International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 13 Issue 8 ǁ August 2025 ǁ PP. 30-35

Relationship between Machine Learning and Big Data: A Review

Wai Mar Hlaing
Department of Software, Info Myanmar College, Yangon, Myanmar
Corresponding Author: Wai Mar Hlaing

Abstract
This paper aims to explore the logical concept of machine learning, big data, and the relationship between them. Machine Learning is creating and implementing the algorithms. Moreover, these algorithms and datasets are applied together to perform a system's goal, such as decision making and prediction. Most businesses and systems apply the massive volume of data to generate accurate results using machine learning tools and techniques. Big data may be in different types of form such as structured data, semi-structured data, unstructured data and streaming data. This paper also contains the definition of big data, data storage system in big data, MapReduce big data analysis model, machine learning techniques, and how to apply the big data using machine learning tools and techniques.

Keywords: *Big Data, Data Storage, MapReduce, Machine Learning Techniques, Relationship*

Date of Submission: 27-07-2025 Date of acceptance: 05-08-2025

I. INTRODUCTION

Nowadays, most businesses and governments provide valuable predictions and decisions using information technology with their transactional and organizational data. Moreover, data of organizations is enormous over time because of increasing connectivity and technological improvements. Therefore, big data is very crucial and it is used to generate valuable insights for different organizations [1]. The characteristics of Big Data were defined by the "6 Vs". These six characteristics help to define what Big Data really means and how it can be managed effectively [2]. Table 1 describes six characteristics of Big Data.

Table 1: Six Characteristics of Big Data

Volume Terabyte Petabyte Records Transaction	**Velocity** Data in motion Data in rest Real time/offline	**Veracity** Authenticity Availability Accountability Trustworthiness
Variety Structured Unstructured Semi-Structured	**Validity** Correct Data Incorrect Data	**Value** Statistical Hypothetical Correlations Modeling

1.1 Types of Data

Data plays a crucial role in understanding the business trends. Many organizations generate and process huge volumes of data. This huge and complex data is referred to as "Big Data". Big data is of three types: structured data, semi-structured data, and unstructured data [3]. By combining these data, some organizations find out valuable insights using data analysis techniques. [4] presents an approach for the integration of structured, semi-structured and unstructured data focusing on the application in Environmental Engineering.

1.1.1 Structured Data

Structured data is data that is organized and stored into a formatted repository. These data can be used for effective analysis of organizations. Structured data can be stored in tables in the form of rows and columns. They have relational keys and can easily be mapped into pre-designed fields. Today, those data are most processed in the development and simplest way to manage information. [5] describes that any data can be stored in systems

like databases or Excel spreadsheets in the form of rows and columns, without losing any information. Examples of structured data include computer logs, Excel files, databases and CSV files.

1.1.2 Unstructured Data

Unstructured data is data that is not organized in a predefined manner or does not have a predefined data model. In sentiment analysis for social media Facebook comments, a data analyst can receive social media data in various forms, such as text in different languages and image data. Examples of unstructured data include email messages, Word preparation records, recordings, photographs, sound documents, presentations, social media data, pages, and numerous different sorts of business archives [6].

1.1.3 Semi-structured Data

Semi-structured Data is not absolutely a predefined schema, but it also has a scalable schema. For example, xml file is semi-structured data format. When the organizations saves their data using xml files, the system needs to write XML schema definition xsd for these xml files. Figure 1 describes a xml file and a partition of xsd file for bookstore system. In xsd schema file, "note" element is defined as optional element. Therefore, note element data is not essentially needed to store at the section of each item in xml file as described in Figure 1.

```
<?xml version="1.0" encoding="UTF-8"?>

<shiporder orderid="889923"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="shiporder.xsd">
  <orderperson>John Smith</orderperson>
  <shipto>
    <name>Ola Nordmann</name>
    <address>Langgt 23</address>
    <city>4000 Stavanger</city>
    <country>Norway</country>
  </shipto>
  <item>
    <title>Empire Burlesque</title>
    <note>Special Edition</note>
    <quantity>1</quantity>
    <price>10.90</price>
  </item>
  <item>
    <title>Hide your heart</title>
    <quantity>1</quantity>
    <price>9.90</price>
  </item>
</shiporder>
```

```
<xs:element name="item" maxOccurs="unbounded">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="title" type="xs:string"/>
      <xs:element name="note" type="xs:string" minOccurs="0"/>
      <xs:element name="quantity" type="xs:positiveInteger"/>
      <xs:element name="price" type="xs:decimal"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
```

Figure 1:XML and XSD Schema for Bookstore Data [8]

1.1.4 Streaming Data

Streaming Data means that the data at some organizations continuously grows over time, and this data connects with the internet. Today is the age of technology and many people in the world uses the internet. The Internet provides many services such as worldwide communication and collaboration, sending and payment money internationally, learning and educating others, forming cross-border social connections, sharing news, and many others. In 2023, sixty-three percent of the world's population used the internet and the number of internet users has gradually increased since then [9]. The organizations apply three types of data: structured data, unstructured data, and semi-structured data, not only historical data, but also streaming data. The analyst analyzes the streaming data to get up-to-date valuable insights. As a case study, pyspark and google colab tools are used for data streaming emulation. In this study, three CSV files with the same attributes are used. While executing the pyspark file, each CSV file is uploaded for data analysis. By the findings, when streaming data is used for data analysis of a system, only the last uploaded file or the last streaming data is executed at runtime.

II. DISCUSSION

This section discusses data storage for big data, MapReduce framework, machine learning techniques and relationships of big data and machine learning.

2.1 Data Storage

The growth of data in current age becomes a main challenge for data storage. For overcoming this problem, Hadoop Ecosystem is emerged. In Hadoop Ecosystem, Hadoop File System (HDFS) can be used for data storage.

HDFS can save big data in different data types as mentioned in above section 1.1. As a case study, Google Colab is used as a cloud service and HDFS is constructed in Colab. There are some steps to work with HDFS on Hadoop in Colab. Step one is downloading the Hadoop tar file from the apache.org website, and this tar file needs to be extracted in Colab. Then, Java jdk is needed to install and to do the environmental setup. After installing jdk, Hadoop can be run in Colab. Then Hadoop file system can be constructed using "mkdir" command. Table 2 presents step-by-step command that works on Colab for installing Hadoop and constructing HDFS.

Table 2: Steps for Constructing HDFS in Colab

Step	Title	Commands
1	Download Hadoop tar file	!wget https://downloads.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz
2	Install ssh (to connect remote server Colab)	!sudo apt-get install ssh
3	Install Java8	!apt-get install openjdk-8-jdk-headless -qq > /dev/null
4	Set Environmental Variables Java Home	import os os.environ['HADOOP_HOME']="/usr/local/hadoop-3.4.1" os.environ['JAVA_HOME']=java_home_text
5	Run Hadoop	!/usr/local/hadoop-3.4.1/bin/hadoop
6	Create a HDFS	!/usr/local/hadoop-3.4.1/bin/hadoop fs -mkdir hapdoopfile

2.2 MapReduce Big Data Analysis Model

MapReduce is one of the components of Hadoop. It is also a programming model that contains mapper and reducer functions. The responsibility of HDFS is to store the data on Hadoop and these data are processed using MapReduce Programming Model. When the data is split into two parts, the mapper size will be equal to two. After ending the mapper function, the scatter results will be combined by the reducer. In a case study, the total number of ordered quantities of each country was investigated for market analysis. After following the steps that are described in section 2.1, dataset is uploaded on HDFS and then MapReduce functions are defined using python programming language. Figure 2 describes the architecture of MapReduce.

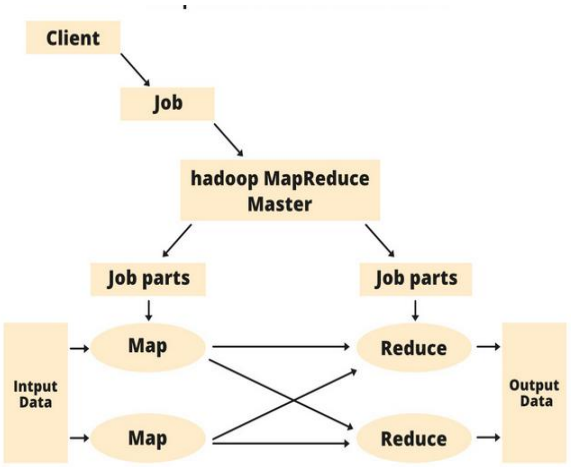

Figure 2: Architecture of MapReduce [10]

2.2.1 Advantages and Disadvantages of MapReduce

MapReduce is a scalable, fault-tolerance and cost-effective programming framework. By using this framework, enormous data can be processed in parallel to get valuable insights during a short time. MapReduce can be used as a machine learning technique with big data. However, this framework is suitable for batch processing. It cannot work for iterative jobs.

2.3 Machine Learning

Machine learning (ML) is a branch of artificial intelligence (AI) focused on enabling computers and machines to imitate the way that humans learn, to perform tasks autonomously, and to improve their performance and accuracy through experience and exposure to more data [11]. There are four types of machine learning techniques such as supervised machine learning, unsupervised machine learning, semi-supervised machine learning and reinforcement learning. These techniques can be used depending on the goal of the application.

The difference between machine learning techniques and MapReduce is that machine learning techniques are used to generate intelligent results using data with repetitive jobs, such as neural networks, whereas MapReduce can be performed for batch processing.

2.3.1 Supervised Machine Learning Technique

Supervised Machine Learning Techniques can be used for dataset that contains the class label attribute. In classification, the model is fully constructed using the training data, and then it is evaluated on testing data before being used to perform predictions on new, unseen data. There are many types of classification and these are binary classification, multi-class classification, multi-label classification and imbalanced classification. In a diabetes medical checking system, medical records of the patients are stored. These datasets contain the symptoms of diabetic patients and their diagnosis results with diabetes type 1, type 2, type 3 and healthy. Whether a new patient has diabetes or not can be known using classification algorithms. There are two types of learners: eager learner, and lazy learner: in machine learning classification.

2.3.1.1 Eager Learner

Eager learners are machine learning algorithms that first build a model from the training dataset before making any prediction on future datasets. They spend more time during the training process because of their eagerness to have a better generalization during the training from learning the weights, but they require less time to make predictions. The algorithms of eager learners are Logistic Regression, Support Vector Machine, Decision Trees, and Artificial Neural Network.

2.3.1.2 Lazy Learner

Lazy learners or instance-based learners, on the other hand, do not create any model immediately from the training data, and this is where the lazy aspect comes from. They just memorize the training data, and each time there is a need to make a prediction, they search for the nearest neighbor from the whole training data, which makes them very slow during prediction. The algorithms of lazy learner are K-Nearest Neighbor, and Case-based reasoning.

2.3.2 Unsupervised Machine Learning Technique

Unsupervised machine learning technique is finding hidden patterns from the dataset without predefined class label. Clustering, Association and Dimensionality Reduction Algorithms are used for unsupervised learning. Clustering algorithms are used to group data points based on similarity. There are two types of clustering, hard clustering and soft clustering. Hard clustering is choosing the most suitable cluster for data points. Soft clustering is calculating the probability of data points in each cluster. For example, when the two cluster is defined for the system, each data point will have probability one for first cluster and probability two for second cluster.

Association is finding the relationships among the items from large datasets. Especially, market basket analysis uses the association algorithms. For example, if one item is bought in the market, what other items can be bought together can be known using association algorithms.

Dimensionality Reduction is the most important feature selection from the input variables or features of dataset. The main goal of dimensionality reduction is to improve model performance, and to reduce computational costs and to be easy in visualization of data analysis.

2.3.3 Semi-supervised Machine Learning Technique

Semi-supervised machine learning technique SSL build a prediction model that contains a small portion of labeled data instances and a lot of unlabeled data instances. SSL combines supervised and unsupervised machine learning techniques. Examples of applications that use semi-supervised machine learning technique are sentiment analysis for movie review system, image classification, and email filtering.

2.3.4 Supervised Vs Semi-Supervised Vs Unsupervised in Machine Learning Techniques

To build a suitable machine learning model for a system, data collection is vital because machine learning techniques learn from data. Table 3 presents how to use the labeled and unlabeled data in each technique.

Table 3: Training Data in Each Technique

Machine Learning Technique	Training data to build Model
Supervised Learning	Labeled data
Semi-supervised Learning	Many Labeled data and a few Unlabeled data
Unsupervised Learning	Unlabeled data

2.3.5 Reinforcement Learning

Reinforcement Learning (RL) is a technique in machine learning. RL allows machines to learn by interacting with an environment and doing decisions and actions through the facts obtained from the environment and receiving feedback based on their actions. This feedback generates two types of form: reward and penalty. In a maze intelligent gaming system, the robot gain the points when the right path is chosen to reach the goal. Otherwise, the points will be lost for moving the wrong direction or path.

2.4 Relationships between Big Data and Machine Learning

Data of an organization comes from various websites in different forms due to the advancement of Internet Technology. Therefore, in current age, Big Data becomes a popular research trend. To get the valuable insights from large amount of data, the analysts need to use machine learning techniques. Google Colab can be used for big data analysis. Google Colab is a free cloud-based platform and it can be used for writing python code and these codes can be executed directly from the browser. In addition, Google Colab can choose CPU, GPU and TPU for computing processes. A lot of data can be uploaded to Google Drive via Colab and the execution time is very short for data analysis.

2.4.1 Python Libraries in Data Science

Python is a popular programming language widely used by data scientists. Python Libraries contain many mathematical functions and analytical tools. There are many python data analysis tools, data visualization tools and machine learning tools. These are briefly itemized in Table 3.

Table 3 Python Libraries

Data Analysis Tools	Data Visualization Tools	Machine Learning Tools
• Pandas • NumPy	• Matplotlib • Seaborn	• Scikit Learn • Keras • PyTorch • OpenCV

Pandas' library can be used for data analyzing, cleaning, exploring and manipulating. Pandas can analyze big data and it can generate the final decision based on statistical data. CSV Data and JSON data can be read using Pandas. Pandas Data Frame can be used to transform data into relevant data for the system. NumPy can also be used for data manipulation, statistical calculation, data transformation and handling missing data in data analysis. Matplotlib and seaborn libraries are used for generating bar chart, pie chart, scatter plot, histogram, line graph and interactive graphs in data visualization. Classification, Association and Clustering Algorithms in Machine Learning can be used to develop a computerized system using Scikit Learn, Keras, PyTorch and OpenCV tools. Figure 4 describes the general steps for implementing an intelligent system. It contains five parts: (1) Data Storage (2) Data Preprocessing (3) Machine Learning Model (4) Evaluation and (5) Testing a user input.

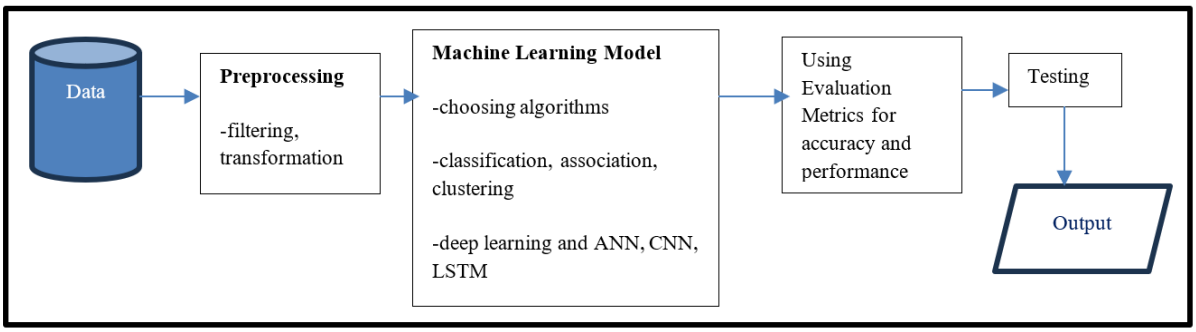

Figure 4: Steps for Implementing an Intelligent System

III. CONCLUSION

This paper describes the emergence of big data in the digital world. Moreover, the characteristics of big data, including data types and their applications, are presented through literature reviews. Hadoop File System and MapReduce Programming Model can be used for batch processing in big data analysis. To get the systems as powerful and intelligent as humans, machine learning algorithms can be used. Types of machine learning algorithms and a brief introduction to Python libraries in data science are described to facilitate the development of machine learning models. In the final section, the paper describes the general steps of an intelligent system in different organizations.

REFERENCES

[1]. Favaretto, M., de Clercq, E., Schneble, C. O., & Elger, B. S. (2020). What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. PLoS ONE, 15(2), 1–20. https://doi.org/10.1371/journal.pone.0228987

[2]. Sabah, Noor.2023. 6 Vs of Big Data.https://doi.org/ 10.13140/RG.2.2.15326.61769

[3]. Rolf Sint, Stephanie Stroka, Sebastian Schaffert, Roland Ferstl.2009. "Combining Unstructured, Fully Structured and Semi-Structured as Information in Semantic Wikis",4th Semantic Wiki Workshop (SemWiki 2009) at the 6th European Semantic Web Conference (ESWC 2009), Hersonissos, Greece, June 1st, 2009. Proceedings

[4]. M. Bărbulescu et al., "Integrating of structured, semi-structured and unstructured data in natural and build environmental engineering," 2013 11th RoEduNet International Conference, Sinaia, Romania, 2013, pp. 1-4, doi: 10.1109/RoEduNet.2013.6511738.

[5]. S. Mishra and A. Misra, "Structured and Unstructured Big Data Analytics," 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore, India, 2017, pp. 740-746, doi: 10.1109/CTCEEC.2017.8454999.

[6]. Das, A. K., & Metkewar, P. (2016). Converting Unstructured Data to Semi-Structured Data. Global Journal For Research Analysis, 5(8), 195–196. https://worldwidejournals.in/ojs/index.php/gjra/article/download/11411/11508

[7]. https://www.w3schools.com/xml/schema_example.asp [online]

[8]. https://ourworldindata.org/internet [online]

[9]. https://www.geeksforgeeks.org/software-engineering/mapreduce-architecture/ [online]

[10]. Verma, P., & Author, C. (2022). A review on Machine Learning: Application and Algorithms. International Journal of Research in Engineering and Science (IJRES) ISSN, 10(10), 274–279. www.ijres.org

[11]. Amer Alnuaimi, Tasnim Albaldawi, 2024, "An overview of machine learning classification techniques", doi:10.1051/bioconf/20249700133, BIO Web of Conferences