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.**
7. **Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Le, Q.V.,& Ng, A.Y.(2012). Large scale distributed deep networks. In advances in neural information processing systems (pp.1223-1231).**
8. **Silver, D.,Huang, A.,Maddison,C.,J.,Guez, A.,Sifre,L.,Van Den Driessche, G.,& Hassabis, D.(2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587).484-489.**
9. **Wang, H.,Huang,L.,Wu,D.J.,& Zhang,K.(2017). Deep reinfecment learning for imitating human driving skills. In 2017 IEEE intelligent Vehicles symposium(IV)(pp.1010-1015).IEEE.**
10. **Konečný, J., McMahan, H. B., Ramage, D., & Richtárik, P. (2016). Federated optimization: Distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610.02527.**
11. **Equations and Maths check these references not sure is referenced correct**
12. **D.chats, “ Parelell computing and petscale comuting” in parallel computing for Data Science, pp.1-10, 2016.**
13. Nielsen, M.(2015). Neural Networks and Deep Learning: Determination Press.

Need more references on mathematical foundation and equatios

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** **and Literature Review**

The integrations of Deep Learning techniques with Big Data has revolutionised analytics, and data driven decision making across various sectors. This section provides an explorations of the key research papers and studies, offering insights into the foundational principles of Deep Learning and the latest advancement in this area. Thes insights are necessary in understanding the significance, challenges and transformative potential of integrating Deep Learning with Big Data.

Deep Learning has rapidly become the pinnacle of data analytics due to its ability to highlight intricate patterns in Big Datasets[1]. Neural network architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), play a central role .CNNs have revolutionised image analysis , recognition tasks, allowing automatic extraction of desired features from diverse data [1][5]. In comparison, RNN’s excel in handling sequential data, like speech recognition, even though they have challenges around long ranger dependencies [5].Transformer models such as BERT and GPT-3 have advanced natural language understanding by their ability to capture relationships in text [5].

In order to manage Big Data characteristics, a number of technologies exist such as distributed file systems, parallel processing frameworks, and other data storage solutions [6][7]. Some examples of distributed file systems include Hadoop and Google File [6]. Parelell processing frameworks like Apache Hadoop and Apache Spark, facilitate the distribution of Big Data tasks, which enables more efficient execution across clusters[6][7]. There are other data storage solutions like Apache Cassandra and Hbase that are also good at handling and retrieving large scale data[6].

Need to add something on gradient decent and cost functions check google scholor ref

Goodfellow

Ruder,S.(2016). An overview of gradient decent optimization algorithms.arXiv preprint arXiv:1609.04747

Zhang

Kingma, D.P., &Ba, J.(2014). Adam: A method for stochastic optimization. asXiv preprint ar Xiv: 1412.6980

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.

Recent advancements in Deep Learning have focused on specialized neural network architectures that are customed to specific data domains. These advances have the potential to leverage industries and give them the competitive edge. Transfer learning, for example, enables models to adapt and optimise their performance on specific data domains, reducing the need for manual feature engineering. [1]. Additionally, BigDL, a distributed Deep Learning framework, has been tailored explicitly for Big Data environments with a focus on scalability and efficiency [1]. 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]. Federated Learning

Federated Learning

Talk about this paper:

1. **Konečný, J., McMahan, H. B., Ramage, D., & Richtárik, P. (2016). Federated optimization: Distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610.02527.**

Talk about Deep reinforcement learning

1. **Silver, D.,Huang, A.,Maddison,C.,J.,Guez, A.,Sifre,L.,Van Den Driessche, G.,& Hassabis, D.(2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587).484-489.**

To facilitate modelling at scale, tools such as Apache Aspark, Apach Hadoopd, TensorFlow, and PyTorch all aid in modelling at scale and data processing. Apache Spark works to reduce data movement between disk and memory. It al also supports a range of machine elarning libraires and data transformation [6].

Add this:

Zaharia et al.’s paper, “ Apache Spark : A Unified Analytics Engine for Bog Data processing”[6], introduces Apache Spark as a versatile tool for processing large scale data. It addresses the challenges associated with Big Data analytics by presenting an engine capable of handling diverse data processing tasks. The paper outlines Apache Spark’s architecture, including its components like Spark Core, Spark SQL, Spark Sreaming, MLlib, and GraphX. It emphasises Spark’s ability to perform in memory processing, resulting in significant performance improvements. The author provides real worls experience where Apache Spark’s has been transformative in the area, such as large scale machine learning, graph processing, and streaming amalytics. These practical examples highlight the broad applicability of Spark’s across various domains.

**Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1). MIT press Cambridge. This is a book so maybe look up academic paper instead maybe leCun??**

1. **Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Le, Q.V.,& Ng, A.Y.(2012). Large scale distributed deep networks. In advances in neural information processing systems (pp.1223-1231).**

Notes on academic papers

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 paper defines 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].

In Chen, Zhang, and Sun’s (2014) paper, titled “Big Data Driven Innovations” discusses the role of Big Data in fostering innovation. The author recognises how the emergence of Bog Data has revolutionised the way business and organisations operate. The paper highlights the significance of data driven decision making and how Big Data analytics can lead to innovative solutions and improvements to business processes. It highlights real world examples of different companies that have benefitted from Big Data analytics to gain a competitive edge. The paper also discusses challenges related to data quality, privacy, and security, stressing the need for proper data management practices. The paper provides an informative overview to the transformative power of Big Data in driving innovation and improving business operations across various industries.

Schmidhuber’s (2015) paper “Deep Learning in Neural Networks: An Overview”, provides an overview of Deep Learning in neural networks [3]. It explores the fundamental concepts and techniques employed in data in deep learning, including neural network architectures, training methods, and applications. Schmidhuber’s work highlights the advancements in deep learning and its potential to extract patterns from large datasets. This paper serves as comprehensive introduction to the field of deep learning and is more focused on presenting the positive aspects and capabilities of deep neural networks.

Zaharia et al.’s paper, “ Apache Spark : A Unified Analytics Engine for Bog Data processing”[6], introduces Apache Spark as a versatile tool for processing large scale data. It addresses the challenges associated with Big Data analytics by presenting an engine capable of handling diverse data processing tasks. The paper outlines Apache Spark’s architecture, including its components like Spark Core, Spark SQL, Spark Sreaming, MLlib, and GraphX. It emphasises Spark’s ability to perform in memory processing, resulting in significant performance improvements. The author provides real worls experience where Apache Spark’s has been transformative in the area, such as large scale machine learning, graph processing, and streaming amalytics. These practical examples highlight the broad applicability of Spark’s across various domains.

Dean et al.’s paper, “Large Scale Distributed Deep Networks”[7] adrresess the challenges of effectivlet processing Deep Learning algorithms on distributed Big Data platforms. The author discusses scalable Deep Learning techniques and the use of TensorFlow to tackle these issues.

Silver et al.’s paper mastering the game of go with Deep Learning NN and tree search demonstrates the potential of deep reinforcement learning in creating intelligent, autonomous systems capable of making informed decisions in dynamic environments. The study highlights how dep neural networks and tree search increasing performance in playing the complex board game of Go. The author discusses the importance of advanced machine learning techniques in solving complex problems and its broader application for developing AI systems for informed decision making in real world scenarios

# Theoretical Foundations

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

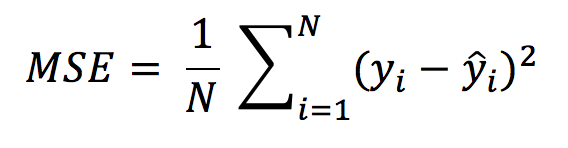
## Mathematical Underpinnings of Deep Learning

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