., cross communication by means. For example, Internet headlines may be categorized as "keyword phrase," "innovation," "recent updates," and so on. Modern machine learning techniques that enable forecasting several sequentially semi labels are included in inter categorization. In contrast to classical text categorization, where classification methods are necessarily eligible, inter framework uses sophisticated machine learning techniques to predict several locally semi categories or tags. A variety of analytical techniques have been proposed in the study on statistical machine learning. Below is a list of the most common and well-known methods, which are widely used in a wide range of applications.

Naive Bayes (NB): The Bayesian classifier (nave Bayes) is a method for categorizing data. This algorithm is centred on the Bayes rule, which implies that each pair of attributes is distinct. It operates well in certain specific circumstances, such as documents or text categorization, spam detection, and so on, and may be used both for the single- and multi groupings [4]. The NB algorithm may be used to efficiently categorize the chaotic examples in the dataset and build strong forecasting models. The main advantage is that, in comparison to more complicated techniques, it just requires a minimal quantity of testing phase to rapidly calculate the actual variables. Nevertheless, because of its heavy expectations on aspect dependency, its efficacy may be harmed. The most popular variations of the Classification algorithm are Gaussian, Multivariable, Complementary, Probabilistic, and Qualitative.

Linear Discriminant Analysis (LDA) is a linearly separable border classification that is generated by matching based on conditional distributions to information and then using Bayes' theorem. This approach is also referred to as the development of the Linear Classifiers of the Fisher, which integrates the projection of datasets onto significantly smaller segments (diminishing dimensions that minimize the complexity of frameworks or the costs of processing). Each category is given a Gaussian distribution, and all sections are assumed to have the same covariance matrix [5]. ANOVA and LDA – variance analysis – are the techniques for the representing a single variable as a component of data or data feature.

Logistic regression (LR) is a way of breaking multivariate regression used to address clustering problems in learning algorithms. LR usually uses an LR framework, also referred to as a theoretically specified nonlinear function, to estimate chances. It can generalize high-dimensional data and information well if the variables are divided evenly [6]. Generalized linear (L1 and L2) methods may be used to decrease the dimensionality in such situations. One of the primary drawbacks of LR is the hypothesis of linear regression between the independent and dependent variables. It may be used to solve predictive data mining issues; however classification is the most typical use.

K-Nearest Neighbors (KNN) is a "case knowledge" or "non-generalizing knowledge" approach that is also referred to as a "slow learning" method. Instead of generating a general internal view, it retains all occurrences in non-dimensional spaces that match training examples [7]. KNN is a machine learning method that classifies new datasets using data and similarity measures (e.g., Euclidean distance function). A democratic majority of each view's KNN is used to classify it. It is fairly resistant to harsh learning algorithms; however efficiency is dependent on data clarity. The most difficult aspect of KNN is determining the best number of neighbors to take into account. Both classification and regression may be accomplished with KNN.

A support vector machine (SVM) is another prominent ML approach for classification, predictive analysis, and other problems (SVM) [8]. A SVM creates a high energy or group of excessive self in high-dimensional or indescribable dimensions. Naturally, in each class, the group that has the larger distance considered from a centralized training set of data attains a significant distinction, because the larger the margin, the low the error for predictive classifier. It works effectively in high-altitude settings and may operate in a myriad of areas dependent on the kernel, which are a collection of mental variables. Linear, exponential, radial basis (RB) function, logistic, and other kernel functions are often employed in SVM classifiers. SVM, on the other hand, performs poorly when the database contains more disturbances, such as duplicate targeted communications.

The decision tree (DT) is the modern non-parametric learning technique that is supervised. Both classification and regression problems are solved using DT learning techniques. DT methods such as ID3, C4.5, and CART are well-known. Furthermore, the recently suggested BehavDT and IntrudTree by I. Sarker, Y. Abushark, F. Alsolami and A. Khan in [9] are successful in the appropriate application areas, such as customer data analysis and cyber data analysis. DT categorises the occurrences by classifying down the tree from the real cause to some tree structure, as shown in Fig 3.

Commencing with the root tree node and going down the tree trunk correlating to the feature values, examples are categorized by verifying the attribute specified by that component. The most common criterion for splitting are "gini" for the Gini inefficiency and "fractal dimension" for the classification algorithm, which may be represented formally as Y. Chen in [10] stipulates.

Random forest (RF): A random forest algorithm is a composite classification method used in data science and ML for a variety of applications. As shown in Fig. 4, this method uses "parallel ensembling," where several architecture of the decision trees are integrated in the parallels of several threads of data, and the results are determine by minority decisions or the majorities. Resultantly, the over-fitting problem may be minimized, while prediction performance and manageability increase. As a result, the RF adaptive learning with the many decision trees is significantly more reliable than a system based on a particular decision tree. It uses a combination of ensemble methods (bagging) and unsupervised feature selections to create a succession of decision trees with controllable variance. It may be used to solve both regression and classification issues, and it works well with both continuous and categorical data.

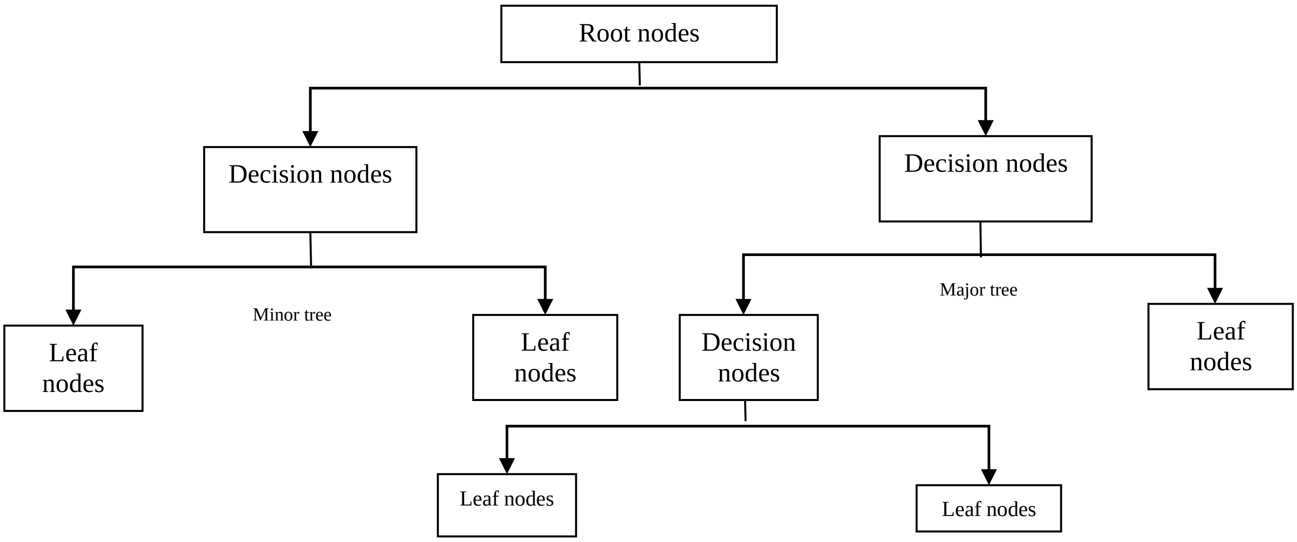

Fig 3. The decision tree structure

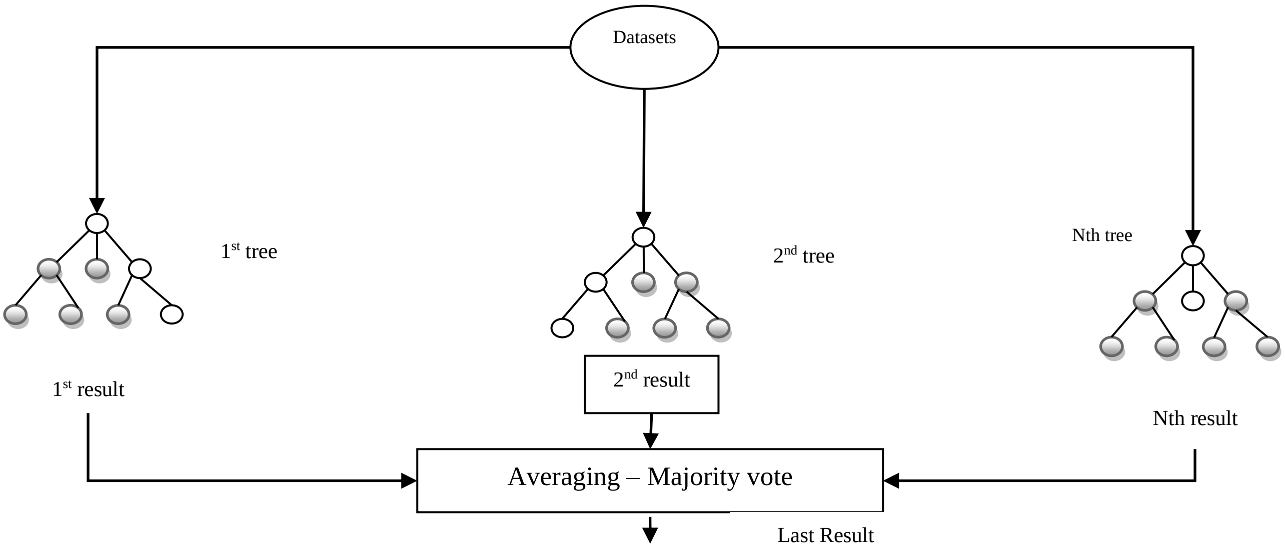

Fig 4. Random forest structure with multiple trees

AdaBoost (Adaptive Boosting): AdaBoost (Adaptive Boosting) is an iterative procedure that uses incremental techniques to possibly enhance classifiers that are poor through learning of mistakes. This is known as "meta-learning," and it was developed by researchers. Adaboost uses "sequence ensembling," while random forest uses "parallel ensembles" it creates significantly powerful classifiers through the integration of different low-performing categories into one. AdaBoost is deemed an adaptive classifier since it significantly increases classifier performance, but it may also induce overfitting in certain situations. On binary classifier tasks, AdaBoost is mostly utilized to enhance the application of the decision tree and the baseline predicted values. Complex data and abnormalities make it vulnerable.

Extreme gradient boosting (XGBoost): Like RF, Gradients Algorithm is an optimization learning technology that provides a final product from a sequence of discrete structures, usually decision trees. In the same way that neural networks utilize learning algorithm to improve weights, the gradients is selected to minimize the gradient descent. When finding the optimum solution, XGBoost (Extreme Gradient Boosting) is a type of learning rate that considers more accurate predictions. It uses sophisticated batch normalization (L1 and L2) to decrease over-fitting and improve model adaptability and execution by computing second-order gradient of a gradient descent to mitigate the adverse effects. XGBoost is simple to use and works well with huge datasets.

Stochastic gradient descent (SGD): SGD is an adaptive algorithm for maximizing an optimal solution with acceptable smoothness characteristics, with the term "random variable" referring to random chance. This decreases the computation complexity, particularly in large optimization algorithms, by providing for quicker iteration in return for a reduced convergence speed. The procedural gradient, which provides the estimations of variables and their amounts of enhancement based on transformation in the dependent variable, is known as gradients. The Gradient Boosting is a linear transformation where return is the partial derivative of a collection of input variables in mathematics. SGD has been effectively used to issues in text categorization and speech recognition in humongous and heterogeneous ML. SGD, on the other hand, is susceptible to features scalability and requires a variety of parameters, like the hyperparameters and iterations count.

IV. CHALLENGES AND FUTURE RESEARCH

Our research into ML algorithms for smart data processing and application raises a number of new research questions in the field. As a result, we outline and analyze the problems experienced, as well as prospective research possibilities and future initiatives, in this section. The type and features of the data, as well as the efficiency of the classification algorithm, determine the efficacy and productivity of ML-based solutions. Even though the present internet permits the creation of a massive quantity of data at a very particular intensity, gathering information in key areas such as security, IoT, universal health care, and agribusiness is vital. As a result, obtaining and maintaining valuable data for ML-based applications, such as green technology application, is important for further research.

As a result, in working with real-world data, a more in-depth analysis of methods of collecting data is required. Furthermore, statistical information may consist of a large number of unclear values, incomplete data, anomalies, and data that has no relevance. The ML method has an influence on the data accessibility and quantity for training, as well as the model that results. As a consequence, cleansing and preprocessing the different datasets retrieved from numbers sources is a challenging obligation. It is essential to modify or improve existing pre-processing methods, or suggest novel feature extraction methodologies, in order to effectively use optimization techniques in the relevant application area.

V. CONCLUSION

Choosing a classification model, which is appropriate for a particular application, might be problematic since the results of various ML algorithms might be different based on the quality of datasets. Choosing an incorrect learning technique will have unintended effects, including a waste of time and resources as well as a decrease in the model's effectiveness and accuracy. The methods described in this study may be instantly applied to a broad variety of real-world issues, such as cybercrime, infrastructure improvements, and healthcare, in accordance with the change creation. The hybrid adaptive learning, for instance, the ensembles of techniques, changing or refining current learning approaches, or developing active learning techniques, might be a future topic of research. As a result, both the data as well as ML are critical to the eventual effectiveness of ML-centred remedies and its advancements. In case the dataset is inappropriate to brainstorm, e.g. inappropriate attributes inefficient learning amounts, poor-quality learning, and non-representation of learning algorithm could potential become decreased or useless and less dependable. Resultantly, formulating a machine-based remedy and constructing smart applications requires quickly processing data and handling different learning methods. As a consequence, the highlighted issues provide intriguing research recommendations in the sector that has to be focussed with practical remedies in various domains of application. In addition, it is believed that our study into ML technologies considers the right locus and might be utilized as a technological basis of future research in the segment of industrial and academic application for decision-makers and practitioners.

References

[1]. J. Daily and B. Gardiner, "Cybersecurity Considerations for Heavy Vehicle Event Data Recorders", SAE International Journal of Transportation Cybersecurity and Privacy, vol. 1, no. 2, pp. 113-143, 2018. Doi: 10.4271/11-01-02-0006.

[2]. D. Nikitin, "Industry 4.0", Scientific Development Trends And Education, 2019. Doi: 10.18411/lj-12-2019-143.

[3]. S. Choudhary and N. Kesswani, "Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 Datasets using Deep Learning in IoT", Procedia Computer Science, vol. 167, pp. 1561-1573, 2020. Doi: 10.1016/j.procs.2020.03.367.

[4]. A. K, "Energy Efficient Network Selection Using 802.16g Based Gsm Technology," Journal of Computer Science, vol. 10, no. 5, pp. 745–754, May 2014.

[5]. D. Toher, G. Downey and T. Murphy, "Semi-supervised linear discriminant analysis", Journal of Chemometrics, vol. 25, no. 12, pp. 621-630, 2011. Doi: 10.1002/cem.1408.

[6]. J. Stoltzfus, "Logistic Regression: A Brief Primer", Academic Emergency Medicine, vol. 18, no. 10, pp. 1099-1104, 2011. Doi: 10.1111/j.1553-2712.2011.01185.x.

[7]. M. HUANG, "Algorithm for finding k-nearest neighbors based on octree segmentation in space", Journal of Computer Applications, vol. 28, no. 8, pp. 2046-2048, 2008. Doi: 10.3724/sp.j.1087.2008.02046.

[8]. J. Li and J. Castagno, "Support Vector Machine (SVM) pattern recognition to AVO classification", Geophysical Research Letters, vol. 31, no. 2, 2004. Doi: 10.1029/2003gl018299.

[9]. I. Sarker, Y. Abushark, F. Alsolami and A. Khan, "IntruDTree: A Machine Learning Based Cyber Security Intrusion Detection Model", Symmetry, vol. 12, no. 5, p. 754, 2020. Doi: 10.3390/sym12050754.

[10]. Y. Chen, "Fractal Modeling and Fractal Dimension Description of Urban Morphology", Entropy, vol. 22, no. 9, p. 961, 2020. Doi: 10.3390/e22090961.