The Scientific Temper (2023) Vol. 14 (3): 950-956
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2023.14.3.60
https://scientifictemper.com/

REVIEW ARTICLE

Machine learning approaches to identify the data types in big data environment: An overview

V. Vijayaraj[1*], M. Balamurugan[2], and Monisha Oberai[3]

Abstract

The digital world has access to a multitude of data in the Fourth Industrial Revolution (4IR, also known as Industry 4.0) era, including Internet of Things (IoT) data, cybersecurity data, mobile data, business data, social media data, health data, etc. These data are produced from various sources. Knowledge of artificial intelligence (AI), specifically machine learning (ML), is essential to intelligently analyse these data and create the associated smart and automated applications. Many different kinds of machine learning algorithms exist in the field, including supervised, unsupervised, semi-supervised, and reinforcement learning. Additionally, deep learning, which belongs to a larger family of machine learning techniques, has the ability to effectively examine a lot of data. In this article, we provide a thorough overview of these machine learning techniques that may be used to improve the functionality and intelligence of an application. Determining the fundamentals of various machine learning approaches and how they can be used in identifying the data types and classify them to be placed in bigdata nodes for the effective storage and retrieval, is thus the core contribution of this work. Based on our findings, we also discuss the difficulties and potential possibilities for future research.

Keywords: Artificial intelligence, Machine learning, Deep learning, Data types and Big data.

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 example, today's electronic world contains a plethora of data types, including Internet of Things (IoT) data, cybersecurity data, smart city data, business data, smartphone data, social media data, health data, COVID-19 data, and many more. The data can be structured, semi-structured, or unstructured, as briefly addressed in Sect. "Types of Real-World Data and Machine Learning Techniques." Insights extracted from these data can be used to develop a variety of intelligent apps in the relevant disciplines. The relevant data types, for example, can be used to develop a data-driven automatic and intelligent big data classification

system. Thus, data management tools and procedures capable of extracting insights or usable knowledge from data in a timely and intelligent manner are required.

The effectiveness and efficiency of a machine learning solution are often determined by the nature and qualities of the data as well as the performance of the learning algorithms. Machine learning algorithms, classification analysis, regression, data clustering, feature engineering, dimensionality reduction, association rule learning, or reinforcement learning approaches are available to develop data-driven systems. Furthermore, deep learning evolved from the artificial neural network, which can be used to intelligently analyse data and is part of a larger family of machine learning technologies. As a result, selecting a decent learning algorithm that is suitable for the intended application in a certain area is difficult. This is because different learning algorithms serve different purposes, and the outcomes of different learning algorithms in the same category may differ depending on the data features. Thus, understanding the principles of various machine

[1]Research Scholar, Cloud Solution Leader - APAC, IBM Global Services, Singapore.

[2]Professor, Department of Computer Science, Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.

[3]Co-Supervisor, Director, Security Services Sales, IBM, Singapore. APAC

***Corresponding Author:** V. Vijayaraj, Research Scholar, Cloud Solution Leader - APAC, IBM Global Services, Singapore, E-Mail: vijay.raj.phd@gmail.com

How to cite this article: Vijayaraj, V., Balamurugan, M. and Oberai, M. (2023). Machine learning approaches to identify the data types in big data environment: An overview. The Scientific Temper, **14**(3): 950-956.

Doi: 10.58414/SCIENTIFICTEMPER.2023.14.3.60

Source of support: Nil

Conflict of interest: None.

Received: 13/07/2023 **Accepted:** 18/08/2023 **Published** : 25/09/2023

learning algorithms and their applicability to 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, bigdata, and many others, is critical.

Based on the importance and potential of "machine learning" to analyse the data indicated above, we present a complete view on many types of machine learning algorithms that can be used to improve an application's intelligence and capabilities in this article. As a result, the main contribution of this paper is to illustrate the concepts and potential of various machine learning approaches, as well as their relevance in datatype detection. This work aims to provide an overview of the various machine learning algorithms that may be used to identify datatypes and store and retrieve them in a bigdata node.

The main contributions of this paper are the capabilities of various learning techniques, providing a comprehensive view on machine learning algorithms that can be used to improve the intelligence and capabilities of a data-driven application, discussing the applicability of machine learning-based solutions in a bigdata environment, and finally summarising potential research directions within the scope of our study. The following section describes the sorts of data and machine learning methods that will be the focus of our investigation. In the next section, we briefly examine and explain several machine learning algorithms, followed by a discussion and summary of various real-world data based on machine learning techniques.

Machine Learning Approaches To Identify The Data Types

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 (Iqbal .H, 2021) (Cao L., 2017) (Sarker IH, 2020).

Types of Real-world Data

Subheadings should be as the above heading "2.1 Subheadings". They should start at the left-hand margin on a separate line.

Structured

It has a well-defined structure, adheres to a data model that follows a regular order, is well-organised and easily accessible, and is used by an entity or a computer programme. Structured data is often stored in a tabular manner in well-defined schemes such as relational databases. Structured data includes things like names, dates, addresses, credit card numbers, stock information, geolocation, and so on Figure 1.

Unstructured

Unstructured data, on the other hand, has no pre-defined format or organization, making it far more difficult to acquire, handle, and analyse, as it primarily consists of text and multimedia information. Unstructured data includes sensor data, emails, blog entries, wikis, word processing documents, PDF files, audio files, videos, photos, presentations, web pages, and many more business documents.

Semi-structured

Semi-structured data, unlike structured data, is not kept in a relational database but contains organisational qualities that make it easier to analyse. Semi-structured data includes HTML, XML, JSON documents, NoSQL databases, and so on.

Metadata

It is not ordinary data, but "data about data." The major distinction between "data" and "metadata" is that data are merely materials that can be used to classify, measure, or even document something in relation to an organization's data attributes. Metadata, on the other hand, explains the pertinent data information, making it more meaningful to data users. The author, file size, date generated by the document, keywords that define the document, and so on are all examples of metadata.

Types Of Machine Learning Techniques

As illustrated in Figure 1, machine learning algorithms are classified into four types: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Following is a quick discussion of each sort of learning strategy and its usefulness to solving real-world challenges (Machine, n.d.) (Types, n.d.) (Common, n.d.).

Supervised

Supervised learning is the process of learning a function that translates an input to an output based on sample input-output pairs. It uses labeled training data and a set of training examples to infer a function. Supervised learning occurs when certain goals are established to be achieved from a specific set of inputs, i.e., a task-driven technique. The most typical supervised tasks are "classification" (data separation)

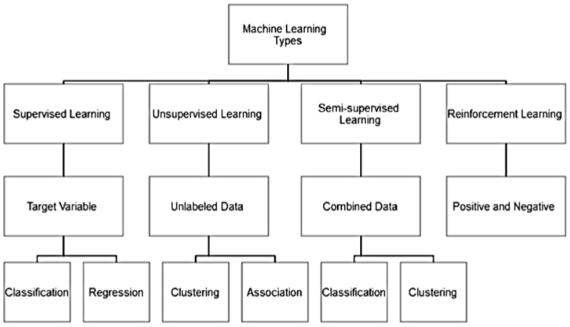

Figure 1: Various types of machine learning techniques

and "regression" (data fitting). For example, predicting a piece of text's class label or sentiment, such as a tweet or a product review, is an example of supervised learning.

Unsupervised

Unsupervised learning is a data-driven method that analyses unlabeled datasets without the need for human intervention. This is commonly used to extract generative features, find relevant trends and structures, organise results, and for exploratory reasons. Clustering, density estimation, feature learning, dimensionality reduction, identifying association rules, anomaly detection, and other unsupervised learning tasks are prevalent.

Semi-supervised

Semi-supervised learning combines the supervised and unsupervised approaches outlined above, as it works with labeled and unlabeled data. As a result, it lies somewhere between learning "without supervision" and learning "with supervision." In the actual world, labeled data may be scarce in some circumstances, but unlabeled data is abundant, making semi-supervised learning helpful. The ultimate goal of a semi-supervised learning model is to offer a better prediction outcome than that obtained by the model utilizing only labeled input. Semi-supervised learning is utilized in a variety of applications, including machine translation, fraud detection, data labeling, and text categorization.

Reinforcement

Reinforcement learning is a machine learning technique that allows software agents and machines to automatically evaluate optimal behaviour in a certain context or environment to enhance efficiency, i.e., an environment-driven approach. This sort of learning is focused on reward or penalty, and its ultimate purpose is to use environmental activists' insights to take action to raise the reward or reduce the danger. It is a powerful tool for training AI models that can aid in the automation or optimisation of the operational efficiency of sophisticated systems such as robotics, autonomous driving tasks, manufacturing, and supply chain logistics; however, it is not recommended for solving simple or straightforward problems.

Machine Learning Algorithms

Classification Analysis

In machine learning, classification is considered a supervised learning method, referring to a problem of predictive modelling in which a class label is predicted for a given sample (Storage, n.d.) (Finding, n.d.) (Alakus TB., 2020).

Binary classification

It refers to classification jobs that have two class labels, such as "true and false" or "yes and no". In such binary classification tasks, one class could be the normal condition, while another class could be the abnormal state. For example, "cancer not detected" is the normal condition of a task involving a medical test, whereas "cancer detected" is the pathological state. Similarly, in the previous example of email service providers, "spam" and "not spam" are considered binary classifications.

Multiclass classification

This traditionally refers to classification jobs with more than two class labels. Unlike binary classification tasks, multiclass classification does not use the idea of normal and abnormal results. Instead, examples are categorised as belonging to one of a set of classes. For example, in the NSL-KDD dataset, classifying various types of network attacks into four class labels, such as denial of service attack (DoS), user to root attack (U2R), root to local attack (R2L), and Probing Attack, can be a multiclass classification task.

Multi-label classification

When an example is associated with multiple classes or labels, this is an essential concern. Thus, it is a generalisation of multiclass classification in which the classes involved in the problem are hierarchically constructed and each example may belong to more than one class in each hierarchical level at the same time, as in multi-level text classification. For example, Google news can be offered under the categories of "city name," "technology," or "latest news," among others. In contrast to classic classification problems in which class labels are mutually exclusive, multi-label classification comprises advanced machine learning methods that support predicting numerous mutually non-exclusive classes or labels.

Many classification algorithms have been proposed in the area of machine learning and data science. Following is a summary of the most frequent and popular approaches that are commonly employed in numerous fields.

Naive Bayes (NB)

The NB method is based on the Bayes theorem, with the premise that each pair of characteristics is independent. It works well and may be utilized in many real-world circumstances, such as document or text classification, spam filtering, and so on, for both binary and multiclass categories. The NB classifier can effectively categorize noisy examples in data and build a robust prediction model. The main advantage is that, compared to more complicated algorithms, it requires a little quantity of training data to rapidly estimate the required parameters. However, its performance may suffer due to its high assumptions on feature independence. The most frequent NB classifier versions are Gaussian, Multinomial, Complement, Bernoulli, and Categorical.

Linear discriminant analysis (LDA) is a linear decision boundary classifier developed by fitting class conditional densities to data and applying Bayes' rule. This method, also known as a generalisation of Fisher's linear discriminant, projects a given dataset into a lower-dimensional space, i.e., a reduction in dimensionality that minimises the model's complexity or reduces the processing costs of the resulting model. The basic LDA model typically assigns a Gaussian density to each class, assuming that all classes share the same covariance matrix. LDA is similar to analysis of variance (ANOVA) and regression analysis in that they both strive to express one dependent variable as a linear combination of other traits or data.

Logistic regression

Logistic regression (LR) is another common probabilistic-based statistical model used to tackle classification problems in machine learning. To estimate probabilities, logistic regression commonly employs a logistic function, often known as the mathematically defined sigmoid function in Eq. 1. It is capable of overfitting high-dimensional datasets and performs well when the data can be split linearly. In such cases, regularisation (L1 and L2) approaches can be utilised to avoid over-fitting. A key disadvantage of logistic regression is the assumption of linearity between the dependent and independent variables. It can be used to solve classification and regression problems but is most typically employed to solve classification difficulties.

g(z)=11+exp(–z).g(z)=11+exp$\boxed{f()}$(–z). (1)

K-nearest neighbors

K-nearest neighbors (KNN) is a "instance-based learning" or non-generalizing learning method, often known as a "lazy learning" algorithm. Instead of developing a broad internal model, it keeps all instances corresponding to training data in n-dimensional space. KNN uses data to classify new data points based on similarity metrics (for example, the Euclidean distance function). Classification is determined by a simple majority vote of each point's k nearest neighbours. It is fairly resistant to noisy training data, and accuracy is dependent on data quality. The most difficult aspect of KNN is determining the ideal number of neighbors to consider. KNN can be used for both classification and regression.

Support vector machine

A support vector machine (SVM) is another prevalent technique in machine learning that can be used for classification, regression, or other tasks. A support vector machine creates a hyper-plane or set of hyper-planes in high- or infinite-dimensional space. In any class, the hyper-plane with the greatest distance from the nearest training data points produces a significant separation since, in general, the higher the margin, the smaller the classifier's generalization error. It works well in high-dimensional spaces and can act differently depending on the mathematical functions known as the kernel. The most common kernel functions employed in SVM classifiers include linear, polynomial, radial basis function (RBF), sigmoid, and so on.

Decision tree

Decision tree (DT): A well-known non-parametric supervised learning method is decision tree (DT). For both classification and regression problems, DT learning algorithms are employed. For DT algorithms, ID3, C4.5, and CART are well recognized. Furthermore, Sarker et al.'s recently introduced BehavDT and IntrudTree are effective in important application domains such as user behavior analytics and cybersecurity analytics, respectively.

Regression Analysis

Regression analysis encompasses numerous machine learning algorithms that allow for the prediction of a continuous (y) result variable based on the value of one or more (x) predictor variables. The most important contrast between classification and regression is that classification predicts distinct class labels, whereas regression predicts a continuous quantity. Financial forecasting or prediction, cost estimation, trend analysis, marketing, time series estimation, drug response modeling, and many more fields now make extensive use of regression models. Some common forms of regression algorithms include linear, polynomial, lasso, and ridge regression, which are briefly detailed below.

Simple and multiple linear regression

Simple and multiple linear regression: This is a well-known regression approach as well as one of the most used ML modeling strategies. The dependent variable is continuous in this technique, the independent variable(s) might be continuous or discrete, and the regression line is linear. Linear regression uses the best fit straight line to build a link between the dependent variable (Y) and one or more independent variables (X). The following equations define it:

y=a+bx+ey=a+bx+e (2)

y=a+b1x1+b2x2+⋯+bnxn+e,y=a+b1x1+b2x2+⋯+bnxn+e, (3)

where an is the intercept, b is the line's slope, and e is the error term. Based on the predictor variable(s), this equation can be used to predict the value of the target variable. Multiple linear regression is a simple linear regression extension that allows two or more predictor variables to model a response variable, y, as a linear function defined in Eq. 3, whereas simple linear regression has just one independent variable defined in Eq. 2 (Anzai Y., 2012).

Polynomial regression

Polynomial regression is a type of regression analysis in which the connection between the independent variable x and the dependent variable y is polynomial in degree n^{th} in x rather than linear. The equation for polynomial regression is derived from the equation for linear regression (polynomial regression of degree 1), which is defined as follows:

y=b0+b1x+b2x2+b3x3+⋯+bnxn+e. y=b0+b1x+b2x2+b3x3+⋯+bnxn+e. (4)

In this equation, y is the predicted/target output, b0,b1,...bn are the regression coefficients, and x is an independent/input variable. In layman's terms, if data is not distributed linearly but rather as 9th degree polynomial, we utilise polynomial regression to obtain the required result (Ardabili SF., 2020).

LASSO and ridge regression

LASSO and ridge regression are well-known as powerful approaches commonly used for generating learning models with many features, due to their ability to prevent over-fitting and reduce model complexity. The L1 regularisation technique is used in the least absolute shrinkage and selection operator (LASSO) regression model, which penalises the "absolute value of magnitude of coefficients" (L1 penalty). As a result, it looks that LASSO reduces coefficients to absolute zero. As a result, LASSO regression aims to determine the subset of predictors that minimises the prediction error for a quantitative response variable. Ridge regression, on the other hand, employs L2 regularisation, which is the "squared magnitude of coefficients" (L2 penalty). Thus, ridge regression forces small weights but never sets the coefficient value to zero, producing a non-sparse solution. Overall, LASSO regression can be used to produce a subset of predictors by removing less important features, whereas ridge regression can be used when a data set has "multicollinearity," which refers to predictors that are connected with other predictors (Baldi P., 2012) (Balducci F., 2018).

Related Works

Analyzing big data requires a lot of computing power. With the use of commonplace computers, distributed computing appears to be a workable solution. However, processing big data effectively on distributed systems presents a number of difficulties, including combining multiple distributed system resources in a seamless manner. Hadoop is one of the most widely utilized systems for processing huge data due to its simplicity. However, it has been noted that in a heterogeneous context, Hadoop's performance suffers. This work suggests Saksham, a block rearrangement technique that reduces processing time in a diverse environment through effective file system management. Big data processing is successfully optimized by the suggested scheme in both homogeneous and heterogeneous environments (Boukerche A., 2020) (De Amorim RC., 2012) (Essien A., 2019).

This work suggests Saksham, a block rearrangement technique that reduces processing time in a diverse environment through effective file system management. The suggested scheme successfully optimizes big data processing in both homogeneous and heterogeneous environments. We focus on file system administration and process management as two key components of heterogeneous distributed computing to improve performance for large data processing. First, file system management essentially manages the placement of the blocks on certain nodes and permits us to rearrange the blocks while taking into account the node's storage and processing capacity. Second, in order to improve process management, we apply the notion of node labeling and scheduling. The outcomes show that the suggested method lowered latency and improved data management while also optimising job execution time (Essien A., 2020) (Fatima M., 2017) (Fujiyoshi H., 2019).

Big data generation and distribution are expanding quickly from a variety of sources, which unavoidably means that data processing must speed up. The task scheduler is in charge of assigning many different jobs to a group of potentially heterogeneous computer nodes in distributed big data processing systems, such as cloud computing, to improve resource efficiency, increase data locality, and shorten makespan. Scheduling techniques that attempt to accomplish these aims in a single pass perform worse than multiple-pass techniques. The suggested MOTS (a hierarchical multi-objective task scheduling scheme) clusters tasks using the K-means algorithm combined with a load balancing equation to maximise resource efficiency before applying evolutionary algorithms to optimise clusters to shorten makespan (Guerrero-Ibáñez J., 2018) (Han J, 2011) (Harmon SA, 2020).

To the latter, physical machines are used, and linked successive tasks are sent to a physical machine in order to reduce data transfer. In Cloudsim, we have tested and simulated our plan. Our tests reveal a 10% makespan reduction and a 4% increase in CPU efficiency when compared to Mai's reinforcement learning strategy and Bugerya's parallel implementation technique. In comparison to Bugerya's methods, the cost of data transfer across related jobs is also 10% lower. It is appropriate for usage in distributed big data processing systems in light of the results and the fact that our suggested task scheduling technique is motivated by the iHadoop method for parallel implementation (Jamshidi M, 2020) (Khadse V, 2018).

Li et al (Li, Y., 2017) provided an overview of various machine-learning techniques used for data type identification in Big Data. It covers methods such as rule-based approaches, machine learning algorithms, and deep learning models. The paper discusses the challenges and future research directions in this field.

Zhang et al (Zhang, L, 2019) presented a comprehensive review of data type identification techniques in the context of Big Data. It discusses traditional methods and focuses on machine learning approaches, including clustering, classification, and deep learning models. The paper also explores the challenges and open research issues in this area.

Gao et al (Gao, X., 2018) focused-on data type identification techniques specifically for big data analytics. It reviews traditional methods, such as statistical analysis and pattern matching, and explores machine learning-based approaches, including feature-based classification and clustering. The paper discusses the strengths and limitations of different techniques and presents future research directions.

Xing (Xing, C., 2018) summarized the state-of-the-art data type identification methods for Big Data. It covers various approaches, including rule-based techniques, machine learning algorithms, and hybrid models. The paper discusses the challenges and future trends in data type identification research.

Wu et al (Wu, H., 2020) focused on deep learning-based approaches for data type identification in Big Data. It provides an overview of deep learning architectures, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers. The paper discusses the application of deep learning in different data types and presents future research directions in this field.

Research Findings

Based on the literature review, most of the big data nodes allow the data to be stored on all the nodes. Searching for specific data is still a challenging task. Keeping the different types of data in separate storage nodes will make it much faster to search and retrieve. In this case, machine learning approaches are suitable to identify the date types and place them in the storage nodes.

Conclusion and Future Work

In this study, we have presented a detailed overview of machine learning strategies for detecting data kinds. We have briefly examined how several types of machine learning approaches can be used to classify data types in accordance with our goal. A successful machine learning model is dependent on both the data and the learning algorithms' performance. Before the system can aid with intelligent decision-making, the sophisticated learning algorithms must be taught using the collected real-world data and knowledge connected to the target application. We also addressed several common machine learning application areas to show its applicability to various real-world situations. Finally, we summarised and discussed the issues encountered, as well as prospective research possibilities and future directions in the field. As a result, the identified issues generate promising research opportunities in the field, which must be addressed with effective solutions in a variety of big data applications. Overall, we anticipate that our investigation of machine learning-based solutions will result in a more accurate identification of data kinds to classify and place in Big Data nodes.

References

Iqbal H. Sarker. (2021). Machine Learning: Algorithms, *Real-World Applications and Research Directions*, Springer Link.

Cao L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*. 50(3):43.

Sarker IH, Hoque MM, MdK Uddin, Tawfeeq A. (2020). Mobile data science and intelligent apps: concepts, ai-based modeling and research directions. *Mobile Networks and Applications*, Springer Link, pages 1–19.

Sarker IH, Kayes ASM, Badsha S, Alqahtani H, Watters P, Ng A. (2020). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*. Springer. 7(1):1–29.

Machine Learning Algorithms for Classification, https://towardsdatascience.com/top-machine learning -algorithms-for-classification-2197870ff501

Types of Machine Learning Algorithms, https://towardsdatascience.com/types-ofmachine-learning-algorithms-you-should-know-953a08248861

Common Machine Learning Algorithms, https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning algorithms/

https://www.sciencedirect.com/science/article/pii/S2666281721001025

Storage and Deep Learning, https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjs_dDh2ez6AhU_tWMGHZtiDFkQFnoECAoQAQ&url=https%3A%2F%2Fwww.computerweekly.com%2Ffeature%2FAI-storage-Machine-learning-deep-learning-and-storage-needs&usg=AOvVaw3O4Yf_f-IZ7hLt5m16HUjk

Finding Standard Data set format for Machine Learning, https://openml.github.io/blog/

openml/data/2020/ 03/ 23/ Finding-a-standard-dataset-format-for machine- learning.html

Alakus TB, Turkoglu I. (2020). Comparison of deep learning approaches to predict covid-19 infection. *Chaos Solitons Fractals*. 140:

Anzai Y. (2012). Pattern recognition and machine learning. *Elsevier*.

Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Rabczuk T, Atkinson PM. (2020). Covid-19 outbreak prediction with machine learning. Algorithms. *MDPI Open Access Journal*. 13(10):249.

Baldi P. (2012). Autoencoders, unsupervised learning, and deep architectures. *Proceedings of ICML workshop on unsupervised and transfer learning*, Pages 37–49 .

Balducci F, Impedovo D, Pirlo G. (2018). Machine learning applications on agricultural datasets for smart farm enhancement. *Machines*. 6(3):38.

Boukerche A. (2020). Wang J. Machine learning-based traffic prediction models for intelligent transportation systems. *Computer Networks*. 181.

De Amorim RC. (2012). Constrained clustering with minkowski weighted k-means. *IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI), pages 13–17. IEEE*.

Essien A, Petrounias I, Sampaio P, Sampaio S. (2019) Improving urban traffic speed prediction using data source fusion and deep learning. *In: 2019 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE*. Pages: 1–8.

Essien A, Petrounias I, Sampaio P, Sampaio S. (2020). A deep-learning model for urban traffic flow prediction with traffic events mined from twitter. *World Wide Web,* Pages:1–24.

Fatima M, Pasha M, et al. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*. 9(01):1.

Fujiyoshi H, Hirakawa T, Yamashita T. (2019). Deep learning-based image recognition for autonomous driving. *IATSS Research*. 43(4):244–52.

Guerrero-Ibáñez J, Zeadally S, Contreras-Castillo J. (2018). Sensor technologies for intelligent transportation systems. *MDPI Sensors*. 18(4):1212.

Han J, Pei J, Kamber M. (2011) Data mining: concepts and techniques. Amsterdam: *Elsevier*.

Harmon SA, Sanford TH, Sheng X, Turkbey EB, Roth H, Ziyue X, Yang D, Myronenko A, Anderson V, Amalou A, et al. (2020) Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. *Nature Communications*. 11(1):1–7.

Jamshidi M, Lalbakhsh A, Talla J, Peroutka Z, Hadjilooei F, Lalbakhsh P, Jamshidi M, La Spada L, Mirmozafari M, Dehghani M, et al. (2020). Artificial intelligence and covid-19: deep learning approaches for diagnosis and treatment. *IEEE Access*. 8:109581–95.

Khadse V, Mahalle PN, Biraris SV. (2018). An empirical comparison of supervised machine learning algorithms for internet of things data. *Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), IEEE*. Pages: 1–6.

Li, Y., & Zhang, C. (2017). Automatic data type identification for big data: A survey. *Journal of Big Data*, 4(1), 32. doi: 10.1186/s40537-017-0097-0.

Zhang, L., Wu, H., & Zou, H. (2019). Data type identification in big data: A comprehensive survey. *IEEE Access*, 7, 140738-140752. doi: 10.1109/ACCESS.2019.2945773.

Gao, X., & Lv, S. (2018). Data type identification for big data analytics: A survey. *IEEE Transactions on Big Data*, 4(2), 147-161. doi: 10.1109/TBDATA.2017.2787413.

Xing, C., Chen, C., & Huang, J. (2018). Data type identification for big data: A systematic literature review. *Information*, 9(3), 57. doi: 10.3390/info9030057

Wu, H., Zhang, L., & Zou, H. (2020). Deep learning-based data type identification in big data: A survey. *ACM Computing Surveys*, 53(4), 1-37. doi: 10.1145/3385412