ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΗΣ ΣΧΕΔΙΑΣΗΣ ΚΑΙ ΠΑΡΑΓΩΓΗΣ
http://www.eee.uniwa.gr
http://www.idpe.uniwa.gr
Θηβών 250, Αθήνα-Αιγάλεω 12241
Τηλ: +30 210 538-1614
Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών
Τεχνητή Νοημοσύνη και Βαθιά Μάθηση
https://aidl.uniwa.gr/

UNIVERSITY OF WEST ATTICA
FACULTY OF ENGINEERING
DEPARTMENT OF ELECTRICAL & ELECTRONICS ENGINEERING
DEPARTMENT OF INDUSTRIAL DESIGN AND PRODUCTION ENGINEERING
http://www.eee.uniwa.gr
http://www.idpe.uniwa.gr
250, Thivon Str., Athens, GR-12241, Greece
Tel: +30 210 538-1614
Master of Science in
Artificial Intelligence and Deep Learning
https://aidl.uniwa.gr/

Master of Science Thesis

Artificial Intelligence in Pharmaceutical Domain (with emphasis on the data quality)

Student: Oikonomidis Georgios
Registration Number: AIDL-0011

MSc Thesis Supervisor
Eleni Aikaterini Leligkou
Associate Professor

ATHENS-EGALEO, September 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΗΣ ΣΧΕΔΙΑΣΗΣ ΚΑΙ ΠΑΡΑΓΩΓΗΣ
http://www.eee.uniwa.gr
http://www.idpe.uniwa.gr
Θηβών 250, Αθήνα-Αιγάλεω 12241
Τηλ: +30 210 538-1614
Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών
Τεχνητή Νοημοσύνη και Βαθιά Μάθηση
https://aidl.uniwa.gr/

UNIVERSITY OF WEST ATTICA
FACULTY OF ENGINEERING
DEPARTMENT OF ELECTRICAL & ELECTRONICS ENGINEERING
DEPARTMENT OF INDUSTRIAL DESIGN AND PRODUCTION ENGINEERING
http://www.eee.uniwa.gr
http://www.idpe.uniwa.gr
250, Thivon Str., Athens, GR-12241, Greece
Tel: +30 210 538-1614
Master of Science in
Artificial Intelligence and Deep Learning
https://aidl.uniwa.gr/

Μεταπτυχιακή Διπλωματική Εργασία

Η εφαρμογή της τεχνητής νοημοσύνης στο φαρμακευτικό τομέα (με έμφαση στην ποιότητα των δεδομένων)

Φοιτητής: Γεώργιος Οικονομίδης
ΑΜ: AIDL-0011

Επιβλέπουσα Καθηγήτρια
Ελένη Αικατερίνη Λελίγκου
Αναπληρώτρια Καθηγήτρια

ΑΘΗΝΑ-ΑΙΓΑΛΕΩ, Σεπτέμβριος 2022

This MSc Thesis has been accepted, evaluated and graded by the following committee:

Supervisor

Eleni Aikaterini Leligkou

Digitally signed by Eleni Aikaterini Leligkou
Date: 2022.10.05 10:13:20 +03'00'

Helen C. Leligou
Associate Professor
Industrial Design and Production Engineering Dept.
University of West Attica

Member

ANTONIOS KARAGEORGOS
28.09.2022 11:03

Antonios Karageorgos
Professor
Department of Forestry, Wood Sciences and Design
University of Thessaly

Member

EFTHYMIOS LALLAS

EFTHYMIOS LALLAS
Sep 28 2022 1:24 PM

Efthimios Lallas
Assistant Professor
Department of Forestry, Wood Sciences and Design
University of Thessaly

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον/την συγγραφέα του και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις θέσεις του επιβλέποντος, της επιτροπής εξέτασης ή τις επίσημες θέσεις του Τμήματος και του Ιδρύματος.

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο/η κάτωθι υπογεγραμμένος Οικονομίδης Γεώργιος του Αριστείδη, με αριθμό μητρώου AIDL0011 μεταπτυχιακός φοιτητής του ΔΠΜΣ «Τεχνητή Νοημοσύνη και Βαθιά Μάθηση» του Τμήματος Ηλεκτρολόγων και Ηλεκτρονικών Μηχανικών και του Τμήματος Μηχανικών Βιομηχανικής Σχεδίασης και Παραγωγής, της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής,

δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της μεταπτυχιακής διπλωματικής εργασίας και κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος. Η εργασία δεν έχει κατατεθεί στο πλαίσιο των απαιτήσεων για τη λήψη άλλου τίτλου σπουδών ή επαγγελματικής πιστοποίησης πλην του παρόντος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του διπλώματός μου.

Επιθυμώ την απαγόρευση πρόσβασης στο πλήρες κείμενο της εργασίας μου μέχρι 30/10/2022 και έπειτα από αίτησή μου στη Βιβλιοθήκη και έγκριση της επιβλέπουσας καθηγήτριας.»

Ο Δηλών
Οικονομίδς Γεώργιος

Γ. Οικονομιδης

(Υπογραφή φοιτητή/ήτριας)

The opinions and the conclusions included in this document express solely the author and do not express the opinion of the MSc thesis supervisor or the examination committee or the formal position of the Department(s) or the University of West Attica.

Declaration of the author of this MSc thesis

I, Georgios Aristeidis Oikonomidis with the following student registration number: AIDL0011 postgraduate student of the MSc programme in "Artificial Intelligence and Deep Learning", which is organized by the Department of Electrical and Electronic Engineering and the Department of Industrial Design and Production Engineering of the Faculty of Engineering of the University of West Attica, hereby declare that:

I am the author of this MSc thesis and any help I may have received is clearly mentioned in the thesis. Additionally, all the sources I have used (e.g., to extract data, ideas, words or phrases) are cited with full reference to the corresponding authors, the publishing house or the journal; this also applies to the Internet sources that I have used. I also confirm that I have personally written this thesis and the intellectual property rights belong to myself and to the University of West Attica. This work has not been submitted for any other degree or professional qualification except as specified in it.

Any violations of my academic responsibilities, as stated above, constitutes substantial reason for the cancellation of the conferred MSc degree.

I wish to deny access to the full text of my MSc thesis until 30/10/2022 following my application to the Library of UNIWA and the approval from my supervisor.

The author
Georgios Oikonomidis

Γ. ΟΙΚΟΝΟΜΙΔΗΣ

(Signature)

This master thesis is dedicated to my father Aristeidis Oikonomidis

I would like to thank certain people whose contribution and support were valuable and decisive in the preparation of this thesis. I would first of all like to express my gratitude and appreciation to my professor supervisor, Ms. Helen C. -Leligou, for trusting me by assigning me this work. I also sincerely thank her for the knowledge she imparted to me during the lessons, the completion of the work and finally for her help and support since she was always available for any difficulties and obstacles that were presented during this period, which also marks our excellent cooperation.

I would also like to thank Dr. Dimitrios Karidas, Mr. Isaak Kasavidis, Prof. Antonios Karageorgos and Prof. Efthymios Lallas for our excellent cooperation in achieving the final results.

Furthermore, I would like to thank every professor in the post-graduate program Artificial Intelligence and Deep Learning.

Lastly, I would like to say a big thank to my family especially to my mother Evgenia.

Abstract

Over the past decades, computer science has been widely developed. From speech recognition to reconstructing brain circuits to even natural language understanding, one can easily claim that Machine Learning and Deep Learning have become part of everyday life (LeCun, et al., 2015). They have been used in most of the industries and the pharmaceutical one is no exception.

During the recent years, data integrity has been an important part of the pharmaceutical industry. It is mandatory to ensure the quality and the safety a drug has from the start of its production till it reaches the customer. In order to achieve this, the American Food and Drug Administration (FDA) introduced the ALCOA principles as guidelines for every pharmaceutical company. The ALCOA acronym stands for Attributable, Legible, Contemporaneous, Original, and Accurate which are values derived from the pharmaceutical industry data collected during the manufacturing process and which ensure the integrity of the data. SPuMoNI is a European funded research project that explores the authenticity of such data using innovative scientific approaches (Leal, et al., 2021).

This thesis presents, an experimental attempt to predict the ALCOA values while using the raw sensor data as an input. To achieve the best regression result, three Deep Learning Recurrent Neural Networks have been used. Long Short-Term Memory, Bidirectional Long-Short Term Memory and Gated Recurrent Unit. More specifically, there are two ALCOA values that are being examined in this thesis, Legible and Accurate. The language used for programming the algorithm was Python through Google Collaboratory and the library imported for the deep learning methods was Google's TensorFlow.

Through the research conducted, it was shown that the prediction's accuracy of the ALCOA principles Legible and Accurate was below 80% which drives us to research a) which additional parameters need to be taken into account and b) which algorithms such as a Transformer Neural Network could lead to accuracy improvements.

Keywords

ALCOA, Deep Learning, Accurate, Legible. Data Integrity, Pharmaceutical Industries

Περίληψη

Από τα μέσα του 21^{ου} αιώνα έως και σήμερα, η επιστήμη των υπολογιστών έχει αναπτυχθεί ευρέως. Από την αναγνώριση ομιλίας έως την ανακατασκευή εγκεφαλικών κυκλωμάτων μέχρι και την κατανόηση της φυσικής γλώσσας, μπορεί κανείς εύκολα να ισχυριστεί ότι η Μηχανική Μάθηση και η Βαθιά Μάθηση έχουν γίνει μέρος της καθημερινής ζωής κάθε ανθρώπου (LeCun, et al., 2015). Χρησιμοποιούνται σε πληθώρα βιομηχανιών και η φαρμακοβιομηχανία δεν αποτελεί εξαίρεση.

Τα τελευταία χρόνια, η ακεραιότητα των δεδομένων αποτελεί σημαντικό μέρος του φαρμακευτικού κλάδου. Είναι απαραίτητο να διασφαλίζεται η ποιότητα και η ασφάλεια που έχει ένα φάρμακο από την έναρξη της παραγωγής του μέχρι να φτάσει στον πελάτη. Για να επιτευχθεί αυτό, η Αμερικάνικη Υπηρεσία Τροφίμων και Φαρμάκων εισήγαγε τις αρχές ALCOA ως κατευθυντήριες γραμμές για κάθε φαρμακευτική εταιρεία. Το ακρωνύμιο ALCOA σημαίνει Attributable, Legible, Contemporaneous, Original και Accurate και χαρακτηρίζει τιμές που εξάγονται από τα δεδομένα (τα οποία συλλέγονται κατά την παραγωγική διαδικασία) της φαρμακευτικής βιομηχανίας και διασφαλίζουν την ακεραιότητά τους. Η SPuMoNI είναι ευρωπαϊκό χρηματοδοτούμενο ερευνητικό πρόγραμμα το οποίο διερευνά την αυθεντικότητα των δεδομένων, με χρήση καινοτόμων επιστημονικών προσεγγίσεων (Leal, et al., 2021).

Στην παρούσα διπλωματική, παρουσιάζεται ένα πείραμα για την πρόβλεψη των τιμών ALCOA χρησιμοποιώντας τα ακατέργαστα δεδομένα αισθητήρων της SPuMoNI ως είσοδο. Για να επιτευχθεί το καλύτερο αποτέλεσμα παλινδρόμησης, έχουν χρησιμοποιηθεί τρία νευρωνικά δίκτυα βαθιάς μάθησης. Πρόκειται για τα Long Short-Term Memory, Bidirectional Long-Short Term Memory and Gated Recurrent Unit. Ειδικότερα, υπάρχουν δύο τιμές ALCOA που εξετάζονται σε αυτή τη διατριβή, η Legible και η Accurate. Η γλώσσα που προγραμματίστηκε ο αλγόριθμος ήταν η Python μέσω του Google Collaboratory και η βιβλιοθήκη που χρησιμοποιήθηκε για τις μεθόδους βαθιάς μάθησης ήταν το TensorFlow της Google.

Μέσα από την έρευνα που διεξάχθηκε, φάνηκε ότι η πρόβλεψη των αρχών ALCOA Legible και Accurate δεν ήταν ιδιαίτερα ακριβής. Ωστόσο, αξίζει να διερευνηθεί α) ποιες επιπλέον παράμετροι πρέπει να ληφθούν υπόψιν ή/και β) ποιο αλγόριθμοι όπως transformers neural network θα μπορούσαν να οδηγήσουν σε αύξηση της ακρίβειας πρόβλεψης.

Λέξεις – κλειδιά

ALCOA, Βαθιά Μάθηση, Accurate, Legible, Ακεραιότητα Δεδομένων, Φαρμακευτικές Βιομηχανίες.

Table of Contents

List of Tables **11**
List of figures **12**
Introduction **14**
The subject of this thesis **14**
Aim and objectives **14**
Methodology **14**
Innovation **14**
Structure **14**

1 CHAPTER 1: Pharmaceutical Industries **16**
1.1 Data Integrity **16**
1.1.1 What is Data Integrity 16
1.1.2 Data Integrity in the Pharmaceutical Industry 17
1.2 ALCOA **17**
1.2.1 ALCOA and Data Integrity 18
1.3 Smart Pharmaceutical MaNufacturIng (SPuMoNI) **19**

2 CHAPTER 2: State of the art of Machine Learning **21**
2.1 Machine Learning **22**
2.2 Regression **22**
2.3 Deep Learning **23**
2.3.1 Convolutional Neural Networks 23
2.3.2 Recurrent Neural Networks 23
2.4 TensorFlow **26**

3 CHAPTER 3: Project Development **28**
3.1 Data and Data Analysis **28**
3.2 Methods Used and Results **29**
3.2.1 Alcoa Principle Legible 29
3.2.2 Alcoa Principle Accurate 36
3.3 Results Analysis **48**

Conclusion **50**

Future Research **50**

Bibliography – References – Online sources **51**

Appendix A **54**

Appendix B **54**

List of Tables

Table 1: Parameters for Bidirection LSTM, ALCOA Principle Legible 30

Table 2: Parameters for GRU, ALCOA Principle Legible..31

Table 3: Parameters for LSTM, ALCOA Principle Legible 33

Table 4: Parameters for Bidirectional LSTM, ALCOA Principle Legible 34

Table 5: Parameters for GRU, ALCOA principle Legible..34

Table 6: Parameters for LSTM, ALCOA principle Legible..35

Table 7: Parameters for Bidirectional LSTM, ALCOA principle Accurate 36

Table 8: Parameters for GRU, ALCOA principle Accurate..38

Table 9: Parameters for LSTM, ALCOA principle Accurate 40

Table 10: Parameters for Bidirection LSTM, ALCOA principle Accurate 42

Table 11: Parameters for GRU, ALCOA principle Accurate..44

Table 12: Parameters for LSTM, ALCOA principle Accurate 46

List of figures

Figure 1: Total Data Quality Managment Cycle (Lee, et al., 2004) 17

Figure 2: Acronym ALCOA principes (Anon., 2020) 18

Figure 3: The SPuMoNI project (González-Vélez, April 2021) 19

Figure 4:SPuMoni's project System Overview (González-Vélez, April 2021) 20

Figure 5: From Artificial Intelligence to Deep Learning (Knox, 2022) 21

Figure 6: Recurrent Neural Networks and different cell units of its hidden layer (Lynn, et al., 2019) 24

Figure 7: LSTM cell unit (Lynn, et al., 2019) 25

Figure 8: Bidirectional Long-Short Term Memory (Graves , et al., 2013) 25

Figure 9: GRU cell unit. (Lynn, et al., 2019) 26

Figure 10: Production Line I1000 ALCOA values 29

Figure 11: Production Line I600 ALCOA values 29

Figure 12: Bidirectional LSTM Legible I1000 30

Figure 13: Bidirectional LSTM Legible using Minmax Scaling I1000 31

Figure 14: GRU Legible I1000 32

Figure 15: GRU Legible using Minmax Scaling I1000 32

Figure 16: LSTM Legible I1000 33

Figure 17: Bidirectional LSTM Legible I600 34

Figure 18: GRU Legible I600 35

Figure 19: LSTM Legible I600 36

Figure 20: Bidirectional LSTM Accurate I1000 37

Figure 21: Bidirectional LSTM Accurate using Minmax Scaling I1000 37

Figure 22: Bidirectional LSTM Accurate using Standard Scaling I1000 38

Figure 23: GRU Accurate I1000 39

Figure 24: GRU Accurate Using Minmax Scaling I1000 39

Figure 25: GRU Accurate Using Standard Scaling I1000 40

Figure 26: LSTM Accurate I1000 41

Figure 27: LSTM Accurate Using Minmax Scaling I1000 41

Figure 28: LSTM Accurate Using Standard Scaling I1000 42

Figure 29: Bidirectional LSTM Accurate I600 43

Figure 30: Bidirectional LSTM Accurate using Minmax Scaling I600 43

Figure 31: Bidirectional LSTM Accurate using Standard Scaling I600 44

Figure 32: GRU Accurate I600 45

Figure 33: GRU Accurate Using Minmax Scaling I600 45

Figure 34: GRU Accurate Using Standard Scaling I600 46

Figure 35: LSTM Accurate I600 47

Figure 36: LSTM Accurate Using Minmax Scaling I600 47

Figure 37: LSTM Accurate Using Standard Scaling I600 48

Acronym Index

AI: Artificial Intelligence

ALCOA: Attributable, Legible, Contemporaneous, Original, and Accurate

ANN: Artificial Neural Network

Bi-LSTM: Bidirectional Long Short-Term Memory

CNN: Convolutional Neural Network FDA: Food and Drug Administration

GDP: Good Documentation Practice

LSTM: Long Short-Term Memory

MAE: Mean Absolut Error

RNN: Recurrent Neural Network

STD: Standard Deviation

SPuMoNi: Smart Pharmaceutical MaNufacturIng

Introduction

This thesis, presents an experiment done in order to strengthen data integrity in pharmaceutical industries. The data used to achieve this was taken from “Smart Pharmaceutical MaNufacturIng (SPuMoNi)” which is a project aiming at ensuring data integrity and quality in the pharmaceutical domain. In order to assure the data integrity, companies use the ALCOA principles. Those consist of five values one for each letter and each one of the letters represents a different approach on data integrity. In recent years, the ALCOA principles grew in number of values and today they consist of nine values in total. Taking those values and predicting them would not only help the FDA but also the pharmaceutical industries to have a better control of their production. The prediction methods that were tried are Bidirectional LSTM, LSTM and GRU.

The subject of this thesis

Data integrity and Good Documentation Practice (GDP) are both mandatory when it comes to pharmaceutical industries. Through this project, there is an effort on reinforcing the FDA tryout on securing the safety and quality of the drugs.

Aim and objectives

The main goal of the algorithm made in this thesis is to anticipate the value of the ALCOA parameters in order to prevent future alarms concerning to the production of the drugs, while at the same time making it easier for the FDA to anticipate the ALCOA values a pharmaceutical company would have.

Methodology

In order to successfully predict the ALCOA values, deep learning methods are being used. More specifically there are three Recurrent Neural Networks developed. Long Short-Term Memory, Bidirectional Long Short-Term Memory and Gate Recurrent Unit. For each ALCOA value examined, we tried all three of these algorithms. Furthermore, for every algorithm we used two scaling methods for the data beyond the raw unscaled data. The scaling methods used where MinMax Scaling and Standard Scaling.

Innovation

Taking into consideration that there is no other bibliography found regarding the ALCOA value prediction neither via Machine Learning nor via Deep Learning, the attempt to combine Artificial Intelligence and the ALCOA principles has the traits of an innovative research. Medical Industrial data is an immense topic and could have many research points to be explored on. Regarding the ALCOA value we believe this project can be a beginning of exploring data integrity through artificial intelligence.

Structure

In this thesis we firstly analyze the subject regarding the pharmaceutical industry. There is a general approach on the data integrity topic with a specialization on the ALCOA principles part. Subsequently, we talk about machine learning and deep learning focusing on the three

Recurrent Neural Networks used in the experiment. Finally, in the third chapter, we present, comment and analyze the results of the algorithm.

1 CHAPTER 1: Pharmaceutical Industries

1.1 Data Integrity

Data

Data is the original records and true copies of them, including meta data and raw data and all reports and transformation of these data. (Ahmad, et al., 2019)

Meta Data

Meta data describe the attributes of other data and permits data to be attributable to an individual.

Raw data

Is the original record retained in the original format or as a "true copy" (Rattan, 2018).

1.1.1 What is Data Integrity

Data Integrity means that data is accurate and consistent throughout its entire lifecycle. For this reason, the companies ensure good documentation and data management practices. Data without integrity has no much value. The better data integrity a company has the more successful it is like to become. In addition, poor data integrity undermines the quality of products. Data integrity applies to both paper and electronic records (Ahmad, et al., 2019).

There are five Data Integrity Constraints (Lee, et al., 2004):

- Entity Integrity: All entries in the database are unique and there are no null values
- Referential Integrity: A foreign key may have either an entry that matches the primary key value in a table to which it is related or a null entry – as long as it is not a part of its table's primary key.
- Domain Integrity: All values of an attribute in the database must be from a specified domain
- Column Integrity. All values of an attribute in the database must be from a specified range within the domain
- User-Defined Integrity: Specific business integrity rules that are not on the above.

Data quality is a key aspect of data integrity. When there are data well defined, their quality is measured, analyzed and the data integrity rules are redefined as time passes. Then the organization would easily achieve data integrity (Lee, et al., 2004).

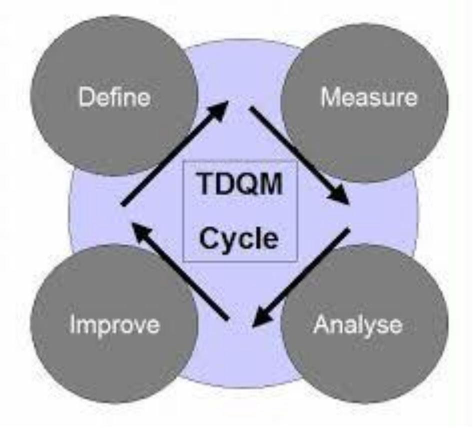


Figure 1: Total Data Quality Managment Cycle (Lee, et al., 2004)

1.1.2 Data Integrity in the Pharmaceutical Industry

When searching Google there should be accuracy in the results. The same principles apply in the drug development circle. Pharmaceutical industries are obliged to ensure that the data entered for the various steps of drug development is accurate and therefore the drugs produced are within some parameters. Data Integrity in pharmaceutical industry is whether an industry produces drugs within the stated parameters. If the inspectors find that the data concerning a drug is modified then the drug is declined approval and is not available in the market for patients. That results to a revenue loss for the industry (Rattan, 2018).

Food and Drug Administration (FDA) published the first guideline for data integrity in 1963. Since then, FDA as well as the European Union (EU) published numerous guidelines related to data integrity for the pharmaceutical industry (Rattan, 2018).

1.2 ALCOA

The acronym ALCOA was introduced in the 1990's by Stan W. Woollen member of the FDA's Office of Enforcement. The acronym stands for:

- **A**ttributable: It should be clear by whom, when and where was the data documented. This can be accomplished by either an "audit trail" of the data entry in the case of electronic source record or manually by initialing and dating a paper record.
- **L**egible: All paper and electronic data should be readable and permanent the record must be accessible and combined with raw as well as with meta data.
- **C**ontemporaneous: The record of the paper or the electronic data should be documented at the time it is performed.
- **O**riginal: Original Record must be preserved including protocol, form, notebooks, spreadsheet, database or software application.
- **A**ccurate: The data must be free of mistakes and errors. It should represent the facts with accuracy and consistency. (Roznoski, 2014)

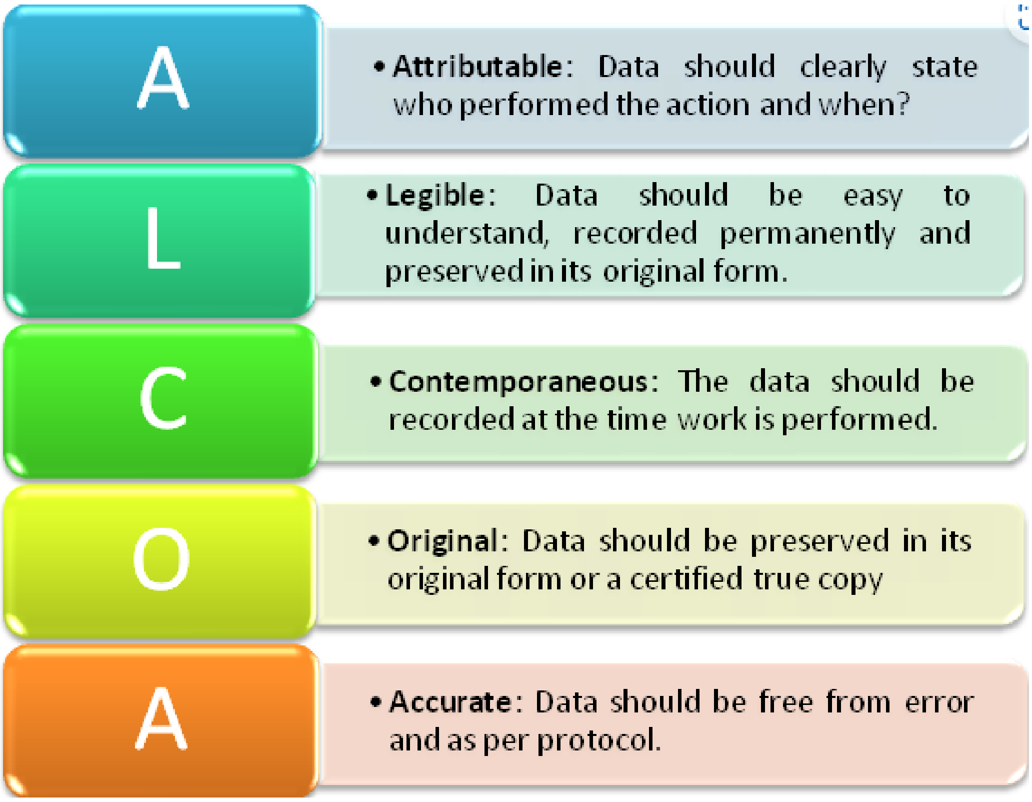

A
• Attributable: Data should clearly state who performed the action and when?
L
• Legible: Data should be easy to understand, recorded permanently and preserved in its original form.
C
• Contemporaneous: The data should be recorded at the time work is performed.
O
• Original: Data should be preserved in its original form or a certified true copy
A
• Accurate: Data should be free from error and as per protocol.

Figure 2: Acronym ALCOA principles (Anon., 2020)

1.2.1 ALCOA and Data Integrity

ALCOA is a framework used by pharmaceutical industries to ensure data integrity and applies for papers and electronic data. ALCOA principles are complying with Good Documentation Practices (GDPs) and good manufacturing practices (GMPs).

In 2010 ALCOA was further expanded to ALCOA-CCEA or ALCOA-C or ALCOA+ by adding four more extra values in order to ensure furthermore the data integrity in the industry. CCEA stand for:

- **C**omplete: The data should not lack anything.
- **C**onsistent: Data is recorded chronologically, with the date and time stamp in the expected sequence.
- **E**nduring: The data is recorded in laboratory notebooks or in validated software systems (spreadsheets, databases).
- **A**vailable: Paper and electronic data should be readily available for review. Therefore, it should be properly labeled to facilitate retrieval (Girard & Watkin, 2020).

To summarize in order to defend data integrity we have to look in to the following components:

I. Develop policies and procedures. The strictest review of process of “audit trails”.

II. Develop people so they can have a better understanding of the data being generated, modified, updated, deleted and stored.

III. Develop better technology and management of the data lifecycle (Rattan, 2018).

1.3 Smart Pharmaceutical MaNufacturIng (SPuMoNI)

SPuMoNI is a 36 months research project funded by the European Commission. It forecasts on developing a semi-autonomous quality decision support system in order to extract transform and control in a variety of data sources within the pharmaceutical manufacturing. SPuMoNI project aims to contribute to patient safety by using quality mechanisms evolving blockchain technology, intelligent agents and data quality mechanisms. (Leal, et al., 2021)

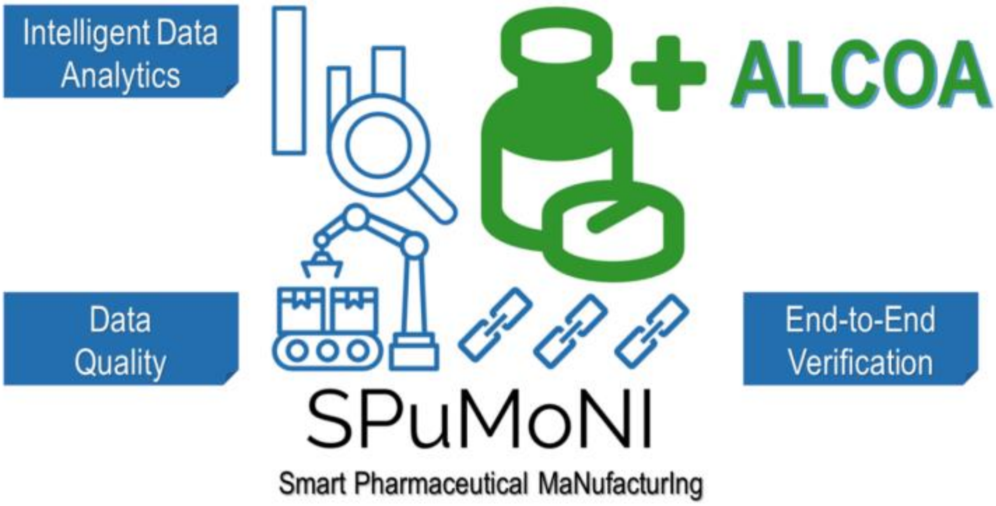

Intelligent Data Analytics
ALCOA
Data Quality
End-to-End Verification
SPuMoNI
Smart Pharmaceutical MaNufacturIng

Figure 3: The SPuMoNI project (González-Vélez, April 2021)

The integrity of pharmaceutical data assets should comply with:

- ALCOA principles
- Food and Drug Administration (FDA) regulations
- European Medicines Agency (EMA)

Nevertheless, there is no guarantee that the current instruments used are not susceptible to falsified data from the pharmaceutical data strings. Thus, the pharmaceutical industry needs autonomous control mechanisms to assure the integrity, authenticity and end-to-end traceability of data.

SPuMoNi's approach and goals on data integrity:

- To assess all the computerize data in a representative pharma environment.
- To design data quality models that include the rules from regulatory documents.
- To identify patterns of data that can violate ALCOA premises.
- To use latest software systems and combine them with data quality control models. (González-Vélez, April 2021)

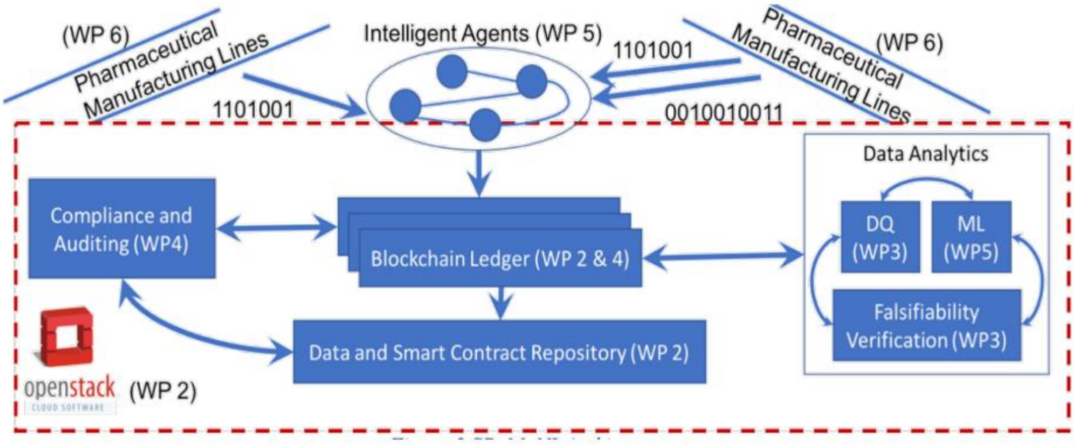

(WP 6)
Pharmaceutical Manufacturing Lines
1101001
Intelligent Agents (WP 5)
1101001
0010010011
Pharmaceutical Manufacturing Lines
(WP 6)
Compliance and Auditing (WP4)
Blockchain Ledger (WP 2 & 4)
Data and Smart Contract Repository (WP 2)
openstack CLOUD SOFTWARE
(WP 2)
Data Analytics
DQ (WP3)
ML (WP5)
Falsifiability Verification (WP3)

Figure 4:SPuMoni's project System Overview (González-Vélez, April 2021)

2 CHAPTER 2: State of the art of Machine Learning

Looking back there many states of the art moments that can be noted as machine learning aroused.

- In 1943 McCulloch and Pitts proposed a new model for solving functions that uses neural networks. (Georgouli, 2015)
- In 1950, Alan Turing, also known as the father of artificial intelligence and computer science, published his work “Computing Machinery”. There he proposed the Turing Test, that is used until today, checking whether a computer is smart. Furthermore, with his work called “Can Computers Think”, he introduced machine learning, genetic algorithms and artificial intelligence. (Christodoulos, 2020)
- In 1956, the first Artificial Intelligence conference was called in Darthmouth College. There, MIT’s John McCarthy & Marvin Minsky and CMU’s Herbert Simon & Allen Newell presented their research called “A Proposal For The Dartmouth Summer Research Project On Artificial Intelligence” that made them the founders of Artificial Intelligence (AI). (McCarthy, et al., 1955) (Georgouli, 2015)
- In 1957 and in 1962 Frank Rosenblatt, through his work “Perceptron”, analyzed the mechanics of neural networks with one or more layers. (Rosenblatt, 1958)
- In 1986, the first multilayer neural network is being created using backpropagation. It is presented by David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams on their paper called “Learning representations by back-propagating errors” (Rumelhart, et al., 1986)
- In 1997, the algorithm “Deep Blue” programmed by an IBM team, managed to beat Garry Kasparov on a chess game. Out of six games played the program won two games and lost one. The other three games ended as a draw. After the games where over Kasparov said: "It was an impressive achievement, of course, and a human achievement by the members of the IBM team, but Deep Blue was only intelligent the way your programmable alarm clock is intelligent. Not that losing to a $10 million alarm clock made me feel any better." (Anon., 1997) (Peter, 2011)
- In 2020, the algorithm DeepMind, solved the protein-folding problem, a problem that scientists were trying to solve for more than 50 years (AlphaFold, 2020).

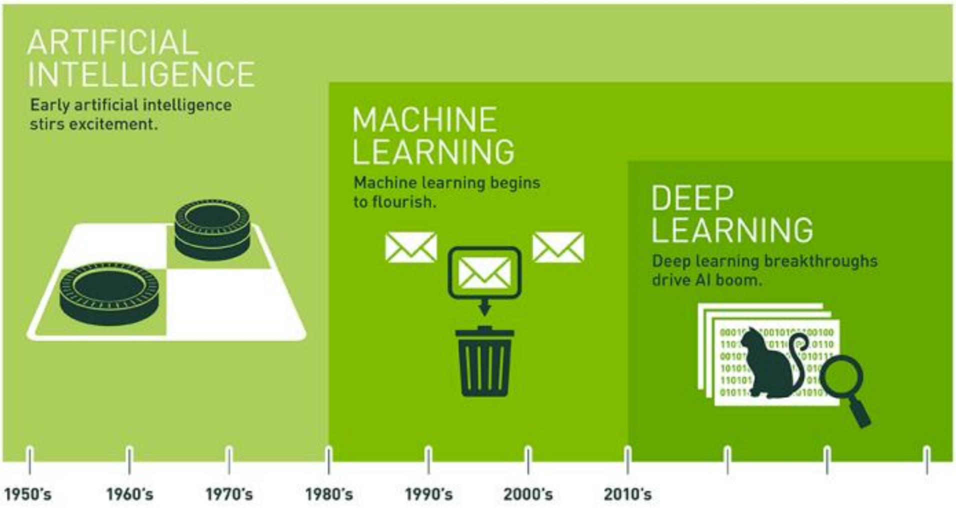

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

Figure 5: From Artificial Intelligence to Deep Learning (Knox, 2022)

2.1 Machine Learning

An algorithm that uses input data in order to achieve a task without being “hard coded” is called a machine learning algorithm. (El Naqa & Murphy, 2015). In order to achieve the desired task, the algorithm, uses data to adapt with the problem and have experience on how to solve it. this procedure is called “training” the algorithm. Afterwards when the algorithm is trained it can generalize to produce the desired outcome from new, previously unseen data, this procedure is called “learning” the algorithm (El Naqa & Murphy, 2015). There are three machine learning models used:

- Supervised Learning: The input and output information are given to the algorithm and it builds the function.
- Unsupervised Learning: The output information is not given to the algorithm so it makes connection between the inputs in order to make some general rules.
- Semi-Supervised Learning: The algorithm is given a variety of the training data and it tries to predict and learn by constructing general rules. (Panagiota, 2019)

The most used machine learning algorithms are: Logistic Regression, Bayesian Network, Support Vector Machines (SVM). Linear Classifier, Decision Tree, Random Forest, k-Nearest Neighbor (k-NN), Artificial Neural Network (ANN) etc (Shinde & Shah, 2018).

2.2 Regression

Through regression analysis, it is easier to statistically explore relationships between species and the environment. In contrast with clustering, where data can be analyzed on all species simultaneously, regression analyses data for each of the species alone (ter Braak & Looman, 1986). In machine learning regression is used to predict the outcome of future events. (W3Scools, n.d.). There are several forms of regression models for machine learning.

1) Linear Regression: The relationship between the input and the outcome is a straight line. An example is the relationship between height and weight.
2) Multiple Regression: There are more inputs that affect the outcome of the model. There can be linear multiple regression models.
3) Non-Linear: The relationship between the input and the outcome is not a straight line.
4) Stepwise Regression Model: Regression Technique used on multiple regression models where every input is added to the model one at a time. (IMSL, 2021)

For the accuracy of the model and the outcome there is a method usually used called **Mean Squared Error (MSE).** MSE, computes the square of the distance between every point from the regression line. Afterwards all the distances are being summed in order to find the mean. Where the lower the MSE the better the forecasting.

$$MSE = \left(\frac{1}{n}\right) * \Sigma(actual - Forecast)^2$$

Where n is the total number of items, actual is the observed y values and forecast is the values from the regression (Panik, 2022).

2.3 Deep Learning

Through the power of machine learning, many aspects of the modern world have been benefited. Commercials, social networks and content filtering are some of the main domains which machine learning has a huge impact. (Kamilaris, et al., 2018) However, when processing raw data, machine learning appears to encounter some limitations. Achieving natural data processing through machine learning will require careful engineering and more expertise. Deep learning methods, help to deal with the above difficulties, they help especially to solve important problems that are difficult to be approached by the AI community. A significant advantage that makes DL more widespread than ML, is feature learning and extracting features from raw data. (LeCun, et al., 2015)

Deep learning neural networks consist of representation-learning methods with multiple levels of representation. (LeCun, et al., 2015) Each level learns a pattern from the previous layer and provides more abstract data for the higher layers. (Kamilaris, et al., 2018). The most important about these levels, is that they are not made by an engineer but they are learned through data processing by using a general-purpose learning procedure. For the majority of the science community, a deep neural network must consist of at least two levels. However, there is no rule that imposes this. (Rusk, 2016)

While using deep learning, either higher accuracy can be achieved in a classification problem, or errors can be reduced in a regression problem. This is due to the large number of different components DL neural networks consist, for example, pooling layers, convolutional layers, connected layers, activation functions, etc. There are many architectures that can be used when making a Neural Network such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). (Kamilaris, et al., 2018)

2.3.1 Convolutional Neural Networks

The applications of the Convolutional Neural Networks have been dated back to the early 1990s (LeCun, et al., 2015). Their name was taken by the mathematical convolutional operation done between linear matrixes. (Albawi & Tareq Abed, 2017). They usually consist of many layers which in combination with the local connections, the shared weights and the pooling, make ConvNets a suitable tool when processing natural signals. The first few layers are often either convolutional or pooling. Afterwards, there is a non -linearity such as a Rectified Linear Unit (ReLU) (LeCun, et al., 2015). The convolutional layers and the fully-connected ones have parameters, on the contrary non-linearity and pooling have not. (Albawi & Tareq Abed, 2017). It has been found that Convolutional Neural Networks have great results when operating in machine learning problems. Due to the small number of parameters used in a CNN, they have been used in many complicated problems that Artificial Neural Network could not solve. The most common problems that CNN are used for are voice recognition, image processing and pattern recognition. (Albawi & Tareq Abed, 2017)

2.3.2 Recurrent Neural Networks

Recurrent neural networks were risen at the time backpropagation was first introduced (Lynn, et al., 2019). Its more effective use is for sequential data such as video, speech and language. There are two types of RNNs, continuous-time and discrete-time RNNs (Xiaosheng, et al., 2019). The main problem of an RNN is said to be the growth or shrink of the backpropagated gradients. This leads to the either the explosion or the vanishment of those gradients. (Lynn, et al., 2019)

We can describe an RNN by using the following equations:

$$h_t = f(W_h h_{t-1} + V_h x_t + b_h)$$

And

$$o_t = f(W_o h_t + b_o)$$

Where, h_t denotes the hidden block of each time step (t), W and V are the weights for the hidden layers in recurrent connection, while b denotes the bias for hidden and output states as f represents an activation function applied on each node throughout the network (Lynn, et al., 2019). As shown from the equations, every node proceeds using a feedback loop for the next node, while the output is given through the current input and a previous hidden state.

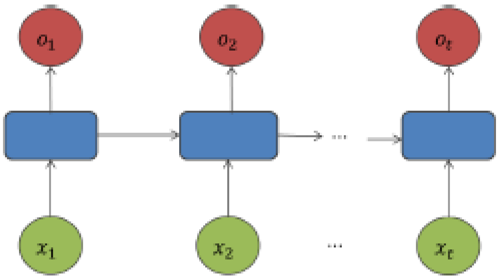


Figure 6: Recurrent Neural Networks and different cell units of its hidden layer (Lynn, et al., 2019)

Long Short-Term Memory

Researchers managed to address the problem RNNs were facing by introducing Long Short-Term Memory (LSTM) models. Nowadays, LSTM is one of the most used methods in the deep learning community and are even more effective than conventional RNNs in every aspect that they have been used. Even some well performing encoder and decoder networks are currently being developed through LSTM (Sak, et al., 2014).

We can describe an LSTM model by using the following equations:

$$f_t = \sigma(W_{xf}^T \cdot x_t + W_{hf}^T \cdot h_{t-1} + b_f)$$

$$i_t = \sigma(W_x^T i \cdot x_t + W_h^T i \cdot h_{t-1} + b_i)$$

$$o_t = \sigma(W_x^T o \cdot x_t + W_h^T o \cdot h_{t-1} + b_i)$$

$$g_t = tanh(W_x^T g \cdot x_t + W_h^T g \cdot h_{t-1} + b_i)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \hat{c}_t$$

and

$$o_t, h_t = g_t \otimes \tanh\,(c_t)$$

Where f_t is the forget gate which decides the part of c_t to be committed. Furthermore, i_t is the input gate which controls the adding part of $\hat{c}_t$ to long-term c_t. The output gate is symbolized as g_t and determines the part of c_t that is read by h_t and output to o_t. The other parameters used, W_xf, W_xi, W_xo, W_xg denote the weight matrices for the corresponding connected input vector, W_hf , W_hi, W_ho, W_hg represent the weight matrices of the short-term state of the previous time step, and b_f , b_i, b_o, b_g are bias (Lynn, et al., 2019).

In the recurrent hidden layer, there are special memory blocks which contain memory cells, gates and an input and output gate. Memory cells, save the temporal state of the networks and gates are normalizing the flow of information. The input gate is linked to the memory cell and it controls the flow of input activation contained in it. The output gate is linked to the rest of the network and it controls the output flow of input activations (Sak, et al., 2014).

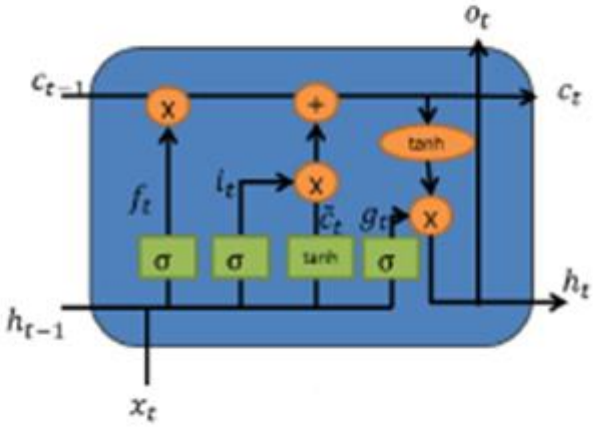


Figure 7: LSTM cell unit (Lynn, et al., 2019)

Bidirectional Long Short-Term Memory

Bidirectional recurrent neural networks (BRNNs) have achieved and improved many RNN such as many sequence learning tasks, natural language and speech processing and they were even used in a protein structure prediction. Particularly in the last example, in 2007, Thireon and Reczko, used an architecture called Bidirectional Long Short-Term Memory (BiLSTM) which outperformed the normal BRNNs and also the feedforward network (Xiaosheng, et al., 2019). The main idea of a BiLSTM model is the separation of two recurrent nets. The first one would present the training sequence forwards and the second one would do it backwards. Both the recurrent nets would be connected with the same output layer (Graves & Jurgen, 2005).

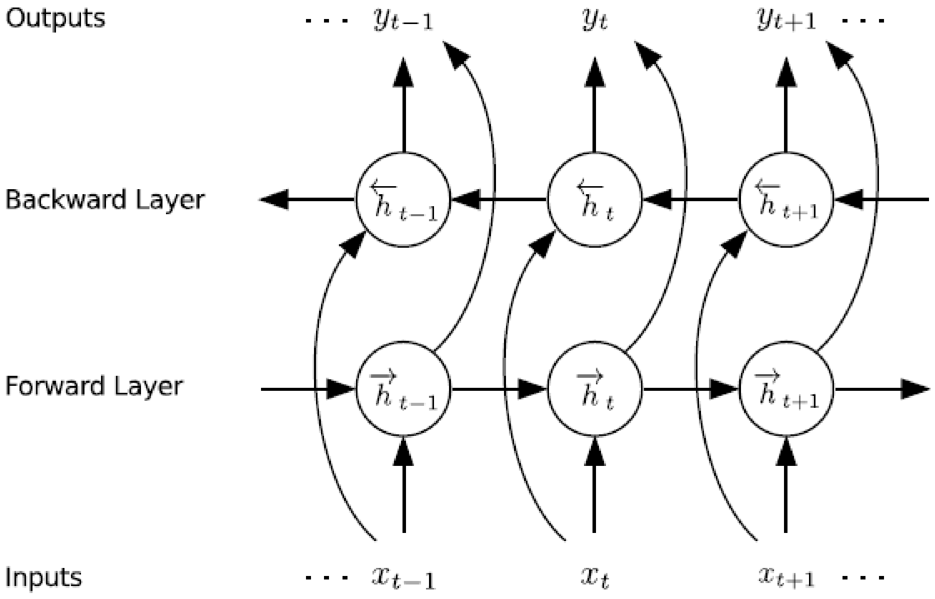


Figure 8: Bidirectional Long-Short Term Memory (Graves , et al., 2013)

Gated Recurrent Neural Networks

Gated Recurrent Neural Networks (Gated RNNs) have shown success in several applications involving sequential or temporal data. For example, they have been applied extensively in speech recognition, music synthesis, natural language processing, machine translation, etc. Long Short-Term Memory (LSTM) RNNs and the recently introduced Gated Recurrent Unit (GRU) RNNs have been successfully shown to perform well with long sequence applications.

Gated RNNs' success is primarily due to the gating network signaling that control how the present input and previous memory are used to update the current activation and produce the current state. These gates have their own sets of weights that are adaptively updated in the learning phase (i.e., the training and evaluation process). While these models empower successful learning in RNNs, they introduce an increase in parameterization through their gate networks. Consequently, there is an added computational expense vis-a-vis the simple RNN model. It is noted that the LSTM RNN employs 3 distinct gate networks while the GRU RNN reduces the gate networks to two. (Rahul & Fathi, 2017)

Gated Recurrent Unit

Based on the LSTM model, there was a model designed called Gated Recurrent Unit (GRU) which has two gate structures only that include the reset gate and the update gate (Wuyan, et al., 2020). The reset gate is symbolized as r_t and the update gate as z_t.
We can describe an GRU model by using the following equations:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$
$$\tilde{h}_t = g(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)$$

with the two gates presented as:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$
$$r_t = \sigma(W_{rz} x_t + U_r h_{t-1} + b_r)$$

An LSTM model can be outperformed by an GRU on specific datasets even on parameter generalization and update and on CPU time (Rahul & Fathi, 2017) (Wuyan, et al., 2020).

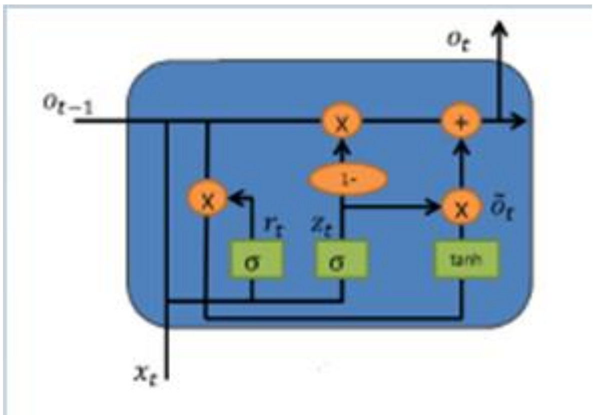


Figure 9: GRU cell unit. (Lynn, et al., 2019)

2.4 TensorFlow

Neural Networks and Machine Learning algorithms have been a state of the art for the last years. In order to achieve a fast and accurate learning of an algorithm, a variety of software libraries have been developed in order to assist programmers. The most popular is TensorFlow which was originally designed by Google Researchers. The core algorithms of TensorFlow were optimized on Cbb and CUDA (Compute Unified Device Architecture) that were created by NVIDIA. There are many programming languages that TensorFlow supports such as Java, JavaScript, C#, Python etc. When programming in TensorFlow, there are two sections in a program. The Construction Phase and the Execution Phase. Construction phase is the making of a graph of computation thus execution phase is the process of running this graph. In order to assist with the programming, TensorFlow also provides many building blocks such as Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Convolutional Layers, Nonlinear Activation Functions etc. (Pang, et al., 2019)

3 CHAPTER 3: Project Development

The main goal of the project was to check if there was a way of predicting the ALCOA values through the sensors data by developing a deep learning algorithm. Furthermore, there was another algorithm developed checking if the ALCOA values had any pattern so as to be predicted using a forecasting algorithm. The programming language used for the development was “Python” through the open source “Google Collab”.

3.1 Data and Data Analysis

The sensor data was collected from a database. It was formed in to two different TSV files where each one contained the sensor values for different production lines the I1000 and the I600. Each production line had different order ID’s. Each order ID corresponded to nine different ALCOA values. All the order ID’s had multiple values of each sensor. They were calculating either the temperature, the velocity or the precure at each state of the process. The ALCOA values were given in different “JSON” type files that we had to adjust through coding to correspond to the desired order ID. The following example presents the data formation we achieved through coding.

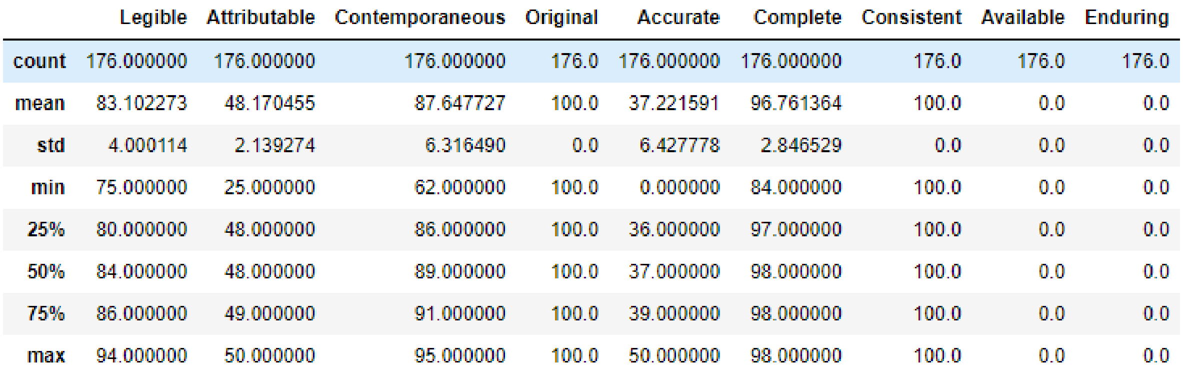
For example: If we had the order A with two sensors the formation would be:

ORDER A	SENSOR 1	ALCOA VALUES
	SENSOR 1	
	SENSOR 2	
	SENSOR 2	

The sensor column would be the input value and the “ALCOA values” column would be the output. This means that the input being used is a three-dimensional table (order ID, value, sensor).

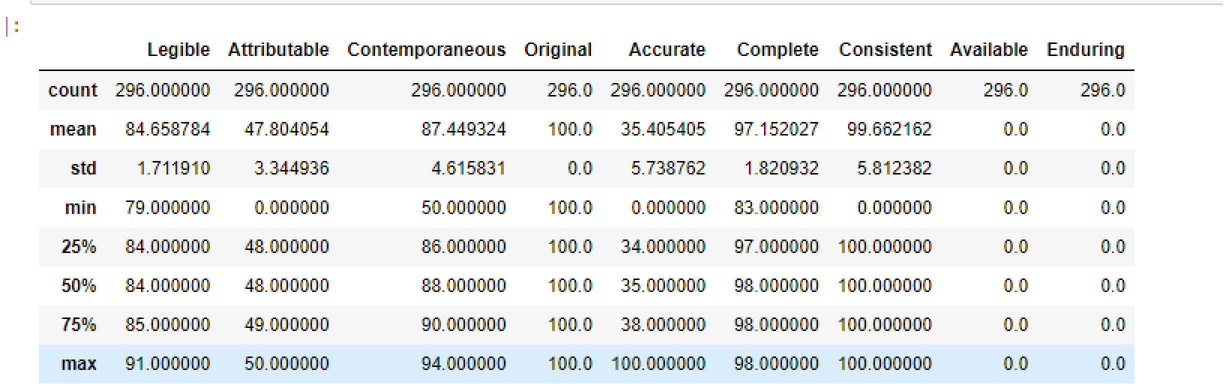
In order to be able to adjust the data to have the desired formation, zero values were added in the value columns so as every table had the same dimensions. The number zero meant that the current sensor had no output at the given time.

When checking every ALCOA principle either for the I1000 or the I600 production line, we had the following results:



	Legible	Attributable	Contemporaneous	Original	Accurate	Complete	Consistent	Available	Enduring
count	176.000000	176.000000	176.000000	176.0	176.000000	176.000000	176.0	176.0	176.0
mean	83.102273	48.170455	87.647727	100.0	37.221591	96.761364	100.0	0.0	0.0
std	4.000114	2.139274	6.316490	0.0	6.427778	2.846529	0.0	0.0	0.0
min	75.000000	25.000000	62.000000	100.0	0.000000	84.000000	100.0	0.0	0.0
25%	80.000000	48.000000	86.000000	100.0	36.000000	97.000000	100.0	0.0	0.0
50%	84.000000	48.000000	89.000000	100.0	37.000000	98.000000	100.0	0.0	0.0
75%	86.000000	49.000000	91.000000	100.0	39.000000	98.000000	100.0	0.0	0.0
max	94.000000	50.000000	95.000000	100.0	50.000000	98.000000	100.0	0.0	0.0

Figure 10: Production Line I1000 ALCOA values



	Legible	Attributable	Contemporaneous	Original	Accurate	Complete	Consistent	Available	Enduring
count	296.000000	296.000000	296.000000	296.0	296.000000	296.000000	296.000000	296.0	296.0
mean	84.658784	47.804054	87.449324	100.0	35.405405	97.152027	99.662162	0.0	0.0
std	1.711910	3.344936	4.615831	0.0	5.738762	1.820932	5.812382	0.0	0.0
min	79.000000	0.000000	50.000000	100.0	0.000000	83.000000	0.000000	0.0	0.0
25%	84.000000	48.000000	86.000000	100.0	34.000000	97.000000	100.000000	0.0	0.0
50%	84.000000	48.000000	88.000000	100.0	35.000000	98.000000	100.000000	0.0	0.0
75%	85.000000	49.000000	90.000000	100.0	38.000000	98.000000	100.000000	0.0	0.0
max	91.000000	50.000000	94.000000	100.0	100.000000	98.000000	100.000000	0.0	0.0

Figure 11: Production Line I600 ALCOA values

By looking at the mean values of each ALCOA and the std, there was a decision made that the only ALCOA principles worthy of predicting through an algorithm where: Legible, Attributable, Contemporaneous and Accurate.

In this paper, there would be an overview of the results for the ALCOA principles Legible and Accurate.

3.2 Methods Used and Results

Because of the table input size, deep learning was the only method capable enough to calculate if the ALCOA value prediction was possible. For each ALCOA value there were used three different deep learning models to achieve the best result. There were many more models tried but the Bidirectional LSTM, the GRU and the LSTM were those that achieved the best results.

Afterwards, two different types of normalization were used on the data. The normalization methods used were: Minmax Scaler and Standard Scaling.

The split was done by using the SK learn library. The test size was set to 0.17 and the random state to 42.

3.2.1 Alcoa Principle Legible

The results that the ALCOA principle Legible had are the following:

For the production line I1000

The value of the Standard deviation (STD) is equal to 4.

Deep Learning Method: Bidirectional LSTM

The parameters used to achieve the best results are the following:

Parameters	**Values**
Optimizer	Adam
Learning Rate	1e-3
Droupout	0.2
LSTM units	100
Epochs	15
Batch size	8
Early Stopping Patience	0.2

Table 1: Parameters for Bidirection LSTM, ALCOA Principle Legible

The results without any normalization:

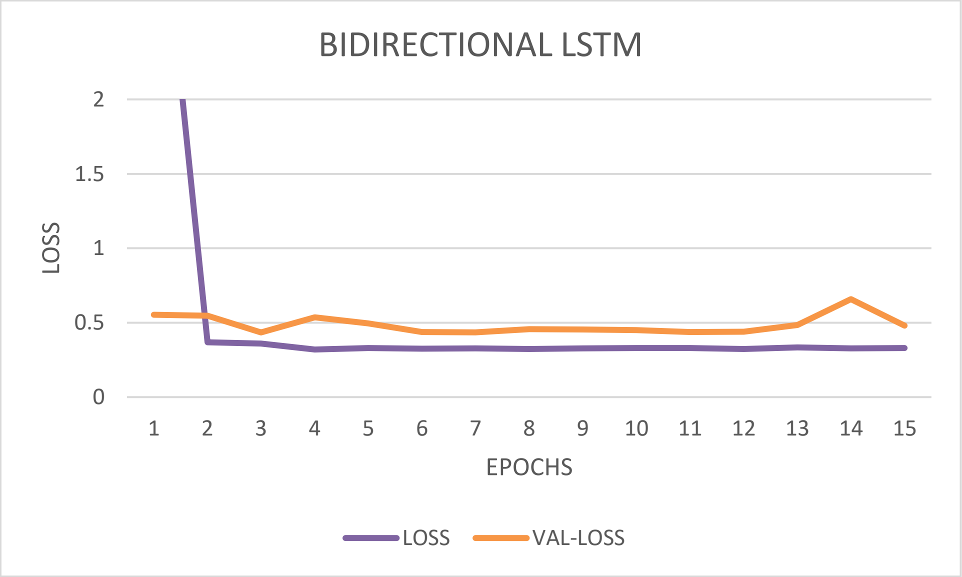


Figure 12: Bidirectional LSTM Legible I1000

The Mean Absolut Error (MAE) was 3.7123 and the Validation Mean Absolute Error (Val-MAE) was 2.3132. In comparison to the STD, the MAE shows a weakness on the algorithms learning process but on the other hand the validation MAE shows that the algorithm can predict the values closer to what was expected. These results cannot be used in a pharmaceutical industry.

The results while using Mimax Scaling Normalization:

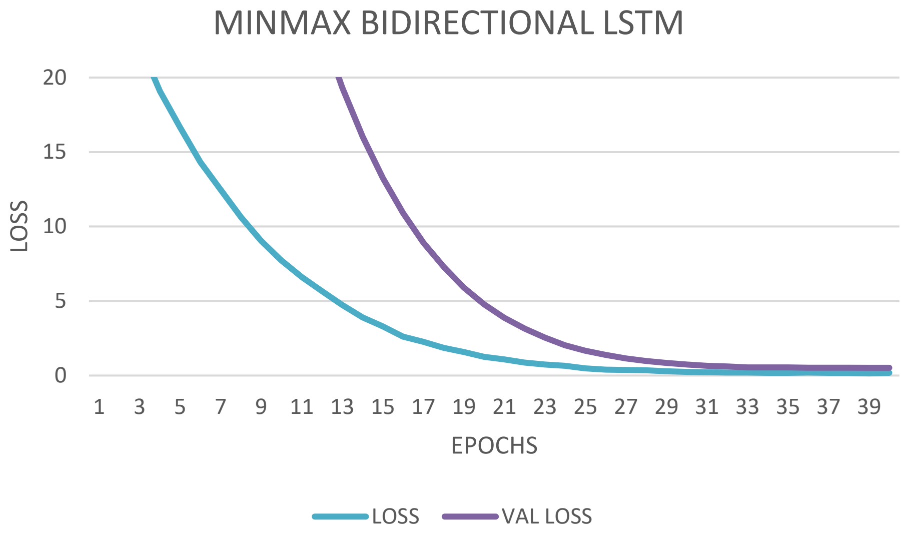


Figure 13: Bidirectional LSTM Legible using Minmax Scaling I1000

The Mean Absolut Error was 4.0723 and the Validation Mean Absolut Error was 3.4002. In comparison to the STD, the MAE shows a weakness on the algorithms learning process and on the validation MAE. These results cannot be used in a pharmaceutical industry.

Deep Learning Method: GRU

Parameters	**Values**
Optimizer	Adam
Learning Rate	1e-4
Droupout	0.2
GRU units	100
Epochs	40
Batch size	8
Early Stopping Patience	0.2

Table 2: Parameters for GRU, ALCOA Principle Legible

The results without any normalization:

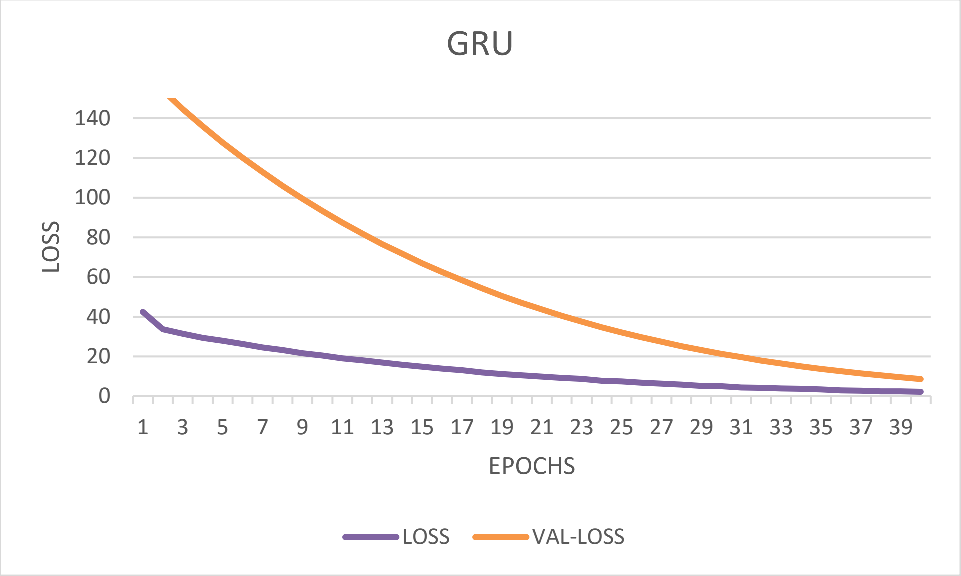


Figure 14: GRU Legible I1000

The Mean Absolut Error was 17.1281 and the Validation Mean Absolut Error was 15.5774. In comparison to the STD, the algorithm shows that by using GRU and without any Normalization, the algorithm is neither learning nor predicting the values.

The results while using Mimax Scaling Normalization:

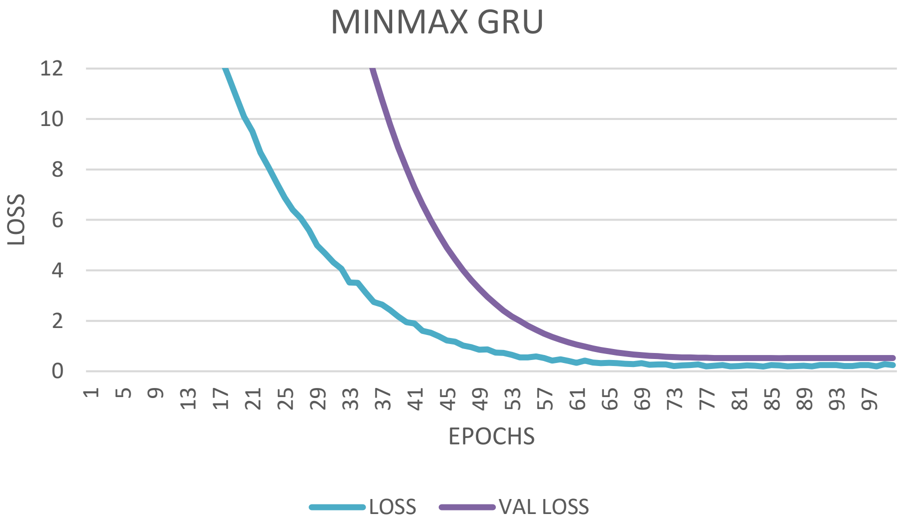


Figure 15: GRU Legible using Minmax Scaling I1000

The Mean Absolut Error was 4.7319 and the Validation Mean Absolut Error was 3.4. In comparison to the STD, the MAE shows a weakness on the algorithms learning process but on the other hand the validation MAE shows that the algorithm can predict the values closer to what was expected. However, it shows that the problem of the GRU model used for the previous experiment was the data input. These results cannot be used in a pharmaceutical industry.

Deep Learning Method: LSTM

Parameters	**Values**
Optimizer	Adam
Learning Rate	1e-4
Droupout	0.2
LSTM units	300
Epochs	20
Batch size	8
Early Stopping Patience	0.4

Table 3: Parameters for LSTM, ALCOA Principle Legible

The results without any normalization:

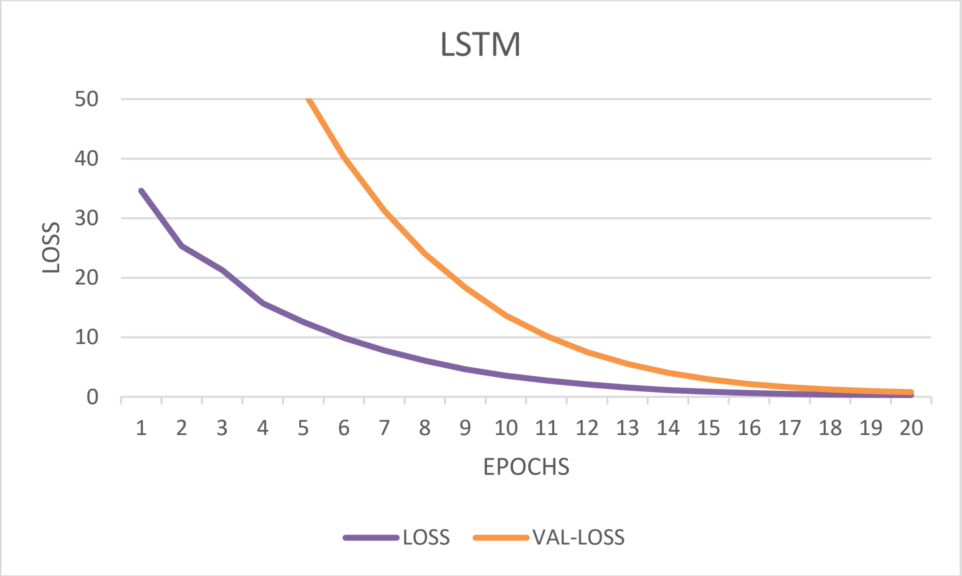


Figure 16: LSTM Legible I1000

The Mean Absolut Error was 5.3069 and the Validation Mean Absolut Error was 3.9677. In comparison to the STD, the MAE shows a weakness on the algorithms learning process and on the validation MAE. These results cannot be used in a pharmaceutical industry.

For the production line I600

The value of the Standard deviation (STD) is equal to 1.71191

Deep Learning Method: Bidirectional LSTM

Parameters	**Values**
Optimizer	Adam
Learning Rate	1e-3

Droupout	0.2
LSTM units	200
Epochs	40
Batch size	8
Early Stopping Patience	0.5

Table 4: Parameters for Bidirectional LSTM, ALCOA Principle Legible

The results without any normalization:

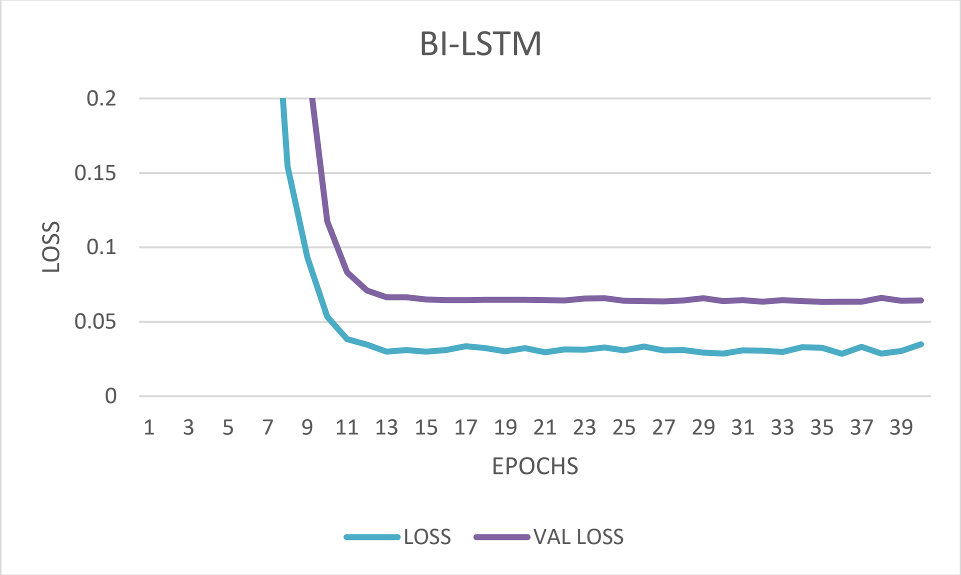


Figure 17: Bidirectional LSTM Legible I600

The Mean Absolut Error was 2.3376 and the Validation Mean Absolut Error was 1.3228. In comparison to the STD, the MAE shows a weakness on the algorithms learning process but on the other hand the validation MAE shows that the algorithm can predict the values closer to what was expected. These results cannot be used in a pharmaceutical industry.

Deep Learning Method: GRU

Parameters	**Values**
Optimizer	Adam
Learning Rate	1e-4
Droupout	0.2
GRU units	200
Epochs	40
Batch size	8
Early Stopping Patience	0.4

Table 5: Parameters for GRU, ALCOA principle Legible

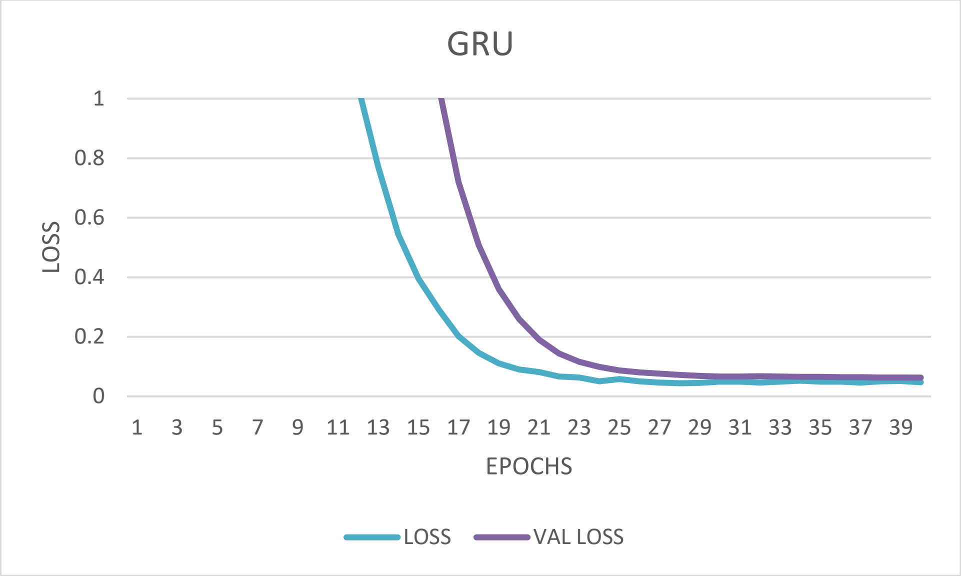


Figure 18: GRU Legible I600

The Mean Absolut Error was 2.6618 and the Validation Mean Absolut Error was 1.3197. In comparison to the STD, the MAE shows a weakness on the algorithms learning process but on the other hand the validation MAE shows that the algorithm can predict the values closer to what was expected. These results cannot be used in a pharmaceutical industry.

Deep Learning Method: LSTM

Parameters	**Values**
Optimizer	Adam
Learning Rate	1e-4
Droupout	0.2
LSTM units	200
Epochs	30
Batch size	8
Early Stopping Patience	0.5

Table 6: Parameters for LSTM, ALCOA principle Legible

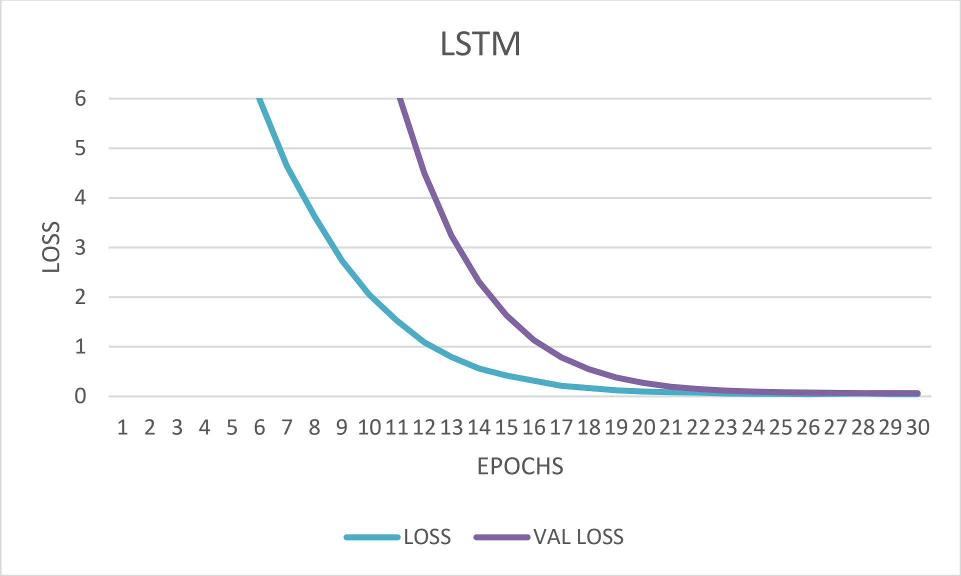


Figure 19: LSTM Legible I600

The Mean Absolut Error was 2.6416 and the Validation Mean Absolut Error was 1.3257. In comparison to the STD, the MAE shows a weakness on the algorithms learning process but on the other hand the validation MAE shows that the algorithm can predict the values closer to what was expected. These results cannot be used in a pharmaceutical industry.

3.2.2 Alcoa Principle Accurate

The results that the ALCOA principle Accurate had are the following:

For the production line I1000

The value of the Standard deviation (STD) is equal to 6,4.

Deep Learning Method: Bidirectional LSTM

Parameters	Values
Optimizer	Adam
Learning Rate	1e-4
Droupout	0.2
LSTM units	200
Epochs	30
Batch size	8
Early Stopping Patience	0.5

Table 7: Parameters for Bidirectional LSTM, ALCOA principle Accurate

The results without any normalization:

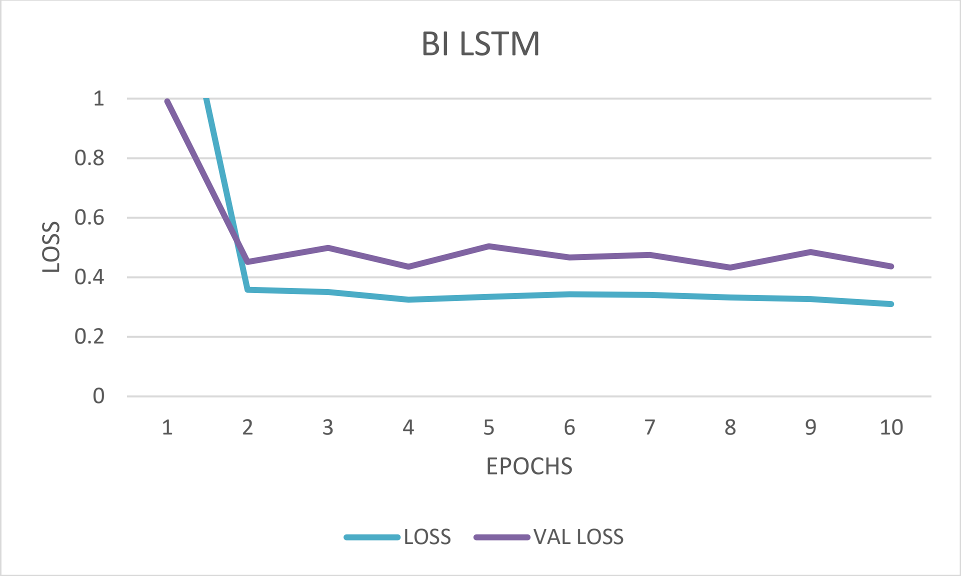


Figure 20: Bidirectional LSTM Accurate I1000

The Mean Absolut Error was 3.5625 and the Validation Mean Absolut Error was 2.2391. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

The results while using Mimax Scaling Normalization:

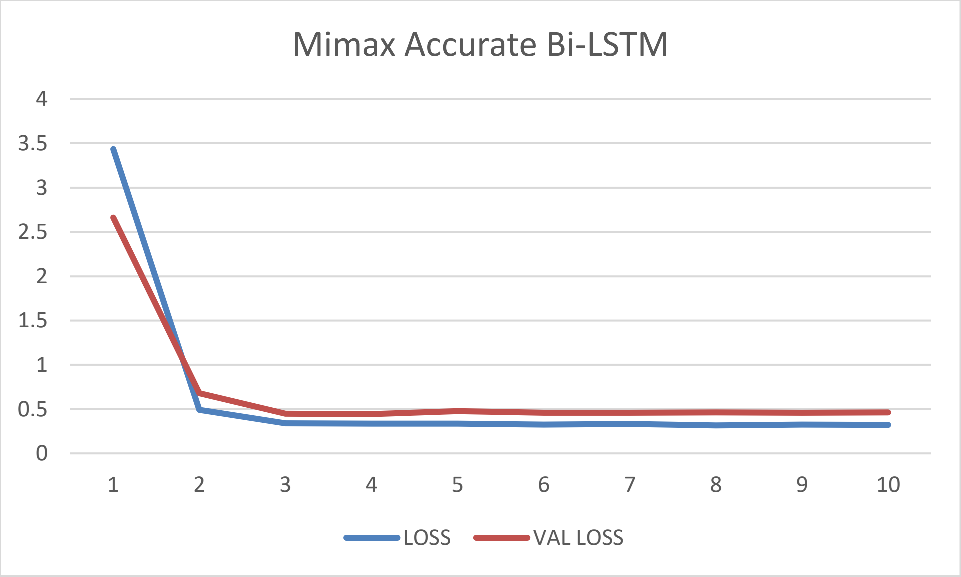


Figure 21: Bidirectional LSTM Accurate using Minmax Scaling I1000

The Mean Absolut Error was 3.5774 and the Validation Mean Absolut Error was 2.2982. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.
The results while using Standard Scaling Normalization:

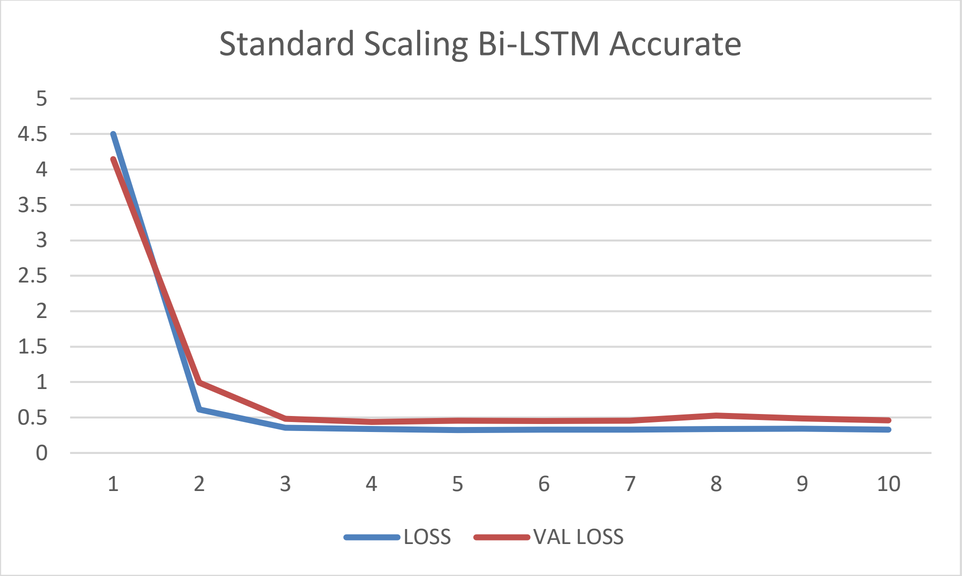


Figure 22: Bidirectional LSTM Accurate using Standard Scaling I1000

The Mean Absolut Error was 3.6161 and the Validation Mean Absolut Error was 2.2927. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

Deep Learning Method: GRU

Parameters	Values
Optimizer	Adam
Learning Rate	1e-4
Droupout	0.2
LSTM units	400
Epochs	25
Batch size	16
Early Stopping Patience	0.4

Table 8: Parameters for GRU, ALCOA principle Accurate

The results without any normalization:

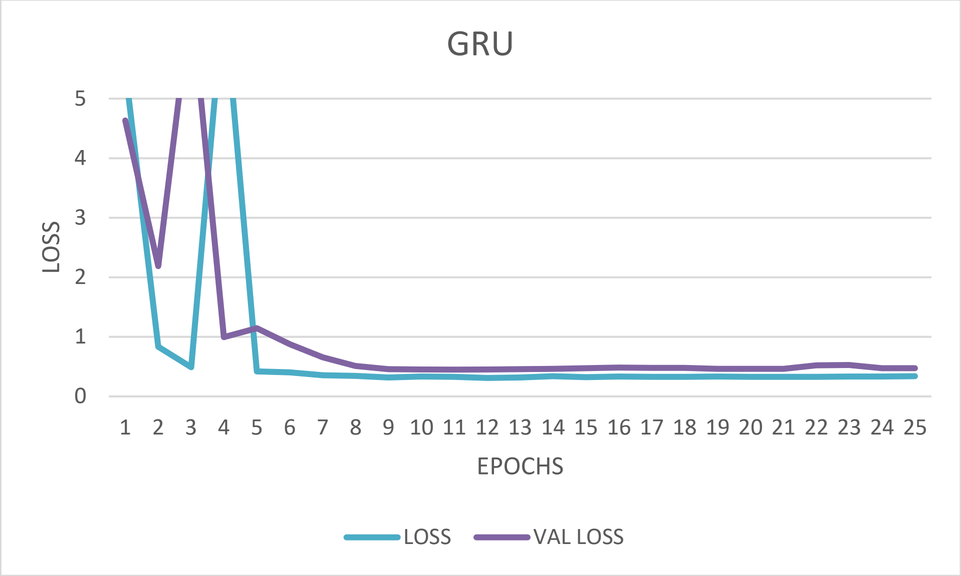


Figure 23: GRU Accurate I1000

The Mean Absolut Error was 3.5783 and the Validation Mean Absolut Error was 2.3065 In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.
The results while using Mimax Scaling Normalization:

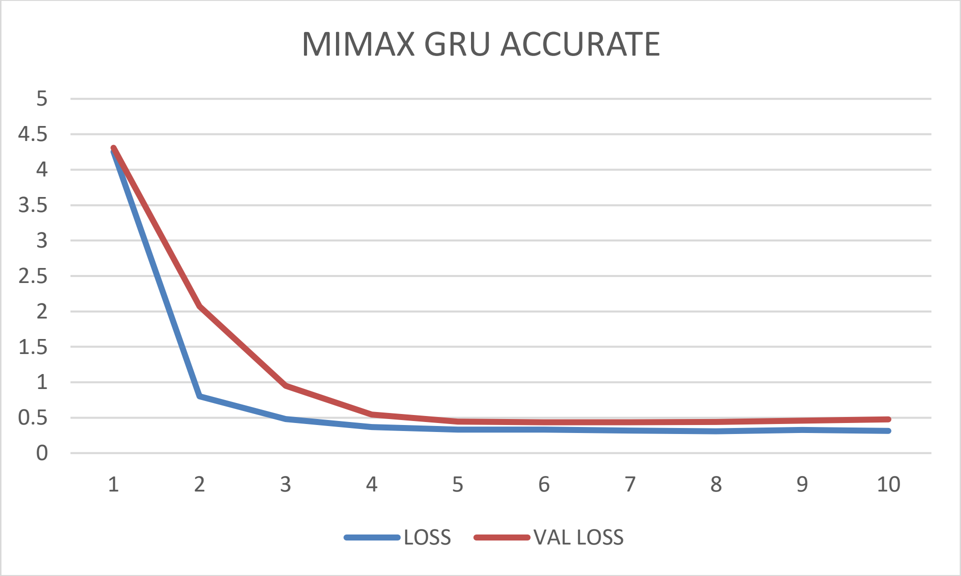


Figure 24: GRU Accurate Using Minmax Scaling I1000

The Mean Absolut Error was 3.4892 and the Validation Mean Absolut Error was 2.309. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

The results while using Standard Scaling Normalization:

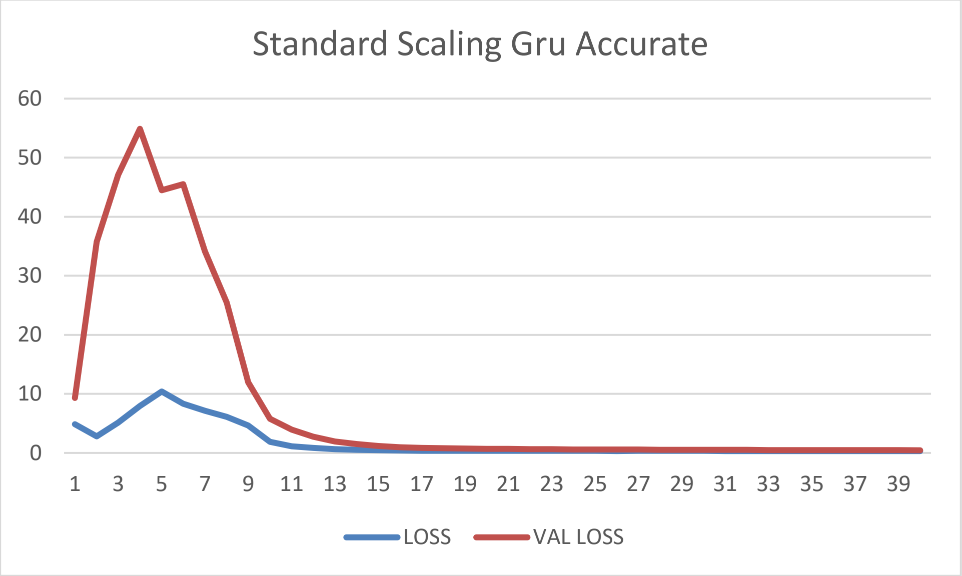


Figure 25: GRU Accurate Using Standard Scaling I1000

The Mean Absolut Error was 3.6919 and the Validation Mean Absolut Error was 2.2927. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

Deep Learning Method: LSTM

Parameters	**Values**
Optimizer	Adam
Learning Rate	1e-4
Droupout	0.2
LSTM units	400
Epochs	20
Batch size	8
Early Stopping Patience	0.4

Table 9: Parameters for LSTM, ALCOA principle Accurate

The results without any normalization:

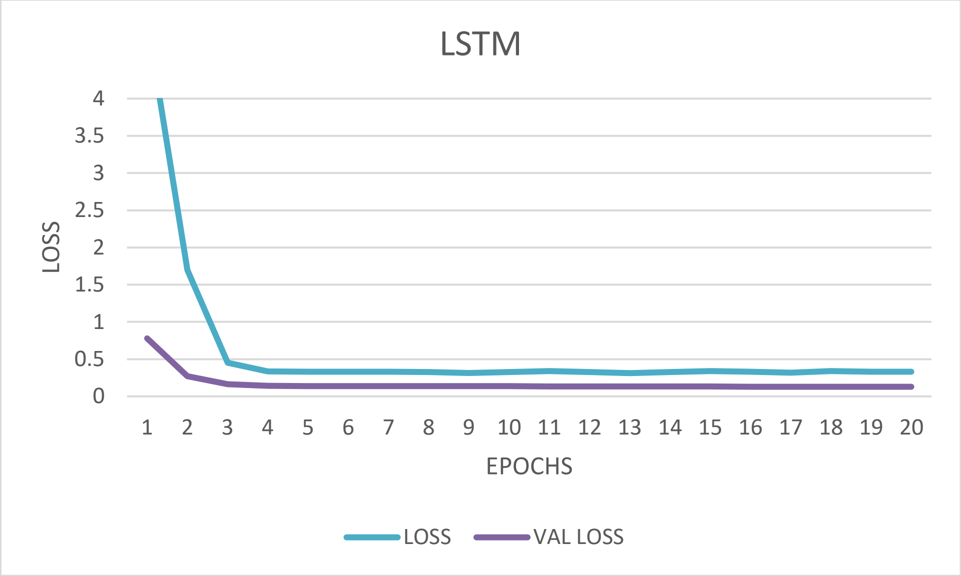


Figure 26: LSTM Accurate I1000

The Mean Absolut Error was 3.6226 and the Validation Mean Absolut Error was 2.4659. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

The results while using Minmax Scaling Normalization:

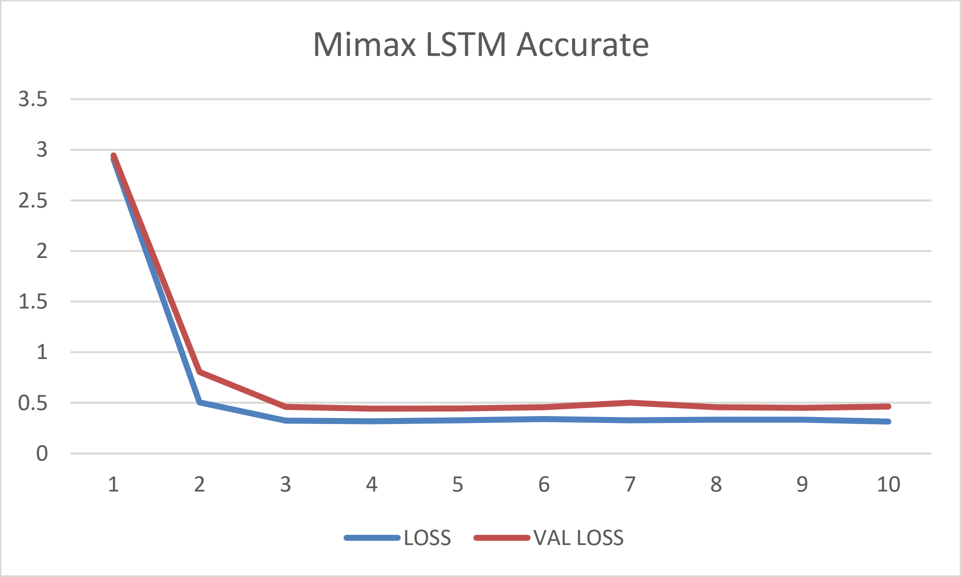


Figure 27: LSTM Accurate Using Minmax Scaling I1000

The Mean Absolut Error was 3.6111 and the Validation Mean Absolut Error was 2.2972. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

The results while using Standard Scaling Normalization:

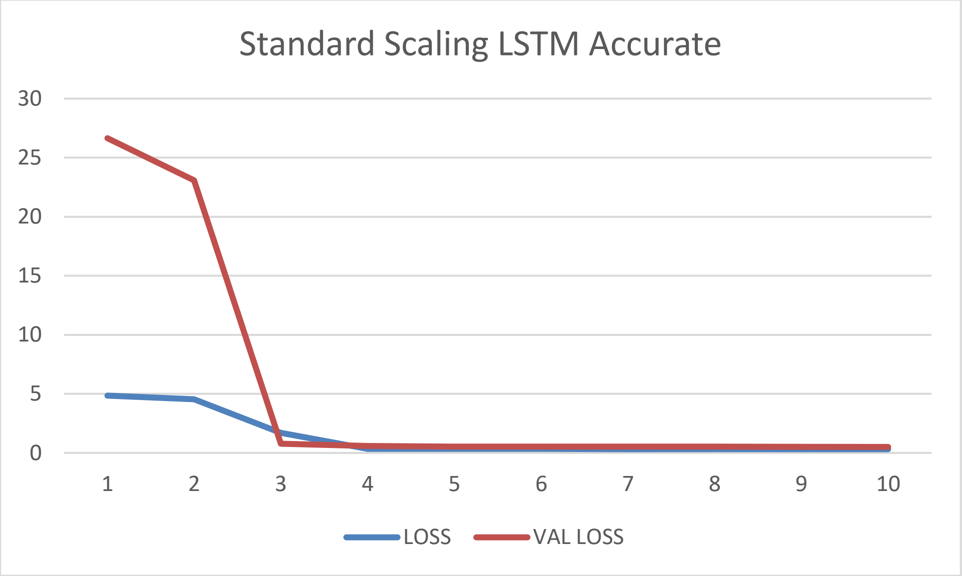


Figure 28: LSTM Accurate Using Standard Scaling I1000

For the production line I600

The value of the Standard deviation is equal to 5.738762.

Deep Learning Method: Bidirectional LSTM

Parameters	**Values**
Optimizer	Adam
Learning Rate	1e-3
Droupout	0.2
LSTM units	200
Epochs	10
Batch size	8
Early Stopping Patience	0.4

Table 10: Parameters for Bidirection LSTM, ALCOA principle Accurate

The results without any normalization:

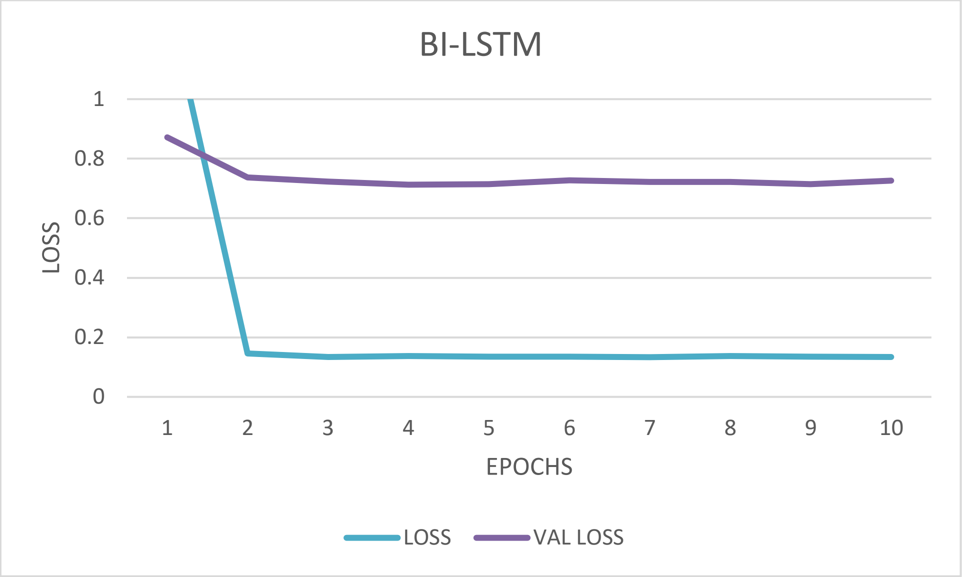


Figure 29: Bidirectional LSTM Accurate I600

The Mean Absolut Error was 3.0026 and the Validation Mean Absolut Error was 3.1362. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

The results while using Mimax Scaling Normalization:

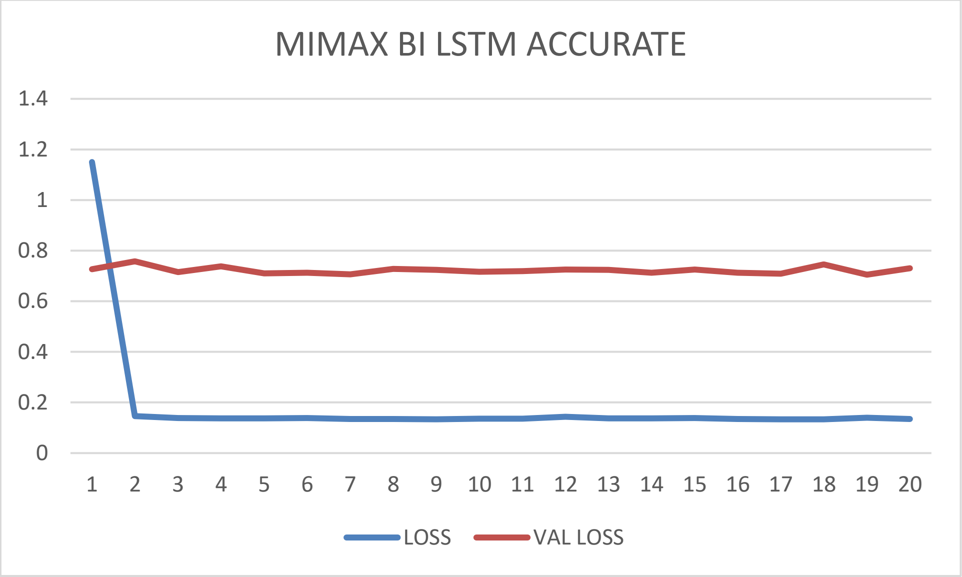


Figure 30: Bidirectional LSTM Accurate using Minmax Scaling I600

The Mean Absolut Error was 3.0126 and the Validation Mean Absolut Error was 3.1382. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

The results while using Standard Scaling Normalization:

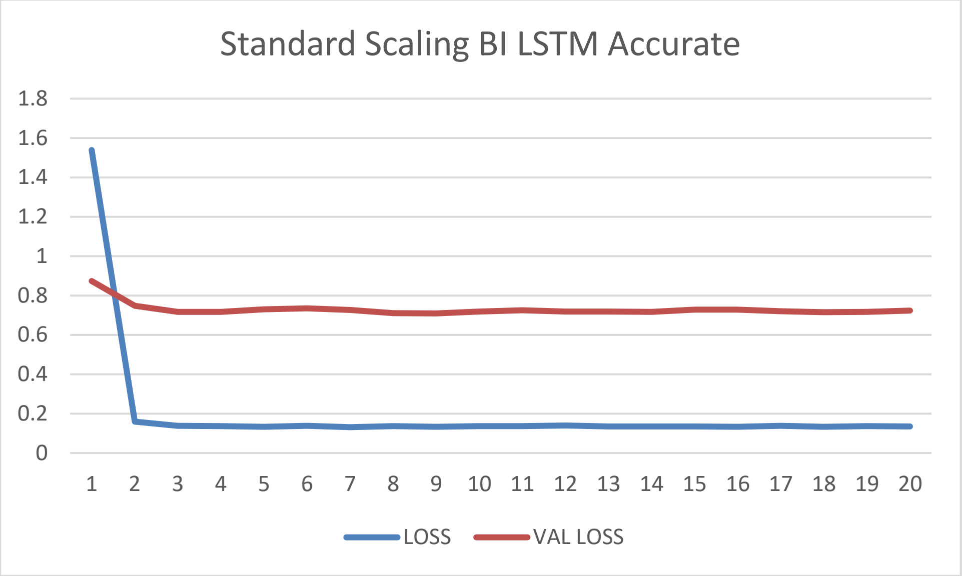


Figure 31: Bidirectional LSTM Accurate using Standard Scaling I600

The Mean Absolut Error was 3.0516 and the Validation Mean Absolut Error was 3.1468. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

Deep Learning Method: GRU

Parameters	**Values**
Optimizer	Adam
Learning Rate	1e-4
Droupout	0.2
LSTM units	200
Epochs	10
Batch size	8
Early Stopping Patience	0.4

Table 11: Parameters for GRU, ALCOA principle Accurate

The results without any normalization:

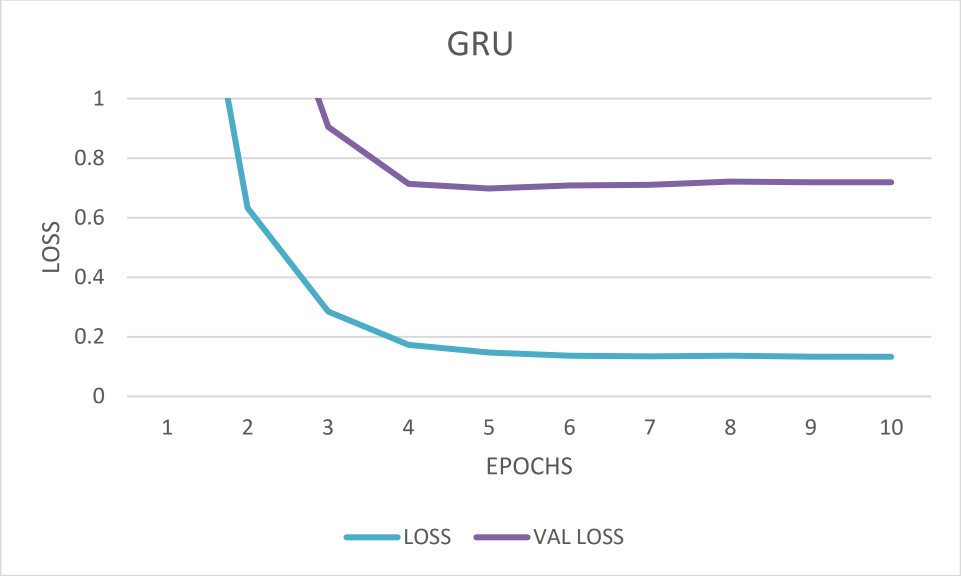


Figure 32: GRU Accurate I600

The Mean Absolut Error was 3.1786 and the Validation Mean Absolut Error was 3.1169. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

The results while using Mimax Scaling Normalization:

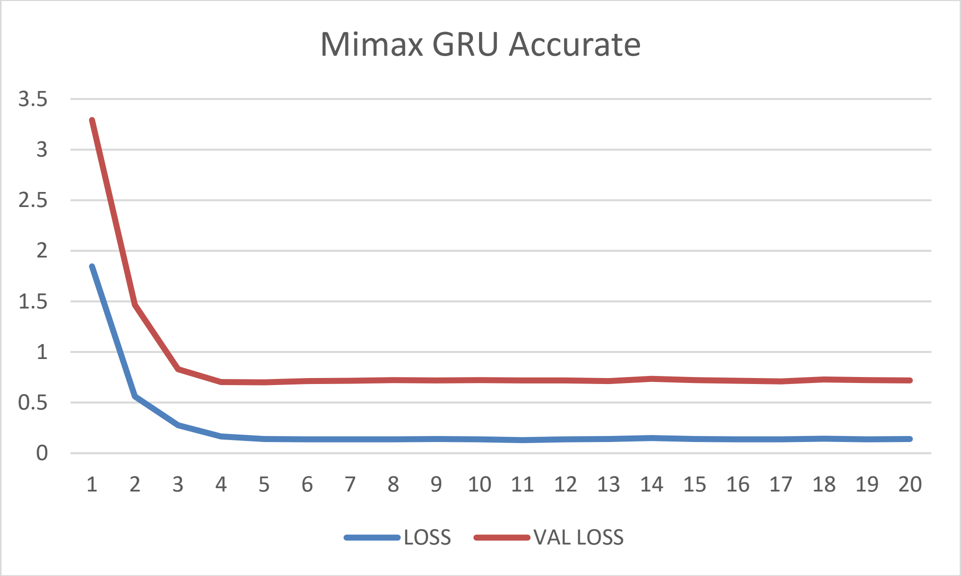


Figure 33: GRU Accurate Using Minmax Scaling I600

The Mean Absolut Error was 3.1886 and the Validation Mean Absolut Error was 3.1269. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

The results while using Standard Scaling Normalization:

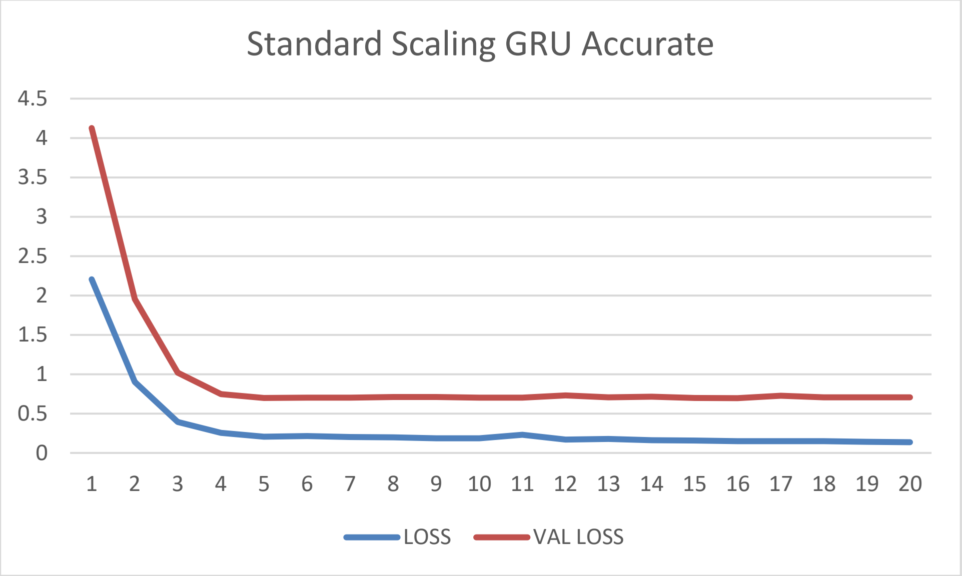


Figure 34: GRU Accurate Using Standard Scaling I600

The Mean Absolut Error was 3.0736 and the Validation Mean Absolut Error was 3.1212. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

Deep Learning Method: LSTM

Parameters	Values
Optimizer	Adam
Learning Rate	1e-4
Droupout	0.2
LSTM units	200
Epochs	40
Batch size	8
Early Stopping Patience	0.4

Table 12: Parameters for LSTM, ALCOA principle Accurate

The results without any normalization:

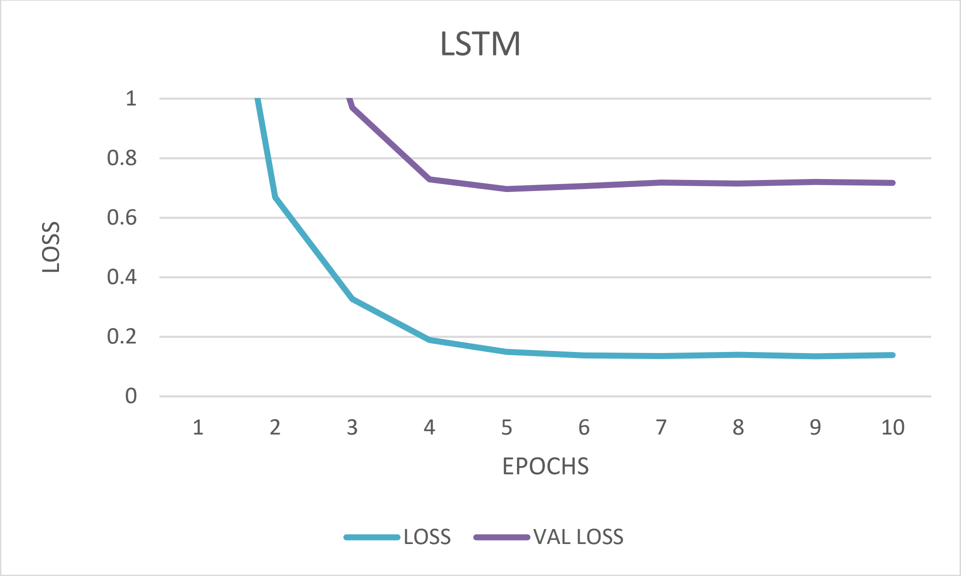


Figure 35: LSTM Accurate I600

The Mean Absolut Error was 3.0747 and the Validation Mean Absolut Error was 3.1085. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

The results while using Mimax Scaling Normalization:

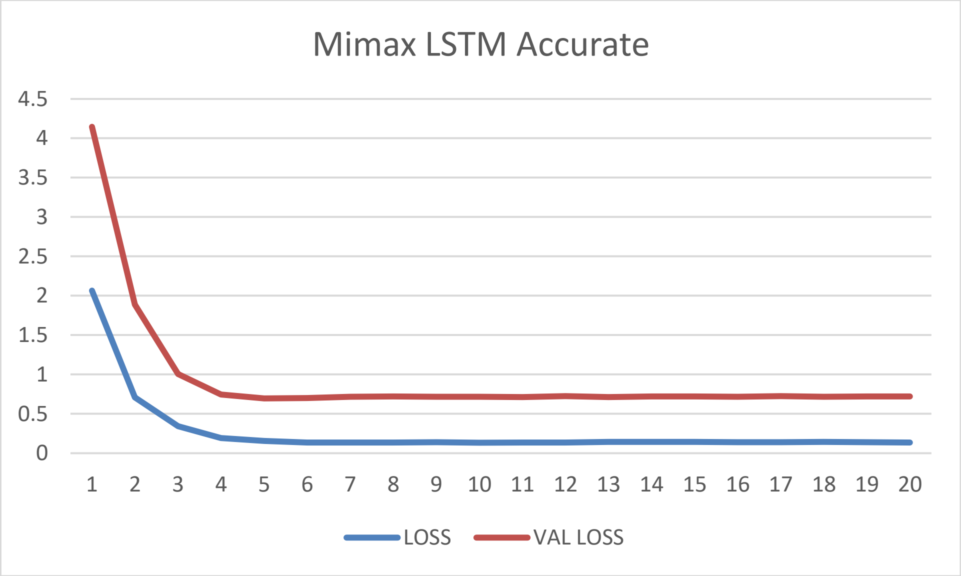


Figure 36: LSTM Accurate Using Minmax Scaling I600

The Mean Absolut Error was 3.0747 and the Validation Mean Absolut Error was 3.1085. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

The results while using Standard Scaling Normalization:

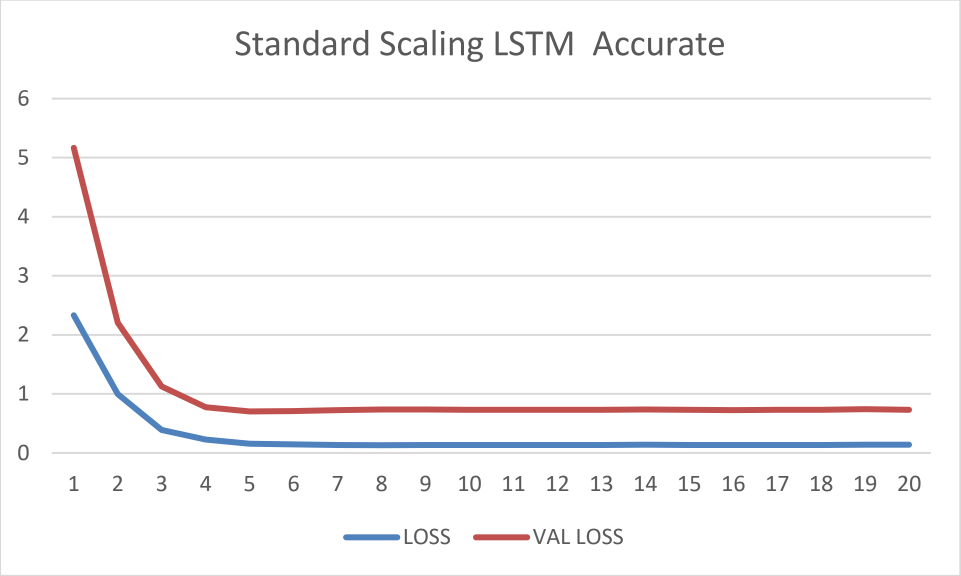


Figure 37: LSTM Accurate Using Standard Scaling I600

The Mean Absolut Error was 3.0736 and the Validation Mean Absolut Error was 3.121. In comparison to the STD, the MAE and the validation MAE shows that the algorithm can both learn and predict the values. These results cannot be used in a pharmaceutical industry.

3.3 Results Analysis

After the evaluation of the results the following conclusions emerged:

Regarding the ALCOA principle Legible, there is not a direct connection between the sensor data and the value of the ALCOA. For the Production line I1000, the STD was 4. The best result that the algorithm managed to get was approximately 3.7 for the Mean Absolut Error and 2.3 for the Validation Mean Absolut Error. This result was without using any normalization on the data and the deep learning method used was Bidirectional LSTM. For the Production line I600, the STD was 1.7. The best result that the algorithm managed to get was approximately 2.3 for the Mean Absolut Error and 1.3 for the Validation Mean Absolut Error. This result was without using any normalization on the data and the deep learning method used was Bidirectional LSTM. There were not many experiments done for the ALCOA principle Legible because it was obvious from the begging that the results could not have an impact on the pharmaceutical industry. In a pharmaceutical industry, when taking drugs into account and the impact they have on the customers, there cannot be such high inaccuracy when it comes to predicting values that affect the drug's data integrity. We believe that the results of ALCOA principle Legible is due to the definition it has as a principle. Having the data either digitally available or in any other form, does not necessary affect the sensor values.

Regarding the ALCOA principle Accurate because of its definition, that has to do with sensor errors, we thought that it had better chances of getting a higher accuracy. Through multiple coding experiments, we managed to get better results for the principle but still they were not accurate enough to immediately be used in an industry. For the Production line I1000, the STD was 6.4. The best result that the algorithm managed to get was approximately 3.7 for the Mean Absolut Error and 2.3 for the Validation Mean Absolut Error. This result was while using standard scaling normalization on the data and the deep learning method used was GRU.

For the Production line I600, the STD was 5.7. The best result that the algorithm managed to get was approximately 3.1 for the Mean Absolut Error and 3.1 for the Validation Mean Absolut Error. This result was by using Minmax scaling normalization on the data and the deep learning method used was LSTM. Due to the high STD and the results we managed to achieve, we believe that by using transformers neural network the results could be even better and could be used by companies or the FDA.

Conclusion

In conclusion, we believe that the subject of the project should be a step for more and more experiments. Data quality and integrity are both necessary when it comes to pharmaceutical industries. Having a variety of tools for securing those domains and preventing accidents can be a turning point for every company.

However, the results presented in this thesis are accurate enough to be used in a pharmaceutical company. Specifically, regarding the ALCOA principle Legible, it was obvious that there could not be an algorithm predicting its value while using sensor data. From the other hand, the ALCOA principle Accurate had better results and we tend to think that there can be better results making it useful for a company to use it. Unfortunately, concerning this thesis and the efforts we made to find a better solution, the outcome of the Neural Networks could not get any lower that the one presented.

Future Research

Since working on an innovative subject, there were plenty of ideas for future experimenting, however, four of them stand out:

- To begin with, we believe that all the ALCOA values predicted in this project can have a better result when using a transformers neural network.
- Furthermore, we would like to try if the ALCOA values have a sequence and can be predicted by taking in mind the previous values of every ALCOA and the data sensor values. For this we made a small experiment and we found out that each ALCOA value has no sequence by itself.
- Moreover, there can probably be an algorithm that can predict the alarms that will take place in a pharmaceutical industry based on the sensor data and the ALCOA values.
- Lastly, there could be a project that tries to predict the ALCOA values not only by using sensor data but also by using data from recipes, note data from employees through Natural Language Processing and production data such as the production time or the automated messages that the machines tend to show in the end of a production stage.

Bibliography – References – Online sources

Ahmad, S., Hafeez, A. & Kumar, A., 2019. Importance of data integrity & its regulation in pharmaceutical industry. *The Pharma Innovation Journal,* 8(1), pp. 306-313.

Albawi, S. & Tareq Abed, M., 2017. Understanding of a Convolutional Neural Network. *IEEE* .

AlphaFold, 2020. *Deepmind.* [Online]
Available at: https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology
[Accessed 20 08 2022].

Anon., 1997. *Deep Blue.* [Online]
Available at: https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/
[Accessed 20 08 2022].

Anon., 2020. *Pharma, Health and Nutrition.* [Online]
Available at: https://pharmanhealth.com/2020/12/17/data-integrity-in-the-pharmaceutical-industry/
[Accessed 08 24 2022].

Christodoulos, L., 2020. *Alan Turing, ο μαθηματικός που μείωσε τον Β' Παγκόσμιο Πόλεμο κατά 2 χρόνια..* [Online]
Available at: http://lazaris.net/lazaris/chris/alan-turing/
[Accessed 20 08 2022].

El Naqa, I. & Murphy, M., 2015. What Is Machine Learning?. *Springer International Publishing Switzerland* , pp. 1-11.

Georgouli, K., 2015. *Τεχνητη Νοημοσυνη Μια Εισαγωγικη Προσεγγιση.* Αθηνα: Σύνδεσμος Ελληνικων Ακαδημαϊκων Βιβλιοθηκών Εθνικό Μετσόβιο Πολυτεχνείο.

Girard, S. & Watkin, A., 2020. *Eurotherm.* [Online]
Available at: https://www.eurotherm.com/life-sciences-cpg/data-integrity-life-sciences/alcoa/
[Accessed 24 08 2022].

González-Vélez, H., April 2021. *The SPuMoNI project.* Praha, European Commission.

Graves , A., Jaitly, N. & Abdel-rahman, M., 2013. Hybrid Speevh Recognition With Deep Bidirectional LSTM. *IEE,* pp. 273-278.

Graves, A. & Jurgen, S., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Elsevier,* p. 602–610.

IMSL, 2021. *IMSL.* [Online]
Available at: https://www.imsl.com/blog/what-is-regression-model
[Accessed 10 08 2022].

Kamilaris, A., X, F. & Prenafeta-Boldu, 2018. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture,* pp. 70-90.

Knox, M., 2022. *NVIDIA.* [Online]
Available at: https://www.msi.umn.edu/sites/default/files/DeepLearningDemystified-UMN.pdf
[Accessed 20 08 2022].

Leal, F. et al., 2021. Smart Pharmaceutical Manufacturing: Ensuring End-to-End Traceability and Data Integrity in Medicine Production. *Big Data Research*, 15 May, p. 24.

LeCun, Y., Bengio, Y. & Hinton, G., 2015. Deep learning. *Nature,* 25 May, pp. 436-447.

Lee, Y., Pipino, L., Strong, D. & Wang, R., 2004. *Process-Embedded Data Integrity.* Hershey: IDEA GROUP PUBLISHING.

Lynn, H. M., Pan, S. B. & Kim, P., 2019. A Deep Bidirectional GRU Network Model for Biometric Electrocardiogram Classification Based on Recurrent Neural Networks. *IEEE,* 21 August.

McCarthy, J., Minsky, M., Herbert , S. & Newell, A., 1955. *A Proposal For The Darthmouth Summer Research Project On Artificial Intelligence.* [Online]
Available at: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html
[Accessed 20 08 2022].

Panagiota, Z., 2019. *Βαθια Μαθηση Απο Ιατρικες Εικόνες.* Πατρα: Πανεπιστήμιο Πατρων.

Pang, B., Nijkamp, E. & Nian Wu, Y., 2019. Deep Learning With TensorFlow: A Review. *Journal of Educational and Behavioral Statistics,* pp. 1-22.

Panik, M., 2022. *Statistics How To.* [Online]
Available at: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-squared-error/
[Accessed 08 10 2022].

Peter, D., 2011. *Kasparov on AI & chess.* [Online]
Available at: https://www.chess.com/news/view/kasparov-on-ai-chess
[Accessed 20 08 2022].

Rahul, D. & Fathi, S., 2017. Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks. *IEEE,* pp. 1597-1600.

Rattan, A., 2018. Data Integrity: History, Issues, and Remediation of Issues. *PDA Journal,* pp. 105-116.

Rosenblatt, F., 1958. The Perception. *Cornell Aeronautical Laboratory*, pp. 386-408.

Roznoski, S., 2014. *How Paper and Electronic Source Data Meet ALCOA Elements.* [Online]
Available at: https://www.advarra.com/blog/paper-electronic-source-data-meet-alcoa/
[Accessed 24 08 2022].

Rumelhart, Hinton & Williams, 1986. Learning representations by back-propagating errors. *Nayure,* 9 10, pp. 533-536.

Rusk, N., 2016. Deep learning. *Nature Methods*, January, p. 35.

Sak, H., Senior, A. & Beaufays, F., 2014. *Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling,* USA: Google.

Shinde, P. & Shah, S., 2018. A Review of Machine Learning and Deep Learning Applications. *Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).*

ter Braak, C. & Looman, C., 1986. Regression. In: *Weighted averaging, logistic regression and the Gaussian response model.* s.l.:s.n.

W3Scools, n.d. *W3Scools.* [Online]
Available at: https://www.w3schools.com/python/python_ml_linear_regression.asp
[Accessed 08 10 2022].

Wuyan, L. et al., 2020. Prediction of dissolved oxygen in a fishery pond based on gated recurrent unit (GRU). *Information Processing in Agriculture,* 10 February.

Xiaosheng, S., Changhua, H. & Jianxun, Z., 2019. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, pp. 1235-1270.

Appendix A

Code For production line I600 ALCOA principles Legible and Accurate

Appendix B

Code For production line I1000 ALCOA principles Legible and Accurate