CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission

|  |  |
| --- | --- |
| Module Title: | Advanced Data Analytics  Big Data Storage and Processing |
| Assessment Title: | Integrated CA1 Sem 2 MSc in Data Analytics |
| Lecturer Name: | David McQuaid Muhammad Iqbal |
| Student Full Name: | Jay Gilbert |
| Student Number: | sbs22066 |
| Assessment Due Date: | 24th Sept 2023\* asking for extension |
| Date of Submission: |  |

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

**Title: Deep learning using Big Data: Innovations, Challenges, and Applications**

**10 Academic papers**

1. **LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.**
2. **Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171-209.**
3. **Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, 61, 85-117.**
4. **Chen, M., Zhang, Y., & Sun, H. (2014). Big data-driven innovations. Production and Operations Management, 23(5), 971-977.**
5. **Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1). MIT press Cambridge.**
6. **Zaharia, M., et al. (2016). Apache Spark: A unified analytics engine for big data processing. Communications of the ACM, 59(11), 56-65.**

**Equations and Maths check these references not sure is referenced correct**

1. **D.chats, “ Parelell computing and petscale comuting” in parallel computing for Data Science, pp.1-10, 2016.**
2. Nielsen, M.(2015). Neural Networks and Deep Learning: Determination Press.

Forestry X Data Science

**Title: Deep learning using Big Data: Advances, Challenges, and Applications**

# **Abstract**

In today’s data rich era, Deep learning particularly Neural Networks, within the scope of Big Data storage and processing is transforming how we extract insights from massive datasets. This research paper explores these technologies, the advancements made, the challenges, and its real world application. A technical demonstration, complete with code and visualisations, illustrates the power of big data analytics and neural networks in shaping the future of European forestry. The results not only validate the potential of this approach but also offer insights into the importance of forest cover classification. The paper explores the implications, limitations, research gaps, including a critical evaluation of these cutting edge technologies. By examining the academic literature and referencing key studies, this research looks at the current knowledge and applies it to a practical example to advance important insights. Deep Learning and Big Data have emerged as powerful tools to address multifaceted challenges.

Contents

[**Abstract** 4](#_Toc146053789)

[**1.** **Introduction** 4](#_Toc146053790)

[**1.1 Significance of Deep Learning in Big Data** 4](#_Toc146053791)

[**1.2 Research Objectives** 4](#_Toc146053792)

[**2.** **State of The Art** 5](#_Toc146053793)

[2.1 Deep Learning Fundamentals 5](#_Toc146053794)

[2.1.1 Neural Network Architecture 5](#_Toc146053795)

[2.1.2 Activation Functions 5](#_Toc146053796)

[2.1.3 Backpropagation Algorithm 5](#_Toc146053797)

[2.2 Big Data Technology 6](#_Toc146053798)

[2.2.1Distributed File Systems 6](#_Toc146053799)

[2.2.2 Parallel Processing Frameworks 6](#_Toc146053800)

[2.2.3 Data Storage Solutions 6](#_Toc146053801)

[2.3 Recent Advancements in Deep Learning 6](#_Toc146053802)

[2.3.1 Convolutional Neural Networks (CNNs) 6](#_Toc146053803)

[2.3.2 Recurrent Neural Networks (RNNs) 6](#_Toc146053804)

[2.3.3 Transformer Models 6](#_Toc146053805)

[2.4 Big Data Analytics Tools 6](#_Toc146053806)

[2.4.1 Apache Spark 6](#_Toc146053807)

[2.4.2 Apache Hadoop 6](#_Toc146053808)

[2.4.3 TensorFlow and PyTorch 7](#_Toc146053809)

[3. Theoretical Underpinnings (1036 words at this point) 7](#_Toc146053810)

[3.1 Mathematical Underpinnings of Deep Learning 7](#_Toc146053811)

[3.1.1 Cost Functions and Optimization 7](#_Toc146053812)

[3.2 Deep learning Architecture 7](#_Toc146053813)

[3.2.1 Deep Feedforward Networks 8](#_Toc146053814)

[3.2.2 Convolutional Neural Networks (CNNs) 8](#_Toc146053815)

[3.2.3 Recurrent Neural Networks (RNNs) 8](#_Toc146053816)

[4. Integration of Deep Learning and Big Data 9](#_Toc146053817)

[Architectural considerations 9](#_Toc146053818)

[Distributed Deep Learning 9](#_Toc146053819)

[Pipelines, 9](#_Toc146053820)

[Scalable training 9](#_Toc146053821)

[Challenges and limitations 9](#_Toc146053822)

[Dat volume and complexity 9](#_Toc146053823)

[Model scalability 9](#_Toc146053824)

[Resource constraints 9](#_Toc146053825)

[Strategies for mitigating challenges 9](#_Toc146053826)

[Data partitioning 9](#_Toc146053827)

[Parallel techniques 9](#_Toc146053828)

[Model optimisation 9](#_Toc146053829)

[5. Real World Application case study/demo/methods in progress fix Hadoop issues, finish code and record 9](#_Toc146053830)

[6. Implications and limitations of Deep Learning in Big Data 21/9 9](#_Toc146053831)

[7. Critical Evaluation22/9 need a ref for limitations chellenges 9](#_Toc146053832)

[8. Conclusion23/9 9](#_Toc146053833)

[References 9](#_Toc146053834)

# **Introduction**

In the age of data- driven decision making, the relationship between Deep Learning and Big Data is advancing the area of analytics. The integrations of Deep Learning, focusing on Neural Networks, within the domain of Big Data storage and processing is significant.

## **1.1 Significance of Deep Learning in Big Data**

Deep learning, which is part of Machine learning, has emerged as a powerful approach for understanding and modelling complex patterns in data [1]. Deep learning has become an indispensable tool [4], due to its capacity to expose the hidden potential within big datasets in various sectors [2]. Its ability to extract patterns and features enhances insights and decision making in healthcare, finance, and environmental monitoring. The fusion of Deep learning and Big Data has fuelled the emergence of data driven innovation at a rapid rate[6]

## **1.2 Research Objectives**

The Primary objective of this research are:

1. To comprehensively explore the integration of Deep Learning, particularly Neural Networks, within the area of Big Data storage and processing.
2. To evaluate the progress, challenges, and opportunities fusing Deep Learning with Big Data analytics.
3. To demonstrate technical ability, with a practical real world application of this fusion in the sustainable Forestry sector.
4. To critically examine the implications, limitations, and research gaps in the integration of Deep Learning and Big Data

**Research Question**

What are the key challenges and opportunities in effectively integrating Deep Learning, with Neural Networks as its focal point, into Big Data storage and preprocessing, and how do these integrations impact data analytics across diverse domains?

# **State of The Art**

## 2.1 Deep Learning Fundamentals

Deep Learning had advanced as an important element of data analytics, through the understanding of Neural network architecture, activation functions, and backpropagations algorithm.

## 2.1.1 Neural Network Architecture

Deep Learning uses computational models ( neural networks) inspired by the human brain’s structure [1]. Similar to the human brain, these networks consist of interconnected layers of neurons, that process data through weighted connections [1]. The Neural Network architecture usually consists of an input layer, multiple layers that are hidden, and an output layer [5]. They are exceptionally good at learning hierarchical representations of data, this is why they are suited to complex tasks. These tasks could include image recognition and natural language processing.

## 2.1.2 Activation Functions

Activation functions play an important role in neural networks. The introduce non-linearity into the model. The choice of activation functions has an effect on the networks capacity to handle complex data patterns [5]. The sigmoid, tanh, and rectified linear unit (ReLU)[1] are common activations functions. For example, ReLU is known for its ability to mitigate the vanishing point gradient which propels training [5].

## 2.1.3 Backpropagation Algorithm

Backpropagation is the backbone of training neural networks. It involves an iterative optimisation algorithm that works to fine tune the network weights, in order to minimize the likelihood of errors between the predictive and actual outputs [1]. The algorithm works by computing gradients through this network, then adjusting weights in the direction that reduces the error [5]. It is an intensive computational process, that uses various techniques to increase optimization such as mini-batch training and parallel processing [5].

## 2.2 Big Data Technology

In order to manage these massive datasets that are characteristic of Big Data, various technologies have emerged like distributed file systems, parallel processing frameworks, and data storage solutions.

## 2.2.1 Distributed File Systems

Distributed file systems such as Hadoop and Google file system can store and manage vast datasets across clusters [6]. They promote data durability, high availability, and efficient parallel processing which in turn enables the efficient execution of data transformations and analytics across these distributed clusters [6].

## 2.2.2 Parallel Processing Frameworks

Parallel processing frameworks like Apache Hadoop and Apache Spark, facilitate with the distribution of data intensive tasks [7]. They assist in the transformation of data and analytics across distributed clusters, enabling a more efficient execution.[7]

## 2.2.3 Data Storage Solutions

Data storage solutions, Apache Cassandra, Hbase….

## Recent Advancements in Deep Learning

In recent times, there has been a significant advancement in Depp Learning , in articular in the domain of specialized neural network architecture. [ add refer]

## 2.3.1 Convolutional Neural Networks (CNNs)

CNNs have greatly improved image analysis and recognition tasks[5]. They can automatically learn relevant features from image data, through their hierachial architecture such as convolutional and pooling layers[5]. CNNs have practical application in the areas of facial recognition, medical imaging and object detection [5].

## 2.3.2 Recurrent Neural Networks (RNNs)

RNNs are more appropriate for sequential data such as speech recognition and language processing. [5]. Temporal dependencies in data can be modelled through their recurrent connections.[5]. However, there are challenges, as they suffer from vanishing gradient problems with handling long range dependencies[5].

## Transformer Models

Transformer model such as BERT and GPT-3, have reshaped natural language understanding [5]. They employ self-attention mechanisms to capture contextual relationships in text [5]. These models have achieves state of the art in various NLP tasks [5].

## Big Data Analytics Tools

Big Data technology, has a number of complementary tools such as Apache Hadoop, Apache Spark, TensorFlow and PYtorch to facilitate modelling at scale and data processing.

## 2.4.1 Apache Spark

Apache Spark’s accelerates data analytics through the minimisation of data movement between disk and memory [7]. Its versatility makes it suitable for Big Data analytics, as supports a wide range of machine learning libraries and data transformations[7].

## 2.4.2 Apache Hadoop

Hadoop’s MapReduce framework is well known for distributed data processing [7]. It separated tasks into smaller sub tasks and processes them in parallel across cluster nodes [7]. Hadoop’s HDFS provides robust storage for large datasets [6].

## TensorFlow and PyTorch

Tensor Flow and PyTorch are known for their extensive libraries and flexibility which makes them popular Deep Learning frameworks [5]. Cutting edge models can be developed through their provision of tools for building and training neural networks[5].

# Literature Review

The integrations of Deep Learning techniques with Big Data has revolutionised analytics and decision making across various sectors. The literature review explores key research papers in the area of Deep Learning and Big Data.

LeCun et al.’s study plays a foundational role in understanding the significance of Deep learning in the context of Big Data Analytics. This paper focuses on deep neural networks, that revel intricate patterns hidden within enormous data sets [1]. The author argues that traditional machine learning techniques regularly have difficulties extracting meaningful insights from data due to their limited architecture and ability to capture hierarchical representations.[1]. In comparison, deep neural networks are known for their multi layered architectures, which is valuable when dealing with Big Data where data sets are large, diverse and may contain many features. Deep learning models excel at automatically identifying relevant features and patterns within these datasets, this reduces the need for manual feature engineering, which can be time consuming and subject to errors [1] LeCun et al.’s research highlights the theoretical foundations of deep learning, such as concepts like backpropagating, enabling deep learning models to adapt and improve over time, making them suited to Big Data. They make a significant contribution to areas such as healthcare diagnosis, financial fraud detection, environmental monitoring ,and many more sectors, by their ability to extract valuable insights automatically from massive datasets[1].

Chen, Mao, and Liu’s (2014) paper “Big data: A survey” provides a detailed overview of Big Data. This paperdefines big data and discusses its fundamental characteristics, with an emphasis on the three V’s ( volume, velocity, and variety [2]The author also reaffirms the significance of big data across different domains highlight its transformative potential. The paper also addresses several challenges associated with Big Data, such as data security, data privacy, and scalability. It discusses the role of cloud computing and distributed systems in managing and processing large datasets effectively. The paper is a valuable resource for gaining a deeper understanding Big Data and its wider applications to various industries and sectors[2].

# Theoretical Underpinnings (1036 words at this point)

The mathematical underpinning of Deep Learning, architectural concepts and handling of Big Data are discussed bellow:

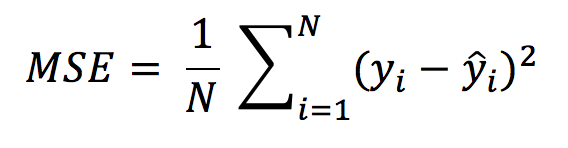
## Mathematical Underpinnings of Deep Learning

Deep Learning is dependent upon a solid mathematical foundation in order to effectively train neural networks effectively and efficiently.

## Cost Functions and Optimization

Cost Functions also referred to as loss functions are at the centre of Deep Learning, they help to quantify discrepancies between predicted values and the actual values or truth [1]. Mean squared Error is commonly used in regression tasks, as shown below:

Insert equiation



Here, represents actual values and Υ represents the predicted values for data point i. [7] In Classification tasks, cross entropy is frequent:

Insert equation

Here, represents binary class labels (0 or 1), and Yi represents predicted probabilities [7]

Gradient decent is an optimization algorithm, it is used to minimize cost functions during the training of neural networks [1]. It computes gradients and updates them in the opposite direction to minimize the cost, iteratively adjusting model parameters ( weights and biases)[5]. The gradient decent rule is:

Insert equation

Where W represents weights, α is the learning rate, C is the gradient of the cost function [7].

## 3.2 Deep learning Architecture

Deep Learning involves different architectural ideas, each with it mathematical and scientific basis.

## 3.2.1 Deep Feedforward Networks

In deep feedforward networks, layers consisting of neurons apply weighted transformations to input data.[1]. Below is the mathematical expression of the output of the neuron in a feedforward network.

Insert equation

Here, aj is the output of the neuron j,xi, iis the input from the neuron I,w,j,I represents weights, bj is the bias term, and o is the activation function [7].

## 3.2.2 Convolutional Neural Networks (CNNs)

CNNs are especially designed for gris like data such as images [5]. They involve convolutional layers that act like filters to detect features in input data [5].

Insert equation

Here, s is the output, I is the input, K is the kernel (filter), and (I,j) represent the pixel coordinates [5]

## 3.2.3 Recurrent Neural Networks (RNNs)

RNNs are specifically designed for sequential data and employ recurrent connenction in oreder to capture temporal dependencies [5]. The hiddem state ht of an RNN at the time t can be expressed as:

Insert equation

Here, xt is the input at time t, whx and whh are weights matrices, bh is the bias, and o is the activation function [5].

# Integration of Deep Learning and Big Data

In this section, we discuss the architectural considerations, challenges, limitations and strategies associated with integrating Deep Learning and Big Data.

## Architectural considerations

In order to gain the maximum benefits of Deep Learning and big data, careful architectural planning is necessary.

## 4.1.1 Distributed Deep Learning

Deep Learning models can be trained on distributed computational clusters in order to accommodate large scale datasets[6]. TensorFlow’s Architecture allows models to be updated across multiple nodes, this allows for faster training and scalability [5].

## 4.1.2 Data Pipelines

Deep learning models require effective data pipelines for feeding in data [6]. These pipelines often include data cleaning, feature extraction, and transformation stages[6]

## 4.1.3 Scalable training

When integrating Deep Learning with Big Data its important to consider scalability [6].Parallelism techniques, such as model and data parallelism, enable the training of large models on distributed clusters [6].

## 4..2 Challenges and limitations

There are may advantages gained from the integration of Deep Learning and Bog Data, however, it also presents some challenges and limitations.

## 4.2.1 Data volume and complexity

Big Data by name, involves massive volumes of data, both unstructured and semi structured data [6]. Handling data of this magnitude can put immense strain on storage and processing resources [6[.

## 4.2.2 Model scalability

Scaling up Deep Learning models to meet Big Data Requirements requires considerable computational resources. Larger models need extra resources which may not readily be available[6].

## 4.2.3 Resource constraints

Organisations may face considerable limitations in terms of hardware infrastructure and computational resource, this directly effects their ability to implement distributed Deep learning effectively [6].

## 4.3 **Strategies for mitigating challenges**

In order to address the challenges of integrating Big Data and Deep learning mentioned previously, various strategies can be employed.

## 4.3.1 Data Partitioning

Data Partitioning can be employed to Large datasets distributed across clusters to improve data access and processing speed[6]. HDFS, for example, uses data partitioning to distribute and replicate data blocks [6].

## 4.3.2 Parallelism techniques

Parallelism techniques such as model and data parallelism enable distributed Deep Learning across multiple nodes [6]. Frameworks like Apache Spark and Tensor Flow facilitate parallel processing [7].

## 4.3.2 Model optimisation

Model optimization techniques such as pruning, quantization, and compression, help to reduce the demand for resources for Deep Learning models[5]. These techniques work to reduce its size while aiming to maintain model performance.

# Technical Example

To demonstrate the applications of Neural Networks for classifying forest cover based on ‘covertype.csv’ dataset.

The steps involved are listed below:

1. Data Loading

# Implications and limitations of Deep Learning in Big Data 21/9

In this section, the implications and limitations of Deep Learning and Big Data are discussed.

* 1. implications of Deep Learning in Big Data
  2. Limitations and challenges
  3. Research Gaps and future Directions

# Critical Evaluation22/9 need a ref for limitations chellenges

In this section, a critical evaluation of the finding of this study in the context of Deep Learning in Big Data is discussed. This research study informs on the integreation of Deep Learning with with Big Data has led to advancements in different domains.including natural language processing, image recognition, and environmental monitoring [1][6]. Deep learning models, such as CNNs and RNNs have shown (limited reliability in certain contexts, accuracy and overfitting, high computational demand, lack of interpretability , data dependency discuss this here

# Conclusion23/9

# References

Look up three forestry papers for real life technical example/demonstration

Make a content list and improve abstract

Needs to be 5-6000 words

Add page numbers before updating contents table

Check if extension allowed

Discuss research gap further- in more detail