ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

Big Data and Deep Learning Analytics

Nipun Tyagi*, Nikita Chauhan, Jaya Ojha and Ayushi Singhal

Department of Computer Science and Engineering ABES Institute of Technology, Ghaziabad, U.P

ABSTRACT

There has been an enormous growth of the Internet, mobile phone, medical facilities, and many more in the 21st century, which can also be known as the beginning of the knowledge era. Knowledge is defined not for what it is, but for what it can do. In this fast-moving technological era, as a result, a huge amount of data is generated in different regions of the world and it is growing day by day, this growing data is known as "Big Data". To extract useful information (analyze) from large unstructured data (like Web, sales, customer contact center, social media, mobile data, and so on) is a complex task, as data being generated is a combination of structured, semi-structured and unstructured data. Traditional systems are not capable to handle semi-structured or unstructured data generated whose volume could range in petabytes or exabytes, as the major challenges are limited memory usage, computational hurdles and slower response time, data redundancy, etc. This problem can be overcome with big data analytics having technologies like Apache Hadoop, Apache Spark, Hive, Pig, etc. which can extract useful information from these large data. Authors are going to explore more on them in these chapters.

Alongside authors will explore "Deep Learning" also known as "Deep Neural Learning" or "Deep Neural Network", which is a class of Machine Learning that progressively extract higher-level features from raw data automatically. It performs 'end-to- end learning' and uses layers of algorithms to process data, understand human speech, and visually recognize objects, which is an important part of it. Feature extraction, self-driving cars, fraud detection, healthcare, neural language processing, etc. are some of the areas where it is applied in daily life. Algorithms like RNN, CNN, FNN, Backpropagation, etc, are some of the algorithms used in deep learning. The authors will explore how Machine learning is different from deep learning.

Deep learning (DL) is also associated with data science in many ways as the DL algorithms work better than older learning algorithms for prediction or feature extraction etc. Which has brought it, more closer towards one of its main objectives i.e., artificial intelligence (AI)? Hence it is immensely advantageous to the data scientists who aim for making predictions and draw useful information to analyze and interpret it for helping the organization in its growth. The processing of Big Data and the evolution of Artificial Intelligence are both dependent on Deep Learning. Deep learning technology came up along with big data analytics. The concept of deep learning is supportive in the big data analytics due to its efficient use for processing huge and enormous data.

This chapter explains about deep learning and big data analytics use in healthcare and alongside authors will study about algorithms used in deep learning and technologies used in big data analytics with its architecture. After reading this chapter, authors must be able to connect deep learning with big data analytics for building new products and contribute to society in a much better way.

***Corresponding author**

Nipun Tyagi, Scholar, Department of Computer Science and Engineering ABES Institute of Technology, Ghaziabad, U.P.

Received: August 07, 2023; **Accepted:** August 18, 2023; **Published:** August 30, 2023

Keywords: Apache Hadoop, Big Data, Deep Learning, CNN, RNN, Yarn, Map Reduce

Introduction

Nowadays, we are living in the Era of Data. Data can be accessed or shared in large amounts between peoples and it is stored into the server or PC. Data is more important in the fields of business and also on a personal basis etc, for the following benefits :

- Data helps in decision making.
- Data gives relevant information about the customer
- Data helps in optimizing the system by analyzing.
- More data gives more analysis and more results that can help in more profits.
- It improves business value. etc.

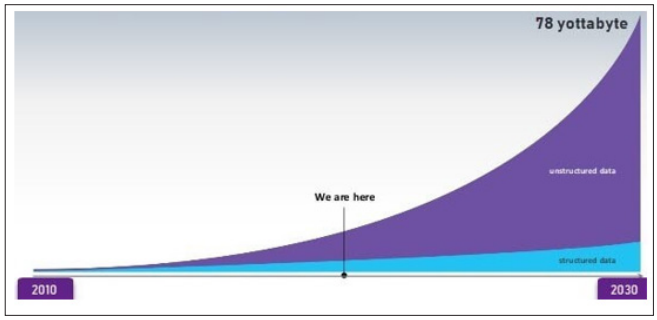

Figure 1: The Following Graph Shows the Increasing Trend of Data In Future [1]

This graph shows that the data after 10 years would be around 78 yottabytes or maybe more. Which would not be easy to manage or to store. So, there was a need to handle or manage that data, therefore there is a concept that can be introduced to solve these problems related about the data i.e Big data and Deep learning is the concept of taking a large number of datasets as input and used to predict the favorable output. By using it, we can remove the difference between our predictions and expected output. In deep learning, there are many hidden layers associated with Input and output layers that are used for computations and it is used in many fields like image recognition, health care, automatic text generation, etc. In deep learning, the patterns and features are automatically identified from a large amount of data and it is used in Big Data Analytics which is an important tool for analysis. If the size of data is large than the models of deep learning are more complex and require more computation to produce more accuracy in results. It makes the devices independent from human knowledge and extracts data directly without human involvement. For example, a large scale image is dealing with one of the approaches for data collection that they can automate the processing for tagging images and extract useful information from images. Also, deep learning is used in disease diagnosis in the medical field. In this chapter, we can more learn about the applications of deep learning and big data in the medical field.

In a large variety of Application domains, data collection and analysis become more important as novel technologies and the need for human-machine interaction with expert systems are growing day-by-day, pushing research towards new information portrayal models and interaction standards. E-commerce to medical diagnostics is the field that is observed by these phenomena. Deep learning is applied in many industries like Google, Facebook, Apple, etc. It can be more explained by using the example of Apple's Siri example, it gives the different variety of facilities like weather reports, reminders, and answers the user's question, etc. by using the deep learning and data can be handled using the Apple's service. Many problems arise in the medical image processing application that has to elaborate on the large volume of data such as strong temporal constraints, security issues, and computational tasks. A data growth model can represent how the clinical data growth increased in both terms such as velocity and volume.

Why Big Data Is Important?

In the today's scenario, Big data is one of the major technology that helps in storing, managing, and manipulating the large volume, large variety and large velocity of data for the Organizations and companies at the required speed, time and give the required Business Value [2].

Why Deep Learning Is Important?

Deep learning achieves higher accuracy levels in recognizing things like humans. It helps to build expectation expectations gap between consumers and electronic companies, such as driverless cars. The present-day improvement is now exceeding the human-level in some of the tasks like recognizing a particular object in an image.

Deep learning involves enormous labeled data as driverless cars are developed by utilizing thousands of videos and millions of images. When combined with cloud computing, it will reduce training time from weeks to hours.

Big Data:

It is a collection of data that is large in volumes and increased by the time. Big data definitions have the main focus on size but there are some other factors in which big data is focused on i.e. data variety and data velocity [3].

In other words, Big data is defined as a technique that solves the non solvable data problems which are not solved by the conventional Databases and tools.

A technique that is used for "Store, Process, Manage, Analysis and Report" large volume of data that allows Real-time Analysis at the required time and speed which is suitable for data handling.

Big data has the following features with the data:

- Data with a tremendously large volume.
- Data with a vast variety.
- Data with very large velocity. There are 5 V's that explain big data are:

Characteristics of Big Data:

(i) **Volume** –The large volume is related to the size that plays an crucial role in finding the value from the data. We can determine the data is Big data or not by depending on the size of data. So the Volume is one of the major characteristics.
(ii) **Variety** – Big data, there is data which may be structured or unstructured or presented as semi-structured. Data that generated in today's era, is in the form of text, images, videos, audios, and PDFs, etc. That is also considered in the analysis process so there is a variety of data.
(iii) **Velocity** –There is another factor i.e data generation speed. It means that at what speed the data is generated or processed to find real potential of data, here speed refers to the Velocity. It is the speed of data flow from sources to destination.
(iv) **Variability** –There is another factor i.e. time. Variability refers to the inconsistency in times which is shown by the data that causes the problems in handling or managing the data effectively.
(v) **Value** –The most important factor is value. Which potential value of big data is huge so it is useless unless and until it can not change it into value [4].

Data has three forms that are as follows :

1. Structured Data
2. Unstructured Data
3. Semi-structured Data

Structured Data: A data which has a fixed format in storage, accessing or processing, and termed as "Structured Data". In which there is a structured format of data like a table in databases. In Today's, there are great techniques that handle such types of data in large volumes. And it can also derive the value of that large amount of data which will grow to a huge extent can being in the range of many zettabytes [1].
Example: An "Student" Table in databases.

Unstructured Data: It is a type of data that does not follow any structure and termed as "Unstructured data" due to its behavior. In which there is data that can contain audio, videos, images, text files, etc [1]. In Today's, organizations have more data that are unstructured type but they do not have any kind of solution to derive the value of the data due to raw format or unstructured format of data.
Example: Audio, Video, etc on a website.

Semi-structured: A combination of both structured and unstrictured is termed as "Semi-structured". It contains both text, video, images, and table format content [1]. It is not defined with the table definition in Relational Databases.
Example: Personal data stored in an XML file.

Big Data Architecture :

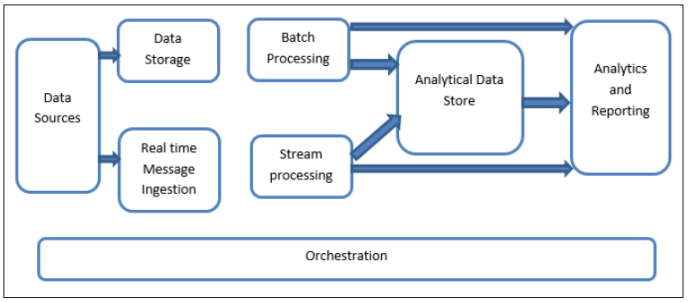

Figure 2: Big Data Architecture [5]

Apache Hadoop:

Apache Hadoop is an open-source framework, used for Big Data analysis given by Apache. It is used to handle the high amount of data for storing and analyzing. It is used for Batch processing and is written in java but is not an OLAP (Online Analytical Processing). Many social media like Facebook, Google, Twitter, Yahoo, LinkedIn, etc. use Apache Hadoop for Data processing.

The concept of **Apache Hadoop** was introduced to resolve two main problems with big data. The first is to store huge amounts of data and second to process that stored data. Since the data is increasing exponentially by the web media and many more sources, so managing load and data becomes very difficult. To manage the data, **Doug Cutting** provided an open-source, scalable, and reliable computing framework known as Hadoop.

Apache Hadoop is the most important framework for Big Data. Scalability is the biggest strength of Hadoop and which improves the performance by working on thousands of nodes rather than a single node without any issue. It has 2 core components HDFS and YARN [6].

Hadoop Architecture:

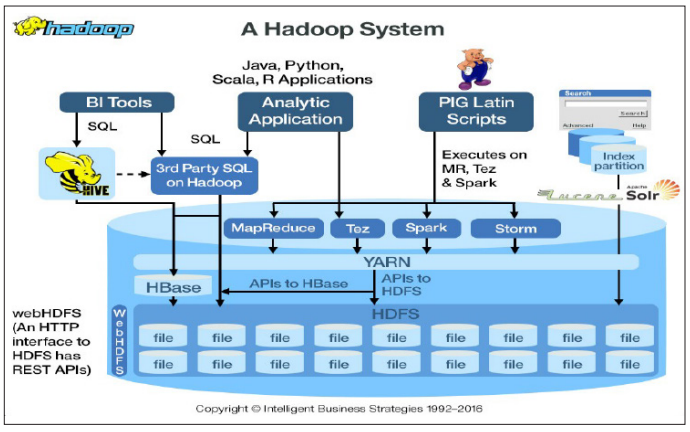

Figure 3: Hadoop Architecture [7]

HDFS (Hadoop Distributed File System): It provides storage for data of Hadoop. HDFS stores data in the smaller units known as blocks in a distributed manner that are produced by the splitting of a single data unit. In which there are two running nodes NameNode (master node) and DataNode (slave nodes) [8].

(i) **NameNode –**
- In the HDFS cluster, there exists only a single master server.
- It is the reason for single-point failure due to a single node existence.
- Many Operations like opening, closing, and remaining of files are executed in file systems that are managed by it.
- It simplifies the architecture of the system.

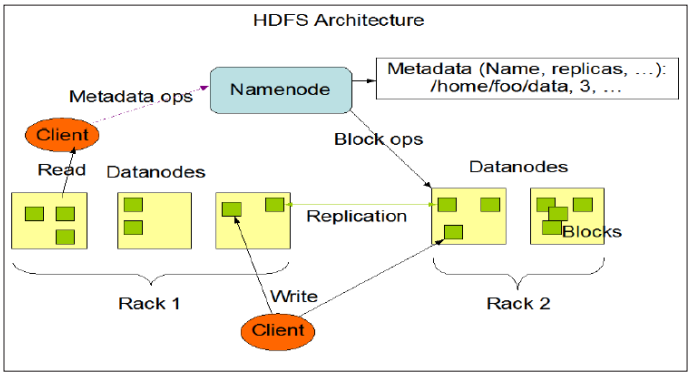

Figure 4: HDFS Architecture [9]

DataNode –
- In which multiple data nodes are present in the HDFS cluster.
- Every data node consists of multiple data blocks.
- Data nodes are used as a storing node.
- For read and write requests from the file system's client, the data node is responsible.
- Creation of replicas & deletion of blocks is performed on getting the instruction from Name Node.

Job Tracker –
- Map Reduce jobs from the client are then sent to the job tracker and it processes the data by the help of Name Node.
- After processing the data, Name Node produced metadata to Job Tracker.

Task Tracker –
- Job Tracker has a slave node, which is known as a task tracker.
- The task of task tracker is to receive the code from the job tracker and test it on the file. This process is known as **Mapper**.

Map Reduce: It is a data processing layer of Hadoop. This allows the user to process a large amount of data by writing applications. These applications run in parallel on clusters at low-end machines by Map Reduce. It is done in a reliable or fault- tolerant way. Map Reduce Jobs contains several maps and several reducer tasks. On data, each task is performed by it, which distributes the load over the cluster. The tasks of Mapper include loading, transforming, parsing, and filtering of data and grouping and aggregations are applied to the transitional data which is the result of map tasks. the task of the reducer is on the subset of output which is generated from Map tasks [8].

HDFS produced the input file for the Map Reduce tasks. How Splitting of the input file into the input splits are done is decided by the input format. Input split is a byte-oriented view of the block of the input file. Loading of input splits data gets done by

the map task. Map tasks run on the node in which important data is present. Here there is no need to move over the network and as it is processed locally.

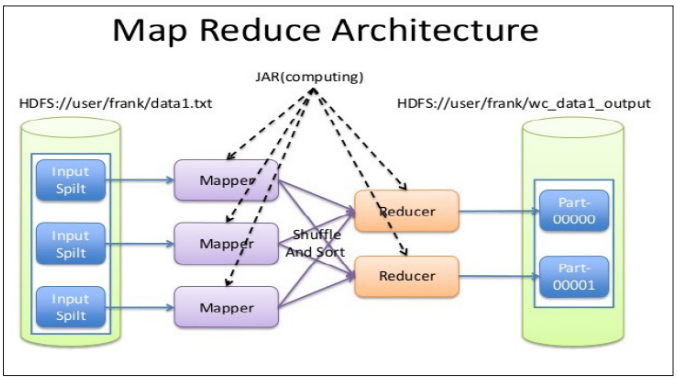

Figure 5: Map Reduce Architecture [10]

Phases of Map Reduce:

Input Reader: In which, the data are splits into the various data blocks of size 64 MB TO 128 MB blocks size and they are associated with the Map function. In which it generates the key-value pair after input reads and those input files are present into the HDFS.

Map Function: In which the output of key-value pairs is the procedure which is included by map function and after that, it produced the corresponding output key-value pairs. Maybe the input-output is different or identical to each other.

Partition Function: Assigning the output of each map function to the reducer is done by the partition function. It produces the index of reducer where the key-value is assigned by this function.

Shuffling and Sorting: To moves the data from map function to the reduced function for the processing, the data is shuffled within the nodes. From time to time the shuffling of data takes so much time for computation. To reduce function, the sorting operation is performed. So the data is correlated using comparison function and arrange data by the sort function.

Reduce Function: In which the unique key is assigned by the reduce function. Already the keys are arranged in sorted order. The output will generate by the value associates with the keys that can iterate the reduction.

Output Writer: Output writer will execute after the flow of data from all the phases. It is used to write the output to stable storage by Reduce.

Yarn: Yet Another Resource Negotiator, it is the resource manager layer of Hadoop. The main function of YARN is to separate resource manager and job scheduling functions. The two managers present inside the yarn are Resource and Node managers [8]. The role of the resource manager is to compete with applications present in the system. The functionality of NodeManager is to monitor the resource usage by the container and reporting to ResourceManger, resources like CPU, memory, disk, network, etc. The task of Application Master is to negotiate the resources with Resource Manager and work with the execution of Node Manager and monitoring of jobs.

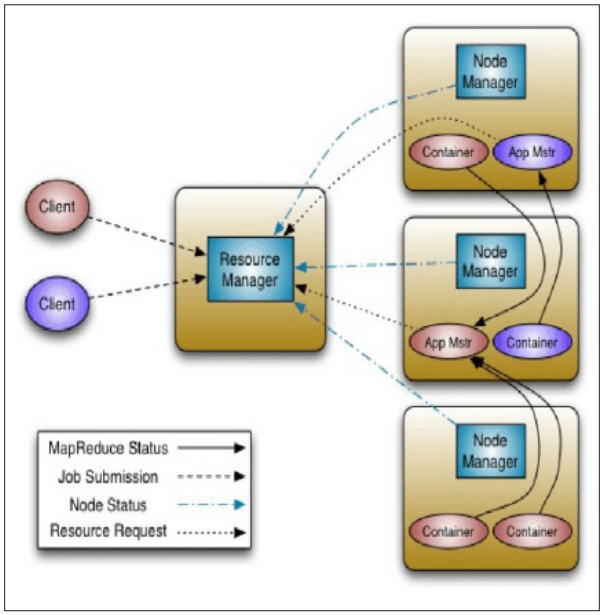

Figure 6: Yarn [11]

Components of YARN:

- Client: client submits the MapReduce jobs.
- Resource Manager: Usage of Resources along with cluster are managed by the Resource Manager.
- Node Manager: In which the computer containers on machine launch and monitor in the cluster by the Node Manager.
- Map Reduce Application Master: MapReduce checks the tasks and application master runs in the container that scheduled by the resource manager and the node manager manages the nodes.

Advantage of Hadoop:

- Fast: Hadoop reduces the processing time and helps in faster retrieval.
- Scalable: By adding the nodes in the cluster, the Hadoop cluster can be extended.
- Cost-Effective: It is cost-effective due to the usage of commodity hardware to store data.
- Resilient to Failure: It is resilient because Hadoop has the property to replicate over the network. The replication factor is configurable. So if one node is dead then Hadoop takes another node i.e the copy of that node [12].

Deep Learning:

Deep learning is the concept of identifying the patterns automatically from a large amount of data and without any human' involvement, it selects the features into the complex unsupervised data, which is useful for Big data analysis. Deep Learning algorithms have been to a great extent undiscovered concerning Big Data Analytics. Most of the Big Data domains, for example, speech recognition and computer vision have seen the utilization of Deep Learning to a great extent to improve classification of displaying results. The capability of Deep Learning to remove a significant level, complex reflections, and data representation from a huge amount of information, particularly from unsupervised data, makes it alluring as an important instrument for Big Data Analytics. All the more explicitly, Big Data issues, like as semantic ordering, data labeling, quick data recovery, and discriminative demonstrating can be better tended to with the guide of Deep Learning.

Deep learning is a subset of machine learning that commands the machines to do in a particular situation that a human mind thinks naturally. Let us try to understand it with an example, deep learning is the technology that is used in robotic cars to recognize the safety signs on-road and helps to differentiate between the pedestrian and lamppost. It has added one of the fundamental innovations behind voice control in different gadgets, for example, cell phones, TVs, tablets, hands-free speakers, and so on and that is why it is getting a large amount of attention and for a substantial explanation. It's enabling us to achieve the results which were quite impossible before deep learning was founded. In deep learning, the computer learns to accomplish image, text, or sound classification directly. Deep learning models can achieve high-level precision, moreover, sometimes it exceeds humans in terms of accuracy. These deep learning models are trained by utilizing an enormous amount of labeled data and neural network architectures containing a large number of layers. In this, we focus on how Deep Learning can help with specific problems in Big Data Analytics [13].

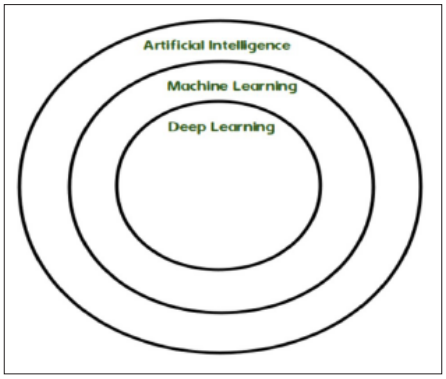

Figure 7: Deep Learning [13]

Deep Learning Architectures

Deep Neural Network: Deep neural network is a neural network that contains many covered layers in between input and output layers. These can model and processing non-linear relationships. It explains the type of machine learning in which the system uses multiple layers of nodes from input information to derive high-level functions. In general, the transforming of data into a more creative and abstract component. In this, the first layer i.e. input layer from where the input is been provided, and the final layer i.e. output layer where the desired output is produced. Between the input and output layer, there is the layer named as a hidden layer that works as an intermediate between the input layer and the output layer. There is a lot of data in the deep neural network so for analysis of the data, we require big data, which enhances the data [13].

Advantages:
- Maximum utilization of unstructured data.
- Removal of the need for feature engineering.
- Capacity to convey excellent outcomes.

Disadvantage:
- It requires large amounts of processing power.

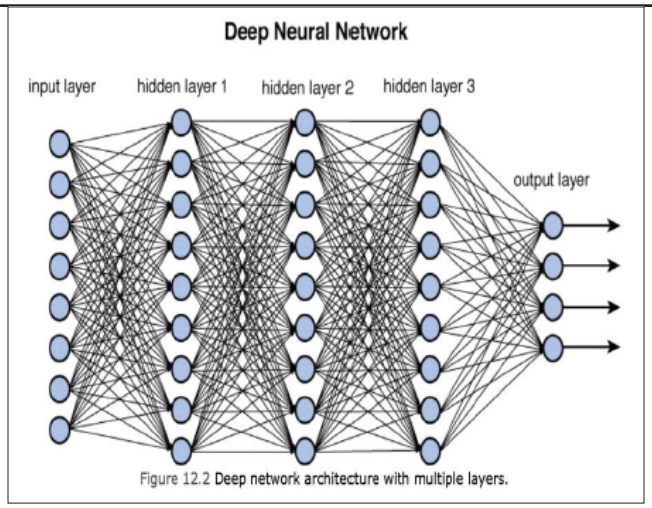

Figure 8: Deep Neural Network [14]

Deep Belief Network(DBN) – It is the category of Deep Neural Network having a multilevel system that includes many unseen layers in which each pair of attached layers is Restricted Boltzmann Machine, which represents a stack of RBMs [13].

In the DBN, the input layer means the fresh sensory inputs, and all unseen layer learns symbolic representations of this input. The network classification is implemented by the output layer.
Procedure to accomplish DBN :
a. Grasp a layer of attributes from noticeable units using the Contrastive Divergence algorithm.
b. Use activations from earlier trained attributes that are noticeable units and then master features.
c. In the end, the complete Deep Belief Network is in position from which the final hidden layer is attained.

Advantages of DBNs:
- It needs a small labeled dataset.
- It is very accurate compared to a shallow net.

Disadvantage of DBNs:
- The quality of the generated speech will be degraded.

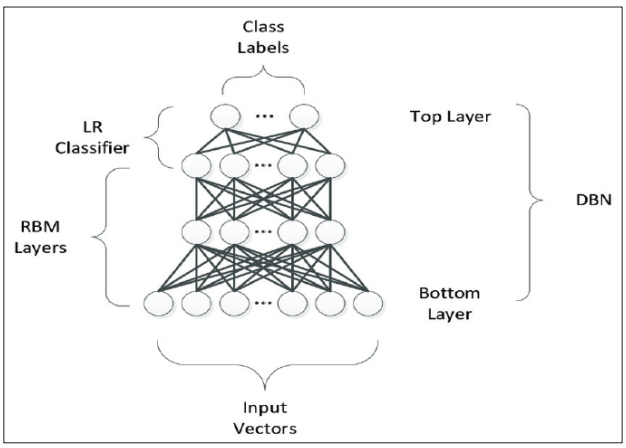

Figure 9: DBNs [15]

Recurrent Neural Network – In RNNs, R stands for 'recurrent' fundamentally because of a uniform task is performed for every single piece of a sequence, in which the output is dependent on the previous computations. Just like the human brain, RNN allows the successive and parallel calculation. RNN allows the additional precise after preparation for manage as a main priority indispensable things concerning the information they received. This network has a memory where the calculated data is stored and produced the output. RNN is perpetual in nature since it performs steady execution for each input while the output of this information relies upon the previous one calculation while manufacturing the output, it's determined and sent into the perennial network. This is done by the large information the big information helps to require the input so offers output. There are unit 2 varieties of RNN:

Bidirectional RNN – The output within the bidirectional RNN depends not solely on the past however conjointly the long-run outcomes.

Deep RNN– In the Deep RNN, many layers are present at every step, allowing for a greater rate of learning and more precision.

Advantages of RNN

- RNN shares constant parameters across all steps. Which greatly reduces the number of parameters that we want to be told.
- RNNs are used with the CNNs for generating the correct descriptions for untagged pictures.

Disadvantages of RNN

- In the case of long sentences and paragraphs, it's troublesome to trace long dependencies.

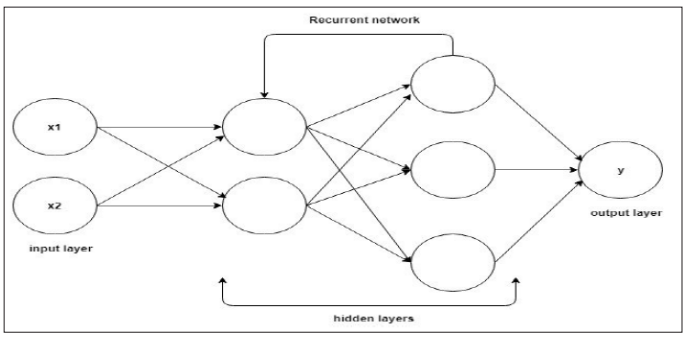

Figure 10: RNN [16]

How Deep Learning Works :

Generally, deep learning procedures utilize neural network structures. The word "deep" refers to the sum total of covered layers present inside the neural network. Beforehand the neural systems just contain 2-3 concealed layers, while profound systems will have as a few as one hundred fifty. Deep learning models are prepared by using a large set of labeled data and neural network architectures that's able to grasp the features automatically from the data with no requirement for manual feature extraction [17].

One of the most popular kinds of deep neural networks is known as CNN or ConvNet. The convolve of CNN understands the features of input data and makes the architecture well appropriate for handling 2D data by using the 2D convolutional layers, such as images, and also used to remove the need for manual feature extraction. CNN works by removing highlights legitimately from images that network to trains on an assortment of images. This computerized highlight extraction makes deep learning models exceptionally exact for computer vision tasks, for example, object grouping.

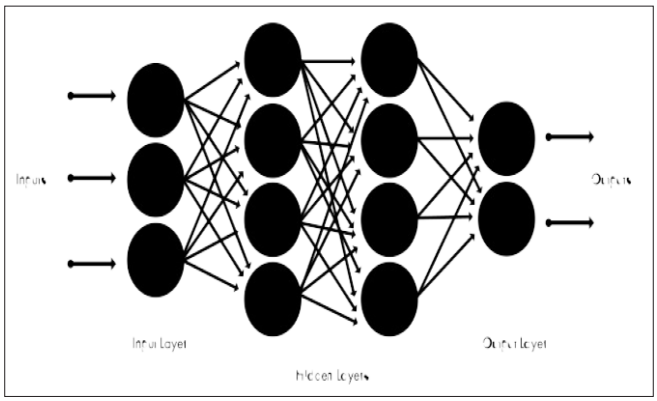

Figure 11: Neural networks, which are organized in layers consisting of a set of interconnected nodes. Networks can have tens or hundreds of hidden layers [17].

Convolutional Neural Networks

CNN is a Deep Learning algorithm that can take an image as an input and determine important aspects and differentiate from each other. CNN learns how to observe different options of an image victimization hundreds or many invisible layers as every invisible layer within the input layer will increase the quality of the grasped image options. The pre-processing that's needed during a CNN is far lower in comparison with an alternative classifying algorithm [18].

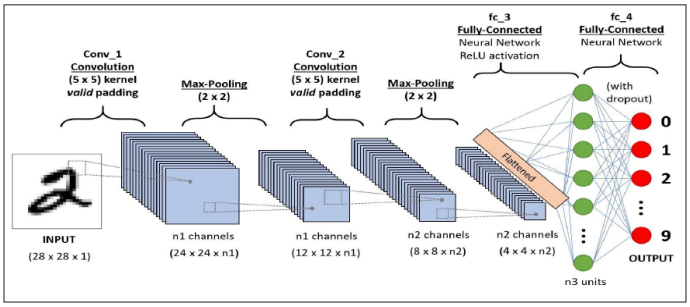

Figure 12: CNN Working Procedure Example [18]

We can apply the neural networks to an image that can use the feedforward neural networks. The input to a ConvNet is organized in a considerable framework that can sustain through layers and maintain these connections. ConvNet comprises different layers of convolutions and activations.

Convolutional layers: In this, the activations from the preceding layers are complex with a set of very small filters, usually of size 3 × 3, collected in a tensor W(j, i), where j stands for filter number and I stands for layer number [18].These all filters contain the equal weights in the complete input field. If the filter is much proficient in detecting horizontal lines, then it helps us to detect all of them anytime they appear. By applying all the CNN filters at the spot of the input to a layer then it turns out into a tensor of feature maps. Convolution is the first layer to draw out features from an input image [15]. It conserves the correlation between pixels by studying image attributes.

Now take a 5 x 5 matrix of image pixel having values as 0 or 1 and 3 x 3 filter matrix as shown in figure 13.

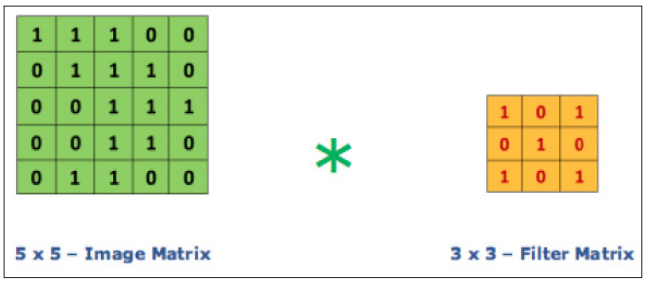

Figure 13: CNN Example [19]

Hence convolution of 5 x 5 (image) matrix multiplied with 3 x 3 (filter) matrix is known as "Feature Map" as convey in the figure 14.

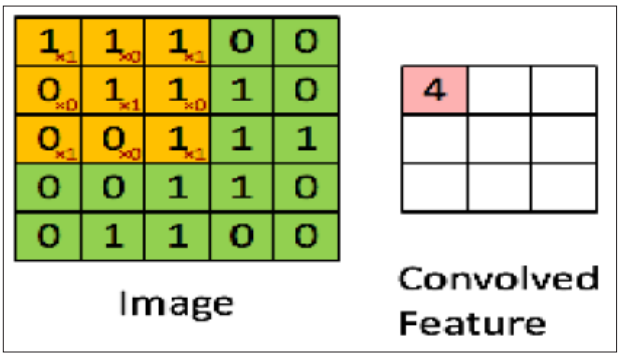

Figure 14: CNN Example Output [19]

Convolution of an image with many different filters can perform operations like edge detection, blur, and sharpen by applying filters.

Activation Layer: It is the second phase where the feature is taken from the first layer and fed through the nonlinear activation functions. The activation functions are a straightforward rectified straight units of ReLUs, outlined as ReLU(z) = max (0, z), or parametric ReLU are used to feed the feature maps via an activation process to generate new tensors, also known as feature maps [19].

Pooling: The feature map that we used is created by the data by using one or more than one convolutional layer is then used to pooled in a pooling layer. The Pooling operations take small grid regions as input and produce the small region. The number is usually calculated by using the max function (max-pooling) or the average function (average pooling) [19]. By removing the pooling layers the network architecture is simplified.

Spatial pooling is also known as subsampling or downsampling which minimize the dimension of each map. It is of different types:

Max Pooling: It takes the largest value from the right feature map.
Average Pooling: Using the biggest element help to take the Average pooling value.
Sum Pooling: It is a sum of all parts present in the feature mapping.

Dropout Regularization: It is used to enhance the performance of CNNs. By using the average of the different models then it tends to increase the performance compared to the single models. Dropout is the term used to take an average which is based on a stochastic sampling of neural networks.

Batch Normalization: This layer is used when the manufacturing normalized activation, activation layer maps by subtracting the mean, and the quality of each derivation is divided.

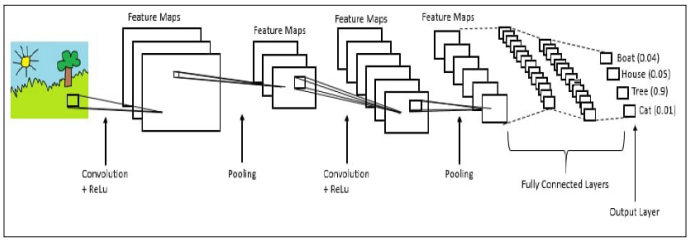

Figure 15: Steps in CNN [19]

CNN include the following things:
CNN process image inside the convolution layer.

- It chooses the parameter, then applies filters to alter the movement of image, padding if requires. Carry out convolution on the figure and initiate ReLU on the matrix.
- Dimension is reduced by pooling.
- Convolutional layers are added until fulfilled.
- Compress the output and sustain into an FC Layer
- Returns the class and apply Logistic Regression with cost functions and categorizes images [19].

Advantage of CNN:

- CNN's is weight sharing.
- CNN's are very good feature extractors.
- CNN sets are very restricted.

Dis-advantage of CNN:

- CNN does not conceal the spot and direction of the object.
- Absence of capability to be spatially constant with the input data.

Examples

As we know the deep learning and big data are two hottest topics growing day by day. The selection of relevant information from big data is not a simple work. Currently, machine learning techniques with improvements and different architectures had played a very salient role in big data analysis and knowledge discovery. In contrast to conventional learning methods, deep learning methods are more successful in industry domains that execute quite well on a large amount of data.

Now, let us understand how deep learning algorithms and big data work together to develop deep patient representation. It is a theoretical explanation to develop deep patient representation. As illustrated by the [Figure1] given below, the health records of the patients are firstly withdrawn out of health data warehouse to preprocess it for identification and normalization and set the patients vectors. The grouped vectors collected from the patients are taken as the input for the feature learning algorithm, as features are ones which help to denote the patients in the data warehouse. The patients are represented with the help of the multi-layer neural networks using suitable deep learning architecture.

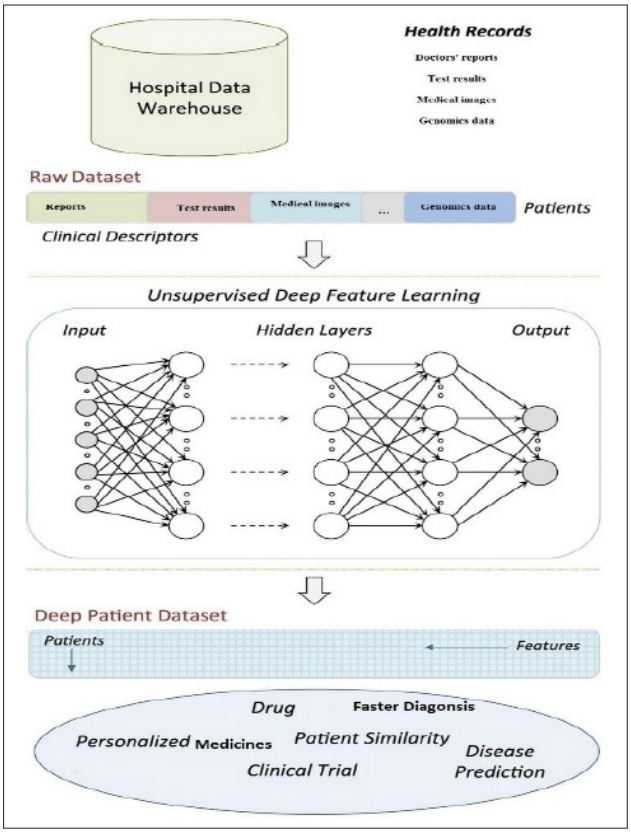

Figure 16: This figure shows the overall process of deep patient's representation on big data diagrammatically [20].

Process:

The large amount of data that is collected from various patient's health records can be structured or unstructured but in most cases, the data is in an unstructured format. This analysis of data is done using big data technology. The unstructured data include text, image, audio, or video. So, it's necessary to preprocess this data into a format understandable by everyone before giving it as an input to the deep learning algorithms. This can be done with the help of Apache Hadoop. This input data is processed using major deep learning architectures like CNN, DBN, and RNN, but the most suitable architecture for medical data and medical images is CNN. The result obtained after processing of input data using deep learning methods is checked by accuracy. If the result is accurate enough then it is used for further diagnosis and treatment while if it is with a low –accuracy rate then input is re-processed again by using other models and architectures for making prediction result better to achieve high accuracy. In the case of medical images, while processing it is difficult to differentiate because it contains many visual features. For that, we consider sensitivity and specificity as a basis for differentiating. The sensitivity and specificity are calculated by the probability of selecting the threshold and prediction of these medical images at the time interval between 0 and 1 [20].

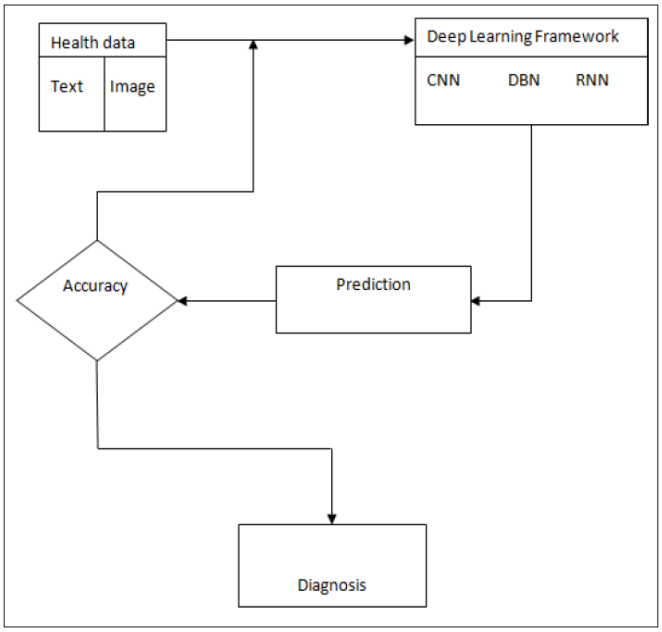

Figure 17: Data Flow chart illustrating the process flow of deep patient's representation [20]

Here, **Sensitivity** is the proportion of patients with the disease who test positive whereas **Specificity** is the proportion of patients without disease who test negative

Sensitivity Formula with probability notation: P(T+|D+) = TP / (TP+FN).**Or**

We can also write it as -Sensitivity = True Positive / Positive

Specificity Formula with probability notation: P(T-|D-) = TN / (TN + FP) **Or**

We can also write it as -Specificity = True Negative / Negative

where,

	Disease Present	Disease Absent
Test Positive	True Positives	False Positive
Test Negative	False Negative	True Negative

Results

Our classification method used here is the deep Convolution Neural Networks (CNN) and its dataflow is from left to right direction. We know that the performance of CNN in the case of medical imaging is far better but when it comes to prediction using medical records and human interventions the error rate is high. Therefore, we use CNN with deep learning architectures to reduce the error rate and increase accuracy.

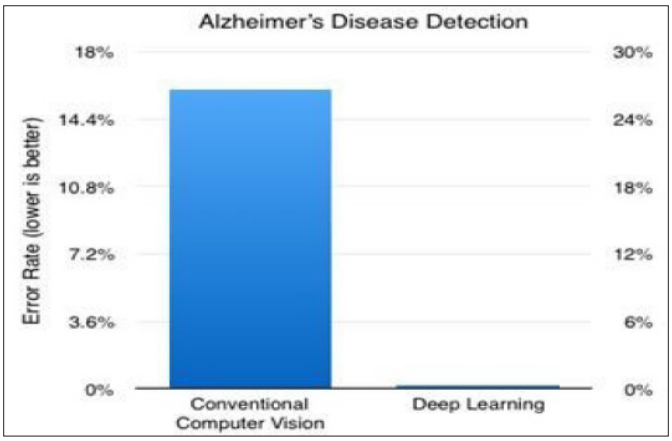

Figure 18: Graph showing CNN Vs Deep Learning [20].

Figure 18 shows the comparison between the CNN algorithm alone with manual interventions Vs CNN algorithm with deep learning and some manual interventions for Alzheimer's Disease Detection. We can see that CNN with deep learning together performs far better as the error rate is approximately 0%.

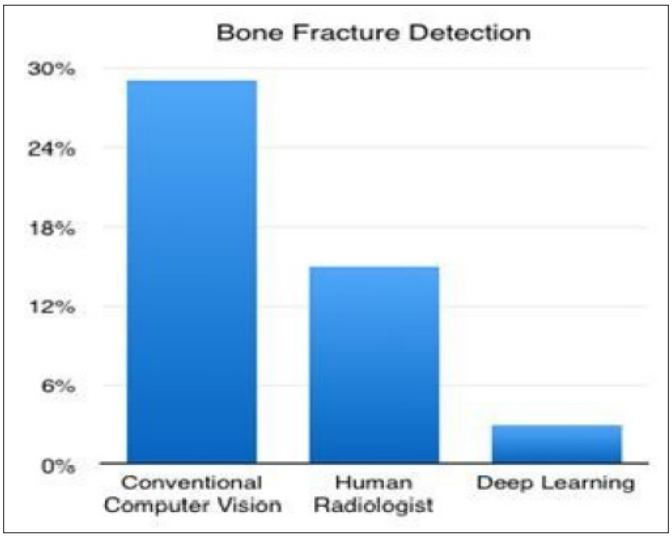

Figure 19: Graph showing CNN Vs Human Radiologist Vs Deep Learning [20].

Figure 19 is the comparison between CNN alone, Human Radiologists, & CNN with Deep Learning for Bone Fracture Detection. We can see in this case also deep learning and CNN together are highly efficient with minimum error rate.

Conclusion

This chapter has discussed deep learning and its methods which are applied to enormous fields of science and engineering like image classification, speech recognition, language processing, etc. Explore Big Data and its architecture as the traditional data processing technique was not efficient enough to process a large amount of data, therefore big data analysis and knowledge discovery were needed [21]. Specific topics studied are:-

- First is, the need for big data analysis in handling and managing the enormous amount of data.
- Second is the need for deep learning methods to process this data and obtain the required result with the help of the CNN algorithm in case of image recognition.
- The third is an illustration of Big data and Deep Learning technologies in the field of medicine to understand how both of them work hand-in-hand to achieve the results with a high accuracy rate.

In this chapter, we have only discussed one main application of deep learning on big data (i.e, image classification) with the help of an example. But there is a wide range of other applications also which is as follows:- Conducting discriminative tasks, Semantic indexing. These technologies are not limited till here, there are several future scopes and some of the future scopes are listed below:-

- Image and fingerprint recognition functions.
- Open source platforms with customer recommendations.
- Banking apps.
- Medical research tool.
- Business trends and outcomes [22-32].

References

1. David Taylor (2023) What is Big Data? Introduction, Types, Characteristics, Examples. GURU99 https://www.guru99.com/what-is-big-data.html.
2. https://www.journaldev.com/8734/introduction-to-bigdata
3. Russom Philip (2011) Big data analytics. TDWI best practices report, fourth quarter 19.4 1-34.
4. Ishwarappa, J Anuradha (2015) A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. Procedia Computer Science 48: 319-324.
5. Big Data Architecture. Heavy AI https://www.heavy.ai/technical-glossary/big-data-architecture.
6. Hadoop. Java T Point https://www.javatpoint.com/what-is-hadoop
7. 32. US Open heralds new era of fan engagement with watsonx and generative AI (2023) IBM https://www.ibm.com/blog/.
8. Hadoop Architecture in Detail – HDFS, Yarn & Map Reduce. Data Flair https://data-flair.training/blogs/hadoop-architecture/
9. HDFS Architecture. Hadoop Apache https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html.
10. https://www.supinfo.com/articles/single/2807-introduction-to-the-mapreduce-life-cycle
11. Apache Hadoop YARN. Hadoop Apache https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html.
12. Kasheeka Goel (2023) Hadoop Introduction - A Beginner's Guide. IntelliPaat https://intellipaat.com/blog/tutorial/hadoop-tutorial/introduction-hadoop/.
13. Introduction to Deep Learning. Geeks for Geeks https://www.geeksforgeeks.org/introduction-deep-learning/.
14. Ravindra Parmar (2018) Training Deep Neural Networks Deep Learning Accessories https://towardsdatascience.com/training-deep-neural-networks-9fdb1964b964.
15. Chenming Li,Yongchang Wang,Xiaoke Zhang, Hongmin Gao (2019) Deep Belief Network for Spectral–Spatial Classification of Hyperspectral Remote Sensor Data 19: 204.
16. RNN or Recurrent Neural Network for Noobs. Hackernoon https://hackernoon.com/rnn-or-recurrent-neural-network-for-noobs-a9afbb00e860.
17. What Is Deep Learning? 3 things you need to know. Math Works https://www.mathworks.com/discovery/deep-learning.html.
18. Sumit Saha (2018) A Comprehensive Guide to Convolutional Neural Networks-the ELI5 way. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

19. Prabhu (2018) Understanding of Convolutional Neural Network (CNN)-Deep Learning. https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148.
20. Balajee Jeyakumar (2017) Big Data Deep Learning in Healthcare for Electronic Health Records. ISROJ 2: 31-35.
21. Naveen Joshi (2017) 3 applications of Deep Learning in Big Data analytics. Allerin https://www.allerin.com/blog/3-applications-of-deep-learning-in-big-data-analytics.
22. What is Deep Learning and its future in 2022? Digi Learnings https://digilearnings.com/deep-learning/.
23. Prajapati Vignesh (2013) Big data analytics with R and Hadoop. Packt Publishing Ltd https://it.dru.ac.th/o-bookcs/pdfs/31.pdf.
24. X Chen, X Lin (2014) Big Data Deep Learning: Challenges and Perspectives. IEEE Access 2: 514-525.
25. https://towardsdatascience.com/deep-learning-algorithms-the-complete-guide-ce020bd47ecc
26. Alexander Selvikvåg Lundervold, Arvid Lundervold (2019) An overview of deep learning in medical imaging focusing on MRI. Journal of Medical Physics 29: 102-127.
27. Weiming Lin, Tong Tong, Qinquan Gao, Di Guo, Xiaofeng Du, et al. (2018) Convolutional Neural Networks-Based MRI Image Analysis for the Alzheimer's Disease Prediction From Mild Cognitive Impairment. Front Neurosci 12: 777.
28. Top 5 Deep Learning Architectures. Packtpub Hub https://hub.packtpub.com/top-5-deep-learning-architectures/.
29. Siddharth Das (2017) CNN Architectures: LeNet, AlexNet, VGG, GoogLeNet, ResNet and more. https://medium.com/analytics-vidhya/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5.
30. Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald (2015) Deep learning applications and challenges in big data analytics. Big Data 2.
31. Bilal Jan, Haleem Farman, Murad Khan, Muhammad Imran, Ihtesham Ul Islam et al. (2019) Deep learning in big data Analytics: A comparative study. Computers & Electrical Engineering 75: 275-287.
32. Basic Concepts and Definitions. https://darwin.unmc.edu/dxtests/reviewof.htm.