# 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 Processing and Storage |
| Assessment Title: | Integrated CA 1 SEM 2 MSc Data Analytics |
| Lecturer Name: | David McQuaid  Muhammad Iqbal |
| Student Full Name: | Umar Iqbal |
| Student Number: | 2023463 |
| Assessment Due Date: | 5 April 2024 |
| Date of Submission: | 7th April 2024 |

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.

Innovative Applications of Deep Learning and Big Data in Healthcare

***Abstract-* With advancement in technology and medicine, the healthcare facilities have improved a lot. The key player in the domain is data, which not only provide insights but also forecast the upcoming trends. The enormous amount of data which is generated each day in healthcare and other fields is so massive that it is impossible for traditional technology to evaluate and study. These type of data that falls in the category of Big Data is processed differently by the data scientists which create advanced algorithms to study it. These specialized algorithms are too complex that they behave like human brain, the scientists have created deep learning models which are known as neural networks which behave like neurons and make decisions appropriately. The intersection of both these fields provide a path for future developments that will not only revolutionize the healthcare but will change the core in all other domains. This paper provides an overview, applications, challenges and future aspects of Deep Learning and Big Data in Healthcare. The data storage management system, the architecture of distributed file system, big data stack and limitations of legacy approaches to big data are evaluated. The various different types of neural networks and set of requirements to determine the type of neural network are also discussed.**

***Keywords- Machine Learning, Deep Learning, Neural Networks, Big Data, Data Analytics, Medicine, Healthcare***

I. INTRODUCTION

In this decade, the healthcare industry has witnessed a drastic change driven by the advancements in technology especially in the areas of Artificial Intelligence and Machine Learning. Deep learning which is subset of Machine Learning and Artificial Intelligence has emerged as a powerful tool for extracting meaningful insights from large and complex databases. Similarly, Big Data, characterized by the massive volume, velocity and variety of healthcare data, presents both challenges and opportunities for transforming healthcare in a great way.

The collaboration of Deep Learning along with Big Data and healthcare with change the various aspects of healthcare including diagnosis of the disease,

treatment therapy, preventive measures, medicine and surgical intervention through bots and AI. The management sector will also be positively impacted by the data driven decisions. The medical sector generates many different categories of data ranging from structured, semi-structured and unstructured. The diagnosis facilities in the healthcare include studying images, graphs and visualizations. The mobile devices and wearable gadgets help keep record of vitals and generate daily logs incorporating huge amount of data which lays the foundation for studying the trends, recognizing patterns and make predictions via modelling which is the core of advanced data analytics.

II. BIG DATA

*A. Understanding the term*

The term Big Data is coined to demonstrate the enormous data which cannot be studied by the traditional data studying tools. To understand the complexity and magnitude of big data, the five core fundamentals volume, velocity, variety, veracity and variability plays an important role. The **Volume** represents the huge amount of data which is generated every single day. The media, sensors and devices creates terabytes of data which needs robust infrastructure to study. The **Velocity** at which the data flows also plays a crucial role in decision making. The transactions, streaming devices, online interactions generate a huge flow which needs to be studied for the insights.

The images, audio files, videos, spreadsheets and text files contribute to the **variety** of data that needs to be processed. This create a challenge for the data analysts to understand, manage and modify the data. The **Variability** refers to the fluctuations in the data generated over the time. The inconsistent nature of the data poses a problem for accurate predictions which needs to be addressed. In the vast sea of data, it is mandatory to check the quality of data being created which constitutes the **veracity** factor. It pertains to accuracy and reliability of data.

*B. Limitations of Legacy approaches to Big Data*

The traditional relational databases and on premises data warehouses have several limitations which include scalability, cost, performance, variety, security, processing speed and flexibility. The exponential growth of data needs horizontal scaling, but as the volume of the data increases, the capacity constraint becomes an issue. The upfront investment in hardware, maintenance cost of scaling the infrastructure to handle big data is expensive as compared to the cloud based services which have pay as you use price models. The traditional systems were optimized for structured data which pose a problem for different variety of data like images, texts and other semi structured or unstructured data. These data types are not supported by the relational databases. Processing these type of data require complex modeling and transformation. Another issue with legacy systems is it could not handle the velocity at which the big data is generated and processed. These systems were not able to handle real time analytics and decision making. The storage and modification of existing data also cause the lack in processing due to the time consuming, inflexible and rigid traditional systems. The system also lack the robust security features. Data stored in on premise centers may be susceptible to breaching and data loss.

*C. Big Data Storage and Management*

There are various processes and technologies involved in analyzing big data. The storage system has to handle enormous amount of data which traditional systems struggle to organize. Distributed storage systems like Hadoop, Apache Cassandra, Apache HBase, Google Cloud Storage and Azure Blob are designed to scale data across multiple servers and reduce fault tolerance. Technologies like Apache Spark, Flink and Kafka provide parallel processing capabilities to handle big data and process it. Machine learning and real time stream processing are some of the core functionalities. Cloud computing and containerization provides scalability by ensuring elasticity and dynamic resource allocation. Data management also includes managing the lifecycle of data by creating guidelines for data retention, deletion and archival. The utilization of storage resources is core of data lifecycle management. The metadata includes information about the structure, lineage and source. For better understanding, search, discovery and performance effective metadata management is necessary. The reason for data storage and management is to study the data to develop useful insights. Various dashboards like Tableau and Power BI are used to visualize, compare, analyze the data and generate reports. Data management and storage is a complex aspect of dealing with big data. Data Warehouses like Amazon Redshift, snowflake and Google Big Query are specialized systems designed to assist big data storage and management.

*D. Architecture*

Distributed file system (DFS) architecture is designed to store large amount of data across multiple nodes. This DFS is stored in a distributed computing environment. The typical architecture of a distributed file system includes Name node, Data node and Client node. The name node is responsible for storing the metadata about the file. It keeps track of which blocks constitute each file and their locations in the cluster. The actual data is stored in the data nodes which are known as the worker nodes. Each data node is responsible for serving read and write requests on its disk. The applications which interact with the distributed file system to read, write and manipulate data is referred as the client node. Client node send operational commands to the name node, which then coordinate with data node to perform the task. The client head may also store data in cache to improve performance and reduce traffic. The data is stored in blocks of fixed size typically 128 MB and there are multiple replicas of each block across the system to ensure there is no data loss and this feature is called replication. It is fundamentally important to make system redundant and fault tolerance.

*E. Big Data Stack*

There are various layers and each layer has many different technologies which form the stack of big data. The layers along with technologies are discussed below:

1. Ingestion and Collection:

* Apache Kafka: It is a distributed stream-processing platform used for building real time data pipelines.
* Apache Flume: Flume is a log collection service for collecting, aggregating and moving data chunks efficiently.
* Apache Nifi: It is a data flow management tool used in the process of automation of data ingestion.

1. Data Storage:

* Hadoop: Hadoop stores large datasets across multiple servers in a cluster. It is distributed file system for big data.
* Apache HBase: It is a scalable NoSQL database which allow real time modifications like read and write in large datasets.
* Amazon Simple Storage Service: S3 is developed by Amazon Web Services AWS for storing objects. It is scalable service for storing and retrieving data as objects.

1. Processing and Analytics:

* Apache Spark: It is a distributed computing engine used to process large scale data. It has various in memory capabilities which makes the computation fast.
* Apache Flink: Flink is a framework used to process live data streams with low latency.
* Apache Hive: Hive is made on top of Hadoop. It uses SQL like queries to analyze datasets. It is a data warehouse infrastructure.

1. Querying and Indexing:

* Apache Hadoop MapReduce: It is programming model which is used for parallel processing of large data sets.
* Apache Impala: Impala is open source SQL query engine used for parallel processing of data store in Hadoop clusters.
* Elasticsearch: It is a search engine that has analyzing capabilities. It is distributed system which offers real time searching.

1. Visualization:

* Tableau: A popular data visualization tool which help users to create interactive dashboards and reports.
* Power BI: It is a business analytics tool developed by Microsoft. It allows sharing insights across the organization. It is also used to visualize the data and generate reports.
* Apache Superset: It is an open source platform used for exploration and visualization. It supports various data sources and generates interactive charts.

1. Machine Learning and AI:

* Tensor Flow: It is developed by Google for building and training neural networks. An open source framework for machine learning and building models.
* Apache Mahout: It is an open source Hadoop library used for scalable algorithms.
* PyTorch: It is developed by Facebook for deep learning. This framework uses python programming language for computation.

1. Governance and Security:

* Apache Ranger: Ranger is a Hadoop framework for security and access control. It also offers centralized administration.
* Apache Atlas: It is used for managing and cataloging metadata. Atlas is management and governance platform.
* Apache Sentry: It is used with Hadoop for authorization and access to Meta data and data on cluster.

III. DEEP LEARNING

*A. Understanding neural networks*

The data analytics and machine learning focus on the detailed and comprehensive study of data which is possible using deep neural networks. The deep learning which is subset of machine learning, is inspired by the functioning of human brain, with layers of interconnected nodes which act as neurons to process the data. Deep neural networks have multiple hidden layers along with input layer and the output layer. Each node has its own weight, threshold and is connected to another node. The weight and threshold are continually adjusted during training to get the similar output as training data. If the output of any node is more than its threshold, the node send the data to another node thus activating the next layer. Neural networks learn by training the data and their accuracy improves over time. They allow us to classify and cluster data at high velocity. Neural networks are powerful tool in artificial intelligence domain.

*B. Applications*

The Computer Vision, speech technology, natural language processing, recommendation systems are some of the applications of the Deep Learning. Object orientation, image classification, facial recognition and segmentation are few examples where neural networks are used in the domain of image processing and computer vision. The security surveillance intelligence systems are also designed based on neural networks. Neural networks are the core of natural language preprocessing which include the tasks such as language translation, text summarization and AI chatbots. Virtual assistants like Apple Siri and Alexa recommends the content using neural networks. The speech to text feature, voice dictation systems, voice security systems all works on the speech recognition technology.

The recommendation systems used by Amazon, Netflix, YouTube and Social media apps like Instagram for suggesting movies, content, posts and music are powered by neural networks which work on real time data. Autonomous vehicles perceives, interprets the environment and collect data which is then analyzed by deep neural networks for object detection, path planning and various other decision making scenarios. Robotics use deep learning algorithms for service bots, industry bots, drones and humanoid bots. They navigate and learn by experiencing their surroundings. The healthcare sector rely on neural networks for disease prediction, medical image analysis like MRI and CT scans. Finance sector also use neural networks for predicting and identifying patterns, fraud detection, risk analysis, credit scoring, algorithmic trading and financial forecasting.

*C. Different types of Neural Networks*

The keys models and architectures of deep learning are Feedforward neural network (FNN), which is also known as multilayer perceptron (MLP) which are used for regression, function approximation and classification. Pattern recognition, computer vision, natural language processing all rely on feedforward neural networks. In this type of network, the information flows only in one direction. There are no cycles or loops. The CNN Convolutional neural networks are used for processing images and grid like data structures. These include convolutional layers, pooling layers and connected layers. Convolutional layer act as a filter for the input data. CNNs use parameter sharing, which in turn reduces the number of parameters in the network.

The Recurrent neural network (RNN) are majorly used for speech technology, time series prediction, sequencing problems and gradient issues. RNNs have connections which make loops or cycles which in turn help them exhibit dynamic behavior. They could also maintain an internal memory which is used for processing of new inputs. The intersection of reinforcement learning and deep learning lay the pavement for the deep reinforcement learning, DRL. Reinforcement learning algorithms learn by interacting with the environment to maximize the rewards. DRL has various applications including games, robotics and autonomous driving.

*D. How to determine the model*

There are various factors which are responsible for determining the type of neural network to use. The nature of the data, specific task or aim which user wants to achieve, resources available and the experience level. First step is to understand the data and its dependencies. The size of the data, the format whether the data is structured or unstructured, dimensionality of the input features are important factors to consider before moving further. The next step is to define the problem statement and considering whether it involves spatial relationships or dependencies. The main step is to balance between the model complexity and the computational resources available. Complex models need large amount of data and such models take longer training times.

The parameters are the critical part of determining the model. The model which has the highest accuracy, precision and F1 score is preferred for classification tasks and for regression, the models which have the lowest mean squared error are preferred. Starting with baseline model and evaluating its performance and then sub sequentially experimenting with different parameters and optimization algorithms. Techniques like cross validation and fine tuning hyper parameters based on the experiment results are performed to find the best fit model. If the data or the resources are limited, then transfer learning techniques are considered.

*E. Challenges*

AI technologies should be developed in a responsible and ethical manner. The biases present in the training data could lead to unfair outcomes. It is important to mitigate bias and ensure fairness to prevent harm to society. Protecting individual’ rights and privacy are also important when data is studied. Neural networks are also known as black boxes because it is difficult to understand how they arrive at decision. Ensuring transparency is important to gain trust and accountability. The neural networks must be designed keeping human centered values in mind. Adopting sustainable AI is important to mitigate environmental impact.

Cybersecurity attacks on AI like hacking, data breaches and adversarial attacks are another serious challenge which need to be taken care of. It is mandatory to protect AI systems from malicious activities through robust security measures and protocols. It is important to adopt a precautionary approach with regards to individual rights, privacy, society values and public trust to address the legal and ethical challenges associated with AI and neural networks.

IV. HEALTHCARE DOMAIN

*A. Predictive Analytics*

Big Data and Deep learning algorithms helps analyze vast amount of healthcare data which helps prevent disease outbreaks, diagnosis of disease and treatment outcomes. Using AI in healthcare sector, helps lower the fatality rate by early detection of cancer in patients. Deep learning can predict the disease based on the test results data and could further assist prescribe more clinical tests to navigate the issue.

*B. Medical Imaging*

Images generated through X-Rays, CT scans and MRIs could be studied more efficiently by the deep learning models which reduces the chance of error very effectively. These models have high accuracy which can help radiologists to generate reports faster and more precise. This process reduces human error and speeds up the diagnosis.

*C. Personalized Medicine*

Advanced data analytics combine genomic data, lifestyle information and health records to train the neural networks which in turn can develop personalized treatment plans which cater to individual needs. This could help in analyzing genetic makeup and health history for customized medicine and diet plans.

*D. Drug Discovery*

The drug creation process is very complex which requires lots of precise monitoring and studying the effects of medicine on human body and their ability to treat the underlying condition. AI could help study the vast data to create potential drugs as well as predict their efficiency and side effects on human body without the trials on animals. This accelerates the development process and reduce costs.

*E. Remote Monitoring*

The sensors which collect the real time data and the IoT devices which create continuous flow of health data both constitute the big data which could be analyzed in real time to monitor patients remotely. This could detect abnormalities and alerts the user to take necessary actions and precautions which could potentially save lives.

*F. Risk Assessment*

AI can study risk factors like genetic buildup, lifestyle choices including the diet and exercise and environmental factors to predict the risk of developing certain disease. This can assist in taking preventing measures. An early precautionary detection can actually prevent the disease from happening. An example f this could be diabetes risk assessment based on multiple factors.

*G. Fraud Detection*

AI can be used to identify the frauds prevailing billing and insurance. This help the healthcare sector by preventing financial losses. The insurance companies can analyzes the data to check whether there were any disease or ailment at the time of taking insurance. Fake billing claims to loot money from insurance providers is a big scam in some nations.

*H. Virtual Heath Assistant*

AI powered Chat bots and virtual assistants in health care applications can guide people based on the data and neural models. They can make recommendations, answer basic queries and help schedule appointments. The remote consultation is also possible through the means of video calling which in turn improve access to healthcare facilities.

*I. Population Heath Management*

The analysis of big data can identify trends in the population data which could save the outspread of diseases. AI can allocate resources accordingly and help develop emergency plans based on the sequencing problems to counter any outburst of disease. This in turn can improve overall healthcare infrastructure by providing the insights in pandemic.

*J. Clinical Decision Support System*

CDSS can assist doctors and medical professionals in terms of developing treatment plans. These evidence based recommendation system can help make decision in diagnoses, treatment and prescriptions. Various situations tests the analyzing capabilities of the medical staff, the AI based decisions can help withstand the pressure and reduce human error.

*K. Surgical Assistance*

The decisions made in the operation theatre are very critical for the patients. The AI based monitoring could save lives by alerting the surgeons about the drastic changes in the vitals. Moreover the surgeries using AI bots are more accurate and precise compared to the in hand surgeries which can be invasive. Many specialized bots function on the deep neural network models and perform noninvasive surgeries.

V. LITERATURE REVIEW

According to the research conducted by Kornelia Batko [1], medical facilities are working on both structured and unstructured data, which comes from databases, transactions, unstructured content of emails and documents, devices and sensors. They use data analytics in administrative, business as well as clinical area. The decisions made are largely data driven. Medical facilities are moving towards data based healthcare and its benefits.

The prediction models help in accurate prediction of Alzheimer, tumors in PET or MRI images using machine learning and big data analytics. Cluster based mechanism is recommended in research paper by Sulaiman Khan [2] for organization purposes to improve big data timely access and easy management capabilities. The results suggest advanced hybrid machine learning and cloud based models should be adapted to reduce treatment cost and improve quality of care.

Cost optimization and information retrieval are major issues of healthcare industry as stated by T.Ramesh [3] in his work. SVM, decision trees and other data mining techniques are used to address the issues. According to insights, health care industry is leading sector where huge revenue is generated as the number of patients increase dramatically with increase in population.

Bioinformatics, medical imaging, sensor informatics, medical informatics, computational biomedicine, electronic health record data and experimental data all contribute to the complex heterogeneous data, which needs integrating, examining and analyzing. This complex data sets are studied using Apache Hadoop Map Reduce because it handle heterogeneous data sets and uses memory and other resources in a more efficient way to reveal hidden patterns and novel knowledge at a great execution speed.[4]

The use of CNN on raw DNA sequence successfully detected relevant patterns. A deep architecture based on CNNs predicts RNA and DNA binding proteins. CNNs were used to predict chromatin marks from DNA sequence. Basset, an open source framework predict DNase I hypersensitivity across different cell types. Sparse AEs were used to classify cancer cases from gene expression profiles. [5]

Neural networks can be applied on all levels of heath care decision making. The hybrid models of neural networks are used to tailor solutions to given problem. [6] ANN based solutions applied on macro and meso level of decision making scenarios ensures that its use in solving complex, unstructured and limited information. The understanding of ethical, societal and economic implications of applying neural networks in healthcare will lead to successful adoption of ANN in organizations.

The deep learning based techniques are powerful tools in dealing with disease detection in preprocessing, feature extraction, feature selection, classification and clustering steps. [7] The hybrid and ensemble methods based on DL yields better results than single models. The process of combination of two or more process is independent of the type of the datasets. As we know deep learning models are memory and time consuming, the designing and implementation of optimized models in healthcare systems is a big challenge.

The current advancements in deep learning in healthcare aim to aid physician in disease detection. The challenges to the current state of art is increasing quality, speed and precision. These challenges can be solved by making improvements in data collection and dissemination. [8] Future applications should aim to redefine patient care using AI.

VI. FUTURE ASPECTS

The research is continuously evolving the field of Deep learning. The transfer learning and multi-model techniques are the future of the Artificial Intelligence and its applications. The intersection of healthcare and deep learning will solve complex problems in the domain. The human machine interactions will improve exponentially due to advancement in natural language processing algorithms. The robustness and security features will decide the adoption of more AI based solutions in every sector.

This is a lifelong process of continual learning, with more data being generated over time, the insights will get more sophisticated. The use of bots will positively impact the market, reducing the costs and human errors. With advancement in data processing power, real time insights, predictive models and recommendation systems will evolve efficiently. The future of AI holds tremendous potential. This will shape the future of industrial revolution. The scientific research and discoveries will change the way we interact with technology in the future.

VII. CONCLUSION

We discussed the notion of big data, the limitations of the legacy systems in handling big data, data storage and management system, architecture and big data technology stack in great depth. The deep learning neural networks, their applications, types, use and challenges were also presented. The applications of big data and deep learning in health care sector were illustrated. The literature review presented the results and observations from various research papers in the field of deep learning, big data analytics and healthcare. The future aspects of big data along with AI lead to the conclusion that more robust, reliable and human centric values based deep learning algorithms will lay the pavement for future development.

The healthcare sector has enormous potential for deep learning and artificial intelligence to evolve. With data based decisions, the healthcare domain will improve gradually. All other sectors also have potential for improvement using data centric approaches.

VIII. REFERENCES

1. Kornelia Batko and Andrzej Slezak, “The use of Big Data Analytics in healthcare”, Journal of Big Data, 2022.
2. Sulaiman khan, Habib Ullah khan and Shah Nazir, “Systematic analysis of healthcare big data analytics for efficient care and disease diagnosing”, Scientific Reports, 2022.
3. T. Ramesh and V. Santhi, “Exploring big data analytics in health care”, International Journal of Intelligent Networks, Vol. I, 2020, pp.135-140.
4. Maria Mohammad Yousef, “Big Data Analytics in Heath Care: A Review Paper”, IJCSIT, Vol.13, No 2, April 2021.
5. Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang and Joel T Dudley, “Deep learning for healthcare: review, opportunities and challenges”, Brief Bioinform, Nov 2018.
6. Nida Shahid, Tim Rappon, Whitney Berta, “Applications of artificial neural networks in health care organizational decision making: A scoping review”, PLOS One, 2019.
7. Shahab Shamshirband, Mahdis Fathi, Abdollah Dehzangi, Anthony Theodore Chronopoulos and Hamid Alinejad Rokny, “A review on deep learning approaches in healthcare systems: Taxonomies, challenges and open issues”, Journal of Biomedical Informatics, Vol 113, Jan 2021.
8. Hafsa Habehh and Suril Gohel, “Machine Learning in Healthcare”, Current Genomics, Benthan Science Publishers Ltd. Dec 2021.
9. James Warren and Nathan Marz, “Big Data: Principles and best practices of scalable realtime data systems”, Manning Publications, April 2015
10. Martin Kleppmann, “Designing Data-Intensive Applications”, O’Reilly Media, Inc. March 2017.
11. Balamurugan Balusamy, Nandhini Abirami R, Seifedine Kadry and Amir H. Gandomi, “Big Data: Concepts, Technology, and Architecture”, Wiley, March 2021.
12. Ian Goodfellow, Yoshua Bengio and Aaron Courville, “Deep Learning”, MIT Press, Nov 2016.
13. Charu C. Aggarwal, “Neural Networks and Deep Learning”, Springer, 2018