**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **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:** | Maria Dominguez Alvarenga |
| **Student Number:** | 2019008 |
| **Assessment Due Date:** | 6th October 2023 |
| **Date of Submission:** | 6th October 2023 |

**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. |

Table of Contents

[Construction Materials for 8 countries of the Eurozone. Performance and comparison analysis. 2](#_Toc136028379)

[1. Abstract 3](#_Toc136028380)

**Deep Learingin using Big Data: Credit Card Fraud Detection System applying Artificial Neural Networks.**

Maria Dominguez  
 CCT College Dublin, Ireland  
2019008@student.cct.ie

**Abstract.**

Two high-focus topics in the data science domain are Deep Learning and Bid Data. Since Private and Public companies have been gathering huge amount of specific information that provides valuable data regarding issues such as marketing, fraud detection, medical issues among others. Considering the previous, since the e-commerce transactions have become more common and helped to the increase of credit card transaction, which could be either online or standard transactions. Due to this, those fraudulent credit card transactions represents significant losses for many companies worldwide every year. Therefore, in order to tackle this issue, the development of appropriate and effective credit card fraud detection models has become critical to counteract this raising issue and therefore, minimize those financial losses. However, since most credit card datasets are highly unbalanced as the amount of legitimate transactions are valid, this becomes a challenge. This academic paper proposes the use of Artificial Neural Network (ANN) to efficiently detect the fraud in credit card transactions, which has produced positive results.

**Introduction.**

As pointed out by Tang et al. [1] Big data is considered to come in volume, variety, velocity, variety and value. Also, big data has been formed as a result of the appear and accumulation of variety of data in unstructured and semi unstructured format. Nevertheless, as outlined by Jan et al. [2] different sizes of input data and formats, quick recovery of information, quality of data, data storage, etc are challenges that Big Data Analytics faces for machine learning and data analysis. Similarly, as stated by Gheisari et al. [3] as Big Data are extremely huge data sets that need to be analysed in order to find trends and patterns, is when Deep learning techniques can be used to find useful and abstract patterns within the data. As Jan et al. Therefore, as outlined by Jan et al. [2] Big Data Analytics main task is to be able to extract valuable patterns that are used to make decisions and predictions, which comes from a huge amount of data. Similarly as outlined by Vangumalli [3], it means that deep learning extract valuable data out from big data [3]. Since nowadays people are using credit card more often, this is the reason why credit card frauds are currently popular. Credit cards frauds can occur in any different organization such as banks, industry, automobile industry among others. Hence, this study is aimed to produce a deep learning model Artificial Neural Network to efficiently predict the fraud in Credit Card transactions, which was compared with the performance with Random Forest predictor for a classification problem.

**Keywords**: Big Data, Deep Learning, Credit Card, Artificial Neural Network, Machine Learning.

**Big Data in Data Analytics.**

**Deep Learning in Data Analytics.**

**Application of Deep Learning Artificial Neural Networks for the detection of Credit Card Fraud**

**Objective Statement:** Analyse the different case studies where deep learning has been used to credit card fraud, specifically Artificial Neural Networks, which is going to be compared with Random Forest machine learning models performed.

The main purpose of this academic paper is to portrait the use of big data analysis for prediction of credit card fraud, which is going to be used the Creditcard dataset, which is public domain. It is intended to use deep learning Artificial Neural Network algorithm as the amount of data,

**Literature Review:**

Different researchers have provided studies that aim to prevent that fraudsters could use the credit card information before any transaction was approved. As pointed out by Asha and Kumar case study [ ] the implementation of supervised Neural Networks was applied and compared with other machine learning models to measure their performance.

As same author goes on [Asha and Kumar, 2021 ] when the stealer uses the other person credit card without the authorization of such person and stealing crucial information as password, PIN and other personal credentials, is when the credit card fraud takes place. Owing to this, though the implementation of deep learning is when it can be identified whether the upcoming transaction is genuine or a result of fraud.

References

[1] Tang, L., Li, J., Du, H., Li, L., Wu, J. and Wang, S., 2022. Big data in forecasting research: a literature review. Big Data Research, 27, p.100289.

[2] Jan, B., Farman, H., Khan, M., Imran, M., Islam, I.U., Ahmad, A., Ali, S. and Jeon, G., 2019. Deep learning in big data analytics: a comparative study. Computers & Electrical Engineering, 75, pp.275-287.

[2] Gheisari, M., Wang, G. and Bhuiyan, M.Z.A., 2017, July. A survey on deep learning in big data. In *2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC)* (Vol. 2, pp. 173-180). IEEE.

[3] Vangumalli, D., (2017). Applications of Deep Learning (CNN) in Big Data Analytics. [online] Available at: <https://medium.com/@dinnuv/applications-of-deep-learning-in-big-data-analytics-c0a8d8b46c51> [Accessed 12 September 2023]

[5 ] Asha, R.B. and KR, S.K., 2021. Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, *2*(1), pp.35-41.

[6] Boulle, A., Chandramohan, D. and Weller, P., 2001. A case study of using artificial neural networks for classifying cause of death from verbal autopsy. International journal of epidemiology, 30(3), pp.515-520.

Kumar and Iqbal (2019) carried out a survey which portraits different techniques applied in the detection of MasterCard fraud through the application of machine learning models such as K-nearest Neighbour, Support Vector Machine, Neural Network among others to measure their performance with the metrics obtained.

A study carried out by Figuerola (2022) points out the issue with the fraud credit card datasets since there are highly unbalance due to genuine number of transactions recorded. As a result of this, the classifier is biased towards the legitimate transactions, which are the popular class which yields a reduced performance for the transactions produced by fraud.

Taken on board the use of imbalanced datasets, a method to identify credit card fraud was proposed in study carried out by Asha and Kumar (2021) using Neural Network deep learning, which was compared with Support Vector Machine and K-Nearest neighbour. The study shown that Artificial Neural Network produced an accuracy of 0.9992%, precision of 0.81% and recall of 0.76 %, which were followed by KNN 0.9982% and SVM 0.93% accuracy respectively. Thus, Ann was proved to be more suitable deep learning technique that tackles the issues produced by the use of imbalance dataset through the pre-processing of the data, normalization, followed by under-sampling of the same.

However, studied performed by Sohony et al. (2018) produced that Random Forest and Neural Networks algorithm produced a higher accuracy, when analysing a large dataset of genuine credit card transactions, where an ensemble learning methodology was proposed for credit card fraud detection as the ratio from fraud transactions to standard transactions is bit suitable.

Nevertheless, on another project conducted by Kumar et al. (2019) shown that Ramdom Forest supervised machine learning model which yield an accuracy of 90% when detecting credit card fraud through the decision tree for classification that is used by Random Forest classifier.

Conversely, on a different study conducted by Bin et al. (2022) states that random forest models are considered to be quite effective when it comes to predicting the class of regression problems, but it has been observed that on credit card fraud detection in real time could establish various limitations. This occurs as a consequence that random forest models perform well on datasets where limited data is available, but in real time scenarios have slower performance. Hence, when predicting credit card fraud detection in real time where a large volume of data is required, the random forest is lacking to effectively trained the dataset and make accurate predictions.

**Further study**

As stated above, a number of research have been conducted in regards with credit card fraud detection. However, there is the need of more efficient systems that can identify the transaction fraud before this takes place.

**References for literature review**

Kumar, P. and Iqbal, F., 2019, April. Credit card fraud identification using machine learning approaches. In *2019 1st International conference on innovations in information and communication technology (ICIICT)* (pp. 1-4). IEEE.

Figuerola Ullastres, E., 2022. *Credit Card Fraud Detection using Ensemble Learning Algorithms* (Doctoral dissertation, Dublin, National College of Ireland).

Asha, R.B. and KR, S.K., 2021. Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, *2*(1), pp.35-41.

Sohony, I., Pratap, R. and Nambiar, U., 2018, January. Ensemble learning for credit card fraud detection. In *Proceedings of the ACM India joint international conference on data science and management of data* (pp. 289-294).

Kumar, M.S., Soundarya, V., Kavitha, S., Keerthika, E.S. and Aswini, E., 2019, February. Credit card fraud detection using random forest algorithm. In *2019 3rd International Conference on Computing and Communications Technologies (ICCCT)* (pp. 149-153). IEEE.

Bin Sulaiman, R., Schetinin, V. and Sant, P., 2022. Review of machine learning approach on credit card fraud detection. *Human-Centric Intelligent Systems*, *2*(1-2), pp.55-68.

**Hadoop and Apache Stark structure for problem resolution.**

A critical factor depicted by Hatua et al. [25] is that the choosing of the appropriate platform to perform the tests is the main step in order to succeed in that experiment. Thus, the amount of that being used along with how to get the best results are the two main concerns to be considered. [26] Cloudlitics outlines Hadoop platform provides a complete distributed file system designed for storing and managing data across clusters of machines, which has been used for a since 2006. In contrast, Spark is considered a newer technology. [26] Cloudlitics stated that being both open sources frameworks designed for big data processing, Spark uses resilient distributed datasets (RDDs). In comparison, Hadoop processing data using MapReduce, where data can be stored across different machines since has a distributed file system (HDSF). [27] Diep states that in terms of performance, Hadoop boots overall performance by accessing the data stored locally on HDFS. Nevertheless, Hadoop can never pair Spark in terms of in-memory processing. [28] Apache Spark is considered easier to use and scale since its streamlined cluster operation and ability to leverage multiple coding languages like Scala, python and Java [Parmar, 2023] . In contrast, Hadoop essentially relies on expertise in MapReduce and Java. Adding to the fact that constant monitoring and maintenance is required for Hadoop due to its nodded architecture. As cited by Guan et al. [29] Hadoop is enjoying a vast market in medical treatment, e-commerce, finance as its great computing power, impressive compatibility as more data is stored at present on the cloud platform based on the Hadoop architecture.

As Hatua, [25] Hadoop platform using MapReduce was applied on study about Early detection of diabetic retinopathy from big data, which provided better results when processing that large datasets in comparison with some of the close competitive state of art techniques analysed in that study.

[30 ] Parmar (2023) outlines that pache Spark was developed to overcome the limitations of Hadoop and introduced the term Resilient Distributed Dataset [RDD] and process the entire data into the memory RAM, which makes Spark 100 faster than Hadoop. That were the reason why Apache Spark became very popular as is very fast and could handle a lot of data.

[25] Hatua, A., Subudhi, B.N., Veerakumar, T. and Ghosh, A., 2021. Early detection of diabetic retinopathy from big data in hadoop framework. *Displays*, *70*, p.102061.

[26] Cloudlitics (2023). Hadoop vs Spark: A Compartative Study. [online] Available at: <https://cloudlytics.com/hadoop-vs-spark-a-comparative-study/> [Accessed 22 September 2023]

[27] Diep, N., (2020). Big Data Analytics: Apache Spark vs. Apache Hadoop. [online] Available at: <https://towardsdatascience.com/big-data-analytics-apache-spark-vs-apache-hadoop-7cb77a7a9424> [Accessed 22 September 2023]

[28] Anderson, S., (2022). Hadoop (MapReduce) vs Apache Spark: A Deep Dive Comparison. [online] Available at: [https://streamsets.com/blog/hadoop-mapreduce-vs-apache-spark/ [Accessed](https://streamsets.com/blog/hadoop-mapreduce-vs-apache-spark/%20%5bAccessed) 22 September 2023]

[29] Guan, S., Zhang, C., Wang, Y. and Liu, W., 2023. Hadoop-based secure storage solution for big data in cloud computing environment. *Digital Communications and Networks*.

[30] Parmar, D., (2023). Learn Apache Spark in 10 Minutes, Step by Step Guide. [online] Available at: <https://www.youtube.com/watch?v=v_uodKAywXA&t=383s> [Accessed 27 September 2023]