IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

# AI Big Data Analytics: Study of various Artificial Intelligence and Machine Learning approaches in Big Data Analytics

**[1] Ms. Swati Kadu, [2] Ashwin Dalvi, [3] Mrs. Sayali Belhe**

[1] Assistant Professor, [2] Student, [3] Assistant Professor
[1]Artificial Intelligence and Data Science,
[1]AISSMS IOIT, Pune, India.

***Abstract:*** Since the last decade, tremendous amounts of data have been created. Such data, however, is insignificant without analytical capability. In order for this data to be valuable, useful information must be extracted from it. Big data analytics is the act of analysing massive data sets with a range of data kinds to find undiscovered linkages, consumer patterns, client preferences, and other vital company details. The AI approach to studying Big Data is examined in this paper, along with several AI techniques including decision-making algorithms, machine learning, search strategies and knowledge-based and reasoning methods that may aid in the extraction of useful information from unstructured as well as structured data. An essential component of artificial intelligence is machine learning. From the raw data, they both together extract important information and provide perceptive recommendations and forecasts.

***Index Terms*** **- Artificial Intelligence, Big Data Analytics, Big Data, Machine Learning, etc.**

## 1.INTRODUCTION

The capacity of technology to imitate human intelligence is known as artificial intelligence. Artificial Intelligence (AI) systems can learn, build logic, and make judgements, much like humans. They may be used to conduct a comprehensive variety of functions, from playing activities to detecting diseases. AI is a perpetually extending science, and there is no limit to what it can carry out. In the future, AI is expected to play an even bigger role in our lives, simplifying activities, making our lives simpler, and assisting us to tackle some of the world's most significant challenges [1]. AI is a continuously developing technology, and there is no limit to what it can accomplish [2]. AI researchers are researching a range of strategies to attain this objective, including machine learning, deep learning, and natural language processing [3]. These techniques enable AI systems to learn from information, understand language, and see the world around them [1]. Big data is an increasingly large amount of information that is too massive and complicated to be handled by conventional processing techniques. Big data originates from a variety of sources, including social media, sensor data, and transaction data. It is often unstructured, meaning that it is not in a format that is readily comprehended by computers. This makes it difficult to process and analyze huge amount of data.

However, big data also presents a wonderful opportunity for enterprises and organizations. By analyzing large data, businesses can obtain knowledge about their customers, operations, and markets. This information can be used to enhance decision-making, increase efficiency, and generate revenue [3]. The purpose of big data analytics using AI is to make sense of enormous datasets and uncover patterns that would be difficult or impossible to identify using conventional approaches.AI may be used to automate the processing of data, rendering the process faster and more efficient. It may also be used to increase the accuracy of data analysis, by discovering patterns that would be neglected by human analysts. Additionally, AI may be used to leverage data analysis, making it feasible to investigate huge databases that would be too vast for conventional approaches. By providing data analysis speedier, more accurate, and more scalable, AI may enable firms to acquire an edge over competitors. Organizations may utilize AI to obtain insights about their consumers, operations, and markets. This information may be utilized to enhance decision-making, boost efficiency, and earn revenue [4]. These tasks can be accomplished in most efficient and expedient method such as TensorFlow, Microsoft Azure, OpenAI, NVIDIA, H2O.ai, Amazon Web Services, Data Robot AI Platform, Flotor and many more.

## 2. Overview of AI and Big Data

### 2.1 Artificial Intelligence

Artificial intelligence does not have one standard definition. Whereas it can simply be described as a versatile field that aim to build self-reliable machines that are currently under the influence of human intelligence. AI is classified into two types:

1. Type I: based on capability.
   (a) Weak AI or Narrow AI
   (b) General AI
   (c) Super AI
2. Type II: based on functionality.
   (a) Reactive Machines
   (b) Limited Memory
   (c) Theory of Mind
   (d) Self-Awareness [5].

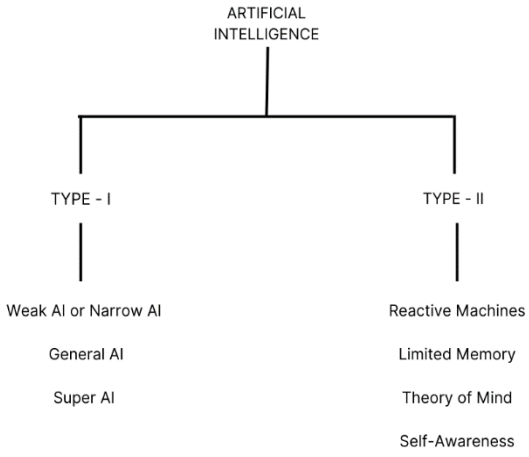


**Fig.1** Classification of AI

Although AI is a multidisciplinary framework with many techniques, advances in machine learning and deep learning are leading to a change in thinking across almost all segments of the IT sector.

### 2.2 Big Data

Michael Cox and David Ellsworth were one of the first to come up with the concept big data, referring to utilising of huge value of records for analyzing and extracting information [6].

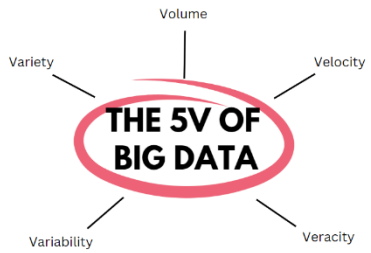


**Fig.2** Characteristics of Big Data

IBM later proposed the 5V characteristics of big data,

i) Volume: The quantity of information that is produced and kept.
ii) Velocity: the rate at which data is produced and processed in order to address the requirements and challenges that lie in development and growth's path.
iii) Variety: The sort and nature of information.
iv) Veracity: The appropriate analysis will depend on the data quality of the gathered data, which might vary greatly.
v) Variability: It refers to information whose value or other qualities are fluctuating in proportion to the situation in which they are being created [7].

### 2.3 Big Data Analytics

Big data analytics is the act of analysing enormous data sets comprising a range of data sorts in order to find hidden linkages, market trends, client preferences, and other vital firm information [10]. Organizations can collect useful information and patterns which may influence business through big data analytics. It also helps in improvising the quality of decision making and boosts the organisation's efficiency.

Many analytical techniques have developed throughout time to extract useful information from raw data, including:

I. **Descriptive analytics** is concerned with examining historical data about a company to explain what happened in past generations.
II. **Predictive analytics** uses a variety of statistical modelling and machine learning approaches to predict potential future outcomes.
III. **Prescriptive analytics** encompass descriptive and predictive analytics to propose the most relevant actions to better company processes [11].

Neural networks, rule-based systems, statistical analysis, machine learning, and data mining are a few more techniques for making better and quicker decisions on huge data sets to find hidden patterns. Big data analytics' main goal is to analyse the data and use it in real-world applications [8].

## 3. AI IN BIG DATA ANALYTICS

Similar to big data, AI is also responsible for the fluctuation in volumes, velocities, and diversity of data. When huge quantities of data need to be managed, AI enables the outsourcing of challenging pattern detection, learning, and other activities to computer- based techniques [6]. We can integrate different data sources using artificial intelligence, additional analytical techniques, and crowdsourcing to get more details about occurrences or pre-events. To determine when an event occurred and how it manifests itself in distinct data sources, each of the many sources of data may be observed through time [9].

The classification of big data analytics methods categorized according to AI subfields is shown in Fig.3.The techniques are classified on the basis of the following parameter:

1. Scalability - The ability of the mechanism to adapt quickly to modifications without compromising the accuracy of the study.
2. Efficiency - It displays the proportion of the process to the overall time and cost needed.
3. Precision - Numerous indicators, such as data mistakes and algorithmic prediction power, are used to detect this.
4. Privacy - It specifies the safeguards in place to ensure that the data is only used for that purpose [11].

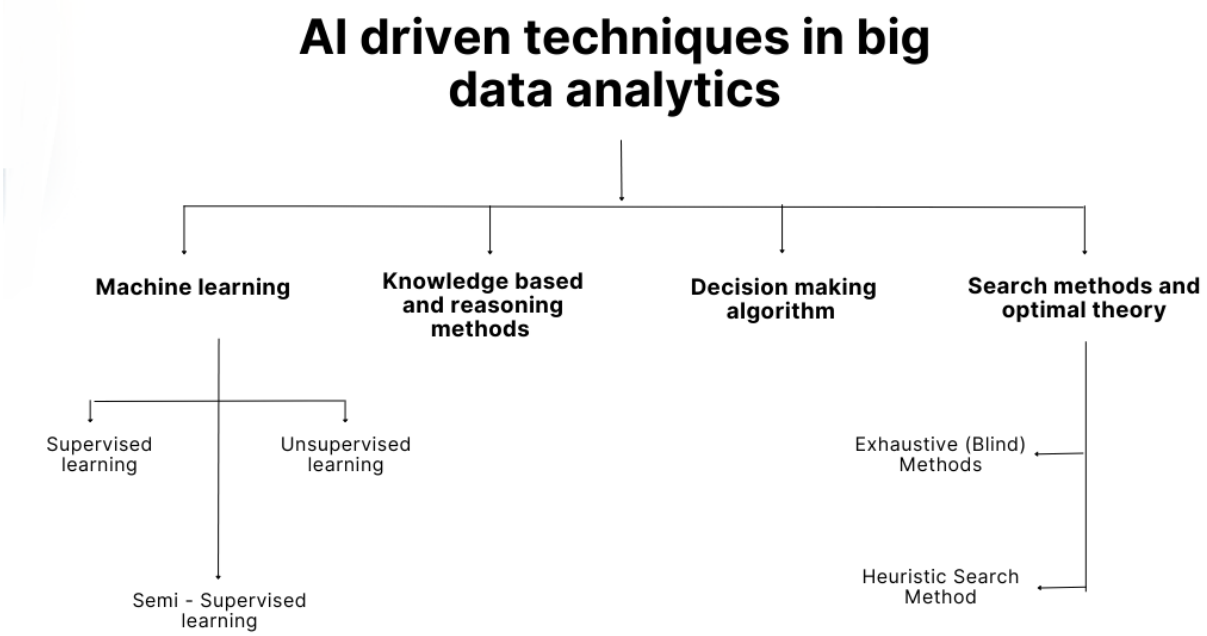


**Fig.3** Schematic diagram of classification of AI

### 3.1 Machine Learning

An essential component of artificial intelligence is machine learning [13]. It is the process of giving computers the ability to learn by using knowledge and data similar to how humans think [12]. ML, however, excels in predictive analytics.

Since they can draw considerable information from raw data and provide intelligent recommendations and predictions, machine learning along with artificial intelligence are crucial for analytics [14]. The main objective of machine learning is to construct models which can educate themselves to develop, perceive the complex patterns, and generate easy solutions to the fresh challenges by using prior data [12].

Supervised, unsupervised, and semi-supervised learning methods may all be grouped under machine learning [13].

### 3.2 Knowledge Based and Reasoning Methods

When using its knowledge base in a particular field, a reasoning machine may do better than a human expert [11].

Knowledge-based Systems (KBS) are systems of computers based on the connection between two primary components: the inference engine and the knowledge base. The total system functions as a problem-solver in a delimitated service area. The inference engine employs a reasoning model and makes use of the information within the knowledge base generate express techniques to input issue descriptions. The knowledge base comprises declarative pieces of information such as rules, object descriptions and facts, related to the issue domain [15].

KBS's primary goal is to increase the effectiveness of time-complexity computation for multiclass classification in huge data sets.

### 3.3 Decision Making Algorithm

These algorithms are designed with the aim to make decision on their own without any human opinion.
In these methods, a utility function decides how appealing a situation is. The agent makes decisions in order to maximise the utility function. The following provides further information on the decision-based strategy that was chosen. When modifications are being made in the cloud while a big data analytics programme is running, it has to be redeployed [11].

Similar to big data, AI is also responsible for the fluctuation in volumes, velocities, and diversity of data. When huge quantities of data need to be managed, AI enables the outsourcing of challenging pattern detection, learning, and other activities to computer-based techniques [16].

### 3.4 Search methods and optimal theory

This is a key concept in artificial intelligence. Search is one of the primary challenges or aims of problem-solving systems. It becomes so anytime the system, through a lack of information, is presented with a decision among a number of possibilities, where each choice leads to the necessity to make other choices, and so on until the issue is addressed. The quantity of search required may be minimised if there is a technique for predicting how effective an operator will be in shifting the original issue state towards a solution. Search methods are broadly classified into two types:

I. Exhaustive (Blind) Methods

- Depth First Search (DFS)
- Means-End Analysis
- Breadth First Search (BFS)
- Bidirectional Search

II. Heuristic Search Method

- Hill Climbing
- Best First Search [17].

## 4. MACHINE LEARNING APPROACHES FOR BIG DATA ANALYTICS

Sequence analysis using machine learning (ML) approaches includes both supervised and unsupervised learning algorithms. While supervised methods primarily address classification issues in this domain, unsupervised techniques are often used for clustering and pattern recognition. The representation of the data items as mathematical values, such as real- or complex-valued integers, vectors, matrices, etc., is often required for machine learning algorithms. Usually, a suitable data transformation and/or feature development are used to get these figures. Following that, object analysis occurs, providing the modelling information as input to the ML-method.

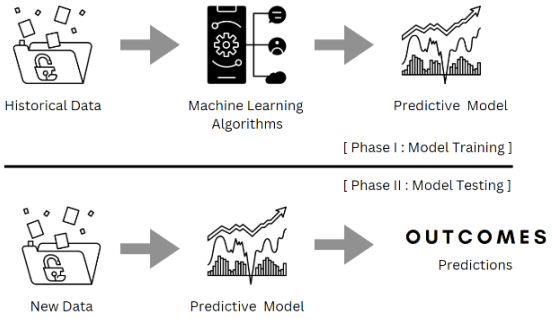


**Fig. 4** A generalized machine learning-based prediction model structure that takes into account both the training and testing phases

A machine learning-based predictive model's general structure is shown in Fig. 4, where the model gets trained using historical data in phase 1 and the outcome is created for brand-new test data in phase 2. Numerous machine learning algorithms may be used for modelling in a particular problem domain depending on their capabilities and learning principles [18].

### 4.1 Decision Tree

Decision trees classify instances by ranking them according to feature values. In a decision tree, each node represents a characteristic of an instance that has to be classified, and each branch represents a value that the node could choose to adopt [20]. It is a model for estimate. A decision tree is a straightforward flowchart-like architecture in which each attribute is tested at each internal node, the results are represented by each branch, and the class designation is provided by each leaf node. The choice is then made once all attributes have been calculated [19]. The whole data set is present at the root node, and as we descend the tree, it continues to divide into smaller subsets. They are created iteratively from training data using a top-down greedy method that successively chooses features [13].

### 4.2 Support Vector Machine

Support vector machine (SVM) is a multimodal classifier that transmits the original data space to a higher-dimensional feature space where it generates a linear classifier. SVM performs very well for data sets of a manageable size. It comes with built-in limitations for big data applications [13]. Since the maximum margin hyper-plane is generated in this space, it may be used to map feature vector onto a higher dimensional vector space. As a result, we choose the hyper-plane such that the maximum distance between it and the nearest data point on each side. The aggregate error categorization decreases with increasing distance between the nearest data of different classes [19].

The number of features included in the training data has no bearing on how computationally complicated an SVM is. Because of this, SVMs are best suited to deal with learning challenges when the quantity of features is excessive compared to the number of training samples [20].

### 4.3 Stochastic Gradient Boosting

A technique to improve a straightforward boosting algorithm is gradient boosting. The weights for the right classification samples and mistake samples are altered in the traditional boosting strategy at each iteration in accordance with the gradient reduction [19]. Boosting the gradient consists of three components:

- A loss function to be minimized.
- A poor learner to produce predictions.
- A model using additive addition for weak learners to lower the loss function.

### 4.4 Naive Bayes classifiers

Naive Bayesian networks (NB) are extremely basic Bayesian networks made up of acyclic graphs that are directed with a single parent (which represents the unobserved node) and multiple children (equivalent to observed nodes), with a strong assumption of independence between child nodes in the context of their parent [20]. Suppose sample X is of type $C_i$. The conditional density of the class is

$$P(C_i|X) = \frac{f(X|C_i)P(C_i)}{P(X)} = \frac{f(X|C_i)P(C_i)}{\sum_{j=1}^{n} f(X|C_i)P(C_i)} \quad - \quad \text{(I)}$$

Where $P(C_i|X)$ is the greatest probability in this case. Features input are X and class type C [19].

### 4.5 k Nearest Neighbors

The K - Nearest Neighbors' algorithm (abbreviated as k-NN) is a technique for classification and regression without using parameters. The k nearest training examples in the feature space make up the input in both circumstances [19]. The underlying assumption of k-Nearest Neighbors (kNN) is that examples within a dataset would often be close to other instances with comparable characteristics. If the instances have classification labels affixed, by taking a quick look at the class of an unclassified instance's closest companions, it is feasible to determine the value of a label for that instance. Since kNN is often thought of as being intolerant of noise, inaccuracies in attribute values may readily distort its similarity measurements, causing it to inaccurately categorize a new instance based on its incorrect closest companions.[20]

**Table 1**. Summary of several machine learning algorithms

| **Algorithm** | Decision tree | Non-linear SVM (based on libsvm) | Linear SVM (based on liblinear) | Stochastic gradient boosting | Naive Bayesian classifier |
|---|---|---|---|---|---|
| **Algorithms type** | Discriminant | Discriminant | Discriminant | Discriminant | Generative |
| **Algorithms characteristic** | Classification tree | Super-plane separation, kernel trick | Super-plane separation | Linear combination of weak classifier (based on decision tree) | Joint distribution of class and feature, conditional independent assumption |
| **Learning policy** | Maximum likelihood estimates with regularisation | Minimising regular hinge loss and maximising soft margin | Minimising the loss of a normal hinge, and maximising the soft margin | Addition minimization loss | Maximum likelihood estimation, maximum posterior probability |
| **Learning algorithms** | Feature selection, generation, prune | Sequential minimal optimization algorithm (SMO) | Sequential dual method | Stochastic gradient descent algorithm | Probabilistic computation |
| **Classification strategy** | If-then logic based on tree spitting | Maximum test sample category | The largest possible test sample | Combining linearly weighted maximum weak classifiers | Maximum posterior probability |

Table 1 compares several machine learning algorithms based on the algorithms' types, traits, learning policies, learning algorithms, and classification schemes. Table 2 compares a few machine learning algorithmic aspects.

| | Decision Trees | Neural Networks | Naïve Bayes | kNN | SVM |
|---|---|---|---|---|---|
| Accuracy in general | @ | $ | * | @ | # |
| Learning rate in relation to the quantity of characteristics and occurrences | $ | * | # | # | * |
| Classification speed | # | # | # | * | # |
| Tolerance to missing values | $ | * | # | * | @ |
| Tolerance to irrelevant attributes | $ | * | @ | @ | # |
| Tolerance to redundant attributes | @ | @ | * | @ | # |
| Tolerance to highly interdependent attributes (e.g., parity problems) | @ | $ | * | * | $ |
| Dealing with discrete/binary/continuous attributes | # | $ (Not discreate) | $ (Not continuous) | $ (Not directly discrete) | @ (Not discreate) |
| Handling the danger of overfitting | @ | * | $ | $ | @ |
| Attempts for incremental learning | @ | $ | # | # | @ |
| Ability of explanation /transparency of knowledge/classifications | # | * | # | @ | * |
| Dealing with model parameters | $ | * | # | $ | * |

**Table 2.** Comparison of machine learning algorithms

(# represent the best performance; $ represent the good performance ;

@ represent the poor performance; *represents the worst performance)

## 5. Conclusion

Big data is an enormous collection of data that may be analysed to obtain critical insights. Decision-making algorithms, machine learning, search strategies and knowledge-based and reasoning methods are the four core categories of AI approaches in big data analytics. Numerous applications, such as corporate intelligence, finance, healthcare, and visual recognition, have showed promise for AI technologies. The utilisation of Big Data technologies to enable accurate decision-making for enhanced performance across industries has immense promise. Decision trees must manage the whole data set of each developing node in order to choose the best separation criteria based on certain quality measures. Because of this, decision trees find it challenging to be used in vast data applications. On data sets of a suitable size, SVM performs well. Applications using large data have limitations by nature. Stochastic gradient boosting, however slower than the multilayer perceptron, attained the maximum classification accuracy across all test data sets.

## REFERENCES


[1] Ms. Neha Saini, "Artificial Intelligence and Its Applications" IJRTI Vol 8 Issue 4. Paper. ID: IJRTI2304061, April-2023.

[2] Anant Manish Singh, Wasif Bilal Haju, Artificial Intelligence. International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 10 Issue VII.ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538, July 2022.

[3] Hussain, Mubashir & Manhas, "Artificial Intelligence For Big Data: Potential And Relevanc", International Acadmey of Engineering and Medical Research, 2016. Volume-1, ISSUE-1, 2016.

[4] Gandomi, Amir & Chen, Fang & Abualigah, Laith, "Big Data Analytics Using Artificial Intelligence. Electronics", 12. 957. 10.3390/electronics12040957, 2023.

[5] Exner-Stöhr, Melanie; Kopp, Alexander; Kühne-Hellmessen, Leonhard; Oldach, Lukas; Roth, Daniela; Zimmermann, Alfred, "The potential of Artificial Intelligence in academic research at a Digital University. Digital Enterprise Computing. Gesellschaft für Informatik", Bonn. PISSN: 1617-5468. ISBN: 978-3-88579-666-4. pp. 61-65. Böblingen. July 11-12, 2017.

[6] D. E. O'Leary, "Artificial Intelligence and Big Data," in IEEE Intelligent Systems, vol. 28, no. 2, pp. 96-99, March-April 2013, doi: 10.1109/MIS.2013.39.

[7] S. Demigha, "The impact of Big Data on AI," 2020 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2020, pp. 1395-1400, doi: 10.1109/CSCI51800.2020.00259.

[8] O'Leary, Daniel, "Embedding AI and Crowdsourcing in the Big Data Lake", Intelligent Systems, IEEE. 29. 70-73. 10.1109/MIS.2014.82, 2014.

[9] Marjani, Mohsen & Nasaruddin, Fariza & Gani, Abdullah & Karim, Ahmad & Hashem, Ibrahim & Siddiqa, Aisha & Yaqoob, Ibrar, "Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges", IEEE Access. 5. 10.1109/access.2017.2689040, 2017.

[10] Rahmani AM, Azhir E, Ali S, Mohammadi M, Ahmed OH, Yassin Ghafour M, Hasan Ahmed S, Hosseinzadeh M, "Artificial intelligence approaches and mechanisms for big data analytics: a systematic study", PeerJ Computer Science 7:e488, 2021.

[11] Çelik, Özer, "A Research on Machine Learning Methods and Its Applications", 10.31681/jetol.457046, 2018.

[12] Wang, Lidong & Alexander, Cheryl, "Machine Learning in Big Data. International Journal of Mathematical, Engineering and Management Sciences", 1. 52-61. 10.33889/IJMEMS.2016.1.2-006, 2016.

[13] Kibria, Mirza & Nguyen, Kien & Villardi, Gabriel & Ishizu, Kentaro & Kojima, Fumihide, "Big Data Analytics and Artificial Intelligence in Next-Generation Wireless Networks", ResearchGate, 2017.

[14] Jean-Claude Latombe, "The Role of Reasoning in Knowledge-Based Systems", Wissensbasierte Systeme, Volume 155. ISBN: 978-3-540-18494-2, 1987.

[15] T. S. Lee, S. Ghosh and A. Neerode, "Asynchronous, distributed, decision-making systems with semi-autonomous entities: a mathematical framework," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 30, no. 1, pp. 229-239, Feb. 2000, doi: 10.1109/3477.826966.

[16] Chijindu, Vincent, "Search In Artificial Intelligence Problem Solving", IEEE African Journal of Computing & ICT, vol 5, 2012.

[17] K. S. Bohnsack, M. Kaden, J. Abel and T. Villmann, "Alignment-Free Sequence Comparison: A Systematic Survey From a Machine Learning Perspective," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 20, no. 1, pp. 119-135, 1 Jan.-Feb. 2023, doi: 10.1109/TCBB.2022.3140873.

[18] Sarker, I.H, "AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems". SN COMPUT. SCI. 3, 158, 2022.

[19] Li, L, "Experimental comparisons of multi-class classifiers". Informatica, 39, 71–85, 2015

[20] Kotsiantis, S. B, "Supervised machine learning: a review of classification techniques", Informatica, 31, 249-268, 2007.