Learning Regulatory Compliance Data for Data Governance in Financial Services Industry by Machine Learning Models

Ka Yee Wong

A Thesis in the Fulfillment of the Requirement for the
Degree of Master of Philosophy

School of Computer Science and Engineering

Faulty of Engineering

The University of New South Wales

September 2020

Thesis Title

Learning Regulatory Compliance Data for Data Governance in Financial Services Industry by Machine Learning Models

Thesis Abstract

While regulatory compliance data has been governed in the financial services industry for a long time to identify, assess, remediate and prevent risks, improving data governance ("DG") has emerged as a new paradigm that uses machine learning models to enhance the level of data management.

In the literature, there is a research gap. Machine learning models have not been extensively applied to DG processes by a) predicting data quality ("DQ") in supervised learning and taking temporal sequences and correlations of data noise into account in DQ prediction; b) predicting DQ in unsupervised learning and learning the importance of data noise jointly with temporal sequences and correlations of data noise in DQ prediction; c) analyzing DQ prediction at a granular level; d) measuring network run-time saving in DQ prediction; and e) predicting information security compliance levels.

Our main research focus is whether our ML models accurately predict DQ and information security compliance levels during DG processes of financial institutions by learning regulatory compliance data from both theoretical and experimental perspectives.

We propose five machine learning models including a) a DQ prediction sequential learning model in supervised learning; b) a DQ prediction sequential learning model with an attention mechanism in unsupervised learning; c) a DQ prediction analytical model; d) a DQ prediction network efficiency improvement model; and e) an information security compliance prediction model.

Experimental results demonstrate the effectiveness of these models by accurately predicting DQ in supervised learning, precisely predicting DQ in unsupervised learning, analyzing DQ prediction by divergent dimensions such as risk types and business segments, saving significant network run-time in DQ prediction for improving the network efficiency, and accurately predicting information security compliance levels.

Our models strengthen DG capabilities of financial institutions by improving DQ, data risk management, bank-wide risk management, and information security based on regulatory requirements in the financial services industry including Basel Committee on Banking Supervision Standard Number 239, Australia Prudential Regulation Authority ("APRA") Standard Number CPG 235 and APRA Standard Number CPG 234. These models are part of DG programs under the DG framework of financial institutions.

ORIGINALITY STATEMENT

☑ I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

COPYRIGHT STATEMENT

AUTHENTICITY STATEMENT

☑ I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in the candidate's thesis in lieu of a Chapter provided:

- The candidate contributed **greater than 50%** of the content in the publication and are the "primary author", i.e. they were responsible primarily for the planning, execution and preparation of the work for publication.
- The candidate has obtained approval to include the publication in their thesis in lieu of a Chapter from their Supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis.

☑ The candidate has declared that **their thesis has publications - either published or submitted for publication - incorporated into it in lieu of a Chapter/s. Details of these publications are provided below.**.

Publication Details #1

Full Title:	Learning Data Quality Analytics for Financial Services
Authors:	Ka Yee Wong, Raymond K. Wong, and Haojie Huang
Journal or Book Name:	Pacific Asia Conference on Information Systems (PACIS)
Volume/Page Numbers:	549
Date Accepted/Published:	18 April 2019
Status:	published
The Candidate's Contribution to the Work:	After identifying the existing issues and predicting potential issues, financial institutions will understand the room for improvement in the quality of risk data in real world. This allows to remediate data as early as possible to mitigate the risk of reoccurrence. Accordingly, their analytical reports can be relied upon. But the key is the measurement of data quality in alignment with the Australia regulatory requirement, Australia Prudential Regulation Authority Standard Number CPG 235, before analytics.
Location of the work in the thesis and/or how the work is incorporated in the thesis:	The work is incorporated in the thesis as Chapter 5.

Publication Details #2

Full Title:	Optimized Sequence Prediction of Risk Data for Financial Institutions
Authors:	Ka Yee Wong, Raymond, K Wong, and Haojie Huang
Journal or Book Name:	Pacific Rim International Conference on Artificial Intelligence (PRACAI)
Volume/Page Numbers:	364-378
Date Accepted/Published:	4 June 2019
Status:	published
The Candidate's Contribution to the Work:	Key contributions are the resolution of current problem by developing a model for sequence predictions with optimizations and the implementation of Long Short Term Memory (LSTM) Recurrent Neural Networks (RNNs) with divergent methodologies (forward, backward and Bi-directional LSTM RNNs) and algorithms (ADAM, SGD, ADADELTA, ADAGRAD & RMSPROP) in alignment with the international regulatory requirement of Basel Committee on Banking Supervision Standard Number BCBS 239.
Location of the work in the thesis and/or how the work is incorporated in the thesis:	The work is incorporated in the thesis as Chapter 3

Publication Details #3

Full Title:	An Efficient Risk Data Learning with LSTM RNN
Authors:	Ka Yee Wong and Raymond, K Wong
Journal or Book Name:	32th Australasian Conference on Artificial Intelligence and 17th Australasian Data Mining Conference
Volume/Page Numbers:	116-128
Date Accepted/Published:	19 September 2019
Status:	published
The Candidate's Contribution to the Work:	We demonstrate an improvement in the performance of machine learning networks for the big risk data prediction – run time, accuracy and loss. This enables financial institutions to find the good and bad quality of data swiftly from the big data. Besides, they can rely on the model to analyse the quality issues with a fraction of data. Given this, they can remediate data earlier for risk management. Additionally, the initial problem stated in the paper is remedied - to analyse and predict them takes tremendous amount of time.
Location of the work in the thesis and/or how the work is incorporated in the thesis:	The work is incorporated in the thesis as Chapter 6

Publication Details #4

Full Title:	Learning System Security Compliance for Banking
Authors:	Ka Yee Wong and Raymond, K Wong
Journal or Book Name:	Pacific Asia Conference on Information Systems (PACIS)
Volume/Page Numbers:	193
Date Accepted/Published:	25 April 2020
Status:	published
The Candidate's Contribution to the Work:	We made contribution by developing a checklist of information security controls to meet the regulatory requirement, Australia Prudential Regulation Authority Standard Number CPG 234 (CPG234), a feasible machine learning model automating the system security compliance process in a real time mode, and analysing security control weaknesses in a regulatory compliance report through a neural network experiment. Whenever banks adopt our model, they can strengthen the capability in achieving a full compliance with the CPG 234.
Location of the work in the thesis and/or how the work is incorporated in the thesis:	The work is incorporated in the thesis as Chapter 7

Publication Details #5

Full Title:	Big Data Quality Prediction on Banking Applications
Authors:	Ka Yee Wong and Raymond, K Wong
Journal or Book Name:	IEEE International Conference on Data Science and Advanced Analytics (DSAA)
Volume/Page Numbers:	20-00145
Date Accepted/Published:	22 August 2020
Status:	published
The Candidate's Contribution to the Work:	We made contribution by the following ways: Proposing a model to detect noises of big data based on the industry standard, Basel Committee on Banking Supervision Standard Number BCBS 239, instead of traditional approaches mainly centering on data accuracy, completeness and timeliness; Weighing data noises in an unsupervised learning based on their correlations and importance, unlike conventional data clustering without taking the overlap of data and their correlations into account; Experimenting neural networks with a multidimensional learning of data quality (DQ) including the feedforwarding learning, the backward learning and the bi-directional learning with a focus on relatively important big data, contrary to previous forecasts primarily on matched or mis-matched data and missing data; and Remediating poor risk data earlier to improve the DQ with predicted DQ. This contributes to the quality of risk reports used for the risk management. In the same way, this brings the value of big data to their businesses. The value can be measured in terms of the compliance level with the BCBS 239 requirements and analysed in dashboards for the management decision making.
Location of the work in the thesis and/or how the work is incorporated in the thesis:	The work is incorporated in the thesis as Chapter 1, 2 and 4

Publication Details #6

Full Title:	Big Data Quality Prediction on Banking Applications
Authors:	Ka Yee Wong and Raymond, K Wong
Journal or Book Name:	International Journal of Data Science and Analytics/ Journal of Data Science and Analytics (JDSA)
Volume/Page Numbers:	JDSA-D-20-00145
Date Accepted/Published:	22 August 2020
Status:	published
The Candidate's Contribution to the Work:	We made contribution by the following ways: Proposing a model to detect noises of big data based on the industry standard, Basel Committee on Banking Supervision Standard Number BCBS 239, instead of traditional approaches mainly centering on data accuracy, completeness and timeliness; Weighing data noises in an unsupervised learning based on their correlations and importance, unlike conventional data clustering without taking the overlap of data and their correlations into account; Experimenting neural networks with a multidimensional learning of data quality (DQ) including the feedforwarding learning, the backward learning and the bi-directional learning with a focus on relatively important big data, contrary to previous forecasts primarily on matched or mis-matched data and missing data; and Remediating poor risk data earlier to improve the DQ with predicted DQ. This contributes to the quality of risk reports used for the risk management. In the same way, this brings the value of big data to their businesses. The value can be measured in terms of the compliance level with the BCBS 239 requirements and analysed in dashboards for the management decision making.
Location of the work in the thesis and/or how the work is incorporated in the thesis:	The work is incorporated in the thesis as Chapter 1, 2 and 4

Candidate's Declaration

I confirm that where I have used a publication in lieu of a chapter, the listed publication(s) above meet(s) the requirements to be included in the thesis. I also declare that I have complied with the Thesis Examination Procedure.

Abstract

While regulatory compliance data has been governed in the financial services industry for a long time to identify, assess, remediate and prevent risks, improving data governance ("DG") has emerged as a new paradigm that uses machine learning models to enhance the level of data management.

In the literature, there is a research gap. Machine learning models have not been extensively applied to DG processes by a) predicting data quality ("DQ") in supervised learning and taking temporal sequences and correlations of data noise into account in DQ prediction; b) predicting DQ in unsupervised learning and learning the importance of data noise jointly with temporal sequences and correlations of data noise in DQ prediction; c) analyzing DQ prediction at a granular level; d) measuring network run-time saving in DQ prediction; and e) predicting information security compliance levels.

Our main research focus is whether our ML models accurately predict DQ and information security compliance levels during DG processes of financial institutions by learning regulatory compliance data from both theoretical and experimental perspectives.

We propose five machine learning models including a) a DQ prediction sequential learning model in supervised learning; b) a DQ prediction sequential learning model with an attention mechanism in unsupervised learning; c) a DQ prediction analytical model; d) a DQ prediction network efficiency improvement model; and e) an information security compliance prediction model.

Experimental results demonstrate the effectiveness of these models by accurately predicting DQ in supervised learning, precisely predicting DQ in unsupervised learning, analyzing DQ prediction by divergent dimensions such as risk types and business segments, saving significant network run-time in DQ prediction for improving the network efficiency, and accurately predicting information security compliance levels.

Our models strengthen DG capabilities of financial institutions by improving DQ, data risk management, bank-wide risk management, and information security based on regulatory requirements in the financial services industry including Basel Committee on Banking Supervision Standard Number 239, Australia Prudential Regulation Authority ("APRA") Standard Number CPG 235 and APRA Standard Number CPG 234. These models are part of DG programs under the DG framework of financial institutions.

Publications

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in the candidate's thesis in lieu of a Chapter provided:

- The candidate contributed **greater than 50%** of the content in the publication and are the "primary author", i.e. they were responsible primarily for the planning, execution and preparation of the work for publication.
- The candidate has obtained approval to include the publication in their thesis in lieu of a Chapter from their Supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis.

☑ The candidate has declared that **their thesis has publications - either published or submitted for publication - incorporated into it in lieu of a Chapter/s. Details of these publications are provided below.**.

Large portions of Chapter 5 have appeared in the following paper:

a. Ka Yee Wong, Raymond K. Wong, and Haojie Huang. 2019. "Learning Data Quality Analytics for Financial Services," *Pacific Asia Conference Information Systems (PACIS)*, Association for Information Systems, number 549.

Large portions of Chapter 3 have appeared in the following paper:

b. Ka Yee Wong, Raymond K. Wong, and Haojie Huang. 2019. "Optimized Sequence Prediction of Risk Data for Financial Institutions," Pacific Rim *International Conference on Artificial Intelligence (PRICAI)*, Springer, pages 364-378.

Large portions of Chapter 6 have appeared in the following paper:

c. Ka Yee Wong, Raymond K. Wong, and Haojie Huang. 2019. "An Efficient Risk Data Learning with LSTM RNN," *32th Australasian Conference on Artificial Intelligence and 17th Australasian Data Mining Conference*, ISBN No. 978-981-15-1699-3_10, pages 116-128.

Large portions of Chapter 7 have appeared in the following paper:

d. Ka Yee Wong, Raymond K. Wong, and Han Tai. 2020. "Learning System Security Compliance for Banking," *Pacific Asia Conference Information Systems (PACIS)*, Association for Information Systems, number 193.

Large portions of Chapter 1, 2 and 4 have appeared in the following papers:

e. Ka Yee Wong, and Raymond K. Wong. 2020. "Big Data Quality Prediction on Banking Applications," *IEEE International Conference of Data Science and Advanced Analytics (DSAA)*.

f. Ka Yee Wong, and Raymond K. Wong. 2020. "Big Data Quality Prediction on Banking Applications," *International Journal of the Data Science and Analytics (IJDSA)* published by Springer.

Acknowledgement

I would like to offer my sincerest gratitude and appreciation to the kind individuals who have helped me throughout my study.

Firstly, I would like to thank my supervisor and the co-supervisor from the University of New South Wales for their assistance and patience.

I feel grateful for the great support from professors and staff of the University of New South Wales in critical moments such as Fethi Rabhi, Rachid Hamadi and Ali Darejeh.

My sincerest thanks and deep appreciation go to a knowledgeable expert, Haojie Huang, for sharing of his deep technical knowledge and insightful guidance on my machine learning models along with experimental programs.

I am grateful for mathematical advices that I have received from Han Tai who has reviewed complex calculations to ensure the accuracy of model formulas.

Similarly, I would like to thank you for resources provided to me from the University of New South Wales, and a strong support from the school staff, Colin Taylor, during my research period by providing valuable instructions and clear direction whenever needed.

Finally, I would like to show my gratitude to the Graduate Research School by taking proactive actions on the resolution of problems in an effective and efficient manner.

Abbreviations

Acc Prediction Accuracy
AI Artificial intelligence
ANN Artificial Neural Networks
APRA Australia Prudential Regulation Authority
ATTN Attention Mechanism
BCBS Basel Committee on Banking Supervision
BCBS 239 Basel Committee on Banking Supervision 239
BD Bi-Directional
BGMM Bayesian GMM
BK Backward
CPG 234 Australia Prudential Regulation Authority CPG 234
CPG 235 Australia Prudential Regulation Authority Standard Number CPG 235
CR Credit Risk
DE Data Element
DEs Data Elements
DG Data Governance
DGF Data Governance Framework
DGP Data Governance Programs
DSR Design Science Research
DSRM Design Science Research Methodology
DM Data Management
DN Data Noise
DNN Deep Neural Networks
DP Dirichlet Process
DQ Data Quality
DQP Data Quality Prediction
DT Decision Tree
DTs Decision Trees
F1 F1-Support
FF Feedforwarding
FIs Financial institutions
FSB Financial Stability Board
GMM Gaussian Mixture Model
Life-Cycle Information Asset Life Cycle
IS Information Security
ISCs Information Security Controls
ISL Information Security Levels
ISR Information Security Rules
KNN K-Nearest Neighbors
Loss Prediction Error
LR Liquidity Risk
LReg Logistic Regression
LSTM Long Short-Term Memory

MBB Memory Between Batches
ML Machine Learning
MR Market Risk
MSE Mean Square Error
NB Naïve Bayes
OR Operational Risk
PDF Probability Density Function
PN Precision
PvB Private Bank
RB Retail Bank
RF Random Forests
RL Recall
RNN Recurrent Neural Network
RNNs Recurrent Neural Networks
sec Seconds
SGD Stochastic Gradient Descent
SVM Support Vector Machines
Systems Systems, Networks and Information Assets
VAR Explained Variances
WB Wholesale Bank

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis is about learning from regulatory compliance data for supporting the data governance ("DG") processes in the financial services industry. In this chapter, we provide information about DG in the industry, and DG coverage along with relevant regulatory requirements. We then explain the thesis motivation by describing opportunities offered by machine learning ("ML") work for DG. Following this, we propose the thesis scope, aims, models, model networks before presenting the research methodology. Finally, we conclude the chapter by introducing upcoming chapters.

1.1 Background and Motivation

1.1.1 Data Governance in the Financial Services Industry

Industry Characteristics

In this financial services industry, regulatory compliance data has been governed for a long time under various DG processes to identify, assess and prevent risks [151, 152, 153].

In this study, DG refers to policies and procedures geared towards the management of usability, availability, integrity and security of data [151, 152]. It is a function owned by the business and executed by business stewards to recognize the value of data to an enterprise and manage it as a company asset. The more data is shared across the company, the more valuable it becomes. This increases the value of data which is the primary goal for DG.

In this industry, massive amounts of data need to be governed [63, 82, 153]. Their significance is classified by business criticality and sensitivity [120]. The classification becomes a benchmark for the determination of information security ("IS") controls. These controls impact IS compliance levels. Generally, critical data is prioritized for early improvements. The data is mostly regulatory compliance related [5, 153] although financial institutions ("FIs") lack powerful tools to aggregate data from various sources [6].

Data in this industry is commonly governed under a data governance framework ("DGF") [153] pursuant to DG policies and procedures. Under such a framework, a series of DG initiatives are launched by FIs as part of meeting their DG objectives. These initiatives are called data governance programs ("DGP").

Data Governance Coverage

In a DG framework, DG rules and policies related to data ownership, data processes, and the level of data quality ("DQ") controls are defined. These rules and policies set out required controls for the whole DG process specifically:

a. lay out data standards for what DQ key performance indicators [153] are required and which data elements ("DEs") are deemed as critical. They include what business rules that are to be adhered to and to be profiled for the quality assessment;
b. set out information security controls ("ISCs") [153] to provide guidance on how to secure data in systems. An example of this is a control over application systems, operating systems and networks;
c. are implemented by a series of DGP to ensure the oversight of DM; and
d. include a business glossary which is a primer to establish the metadata for achieving common data definitions. The glossary is used in the DQ management.

A DGF can be used to maintain and improve the level of DQ and IS during DG processes. It is commonly implemented by launching various DGP.

Data Governance Regulatory Requirements

In launching DGP as part of a framework, FIs are obligated to meet three regulatory requirements related to DG. These requirements defining quantitative DQ in terms of DQ metrics are issued by international and local regulators in the financial services industry.

The first requirement is that from an international perspective, FIs are expected to meet DQ principles in the process of aggregating risk data into risk reports. These principles include the principle number 3 (accuracy and integrity), the principle number 4 (completeness) and the principle number 5 (timeliness) under Basel Committee on Banking Supervision ("BCBS") Standard Number 239 ("BCBS 239") [3]. FIs are recommended to provide forward-looking capabilities of DQ for the improvement of DQ.

In meeting these principles, DQ issues, namely data noise ("DN"), need to be minimized including an omission of translation, a negligence in the data format transformation, and retention of data which is redundant, duplicated, stale, unreasonable, invalid, mis-matched, incomplete or null. These issues are common [3].

The second requirement is that from a local standpoint, FIs are required to manage data risks through the assessment and management of DQ by six dimensions including the dimension of (a) accuracy; (b) completeness; (c) consistency; (d) timeliness; (e) availability; and (f) fitness for use under Australia Prudential Regulation Authority ("APRA") Standard Number CPG 235 ("CPG 235") [120].

The third requirement is that from another local point of view, FIs are expected to manage IS risks under APRA Standard CPG 234 ("CPG 234") [75] by including thirteen IS controls, called ISCs, during the information asset life-cycle ("Life Cycle") under IS policy frameworks.

These regulatory requirements reveal the importance of meeting DG objectives including an improvement of DQ [3], the management of data risk [120] and an enhancement of IS [75] respectively.

1.1.2 Opportunities for Using Machine Learning in Data Governance

In recent years, DG has emerged as a new paradigm that uses ML models to enhance the level of data management ("DM") [151, 153]. Using these models, FIs can strengthen their capabilities to adapt to the changing regulatory compliance data environment [153].

Applying ML models to DG is indispensable for multiple reasons listed below.

a. Tera scale ML is an important ingredient in the quality modeling and monitoring of big data [38]. This covers large scale applications of ML and their computational models;
b. Applying ML by FIs benefits risk management significantly by addressing specific problems including data risk such as data quality risk [141];
c. ML can create accurate methods for data analysis, modeling and prediction by identifying complex and non-linear patterns in huge data sets [142];
d. Making efficient performance prediction for large-scale advanced analytics minimizes the amount of training data required [143]; and
e. ML can play an essential role in the regulatory reporting of the financial services industry by improving the compliance processes [142].

ML models have become critically important in supporting decision making in real time in the financial services industry [53]. In an examination of existing research relating to the application of ML models to DG, these models have not been widely deployed to DG processes in the industry despite the availability of many models.

a. In 2018, the regulator, Financial Stability Board ("FSB") monitoring the global financial system, in this financial services industry recognized the need to leverage ML models for DQ improvement [6]. FSB recommended FIs to provide an assurance on DQ through additional checks with these models. This is similar to the suggestion from

BCBS highlighting the importance of providing forward looking capabilities of DQ [3]. A typical example of this is prediction of DQ;

b. In 2017, many FIs have been striving to build an analytics hub to implement ML for high quality analytics [80]. They recognized the need to exert more efforts on developing advanced analytical capabilities [154]. These events show the significance of leveraging ML for an analysis of DQ including current and future DQ for the compliance purpose;
c. In 2016, FIs have been putting much efforts on the application of artificial intelligence ("AI") to reduce FI compliance costs, called RegTech or Regulatory Technology, by utilizing the power of big data [81]. They trust that AI or ML would be able to swiftly filter bad quality from big data and slide good quality for use [82]. They recognize the need to improve the network efficiency of learning DQ with ML; and
d. In 2019, APRA revised a standard, CPG 234 [75], to guide the management of IS through ISCs in the financial services industry. Until 2020, many banking systems have been attacked by hackers impacting numerous customers [83]. To address this, information security levels ("ISL") need to be enhanced. ML can be used to predict potentially anomalous patterns of threats for the risk management purpose [190]. An example is to forecast IS compliance levels for preventing IS issues from happening [40].

The above discussion reveals problems of DG in the financial services industry and the opportunity to use ML to improve data quality prediction ("DQP"), DQ and DQP analytics capabilities, an improvement in network efficiency of DQP, and prediction of IS compliance levels during DG processes. These problems motivate us for using ML in DQP and IS compliance prediction under DG of the industry.

1.2 Scope, Aims and Models

In facing these DG problems, we describe scope, aims and models of this thesis.

1.2.1 Scope

The scope of the thesis centres on data analysis in connection with DG regulatory requirements: a) risk data as set out in BCBS 239 consisting of market risk ("MR"), credit risk ("CR"), operational risk ("OR") and liquidity risk ("LR"); b) business and operations data as defined in CPG 235 including data used for analytics and intelligence purposes; and c) IS data as specified in CPG 234. As such, data in this thesis is regulatory compliance data. This belongs to big data due to the large volume and a wide variety of data and a high degree of veracity [165, 166].

1.2.2 Aims

The aims of the thesis are to address the data governance problems by applying ML models to DG mentioned in Section 1.1.2. The specific aims of the thesis are listed below correspondingly for the enhancement of DG for FIs:

a. Improving DQ through the initiation of an action plan to remediate deficient DQ [4]. DQP under DG in supervised learning enables to identify poor data in advance;
b. Managing data risk through DQ assessment and management [120]. DQP analysis by multi-dimensions under DG provides a notion on what kinds of data are of low quality;
c. Managing bank-wide risk through building an effective operating model for risk data aggregation and reporting practices [4]. The network efficiency improvement in DQP under DG helps to identify potential poor data swiftly; and
d. Enhancing IS through ISCs implementation [75]. Prediction of ISL under DG provides an early alert to which ISCs are inadequate.

Our main research question is: "Whether DQ and IS compliance levels during DG processes of FIs can be accurately predicted through our ML models by learning regulatory compliance data from both theoretical and experimental perspectives".

1.2.3 Models

To meet the aims, we propose five models which are briefly described below.

a. A DQP model using supervised learning under DG to meet the regulatory requirement of DG. This model considers sequential learning of DN by taking temporal sequences and correlations of DN into account. This DQP enables to identify poor data in advance, as mentioned in item number (a) of the Section 1.2.2;
b. A DQP model using unsupervised learning under DG to meet the regulatory requirement of DG. This model considers the importance of DN on top of the temporal sequences and correlations of DN collectively in DQP. Additionally, this model takes temporal sequences and correlations of DN into account in DQ measurement. This DQP enables to identify poor data in advance, as mentioned in item number (a) of the Section 1.2.2;
c. A DQP analytical model under DG to meet the regulatory requirement of DG. This DQP analysis provides a notion on what kinds of data are of low quality, as mentioned in item number (b) of the Section 1.2.2;
d. A DQP network efficiency improvement model under DG to meet the regulatory requirement of DG by measuring network run-time saving. This network efficiency improvement in DQP helps to identify potential poor data swiftly, as mentioned in item number (c) of the Section 1.2.2; and
e. An ISL prediction model considering sequences, correlations and importance of IS factors collectively for IS compliance prediction under DG to meet the regulatory requirement of DG. This ISL prediction provides an early alert to which ISCs are inadequate, as mentioned in item number (d) of the Section 1.2.2.

These models address limitations of ML work related to DG mentioned in Section 2.2.

Models 1 - Data Quality Prediction in Supervised Learning

In literature discussed in Section 2.2.2.1 to 2.2.2.3 and 2.2.3, there are limitations of ML work related to DQP in supervised learning under DG: a) ML techniques have not been applied in DQP using supervised learning extensively. In particular, a ML model has not

been proposed for DQP in supervised learning; b) ML techniques have not been applied in the forecast of DQ during DG processes by taking temporal sequences and correlations of DN into consideration. In particular, a ML model considering temporal sequences and correlations of DN has not been proposed for DQP; and c) ML techniques have not been applied to DQP during DG processes for meeting DG regulatory requirements. These are common issues from research results, as summarized below.

a. For DQP in supervised learning, [31] predicted DQ in a ML model, [34] focused on outliner detection and data discretization without confirming the model effectiveness from theoretical and experimental aspects, and [36] implemented an international requirement, BCBS 239, without running any experiments;
b. DN are time-series dependent [167] and co-relate with another or others. Temporal sequences of DN need to be considered. However, DQP are yet to be extended to sequential learning in deep neural networks ("DNN") [171, 172]. These did not apply ML models to learn sequential dependencies of DN for DQP. Existing methods focus on matched and mis-matched data [173] and missing data [58]; and
c. [32] utilized ML tools to improve DQ and [87] classified speech signals by labelling noisy data. Both did not consider any international or local standards [3, 120].

In this thesis, we intend to tackle the problem with a DQP model using supervised learning under DG to meet the regulatory requirement of DG. This model considers sequential learning of DN by taking temporal sequences and correlations of DN into account.

Model 2 - Data Quality Prediction in Unsupervised Learning

In literature discussed in Section 2.2.2.1, 2.2.2.2, 2.2.2.4 and 2.2.3, there are limitations of ML work related to DQP in unsupervised learning under DG: a) ML techniques have not been applied in DQP using unsupervised learning. In particular, a ML model has not been proposed for DQP in unsupervised learning; b) temporal sequences and correlations of DN have not been considered in DQ measurement. In particular, a ML model has not been proposed for DQ measurement under DG taking temporal sequences and correlations of DN into account; c) ML techniques have not been applied in the forecast of DQ during DG

processes by taking temporal sequences, correlations and importance of DN into account. In particular, a ML model considering temporal sequences, correlations and importance of DN has not been proposed for DQP; and d) ML techniques have not been applied to DQ learning such as DQP during DG processes for meeting DG regulatory requirements. These are generic from research results, as summarized below.

a. Unsupervised learning is yet to be applied in DQP. [102] leveraged Gaussian Mixture Model ("GMM") to exploit a connection between the statistical estimation and clustering problems and [84] utilized Bayesian GMM ("BGMM") to learn new topics in a set of conversations. Both were not applied to the estimation of DN weights. Others estimated the density of paper currency [85] and the sensitivity of data [190];
b. DN are time-series dependent [167]. Temporal sequences of DN need to be considered. However, current scientific measurement methods fail to capture the overlap of data including DN and cannot weigh DN or their relations: [168] fitted a mixture model to mainly capture the interdependence of numerical data attributes and [110] reiterated that data clustering could not identify feature dependencies for mixed data;
c. DQP are yet to be extended to sequential learning [171, 172] and made with an ATTN [171, 178] by learning different data weights based on data importance though sequential data demands for a temporal attention to learn its order dependences [30];
d. [32] utilized ML to improve DQ and [87] classified speech signals by labelling noisy sequence data. These did not consider any international or local standards [3, 120].

In this thesis, we intend to tackle the problem with a DQP model using unsupervised learning under DG to meet the regulatory requirement of DG. This model considers the importance of DN in DQP on top of temporal sequences and correlations of DN. Also, this model takes temporal sequences and correlations of DN into account in DQ measurement.

Model 3 - Data Quality Prediction Analytics

In literature discussed in Section 2.2.2, there are limitations of ML work related to DQP analytics under DG: a) ML techniques have not been applied in DQP analytics under DG

to meet DG regulatory requirements. A ML model for analyzing DQP during DG processes has not been proposed; and b) ML techniques have not been applied to DQ learning such as DQP analytics during DG processes for meeting DG regulatory requirements. These are common issues from research results, as summarized below.

The most relevant research on ML work related to DQP analytics for complying with DG regulatory requirements of the financial services industry are that: [88] used two networks, Multi-Layer Perceptron and Bayesian Networks, to measure and predict liquidity risk whereas [63] used ML to predict bank credits. [195] analysed flood risks with an AHP method and [196] classified credits with two neural networks instead.

In this thesis, we intend to tackle the problem with a DQP analytical model under DG to meet the regulatory requirement of DG.

Model 4 – Network Efficiency Improvement in Data Quality Prediction

In literature discussed in Section 2.2.2 and 2.2.3, there are limitations of ML work related to the network efficiency improvement in DQP under DG: a) ML techniques have not been applied in the network efficiency improvement in DQP. DQP network run-time saving has not been measured and a ML model for the network efficiency improvement in DQP during DG processes has not been proposed; and b) ML techniques have not been applied to DQ learning such as DQP during DG processes for meeting DG regulatory requirements. These are generic issues from research results, as summarized below.

a. The most relevant research related to network efficiency improvement in DQP are that: [197] used neural networks to predict consumption including the training speed while [198] assessed simulated trades for prediction in a reasonable time. These did not focus on DQP. There have been many DQP works but they are yet to be extended to learning the network efficiency improvement in DQP: [199] measured the quality on a large dataset and [200] enhanced the network predictive power with ML. These are yet to measure the network efficiency in terms of the network run-time saving; and

b. [32] utilized ML to improve DQ and [87] classified speech signals by labelling noisy sequence data. These did not consider any international or local standards [3, 120].

In this thesis, we intend to tackle the problem with a DQP network efficiency improvement model under DG to meet the regulatory requirement of DG by measuring network run-time saving.

Model 5 - Information Security Compliance Prediction

In literature discussed in Section 2.2.2 and 2.2.4, there are limitations of ML work related to IS compliance prediction under DG: a) ML techniques have not been applied in the prediction of ISL and IS compliance levels under DG. ISL during DG processes have not been predicted. A ML model considering sequences, correlations and importance of IS factors collectively for IS compliance prediction has not been proposed; and b) ML techniques have not been applied to IS learning during DG processes for meeting DG regulatory requirements. These are common from research results, as summarized below.

a. IS factors are inter-dependent [250]. Their sequences need to be considered. Certain works have been performed for prediction of compliance levels: [125] proposed ML for the privacy policy and [252] automated the evaluation of data privacy. These were not extended to ISL or IS compliance prediction under DG. Instead, there have been numerous research applying ML models to predict financial data: stock prices were predicted in Long Short-Term Memory ("LSTM") DNN with an Attention Mechanism ("ATTN") [207] and liquidity risks were forecasted with DNN [126]. Also, ML models making prediction by considering sequences, correlations and importance of data have been proposed in ample research. These models were applied with sequential learning and ATTN [94, 95, 96, 203]. However, sequential learning and ATTN have not been extended to ISL or IS compliance prediction; and
b. The most relevant IS learning research are that: [252] automated the evaluation of data privacy for the compliance purpose, [253] suggested an automated mean to process personal data relating to the General Data Protection Regulations with ML, and [125] proposed a ML approach to make forecasts for the privacy policy from a risk-based

perspective. These did not consider IS or DG related regulatory requirements.

In this thesis, we tackle the problem with an ISL prediction model for IS compliance prediction under DG to meet the regulatory requirement of DG. This model takes temporal sequences, correlations and importance of IS factors collectively into prediction.

Above models are proposed to address limitations of ML work related to DG in Section 2.2. These can be launched as DGP under the DGF to improve the level of DM for FIs.

1.3 Deep Neural Networks for Machine Learning Models

ML models are implemented with DNN in which LSTM Recurrent Neural Networks ("RNNs") are trained to learn regulatory compliance data. Three types of LSTM RNNs are trained: feedforward ("FF"), backward ("BK") and bi-directional ("BD"). These are applied with an ATTN to meet regulatory requirements of the financial services industry.

In the industry, data is usually retained over years, which may more than three decades [63]. A massive network is required. The data structure is unbalanced for aggregate bank and individual bank data [63]. Hence, to analyze data is a challenge for FIs [156].

For data co-relations, DEs that are most strongly correlated in the aggregate bank dataset are not the same DEs that are mostly correlated in the individual bank dataset [63]. For instance, the correlation between deposits and net interest-bearing liabilities is negative for the individual bank data. In contrast, this is almost zero for the aggregate bank data. Instead, the correlation between loans and net interest-bearing liabilities is positive. In real world, the learning of data is sophisticated [246]. DN are time-series dependent. The probability of an issue occurrence for next time would be high after issues have been identified last time. In contrast, the probability would be low if issues were remediated last time. Similarly, the probability of re-occurrence would be high in case new issues emerged.

In the industry, an aggregation of risk data requires the processing of enormous amounts of data [153, 155]. Historical and future scenarios ought to be considered in prediction over

years [3]. The prediction should be justified by the history of a sequence since reports are submitted to regulators [247]. DQP is one of the reports to be run with ML models [243].

In the consideration of these requirements in the industry, LSTM RNNs are suitable: a) RNNs can build an immerse network [11, 210] to process huge amounts of data. They cyclically pass states in networks to accept a wider range of time series related data [211]. Also, they can encode sequential correlations between instances before decoding sequences for prediction; b) LSTM RNNs [7, 8] can model long term temporal dependencies automatically. BD RNNs have forward network preserving past information [212]. This can predict data over time. In contrast, BK networks can preserve future information by propagating output error backwards. They aid in solving complex problems. With the merge, BD networks can exploit information from the past and the future for prediction [11] by introducing a 2nd layer to make connections flow in an opposite temporal order [8]; and c) ATTN can be applied to LSTM RNNs to compute a response at a position by attending to all positions [94] saving training time, to improve model dependencies [95], and to consider co-occurrence dependencies of attributes for time series prediction [96].

1.4 Research Methodology

In this thesis, a design science methodology is used to design proposed artifacts. Design science is a problem-solving paradigm [248] and has its basis in engineering. It is an approach to create guidelines, new ideas, and a set of practices that do this efficiently. The Design Science Research ("DSR") approach uses design as a research technique and primarily uses structure design artifacts to address the research problem.

The final product of DSR is an artifact that relates directly to the problem, which needs to be a "verifiable contribution" to the problem area [248]. Finally, the contribution needs to be described in a way that both the technical and audiences understand.

The adaptation of the DSR methodology ("DSRM") process for this thesis consists of the following steps:

a. Identify the problem and motivation: This addresses the lack of research work to predict DQ and IS compliance levels by learning regulatory compliance data;
b. Define aims of a solution: The aim is to develop ML models to predict DQ and IS compliance levels, define a model approach to measure DQ and compliance levels, and make a program to compile new datasets (risk data and IS). This allows FIs to make prediction in experiments with a synthesized dataset or a realistic banking dataset, resulting in the enhancement of DG including the improvement of DQ [4], management of data risks [120], management of bank-wide risks [4] and enhancement of IS [75];
c. Design and implement: The artifacts developed in this thesis are made in the research proposal. These include ML models, model approach and data synthetization program;
d. Demonstration: Experiments are made for the implementation of DNN with different learning methodologies or algorithms;
e. Evaluation: Cross-validation techniques are used with validation data to evaluate DNN in terms of the prediction accuracy and loss; and
f. Communication: This thesis and DNN are available to demonstrate the concepts presented.

Above steps are implemented in the DSRM, as visualized in **Fig. 1.1**.

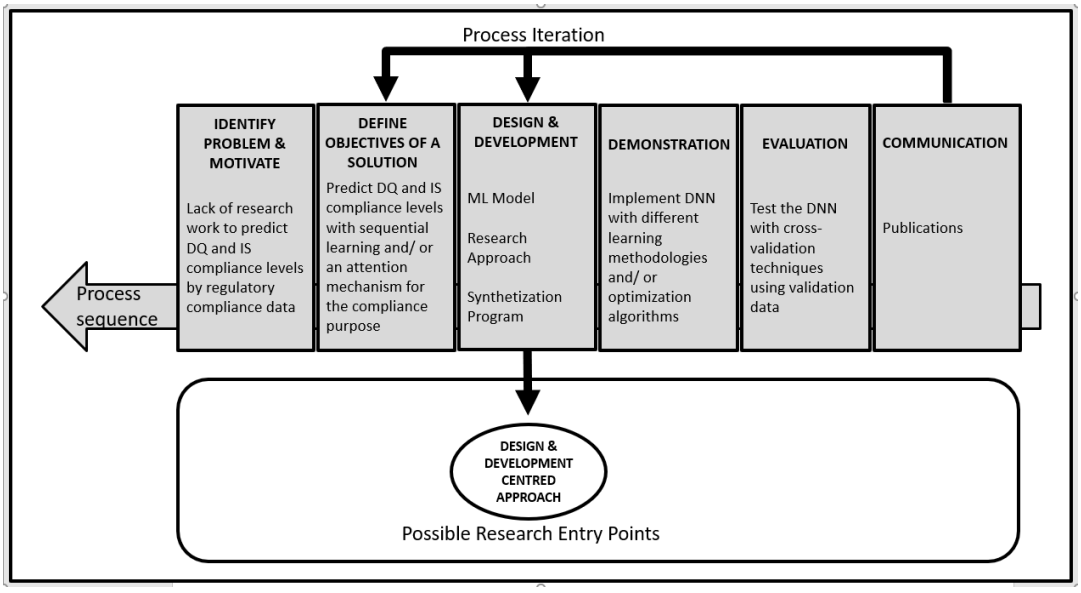


Fig. 1.1 Implementation of the DSRM

1.5 Thesis Outline

The rest of this thesis is organized as follows:

Chapter 2 Literature Review: This chapter lists all research topics that have been studied including ML techniques, ML models, model networks, ML work related to DG, summary on the limitations of research work, and our research proposal together with the research approach.

Chapter 3 DQP in Supervised Learning: This chapter demonstrates how to predict DQ under DG in supervised learning with a ML model. In the model, we label data based on an international requirement, BCBS 239. The model is implemented with networks, LSTM RNNs, for DQP including FF LSTM RNN, BK LSTM RNN and BD LSTM RNN. Then, we direct networks to learn temporal sequences and correlations of DN. The model is tested with a synthesized dataset and validated with a realistic banking dataset in experiments.

Chapter 4 DQP in Unsupervised Learning: In this chapter, we present how to predict DQ under DG in unsupervised learning with a ML model. At first, we detect DN from a dataset pursuant to an international requirement, BCBS 239. In our ML model, detected DN impacts are estimated in two generative mixture methods as weights in unsupervised learning. The weights are input into networks for DQP. Networks implemented are LSTM RNNs applying sequential learning along with an ATTN to pay attention to important DN. In experiments, network performance is examined at integrated and individual levels.

Chapter 5 DQP Analytics: In this chapter, we show how to analyze DQP under DG with a ML model. In the model, DQP are analyzed in accordance with a local requirement, CPG 235. The model is implemented by LSTM RNNs applying complex learning methods such as windows, timesteps and memory between batches. In experiments, the model predicts DQ and at the same time analyzes DQ by divergent dimensions.

Chapter 6 Network Efficiency Improvement in DQP: This chapter proposes how to improve DQP network efficiency under DG with a ML model. Before predicting DQ based on a local requirement, CPG 235, we propose a data profiling approach to slide divergent portions of data from a dataset for learning in networks. After this, we present a ML model to train LSTM RNNs applying memory between batches for measuring network run-time saving. In experiments, networks are tested with various algorithms and evaluated by validation data. Experimental results show the network efficiency improvement in DQP.

Chapter 7 IS Compliance Prediction: This chapter proposes how to predict IS compliance levels under DG with a ML model. At first, we develop a compliance approach. Then, we define IS rules according to a local requirement, CPG 234, for detecting ISCs in a ML model. Detection results are aggregated in a scoring function for ranking IS compliance levels. The levels are input into networks, LSTM RNNs, with sequential learning and an ATTN for prediction. These networks generate analytical reports showing the levels under the Life Cycle. Experimental results of our model are compared with that of other DNN.

Chapter 8 Conclusion: This chapter summarizes the outcomes of this thesis and contributions. We also present possible future research directions, and finish the chapter by discussing untouched, but interesting topics in this research area.

Chapter 2

Literature Review

There are many ML techniques for making predictions that have been developed to date. This chapter reviews ML work related to DG, discusses limitations of research work, and presents the research proposal, research approach, and research methodology. First, in Section 2.1, we provide an overview of ML types, their model components, and their various networks. Heterogenous network learning methods are explored. In Section 2.2, we survey ML work related to DG. Survey results are analyzed in four dimensions. Next, Section 2.3 summarizes limitations of current research work. In Section 2.4, we introduce our research proposal, research approach and research methodology. Finally, the literature review outcomes in this chapter are summarized in Section 2.5.

2.1 Machine Learning

The literature review starts with an introduction to ML techniques, model components, and model networks.

2.1.1 Techniques

Unsupervised Learning

Unsupervised learning generally amounts to discovering a number of patterns, subsets, or segments (clusters) within the data, without any prior knowledge of the target classes or concepts [145]. This saves time in labeling [161]. In the absence of labeling data, networks will learn the inherent structure of training data. Some researchers use this for clustering whereas other researchers highlight the use of representation learning and density estimation.

In unsupervised learning, an algorithm learns on its own using the provided data. The algorithm is frequently used to generate labeled instances automatically [144]. These instances are used in supervised learning. The learning can be taken for an exploratory analysis and the dimensionality reduction. It makes use of mostly unlabeled data for training. An example of these is an unsupervised clustering technique to segment customers for an analysis of the relationship learning between individual features with less features or latent features that are interrelated with initial features. Algorithms that can be used are k-means clustering, principal component analysis and autoencoders [146].

Supervised Learning

A supervised learning technique deals with how to input data into networks and how to label data. Some researchers make inputs in the form of a vector or matrix and reduce the number of inputs by MapReduce method [45]. This provides the flexibility that input data is transformed before they are fed to the model. Other researchers highlight the use of structured labeling for the ML model [46]. Rules can be defined for data by labelling. This avoids any unexpected or messy prediction or analysis which always occurs in unsupervised learning. In supervised learning, an algorithm learns from training data which is labeled. The result of learning is a model. The model can be used to predict a response when it is presented with input data which has not been seen before by the model [144].

For supervised learning algorithms, values for the model parameters are chosen and determined with experimentation and hyperparameter tuning. There is an algorithm which analyzes training data and produces an inferred function that can be used for mapping new examples. For instance, this provides a mapping from attributes to specified classes or concept groupings. Classes are identified and prelabelled in data prior to learning [145]. In case an optimal scenario is designed, the network algorithm can be used to correctly determine class labels for unseen instances. Common algorithms are Logistic Regression (“LReg”), Naïve Bayes (“NB”), K-Nearest Neighbors (“KNN”), Decision Tree (“DT”), Support Vector Machines (“SVM”), Artificial Neural Networks (“ANN”) and Random Forests (“RF”).

Reinforcement Learning

This is a kind of learning from mistakes. To implement this, networks will make a lot of inaccurate predictions in the beginning. Reinforcement learning is commonly used in deconstructing a task into a hierarchy of subtasks [148], learning with higher-level temporally abstract or actions [149] and efficiently abstracting over the state space through function approximation.

Whenever the network algorithm learns good behaviors with positive signal and bad behaviors with a negative one, networks will reinforce the algorithm to prefer good behaviors over bad ones. Over time, networks are training with the algorithm to make less mistakes. In the process of learning, networks require an agent and an environment [147], both need to be connected through a feedback loop. With the loop, a set of actions taken in networks can influence the environment. By continuously providing a signal of an updated state and reward to the agent, networks will be able to learn the reinforcement signal of behaviors. For example, an agent can learn to play a game by being told whether it wins or loses. However, it is never given the "correct" action at any given point in time [150].

In comparison of these learning methods, supervised learning and unsupervised learning can be used to resolve the research problem stated in the thesis. The supervised learning requires labeling of data inputs to train networks and the design of data labels are described in the following chapters. On the contrary, the unsupervised learning demands for an additional training of data inputs before importing trained outputs into networks for scientific computation with ML algorithms.

2.1.2 Model Components

During ML model development, network components which contain inputs, processing and outputs [7] are defined. Apart from the dataset, data feature selection and optimizer will be selected. In the network setup, some factors are to be considered such as input layer, trainable layer and output layer. In each layer, the activation function along with classifiers will be determined.

In the model implementation, some networks are trained. Before the training, the number of epochs for the network training is determined [22]. This number influences the network prediction performance such as accuracy and error. Apart from this, the network run-time can be recorded to see how efficient the network is [10].

After the implementation, the network performance is evaluated in terms of metrics such as prediction error which is called a loss [51]. The error is estimated by training data as well as testing data [52]. To confirm the model effectiveness, the network is evaluated by validation data.

2.1.3 Potential Networks for Machine Learning Models

A wide range of networks are available for ML model implementation. They are generally classified into three types including a) LReg, DT, SVM, NB, RF [53]; b) KNN [55]; and c) ANN [54].

The first type networks have their merits. For example, LReg takes inputs with two possible values based on data input relationships by optimizing the conditional likelihood [213]. Decision Trees ("DTs") are trees classifying instances by filtering them based on feature values. Each node in a DT represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values [53]. SVM are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. It includes some components including regularized linear learning models (such as classification and regression), theoretical bounds, convex duality and the associated dual-kernel representation, and sparseness of the dual-kernel representation [159]. NB networks are very simple Bayesian networks composing of directed acyclic graphs with only one parent (representing the unobserved node) and several children (corresponding to observed nodes) with a strong assumption of independence among child nodes in the context of their parent [53, 158]. RF are ensemble learning methods for

classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the model of the classes or mean/ average prediction of the individual trees [160].

All these can be used for analytical and prediction purposes. But they could not be used for sequential learning by considering temporal sequences and correlations of data in the network training.

The second type is KNN. It can be used to construct a divergent approximation of the expected function for each instance [214] while the nearest neighbors of an instance are Euclidean geometry.

The third type is ANN. There has been a key research comparing functionalities of this network against above-mentioned networks. The comparative results showed that ANN was the most common method for prediction and analysis [56]. This network type has been well developed in prior years [57]. It covers FF networks and RNNs.

From the network architecture and algorithm viewpoint, feedforwarding networks are classified into deviating types encompassing auto-encoder, probabilistic, time delay and convolutional whereas RNNs contain simple, complex, Elman, Jordan, Bidirectional and LSTM networks [57].

In comparison, RNNs can be used to model sequential data by loops to capture temporal evolution of data [11]. They can train data that occur in current state, previous state and future state, dissimilar to the first and second types of networks.

In an in-depth examination of RNNs, network structures are different. They can be exploited to encode sequences for application as the encoder-decoder framework [215]. Apart from this, sequence-to-sequence models with an encoder-decoder framework have been successful in experiments [216, 217]. These models can be trained from both theoretical and experimental perspectives.

In the comparison of network functionalities, RNNs can establish a temporal relation of input data and at the same time build an immense network (with 1,000 input size) with the highest memory (100) when compared with other neural networks [56]. RNNs can help to process huge amount of regulatory compliance data across years or big data.

Out of RNNs, there are multiple networks that can be considered. Further research on RNNs reveal that networks can be designed in a FF run or a BK run depending on learning requirements. The FF run network can be combined with the BK run network to form a hybrid network for learning prediction or analyzing patterns differently. The hybrid one is a BD network introducing a 2nd layer to make hidden-to-hidden connections flow in an opposite temporal order [10].

There have been myriad models comparing BD networks against FF and BK networks in experiments. Their experimental results vary. For example, [222] used BD LSTM sentence representations to model a parser state with only 3 sentence positions and compared the network against BK and FF LSTM networks. Furthermore, [251] proposed a FF LSTM language model, a BK LSTM language model and a BD LSTM based gap completion model to investigate the estimation of sentence probability while [219] presented BD LSTM networks to compare with FF and BK LSTM networks for the classification of framewise phoneme. Additionally, [225] developed an approximate inference algorithm for 1-Best (and M-Best) decoding in BD neural sequence models to reason about both FF and BK time dependencies. Its experiment included FF, BK and BD network results.

We survey how networks learn differently, as elaborated below.

Recurrent Neural Network ("RNN") computes the sequence of hidden state vector ($h = h_1, \dots, h_T$) to generate vector sequence ($y = y_1, \dots, y_T$) for a given input vector sequence ($x = x_1, \dots, x_T$), iterating the equations from t = 1 to T [26]:

$$h_t = \mathcal{H}(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (2.1)$$

$$y_t = \left(W_{hy}h_t + b_y\right) \quad (2.2)$$

where W is weight matrices, W_{xh} is weight matrix between input and hidden vectors, b is bias vector, b_h is bias vector for hidden state vector, and $\mathcal{H}$ is an activation function (of Sigmoid) for hidden nodes.

LSTM RNN model temporal sequences and dependencies by replacing traditional nodes with memory cells such that it has an internal and outer recurrence. The cell is influenced by gates (forget, input modulation, internal state and hidden state gates) other than the input and output. The activations of units for LSTM neurons in layers at time t are computed below:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (2.3)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (2.4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (2.5)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (2.6)$$

$$h_t = o_t \odot \tanh(c_t) \quad (2.7)$$

where input gate is i_t with weight matrix of W_{xi}, W_{hi}, W_{ci}, forget gate is f_t with weight matrix is W_{xf}, W_{hf}, W_{cf}, the cell is c_t (generated by calculating the weighted sum using previous cell state and current information generated by the cell), σ is the Logistic Sigmoid function, b is bias, output gate is o_t with a weight matrix of W_{xo}, W_{ho}, W_{co} and output response is h_t. The input gate and forget gate govern the information flow into and out of the cell. The output gate controls how much information from the cell is passed to the output h_t. With current input x_i, the state h_{i-1} of a previous step is generated and the current state of the cell c_{i-1} decides whether to take inputs, forget memory stored, and output the state generated latter [8, 212].

In a BD network, two LSTMs based on timesteps of the input sequence are trained - on the existing input sequence and then on the reversed input sequence. This combines forward and backward outputs in the network.

BD RNN computes FF hidden sequence $\vec{h}$ and BK hidden sequence $\overleftarrow{h}$ respectively before combining $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ to generate outputs y_t:

$$\overrightarrow{h_t} = H(W_{x\vec{h}}\, x_t + W_{\vec{h}\vec{h}}\overrightarrow{h_{t-1}} + b_{\vec{h}}) \tag{2.8}$$

$$\overleftarrow{h_t} = H(W_{x\overleftarrow{h}}\, x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h_{t+1}} + b_{\overleftarrow{h}}) \tag{2.9}$$

$$y_t = W_{\vec{h}y}\overrightarrow{h_t} + W_{\overleftarrow{h}y}\overleftarrow{h_t} + b_y \tag{2.10}$$

In a FF pass [21], LSTM runs forwards from time to time when the input is feed. In the process, activations are updated and network stores all hidden layer and output activations at each time step. For each memory block, activations are updated below:

$$\text{Input gates: } x_i = \sum_{j\in N} W_{ij} y_j(\mathcal{T}-1) + \sum_{c\in C} W_{ic} s_c(\mathcal{T}-1) \tag{2.11}$$

$$y_i = f(x_i) \tag{2.12}$$

$$\text{Forget gates: } x_\emptyset = \sum_{j\in N} W_{\emptyset j} y_j(\mathcal{T}-1) + \sum_{c\in C} W_{\emptyset c} s_c(\mathcal{T}-1) \tag{2.13}$$

$$y_\emptyset = f(x_\emptyset) \tag{2.14}$$

$$\text{Cells: } \forall_c \in C, x_c = \sum_{j\in N} W_{cj} y_j(\mathcal{T}-1) \tag{2.15}$$

$$s_c = y_\emptyset s_c f(\mathcal{T}-1) + y_i g(x_c) \tag{2.16}$$

$$\text{Output gates: } x_w = \sum_{j\in N} W_{wj} y_j(\mathcal{T}-1) + \sum_{c\in C} W_{wc} s_c(\mathcal{T}) \tag{2.17}$$

$$y_w = f(x_w) \tag{2.18}$$

$$\text{Cell outputs: } \forall_c \in C, y_c = y_w h(s_c) \tag{2.19}$$

In a BK pass, the LSTM propagates output errors backwards through unfolded net after resetting all partial derivatives to 0. For each LSTM block, δ's is calculated as:

$$\text{Cell outputs: } \forall_c \in C, \in_c = \sum_{j\in N} W_{jc}\delta_j(\mathcal{T}+1) \tag{2.20}$$

$$\text{Output gates: } \delta_w = f'(x_w) \sum_{c\in C} \in_c h(s_c) \tag{2.21}$$

$$\text{States: } \frac{\partial E}{\partial s_c}(\mathcal{T}) = \in_c y_w h'(s_c) + \sum_{c\in C} \frac{\partial E}{\partial s_c}(\mathcal{T}+1) y_\emptyset(\mathcal{T}+1) + \delta_{\mathrm{i}}(\mathcal{T}+1) w_{\mathrm{ic}} + \delta_{\mathrm{i}}(\mathcal{T}+1) w_{\emptyset \mathrm{c}} + \delta_{\mathrm{w}} w_{\mathrm{wc}} \quad (2.22)$$

$$\text{Cells: } \forall_c \in C, \delta_c = y_i\, g'(x_c) \frac{\partial E}{\partial s_c} \quad (2.23)$$

$$\text{Forget gates: } \emptyset = f'(x_\emptyset) \sum_{c\in C} \frac{\partial E}{\partial s_c} s_c(\mathcal{T}-1) \quad (2.24)$$

$$\text{Input gates: } \delta_i = f'(x_i) \sum_{c\in C} \frac{\partial E}{\partial s_c} g(x_c) \quad (2.25)$$

Network Memory

LSTM can be applied to RNNs as training networks have difficulty in capturing long term dependency because of vanishing gradients [219]. It directs networks to model long term temporal dependencies as sequences automatically as it exploits information from the past and future to build a giant network. As such, networks remember memory across long sequences [8] to obtain a control over when an internal state is cleared for a forecast [7, 171] or for a data sequence modelling [104]. Sequences are learnt based on the history of sequences of input data in the context of time series. This replaces traditional nodes with memory cells leading to an internal and outer recurrence [9]. As a consequence, LSTM RNNs are able to learn time sequence features for prediction or analysis [25]. They store long term data in additional cells and to use gates to control information flow [220].

Further literature analysis has revealed that there have been myriads of similar models comparing BD networks against FF and BK networks in experiments and their experimental results vary. For example, [221] used BD LSTM sentence representations to model a parser state with only 3 sentence positions and compared the network against BK and FF LSTM networks. Also, [222] proposed a FF LSTM language model, a BK LSTM language model and a BD LSTM based gap completion model to investigate the estimation of sentence probability while [223] presented BD LSTM networks to compare with FF and BK LSTM networks for the classification of framewise phoneme. Additionally, [219] developed an approximate inference algorithm for 1-Best (and M-Best) decoding in BD neural sequence models to reason about both FF and BK time dependencies. Its experiment included FF, BK and BD network results.

Network Wrapper

TimeDistributed Wrapper is an example of a Keras wrapper. The wrapper layer applies the same dense (neural network layer), which is a fully-connected, operation to every timestep of a three-dimensional data input to allow to gather the output at each timestep. This leads to an effective sequence learning. The sequence is split into input-output pairs and for the sequence to be predicted one step at a time.

The output layer is a Keras dense layer which is a regular densely connected network layer with the activation function to predict a probability distribution for the next value in the sequence [106]. The 3-dimensional data input is TimeDistributed as interpreted below:

$$(X, T, F) \xrightarrow{TimeDistributed(LSTM)} m_{1,\ldots,}m_n$$

The input shape is a triple (samples, timesteps and features) represented by (X, T, F) in LSTM networks in which the TimeDistributed wrapper enables the network training by applying the component of memory network, M, to each time-step [122].

Network Activation

For activation dynamics [33, 52], data inputs to each memory block are multiplied with input gates for that block, and the final output of a block is the activation of the cell status multiplied by the output gate, as computed below:

$$x^{\lambda} = w^{\lambda n}\, y_{t-1}^{n}, \qquad for\ \lambda \epsilon \{l, c, w, o\} \tag{2.26}$$

$$y^{\lambda} = f^{\lambda}(x_{t-1}^{\lambda}), \qquad for\ \lambda \epsilon \{l, c, w, o\} \tag{2.27}$$

$$s = s_{t-1} + y^{l} \circ y^{c}, \tag{2.28}$$

$$y^{\theta} = y^{w} \circ f^{s}(s) \tag{2.29}$$

There are divergent layers: inputs layer (c), input gates (l), memory cells (s), output layer (o), output gates (w), cell outputs (θ) and layer connecting with nodes between timesteps (n). At layer λ, $x^{\lambda}$is input and f^{λ} is an activation function while y^{λ} is output, t is time, $\circ$

is entry wise product and f(x) is a vector of function values when applying f to each element of the vector x. The equation for x^{λ} and y^{λ} are similar in networks but activation dynamics for memory cell statuses and outputs are divergent. Thereupon, network inputs, activations and partial derivatives are evaluated at time t such as $y \equiv y_t$.

Network Sequence Prediction

In the research on sequential learning in DNN, numerous works are related to prediction. They centred on many domains: some researchers investigated the application of sequential learning in speech recognition [88], video captioning [89], reading comprehension [90], ads recommendation [91] and natural language processing [92]. None of them was learnt for DQP.

Sequential learning has been applied in various research projects to explore dependencies of an image content problem [93]. For example, [94] presented a self-attention network to compute response at a position in a sequence by attending to all positions. The most recent one [96] predicted attributes of images by taking the co-occurrence dependencies among attributes into account. Thus, sequential learning can be applied to a model to explore DQ correlations.

Network Optimization – ATTN

To optimize sequential learning, we study an ATTN to see whether it improves the network performance for prediction. ATTN is one of the predominant mechanisms that has been applied to DNN recently. It assigns different weights to various data to enable networks to focus on important data [97]. The application was on a few domains: An earlier study [89] presented a temporal ATTN for video caption generation. Another study [182] introduced a deep attention selective network for image classification. Others used it to recognize 3D action [99]. Some of them leveraged it for machine translation [92, 100] and document classification [101]. [102] introduced temporal attention on different time steps for electronic health records. In last few years, ATTN has been introduced to use encoder to reference records dynamically in the decoder [100]. This is yet to be applied to DQP or

prediction of IS compliance levels to pay attention to important DN or IS factors respectively.

Network Algorithms

Back Propagation Through Time is used as an algorithm in RNNs to update cell weights [225]. It calculates data sequentially. When outputs are estimated, an error is back propagated to obtain error responsibilities. The error of next time step is back-propagated with the error of this time step. The sum of errors exploits the information of recent input sequence and put more importance on the latest input. Thereupon, prediction are sequentially dependent [226].

To optimize networks, network optimization algorithms such as ADAM and ADAGRAD can be used.

ADAM is derived from adaptive moment estimation and computes dual adaptive learning rates for multiple parameters from the estimates of the first and second moments of the gradients [29, 227]. As the algorithm updates exponential moving averages of gradient and squared gradient when-ever hyper-parameters control the exponential decay rates of moving averages, it can train networks efficiently. The ADAM algorithm is computed in the following [227]:

$$x_t = x \cdot \frac{\sqrt{1-\beta_2^t}}{(1-\beta_1^t)} \tag{2.30}$$

$$x_t = \theta_t \leftarrow \theta_{t-1} - x_t \cdot m_t / \left(\sqrt{v_t} + \widehat{\in}\right) \tag{2.31}$$

where β is a delay rate, t is a time step, m_t is the moving average of gradient, v_t is a squared gradient and θ is a parameter, Assuming $f(\theta)$ is an objective function, the stochastic scalar function is differentiable with regards to the parameter. To minimize the expected value of this, E[$f(\theta)$], the realization of stochastic function at timesteps 1,…, t and the gradient (vector of partial derivatives of ft) at timestep t can be defined. Then, the algorithm updates exponential moving averages of the gradient and the squared gradient

whenever hyper-parameters β1, β2 ∈ [0,1] control exponential decay rates of these moving averages. ADAM has an update rule: a choice of step sizes. Assuming ∈, an effective step taken in a parameter space at timestep is Δ_t equalling to the equation of $x \cdot \widehat{m}_t / \sqrt{\widehat{v}_t}$.

The effective size has two upper bounds, and the computation is shown below:

$$|\Delta_t| \leq x \cdot (1-\beta_1)/\sqrt{1-\beta_2} \text{ if } (1-\beta_1) > \sqrt{1-\beta_2} \text{ and } |\Delta_t| \leq x \quad (2.32)$$

$$\left|\widehat{m}_t / \sqrt{\widehat{v}_t}\right| < 1 \text{ if } (1-\beta_1) = \sqrt{1-\beta_2} \quad (2.33)$$

In most scenarios, the $\left|\frac{\mathbb{E}}{[g]} / \sqrt{\mathbb{E}[g^2]}\right| \leq 1$ gives the following:

$$\left|\widehat{m}_t / \sqrt{\widehat{v}_t}\right| \approx \pm 1 \quad (2.34)$$

where g is gradient. The effective magnitude of the steps taken in the parameter space at each time step are bounded by the step size setting x. That is $|\Delta_t| < \& \approx x$.

ADAGRAD is another algorithm which is a variant of stochastic gradient descent [227].

The computation of this algorithm is:

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t+1,i} + \epsilon}} \bullet g_{t,i} \bullet \quad (2.35)$$

where η is learning rate at each time step t for every parameter θ_i based on the past gradients computed for θ_i and $g_{t,i}$ is the partial derivative of the objective function. $G_t \in \mathbb{R}^{dxd}$ is a diagonal matrix where each element i is a sum of squares of gradients up to the timestep whereas ϵ is a smoothing term avoiding division by 0 (on order of $1e$ -8).

ADAGRAD uses a different learning rate for every parameter θ_i at each time step [29] and converges in batch [234]. It sets the value as 0.01, and adapts the rate to parameters, performing low rates for parameters with frequently occurring features and high rates for

those with infrequent features. Then, the implementation is vectorized by performing a matrix vector product $\odot$ between G_t and g_t (denoting gradient at t):

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} \odot g_t. \tag{2.36}$$

As we can see, these algorithms have their merits for the optimization of network performance.

Network Performance Metrics

A number of metrics can be used to measure the performance of networks.

Firstly, the accuracy of networks can be evaluated as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.37}$$

where $TP, TN, FP\ and\ FN$ denote true positive, true negative, false positive and false negative respectively [105].

The loss can be defined as an averaged cross entropy to maximize the likelihood of correct prediction. The computation is:

$$L = \frac{1}{M}\sum_{i=1}^{M}(-y_i\, log(p_i) - (1-y_i)\log(1-p_i)) \tag{2.38}$$

where M represents the number of training samples in a dataset and y is the target output. The ith sample is labelled as $y_i \in \{0,1\}$ and p_i is the predicted probability based on the input sequence [48].

The precision, recall and F1-Support can be calculated based on the following equations:

$$\text{Precision (P)} = \frac{TP}{TP+FP} \tag{2.39}$$

$$\text{Recall (R)} = \frac{TP}{TP+FN} \tag{2.40}$$

$$\text{F1-Support} = \frac{2PR}{P+R} \tag{2.41}$$

where $TP, FP, FN\ and\ TN$ represent true positive, false positive, false negative and true negative respectively [249].

2.2 Machine Learning Work Related to Data Governance

From Section 1.1.2, there are opportunities for using ML in DG for the financial services industry. In fact, statistical researchers and financial regulators recently suggested that the quality of big data could be statistically computed by ML techniques [243].

We analyse literature for ML work related to DG. Relevant results are summarized in Section 1.2.3 of Chapter 1. The results show that following considerations need to be addressed:

2.2.1 Data Governance and Regulatory Data

In this financial services industry, regulatory compliance data has been governed for a long time under various DG processes to identify, assess and prevent risks [151, 152, 153]. DG refers to policies and procedures geared towards the management of usability, availability, integrity and security of data [151, 152]. In this industry, FIs need to meet three requirements defining quantitative DQ in terms of DQ metrics. These requirements are issued by international and local regulators in the industry.

The first requirement is that from an international perspective, FIs are expected to meet DQ principles in the process of aggregating risk data into risk reports. These principles include the principle number 3 (accuracy and integrity), the principle number 4

(completeness) and the principle number 5 (timeliness) under BCBS 239 [3]. FIs are recommended to provide forward-looking capabilities of DQ for the improvement of DQ.

In meeting these principles, DQ issues, DN need to be minimized including an omission of translation, a negligence in the data format transformation, and retention of data which is redundant, duplicated, stale, unreasonable, invalid, mis-matched, incomplete or null. These issues are common [3].

In assessing the compliance level with this requirement, many FIs have not been able to meet certain principles till 2018 [4]. Their compliance levels are summarized in **Table 2.1**.

Table 2.1 Compliance Status

	Governance & infrastructure		Risk data aggregation capabilities				Risk reporting practices				
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
2017	2.90	2.73	2.60	2.90	2.87	2.9	2.73	3.03	3.03	2.97	3.33
2016	2.83	2.60	2.60	2.93	2.73	2.9	2.77	3.00	3.10	2.97	3.37
Diff.	0.07	0.13	0.00	-0.03	0.13	0.0	-0.03	0.03	-0.07	0.00	-0.03

This table showed that some principles should be improved. For instance, the principle number P4 (completeness) and P7 (accuracy) could not be met by FIs despite their best efforts for the compliance. To accelerate the compliance, the financial regulator took action by performing tests over FIs for the compliance level checking in 2018 [4].

The second requirement is that from a local standpoint, FIs are required to manage data risks through the assessment and management of DQ by six dimensions including the dimension of (a) accuracy; (b) completeness; (c) consistency; (d) timeliness; (e) availability; and (f) fitness for use under APRA CPG 235 [120]. When evaluating the compliance level with these dimensions, a few FIs have not met this requirement until 2019. In Australia, the government demanded for the restoration of trust in the financial system after several cases of misconducts [76]. This is attributable to recent scandals. For instance, the Prudential Regulator refused a bank's corporate risk data due to data inaccuracy and incompleteness in April 2018 [77]. This month, the Royal Commission challenged FIs for financial misconducts arising from the poor quality of risk data [78].

The third requirement is that from another local point of view, FIs are expected to manage IS risks under APRA CPG 234 [75] by including thirteen IS controls, called ISCs, during the information asset life-cycle ("Life Cycle") under IS policy frameworks. In terms of the compliance level of CPG 234, FIs are yet to enhance IS [83]. In January 2020, there was a data breach resulting from a cyber-attack for P&N Bank. In August 2019, PayIDs of customers have been stolen across Australian big four banks [79]. In March 2019, the Bank of Queensland experienced a personal data breach by a third-party provider.

From the above compliance assessments, it is onerous for FIs to meet these three requirements. Many FIs found it challenging due to multiple problems: insufficient controls over risk data [5], lack of a DG framework [24], inadequate quality data [6] and lack of a robust IT infrastructure [12]. In making preparation for the compliance, 40% of domestic systemically important banks worldwide have performed an independent validation of their capabilities to meet the first requirement, BCBS 239 [4].

Regulatory data can cover a) risk data as set out in BCBS 239 consisting of MR, CR, OR and LR; b) business and operations data as defined in CPG 235 including data used for analytics and intelligence purposes; and c) IS data as specified in CPG 234.

2.2.2 Data Quality Measurement and Prediction

Applying ML models to DQ measurement and prediction is critical. Tera scale ML is an important ingredient in the quality modeling and monitoring of big data [38]. This covers large scale applications of ML and their computational models. Applying ML by FIs benefits the risk management significantly by addressing specific problems such as data risk [141]. These are mentioned in Section 1.1.2 of Chapter 1.

We now examine existing data science research on ML work related to DQ measurement and prediction during DG processes. We identify the following studies:

2.2.2.1 Data Quality Characteristics

In real life, data quality may not be high. Data may be inaccurate or has inherent noise [162]. High quality data may not be available. This, however, is a pre-condition to add value by utilizing the data. The analysis of data by developing dashboards aids in the management decision making [71]. When the decision is applied to risk data, enterprises understand the risky areas of their business. They can start to develop a plan for mitigating risks [72, 73]. An ideal case is when risks are forecasted. Then, potential risks can be minimized. This risk management practice is common in the financial services industry [74]. Unless the quality of data is verified and is well managed, the value of data cannot be assured [151, 153].

DN in business applications have different impacts and they are time-series dependent [167]. On the one hand, DN may co-relate with another or others. Their relationships are many-to-many. This makes the weighing DN impact complicated. On the other hand, DN are time-series dependent. For instance, current DN usually reoccur in next period, but remediated DN rarely occur in future. Similarly, emerged DN contaminate subsequent data recurrently unless they are rectified. Temporal sequences, correlations and importance of DN have not been considered. These are summarized in Section 1.2.3 of Chapter 1.

2.2.2.2 Data Quality Measurement and Prediction with Machine Learning Models

In DQ measurement work, we study existing ML models.

a. Prior research commonly measured DQ issues based on three dimensions: accuracy, completeness and timeliness [238, 239]. Some used metrics [132]. Other leveraged missing data [58]. All are yet to consider DG regulatory requirements of the financial services industry; and
b. Current scientific methods fail to capture the overlap of data including DN. [168] fitted a mixture model to mainly capture the interdependence of numerical data attributes. [110] reiterated that data clustering could not identify feature dependencies for mixed data. [170] found that errors of model-based clustering are neglected due to the difficult

computation. It is onerous to use these methods for weighing DN and their relations or correlations.

These results show that ML techniques have not been applied in the measurement of DQ during DG processes. In summary, temporal sequences and correlations of DN have not been considered in DQ measurement. In particular, a ML model has not been proposed for DQ measurement under DG taking temporal sequences and correlations of DN into account. These are mentioned in Section 1.2.3 of Chapter 1.

In DQ prediction work, we study existing ML models.

a. Existing DQP methods focus on matched and mis-matched data [173] as well as missing data [58];
b. In recent years, [33] identified DQ by rules in an energy industry to predict data that is to be corrected with statistical relational learning;
c. Another work [167] forecasted DQ on Apache Spark with ML which allowed users to define verification codes before the declarative API was leveraged for forecasts;
d. Although DQP are gaining significant amount of interest in recent literatures, they have not been extended to sequential learning in DNN [171, 172]. Data sequences and correlations were not considered in prediction;
e. DQ is yet to be predicted with an ATTN [171, 178] to pay attention to important data. However, sequential data demands for a temporal attention to learn order dependences of data inputs [30]. ATTN can aid in learning different data weights based on their importance; and
f. Instead, ATTN has been experimented in a few domains: a research [103] presented a temporal ATTN for video caption generation and another research [89] introduced a deep attention selective network for the image classification. In recent years, research leveraged ATTN for 3D action recognition [182], machine translation [100] and [184], document classification [92] and electronic health recording [101]. All these have not been applied with a ML model to learn temporal sequences, correlations and importance of DN collectively in DQP during DG processes.

These results show that ML techniques have not been applied in the forecast of DQ during DG processes by taking temporal sequences, correlations and importance of DN into consideration. In particular, a ML model considering temporal sequences, correlations and importance of DN has not been proposed for DQP under DG. These are mentioned in Section 1.2.3 of Chapter 1.

2.2.2.3 Data Quality Prediction Using Supervised Learning

We analyse the application of ML techniques in DQP during DG processes using supervised learning.

a. [31] made prediction in a ML model for an estimation of DQ. This proposed a solution, DQ-Long Short-Term Memory. It is yet to take any regulatory requirements into account;
b. Another work is [32] utilizing ML tools to improve DQ. This identified DQ and defined a DQ rating algorithm and grading system, DataIQ, to rank data before making prediction. DQ scores were analyzed in terms of simple statistics such as median, mean, std, max and min, as opposed to international and local standards (considering integrity and accuracy, completeness and timeliness, availability, and fitness for use) [3, 120];
c. One work is a research [37] classifying DQ issues. This used a semi-auto classification method to identify simple DQ issues such as matched or mis-matched data;
d. One more work is a research [34] focusing on outliner detection, data discretization and feature construction. It did not consider regulatory requirements or run experiments to confirm the model effectiveness from theoretical and experimental aspects;
e. An additional work is a classification of speech signals in RNNs by labelling noisy and unsegmented sequence data [87]; and
f. Last work is a research [36] recognizing the importance of an international requirement, BCBS 239, for the Industry. This described the implementation of BCBS 239 and interpreted two cases (insider trading and FIBO) for the application of the requirement. However, it was not tested in experiments to confirm its practicability.

These results show that ML techniques have not been applied in DQP during DG processes using supervised learning extensively. In summary, DQ have not been predicted under DG in supervised learning. In particular, a ML model has not been proposed for DQP during DG processes in supervised learning. These are mentioned in Section 1.2.3 of Chapter 1.

2.2.2.4 Data Quality Prediction Using Unsupervised Learning

We analyse the application of ML techniques in DQP during DG processes using unsupervised learning.

a. One study [102] has leveraged GMM to exploit a connection between the statistical estimation and clustering problems in computational geometry and another one [84] as a mixture of Gaussian Distribution with Dirichlet Process ("DP") has utilized BGMM to learn new topics in a set of conversations. Both are yet applied to the estimation of DN weights. Weighing DN has not been proposed in prior research studies, as discussed in Section 2.2.2.2;
b. Other methods estimated the density of paper currency [85] and the sensitivity of data from an Austrian bank [190]; and
c. A recent study [191] showed that a GMM was used to build a semi-supervised model in multi-mode processes for forecasting the quality of big data.

These results show that ML techniques have not been applied in DQP during DG processes using unsupervised learning. In summary, DQ have not been predicted under DG in unsupervised learning. In particular, a ML model has not been proposed for DQP during DG processes in unsupervised learning. These are mentioned in Section 1.2.3 of Chapter 1.

2.2.3 Data Governance Requirement Compliance

Applying ML models to DG requirement compliance is indispensable. ML is essential in the regulatory reporting of the financial services industry by improving the compliance

processes through the organization of structured and unstructured data [142]. In fact, ML can create accurate methods for data analysis, modeling and prediction by identifying complicated and non-linear patterns in huge data sets [142]. These are mentioned in Section 1.1.2 of Chapter 1.

We examine existing data science research on ML work to meet DG regulatory requirements of the financial services industry. There are two types of compliance including DQ and IS.

Firstly, we survey the application of ML models in DQ learning for complying with DG regulatory requirements of the financial services industry. The most relevant research include the following:

a. A research [88] leveraged Multi-Layer Perceptron and Bayesian Networks to measure and predict LR respectively. Experiments output relatively low error rates (8.0e-3 for GA and 1.7e-10 for LMA) and low RMSE (less than 0.2). Financial data has been forecasted for analysis rather than the quality of data;
b. Another one [63] used ML to predict bank credits with twenty-three features achieving a prediction accuracy of 80%. Consistently, this prediction was on financial data;
c. One more [195] analysed flood risks with AHP method by defining the importance of risks and dividing hazards into 5 risks. This analysis was on flood risks instead of the quality of data; and
d. Last more [196] classified credit with two networks such as logistic regression and support vector machine. The network accuracy was 75% but reduced to 43.5% for critical regions. In the same way, this focused on financial data which is unrelated to DQ.

These results show that ML techniques have not been applied to DQ learning such as DQP or DQP analytics during DG processes for meeting DQ regulatory requirements. In particular, DQP analytics during DG processes have not been proposed with a ML model. These are mentioned in Section 1.2.3 of Chapter 1.

Secondly, we survey the application of ML models in IS learning for complying with DG regulatory requirements of the financial services industry. The most relevant research include the following:

a. A research automated the evaluation of data privacy for the compliance purpose and generated a report on the evaluation result [252];
b. Another research suggested an automated mean to process personal data relating to the General Data Protection Regulations [253]. It utilized Semantic Analytical Stack on top of the Apache Spark to provide a stack of functional layers from RDF/ OWL data representation to ML algorithms; and
c. One recent research proposed a ML based approach for the privacy policy from a risk-based perspective [125]. It used Github to test the approach and leveraged a market available tool (applying ML algorithms including Naïve Bayes, Support Vector Machine, Decision Tree and Random Forest) to make forecasts. All these did not consider IS or DG related regulatory requirements.

These results reveal that ML techniques have not been proposed to IS learning or IS prediction for meeting DG regulatory requirements. Prior work targeted on other regulatory requirements with ML models. In summary, IS compliance levels have not been predicted during DG processes with a ML model. These are summarized in Section 1.2.3 of Chapter 1.

Above results show that the application of ML models in DQ learning and IS learning for complying with DG regulatory requirements of the financial services industry is limited.

2.2.4 Network Efficiency Improvement in Data Quality Prediction

Applying ML models to network efficiency improvement in DQP is essential. Making efficient performance prediction for large-scale advanced analytics minimizes the amount of training data required [143], as mentioned in Section 1.1.2 of Chapter 1.

We now discuss existing data science research on ML work relating to the network efficiency of DQP during DG processes. The most relevant research studies are limited to the following:

a. A research study [197] used metrics for neural networks aiming at the consumption prediction including the training speed and network accuracy. Its networks were Support Vector Regression (SVR), local SVR and H2O deep learning. Its prediction was yet to be extended to DQ or DQP;
b. Another research [198] back-tested a strategy to assess simulated trades and checked the accuracy and loss of network prediction by cross-validations. It demonstrated the success of processing a cluster of big data for prediction within a reasonable time of few hours. Equally, its focus was not on DQ; and
c. Some research predicted DQ but did not learn the network efficiency improvement in DQP. One work [199] measured the quality on a large dataset with networks in terms of the accuracy, completeness and consistency. It was yet to be used in the measurement of the network efficiency in terms of network run-time. Another work [200] studied how ML enhanced the network performance in terms of the predictive power and classification accuracy. Both are yet to measure the network efficiency in terms of the network run-time saving.

These results show that ML techniques have not been proposed to address the network efficiency improvement of DQP. In summary, DQP network efficiency under DG has not been improved with ML models. In particular, DQP network run-time saving has not been measured and a ML model for DQP network efficiency improvement during DG processes has not been proposed. These are mentioned in Section 1.2.3 of Chapter 1.

2.2.5 Information Security Compliance Prediction

Applying ML models to IS compliance prediction is crucial. ML is essential in the regulatory reporting of the financial services industry by improving the compliance processes [142], as mentioned in Section 1.1.2 of Chapter 1.

We now examine existing data science research on ML work relating to the prediction of ISL and IS compliance levels during DG processes. We identify the following studies:

2.2.5.1 Information Security Characteristics

IS factors in computer systems and networks have different impacts and they are inter-dependent [250]. IS risks co-relate with another or others. Their sequences, correlations and importance need to be considered.

2.2.5.2 Information Security Compliance Prediction

In IS prediction work, we study existing ML models.

a. A research proposed a ML based approach for the privacy policy from a risk-based perspective [125]. It used Github to test the approach and leveraged a market available tool (applying ML algorithms) to make forecasts. Another research automated the evaluation of data privacy for the compliance purpose [252]. They were not extended to ISL or IS compliance prediction. Although there were some ML work for regulatory requirement compliance, ML techniques have not been applied in ISL prediction or relevant compliance prediction. Prior ML research centred on the prediction of other requirement levels;
b. For ML models predicting data of the financial services industry, there have been abundant work. In last two years, stock prices have been predicted in LSTM DNN with an ATTN to learn temporal sequences and pay attention to more important stock prices [207] and LR have been forecasted with DNN [126]. These predictions have not been extended to the prediction of ISL or IS compliance levels. Earlier ML research focus on the prediction of financial data; and
c. ML models making prediction by taking sequences, correlations and importance of data into account have been proposed in ample research. Over a decade, a significant amount of studies used and directed sequential learning with an ATTN in a ML model for prediction, such as forecast of dependencies for an image content problem [203], computation of responses at a position by attending to all positions in a self-attention

network [94], improvement in model dependencies with DNN applying a self-attention [95], and prediction of image attributes by considering attribute co-occurrence dependencies [96]. However, sequential learning and ATTN have not been applied in ISL or IS compliance prediction by considering sequences, correlations and importance of IS factors collectively.

These results show that ML techniques have not been applied in the prediction of ISL and IS compliance levels during DG processes. In summary, ISL during DG processes have not been predicted. In particular, a ML model considering sequences, correlations and importance of IS factors collectively for IS compliance prediction has not been proposed. These are mentioned in Section 1.2.3 of Chapter 1.

2.3 Summary on the Limitations of Research Work

From above data science results, ML techniques have not been largely deployed to DG including the following:

a. From Section 2.2.2, key limitations are: a) temporal sequences and correlations of DN have not been considered in DQ measurement. In particular, a ML model has not been proposed for DQ measurement under DG taking temporal sequences and correlations of DN into account; b) ML techniques have not been applied in the forecast of DQ during DG processes by taking temporal sequences, correlations and importance of DN into consideration. In particular, a ML model considering temporal sequences, correlations and importance of DN has not been proposed for DQP under DG; c) ML techniques have not been applied in DQP during DG processes using supervised learning extensively. In particular, a ML model has not been proposed for DQP during DG processes in supervised learning; d) ML techniques have not been applied in DQP during DG processes using unsupervised learning. In particular, a ML model has not been proposed for DQP during DG processes in unsupervised learning; and e) ML techniques have not been applied to DQ learning such as DQP during DG processes for meeting DG regulatory requirements;
b. From Section 2.2.3, ML techniques have not been applied to DQ and IS learning during

DG processes DG to meet regulatory requirements of DG;

c. From Section 2.2.4, a) ML techniques have not been applied in network efficiency improvement for DQP. DQP network run-time saving has not been measured and a ML model for the network efficiency improvement in DQP during DG processes has not been proposed; and b) ML techniques have not been applied to DQ learning such as DQP during DG processes for meeting DG regulatory requirements; and
d. From Section 2.2.5, a) ML techniques have not been applied in the prediction of ISL and IS compliance levels under DG. ISL during DG processes have not been predicted. A ML model considering sequences, correlations and importance of IS factors collectively for IS compliance prediction has not been proposed; and b) ML techniques have not been applied to IS learning during DG processes for meeting DG regulatory requirements.

All these are limitations of ML works related to DG although ML techniques were highly recommended to support DG. These are mentioned in Section 1.2.3 of Chapter 1.

2.4 Research Proposal

To address above limitations, we propose five ML models in the research proposal.

2.4.1 Proposed Models

Our proposed models are listed below, as mentioned in Section 1.2.3 of Chapter 1.

a. A DQP model using supervised learning under DG to meet the regulatory requirement of DG. This model considers sequential learning of DN by taking temporal sequences and correlations of DN into account;
b. A DQP model using unsupervised learning under DG to meet the regulatory requirement of DG. This model considers the importance of DN on top of the temporal sequences and correlations of DN collectively in DQP. Additionally, this model takes temporal sequences and correlations of DN into account in DQ measurement**;**
c. A DQP analytical model under DG to meet the regulatory requirement of DG;
d. A DQP network efficiency improvement model under DG to meet the regulatory

requirement of DG by measuring network run-time saving; and

e. An ISL prediction model considering sequences, correlations and importance of IS factors collectively for IS compliance prediction under DG to meet the regulatory requirement of DG.

We design these five models, Model 1 to Model 5, by taking DG regulatory requirements into account, as visualized in **Fig. 2.1**. These models can be framed as DGP under DGF to improve the level of DM for the governance purpose. **Fig. 2.1** illustrates procedures for the design of five ML Models, mentioned in Chapter 3 to 7 correspondingly. This spans steps from broad assumptions to detailed methods of data collection, analysis, and interpretation of the model. These models are illustrated below:

a. For the research limitation in Chapter 3 – DQP in supervised learning, a ML model in alignment with the international requirement, BCBS 239, implemented by FF, BK and BD LSTM RNNs applying sequential learning is proposed. The model is experimented with two input data including the synthesized risk dataset and realistic banking dataset;
b. For the research limitation of Chapter 4 – DQP in unsupervised learning, a ML model in alignment with the international requirement, BCBS 239, implemented by LSTM RNNs applying sequential learning and an ATTN is proposed. The model is experimented with the synthesized risk dataset.
c. For the research limitation of Chapter 5 – DQP analytics, a ML model in alignment with the local requirement, CPG 235, implemented by LSTM RNNs applying sequential learning and complex learning methods is proposed. The model is experimented with the synthesized risk dataset;
d. For the research limitation in Chapter 6 – Network efficiency improvement in DQP, a ML model in alignment with the local requirement, CPG 235, implemented by LSTM RNNs applying sequential learning and memory between batches is proposed. The model is experimented with the synthesized risk dataset; and
e. For the research limitation in Chapter 7 – IS compliance prediction, a ML model in alignment with the local requirement, CPG 234, implemented by LSTM RNNs applying sequential learning and an ATTN is proposed. The model is experimented with the IS dataset.

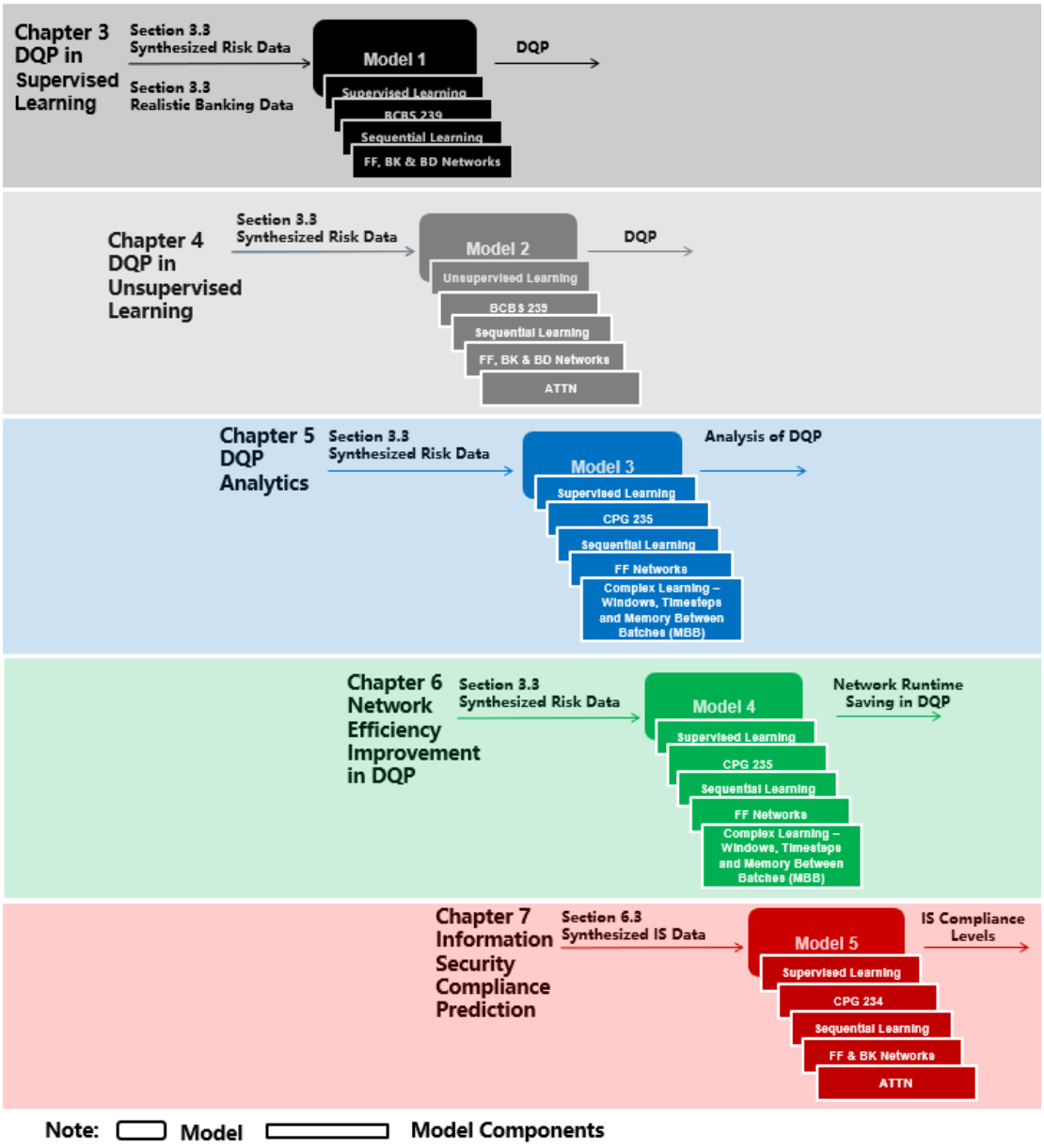


Fig. 2.1 Model 1 to Model 5

This model approach shows inputs and outputs of each model and the connections between the five models. This stands out how models are developed for making contributions of each chapter, as described in Chapter 8.

2.4.2 Research Approach

For this thesis research approach, we cover three artifacts:

a. Data synthetization program is given in **Fig. 2.2**. This is used to create new datasets (risk data and IS). DEs are labelled for the measurement of DQ and compliance levels;
b. Model approach is mentioned in **Fig. 2.1**. This is adopted to solve problems with the calculation of DQ and IS compliance scores after DQ and ISL are measured; and
c. ML models are depicted in Section 2.4.1. They train DNN to predict DQ and IS compliance levels for regulatory compliance data based on the scores calculated.

The first artifact, data synthetization program, in the approach is shown in **Fig. 2.2**.

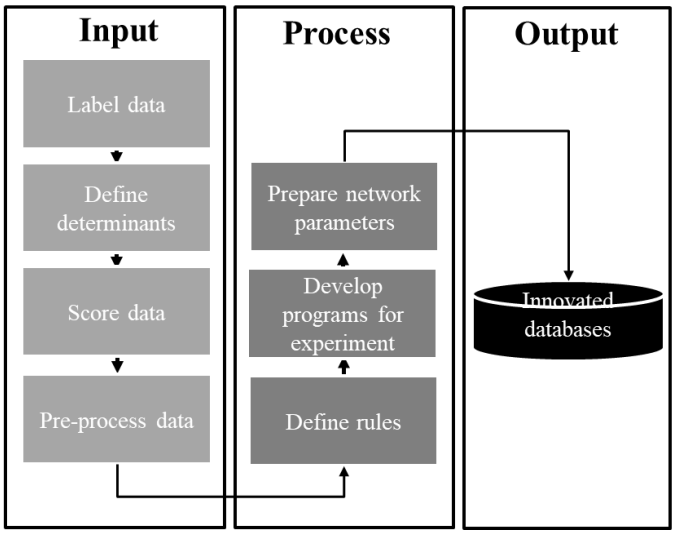


Fig. 2.2 Data Synthetization Program

Fig. 2.2 depicts steps for the synthetization of new datasets. In the data synthetization program, we label data, define determinants and score data before pre-processing data. In the progress of synthetization, we define rules and develop programs for synthetization experiments. When networks are developed, relevant parameters are pre-set for generating new datasets.

The second artifact, model approach, is visualized in **Fig. 2.1** and the third artifact in the research approach is ML Models, as described in Section 2.4.1.

2.5 Summary

In this chapter, we provide an overview of ML techniques including ML methods, their model components along with their various networks. Then, we survey current ML work related to DG from four perspectives including DQ measurement and prediction, DG Requirement Compliance, network efficiency improvement in DQP and IS compliance prediction. In a nutshell, there have been numerous ML models developed. These could be applied to DG including learning of DQ or IS compliance levels. However, the extent of application to DG was not extensive.

Following this, we summarize limitations of current research work. These limitations motive us to propose our research work. In view of this, we introduce our research proposal and research approach. In the research proposal, we propose five ML models in Section 2.4.1 addressing limitations mentioned in Section 2.3. Under the research approach, we cover three artifacts containing data synthetization program, model approach and ML models.

Chapter 3 proposes a ML model to predict DQ in supervised learning. Chapters 4 and 5 present a ML model to predict DQ in unsupervised learning and analyze DQP correspondingly. Chapters 6 and 7 propose a ML model improving the network efficiency in DQP and predicting IS compliance levels respectively. Finally, Chapter 8 concludes the thesis, and discusses contributions as well as possible future works.

Chapter 3

Data Quality Prediction in Supervised Learning

In Chapter 3, we propose a ML model that can be used to predict DQ under DG in supervised learning. This is part of DGP under the DG framework. This chapter proposes how to label data in accordance with the international requirement, BCBS 239. Herein, we utilize LSTM RNNs with sequential learning to learn temporal sequences and correlations between DN for DQP.

In this chapter, Section 3.1 is an introduction of the proposed model. In Section 3.2, we first explain data labeling based on BCBS 239. The labelled data is then used for sequence prediction in divergent LSTM RNNs including FF, BK and BD networks. The network learning is explained with a system architecture. We then discuss the utilized data and the conducted experiments in Section 3.3. Finally, Section 3.4 concludes and summarizes this chapter.

3.1 Introduction

From research results summarized in Section 1.2.3 and elaborated in Section 2.2, there are limitations of ML work related to DQP in supervised learning under DG: a) ML techniques have not been applied in DQP during DG processes using supervised learning extensively. In particular, a ML model has not been proposed for DQP during DG processes in supervised learning; b) ML techniques have not been applied in the forecast of DQ during DG processes by taking temporal sequences and correlations of DN into consideration. In particular, a ML model considering temporal sequences and correlations of DN has not been proposed for DQP under DG; and c) ML techniques have not been applied to DQ learning such as DQP during DG processes for meeting DG regulatory requirements.

To address these, we leverage a ML model in supervised learning to predict DQ according to an international requirement, BCBS 239 [3] in this chapter. This improves DQ [4, 6] during DG processes.

For DQP, we develop a ML model to train DNN for learning regulatory compliance data. With the prediction, FIs understand what DQ are going to be.

In total, three networks are implemented including FF LSTM RNNs, BK LSTM RNNs and BD LSTM RNNs. In these networks, DN form a sequence over timeline to predict DQ by learning correlations between sequential data.

3.2 Proposed Model

DQ are predicted in a ML model with DNN. In our approach, we label regulatory compliance data in supervised learning for the measurement of DQ in terms of DQ scores.

3.2.1 Data Labelling in Supervised Learning

In the model, we label data for scoring before DQP. Data labelling is depicted in **Fig. 3.1**.

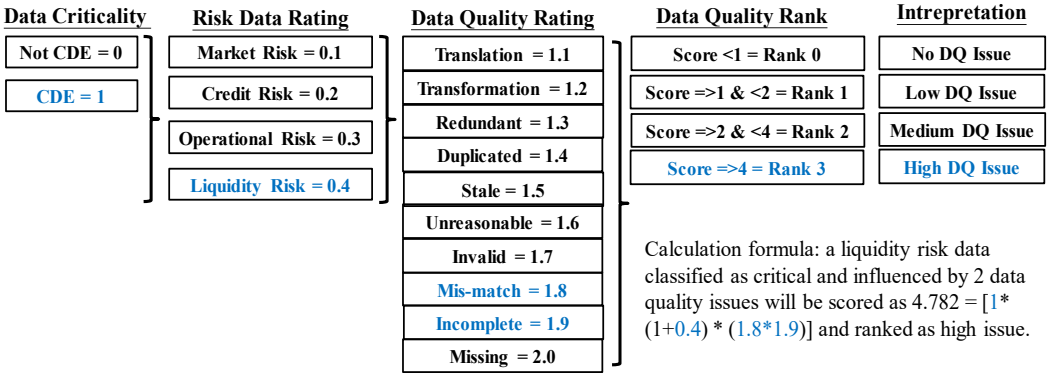


Fig. 3.1 Data Labelling

We: a) label data as 1 or 0 to indicate whether they are critical; b) assign a risk data rating (0.1 to 0.4) and a DQ rating (1.1 to 2) based on types of risk data and quality issues respectively; c) classify DQ scores (<1, =>1 & <2, =>2 & <4 and =>4) into four (4) ranks

(no, low, medium, or high DQ issue or DN) as actual outputs to be compared with the prediction made in experiments. These ratings are justified:

a. Data Criticality Factors: Data is classified based on business criticality and sensitivity [120]. Referring to a research defining factors impacting DQ [98], we make a similar assumption: $DE_{Criticality} = 0$ if a DE is not used in the data aggregation and $DE_{Criticality} = 1$ if a DE is used in the aggregation influencing DQ.

b. Risk Data Rating: different risk types are inherited with different levels of risk. MR is approximated at 10% ($E(R_i)$) under a CAPM [228], CR is assumed to be 20% (CVaR) under a confidence level of 99.5% to 99.99% for the finance sector using Monte-Carlo simulation [224], OR is set to 30% due to VaR between 27.84% and 37.71% [208, 209] and LR is defined as 40% (R^2) under a liquidity measure of Depth [194, 204]. Rating for these risks are used together for the risk data aggregation under BCBS 239 [3];

c. DQ Rating: min. or max. operation (from 0 to 1) is applied to aggregate multiple quality issues [193]. We define DQ ratings (1.1 to 2.0) by normalizing 10 issues. These are intuitive since there is no empirical research available; and

d. DQ Scores: We classify scores after referencing to a research ranking quality to allow the management to understand which ones are crucial to DQ [98].

DQ scores are a multiplication of the ratings for each DN. The overall score is computed as: Data Criticality Factors*(1 + Risk Data Rating)*DQ Rating. The ratings and ranks are usually assigned by risk experts in real life.

Data is pre-processed before they are imported into networks. They are input into networks to find unusual records which are different from the standard (e.g. null values) such as BCBS 239. To pre-process data, data is scored and then normalized to a binary value of either 1 or 0 by a min-max scaler.

3.2.2 Feedforward, Backward, and Bi-Directional Networks

Three networks including FF, BK and BD LSTM RNNs are implemented in the model. Relevant network methodologies are described in **Table 3.1**.

Table 3.1 Network Methodologies

Networks	Methodologies
FF LSTM RNNs	We build a 3-dimensional LSTM based on factors of samples, timesteps and features for sequence classification prediction. At an input layer, we take 1 million samples (as sequences), leverage timesteps and input the number of features. Then, we determine the number of memory units for the hidden layer. At an output layer, we generate a value for each timestep using an activation function to predict DQ. In this layer, the network forms a time-distributed wrapper layer based on input sequences to forecast outputs. The network weights are found by ADAM algorithm [10] and the accuracy of outputs is computed. Afterwards, the network generates new input sequences to predict DQ. Those exceeding thresholds are classified. At the end, outputs turn out to be 0 or 1. This layer is the input for BD LSTM RNNs.
BK LSTM RNNs	This network wraps the LSTM hidden layer with the backward layer to construct two sets of hidden layers – one fits in the input sequence and another one fits with a reversed input sequence. Then, the time-distributed wrapper layer around the output layer will receive dual input sequences for merging before forecasting DQ [21]. This wrapper is unlike the block processing and look-ahead convolution layer utilized in a bidirectional network [218].
BD LSTM RNNs	In training this network, we revise input sequences and import them into the LSTM as the backward input sequences before merging them. After fitting the combined one into the model, we measure the performance between them in terms of a log loss [23] over epochs.
Other Methods for Analysis	Apart from the concatenation method (con), we incorporate three methods into our model to analyze the outcome of our hybrid network including multiply (mul), average-out (ave) and summation (sum) giving rise to divergent outcomes in our case studies.

These networks forecast DQ according to the past and future data. They include sequence prediction with ADAM algorithm for optimizing the learning performance. ADAM is selected due to its merits mentioned in Section 2.1 of Chapter 2.

In LSTM RNNs, a Back Propagation Through Time algorithm is calculated sequentially. When outputs are estimated, RNNs will back propagate an error to obtain error responsibilities. The error of the next time step is backpropagated with the error of this time step. The sum of errors exploits the information of the recent input sequence and puts more importance on the latest input. Accordingly, DQP are sequentially dependent. This is estimated as the P(Y|X) which is computed by the following:

$$\prod_{t\in[1,n_y]} \frac{\exp\left(f\left(h_{t-1},\ e_{yt}\right)\right)}{\sum_{y'}\exp\left(f\left(h_{t-1},\ e_{y'}\right)\right)} \tag{3.1}$$

This equation is from $\prod_{t\in[1,n_y]} P(y_t|\, x_1, x_2, \ldots, x_t, y_1, y_2, \ldots, y_{t-1})$ where f(h_{t-1}, e_{yt}) is an activation function between e_{h-1} and e_{yt}, and h_{t-1} is an output at a previous time t−1.

In this chapter, we apply regularization to networks. We try to dropout (10% of the activations) on the LSTM layer and make regularization to see whether the problem of overfitting can be alleviated. Dropout prevents the co-adaptation of hidden units by randomly omitting feature detectors from the network during the forward propagation. Constraining L2-norms of weight vectors (w) by rescaling w after a gradient descent step, we obtain a cost function [212] as computed below:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} t_i \log\left(y_i\right) + \lambda\|\theta\|_F^2 \tag{3.2}$$

This is a negative log-likelihood of true class label y where t is time, m is number of target classes, λ is L2 the regularization hyperparameter and y_i are predicted outputs.

3.2.3 Sequence Prediction

Sequence prediction in RNNs can be from one to many, many to one or many-to-many. Correspondingly, the data input can be mapped to a sequence with multiple steps as an output, a sequence of multiple steps as data inputs can be mapped to class the prediction or a sequence of multiple steps are data inputs which can be mapped to a sequence with multiple steps as outputs. Accordingly, RNNs can be used for text data, speech data, classification prediction, regression prediction and generative models.

In BD LSTM RNNs, sequences of received signals are fed in the FF direction into the LSTM cell resulting in an output $\overrightarrow{a_k}$, and fed in the BK into another LSTM cell resulting in an output $\overleftarrow{a_k}$ [28]. Assuming that $\overrightarrow{a}_k^{(l)}$ and $\overleftarrow{a}_k^{(l)}$ are outputs of the final BD layer, the layer with an activation function is used to gain $\hat{x}_k$ as:

$$\hat{x}_k = \emptyset\left(W_{\vec{a}}\,\overrightarrow{a}_k^{(l)} + W_{\overleftarrow{a}}\,\overleftarrow{a}_k^{(l)} + b_x\right) \tag{3.3}$$

This estimated $\hat{x}_k$ is given by the following:

$$\begin{bmatrix} \hat{p}_{model}\left(\hat{x}_k = s_1 \mid y_k, \overrightarrow{a}_{k-1}^{(l)}, \overrightarrow{c}_{k-1}^{(l)}, \overleftarrow{a}_{k-1}^{(l)}, \overleftarrow{c}_{k-1}^{(l)}\right) \approx & \hat{p}_{model}\left(\hat{x}_k = s_1 \mid y_{\mathcal{T}}, y_{\mathcal{T}-1}, \dots, y_1\right) \\ \hat{p}_{model}\left(\hat{x}_k = s_m \mid y_k, \overrightarrow{a}_{k-1}^{(l)}, \overrightarrow{c}_{k-1}^{(l)}, \overleftarrow{a}_{k-1}^{(l)}, \overleftarrow{c}_{k-1}^{(l)}\right) \approx & \hat{p}_{model}\left(\hat{x}_k = s_m \mid y_{\mathcal{T}}, y_{\mathcal{T}-1}, \dots, y_1\right) \end{bmatrix} \tag{3.4}$$

where $\hat{p}_{model}$ is the probability of estimation, y_k is input to the layer, k is a sequence of transmission, W is weight, b_x is bias parameter and $\mathcal{T}$ is a sequence of length. The output of layer is of the length m and the network considers data from previously received signals (encoded in $a_{k-1}^{(l)}$ and $c_{k-1}^{(l)}$) and from current signals.

3.2.4 System Architecture

For the network topology, the architecture is displayed in **Fig. 3.2**. The FF LSTM RNN is at the 1st layer. Turning to the BK one, one layer is added such that the network is BD.

Then they are concatenated to generate an output at another layer, "2in1" circles, inside **Fig. 3.2** constructing a hybrid network [192].

This model is implemented with a scoring approach for classifying DQ (exceeding threshold or not) in BD LSTM RNNs. This is not equivalent to the model [207] experimenting neural networks in unidirectional and bidirectional structures - built with deep bidirectional LSTM network-based sequences for classifying signals. Instead, it is similar to a model [25] learning time-sequence features in a BD LSTM RNNs. The variance is that we explore other methodologies to ascertain the effectiveness of BD LSTM RNNs.

Model output sequences are dynamic depending on input sequences and the number of predictions at each time step. Output values initially are set at 0. When the cumulative sum of input attributes in the sequence exceeds a threshold, values will turn to 1 while the threshold is a portion of the length of the input sequence [189]. The threshold is trained in networks based on a sequence length [0:1]. The output sequence of 1 reveals that a DQ exceeds the threshold which is the data risk threshold.

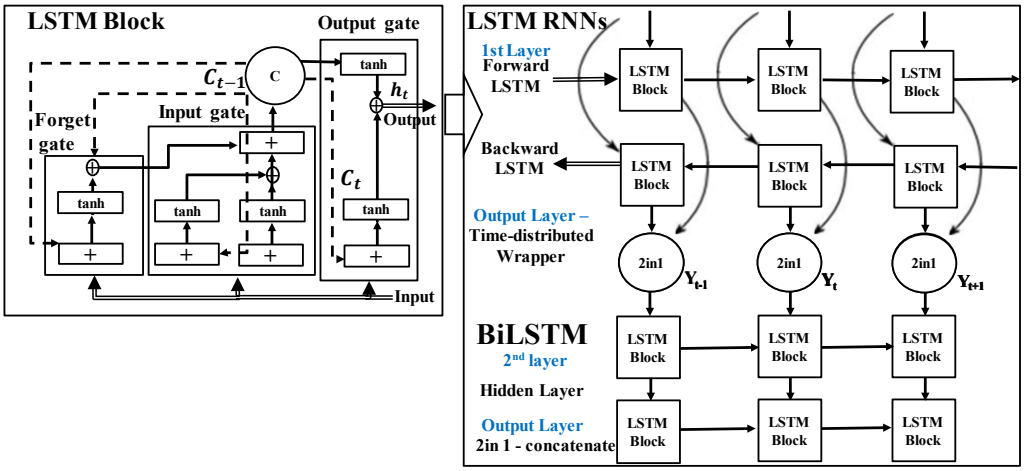


Fig. 3.2 System Architecture

3.3 Experiments

Python v3.5 with Keras and TensorFlow backend is used on a system with the processor

of i.7-7500U CPU@2.9GHz, OS of 64-bit and Win 10 Pro. The dataset is divided into two parts. 70% of data is fitted to networks and 30% is used for the network evaluation. This system setup and data split are used in the models of other chapters of this thesis.

The purpose of these experiments is to demonstrate the effectiveness of our model on DQP in supervised learning by conducting a set of experiments over the synthesized dataset as well as the realistic banking dataset.

Dataset

Under BCBS 239, the risk data quality is to be assessed including MR, CR, OR and LR. In real world, there was no dataset for the financial services industry containing these four types of risk data collectively. Instead, there have been a plethora of datasets for a single risk type from the internet or public domain.

In the industry, diverse risk data would be compared to find whether data is matched or not before the identification of DN. In the absence of different risks, data mis-matches cannot be identified to meet regulatory requirements as required by regulators of the financial services industry.

After searching publicly available datasets, we discover abundant banking datasets. Notwithstanding, none of them includes four types of risk data jointly. The market available datasets are limited to the following:

a. Direct marketing campaigns (phone calls) in a Portuguese bank institution [41];

b. Eighteen datasets from data.world [42] – The data is up-to-date ranging from 1999 to Sep 2018. These mainly cover assets, ATM locations, credit card complaints, federal reserve system, failed banks, UK economy data, interest rates, bank suspension data, real anonymized transactions, loan data, bank institution history, financial empowerment centres, bank statistics of the US, and check cashing locations as well as financial services. These datasets did not cover four types of risks collectively;

c. Over one thousand datasets (1,197 as of Apr 2019) from World Bank [43] – The data date is close to recent years. Correspondingly, none of them has a full set of risk data. The dataset focused on liquid reserves to bank assets ratio, capital to asset ratio, nonperforming loan to total gross loans, merchandise imports, external debt stocks and concessional, merchandise exports, private creditors, IBRD loans and IDA credits, risk premium on lending, social protection rating, lending, IBRD, multilateral, tariff, gross fixed capital, net taxes on product, bonds, debt buyback, inflation (consumer prices), deposit interest rate, lending interest rate and losses due to theft; and

d. Plenty banking dataset from biml [44] – the data update date is recent. The situation is the same. A full set of risk data cannot be found. The dataset solely covered loan risk data, loan status, marketing, BBVA cards, credit, currency exchange, card cubes, apply shares, IPOs, churn, tech share values and volume, crunchbase data, project assessment, financial intermediary funds, average value weighted returns and Europe debts.

In view of the above, networks in this chapter are trained with a synthesized dataset. This dataset is synthetized based on real-world risk DN. In risk publications, key DEs have been announced by risk experts regularly. They have been published in publications of MR [13, 14], CR [15], OR [16, 17] and LR [18]. After making reference to these, we: a) summarize characteristics of DEs; b) capture their commonalities; and c) synthesize a set of data simulating realistic data features. Inside the dataset, data is implanted with common quality issues from DQ publications [128, 129, 193]. The dataset is depicted in **Table 3.2**.

Table 3.2 Integrated Dataset

Total Number of DEs	DEs for Each Risk Data Type	Data Nature	DN
132	33 Market Risk	8 Static Data (seldom changed after being recorded) and 25 Dynamic Data (which may change continually) in Each Risk Database	10 Classes
	33 Credit Risk		
	33 Operational Risk		
	33 Liquidity Risk		

This dataset has 132 DEs belonging to four (4) risk databases. Each database contains 33 features in which 8 are static and 25 are dynamic.

The dataset covers one (1) million banking customer records. It includes corporate and individual data, and values are discrete instead of continuous. Some features are extracted to **Table 3.3**.

Table 3.3 Data Features (Examples)

MR	CR	OR	LR
Asset Maturity (1945 days, tbc, na)	Loan ID (385623, 0, tbc, na)	Loss Income Ratio (1.15%, 92.04%, 54.6%)	Liquidity Rate (10.39%, 65.29%)
NPV (425543, 0, tbc, na)	Weighted Avg PD (6.31%, 19.48%)	Residual Legal Liability ($1385, 12, 0)	Instrument (TBC, Forward, Equity)

Table 3.3. shows that data features are embedded with heterogenous quality issues such as MR's NPV (0, tbc and na), CR's loan ID (0, tbc or na), OR's legal liability (0) and LR's instrument (TBC).

To drill down into a database, MR data features are extracted to **Table 3.4** for reference.

Table 3.4 MR Data Features (Sample)

MR	Discount Rate	Cost Of Equity	Return On Equity	Risk Free Rate	Systematic Risk	Mkt Risk Premium	Equity Risk Premium
Mean	0.50	2.99	1.00	0.50	33.58	0.50	1.47
Min	0	0	0	0	0	0	0
Max	1	414	83	1	558	1	164

MR	AssetAmt	Nationality	CustID	AssetMaturity	CashFlow
Count	993350	993991	990227	993286	993268
Unique	625185	34	109769	3653	624452
Top	tbc	CAD	ABC61595	tbc	na
freq	6589	31292	26	6761	6563

In this MR database, the number of DN by DQ ranks are listed in **Table 3.5**.

Table 3.5 Number of DN by Ranks ("Rating" in Table)

MR_Segments	MR_Address_Rating		MR_Segments	MR_AssetAmt_Rating		MR_Segments	MR_CashFlow_Rating	
0	0	6075	0	0	6075	0	0	6075
	1	13629		1	13512		1	13504
	2	150		3	267		3	275
1	0	100747	1	0	100747	1	0	100747
	1	224211		1	221979		1	221876
	2	2222		3	4454		3	4557
2	0	100665	2	0	100665	2	0	100665
	1	223551		1	221344		1	221400
	2	2310		3	4517		3	4461
3	0	100301	3	0	100301	3	0	100301
	1	223840		1	221730		1	221742
	2	2299		3	4409		3	4397

Note: 0 – no DN, 1 – low DN, 2 – medium DN, 3 – high DN in the rating while business segments are classified into 0, 1, 2 and 3 representing no segment, the private bank ("PvB"), the wholesale bank ("WB") and the retail bank ("RB").

DN are aligned with DQ principles of the BCBS 239 including the principle # 3 - accuracy and integrity, principle # 4 - completeness and principle # 5 – timeliness, as defined in **Table 3.6**.

Table 3.6 DN Mapped to BCBS 239

	Principle # 3	Principle # 4	Principle # 5
DN	1.1 Translation: a bank balance in foreign currency not yet converted into a local currency 1.2 Transformation: the birthday format of a banking system not yet synchronized with other systems 1.3 Redundant: potential customers not yet on-board as a true bank customer (due to a failure in the bank's customer due diligence approval process) are retained 1.4 Duplicated: an extra customer ID for the same clients not yet verified 1.6 Unreasonable: undesirable clients with a very poor credit are kept 1.7 Invalid: customers over age 150 hold an account in a bank without investigation 1.8 Data mis-match: a master data cannot be reconciled to other banking systems	1.9 Incomplete: passport number deviates from the standard (e.g. required digits) 2.0 Missing: an amount cannot be shown in a statement for equity trading	1.5 Stale: obsolete records over the data retention period of a bank not yet purged

Results

Applying ADAM algorithm to networks for DQP, we note that the accuracy of prediction of the integrated dataset for the BD LSTM RNN (100%) is superior to that of the FF LSTM RNN (0) and the BK LSTM RNN (31%), as depicted in **Table 3.7**. The high accuracy is similar to a BD LSTM RNN in a research [188] solving a current problem - customers' purchase decisions by process automation.

The prediction error in terms of a loss for the BD LSTM RNN is constantly lower than that of other two networks. The loss for the former is close to 0.67% whereas that of other two networks are 0.78% and 0.77% respectively. Accordingly, the BD LSTM RNN is better other two RNNs.

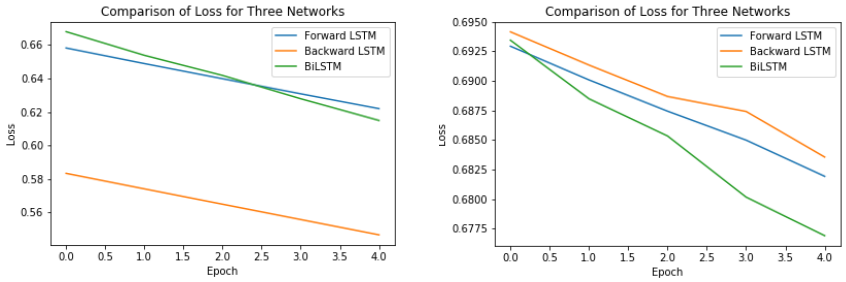
Table 3.7 Loss and Accuracy in 3 LSTM RNNs: FF, BK and BD

RNN	FF		BK		BD	
Epoch	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss
1	0	0.807	30.78	0.771	52.3	0.695
2	0	0.799	30.78	0.770	69.4	0.698
3	0.1	0.791	30.78	0.770	90.2	0.682
4	0.2	0.783	30.78	0.769	100	0.673
5	0.4	0.775	30.78	0.768	100	0.665

Among them, we further compare the loss, as shown in **Fig. 3.3**. The loss of the BK network is the lowest (0.55%), lower than that of the FF and BD networks by 0.07% to 0.08%. Nonetheless, the loss of the BD network is lower than that of the FF network.

In order to confirm whether the BK network is consistently better than other two networks, we verify the network using an OR database. The result differs, as visualized in **Fig. 3.3** - the loss of FF and BK networks is similar (0.683%) and both are higher than that of the

BD network (0.005% to 0.0075%). Thereupon, the BD LSTM RNN is superior to other networks for this OR database.



Integrated Dataset OR Database

Fig. 3.3 Loss for 3 LSTM RNNs: FF, BK and BD

Evaluation

We test the effectiveness of the BD LSTM RNN by cross-validating the accuracy and loss. Results show that the accuracy of training and testing data sets for the FF and BD LSTM RNNs is the same (100%) while that for the BK network is much lower (30.78%). This is consistent with the loss, as shown in **Table 3.8**. Training and testing loss for the former two networks decreases consistently and stabilizes at the same point, unlike the BK network. This demonstrates a good fit that training loss meets with testing loss at the end.

Table 3.8 Cross Validated Loss for 3 LSTM RNNs

Epochs	1	2	3	4	5	6	7	8	9	10
FF	0.607	0.598	0.588	0.578	0.568	0.559	0.549	0.539	0.529	0.519
BK	0.765	0.764	0.763	0.763	0.762	0.761	0.760	0.759	0.758	0.758
BD	0.598	0.588	0.578	0.568	0.559	0.549	0.539	0.529	0.519	0.509

This cross-validation technique deviates from a studying training LSTM RNNs [192] (unidirectional, bidirectional, and cascaded architectures based) to benchmark other models (e.g. SVM) for prediction.

Case Studies

To drill down into the network performance, we study the following three cases.

Case 1 - We select OR data to justify the accuracy and loss of the integrated dataset. The accuracy is found high (over 99%) for the BK and BD LSTM RNNs whereas the loss of them is as low as that of the integrated dataset - 0.68% for the FF LSTM RNN, 0.69% for the BK LSTM RNN and 0.66% for the BD LSTM RNN, as shown in **Table 3.9**. However, the loss of the BD LSTM RNN is the lowest (0.664%). In view of this, this LSTM RNN is more favorable in DQP than other networks.

Table 3.9 Accuracy and Loss of 3 LSTM RNNs for OR Data

	FF		BK		BD	
Epoch	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss
1	81.8	0.693	99.3	0.694	99.9	0.678
2	99.3	0.690	99.3	0.694	99.3	0.675
3	99.9	0.687	99.3	0.693	99.3	0.672
4	99.3	0.684	99.3	0.693	99.3	0.667
5	99.3	0.680	99.3	0.691	99.3	0.664

To find the lowest loss, we inspect four (4) loss estimation methods, as compared in **Fig. 3.4**. The loss is minimized in the summation method (0.673%) whereas the highest loss is estimated by the multiply method (0.691%). Other two methods, including concatenate and average, estimate that the loss is reduced to 0.678% to 0.683% respectively.

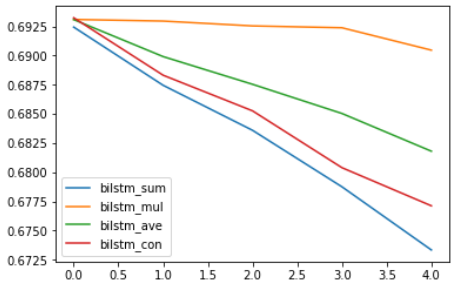


Fig. 3.4 BD LSTM RNNs' Performance by Four Methods

With these results, the summation method is preferable. From the literature review, we notice that these methods deviate from a study implementing a stacked bidirectional and BD LSTM network [90] to measure the BK dependency.

Case 2 – We conduct an independent check to see whether our prediction are convincing. To achieve this, we explore the problem of under-fitting or over-fitting for networks with a set of training data and a set of validation data. The outcome is visualized in **Fig. 3.5**.

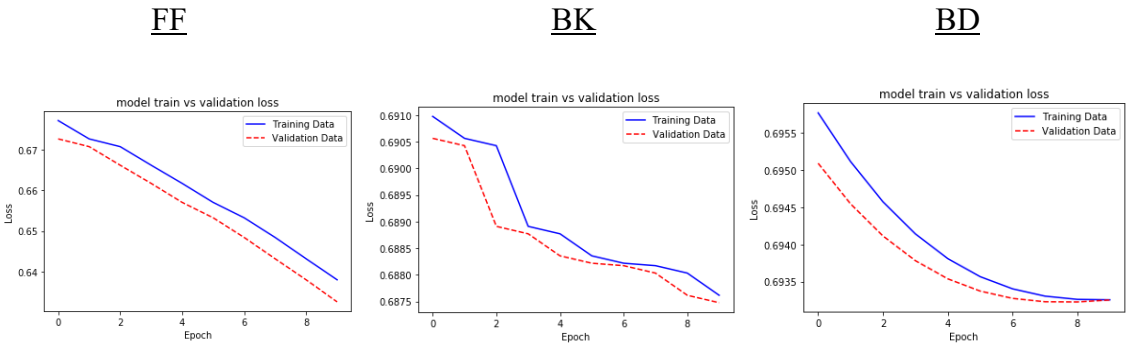


Fig. 3.5 Validated Loss of 3 LSTM RNNs for OR Data

For all networks, the loss of the validation data is found lower than that of the training data. There is no under-fit or over-fit issue. This occurs when the network generalizes well or the training set is large. In the result, there is a good fit for the BD LSTM RNN, not the other two RNNs. This is the same as that of the result for the integrated dataset – the loss of training and validation data reduces steadily and meets at the end. In view of this, we are confident of the model for DQP.

Case 3 – From the integrated dataset, the accuracy and loss of prediction for the BD LSTM RNN is 100% and 0.665% respectively utilizing the algorithm of ADAM. In order to explore whether the loss can be improved, we apply another algorithm, Stochastic Gradient Descent ("SGD"), to networks due to its merit. SGD can be used to estimate the probability of output based on a randomly selected subset of inputs with the stochastic approximation of gradient descent optimization. The output can be estimated using least squares where the objective function is minimized [235]. Results are displayed in **Table 3.10**.

Utilizing SGD, we observe that the highest accuracy occurs in the BD LSTM RNN (>99.9%) which is lower than that in the integrated dataset (100%). There is a contradictory result for the loss. Utilizing SGD, the loss for the BD LSTM RNN (0.68%) is higher than that of the BD LSTM RNN with ADAM (0.67%). In comparison, ADAM algorithm is preferable. When compared three networks applying SGD, we can see that the highest accuracy is found in the BD LSTM RNN but the loss for this is higher than that of the FF network (0.658%) and lower than that of the BK network (1.427%).

Table 3.10 Accuracy and Loss in 3 LSTM RNNs (SGD)

FF			BK		BD	
Epoch	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss
1	64.37	0.693	30.78	1.460	95.09	0.690
2	98.56	0.685	30.78	1.453	97.20	0.688
3	99.98	0.675	30.78	1.444	98.55	0.687
4	99.32	0.667	30.78	1.433	99.27	0.683
5	99.30	0.658	30.78	1.427	99.96	0.680

To confirm whether ADAM is the best for the network performance, we benchmark other algorithms. Outcomes are shown in **Fig. 3.6**. The accuracy is found high for ADAM (100%), SGD (99.96%) and ADAGRAD (100%) but the loss is the lowest for ADAGRAD (0.452%) when compared with ADAM, SGD, ADADELTA and RMSPROP. Consequently, ADAGRAD is better than ADAM in our experiments with respect to DQP.

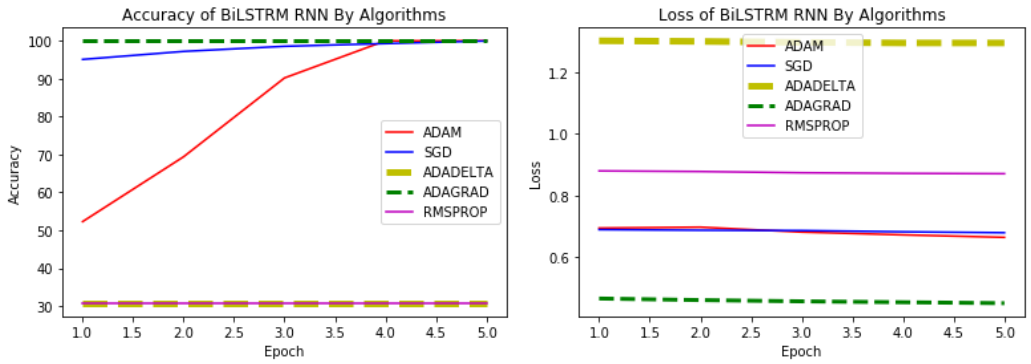


Fig. 3.6 Accuracy and Loss for BD LSTM RNN by Algorithms

Model Re-testing

To ensure that our model is effective in real life, we utilize a realistic banking dataset to predict DQ in supervised learning although this dataset covers single risk type. Re-testing results are summarized in the following paragraphs. Before an illustration of the testing, we introduce the dataset first.

This dataset contains the International Development Association (IDA) credits which are public and publicly guaranteed debt extended by the World Bank Group. Data can be found from the link of https://finances.worldbank.org/browse?category=Loans+and+Credits&limitTo=datasets

IDA provides development credits, grants and guarantees to its recipient member countries to meet their development needs. Credits from IDA are at concessional rates and data is in U.S. dollars calculated using historical rates. This dataset contains historical snapshots of the IDA Statement of Credits and Grants including the latest available snapshot. As the World Bank complies with all sanctions applicable to World Bank transactions.

In total, this dataset covers thirty (30) DEs. They are similar to data fields of our synthesized dataset. Some data features are extracted to **Table 3.11** for reference.

Table 3.11 Realistic Banking Data Features (Samples)

Attributes	Sample Values	DN Examples	Features
Borrower	STATE PLANNING COMM.	Null value	Unique=328, top=MINISTRY OF FINANCE, freq=130338
Closed Date	12/31/1984 12:00:00 AM	Date format inconsistent with %d/%m/%Y %I:%M:%S %p	Unique=1210, top=06/30/2013 12:00:00 AM, freq=12452
Project ID	P007353, P009314	89100192, PSW, null value	Unique=6818, top=P079736, freq=856
Project Name	LIVESTOCK MARKETING	na, n/a, nil, tbc, any, all	Unique=7373, top=EDUCATION II, freq=4329
Credit Number	IDA07770, IDA07780	Repeated numbers	Unique=9115, top=IDA06250, freq=111
Agreement Signing Date	4/06/1978 12:00:00 AM	Date format inconsistent with %d/%m/%Y %I:%M:%S %p	Unique=4756, top=06/23/2003 12:00:00 AM, freq=1443
Cancelled Amount	24276.41, 21912427.15	Null value	Mean= 2733973.78894761, SD= 17323842.8046962
Country Code	HN, AF, JO	4P, 6C, 8S	Unique=128, top=IN, freq=46074
Region	EUROPE AND CENTRAL ASIA	OTHER	Unique=7, top=AFRICA, freq=437374
Currency of Commitment	USD	EUR, XAF or JPY	Unique=5, top=XDR, freq=704864
Exchange Adjustment	0	-, null value	Unique=3, top=0, freq=810201
Service Charge Rate	0.75, 1.11, 1.3	0, null value	Unique=NaN, top= NaN, freq= NaN

Model Re-Testing Results

To verify the effectiveness of our ML model, we re-test the model with a "realistic banking dataset" under the same experiment settings. The testing networks and parameters (such as algorithm of ADAM) are the same as that of the model that we have experimented in this chapter. Additionally, data fields in the realistic dataset are more or less the same as that of

our synthesized dataset. More importantly, the size of this realistic dataset is large (with 842,969 records) which is comparable to our synthesized dataset (1 million).

In this re-testing, we train three networks to predict DQ including FF, BK and BD LSTM RNNs. Prediction results are extracted to **Table 3.12**. It shows the prediction accuracy and error (in terms of a loss) which are comparable to our initial experimental results tested with the synthesized dataset.

Table 3.12 (Re-Testing)
Loss and Accuracy in Three LSTM RNNs: FF, BK and BD

RNN	FF		BK		BD	
Epoch	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss
1	33.41	0.723	33.41	1.253	66.85	0.658
2	33.41	0.719	2.72	1.253	66.59	0.653
3	33.38	0.716	2.72	1.250	66.59	0.648
4	33.25	0.712	2.72	1.249	66.59	0.644
5	32.80	0.709	2.72	1.249	66.59	0.640

From these results, we note the following:

a. The accuracy for two networks (BK LSTM RNN and BD LSTM RNN) is lower (2.72% and 66.59%) in re-testing results as the accuracy for these networks in initial results is as good as 30.78% and 100% respectively; and
b. However, the accuracy for the FF LSTM RNN is better (32.80%) when compared with initial results in which the accuracy for this network is only 4%. Regardless of these, the highest accuracy in re-testing and initial results is found in the BD LSTM RNN at the end.

Unlike accuracy results, losses for the FF LSTM RNN (0.709) and BD LSTM RNN (0.640) in re-testing results are better than that of our initial experimental results (0.775 for the FF LSTM RNN and 0.665 for the BD LSTM RNN) except for the BK LSTN RNN. Finally, the lowest loss in re-testing and initial results is found in the BD LSTM RNN.

Model Re-Testing Evaluation

In the re-testing, we validate the effectiveness of three RNNs and compare their results with actual outputs. The validated accuracy for the FF LSTM RNN, BK LSTM RNN and BD LSTM RNN corresponds to 66.59%, 66.59% and 2.72%. These differ from our initial experimental results in which these figures were 100%, 100% and 30.78% respectively.

Apart from this, we validate the prediction error in terms of the validated loss, as shown in **Table 3.13**. This is equivalent to the validation made in our initial experiments of this chapter.

Table 3.13 (Re-Testing)
Cross Validated Loss for Three LSTM RNNs

Epoch	1	2	3	4	5	6	7	8	9	10
FF	0.703	0.700	0.697	0.694	0.691	0.688	0.685	0.683	0.680	0.677
BK	1.248	1.245	1.244	1.244	1.244	1.243	1.240	1.241	1.240	1.240
BD	0.632	0.629	0.626	0.624	0.621	0.619	0.618	0.616	0.615	0.614

By comparing re-testing results with initial results, we find that the lowest validated loss is found in the BD LSTM RNN. The lowest validated loss in the re-testing is 0.614 whereas that in initial results is 0.509. They are relatively similar. In the re-testing, the validated loss for the FF LSTN RNN (0.677) and BK LSTM RNN (1.240) is much higher than that of the same networks in our initial experimental results (0.519 and 0.758).

Above performance measurements such as accuracy and loss are defined in equations 2.37 and 2.38 respectively inside Chapter 2.

As we can see, our experiments show the effectiveness of our model on DQP in supervised learning by conducting a set of experiments over the synthesized dataset as well as the realistic banking dataset.

Related Works

How to decide DN has been discussed in divergent fields of academia. Prior research commonly decided them based on three dimensions: accuracy, completeness and timeliness [238, 239]. Some used metrics [132]. Other researchers addressed the same problem such as missing data [58]. Their dimensions did not consider regulatory requirements. Additionally, we introduce other factors to decide data noise including, omission of translating data, negligence in data transformation, and data which are redundant, duplicated, unreasonable, invalid, mis-match, incomplete, missing and stale.

Factors are derived from an international standard, BCBS 239 [4] requiring an aggregation of risk data and the measurement of their DQ according to DQ principles. In the model, we define 10 factors. This is novel in comparison with prior researchers addressing the same problem [240, 241, 242].

Reviewing the literature of sequential learning in DNN, we notice that the importance is continuously growing. Many earlier works were related to prediction. However, none of them was learnt for DQP. They centred on other domains: Some researchers have investigated the application of sequential learning in speech recognition [88], video captioning [89], reading comprehension [90], ads recommendation [91] and natural language processing [92].

3.4 Summary

In this chapter, we propose a ML model for DQP in supervised learning. In supervised learning, we label data based on DQ principles of the international requirement, BCBS 239. **Table 3.6** shows these principles including the principle # 3 - accuracy and integrity, principle # 4 - completeness and principle # 5 – timeliness.

With labelled data, we input the labelled data in LSTM RNNs for DQP. Three networks are implemented in the model including FF, BK and BD LSTM RNNs, as mentioned in **Table 3.1**. Their learning methodologies, sequence prediction and system architecture are explained in Section 3.2.2, 3.2.3 and 3.2.4 respectively.

In experiments, we firstly direct networks to learn temporal correlations between DN sequences with a synthesised dataset, as mentioned in Section 3.3. Afterwards, we re-test the model with a realistic banking dataset. Both experimental results demonstrate that our model is effective in training synthesized and realistic datasets for DQP. The effectiveness is confirmed in BD LSTM RNN results. The prediction accuracy in our initial results and in the re-testing is relatively high and the prediction error (in terms of a loss) is low. The network performance is further elaborated with some case studies.

Prediction of DQ accurately in LSTM RNNs are advantageous to the financial services industry. FIs can understand what DQ are going to be with a sceientific computational method. This enhances their forward-looking capabilities of DQ by providing any potential violations of risk limits over thresholds in advance. Thereupon, DQ can be improved in long term [4] which is consistent with the expectation of financial regulators.

The next chapter presents DQP using unsupervised learning with a ML model. Model networks learn the importance of DN on top of the temporal sequences and correlations of DN collectively in DQP.

Chapter 4

Data Quality Prediction in Unsupervised Learning

Chapter 3 focuses on DQP under DG in supervised learning with ML models. In this chapter, we make prediction with a ML model using a more advance learning method - unsupervised learning. This is part of DGP under the DG framework.

Although model networks in this chapter are equivalent to that of Chapter 3, they are additionally applied with an ATTN mentioned in Section 2.1.3 of Chapter 2. The model in this chapter considers the importance of DN on top of the temporal sequences and correlations of DN collectively in DQP. Additionally, this model takes temporal sequences and correlations of DN into account in DQ measurement before DQP.

In this chapter, Section 4.1 provides an introduction. Section 4.2 describes the way to detect DN based on rules deriving from an international requirement, BCBS 239. Detected DN impacts are estimated in terms of weights by generative mixture methods in unsupervised learning. These are input into networks for DQP. In Section 4.3, we report experiments and network evaluation results. Finally, we summarize this chapter in Section 4.4.

4.1 Introduction

From research results summarized in Section 1.2.3 and elaborated in Section 2.2, there are limitations of ML work related to DQP in unsupervised learning under DG: a) c) ML techniques have not been applied in DQP using unsupervised learning. In particular, a ML model has not been proposed for DQP in unsupervised learning; b) temporal sequences and correlations of DN have not been considered in DQ measurement. In particular, a ML model has not been proposed for DQ measurement under DG taking temporal sequences and correlations of DN into account; c) ML techniques have not been applied in the forecast of DQ during DG processes by taking temporal sequences, correlations and importance of

DN into account. In particular, a ML model considering temporal sequences, correlations and importance of DN has not been proposed for DQP; and d) ML techniques have not been applied to DQ learning such as DQP during DG processes for meeting DG regulatory requirements.

To address these, we leverage a ML model in unsupervised learning to predict DQ in accordance with the international requirement, BCBS 239 [3] in this chapter. This model improves DQ [4, 6] during DG processes.

In unsupervised learning, DN impacts are estimated in two generative mixture methods in terms of weights by taking their correlations into account after the detection of DN. These methods are GMM and BGMM.

To predict DQ, the model is implemented by LSTM RNNs, similar to that in Chapter 3. On top of the sequential learning, we apply a new mechanism, ATTN, to these networks in this chapter. An ATTN directs network learning to pay attention to more important DN.

4.2 Proposed Model

DQ are predicted in a ML model with DNN. Before DQP, we approach the weighing of DN impacts by two methods, GMM and BGMM, in unsupervised learning after DN are detected. How DN are detected and weighed before the DQP is outlined in **Fig. 4.1**.

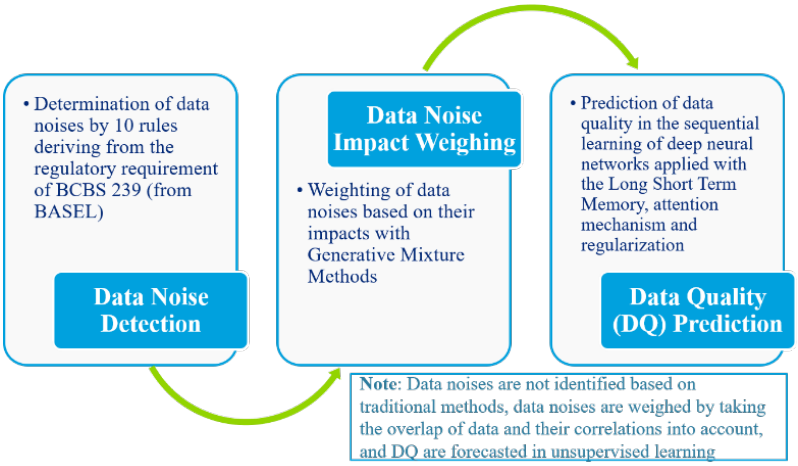


Fig. 4.1 Model Overview

In the model, we take three steps as elaborated below.

4.2.1 Data Noise Detection

First Step

DN are detected as anomalies as they deviate from normal attribute values. The detection for each DE is made in a fitness function [167] based on ten rules [130, 131] (below) defined as DQ Ratings (DQR_i) [179] in the model. DQR_i in the following paragraphs are classified into three criteria (DQC_i) according to DQ principles of BCBS 239 [4].

DQR_1 Interpretability: The degree to which DEs are interpreted in appropriate languages [173]. An actual banking value may not be translated from a foreign currency (such as USD) into a standard currency (such as EUR):

$$f\left(l_{actual,}l_{converted}\right) = \begin{cases} 1, l_{actual} = l_{converted} \\ 0, l_{actual} \neq l_{converted} \end{cases} \quad (4.1)$$

DQR₂ Conformity: The degree to which DEs conform to a format [133]. E.g. DD/MM/YY. The value decreases with the number of mis-matches with a data dictionary.

$$f(o_{actual,}o_{dictionary}) = \begin{cases} 1, o_{actual} = o_{dictionary} \\ 0, o_{actual} \neq o_{dictionary} \end{cases} \quad (4.2)$$

DQR₃ Indispensability: The degree to which DEs are critical [133]. A client incapable of passing a bank due diligence check due to anti-money laundering may be wrongly deemed as a true essential client:

$$f(r_{actual,\ } r_{due-diligence}) = \begin{cases} 1, r_{actual} = r_{due-diligence} \\ 0, r_{actual} \neq r_{due-diligence} \end{cases} \quad (4.3)$$

DQR₄ Uniqueness: The degree to which a DE is unique [175]. A DE with extra instances is non-unique. Duplicated customer IDs in a banking dataset reduce the value of uniqueness:

$$f(d_{actual,\ } d_{non-unique}) = \begin{cases} 1, d_{actual} \notin d_{non-unique} \\ 0, d_{actual} \in d_{non-unique} \end{cases} \quad (4.4)$$

DQR₅ Timeliness: The degree to which DEs are kept up to date. It is a lag between present time and last update time over a data retention period dependent on a bank's policy. The period is an expiry period which is a time limit of data retention [134].

$$f(t_{now}, t_{previous}, t_{retention}) = \begin{cases} 1, if\ t_{now} - t_{previous} \leq t_{retention} \\ 0, if\ t_{now} - t_{previous} > t_{retention} \end{cases} \quad (4.5)$$

DQR₆ Believability: The degree to which DEs are regarded as credible [133]. A banking client classified with a risk level other than low is an unbelievable customer.

$$f(b_{actual,}\ b_{low-risk}) = \begin{cases} 1, b_{actual} = b_{low-risk} \\ 0, b_{actual} \neq b_{low-risk} \end{cases} \quad (4.6)$$

DQR$_7$ Validity: The degree to which DEs have a right age over a valid period [169, 173]. A client younger than a valid age, $v_{right-age}$ =12, holding a bank account is invalid:

$$f(v_{actual,}\ v_{right-age}) = \begin{cases} 1, v_{actual} \leq v_{right-age} \\ 0, v_{actual} > v_{right-age} \end{cases} \quad (4.7)$$

DQR$_8$ Consistency: The degree to which contents are matched for a DE from different perspectives. When the value of a DE in a banking database mismatches with that of other databases ($s_1, \ldots, s_4$), there is a distance (D) [137, 173], reducing the value of consistency.

$$f(s_{actual,}\ s_{D1,\ldots,D4}) = \begin{cases} 1, & if\ s_{D1} = s_{D2} = s_{D3} = s_{D4} \\ 0, & otherwise \end{cases} \quad (4.8)$$

DQR$_9$ Completeness: The degree to which DEs are complete. A phone digit deviating from a standard, such as 11 digits, in a banking dataset is incomplete due to data corruption [137, 138]:

$$f(c_{actual,}\ c_{digit}) = \begin{cases} 1, & if\ c_{actual}\ is\ standard \\ 0, & if\ c_{actual}\ is\ NOT\ standard \end{cases} \quad (4.9)$$

DQR$_{10}$ Availability: The degree to which a value of DEs is available [175]. An existing DE containing a null value [139] is deemed as unavailable:

$$f(u_{actual,}\ u_{blank}) = \begin{cases} 1, if\ c_{actual}\ is\ NOT\ blank \\ 0, \quad if\ c_{actual}\ is\ blank \end{cases} \quad (4.10)$$

Above rules are defined according to an inductive logic programming and rule mining approach [140] and run by an operator, Python. With current DN, we can infer and determine target noise.

4.2.2 Data Noise Impact Analysis in Unsupervised Learning

Second Step

Detected values are used as data points (x_i) to estimate impacts of DN on each DE. Impacts are estimated by the probability of attribute values in two methods in terms of the probability density function ("PDF"). All PDFs are aggregated [135] before they are scored as the joint occurrence probability as follows:

$$P(x_1 = d_1) \cdot P(x_2 = d_2) \cdot \ldots \cdot P(x_{132} = d_{132}) \tag{4.11}$$

This is a weighted sum of DQR scores (DQR_i) [130, 164] which are input into networks for prediction. The sum is the DQ level for each DQC_i [157, 163]. In the model, DQR_i are categorized into three DQ dimensions [124, 127, 238]: $DQC_1 = DQR_1 \cdot DQR_2 \cdot DQR_3 \cdot DQR_4 \cdot DQR_6 \cdot DQR_7 \cdot DQR_8$, $DQC_2 = DQR_9 \cdot DQR_{10}$ and $DQC_3 = DQR_5$.

To aggregate DN for scoring, we customize an aggregate quality scoring algorithm in **Fig. 4.2**. This algorithm shows the pseudo code for scoring aggregate DQ. Before scoring, we detect DN. After scoring, we analyze DN impacts prior to forecasting DQ. This algorithm provides guidance on the measurement of DQ by a new DQ scoring method under DG. This method has not been proposed in earlier ML research related to DG, as mentioned in Section 1.2. A ML model is yet to be developed for DQP during DG processes in supervised and unsupervised learning.

ALGORITHM: The Pseudo-code of the Aggregate Quality Measurement

Input: Attribute values $(x_1,\ldots,x_{132})$ may be embedded with data noises.
Output: Aggregate DQ (y_i) of each attribute.
1: Let the number of DQ dimensions (d_m) as three: {accuracy, completeness, timelineness} representing the DQC_i.
2: Let the number of DQR_i as ten (10) under the DQC_i.
3: Let the number of DEs (N') as 132.
4: Let the number of instances (i) as one (1) million.
5: Find attribute values from a set of N' DEs.
6: **For each** i **do**
7: Detect attribute values for each i based on rules, DQR_i.
8: Find DQ noise(s) under the fitness function:

$$f(x_i) = \begin{cases} 0, & x_i < 1 \\ 1, & x_i \geq 1 \end{cases}$$

9: **end for**
10: **For each** DQR_i **do**
11: Compute the probability of attribute value for each i by weight for each Gaussian π_k, mean of the Gaussians, u_k and covariance of each Gaussian $\sum_k$.
12: Estimate the probability of a Gaussian j generating an attribute for each i with update a new weight, a new mean and a new covariance.
13. **end for**
14: **For each** DQC_i **do**
15: Aggregate new weights π_k as PDFs for each DQR_i.
16: Calculate PDFs under DQC_i for each attribute to generate DQ (y_i): $P(x_1 = d_1) \cdot P(x_2 = d_2) \cdot \ldots \cdot P(x_{132} = d_{132})$.
17. **end for**
18: Compute the frequency of all DQ dimensions for DEs:

$$x_1 \in \{0,1\}^{d_1} \cdot x_2 \in \{0,1\}^{d_2} \cdot x_3 \in \{0,1\}^{d_3}$$

19: **return** $f(y_i)$.

Fig. 4.2 Algorithm for Aggregate Quality Scoring

For measuring impacts of DN, we use two methods including GMM [102] and BGMM. Their outputs are multivariate PDFs as weights of DN. BGMM is a mixture of Gaussian Distribution with a DP. But it is an extension from the GMM. Both methods help to find weights [84].

Take a set of latent groups as an example, we observe that there are three scenarios: One of the scenarios is that a Gaussian is centered at (DQR_1 = 1, DQR_2 = 1, . . . , DQR_{10} = 1) and so data points are free from DN. Another scenario is that a Gaussian is centered at (DQR_1 = 0, DQR_2 = 1, . . . , DQR_{10} = 1) indicating that some DN are embedded. Last scenario is that a Gaussian is centered at (DQR_1 = 0, DQR_2 = 0, . . . , DQR_{10} = 0) where data points are full of noise.

In estimating impacts, we take the following steps in the two generative mixture methods.

GMM. The probability of a data point is the weighted sum of k Gaussians while $k = 1$:

$$p(x_n|parameters) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|u_k, \sum_k) \tag{4.12}$$

where π_k represent weights for each Gaussian, u_k denote means of Gaussians and $\sum_k$ indicate covariances of each Gaussian. With the combination of k Gaussians, we compute means and covariances. In order to learn weights, means, covariances for each Gaussian, we reform the GMM by incorporating a new variable z and a uniform number generator, and apply an Expectation Maximization algorithm. In the step of Expectation, we determine the probability of a Gaussian $\mathbb{E}(z_{ik}) = r_{ik}$ to generate a data point x_i by calculating the probability of a Gaussian, $\pi_k \mathcal{N}(x_i|u_k, \sum_k)$. Following this, we normalize it through $\sum_j \pi_j \mathcal{N}(x_i|u_j, \sum_j)$. This assesses responsibilities (r) based on current Gaussian. In the step of Maximization, we estimate a new weight, mean and covariance for each data point with an equation $\pi_k = \frac{N_k}{N}$ and then another equation $u_k = u_k = \frac{1}{N_k} \sum_{i=1}^{N} r_{ik} x_i$. Afterwards, we find outputs $\sum_k$ using the following equation:

$$\sum_k = \frac{1}{N_k} \sum_{i=1}^{N} r_{ik} (x_n - u_k)\,(x_n - u_k)^T \tag{4.13}$$

In this equation, the new weight is a sum of probabilities of a Gaussian k divided by the number of points. A new mean is a multiplication of probabilities for that cluster and a new covariance is multiplied by probabilities for that cluster. By utilizing these equations, we take data correlations into account. This approach is similar to a study [86] assessing weights by estimating their distributions with a GMM and capturing the relations between context information and DQ.

BGMM. The probability $p(\mu, \sum)$ is sampled from a Dirichlet distribution and is computed with the following equation:

$$\mathcal{N}(\mu|m_0, (\beta_0 \sum^{-1})^{-1}) W(\sum^{-1}|, W_0 v_0) \tag{4.14}$$

From this equation, W_0 is a general shape determining the variability of samples v_0, a center m_0 and a constant β_0. This indicates how far the mean should be from m_0 on average. The probability of the BGMM is $p(x_n|\pi, \mu, \sum)$. This is similar to the GMM:

$$\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Lambda_k) \text{ where } \Lambda = \sum)^{-1} \tag{4.15}$$

However, the joint probability, $p(X, Z, \mu, \sum, \pi)$, is more complex: $p(X|Z, \mu, \sum) p(Z|\pi) p(\pi) p(\mu|\sum) p(\sum)$. To integrate unobserved variables, we can find the probability $p(X)$ as:

$$\int \int \int \int p(X|Z, \mu, \Lambda)\, p(Z|\pi) p(\pi) p(\mu|\Lambda)\, p(\Lambda) dZ d\mu d\Lambda d\pi \tag{4.16}$$

This integration is intractable. For the computation of the weight estimation error, we adopt a weighted algorithm to minimize MSE in an objective function [111] below:

$$\frac{1}{2} \sum_{x^{(k)} \in D^k} \sum_{y} \left(P(y|x^{(k)}) - \hat{P}(y|x^{(k)}) \right)^2 \tag{4.17}$$

In this equation, D represents a dataset and w refers to a weight of an attribute. On top of this, we estimate errors in terms of explained variances [123]. This is similar to a study [121] adopting Bayesian's weighing of high-quality data and with incomplete data for testing. This study identified that the smallest weight of solid-state data was 0.3 but the largest weight was 1. The key was that the optimal weight of high-quality data selected a large value and decreased sharply whenever noise and sparseness increased.

As we can see, these two methods estimate DN weights in unsupervised learning. These weights are used to predict DQ in experiments.

4.2.3 Data Quality Prediction with an Attention Mechanism

Third Step

To predict DQ, we input PDFs into LSTM RNNs [49, 59, 188] for learning. These networks are the same as that in Chapter 3.

In total, six networks are implemented. They are three LSTM RNNs with and without an ATTN. Three LSTM RNNs are FF, BK and BD LSTM RNNs.

The network learning is directed by an ATTN while the network performance is optimized by two regularizations including regularization 1 and regularization 2.

Network Learning

In sequential learning of LSTM RNNs, we direct the network to pay attention to more important DN for DQP by applying an ATTN.

An ATTN enables networks to focus on selective or specific information [50, 103, 172]. It assigns different weights to different data. Consequently, the network pays attention to important data [97]. It can be used to capture inter-relationships among data. The intensity of each DE is affected by others. Accordingly, attentions for some data could be distracted and others could be enhanced. Outputs turn out to be sequences taking the impacts of others into account. They are calculated by summarizing input representations with different attention weights.

In the model, ATTN is applied to LSTM layer. If data inputs at a given time are important, the network learning algorithm updates the memory cell of the LSTM layer by importing

more information. If not, the network suppresses its impact on the memory and takes more historical information. Consequently, the network selectively attends to the most related DN. This enables the model to adaptively attend to important DN [104].

DQ learn differently in divergent LSTM RNNs with an ATTN, as described below.

Feedforward Sequential Learning. The FF network learns from the past and future prediction of a timestep to predict DQ in a feedforwarding (*f*) way:

$$y_{t+1}^{f1}, h_{t+1}^{f1}, c_{t+1}^{f1} = LSTM^{f1}(c_{t-1}^{f1}, h_{t-1}^{f1}, x_t;\ W^{f1}) \tag{4.18}$$

At first, the network hidden state and cell state of each layer are initialized at 0. In the network, the 1st layer uses an input at a time (x_t), previous hidden state (h_{t-1}^{1}) and previous internal hidden state (c_{t-1}^{1}) to generate an output (y_{t+1}^{f1}) with a weight (W^{f1}).

The LSTM computes the forward hidden sequence $\vec{h}$ with a bias (b) to generate an output below:

$$y_{t+1} = W_{\vec{h}_y}\,\vec{h}_t + b_y \tag{4.19}$$

The input (x_t) is a sequence passing to the 1st layer (1) at a time (t-1, t to t+1). Input attributes in the network generating outputs are read from left to right. How to generate outputs is described in **Fig. 4.3** showing an algorithm of the FF LSTM RNN:

ALGORITHM The pseudo-code of the FF LSTM RNN

Input: The value of each attribute (x_1,…,x_{132}) as the weight of each DE (PDF).
Output: The predicted output sequence (y_{t+1}).
1: Train value of attribute to generate vectors.
2: **for** n =1 → N (N represents the number of PDFs) **do**
3: **for** m =1 → M (M represents the number of DN or DQ issues) **do**
4: Do encoding to obtain the vector of each input sequence.
5: Employ LSTM to obtain the sequence representations and compute forward hidden sequence $\vec{h}$.
6: **for** k =1→ K (K represents the number of LSTM layers) **do**
7: The output is a sequence of hidden states at each time step t (t = *1*, *2*,…t) in TimeDistributed wrapper layer.
8: The output of previous layer is input into the network to obtain a weight ($W_{\vec{h}_y}$).
9: The output of step 7 would calculate the element-wise product with the output of step 8.
10: The output of step 9 would add with the input of step 7 as new sequence representations.
11: **end for**
12: **end for**
13: Employ classifier to obtain the output sequence of each attribute.
14: Update parameters by the ADAM algorithm.
15: **end for**

Fig. 4.3 Algorithm for FF LSTM RNN

Backward Sequential Learning. In training the BK network, LSTM computes backward hidden sequences $\overleftarrow{h}$ instead:

$$y_{t+1} = W_{\overleftarrow{h}_y} \overleftarrow{h}_t + b_y \tag{4.20}$$

where input attributes generating outputs are read from right to left. This algorithm for BK network is customized in **Fig. 4.4**:

ALGORITHM The pseudo-code of the BK LSTM RNN

Input: The value of each attribute (x_1,…,x_{132}) as the weight of each DE (PDF).
Output: The predicted output sequence (y_{t+1}).
1: Train value of attribute to generate vectors.
2: **for** n =1 → N (N represents the number of PDFs) **do**
3: **for** m =1 → M (M represents the number of DN or DQ issues) **do**
4: Do encoding to obtain the vector of each input sequence.
5: Employ LSTM to obtain the sequence representations and compute backward hidden sequence $\overleftarrow{h}$.
6: **for** k =1→ K (K represents the number of LSTM layers) **do**
7: The output is a sequence of hidden states at each time step t (t = *1*, *2*,…t) in TimeDistributed wrapper layer.
8: The output of previous layer is input into the network to obtain a weight ($W_{\overleftarrow{h}_y}$).
9: The output of step 7 would calculate the element-wise product with the output of step 8.
10: The output of step 9 would add with the input of step 7 as new sequence representations.
11: **end for**
12: **end for**
13: Employ classifier to obtain the output sequence of each attribute.
14: Update parameters by the ADAM algorithm.
15: **end for**

Fig. 4.4 Algorithm for BK LSTM RNN

Bi-directional Sequential Learning. Two LSTMs are trained based on timesteps of input sequences on current input sequence and then on a reversed input sequence. The BD network combines forward and backward LSTM outputs to predict DQ.

As the network is BD, the network computes a forward hidden sequence $\vec{h}$ and a backward hidden sequence $\overleftarrow{h}$ and then combines them to generate an output:

$$y_t = W_{\vec{h}_y} \vec{h}_t + W_{\overleftarrow{h}_y} \overleftarrow{h}_t + b_y \tag{4.21}$$

An input attribute in the FF network is read from left to right while that in the BK network is read from right to left. A combination of these forms a BD network. An algorithm for the BD LSTM RNN is given in **Fig. 4.5**:

ALGORITHM The pseudo-code of the BD LSTM RNN

Input: The value of each attribute (x_1,…,x_{132}) as the weight of each DE (PDF).
Output: The predicted output sequence (y_{t+1}).
1: Train value of attribute to generate vectors.
2: **for** n =1 → N (N represents the number of PDFs) **do**
3: **for** m =1 → M (M represents the number of DN or DQ issues) **do**
4: Do encoding to obtain the vector of each input sequence.
5: Employ LSTM to obtain the sequence representations and incorporate gates to manage the state of cells (c_t).
6: **for** k =1→ K (K represents the number of LSTM layers) **do**
7: The output is a forward hidden sequence of hidden states $\vec{h}_t$ where the sequence of hidden previous layer is input into the network to obtain a weight ($W_{\vec{h}_y}$).
8: The output is a backward hidden sequence of hidden states $\overleftarrow{h}_t$ where the sequence of hidden previous layer is input into the network to obtain a weight ($W_{\overleftarrow{h}_y}$).
9: The output of step 7 is combined with the output of step 8 as new sequence representations (y_{t+1}).
10: **end for**
11: **end for**
12: Employ classifier to obtain the output sequence of each attribute.
13: Update parameters by the ADAM algorithm.
14: **end for**

Fig. 4.5 Algorithm for BD LSTM RNN

All these algorithms provide guidance on DQP by different DQ learning methods under DG. These methods have not been proposed in earlier ML research related to DG, as mentioned in Section 1.2.3 of Chapter 1 and 2.2 of Chapter 2. A ML model has not been proposed for DQP during DG processes in supervised and unsupervised learning.

Following this, the network learning is optimized by two regularizations including regularization 1 and regularization 2.

Regularization 1. To increase the predictive power of the model, we select a network with an outstanding performance in experiments for testing. It is trained separately on four data sub-sets including MR, CR, OR and LR. These data sets [206] or recaptured data sub-sets [207] aid in refining the model by exploring which parts of data can be trained efficiently.

Regularization 2. To alleviate an overfitting problem in prediction, we apply regularization [100] to the worst prediction result. The regularization improves the prediction error in terms of a loss and MSE. We continue to follow a pioneering work [205] to add dropout to avoid co-adaption of hidden units through the omission of features in network propagation.

The dropout is applied on the LSTM layer. On top of this, we add $L2-$ norms of weight vectors to scale weights to a level equivalent to DQ.

4.3 Experiments

The experiment setup and data split are the same as that in Chapter 3. The purpose of these experiments is to demonstrate the effectiveness of our model on DQP in unsupervised learning by conducting a set of experiments over a dataset.

Dataset

The dataset in this chapter is from Chapter 3.

Results

First Result - Detected DN: Detected DN from the dataset are visualized in **Fig. 4.6**. They are classified into four categories including MR, CR, OR and LR. Take OR as an example, the believability value of an attribute, LossFrequency, is 51% whereas its validity and consistency values correspond to 81% and 100%. Instead, values of believability, validity and consistency of an attribute, Loss-Severity, are 100%, 73% and 100% respectively while that of an attribute, OpsLoss, are 100%, 97% and 100% respectively.

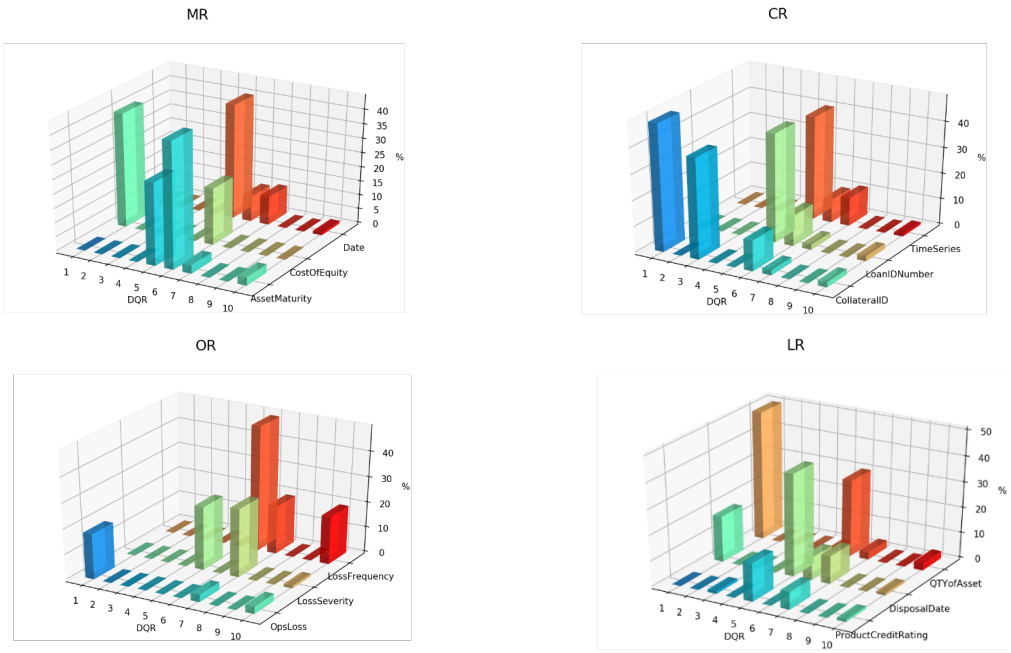


Fig. 4.6 Detected DN

As we can see, the value of attributes is measured in terms of DQ metrics by a percentage.

Second Result – Weighed DN Impacts: Impacts of DN are estimated as weights. Impacts for twelve attributes by DQR_i under two methods including GMM (G) and BGMM (B) are listed in **Table 4.1**.

Table 4.1 Impacts of DN in Terms of PDFs

Attribute	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
DQR_i	1	3	6	3	5	7	1	7	10	3	1	6
B	0.49	**0.20**	**0.56**	**0.39**	0.41	0.10	0.83	**0.26**	0.82	0.86	0.51	0.92
G	0.49	**0.80**	**0.00**	**0.61**	0.41	0.10	0.83	**0.74**	0.82	0.86	0.51	0.92

From **Table 4.1**, PDFs are the same in both methods except four attributes including x_2, x_3, x_4 and x_8. To understand their estimation effectiveness, we measure prediction errors in terms of MSE and explained variances ("VAR"), as listed in **Table 4.2**. MSE can relax the

zero-bias assumption [236] while VAR in both positive and negative vectors can still be used to explain the association of data features [237] since the variance is not due to error variance. VAR also gives the percentage of variance explained by the regression.

Table 4.2 Impact Estimation Errors: MSE vs VAR

Attribute	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
B – MSE	0.00	0.00	1.00	0.00	5.71	3.69	1.00	**0.00**	1.00	1.00	31.91	0.34
G – MSE	0.00	1.00	4.55	1.00	5.71	3.69	1.00	**1.00**	1.00	1.00	31.91	0.34
B – VAR	1.00	1.00	-3.00	1.00	-15.00	-8.00	-3.00	**-1.00**	-3.00	-3.00	-79.80	-3.00
G - VAR	1.00	-3.00	-3.00	-3.00	-15.00	-8.00	-3.00	**-3.00**	-3.00	-3.00	-79.80	-3.00

MSEs are the same for all attributes under both methods except four attributes (x_2, x_3, x_4 and x_8) whose MSEs are consistently lower in the BGMM relative to the GMM. Similarly, VARs are found identical for all attributes except three attributes (x_2, x_4 and x_8). In consideration of these, BGMM method is superior. This superior method's results are used as inputs into networks for DQP.

Third Result – DQP: In training three networks with and without an ATTN, we output predicted DQ. The network learning performance is measured in terms of the PN, RL and F1, as summarized in **Table 4.3**.

Table 4.3 PN, RL and F1 by Networks

LSTM RNNs	**PN**	**RL**	**F1**
FF	64%	**80%**	**71%**
FF+ATTN	**74%**	**80%**	**71%**
BK	68%	47%	52%
BK+ATTN	64%	**80%**	**71%**
BD	68%	72%	70%
BD+ATTN	64%	**80%**	**71%**

From **Table 4.3**, networks with an ATTN achieve better results. The maximum PN is 74%, average RL is 80% and F1 is 71% in general. The best performance is found in the network of FF+ATTN.

Fourth Result – DQP Analysis: To analyze DQP at a granular level, we disaggregate twelve attributes into four parts: MR, CR, OR and LR. Each is input into the best network, FF LSTM RNN with an ATTN, for learning. Results are visualized in **Fig. 4.7**.

Fig. 4.7 PN, RL and F1 by Databases

Fig. 4.7 shows the average score for the PN, RL and F1. Specifically, the PN is the highest in the LR (74%), better than that of the MR (64%), CR (35%) and OR (54%). This PN is apparently higher than the average PN by 23%. Consistently, the best RL and F1 are found in this risk type (86% and 80% respectively). In comparison, the RL in the LR (86%) is higher than that of others by 6%, 27% and 12% respectively while the F1 in the LR (80%) is superior to others by 9%, 36% and 18% respectively. The best RL and F1 are significantly better than the average RL and F1 by 13% and 22% respectively.

Evaluation

To test the model effectiveness, we verify networks with validation data. The validation is measured in terms of the accuracy, loss and MSE, as listed in **Table 4.4**, which is compared with the initial network performance (as shown in **Table 4.5**).

Table 4.4 Validated (V) Accuracy, Loss and MSE

Databases	V. Accuracy	V. Loss	V. MSE
MR	80.08%	**0.537**	**0.175**
CR	58.86%	0.680	0.244
OR	73.56%	0.586	0.198
LR	**86.05%**	0.544	0.177

Table 4.5 Prediction Accuracy, Loss and MSE

Databases	Accuracy	Loss	MSE
MR	80.08%	**0.542**	**0.177**
CR	58.86%	0.681	0.244
OR	73.56%	0.589	0.199
LR	**86.05%**	0.552	0.180

From **Table 4.5**, the network of FF+ATTN predicts DQ accurately for four risk types. The accuracy for all is high ranging from 58.86% to 86.05% while their prediction errors in terms of losses are limited to a range of 0.542 to 0.681. Furthermore, all MSEs are kept at a comparably low level from 0.537 to 0.680.

When comparing these with validated results (**Table 4.4**), we can see that lowest validated loss and validated MSE are found in the MR (0.537 and 0.175 respectively), lower than that of other three risks (by 27% and 39% at most). In consideration of the accuracy, loss, MSE and PN/ RL/ F1 collectively, the poor performance occurs in the CR in this experiment.

DQP Improvement

To further improve the performance, we apply regularization to the best network (FF+ATTN) with the CR database for testing. CR database is selected due to its poor performance found in last two tables. Outcomes are summarized in **Table 4.6**.

Table 4.6 Regularized Network Prediction Improvement

Database	CR
Accuracy	58.86%
Loss	**0.679**
MSE	**0.243**
PN	35%
RL	59%
F1	44%

By adding regularization, we observe that the network of FF+ATTN attains the same accuracy, PN, RL and F1. Instead, both loss and MSE improve from 0.681 to 0.679 and from 0.244 to 0.243 respectively.

DQP Improvement Evaluation

For ascertaining the effectiveness of the model, we verify the prediction with validation data. Outcomes are similar to that of prediction outcomes, as shown in **Table 4.7**.

Table 4.7 Regularized Network Prediction Improvement Validation

Database	CR
V. Accuracy	58.86%
V. Loss	**0.679**
V. MSE	0.243

Table 4.7 shows that validated loss and validated MSE are lower from 0.680 to 0.679 and 0.244 to 0.243 respectively. These are equivalent to [184] regularizing an ATTN in an entropy term to encourage attention weights to be uniform and to penalize excessive attention to a certain region. Accordingly, the prediction error is improved due to the gradual learning from the ATTN.

Above performance measurements such as accuracy, loss, precision, recall and F1-Support are defined in equations 2.37 to 2.41 inside Chapter 2.

As we can see, our experiments over the prepared dataset show the effectiveness of our model on DQP in unsupervised learning by conducting a set of experiments over a dataset.

Related Works

The way to determine DN has been discussed in different fields of academia. Earlier research generally determined them according to 3 dimensions: accuracy, completeness and timeliness [238, 239]. Some used metrics [132]. Other researchers addressed the same problem such as missing data [58]. Their dimensions did not consider regulatory requirements, unlike us.

DQP are yet extended to unsupervised learning to a great extent. This can be seen from current situation: a) GMM [102] has been used to exploit a connection between the statistical estimation and clustering problems in computational geometry; b) BGMM [84] has been used for learning new topics in a set of conversations. These unsupervised learning are yet applied to the estimation of DN weights; c) Other unsupervised learning methods estimated the density of paper currency [85] and the sensitivity of data from an Austrian bank [190] instead of DN; and d) a recent study [191] showed that a GMM was used to build a semi-supervised model in multi-mode processes for forecasting the quality of big data. This is far from unsupervised learning of DN or DQ. Furthermore, there was limited prior research on DQP according to regulatory requirements in supervised learning. In the past, there was a prediction of missing data [58] which did not take regulatory

requirements into account. In contrast, there have been ample research learning non-DQ domains in supervised learning. An example of this is a classification of speech signals in RNN by labelling noisy and unsegmented sequence data [87].

In the literature review of sequential learning for DNN, we find that the importance is continuously growing. Plenty of earlier works were associated with prediction. However, none of them was learnt for DQP. Instead, they focused on other domains: Some researchers have investigated the application of sequential learning in speech recognition [88], video captioning [89], reading comprehension [90], ads recommendation [91] and natural language processing [92].

Despite this, sequential learning has been successfully applied to various research to explore dependencies of an image content problem [93]. Also, [95] proposed a generative adversarial network with self-attention to achieve an improvement in the balance between the ability to model dependencies. The most recent one [96] predicted attributes of images by taking co-occurrence dependencies among attributes into account. Thereupon, sequential learning can be applied to our model to explore DN correlations.

In recent years, ATTN is one of the major mechanisms that has been applied to DNN. It assigns divergent weights to various data to direct networks to pay attention to important data [97]. The application was on a few domains: A previous study [89] presented a temporal ATTN for video caption generation whereas a study [182] introduced a deep attention selective network for image classification. Some used it to recognize 3D action [99]. Others leveraged it for machine translation [92, 100] and document classification [101]. [102] presented temporal attention on different time steps for electronic health records. In recent few years, ATTN has been introduced to use encoder to reference records dynamically in the decoder [100]. All these are yet applied to DQP to pay special attention to important DN.

From above results, prior research on DQP with ML models were rare. They are yet to propose any algorithms relating to the scoring of aggregate DQ or heterogenous learning

methods for DQP.

In the financial services industry, to predict DQ is challenging. [6] has emphasized that many FIs are yet to comply with DQ principles as stated in BCBS 239. In a later publication, statistical researchers and the financial regulators suggested that big data quality can be statistically computed by ML techniques [243]. From the literature review, there have been restricted research on DQP under DG for the industry. On the contrary, there have been abundant prediction of non-DQ domains for the industry. For instance [244] used neural network model to forecast stock market and [245] deployed RNN to calculate stock returns. Furthermore, [207] experimented ATTN LSTM to predict stock price. Besides, some centred on financial prediction such as LR of banks with ANN and Bayesian networks [126], bank failure with SVM [66] and bad customers with low CR using LReg and DNN [69]. None of them targets on DQ. This motivates us to widen a horizon to DQP to the industry.

4.4 Summary

This chapter demonstrates how to leverage a ML model to predict DQ accurately in unsupervised learning.

In unsupervised learning, labelling DN is not required. Before DQP, DN are detected from a dataset in Section 4.2.1. Detected DN impacts are estimated in terms of PDF in two generative mixture methods. GMM and BGMM methods are mentioned in Section 4.2.2. The PDF are probabilistic values representing relative impacts of each noise to an attribute. These as weights are aggregated in a scoring function before they are input into networks for prediction. **Fig. 4.2** shows an aggregate quality scoring algorithm guiding on a new DQ scoring method under DG. The input data is trained in LSTM RNNs inside Section 4.2.3. The prediction can help to meet the international requirement, BCBS 239.

In the model, networks trained are LSTM RNNs with three learning methodologies including FF, BK and BD. Their learning algorithms are provided in **Fig. 4.3, 4.4 and 4.5**

correspondingly. They guide on different DQ learning methods under DG. These networks are additionally applied with an ATTN. Thereupon, the model prediction not only take temporal sequences and correlations of DN into account but also the importance of DN, unlike Chapters 3.

Experimental results are remarkable. The network prediction accuracy, PN, RL and F1 are high while the prediction error is low. The network performance is examined at two levels including an integrated level and an individual level.

This model for DQP incontrovertibly brings values to the financial services industry. It creates a method to predict DQ in an unsupervised learning by saving time on labelling data [161] which demands for numerous resources. Time saving is indispensable in the industry where there are massive amounts of data [63]. Other than this, prediction of DQ precisely with LSTM RNNs enable FIs to understand what DQ are going to be with a sceientific computational method. This enhances their forward-looking capabilities of DQ [3] by providing any potential violations of risk limits over thresholds. Hence, DQ can be improved in long term [3] which is consistent with the expectation of financial regulators.

The next chapter presents DQP analytics with a ML model. The analysis is made by multi-dimensions including the dimension of risk types and business segments.

Chapter 5

Data Quality Prediction Analytics

In this chapter, we further explore the analysis of DQP under DG by multi-dimensions with a ML model on top of DQP mentioned in Chapter 3 and 4. This is part of DGP under the DG framework.

Relying on the ML model with sequential learning using supervised learning in Chapter 3, we apply more complicated network learning methods to direct DQ learning in this chapter. Apart from this, the focus of this chapter is on the compliance with a local regulatory requirement, CPG 235, instead of BCBS 239. This tests the applicability of our ML model in complying with other DG regulatory requirements. In order to meet another requirement, we label data differently based on six DQ dimensions. Consequently, the model generates visualized prediction by different dimensions such as risk types and business segments.

In this chapter, Section 5.1 provides an introduction. In Section 5.2, we present a proposed model for DQP analytics. In the model, data is labelled based on six DQ dimensions. Then, the model is implemented with LSTM RNNs, similar to Chapter 3. These networks are further applied with more complex network learning methods to understand how networks learn differently. Following this, we demonstrate performed experiments in Section 5.3. Finally, Section 5.4 concludes this chapter.

5.1 Introduction

From research results summarized in Section 1.2.3 and elaborated in Section 2.2, there are limitations of ML work related to DQP analytics under DG: a) ML techniques have not been applied in DQP analytics under DG to meet DG regulatory requirements. A ML model for analyzing DQP during DG processes has not been proposed; and b) ML techniques have not been applied to DQ learning such as DQP analytics during DG processes for meeting DG regulatory requirements.

To address these, we leverage a ML model to analyze DQP in accordance with a local requirement of CPG 235 [120] in this chapter. This manages data risk [120] during DG processes.

The model data is labelled based on six DQ dimensions in supervised learning before DQP analytics. These dimensions include accuracy, completeness, consistency, timeliness, availability, and fitness for use.

The model is implemented with multiple LSTM RNNs to find the best network for DQP analytics. These networks are applied with more complex learning methods such as windows, time-steps and memory between batches ("MBB"). These networks generate DQP analytical reports by risk types such as MR, CR, OR and LR, and by business segments such as PvB, WB and RB. With DQP analysis, FIs understand what kinds of data are of low quality.

5.2 Proposed Model

DQ are predicted in a ML model with DNN. In our approach, data labelling and scoring methods are the same as that in Chapter 3 except the regulatory requirement to be complied.

5.2.1 Regulatory Requirement CPG 235 Mapping

In this chapter, FIs are expected to meet the regulatory requirement of CPG 235. According to CPG 235, DQ are assessed by six dimensions: (a) accuracy; (b) completeness; (c) consistency; (d) timeliness; (e) availability; and (f) fitness for use [120]. These are illustrated with an example of DN, as defined in **Fig. 5.1**. In total, ten (10) DN [128, 129] are mapped to these dimensions. The fewer the DN, the higher the quality.

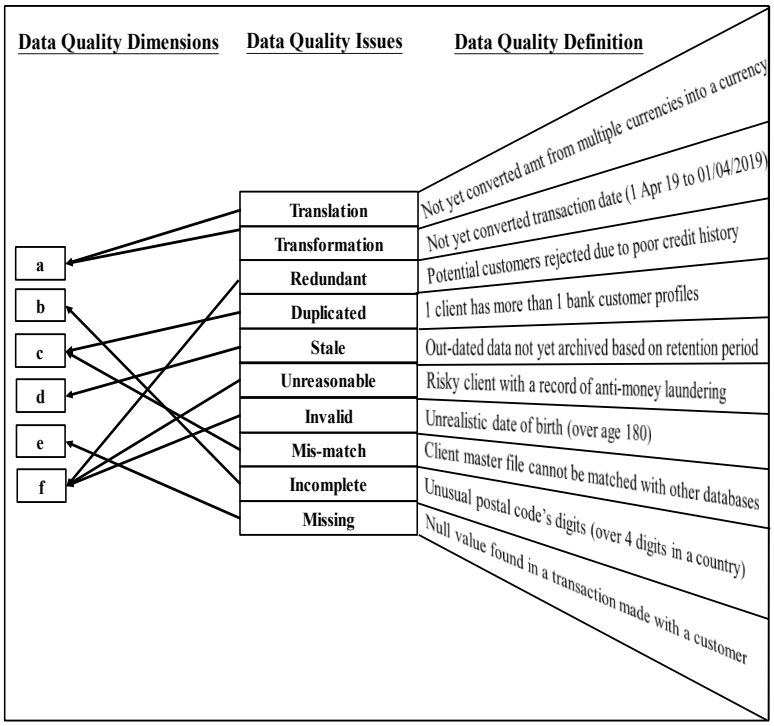


Fig. 5.1 DN (Data Quality Issues) Mapped to CPG 235 DQ Dimensions

5.2.2 Networks with Windows, Timesteps and Memory between Batches

The model consists of four (4) networks. These networks are LSTM RNNs [187] analyzing DQP, as depicted in **Table 5.1**.

Table 5.1 Networks and Relevant Methodologies

Networks	Methodologies
LSTM RNN	Input data (x) is in the form of: samples, time steps, features. There is one sample and feature. Given DN for each DE now (t), we predict the problem for next time ($t+1$). For these data, we prioritize sequences of values and define look_back – the number of previous time steps as input variables to predict next result. In case the number is 1, next result will be $t+1$. Also, we define a layer with 1 input, a hidden layer with four LSTM blocks and 1 output layer. The activation function is Sigmoid and the number of epochs is 10 while the size of batch is 1. After fitting data into the network, we predict DQ based on training and testing data. Then, we test the network for unforeseen data by cross-validation techniques.
LSTM RNN Using Windows	DQ is predicted at next time ($t+1$) by utilizing current time (t) and two recent timesteps ($t-1$ and $t-2$) as inputs. The number of previous timesteps is a window and the size of it is tuned for each problem. By looking back, network error may increase, and so the window size and network architecture will be tuned.
LSTM RNN Using Time Steps	Previous time steps are taken as inputs to predict outputs at next step instead of treating past observations as separate input features. As such, different numbers of timestep are used – from a point of failure or a point of surge. This shows whether the problem is framed accurately or not.

LSTM RNN Using MBB	Utilizing memory to make prediction aids in remembering long sequences. When we fit data into the network, the state of network will be reset after each batch. This allows to manage as to when the internal state of the LSTM network is cleared. As a result, a stateful layer is formed. At the end, the state for the complete sequence is developed. In training the network, no data is reshuffled, and the network state is reset after each epoch. Once the network is built, the stateful parameter is set to true. In setting the batch input shape, we hard-code the number of samples in a batch, the number of timesteps in a sample and the number of features in each time step. Hence, we forecast DN to see whether DQ exceed the risk threshold.

We train these four LSTM RNNs to find the most favorable network. These networks model varying length sequences and capture long range dependencies in DQP analytics.

The model architecture, network equations and learning method are the same as that in Chapter 3.

5.3 Experiments

The purpose of these experiments is to investigate the DQP performance of our model by risk types and business segments with separate experiments over individual datasets. The experiment setup and data split are the same as that in Chapter 4.

Dataset

The dataset is the same as that in Chapter 3. Some data features for four risk types are extracted to **Table 5.2**.

Table 5.2 Data Features (Examples)

MR	CR	OR	LR
Asset Maturity (1945 days, tbc, na)	Loan ID (385623, 0, tbc, na)	Loss Income Ratio (1.15%, 92.04%, 54.6%)	Liquidity Rate (10.39%, 65.29%)
NPV (425543, 0, tbc, na)	Weighted Avg PD (6.31%, 19.48%)	Residual Legal Liability ($1385, 12, 0)	Instrument (TBC, Forward, Equity)

Table 5.2 shows that some data features are embedded with DN such as MR's NPV (0, tbc and na), CR's loan ID (0, tbc or na), OR's legal liability (0) and LR's instrument (TBC).

Results

At first, we predict DQ with an integrated dataset. We train four networks with the algorithm of ADAM to output the prediction accuracy ("Acc") and prediction error ("Loss"), as displayed in **Table 5.3**. ADAM is selected due to its merits mentioned in Section 2.1 of Chapter 2.

The accuracy for the four networks is similar (at a level of 69%) in the 10^{th} epochs but the level is consistently high only for the LSTM RNN using MBB. With regards to the loss, this LSTM RNN using MBB is as good as the LSTM RNN using time steps. The loss for the former is minimized at the 1^{st} and end of the epoch whereas that for the latter reaches a minimal level in last 3 epochs in comparison with others. Both are comparable for DQP.

Table 5.3 Accuracy and Loss for Four LSTM RNNs

LSTM	RNN		RNN using Win		RNN w Time Steps		RNN using MBB	
Epoch	Acc	Loss	Acc	Loss	Acc	Loss	Acc	Loss
1	0.6913	0.6198	0.6877	0.6366	0.6918	0.6191	0.6921	0.6188
2	0.6921	0.6183	0.6921	0.6183	0.6921	0.6183	0.6921	0.6184
3	0.6921	0.6181	0.6921	0.6183	0.6921	0.6182	0.6921	0.6183
4	0.6921	0.6179	0.6921	0.6183	0.6921	0.6181	0.6921	0.6182
5	0.6921	0.6179	0.6921	0.6183	0.6921	0.6181	0.6921	0.6181
6	0.6921	0.6179	0.6921	0.6183	0.6921	0.6179	0.6921	0.6181
7	0.6921	0.6179	0.6921	0.6183	0.6921	0.6179	0.6921	0.6180
8…10	0.6921	0.6179	0.6921	0.6183	0.6921	0.6179	0.6921	0.6179

To confirm the prediction error, we estimate mean square error ("MSE"), as exhibited in **Fig. 5.2**. The lowest MSE is achieved by three RNNs including LSTM RNN, LSTM RNN using time steps and LSTM RNN using MBB. These three RNNs minimize the loss to 0.2133 over 10 epochs.

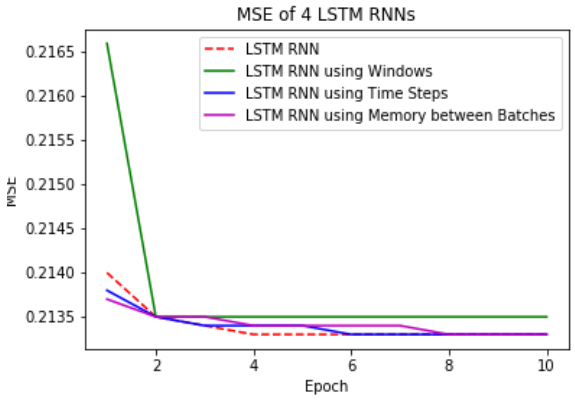


Fig. 5.2 MSE for Four LSTM RNNs

Evaluation

We test the effectiveness of networks and compare their results with actual outputs. The validated accuracy is equal for all networks (69.25%) and validated losses are compared in **Fig. 5.3**. The loss is the lowest in the LSTM RNN using Windows (0.6174). Similarly, the validated MSE is minimized in this RNN (0.2131) out of all networks, as visualized in **Fig. 5.4**. In view of this, the LSTM RNN using Windows is superior to networks. The MSE for others is that: 0.2132 for the LSTM RNN, 0.2133 for the LSTM RNN using time steps, and 0.2134 for the LSTM RNN using MBB.

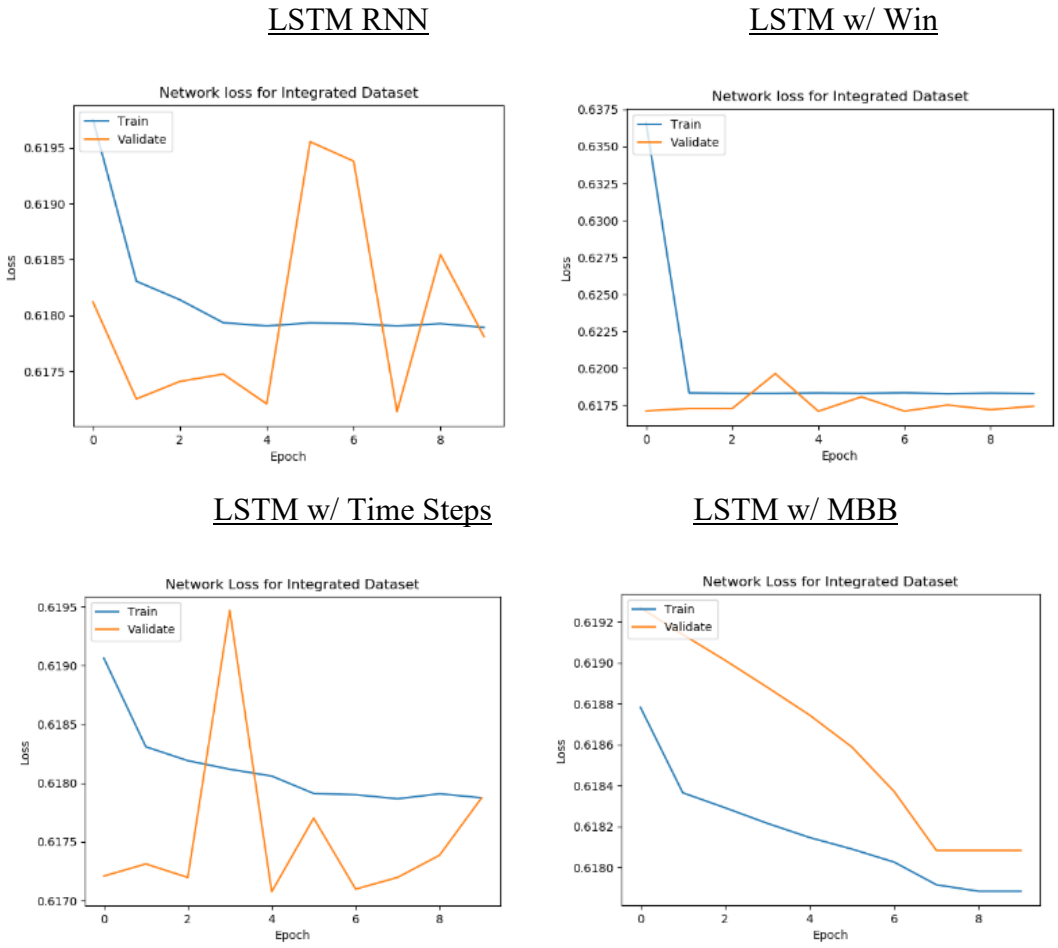


Fig. 5.3 Validated Loss for Four LSTM RNNs

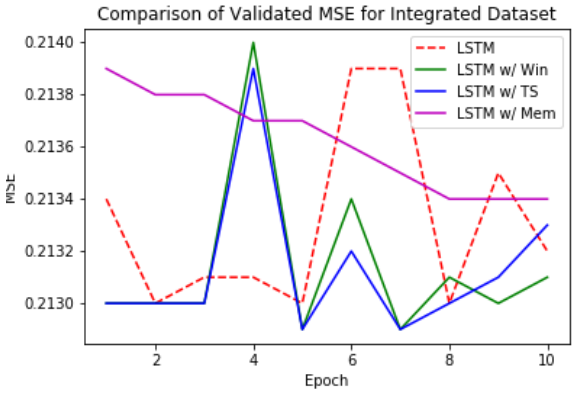


Fig. 5.4 Validated MSE for Four LSTM RNNs

DQP Analytics

For analytics of DQP, we study three cases below.

Case 1 – We select the LSTM RNN using MBB and the LSTM RNN with time steps for a further study as a result of their similar excellent performance. To maximize the accuracy and minimize the loss, we analyze prediction with three more algorithms, as listed in **Tables 5.4** and **5.5**. The LSTM RNN using MBB with ADAGRAD achieves the highest accuracy (69.21%) and the lowest loss (0.6174). In view of this, this LSTM RNN is preferrable.

Table 5.4 Accuracy and Loss of RNNs with Time Steps by Algorithms

RNN	ADAM		RMSPROP		ADADELTA		ADAGRAD	
Epoch	Acc	Loss	Acc	Loss	Acc	Loss	Acc	Loss
1	0.6918	0.6191	0.6919	0.6267	0.6919	0.6218	0.6918	0.6183
2	0.6921	0.6183	0.6921	0.6219	0.6921	0.6222	0.6921	0.6177
3	0.6921	0.6182	0.6921	0.6217	0.6921	0.6228	0.6921	0.6176
4	0.6921	0.6181	0.6921	0.6216	0.6921	0.6221	0.6921	0.6176
5	0.6921	0.6181	0.6921	0.6216	0.6921	0.6221	0.6921	0.6175
6	0.6921	0.6179	0.6921	0.6216	0.6921	0.6221	0.6921	0.6175
7	0.6921	0.6179	0.6921	0.6217	0.6921	0.6221	0.6921	0.6175
8	0.6921	0.6179	0.6921	0.6216	0.6921	0.6213	0.6921	0.6175
9	0.6921	0.6179	0.6921	0.6216	0.6921	0.6214	0.6921	0.6175
10	0.6921	0.6179	0.6921	0.6216	0.6921	0.6215	0.6921	0.6175

Table 5.5 Accuracy and Loss of RNNs using MBB by Algorithms

RNN	ADAM		RMSPROP		ADADELTA		ADAGRAD	
Epoch	Acc	Loss	Acc	Loss	Acc	Loss	Acc	Loss
1	0.6921	0.6188	0.6921	2.8424	0.6921	0.6213	0.6921	0.6178
2	0.6921	0.6184	0.6921	4.9079	0.6921	3.9912	0.6921	0.6175
3	0.6921	0.6183	0.6921	4.9079	0.6921	4.9079	0.6921	0.6175
4	0.6921	0.6182	0.6921	4.9079	0.6921	4.9079	0.6921	0.6175
5…6	0.6921	0.6181	0.6921	4.9079	0.6921	4.9079	0.6921	0.6174
7	0.6921	0.6180	0.6921	4.9079	0.6921	4.9079	0.6921	0.6174
8…10	0.6921	0.6179	0.6921	4.9079	0.6921	4.9079	0.6921	0.6174

Besides, we examine the MSE of these two LSTM RNNs, as visualized in **Fig. 5.5**. The loss is minimized in ADAGRAD (0.2131) for both RNNs when compared with other networks: a) the LSTM RNN with time steps applying ADAM (0.2133), RMSPROP (0.2147) or ADADELTA (0.2147); and b) the LSTM RNN with MBB applying ADAM (0.2133), RMSPROP (0.4105) or ADADELTA (0.3369).

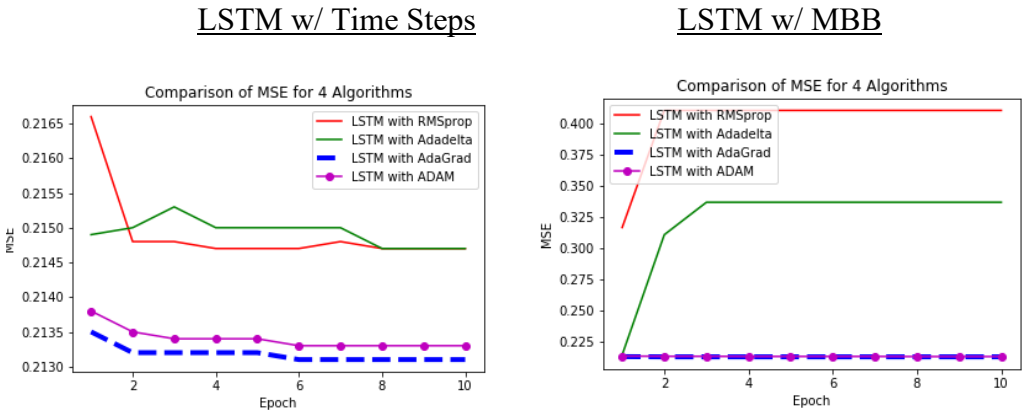


Fig. 5.5 MSE of Two LSTM RNNs under Four Algorithms

Case 2 – For DQP analysis by risk types, we train four databases separately including MR, CR, OR and LR. Then, we input data into LSTM RNNs using MBB and applying the algorithm of ADAGRAD for DQP. Prediction outcomes are visualized in **Fig. 5.6**. This figure shows that different data characteristics lead to various predictive powers under the same algorithm. Relevant data features are described in **Table 3.2, 3.3** and **3.4**.

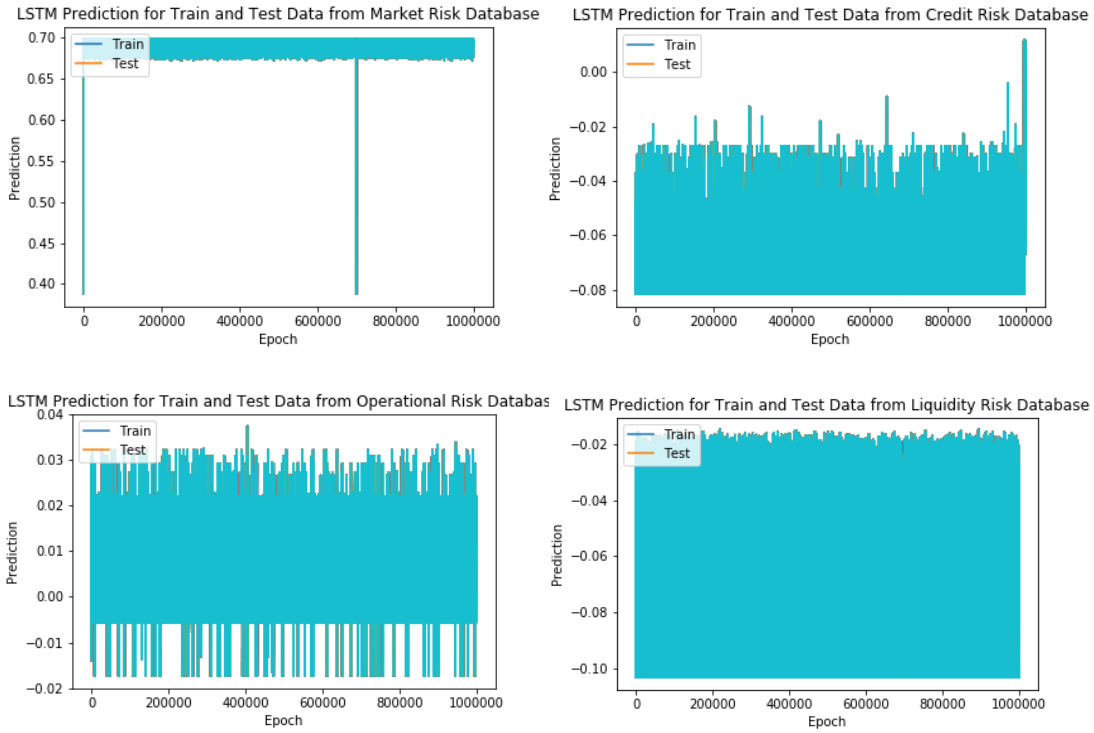


Fig. 5.6 Prediction of LSTM RNNs for Four Risk Types

Applying ADAGRAD to networks, we estimate the precision ("PN"), recall ("RL") and F1-Support ("F1"), as listed in **Table 5.6**. Results show that the PN for CR and OR is high (96% and 99% respectively) but the lowest one belongs to MR (9%). LR is in between (58%). The RL and F1 outcomes are similar. Thus, different data characteristics lead to divergent precision and recall rates under the same algorithm. Data features are described in **Table 3.2, 3.3** and **3.4**.

Table 5.6 PN, RL and F1 of LSTM RNNs for Four Risks

LSTM RNN	MR	CR	OR	LR
PN/ RL/ F1	0.09/ 0.31/ 0.51	0.96/ 0.98/ 0.97	0.99/ 0.99/ 0.99	0.58/ 0.76/ 0.66

Additionally, we measure the prediction error in terms of MSE, as visualized in **Fig. 5.7**. Consistently, the MSE for the OR (0.0368) and CR (0.0448) is the lowest whereas the highest MSE occurs in the MR (0.2133).

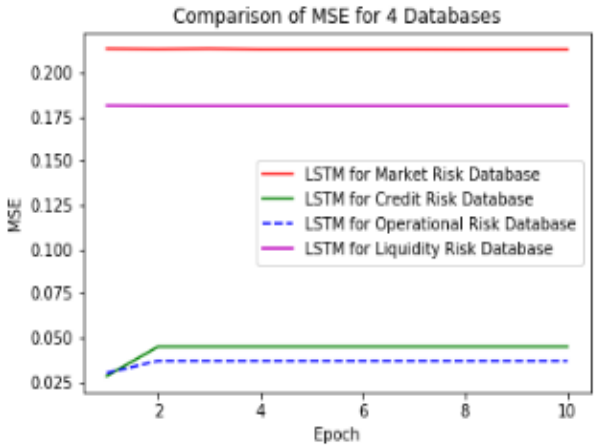


Fig. 5.7 MSE of LSTM RNNs for Four Risk Types

Case 3 – For the analytics by business segments, we leverage the RB database as an example for the demonstration purpose. Prediction are made in LSTM RNNs with MBB applying the algorithm of ADAGRAD, as revealed in **Fig. 5.8**. By comparing databases, we note that the prediction outcomes focus on a limited range for the MR (0.33-0.365), OR (0.265-0.30) and LR (0.29-0.325) but they cover a wider range for the OR (0.16-0.20) due to more DN as inputs generated by Python. The range of output sequence length for MR is the shortest as a result of fewer DQ issues. Converting visualized prediction into DQ metrics, we note that the percentage of aggregated DQ for MR ranges from 32% to 36%.

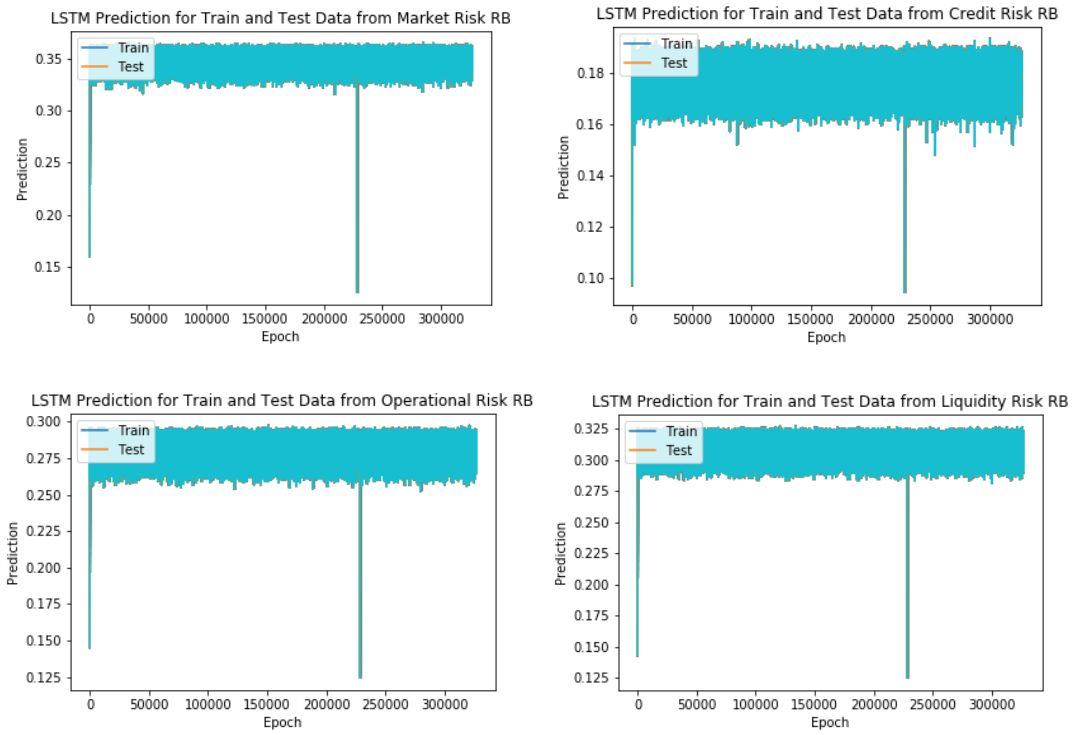


Fig. 5.8 Prediction of the RB Segment for Four Databases

To verify the prediction error, we test LSTM RNNs by cross-validation. The validated loss results are shown in **Fig. 5.9**. The validated loss for the OR (0.4693) and OR (0.5939) is the lowest, similar to the MSE in **Fig. 5.10**. The lowest occurs in the CR (0.0177) and OR (0.0512).

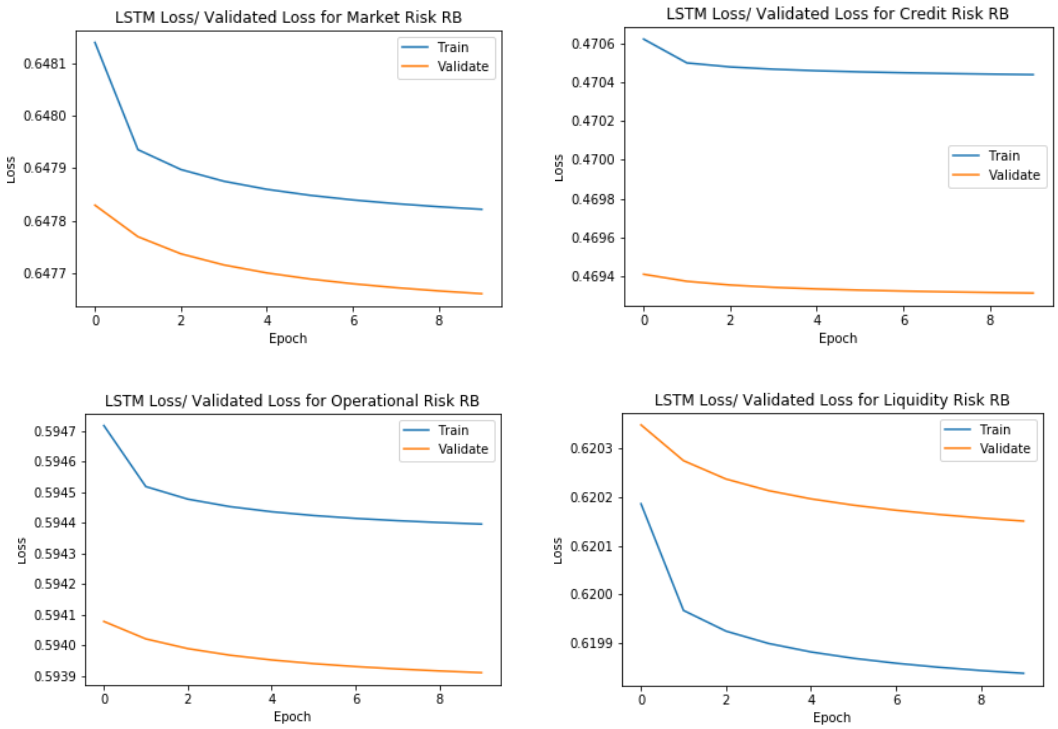


Fig. 5.9 Loss and Validated Loss of the RB Segment by Databases

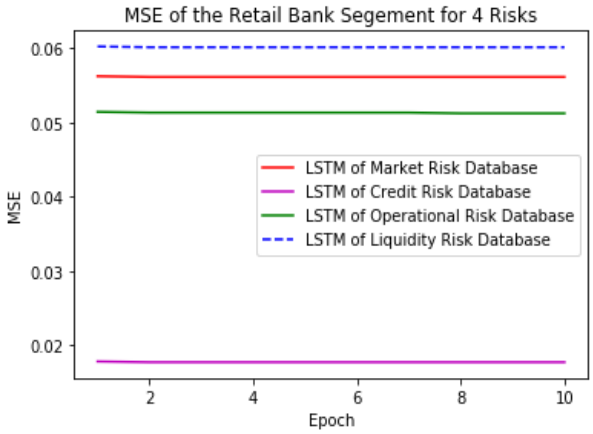


Fig. 5.10 MSE of the RB Segment for Four Databases

Above performance measurements such as accuracy, loss, precision, recall, F1-Support are defined in equations 2.37 to 2.41 inside Chapter 2.

As we can see, our experiments over individual datasets show the DQP performance by risk types and business segments in separate experiments.

Related Works

To the best of our knowledge, there was no previous work on ML to predict DQ for the compliance with a local requirement, CPG 235. Analytics of DQP were limited. Related works are that: a) [88] leveraged MLP and Bayesian Networks to measure and predict LR respectively. Error rate was low (8.0e-3 for GA and 1.7e-10 for LMA) while the RMSE was < 0.2.

Instead, our model predicts DQ of 4 risks; b) [63] used ML to predict a bank credit with 23 features achieving an accuracy of 80%. Instead, we measure DQ with 132 features; c) [196] classified credit with logistic regression and SVM. The accuracy was 75% but reduced to 43.5% for critical region, unlike our experimental results; and d) [195] analyzed flood risk with an AHP method. The importance has been defined and the hazard has been divided into 5 risks, similar to the data criticality and data quality ranking in our model.

5.4 Summary

In this chapter, we propose a model to analyze DQP based on a local regulatory requirement, CPG 235. In supervised learning, data is labelled according to six DQ dimensions including accuracy, completeness, consistency, timeliness, availability, and fitness for use, as illustrated in Section 5.2.1.

With labelled data, we input the labelled data into LSTM RNNs for analytics of DQP. To understand how networks learn differently, we apply complex learning methods to these networks such as windows, time-steps and MBB in Section 5.2.2. These networks are tested by divergent algorithms and evaluated by cross-validation techniques to confirm the model effectiveness.

We also implement the model with a set of experiments in Section 5.3. The overall experiment results show that our model is effective in DQP analytics under supervised

learning. The model accurately predicts DQ and at the same time analyzes DQ at a granular level by risk types (such as MR, CR, OR and LR) and by business segments (including PvB, WB and RB).

Performing multi-dimensional analytics is beneficial to the financial services industry. With granular analytical reports, FIs understand which types of data should be prioritized for data risk management [120]. In addition, analytics of DQP in alignment with the regulatory requirement aids in meeting the expectation of financial regulators.

The next chapter presents our attempt to improve the network efficiency for DQP, where we focus on the network run-time saving for the improvement purpose.

Chapter 6

Network Efficiency Improvement in Data Quality Prediction

In this chapter, we further explore the network run-time saving of DQP under DG on top of DQP and DQP analytics. Since Chapter 5 model accurately predicts DQ and at the same time analyzes DQ by divergent dimensions. The model in Chapter 5 is further improved in this chapter in terms of the network efficiency improvement, as part of DGP under the DG framework.

For improving the network efficiency, we propose a ML model to profile different portions of data from a dataset and test them in networks for DQP. In the model implementation, we adopt the most efficient network in Chapter 5 for this chapter. They are LSTM RNNs with MBB predicting DQ in accordance with a local requirement, CPG 235. Section 6.1 provides an introduction of this chapter. In Section 6.2, we discuss how to perform data profiling systemically and how DQP are learnt in LSTM RNNs with MBB. We demonstrate the performed experiments in Section 6.3. Finally, Section 6.4 concludes this chapter.

6.1 Introduction

From research results summarized in Section 1.2.3 and elaborated in Section 2.2, there are limitations of ML work related to the network efficiency improvement in DQP under DG: a) ML techniques have not been applied in the network efficiency improvement in DQP. DQP network run-time saving has not been measured and a ML model for the network efficiency improvement in DQP during DG processes has not been proposed; b) ML techniques have not been applied to DQ learning such as DQP during DG processes for

meeting DG regulatory requirements.

To address these, we improve the network efficiency of DQP with a ML model in this chapter. This manages bank-wide risk [4] during DG processes.

In the financial services industry, data is stored in multiple repositories. To analyze and predict them takes tremendous amount of time [186]. Facing this, we propose a ML model to perform data profiling for network learning by extracting different portions of data from a dataset. Profiled data is imported into networks for DQP. In the prediction process, we measure network run-time saving.

In this model, how to justify data as good or bad is referenced to six DQ dimensions as set out in a local requirement, CPG 235 [120], similar to that in Chapter 5.

The model is implemented with LSTM RNNs, similar to that in Chapter 5. These networks with MBB are applied with the algorithm of ADAGRAD to find the most efficient network for DQP while maintaining accurate DQP.

6.2 Proposed Model

DQ are predicted in a ML model with DNN. In our approach, data labelling and scoring methods are the same as that in Chapter 5. In this chapter, DN are mapped to the regulatory requirement, CPG 235. After the mapping, the model is implemented by networks with profiled data.

6.2.1 Data Profiling

In performing data profiling, we select data clusters for testing. In total, four databases are profiled.

1st Database - the data element ("DE") of "Asset Amount" in MR database is selected.

- Asset Amount (in number) is classified into 4 categories (1, 2, 3 and 4 corresponding to the amount of <85,000, between 85,000 & 385,000, over 385,000 & all amounts). Out of these categories, the number of records is around 10%, 30%, 60% and 100% of the database correspondingly.

Besides, the DE of "Nationality" is selected.

- Nationality (in several options) is classified into 4 categories: 1, 2, 3 and 4 corresponding to a group of countries (CAD, SGD & EUR), another group of countries (AUD, CAD, CNY, CZK, EUR, GBP, HKD, JPY, MYR, NZD & SGD), a group of countries (excluding category 1 & 2) and all countries. Out of these, the number of records is 9%, 34%, 65% and 100% of the database respectively.

2nd Database - the DE of "CollateralAmt" in CR database is selected.

- CollateralAmt (in amount) is classified into 4 categories (1, 2, 3 and 4 corresponding to the amount of <85,000, between 85,000 & 385,000, over 385,000 and all amounts). Out of these categories, the number of records corresponds to 10%, 30%, 60% and 100% of the database.

3rd Database - the DE of "EventDate" in OR database is chosen.

- EventDate (in date format of mm.dd.yyyy) is classified into 4 categories (1, 2, 3 and 4 corresponding to the dates later than 19 Feb 2017, the dates later than 19 Feb 2015, the dates later than 19 Feb 2012 and all dates). Out of these categories, the number of records is 16%, 36%, 66% and 100% of the database respectively.

4th Database - the DE of "SettlementDate" in LR database is chosen.

- SettlementDate (in date format of DD/MM/YYYY) is classified into 4 categories (1, 2, 3 and 4 corresponding to aging period: between 1 & 2 years, 1 & 4 years, 1 & 7 years

and all years). Out of these categories, the number of records is 10%, 30%, 60% and 100% of the database correspondingly.

These four databases are consolidated into an integrated dataset. After performing data profiling, we use the least percentage of data from the entire database (either 9%, 10% or 16%) as a base for comparison of the network run-time saving.

6.2.2 LSTM Networks with Memory between Batches

In the implementation of the model, we input profiled data into LSTM RNNs with MBB. These network equations, learning method and network setups are the same as that in Chapter 5.

In LSTM networks, the MBB enables networks to remember the content of previous batches: the last state for each sample in a batch is used as an initial state for the same in the next batch. This propagates previous states for each sample across batches. Thereupon, the data feature is standardized as:

$$\hat{x}_k = \frac{x_k - \bar{x}_k}{\sqrt{\delta_k^2 + \epsilon}} \tag{6.1}$$

where ϵ is a positive constant to improve the numerical stability. This feature standardization is a procedure that can be used to reduce convergence rates.

The optimization of batch is normalization (*BN*) introducing learnable parameters γ and β to scale and shift data correspondingly resulting in the form of:

$$BN(x_k) = \gamma_k \hat{x}_k + \beta_k \tag{6.2}$$

Setting $\gamma_k$to σ_k and $\beta_k$to $\bar{x}_k$, networks recover its initial layer. For this, the layer in the network becomes:

$$y = \phi\ (W_x + b) \tag{6.3}$$

where W is weights matrix, b is bias vector, x is input and ϕ is an activation function. Hence, the batch normalization is:

$$y = \phi\ (BN(W_x)) \tag{6.4}$$

Given the standardization, the effect of bias vector is cancelled. By normalization, the backpropagation needs to be adapted to propagate gradients.

6.3 Experiments

In experiments, we direct sequence learning in LSTM RNNs without MBB first. Upon completion, we compare these network results with the outcomes of LSTM RNNs with MBB. The experiment setup and data split are the same as that in Chapter 5.

The purpose of these experiments is to achieve convincing network runtime saving in DQP with our model by running experiments with various DEs on a dataset.

Dataset

The dataset is from Chapter 3. It contains four risk types including MR, CR, OR and LR. Some data features with DN are extracted to the following (with examples).

a. MR – Asset Amount (251527, na, ‘ ’, 838); Nationality (‘ ’, tbc, JPY, GBP, AUD); MR Segments (Retail Bank, Private Bank, Wholesale Bank); Customer Risk Rating (H, M, L, OnBoarding, P, Q);

b. CR – Collateral Amount (29397, 6727, tbc, ‘ ’); TimeStamp (3/10/2015 6:09, 6/02/2008 3:15, 16/06/2018 4:34); Guarantor ID Number (123272, 32416, tbc); Product Price (725, 85, 3089, na);

c. OR – Event Date (06.25.2009, 10.05.2012, 1.29.2018); Residual Legal Liability (1385, 12, 3307, 715); Control Factors (system control, na, others, regular review); Loss Multiplier (1, ‘ ’, 1.4, 2); and

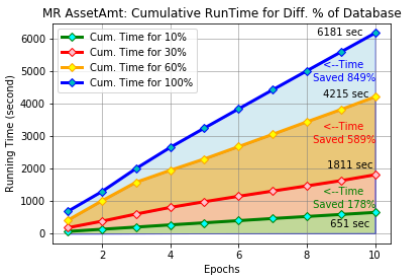
d. LR – Settlement Date (4/11/2014, 13/12/2009, 19/08/2018); NAV (tbc, 871942, 17914, ‘ ’); Liquidity Rate (0.1039, 0.5103, 0.9975); Number of Trades (1685, 13, na, ‘’).

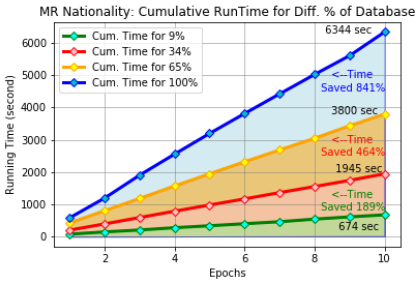
These data features with DN are assigned with quality scores before DN are classified into quality rankings, as inputs into networks for DQP.

Results

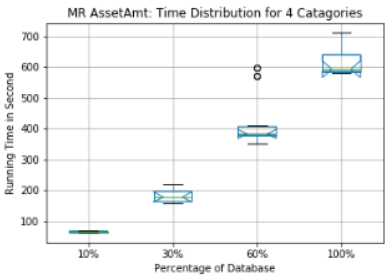
Utilizing the algorithm of ADAGRAD, we select DEs from MR to train LSTM RNNN without MBB with the dataset of 10%, 30%, 60% and 100% of the entire database. DEs are asset amount and nationality. Testing results are shown in **Fig. 6.1** and **Fig. 6.2**.

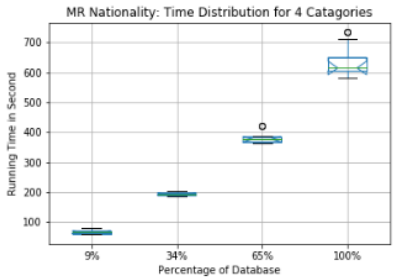
(a) Cumulative Runtime for Diff. % of Database with ADAGRAD



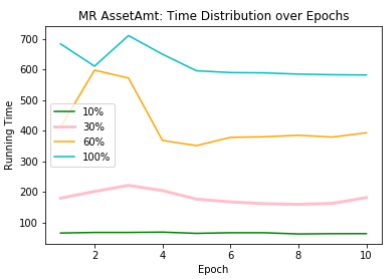


(b) Runtime Range with ADAGRAD





(c) Runtime Trend with ADAGRAD



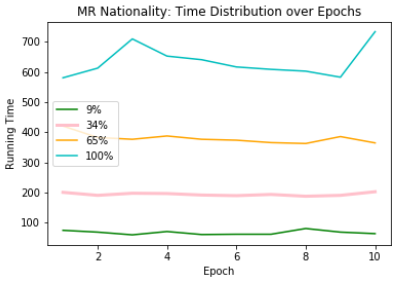


Fig. 6.1 Runtime - MR Asset Amount (ADAGRAD)

Fig. 6.2 Runtime - MR Nationality (ADAGRAD)

a. Regarding the DE of MR asset amount in **Fig. 6.1**, much run time is saved in the network with a dataset which is 10%, 30%, 60% of the entire database. Using 10% of the network running time as a base, we notice that the percentage of time saving is 178%, 589% and 849% for the dataset having 30%, 60% and 100% of the data from entire database. It is attributable to the total runtime of 1811, 4215 and 6181 seconds ("sec") respectively. By selecting a small fraction of data for prediction (by 10%), we can reduce a huge amount of network run time. It simply requires 651 sec.

b. Inside the network, the run time on average is stable over 10 epochs for the dataset with 10% and 30% of the entire database (over 65 sec and 181 sec) except that with 60% and 100%. The run time for the dataset with 60% of data from entire database rises to a high level (598 sec) at 2nd epochs but returns to a normal level till the end of epoch (393 sec) whereas that for the dataset having 100% of the entire database is unstable – dropping at 2nd epochs (611 sec) and bouncing back to a high level at 3rd epochs (711 sec) before decreasing at an average level (600 sec) at the 4th epoch.

c. The average runtime results are visualized in **Fig. 6.1**. They are for the dataset sourcing 10%, 30%, 60% and 100% of the data from the entire database. The widest range of run time lies in 100% dataset sourcing all data from the entire database.

d. Concerning the DE of nationality in **Fig. 6.2**, we achieve a similar result. Significant amount of runtime is saved with the dataset of 9% data sourcing from the entire database. The network runtime is limited to 674 sec. This is also saved for the dataset having 34% and 65% of the database – only 1945 and 3800 sec in comparison with the lengthy time (by 6344 sec) required for the entire database. Using 9% of the network running time as a base, we note that saving corresponds to 189%, 464% and 841% for the dataset containing 34%, 65% and 100% of the database.

e. In this network, the run time is stable for datasets with 9%, 34% and 65% of the entire database (by 67, 195 and 380 sec respectively) except that with the dataset having 100% of the data.

f. We can also see that different data characteristics lead to various network saving under the same algorithm. The data features are described in **Table 3.2** and **3.3**.

Test Scenarios

To check if the network performance can be further improved, we conduct additional tests. Scenario 1 - We train the network with another algorithm, SGD, due to its merit [235]. SGD can be used to estimate the probability of output based on a randomly selected subset of inputs with the stochastic approximation of gradient descent optimization. The output can

be estimated using least squares where the objective function is minimized. The run time for DE of MR asset amount is depicted in **Fig. 6.3** while that for the DE of MR nationality is made in **Fig. 6.4**.

Cumulative Runtime for Diff. % of Database
in Network without MBB Applying SGD

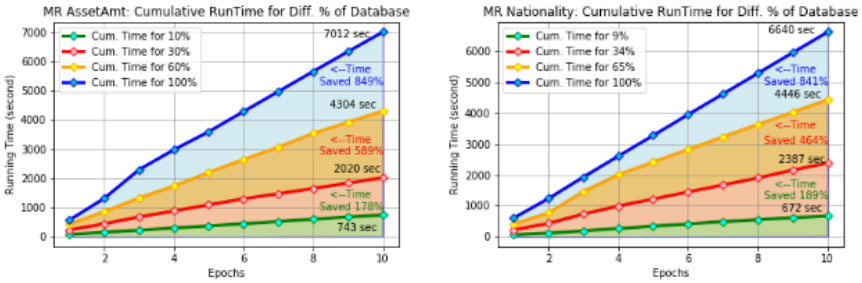


Fig. 6.3 Runtime - MR Asset Amount (SGD) Fig. 6.4 Runtime - MR Nationality (SGD)

Consistently, much time is saved by applying the algorithm of SGD. The run time for the dataset with 10%, 30% and 60% of the entire database corresponds to 743, 2020 and 4304 sec when compared with the total time of 7012 sec. In view of this, the percentage of extra run time is 178%, 589% and 849% for the dataset of 30%, 60% and 100% of the data from the entire database respectively assuming the runtime of the dataset with 10% of the entire database is used as a base. Additionally, this situation is the same as that of the DE of nationality. The variance is the percentage of extra time – 189%, 464% and 841% for the dataset sourcing 34%, 65% and 100% of the data from the entire database.

Scenario 2 – We compare the initial LSTM RNN with another network, the LSTM RNN with MBB. To test it, we select the DE of asset amount from MR to apply different algorithms (ADAGRAD and SGD) to this network, as given in **Fig. 6.5** and **Fig. 6.6**.

Cumulative Runtime for Diff. % of Database in Network with MBB

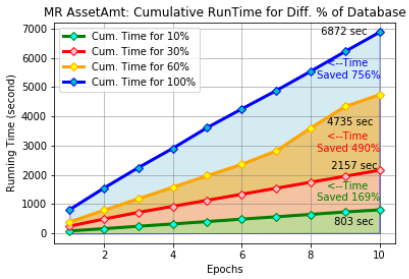


Fig. 6.5 Runtime for MR Asset Amount in Network with MBB (ADAGRAD)

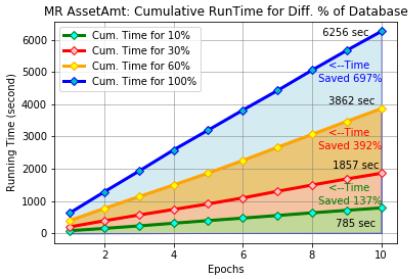
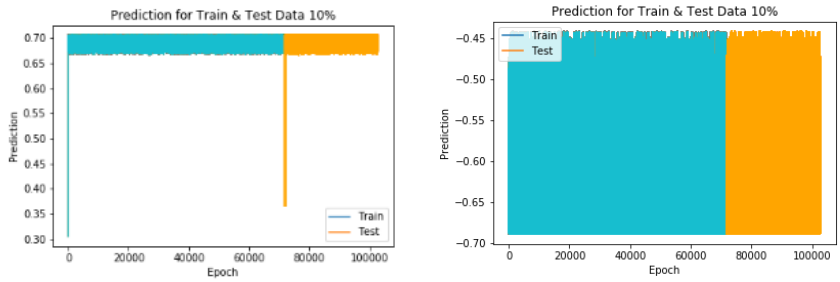


Fig. 6.6 Runtime for MR Asset Amount in Network with MBB (SGD)

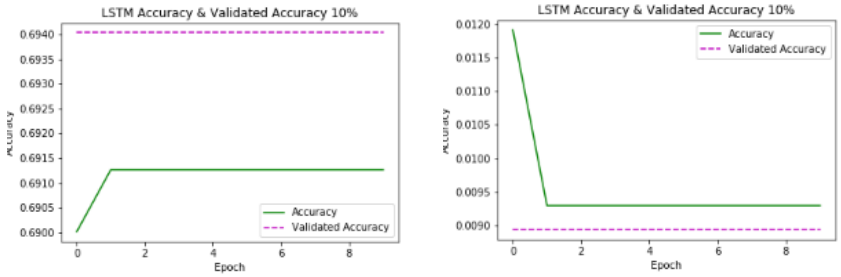
To train the LSTM RNN with MBB with another algorithm, ADAGRAD, we note that the saved time is 169%, 490% and 756% for the dataset having 30%, 60% and 100% of the entire database correspondingly. When compared with the initial network, we observe that the saving is similar. By applying SGD, we find that the saving is explicit - 137%, 392% and 697%. Then, we check prediction outcomes, prediction accuracy and prediction error (in terms of a loss) of the network with MBB.

Scenario 3 – Selecting a dataset with 10% of data for testing, we notice that the accuracy is high (69.40%) and loss is low (0.616) with respect to the network applying ADAGRAD, as shown in **Fig. 6.7**. For the network with MBB applying SGD, the accuracy is highly low (0.89%) whereas the loss is high (11.187), as indicated in **Fig. 6.8**. As a result, the algorithm of ADAGRAD is preferrable.

(a) Prediction in Network with MBB



(b) Accuracy & Validated Accuracy



(c) Loss & Validated Loss

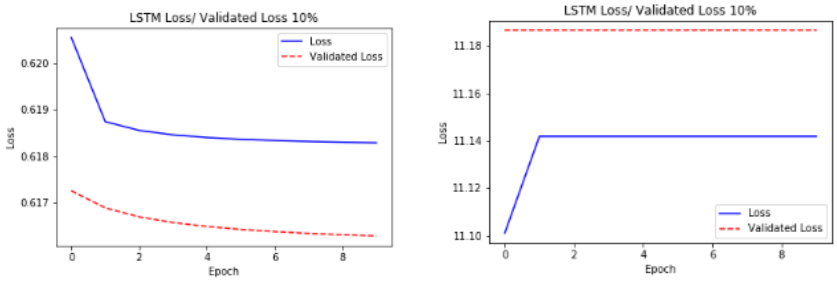


Fig. 6.7 Network Performance (ADAGRAD) Fig. 6.8 Network Performance (SGD)

Above performance measurements such as accuracy and loss are defined in equations 2.37 and 2.38 respectively inside Chapter 2.

As we can see, our experiments achieve convincing network runtime saving in DQP with our model by running experiments with various DEs on a dataset.

Related Works

DQP. [201] studied how ML enhanced the network performance in terms of the predictive/ classification accuracy. Our experiments demonstrate how to drive the efficiency improvement with a DNN.

Network Architecture. [201] applied a deep architecture model using auto-encoders to represent traffic flow features for prediction. Our model is a LSTM RNN modelling long-term temporal dependencies and remembering memory across long sequences. It successfully discovered latent traffic flow feature representation but ours find out a superior architecture (with ADAGRAD algorithm) for the quality prediction.

Network algorithm. [202] innovated a shallow neural network model to detect colon cancer but our model compares heterogeneous algorithms (ADAGRAD and SGD) in LSTM RNNs to rapidly predict DQ. Both achieved excellent results.

Performance Measurement. [197] utilized metrics for its networks aiming at the consumption prediction – training speed and accuracy of networks including Support Vector Regression (SVR), local SVR & H2O deep learning. The measurement is equivalent to us, but ours are LSTM RNN and LSTM RNN with MBB.

DQ Dimensions. [200] measured the quality on a large dataset in terms of the accuracy, completeness and consistency. Ours include these as well as other dimensions sch as translation, transformation, redundancy, duplication, obsolescence, reasonableness and validity. Both compared the network with various algorithms.

DQ Score Calculation. [123] calculated the accuracy by dividing the number of correct values from the number of observations based on the ISO 25012 standard 11 efficient risk data learning while our calculation computes data scores under a scientific method – taking the risk of quality issues into account after alignment with CPG 235.

Network Performance Evaluation. [198] back-tested a strategy to assess simulated trades. This does not deviate from us – utilizing cross-validations to check the accuracy and loss

of network prediction. Both demonstrated the success of processing a cluster of the big data for a prediction within a reasonable time of few hours.

6.4 Summary

In this chapter, we present a ML model to improve the network efficiency in DQP. Before the network learning, we adopt a systematic approach to profile input data, as described in Section 6.2.1.

Given an accurate prediction of DQ in Chapter 5, we leverage the same network, LSTM RNNs with MBB, for this chapter. Section 6.2.2 depicts how MBB is leant in networks. In this chapter, we additionally apply various algorithms to test networks with varying fractions of data. The network saving time is measured in various test scenarios inside our experiments. Consequently, we identify the most efficient network for DQP.

The experiment results demonstrate a significant improvement in the network efficiency for DQP in terms of the network run-time saving while maintaining a high prediction accuracy.

Saving network run-time to predict DQ is helpful for FIs relying heavily on enormous sets of data for analytics. This enables FIs to identify poor DQ earlier for the bank-wide risk management [4].

The next chapter presents IS compliance prediction with a ML model. The prediction focus is IS compliance levels based on ISCs under the Life Cycle as set out in CPG 234.

Chapter 7

Information Security Compliance Prediction

In previous chapters, we predict DQ as part of DGP under the DGF. In this chapter, another DG initiative as part of DGF to meet DG objectives is to predict IS compliance levels during DG processes. This tests the applicability of our ML model in complying with non-DQ regulatory requirements.

Similar to Chapter 3, we make prediction with a ML model applying sequential learning in supervised learning. In this chapter, we extend the ML model in Chapter 3 by applying an ATTN for prediction. This not only captures temporal sequences and correlations of IS factors but also the importance of them.

Our model starts with a compliance approach development followed by defining information security rules ("ISR") for the determination of IS scores. Afterwards, the scores are aggregated into a scoring function for ranking ISL. Following this, the ISL are input into networks for prediction. Section 7.1 provides an introduction of this chapter. The proposed model with sequential learning and an ATTN is explained in Section 7.2. Section 7.3 reports the utilized data, the performed experiments, and the achieved results. Finally, Section 7.4 summarizes this chapter.

7.1 Introduction

From research results summarized in Section 1.2.3 and elaborated in Section 2.2, there are limitations of ML work related to IS compliance prediction under DG: a) ML techniques have not been applied in the prediction of ISL and IS compliance levels under DG. ISL during DG processes have not been predicted. A ML model considering sequences, correlations and importance of IS factors collectively for IS compliance prediction has not been proposed; and b) ML techniques have not been applied to IS learning during DG

processes for meeting DG regulatory requirements.

To address these, we leverage a ML model to predict IS compliance levels according to a local requirement, CPG 234 [75] in this chapter. This enhances IS during [75] DG processes.

CPG 234 sets out a guideline on the management of IS by implementing thirteen controls, called ISCs. These ISCs influence ISL over systems, networks and information assets (collectively named as "Systems"). FIs are obliged to consider these controls inside the Life Cycle to guard against any cyber-attacks and IS threats or vulnerabilities.

To assist FIs in making preparation for the compliance of CPG 234, we propose a compliance approach before presenting a ML model in supervised learning to predict IS compliance levels in a real time mode by automating the IS compliance process in LSTM RNNs.

The model networks are the same as that in Chapter 3 – LSTM RNNs with sequential learning. In this chapter, networks are additionally applied with an ATTN. Consequently, they take the importance of IS scores into account during IS compliance prediction on top of the temporal sequences and correlations of IS scores. In experiments, they are trained to generate compliance reports for an analysis of IS scores by thirteen ISCs.

7.2 Proposed Model

IS compliance levels are predicted in a ML model with DNN. Before the prediction, we develop a compliance approach to evaluate ISL by leveraging ISCs which are dependent on ISR.

7.2.1 Compliance Approach

In order to meet CPG 234, we make reference to the industry best practice [183, 185] in the approach development. The approach is illustrated in **Fig. 7.1**.

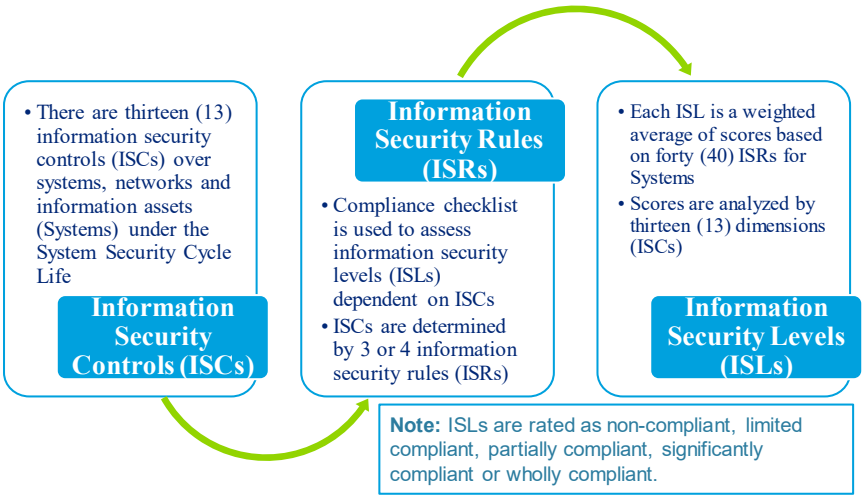


Fig. 7.1 Compliance Approach

This approach is to leverage ISCs for evaluating ISL in Systems while ISCs are determined by multiple ISR, defined as *ISR* in equations (below). Accordingly, we design a compliance checklist to cover these *ISR*. All rules are applied to six systems which are tested in the model.

7.2.2 Information Security Rules

In total, Forty *ISR* are defined. They are presented in the form of questions in a compliance checklist, as listed in **Table 7.1**. These rules are categorized by thirteen controls of the Life Cycle.

Table 7.1 ISR Checklist

	ISR_1 to ISR_3 Under Process 1 – Change Management
1.1	Are changes to systems reviewed for the changes in risk profiles? [254]
1.2	Is data kept confidential and private in new e-banking technologies? [255]
1.3	Is the data model adjusted to the changed system? [180]
	ISR_4 to ISR_6 Under Process 2 – Configuration Management
2.1	How well is system configured to protect against vulnerability? [256]

2.2	Are uses restricted from accessing server configuration files to avoid directory traversal attacks? [257]
2.3	Is virtualization focused configuration management tool deployed? [258]
	ISR_7 to ISR_9 Under Process 3 – Deployment and Environment Management
3.1	Is banking software development segregated from software testing? [259]
3.2	Is cloud computing in bank systems segregated by logical storage? [260]
3.3	Is a regular review performed to confirm who manages and administers data, and controls to detect and react to security breaches? [261]
	ISR_{10} to ISR_{12} Under Process 4 – Access Management Controls
4.1	Is system access assigned based on user roles which are constantly updated? [262]
4.2	Are access controls implemented in banking biometrics systems? [263]
4.3	Are access controls for outsourced vendors defined in SLA? [264]
	ISR_{13} to ISR_{15} Under Process 5 – Hardware & Software Asset Controls
5.1	Is an external machine authenticated and authorized based on cyber banking security protocols and standards? [265]
5.2	Does IP packet filtering protect networks against intruder attacks? [266]
5.3	Are firewalls configured to protect against unauthorized access? [266]
	ISR_{16} to ISR_{18} Under Process 6 – Network Design
6.1	Is data encrypted to prevent hacker sniffing e-banking networks? [267]
6.2	Is penetration testing used for identifying network vulnerabilities? [268]
6.3	Are networks configured to guard against physical attack and unauthorized network intrusion? [254]
	ISR_{19} to ISR_{21} Under Process 7 – Vulnerability Management Controls
7.1	Are malware protection technologies deployed to protect systems (e.g. encryption of code, polymorphism or obfuscation)? [269]
7.2	Is an intrusion prevention system used to analyze traffic control? [270]
7.3	Are web applications scanned to identify vulnerable instances? [271]
	ISR_{22} to ISR_{24} Under Process 8 – Patch Management Controls
8.1	Are security patches updated regularly for online banking? [254]

8.2	Does patch management include the collection of the latest patches and the management of post-patch conflicts? [272]
8.3	Are event logs reviewed to confirm the latest patches applied? [273]
	ISR_{25} to ISR_{27} Under Process 9 – Service Level Management (SLA)
9.1	Are SLAs with each infrastructure provider customized? [229]
9.2	Can vendors quickly reallocate computing resources without any downtime based on the SLA? [230]
9.3	Are metrics used to measure the service level of vendors? [231]
	ISR_{28} to ISR_{30} Under Process 10 – Monitoring Controls
10.1	Are transactions in online banking systems monitored to detect fraud patterns with artificial intelligence or transaction history analysis? [232]
10.2	Are network traffic for mobile banking apps monitored to inspect deep packets and their flow for vulnerability assessment? [233]
10.3	Are removable device and email system vulnerability detected by security monitoring systems? [27]
	ISR_{31} to ISR_{33} Under Process 11 – Response Controls
11.1	Are cyber security incidents detected, prevented and responded by a computer emergency response team? [35]
11.2	Are cyber-attacks well communicated and documented? [60]
11.3	Are cyber security incidents reported and forecasted? [61]
	ISR_{34} to ISR_{36} Under Process 12 – Capacity and Performance Management Controls
12.1	Is system service capacity optimized? [62]
12.2	Is network and web performance tracked to manage system events? [64]
12.3	Is controller performance on server clusters analyzed? [65]
	ISR_{37} to ISR_{40} Under Process 13 – Service Provider Management Controls
13.1	Are outsourced systems managed with a degree of control? [264]
13.2	Are vendor services designed based on the size of organization? [67]
13.3	Are vendor service failures and service recoveries analyzed? [68]
13.4	Are vendor service level guarantees specified in SLA? [70]

By leveraging these rules, FIs understand which process in the Life Cycle lacks ISCs [181].

7.2.3 Information Security Scoring Function

In the model, the level of all *ISR* is computed as:

$$\frac{\sum_{i=1}^{6} ISR_{1_i}\, k_{1_i} + \sum_{i=1}^{6} ISR_{2_i}\, k_{2_i} + \cdots + \sum_{i=1}^{6} ISR_{40_i}\, k_{40_i}}{\sum_{i=1}^{6} ISR_{1_i} + \sum_{i=1}^{6} ISR_{2_i} + \cdots + \sum_{i=1}^{6} ISR_{40_i}} \tag{7.1}$$

where k_i is the score of ISR_i within a range (of one to five) representing the level of each of the *ISR* - very low, low, average, high and very high:

$$k_i = \begin{cases} ISR_{verylow}, k_i = 1 \\ ISR_{low}, k_i = 2 \\ ISR_{average}, k_i = 3 \\ ISR_{high}, k_i = 4 \\ ISR_{veryhigh}, k_i = 5 \end{cases} \tag{7.2}$$

Forty (40) *ISR* are defined. Subsequently, the total score of Systems for *ISR* is computed as:

$$ISR_1k_1 + ISR_2k_2 + \ldots + ISR_{40}k_{40} \tag{7.3}$$

The lower the score, the lower the level of ISCs. Let inputs are x_i for Systems:

$$f(x_i) = x_i(ISR_1k_1 + ISR_2k_2 + \ldots + ISR_{40}k_{40}) \tag{7.4}$$

This is applied to a scoring function $f(x_i)$ with a dataset. The function is a weighted average of scores for Systems. Scores are determined by questions defined in *ISR*, similar to the function in a previous research [180]. Upon confirmation of these scores, we aggregate them to evaluate the collective level of controls.

With all scores, the probability (P) of the joint occurrence for ISR_i is computed as:

$$P(ISR_i) = P(ISR_1(x_i)) \cdot P(ISR_2(x_i)) \cdot \ldots \cdot P(ISR_{40}(x_i)) \quad (7.5)$$

The probability is classified into a rank (r_i) out of five ranks: the level of non-compliance (C_{non}), limited compliance ($C_{limited}$), partially compliance ($C_{partially}$), significantly compliance ($C_{significantly}$) and wholly compliance (C_{wholly}):

$$f(r_i) = \begin{cases} C_{non}, r_i = 1 \\ C_{limited}, r_i = 2 \\ C_{partially}, r_i = 3 \\ C_{signficantly}, r_i = 4 \\ C_{wholly}, r_i = 5 \end{cases} \quad (7.6)$$

In case an overall information security control over Systems is adequate, the compliance level (C) will be ranked as 5. Otherwise, the level will scale from one to four dependent on the adequacy of controls (measured in terms of scores for all ISR). As a result, this is a five-ranking function.

7.2.4 LSTM Networks with an Attention Mechanism

In the model, we take a series of steps to train LSTM RNNs for predicting IS compliance levels, as described in **Fig. 7.2**.

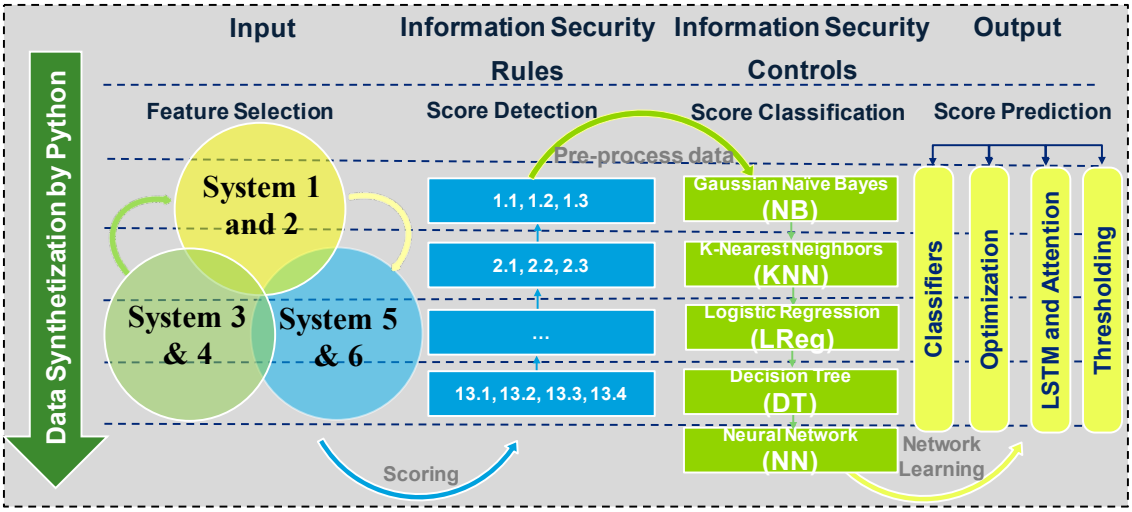


Fig. 7.2 Network Training Procedures

In the beginning, LSTM RNNs are imported with a dataset. Data features are *ISR'* scores of six systems. Scores are classified into ISCs. All ISCs are ranked by IS compliance levels before networks forecast the levels.

In total, we train four networks including FF LSTM RNNs without an ATTN, BK LSTM RNNs without an ATTN, FF LSTM RNNs with an ATTN, and BK LSTM RNNs with an ATTN. These networks learn temporal sequences and correlations of ISCs with an activation function, Sigmoid. They predict IS compliance levels to ascertain whether output sequences exceed thresholds. This threshold refers to the acceptance level of IS. Apart from the sequence prediction, we apply an ATTN to networks to direct the learning to pay attention to important ISCs.

In experiments, these network outcomes are compared with other networks such as NB, KNN, LReg and DT. Regardless of network types, all networks use classifiers to categorize IS compliance levels before making prediction.

7.3 Experiments

The purpose of these experiments is to study the IS compliance prediction performance of our proposed model by running experiments with an IS dataset.

The experiment setup and data split are the same as that in Chapter 3.

Dataset

In real life, no real banking data related to IS is publicly available. These are confidential information which cannot be disclosed. In view of this, we synthesize a dataset with Python. To create this, we pre-define rules including a) scores with a range from one to five; and b) forty instances (equivalent to ISR).

In this dataset, we assume six systems, forty instances, and three data inputs such as question numbers, questions and scores. The assumption is made to meet CPG 234 requirement. Under CPG 234, multiple systems need to be considered and a range of ISCs should be proposed [75]. These parameters in this chapter are set for the demonstration purpose. These vary depending on the real situation of a company.

For these six systems, answers to questions are stated in terms of IS scores. Some data features are extracted to **Table 7.2**.

Table 7.2 IS Data Features (Samples)

ISR No	ISR	Question No	Questions	System 1	System 2	System 3	System 4	System 5	System 6
1	1	1.1	.	5	1	5	3	1	1
2	1	1.2	.	1	3	2	4	3	3
3	1	1.3	.	1	3	5	2	1	1
4	2	2.1	.	4	5	5	1	5	5
5	2	2.2	.	4	5	5	4	3	3
6	2	2.3	.	1	5	4	3	1	1
7	3	3.1	.	1	2	5	1	4	4
8	3	3.2	.	4	4	2	5	4	4
9	3	3.3	.	1	5	4	3	5	5
10	4	3.4	.	4	5	5	2	1	1

The statistics of scores are listed in **Table 7.3**. The scores are binarized before they are

input into networks for prediction.

Table 7.3 Score Statistics (S.D. – Standard Deviation)

Scores	System 1	System 2	System 3	System 4	System 5	System 6
Mean	2.8250	3.5500	3.3500	3.0750	2.9500	2.9500
S.D.	1.5340	1.2598	1.5115	1.3085	1.4841	1.4841

Results

We compare different LSTM RNNs to check which one outputs the best prediction result. Experimental outcomes are compared in **Table 7.4**.

Table 7.4 Prediction Results by Networks

LSTM RNNs	Accuracy	Loss	PN	RL	F1
FF	0.9500	0.4120	0.9025	0.9500	0.9256
BK	0.3000	0.7167	0.8726	0.3000	0.4213
FF+ATTN	0.9500	0.3603	0.9025	0.9500	0.9256
BK+ATTN	0.9500	0.4543	0.9025	0.9500	0.9256

From **Table 7.4**, the highest accuracy is found in the network of FF, FF+ATTN and BK+ATTN. But the lowest loss occurs in the network of FF+ATTN (0.3603). In addition, these three networks have similar PN, RL and F1, significantly higher than that of the network, BK. Overall, the FF+ATTN attains the highest prediction accuracy and minimalizes the loss. ATTN helps BK improve the network predictability from 30% to 95%. Also, it optimizes the performance of FF and BK by reducing the loss from 0.4120 to 0.3603 and increasing the accuracy from 30% to 95% respectively.

Result Evaluation

We evaluate networks by cross-validation (CV). Results are listed in **Table 7.5**. Validated (V.) accuracy for all networks except the BK is the highest (95%) while validated loss for the FF+ATTN is the lowest (0.0922).

Table 7.5 Evaluation Results by Networks

LSTM RNNs	V. Accuracy	V. Loss
FF	0.9500	0.4016
BK	0.3000	0.2614
FF+ATTN	0.9500	0.0922
BK+ATTN	0.9500	0.1375

Case Studies

We make an in-depth analysis on the prediction of IS compliance levels as follows.

Case 1 – We analyze how prediction look like, as visualized in **Fig. 7.3**. Prediction are measured in terms of output sequence length (from 0 to 1). From the BK LSTM RNN, prediction deviate from actual outputs significantly when compared with other networks over timesteps.

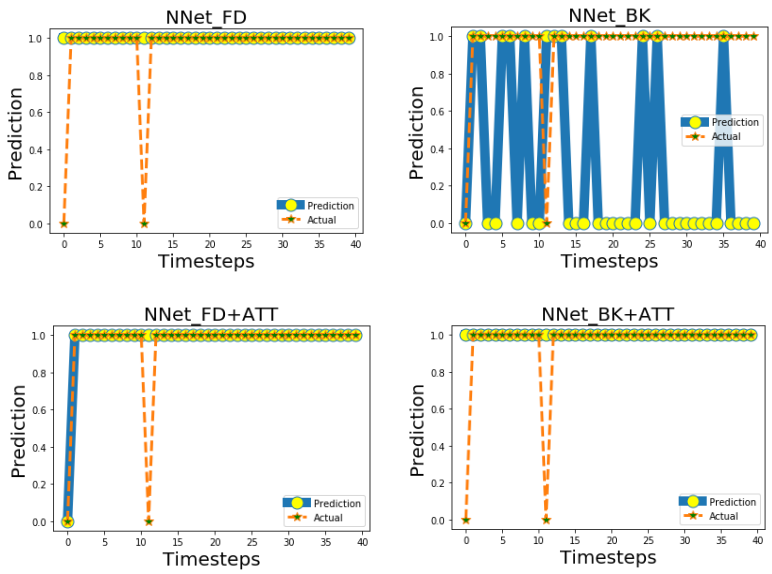


Fig. 7.3 Prediction of Compliance Levels by Networks

Note to **Fig. 7.3** and **7.4**: FF LSTM RNN is NNet_FD, BK LSTM RNN is NNet_BK, FF LSTM RNN with ATTN is NNet_FD+ATT and BK LSTM RNN with ATTN is NNet_BK+ATT

Case 2 - To confirm if there is an overfit or underfit issue for networks, we compare training losses against validated losses, as given in **Fig. 7.4**. Both training and validated data cannot meet for these networks. This can be avoided using more data. When training data and validated data meet at an inflection point, we can stop training the network. Note that the loss of the FF LSTM RNN (0.4120), BK LSTM RNN (0.7167) and BK LSTM RNN with an ATTN (0.4543) are not comparable to that of the FF LSTM RNN with an ATTN (0.3603).

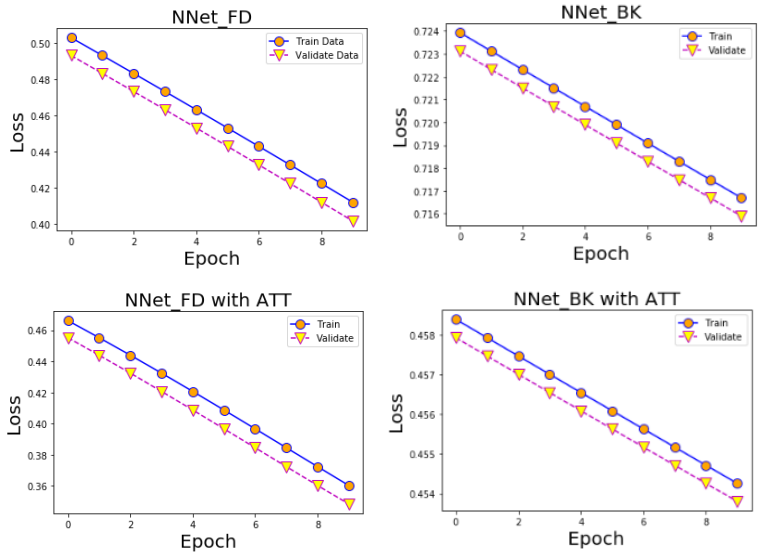


Fig. 7.4 Train Loss and Validated Loss by Networks

Case 3 - To compare MSE and V. MSE of networks, we extract results to **Fig. 7.5**. Consistently, the FF LSTM RNN with an ATTN minimizes the MSE and V. MSE out of all networks.

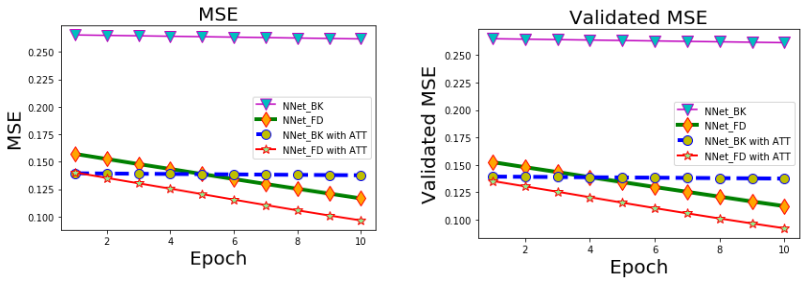


Fig. 7.5 MSE and Validated MSE by Networks

Case 4 - We compare LSTM RNNs with other networks. Comparison results are listed in **Table 7.6**.

Table 7.6 Comparison of Different Networks

Networks	Accuracy	Loss	PN	RL	F1
NB	0.4286	1.5109	0.3389	0.3333	0.2976
KNN	0.5357	20.6675	0.0208	0.0833	0.0333
LReg	0.5357	2.2851	0.5139	0.4167	0.3556
DT	0.9286	23.1414	0.3833	0.2500	0.2847
FF LSTM RNN +ATTN	0.9500	0.3603	0.9025	0.9500	0.9256

The FF LSTM RNN with an ATTN achieves the highest accuracy (95.00%) and the lowest loss (0.3603) in comparison with NB, KNN, LReg and DT. At the same time, it attains the highest PN (90.25%), RL (95.00%) and F1 (92.56%).

Case 5 - We evaluate other networks by CV. Results are listed in **Table 7.7**. V. accuracy and loss in the FF LSTM RNN with an ATTN (95.00% and 0.0922 respectively) are superior to that in NB, KNN, LReg and DT. In consideration of the accuracy and loss collectively, the FF LSTM RNN with an ATTN is the most efficient network relative to others in this experiment.

Table 7.7 Evaluation of Different Networks

Networks	V. Accuracy	V. Loss
NB	0.3333	4.6189
KNN	0.0833	6.8511
LReg	0.4167	2.6536
DT	0.2500	4.9994
FF LSTM RNN +ATTN	0.9500	0.0922

Model Reports

In order to prepare reports for the compliance purpose, we compute the distribution of scores for six systems. Outputs are extracted to **Fig. 7.6**. This shows the overall compliance level of all systems which are to be attested by financial regulators (when needed).

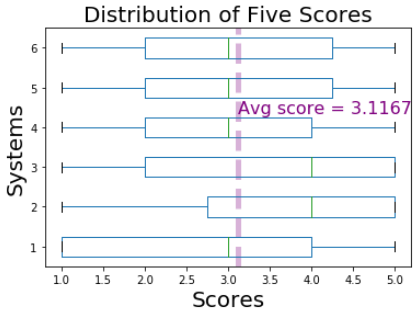


Fig. 7.6 Distribution of Five Scores for Six Systems

It can be observed that most 'score 1' are found in system 1. They should be improved as soon as possible due to significant control weaknesses. **Fig. 7.6** shows the measurement of compliance levels in terms of metrics such as 60% of compliance for score 6, 60% of compliance for score 5, 60% of compliance for score 4, 80% of compliance for score 3, 80% of compliance for score 2, and 60% of compliance for score 1.

Apart from this high-level report, we analyze scores by ISCs under the Life Cycle. These scores in detailed reports are extracted to **Fig. 7.7**. Two controls out of all ISCs are found with the highest number of 'score 1'. They are change management and response controls. Take the change process as an example, 7 *ISR* are rated as 'score 1'. Hence, we notice that system 1, 5 and 6 have two *ISR* rated as 'score 1' while system 2 has one *ISR* rated as 'score 1'. To avoid violating the regulatory requirement, FIs can prioritize these controls for IS enhancement purpose earlier. **Fig. 7.7** can be interpreted as the measurement of compliance levels in terms of metrics. Take the information security control of change management as an example, system 3 is 100% compliant with control 1 and 3 but only 60% compliant with control 2.

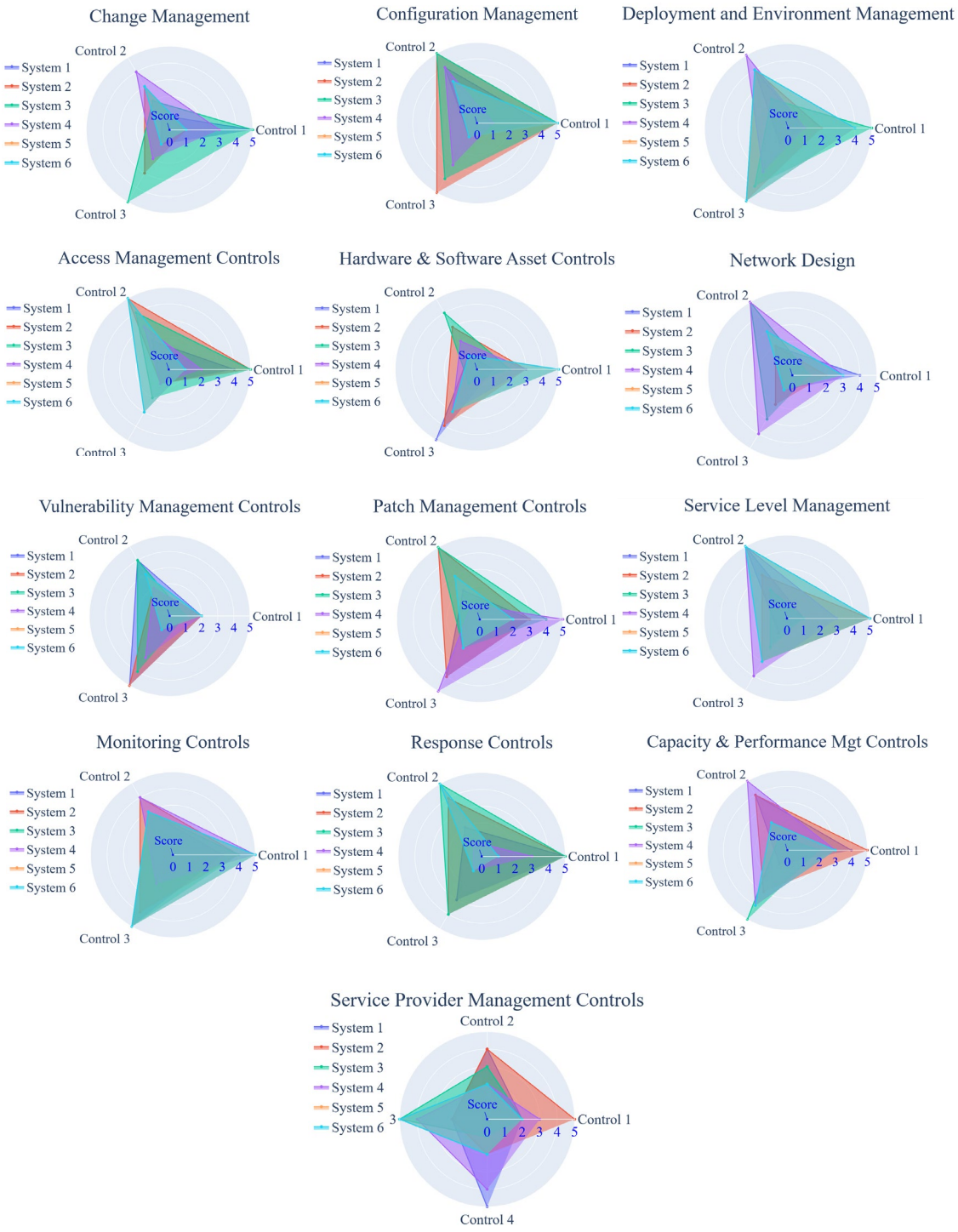


Fig. 7.7 Compliance Report by Controls under the Life Cycle

Above performance measurements such as accuracy, loss, precision, recall and F1-Support are defined in equations 2.37 to 2.41 inside Chapter 2.

As we can see, our experiments show an outstanding IS compliance prediction performance of our proposed model by running experiments with an IS dataset.

Related Works

There has been limited research on the application of LSTM networks with an ATTN to automate the IS compliance process. We are the first to present a ML model predicting IS compliance levels based on IS score correlations and their importance for the financial services industry.

Instead, a significant amount of previous studies directed sequential learning with an attention in other domains such as dependencies of a problem [203], computation of response at a position in a sequence by attending to all positions in a self-attention network [94], improvement in the balance between the ability to model dependencies in a generative adversarial network with a self-attention [95], and prediction of attributes for images by considering co-occurrence dependencies of attributes [96].

In the industry, there have been a few research studies applying a similar ML model such as prediction of stock prices in attention LSTM DNN [207] and forecast of LR with DNN [126]. These are yet extended to predict compliance levels.

7.4 Summary

We propose a ML model to overcome the problem of predicting IS compliance levels. The model networks can be leveraged to automate the compliance process in a real time mode. The model prediction in accordance with a local requirement, CPG 234, facilitates FIs to meet the expectation of financial regulators in the financial services industry.

We firstly develop a compliance approach in **Fig. 7.1**. Next step is to define ISR to determine IS scores while rules are listed in **Table 7.1**. The scores are aggregated into a scoring function for ranking IS compliance levels. The levels are input into networks for prediction.

Section 7.2.3 describes the procedure of network training where networks in this chapter are LSTM RNNs which are the same as that in Chapter 3. These networks in this chapter are applied with an ATTN. Our model goes beyond the traditional method by exploiting sequential correlations of IS scores and paying special attention to important IS scores.

Overall experiment results over an IS dataset demonstrate that our model achieves higher prediction accuracy, PN, RL and F1 in ISL prediction in comparison with other DNN, such as NB, KNN, LReg and DT.

Predicting IS compliance levels with analytical reports is beneficial to the industry. Relying on the reports, FIs understand which controls under the Life Cycle should be prioritized for the IS enhancement purpose [75].

Next chapter concludes all of the chapters mentioned above and key contributions made with these research works followed by a discussion on future works.

Chapter 8

Conclusion

The aim of this thesis mentioned in Section 1.2.2 of Chapter 1 is to enhance DG of FIs by: a) improving DQ [4]; b) managing data risks [120]; c) managing bank-wide risks [4]; and d) enhancing IS [75].

For this purpose, we propose five ML models to learn regulatory compliance data:

a. In Chapter 3, a DQP model using supervised learning under DG to meet the regulatory requirement of DG. This model considers sequential learning of DN by taking temporal sequences and correlations of DN into account;
b. In Chapter 4, a DQP model using unsupervised learning under DG to meet the regulatory requirement of DG. This model considers the importance of DN on top of the temporal sequences and correlations of DN collectively in DQP. Additionally, this model takes temporal sequences and correlations of DN into account in DQ measurement;
c. In Chapter 5, a DQP analytical model under DG to meet the regulatory requirement of DG;
d. In Chapter 6, a DQP network efficiency improvement model under DG to meet the regulatory requirement of DG by measuring network run-time saving; and
e. In Chapter 7, an ISL prediction model considering sequences, correlations and importance of IS factors collectively for IS compliance prediction under DG to meet the regulatory requirement of DG.

These models address research limitations mentioned in Section 1.2.3 that most of the state-of-the-art research works have not covered, as elaborated in Section 2.2. These limitations are summarized in Section 8.1 of this chapter.

Chapter 3 to 7 show how our ML models accurately predict DQ and IS compliance levels during DG processes of FIs by learning regulatory compliance data from both theoretical and experimental perspectives. Experimental results demonstrate that these models under DG succeed in accurately predicting DQ in supervised learning, precisely predicting DQ in unsupervised learning, analyzing DQP by divergent dimensions such as by risk types and business segments, saving significant network run-time in DQP for improving the network efficiency, and accurately forecasting IS compliance levels.

This chapter first summarizes proposed models, findings and research aims of each chapter of this thesis. Then, Section 8.2 discusses contributions of our models and Section 8.3 outlines our future work.

8.1 Summary of Chapters

In Chapter 1, we introduced thesis background, scope, aims, models, model networks, research methodology and thesis outline. Prior research showed that ML models have not been extensively applied to DG problems. In view of these ML research limitations, five ML models were proposed: a) DQP model using supervised learning under DG to meet the regulatory requirement of DG. This model considers sequential learning of DN by taking temporal sequences and correlations of DN into account; b) A DQP model using unsupervised learning under DG to meet the regulatory requirement of DG. This model considers the importance of DN on top of the temporal sequences and correlations of DN collectively in DQP. Additionally, this model takes temporal sequences and correlations of DN into account in DQ measurement; c) A DQP analytical model under DG to meet the regulatory requirement of DG; d) A DQP network efficiency improvement model under DG to meet the regulatory requirement of DG by measuring network run-time saving; and e) an ISL prediction model considering sequences, correlations and importance of IS factors collectively for IS compliance prediction under DG to meet the regulatory requirement of DG**.**

In Chapter 2, we listed all research topics that have been studied including ML techniques, ML models, model networks, ML work related to DG, summary on the limitations of

research work, and our research proposal together with the research approach and research methodology. There have been many ML methods including unsupervised, supervised and reinforcement learning. In the ML model development, many components need to be considered including inputs, processing and outputs. For the model implementation, numerous networks could be leveraged including a) LReg, DT, SVM, NB, RF; b) KNN; and c) ANN. These networks were commonly used for prediction and analytical purposes. In networks, there have been various learning techniques. We introduced network methodologies, network memory, network wrapper, network activation, network sequence prediction, network optimization, network algorithms and network performance metrics. Following this, we surveyed ML works related to DG. There have been numerous ML models developed. These could be applied to DG including learning of DQ or IS compliance. Research results showed that the extent of ML application to DG was not extensive from four perspectives including DG measurement and prediction, DG requirement compliance, network efficiency improvement in DQP and IS compliance prediction. Then, we summarized limitations of current research work. These limitations motived us to propose our research work. In the research proposal, we proposed five ML models. Under the research approach, we covered three artifacts. To design proposed artifacts, we used a design science methodology "DSRM".

In Chapter 3, we proposed a ML model to predict DQ under DG in supervised learning. This model tackled a real DG issue in the financial services industry that most of the state-of-the-art research works have not covered: i.e., a) ML techniques have not been applied in DQP during DG processes using supervised learning extensively. In particular, a ML model has not been proposed for DQP during DG processes in supervised learning; b) ML techniques have not been applied in the forecast of DQ during DG processes by taking temporal sequences and correlations of DN into consideration. In particular, a ML model considering temporal sequences and correlations of DN has not been proposed for DQP under DG; and c) ML techniques have not been applied to DQ learning such as DQP during DG processes for meeting DG regulatory requirements. In our ML model, we firstly labelled data based on an international requirement of BCBS 239. The model was implemented with LSTM RNNs for DQP including FF LSTM RNN, BK LSTM RNN and

BD LSTM RNN. Then, we directed networks to learn temporal sequences and correlations between DN sequences with a synthesised dataset. Following this, the model was validated with a realistic banking dataset. Experimental results demonstrated that our model was effective in DQP in terms of the prediction accuracy and error. This model could be used to improve DQ [4, 6] by predicting DQ according to an international requirement, BCBS 239, in supervised learning.

In Chapter 4, we proposed a ML model to predict DQ under DQ in unsupervised learning. This model tackled a real DG issue in the financial services industry that most of the state-of-the-art research works have not covered: i.e., a) ML techniques have not been applied in DQP during DG processes using unsupervised learning. In particular, a ML model has not been proposed for DQP during DG processes in unsupervised learning; b) temporal sequences and correlations of DN have not been considered in DQ measurement. In particular, a ML model has not been proposed for DQ measurement under DG taking temporal sequences and correlations of DN into account; c) ML techniques have not been applied in the forecast of DQ during DG processes by taking temporal sequences, correlations and importance of DN into consideration. In particular, a ML model considering temporal sequences, correlations and importance of DN has not been proposed for DQP under DG; and d) ML techniques have not been applied to DQ learning such as DQP during DG processes for meeting DG regulatory requirements. In our model, unsupervised learning method saved time in data labelling. At first, we detected DN from a dataset pursuant to an international requirement, BCBS 239. Detected DN impacts were estimated in two generative mixture methods as weights in unsupervised learning. The weights were input into networks for DQP. Networks implemented were LSTM RNNs applying sequential learning along with an ATTN to predict DQ taking not only temporal sequences and correlations of DN into account but also the importance of DN. Network performance was further optimized by dual regularizations. Consequently, the network prediction accuracy, PN, RL and F1 were high while the prediction error was low. The network performance was examined at two levels including an integrated level and an individual level to analyze the model predictive power. This model could be used to improve DQ [4, 6] by predicting DQ in accordance with the international requirement,

BCBS 239, in unsupervised learning.

In Chapter 5, we presented a ML model to analyze DQP under DG. This model tackled a real DG issue in the financial services industry that most of the state-of-the-art research works have not covered: i.e., a) ML techniques have not been applied in DQP analytics under DG to meet DG regulatory requirements. A ML model for analyzing DQP during DG processes has not been proposed; and b) ML techniques have not been applied to DQ learning such as DQP analytics during DG processes for meeting DG regulatory requirements. In the model, DQP were analyzed in accordance with a local requirement, CPG 235. The model was implemented by LSTM RNNs applying more complicated learning methods such as windows, timesteps and MBB to learn DQP differently. Experimental results showed that our model was effective in DQP analytics. The model accurately predicted DQ and at the same time analyzed DQ at a granular level by risk types (such as MR, CR, OR and LR) and business segments (including PvB, WB and RB). Networks were tested with divergent algorithms. This model could be used to manage data risk [120] by analyzing DQP according to a local requirement of CPG 235.

In Chapter 6, we proposed a ML model to improve the network efficiency in DQP under DG. This model tackled a real DG issue in the financial services industry that most of the state-of-the-art research works have not covered: i.e., a) ML techniques have not been applied in the network efficiency improvement in DQP. DQP network run-time saving has not been measured and a ML model for the network efficiency improvement in DQP during DG processes has not been proposed; and b) ML techniques have not been applied to DQ learning such as DQP during DG processes for meeting DG regulatory requirements. Before predicting DQ based on a local requirement, CPG 235, we proposed a data profiling approach to slide divergent portions of data from a dataset for learning in networks. After this, we proposed a ML model to train LSTM RNNs applying MB for measuring network run-time saving. Networks were tested with various algorithms to find the most efficient network. They have been evaluated by validation data to confirm the model effectiveness. Experimental results demonstrated a significant improvement in the network efficiency for DQP in terms of the network run-time saving while maintaining a high prediction accuracy.

This model could be used to manage bank-wide risk [4] by improving the network efficiency of DQP.

In Chapter 7, we proposed a ML model to predict IS compliance levels under DG. This model tackled a real DG issue in the financial services industry that most of the state-of-the-art research works have not covered: i.e., a) ML techniques have not been applied in the prediction of ISL and IS compliance levels under DG. ISL during DG processes have not been predicted. A ML model considering sequences, correlations and importance of IS factors collectively for IS compliance prediction has not been proposed; and b) ML techniques have not been applied to IS learning during DG processes for meeting DG regulatory requirements. At first, we developed a compliance approach. Then, we defined ISR according to a local requirement of CPG 234 for detecting ISCs in a ML model. After this, detection results were aggregated in a scoring function for ranking ISL. The ISL were input into networks, LSTM RNNs, for predicting IS compliance levels. LSTM RNNs were applied with sequential learning and an ATTN to learn sequential correlations of IS scores and pay special attention to important IS scores. These networks automated the compliance process to generate analytical reports showing IS compliance levels under the Life Cycle. Experimental results demonstrated that our model achieved higher prediction accuracy, PN, RL and F1 in ISL prediction when compared with other DNN, such as NB, KNN, LReg and DT. This model could be used to enhance IS [75] by predicting IS compliance levels according to a local requirement, CPG 234.

8.2 Contributions

From literature results summarized in Section 1.2.3 and elaborated in Section 2.2, ML techniques have not been largely deployed to DG. ML models have not been widely applied to and experimented in DG processes. These motivate us to make contributions towards the application of ML models to DG for the financial services industry, as elaborated below.

8.2.1 Data Quality Prediction in Supervised Learning

We propose an effective ML model to tackle a real DG issue in the financial services

industry that most of the state-of-the-art research works have not covered: i.e., a) ML techniques have not been applied in DQP during DG processes using supervised learning extensively. In particular, a ML model has not been proposed for DQP during DG processes in supervised learning; b) ML techniques have not been applied in the forecast of DQ during DG processes by taking temporal sequences and correlations of DN into consideration. In particular, a ML model considering temporal sequences and correlations of DN has not been proposed for DQP under DG; and c) ML techniques have not been applied to DQ learning such as DQP during DG processes for meeting DG regulatory requirements. These are summarized in Section 1.2.3 and elaborated in Section 2.2.

The contributions of this model can be summarized as follows: a) we take advantage of DQ principles from an international regulatory requirement of BCBS 239 [3] under DG to measure DQ in supervised learning with a ML model; b) we exploit a ML model considering DN in the current state, the previous state and the future state to predict DQ during DG processes by applying sequential learning to DNN; c) we train multiple LSTM RNNs (FF, BK and BD LSTM RNNs) in the model to test the predictability of divergent learning methodologies, improve the network predictive power by applying heterogenous algorithms (ADAM, SGD, ADADELTA, ADAGRAD and RMSPROP) to networks, and find the most efficient network for DQP; and d) we construct a dataset made of four risk types which are not available in the public domain. This dataset is used to test networks for confirming the model effectiveness, and can be used in other research works. To our knowledge, this presents the first attempt to develop a ML model in alignment with the international regulatory requirement of BCBS 239 to predict DQ and the model is designed with sequential learning taking temporal sequences and correlations of DN into account in DQP. Accordingly, the learning of complex DN relations is close to real word scenarios for DQP under DG. At the end, experiments show accurate prediction of DQ. Thereupon, the model is helpful for the improvement of DQ [4] (as mentioned in Section 1.2.2 Aims of this thesis) earlier, as part of the DG.

8.2.2 Data Quality Prediction in Unsupervised Learning

We propose an effective ML model to tackle a real DG issue in the financial services industry that most of the state-of-the-art research works have not covered: i.e., a) ML techniques have not been applied in DQP during DG processes using unsupervised learning. In particular, a ML model has not been proposed for DQP during DG processes in unsupervised learning; b) temporal sequences and correlations of DN have not been considered in DQ measurement. In particular, a ML model has not been proposed for DQ measurement under DG taking temporal sequences and correlations of DN into account; c) ML techniques have not been applied in the forecast of DQ during DG processes by taking temporal sequences, correlations and importance of DN into consideration. In particular, a ML model considering temporal sequences, correlations and importance of DN has not been proposed for DQP under DG; and d) ML techniques have not been applied to DQ learning such as DQP during DG processes for meeting DG regulatory requirements. These are summarized in Section 1.2.3 and elaborated in Section 2.2.

We propose a new unsupervised learning model that does not require data labelling. The unsupervised learning saves significant data labeling time [161]. Our ML model firstly exploits DN weighing in two generative mixture methods rather than labelling data manually after detection of DN pursuant to an international regulatory requirement of BCBS 239 [3]. DN weights are aggregated in a scoring function before they are input into networks for prediction. Relevant aggregate quality scoring algorithm is proposed to provide guidance on the measurement of DQ by a new DQ scoring method under DG. This method has not been proposed in earlier ML research related to DG, as mentioned in Section 1.2.3 and Section 2.2. Since a ML model is yet to be developed for DQP during DG processes in supervised and unsupervised learning. The model then takes advantage of sequential learning of DNN to predict DQ during DG processes by considering DN in the current state, the previous state and the future state. Multiple networks are tested including FF, BK and BD LSTM RNNs and their learning algorithms are proposed to provide guidance on DQP by different DQ learning methods under DG. These methods have not been proposed in earlier ML research related to DG, as mentioned in Section 1.2.3 and

Section 2.2. A ML model is yet to be developed for DQP during DG processes in supervised and unsupervised learning. The networks are applied with an ATTN, not only taking temporal sequences and correlations of DN into account in the DQP but also the importance of DN. This deviates from Chapter 3. The experiment results show that our model provides accurate estimates of DQ at both integrated and individual levels when supervised learning is not practicable or labelling data is costly. These accurate estimates enable FIs to understand what DQ are going to be with a sceientific computational method under DG. Thus, FIs can enhance their forward-looking capabilities of DQ [3] by providing any potential violations of risk limits over thresholds. They can also improve DQ in long term [3] to satisfy the expectation of regulators in the industry. Accordingly, our model is useful for an early improvement of DQ [4] (as mentioned in Section 1.2.2 Aims of this thesis), as part of the DG.

8.2.3 Data Quality Prediction Analytics

We propose an effective ML model to tackle a real DG issue in the financial services industry that most of the state-of-the-art research works have not covered: i.e., a) ML techniques have not been applied in DQP analytics under DG to meet DG regulatory requirements. A ML model for analyzing DQP during DG processes has not been proposed; and b) ML techniques have not been applied to DQ learning such as DQP analytics during DG processes for meeting DG regulatory requirements. These are summarized in Section 1.2.3 and elaborated in Section 2.2.

Given this research gap, we are motivated to analyze DQP under DG with a ML model. We take a series of steps to achieve this by: a) labelling data according to a local regulatory requirement of CPG 235 [120] in terms of six DQ dimensions including accuracy, completeness, consistency, timeliness, availability, and fitness for use in supervised learning; b) directing networks, LSTM RNNs, in a ML model with complex learning methods to learn DQP differently for the multi-dimensional analytical purpose while the methods contain windows, time-steps and memory between batches; c) testing networks with divergent algorithms to find the most efficient performance; and d) generating

granular analytical reports on DQP during DG processes by risk types (such as MR, CR, OR and LR) and by business segments (such as PvB, WB and RB) for use by FIs. These analytics can be used to satisfy the expectation of regulators in the industry and help to strengthen the capabilities of FIs to understand where potential poor data is under DG. The model effectiveness is demonstrated by conducting a set of experiments over the risk dataset. Accordingly, this model is advantageous for an early data risk management [120] (as mentioned in Section 1.2.2 Aims of this thesis), as part of the DG.

8.2.4 Network Efficiency Improvement in Data Quality Prediction

We propose an effective ML model to tackle a real DG issue in the financial services industry that most of the state-of-the-art research works have not covered: i.e., a) ML techniques have not been applied in the network efficiency improvement in DQP. DQP network run-time saving has not been measured and a ML model for the network efficiency improvement in DQP during DG processes has not been proposed; and b) ML techniques have not been applied to DQ learning such as DQP during DG processes for meeting DG regulatory requirements. These are summarized in Section 1.2.3 and elaborated in Section 2.2.

The issue is addressed by a) adopting a systematic approach to measure the network run-time saving for DQP during DG processes; b) improving the network efficiency for DQP based on a local regulatory requirement, CPG 235 [120], in a ML model; c) reporting network run-time saving by testing different portions of data in the model; and d) showing a significant percentage of the DQP run-time saving in experimental results. We believe that the approach and model can be used by FIs to swiftly find future good and bad quality of data from the massive amount of data [82] with the use of a small fraction of data while maintaining accurate DQP under DG. This model is scalable and has the potential to expand its applicability to other network efficiency improvement research works. Thus, the model helps to efficiently manage bank-wide risks [4] (as mentioned in Section 1.2.2 Aims of this thesis) earlier, as part of the DG.

8.2.5 Information Security Compliance Prediction

We propose an effective ML model to tackle a real DG issue in the financial services industry that most of the state-of-the-art research works have not covered: i.e., a) ML techniques have not been applied in the prediction of ISL and IS compliance levels under DG. ISL during DG processes have not been predicted. A ML model considering sequences, correlations and importance of IS factors collectively for IS compliance prediction has not been proposed; and b) ML techniques have not been applied to IS learning during DG processes for meeting DG regulatory requirements. These are summarized in Section 1.2.3 and elaborated in Section 2.2.

We propose a ML model to overcome limitations of these research works. Considering the automation of IS compliance process in a real time mode, we leverage networks, LSTM RNNs, to predict IS compliance levels during DG processes. The prediction are made in accordance with a local requirement, CPG 234 [75]. At first, we develop a compliance approach and define ISR to determine IS scores. The scores are aggregated into a scoring function for ranking ISL which are input into networks for prediction. The networks are applied with sequential learning and an ATTN. This goes beyond the traditional method by exploiting temporal sequences and correlations of IS scores and paying attention to important IS scores under DG. Experimental results over the dataset indicate that performance of our model significantly outperforms that of other networks such as NB, KNN, LReg and DT in terms of prediction accuracy, precision, recall and F1-Support. Beyond comparing the predictability of our model, we report weaknesses of ISCs by thirteen controls under the Life Cycle to satisfy the expectation of regulators in the industry. These help to strengthen the capabilities of FIs to understand where potential weaknesses of ISCs are. Accordingly, this model is beneficial to an enhancement of IS [75] (as mentioned in Section 1.2.2 Aims of this thesis) earlier, as part of the DG.

Thereupon, all these models strengthen the capabilities of FIs in DG by improving DQ [4], managing data risks [120], managing bank-wide risks [4], and enhancing IS [75] based on regulatory requirements of the financial services industry including BCBS 239, CPG 235 and CPG 234. These improvements can be included in DGP under DGF to enhance the

governance of DM.

8.3 Future Work

This section first outlines potential short-term extensions to our proposed models in each chapter, and then returns to discuss untouched directions for future research.

8.3.1 Short Extensions

In Chapter 3, we proposed a DQP sequential learning model considering DN in the current state, the previous state and the future state during supervised learning under DG. However, data remediation process is yet to be automated [4]. For example, data can be rectified by the imputation of values for the quality improvement. With predicted DQ in Chapter 3, data of low quality can be improved with another ML model.

In Chapter 4, we presented a DQP sequential learning model with an ATTN in unsupervised learning under DG. This work exploited temporal sequences and correlations of DN as well as the importance of DN in DQP. This work can be extended to non-regulatory compliance data in the financial services industry such as big data. An example of big data is equities [178].

In Chapter 5, we presented a DQP analytical model taking the dimension of risk types and business segments into account under DG. For meeting regulators' expectation [4], we can automate data remediation and analyze the remediation by different dimensions with a new ML model.

In Chapter 6, we presented how to measure network run-time saving with divergent portions of data in a DQP network efficiency improvement model under DG. We selected some data features for testing such as MR. More data features such as CR, OR and LR can be used to test the network saving time for finding the most efficient network by risk types.

In Chapter 7, we proposed an ISL prediction model under DG to predict IS compliance

levels based on IS scores. In order to demonstrate the applicability and scalability of this model, we will apply the ML model in Chapter 7 to predict compliance levels of General Data Protection Regulations. This facilitates to identify data privacy risks [174].

8.3.2 Future Direction

Along with the proposed improvements above, there are some related areas that we did not touch on in this thesis. This section explains these areas.

Constructing Datasets for High Risk Data

In Chapter 3, we constructed a huge dataset made of four risk types. However, we did not focus on constructing datasets containing high risk data such as anti-money laundering [176] for the purpose of DQP. However, current datasets from the Internet did not cover this financial crime risk data due to the data sensitivity. This hampers the progress of ML models to identify and predict DN. Simulating datasets for this data is a crucial future direction that will help to improve DQ and manage data risks from the perspective of the anti-money laundering.

Harnessing ML for the Non-Financial Services Industry

Harnessing ML models for learning data in other industries is also an important direction which is untouched in this thesis. For example, there have been multiple data issues in the telecommunication industry [177]. We will improve the network efficiency of DQP for this industry by applying the ML model in Chapter 6. This helps to solve other industry issues.

All these become the follow-up research of this thesis. The aim is to help resolve more real problems in the financial services industry as well as in the non-financial services industry.

Bibliography

[1] T. M. Mitchell., "The discipline of machine learning," *Carnegie Mellon University, School of Computer Science, Machine Learning Department* (2006)

[2] T. Hey., "The next scientific revolution," *Harv Bus Rev*, 88(11):56–63 (2010)

[3] Bank for International Settlemenst (BIS), "BCBS - Principles for Effective Risk Data Aggregation and Risk Reporting," *BIS*, pp. 8 to 23 (2013)

[4] Bank for International Settlemenst (BIS), "BCBS - Progress in Adopting the Principles for Effective Risk Data Aggregation and Risk Reporting," *BIS*, pp. 4-13 (2018)

[5] Harreis, H., Ho, T., Machado, J., Merrath, P., Rowshankish, K., and Tavakoli, A., "Living with BCBS 239, IIF Survey," *Mckinsey Institute of International Finance*, pp. 1-5 (2017)

[6] Financial Stability Board (FSB), "Report - Artificial Intelligence and Machine Learning in Financial Services, Market Developments and Financial Stability Implications," *FSB*, pp. 3-9 (2017)

[7] Weninger, F., Bergmann, J., and Schuller, B., "Introducing CURRENNT: The Munich Open-Source CUDA RecurREnt Neural Network Toolkit," *Journal of Machine Learning Research*, pp. 547-551 (2015)

[8] Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., and Xie, X., "Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM Networks," *in proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, Assocation for the Advancement of Artifical Intelligence, pp. 3697-3702 (2016)

[9] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B., "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification," *in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 207-212 (2016)

[10] Bianchi, F.M., Scardapane, S., LØkse, S., and Jenssen, R., "Bidirectional Deep-readout Echo State Networks," *in ESANN 2018 proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 425-430 (2018)

[11] Baran, R., and Zeja, A., "The IMCOP System for Data Enrichment and Content Discovery and Delivery," *2015 International Conference on Computational Science and Computational Intelligence*, IEEE, pp. 143-146 (2015)

[12] Deloitte, "A Guide to Assessing Your Risk Data Aggregation Strategies: How Effectively Are You Comply with BCBS 239," *Deloitte*, pp. 9 (2016)

[13] KPMG, "Equity Market Risk Premium – Research Summary," *KPMG*, pp. 3-7 (2018)

[14] Corporate Finance Institute (CFI), "Market Risk Premium," *CFI*, pp. 2-5 (2018)

[15] Moody's Analytics, "Credit Risk Calculator," *Moody's*, pp. 3-7 (2018)

[16] KPMG, "Basel 4: the Way Ahead, Operational Risk, The New Standardized Approach," *KPMG*, pp. 3-9 (2018)

[17] Migueis, M., "Is Operational Risk Regulation Forward-looking and Sensitive to Current Risks?," *Board of Governors of the Federal Reserve System*, pp. 1-7 (2018)

[18] The Board of the International Organization of Securities Commissions (BIOSC), "Recommendations for Liquidity Risk Management for Collective Investment Schemes," *BIOSC*, pp. 1-20 (2018)

[19] João Marcelo Borovina Josko, Marcio Katsumi Oikawa and João Eduardo Ferreira, "A Formal Taxonomy to Improve Data Defect Description," *International Conference on Database Systems for Advanced Applications (DASFAA)*, Springer, pp. 307-320 (2016)

[20] Ergen, T., and Kozat, S.S., "Efficient Online Learning Algorithms Based on LSTM Neural Networks," *Journal of IEEE Transactions on Neural Networks and Learning Systems*, IEEE, vol. 4, no. 2, pp. 1-12 (2017)

[21] Ruder, S., Ghaffari, P., and Breslin, J. G., "A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis," *Conference on Empirical Methods in Natural Language Processing – Association for Computational Linguistics*, arXiv preprint arXiv: 1609.02745 (2016)

[22] Graves, A., and Schmidhuber, J., "Framewise Phoneme Classification with Bidirectional LSTM Networks," *in Proceedings of International Joint Conference on Neural Network*, IEEE, pp. 2047-2051 (2005)

[23] Taylor, A., Leblanc, S., and Japkowicz, N., "Anomaly Detection in Automobile Control Network Data with Long Short-Term Memory Networks," *2016 IEEE International Conference on Data Science and Advanced Analytics*, IEEE, pp. 130-138

(2016)

[24] PWC, “Data Governance Survey Results: A European Comparison of Data Management Capabilities in Banks,” *PWC*, pp. 11 (2016)

[25] Yu, Z., Ramanarayanan, V., Suendermann-Oeft, D., Wang, X., Zechner, K., Chen, L., Tao, J., and Qian, Y., “Using Bidirectional LSTM Recurrent Neural Networks to Learn High-level Abstractions of Sequential Features for Automated Scoring of Non-Native Spontaneous Speech,” *in 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE, pp. 338-345 (2015)

[26] Fan, Y., Qian, Y., Xie, F., and Soong, F. K., “TTS Synthesis with Bidirectional LSTM based Recurrent Neural Networks,” *Fifteenth Annual Conference of the International Speech Communciation Association*, pp. 1964-1968 (2014)

[27] Kim, K., and Kim, J., “A Study on Analysing Risk Scenarios About Vulnerabilities of Security Monitoring System: Focused on Information Leakage by Insider,” *Information Workshop on Information Security Applications*, pp. 159-170 (2018)

[28] Farsad, N., and Goldsmith, A., “Detection Algorithms for Communication Systems Using Deep Learning,” *Stanford University*, *arXiv preprint arXiv:1705.08044* (2017)

[29] Kingma, D. P., and Lei Ba, J., ADAM: “A Method for Stochastic Optimization,” *International Conference on Learning Representation*, *arXiv preprint arXiv*:1412.6980 [cs.LG] (2014)

[30] Zhou, X., Wan, X., and Xiao, J., “Attention-based LSTM Network for Cross-Lingual Sentiment Classification,” *in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, pp. 247-256 (2016)

[31] Seo, S., Mohegh, A., Weiss, G. B., Liu, Y., “Automatically Inferring Data Quality for Spatiotemporal Forecasting,” *ICLR in US* (2018)

[32] Kontokosta, C., and Bonczak, B., “DataIQ – A Machine Learning Approach to Anomaly Detection for Energy Performance Data Quality and Reliability,” *in Proceedings of the 2016 ACEEE Summer Study on Energy Efficiency in Buildings in Asilomar* (2016)

[33] Visengeriyeva, L., Akbik, A., Kaul, M., Rabl, T., and Markl, V., “Improving Data Quality by Leveraging Statistical Relational Learning,” *ICIQ Spain*, pp. 2-15 (2016)

[34] Kotsiantis, S. B., Kanellopoulos, D., and Pintelas, P. E., “Data Preprocessing for

Supervised Learning," *International Journal of Computer Science* 2006; 1(2):111-7 (2006)

[35] Aloul, F. A., "The Need for Effective Information Security Awareness," *Journal of Advances in Information Technology* (2012)

[36] Chen, J., "Evaluation of Application of Ontology and Semantic Technology for Improving Data Transparency and Regulatory Compliance in the Global Financial Industry," *MIT* (2015)

[37] Schid, C. J., and Schultz, S., "Linking Deutsche Bundesbank company data using machine-learning-based classification," *in proceedings of the Second International Workshop on Data Science for Macro-Modelling*, Article 10, pp. 1-3 (2016)

[38] Condie, T., Mineiro, P., Polyzotis, N., and Weimer, M., "Machine Learning on Big Data," *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pp. 1242-1244 (2013)

[39] Marlin, S., "Bank Data Executives Split on BCBS 239 Sanctions," *Risk.net*, pp. 1-2 (2016)

[40] Yu, Q., and Shen, Y., "Research of Information Security Risk Prediction Based on Grey Theory and ANP," *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference* (IMCEC) (2017)

[41] Moro, S., Cortez, P., and Rita, P., "A Data-Driven Approach to Predict the Success of Bank Telemarketing," *Decision Support Systems, Elsevier*, 62:22-31 (2014)

[42] https://data.world/datasets/banking

[43] https://data.worldbank.org/

[44] https://data.world/datasets/banking

[45] Ghoting, A., Krishnamurthy, R., Pednault, E., Reinwald, B., Sindhwani, V., Tatikonda, S., Tian, Y., and Vaithyanathan, S., "SystemML: Declarative Machine Learning on MapReduce," *2011 IEEE 27th International Conference on Data Engineering*, pp. 231-242 (2011)

[46] Kulesza, T., Amershi, S., Caruana, R., Fisher, D., and Charles, D., "Structured Labeling to Facilitate Concept Evolution in Machine Learning," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3075-3084 (2014)

[47] Kotsiantis, S. B., Kanellopoulos, D., and Pintelas, P. E., "Data Preprocessing for Supervised Leaning," *International Journal of Computer Science*, Volume 1: Number 1,

2006 ISSN 1306-4428, pp. 111–117 (2006)

[48] Zhang, Y., Dai, H., Xu, C., Feng, J., Wang, T., Bian, J., Wang, B., and Liu, T.-Y., "Sequential click prediction for sponsored search with recurrent neural networks," In C. E. Brodley and P. Stone, editors, *AAAI*, pp. 1369-1375 (2014)

[49] Murad. A., and Pyun, J.Y., "Deep recurrent neural networks for human activity recognition," *Sensors* 17(11):2556 (2017)

[50] Weterings, K., Bromuri, S., and Eekelen, M. V., "Explaining Customer Activation with Deep Attention Models," *European Conference on Information Systems* (2019)

[51] Klaus, G., Rupesh, K., Srivastava, J. K., Bas, R. S., and Jürgen, S., "LSTM: A Search Space Odyssey," *IEEE Journal of Transactions on Neural Networks and Learning Systems*, Volume: 28, Issue: 10, IEEE, pp. 2222-2232 (2017)

[52] Lyu, Q., and Zhu, J., "Revisit Long Short-Term Memory: An Optimization Perspective," *Advances in Neural Information Processing Systems Workshop on Deep Learning and Representation Learning* (2014)

[53] Kotsiantis, S. B., "Supervised Machine Learning; A Review of Classification Techniques," *An International Journal of Computing and Informatics*: Volume 31, No. 3 and in: *Proceeding of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pp. 3-24 (2007)

[54] Guresen, E., and Kayakutlu, G., "Definition of artificial neural networks with comparison to other networks," *Journal of Procedia Computer Science*, Volume 3, pp. 426-433 (2011)

[55] Zhang, S., Li, X., Zong, M., Zhu, X., and Cheng, D., "Learning k for kNN Classification," *Journal of ACM Transactions on Intelligent Systems and Technology (TIST)*, Volume 8, Issue 3, Article No. 43 (2017)

[56] Längkvista, M., Karlssona, L., and Louta, A., "A Review of Unsupervised Feature Learning and Deep Learning for Time-Series Modeling," *Journal of Patter Recognition Letters*, Volume 42, pp. 11-24 (2014)

[57] Grossi, E., and Buscema, M., "Introduction to artificial neural networks," *European Journal of Gastroenterology & Hepatology*: Dec 2007: Volume 19: Issue 12, pp. 1046-1054 (2007)

[58] Shi, W., Zhu, Y., Zhang, J., Tao, X., Sheng, G., Lian, Y., Wang, G., & Chen, Y., "Improving power grid monitoring data quality: An efficient machine learning framework for missing data prediction," *IEEE 17th International Conference on High Performance Computing and Communications*, IEEE Computer Society, pp. 417-422 (2015)

[59] Mechelke, M., and Habeck, M., "Bayesian weighting of statistical potentials in nmr structure calculation," *PloS one* 9(6):e100197 (2014)

[60] Miller, B., and Rowe, D., "A Survey SCADA of and Critical Infrastructure Incidents," *Annual Conference on Research in Information*, pp. 51-56 (2012)

[61] Liu, Y., Sarabi, A., Zhang, J., and Naghizadeh, P., Karir, M., Bailey, M., and Liu, M., "Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents," *USENIX Security Symposium* (2015)

[62] Xiao, H., and Zhang, G., "The Queuing Theory Application in Bank Service Optimization," *International Conference on Logistics Systems and Intelligent Management* (2010)

[63] Tavana, M., Abtahi, AR., Di Caprio, D., and Poortarigh, M., "An Artificial Neural Network and Bayesian Network model for liquidity risk assessment in banking," *Journal of Neurocomputing 275*, pp. 2525-2554 (2018)

[64] Sun, J., and Chen, Y., "Building a Common Enterprise Technical Architecture for an Universal Bank," *International Conference on Management and Service Science* (2010)

[65] Kusic, D., Kephart, J. O., Hanson, J. E., Kandasamy, N., and Jiang, G., "Power and Performance Management of Virtualized Computing Environments via Lookahead Control," *Cluster Computing* (2009)

[66] Gogas, P., Papadimitriou, T., and Agrapetidou, A., "Forecasting bank failures and stress testing: A machine learning approach," *International Journal of Forecasting, Volume 34, Issue 3*, pp. 440-455 (2018)

[67] Tallon, P. P., "A Service Science Perspective on Strategic Choice, IT, and Performance in U.S. Banking," *Journal of Management Information Systems*, Vol 26: 4 (2010)

[68] Michel, S., "Analyzing service failures and recoveries: a process approach," *International Journal of Service Industry Management* (2001)

[69] Attigeri, G. V., Pai, M., and Pai, R. M., "Credit Risk Assessment Using Machine Learning Algorithms", *Advanced Science Letters, 23(4)*, pp. 3649–3653 (2017)

[70] Bhoj, P., Singhal, S., and Chutani, S., "SLA Management in Federated Environments", *Computer Networks*, Vol 35, Issue 1, pp. 5-24 (2001)

[71] Bumblauskas, D., Nold, H., Bumblauskas, P., and Igou, A., "Big data analytics: transforming data to action," *Business Process Management Journal* (2017)

[72] Choi, T. M., Chan, H. K., and Yue, X., "Recent development in big data analytics for business operations and risk management," *In: IEEE Trans. Cybern*, 47(1), pp 81–92 (2017)

[73] Krishna D., "Big data in risk management," *Journal of Risk Management in Financial Institutions* 9:46–52 (2016)

[74] Aebi, V., Sabato, G., and Schmid, M., "Risk management, corporate governance, and bank performance in the financial crisis," *Journal of Banking and Finance* 32:3213–3226 (2012)

[75] APRA, "Prudential Practice Guide, Draft CPG 234 Informaiton Security," *Australian Prudential Regulation Authority (APRA)* (2019)

[76] Frydenberg, J, "Restoring Trust in Australia's Financial System", *Australian Government, The Treasury*, pp. 3-42 (2019)

[77] Frost, J, "APRA Rejected CBA Home Loan Data as Inaccurate and Incomplete," *Financial Review: Business, Banking & Finance*, pp. 1-1 (2018)

[78] Yeates, C., "Banks Dive as UBS Raises Home Loan Concerns," *Sydney Morning Herald: Banking & Finance*, pp. 1-2 (2018)

[79] Grubbb, B., and Yeates, C., "Almost 100,000 Australians' Private Details Exposed in Attack on Westpac's PayID," *Sydney Morning Herald* (2019)

[80] Crozier, R., "NAB is Building a Central Analytics Hub," *IT NEWs*, pp. 1-1 (2017)

[81] Eyers, J., "CBA Want to Use AI to Tackle Fraud and Cyber Attacks," *Business Insider*, pp 1-2 (2016)

[82] H*ä*rle, P., Havas, A., Kremer, A., Rona, D., and Samandari, H., "The Future of Bank Risk Management," *Mckinsey&Company*, pp 1-32 (2015)

[83] Webber Insurance Services, "The Complete List of Data Breaches in Australia for 2018, 2019 and 2020," https://www.webberinsurance.com.au/data-breaches-list, Webber Insurance Services (WIS) (2020)

[84] Lucic, M., Faulkner, M., Krause, A., Feldman, D., "Training Gaussian Mixture Models

at Scale via Coresets," *Journal of Machine Learning Research* 18(1):5885–5909 (2018)

[85] Nguyen, V. A., Boyd-Graber, J., Resnik, P., Cai, D. A., Midberry, J. E., Wang, Y.: "Modeling topic control to detect influence in conversations using nonparametric topic models," *ML*, 95(3): p.p. 381–421 (2013)

[86] Liu, S. Z., Zheng Z. Z., Wu F., Tang, S. J., Chen, G. H., "Context-Aware Data Quality Estimation in Mobile Crowdsensing," *IEEE Conference on Computer Communications* (INFOCOM) (2017)

[87] Yao, L., and Ge, Z. Q. "Scalable Semisupervised GMM for Big Data Quality Prediction in Multimode Processes", *IEEE Transactions on Industrial Electronics* 66(5), pp. 3681-3692 (2018)

[88] Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J., "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," *In International Conference on Machine Learning* (2006)

[89] Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., and Courville, A., "Describing videos by exploiting temporal structure," In*: Proceedings of the IEEE International Conference on Computer Vision*, pp 4507–4515 (2015)

[90] Cui, Z. Y., Ke, R. M., and Wang, Y. H, "Deep Stacked Bidirectional and Unidirectional LSTM Recurrent Neural Network for Network-wide Traffic Speed Prediction," *International Workshop on Urban Computing (UrbComp) with the ACM SIGKDD*, *arXiv preprint arXiv*, pp. 2-8 (2016)

[91] Zhai, S., Chang, K.-h., Zhang, R., and Zhang, Z. M., "Deepintent: Learning attentions for online advertising with recurrent neural networks," *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1295–1304 (2016)

[92] Luong, M. T., Pham, H., and Manning, C. D., "Effective approaches to attention-based neural machine translation," *EMNLP* (2015)

[93] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y., "Show, attend and tell: Neural image caption generation with visual attention," *In ICML*, pp. 2048-2057 (2015)

[94] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I, "Attention is all you need," *In NIPS*, pp. 6000–6010 (2017)

[95] Zhang, H., Goodfellow, I. J., Metaxas, D. N., and Odena, A., "Self-attention generative adversarial networks," *CoRR* abs/1805.08318 (2018)

[96] Chen, T., Yin, H. Z., Chen, H. X., Wu., L., Wang, H., Zhou, X. F., and Li, X., "TADA: Trend Alignment with Dual-Attention Multi-task Recurrent Neural Networks for Sales Prediction," *International Conference on Data Mining* (2018)

[97] Li, X. P., Song, J. K., Gao, L. L., Liu, X. L., Huang, W. B., He, X. N., and Gan, C., "Beyond RNNs: Positional Self-Attention with Co-Attention for Video Question Answering", *The 33rd AAAI Conference on Artificial Intelligence* (2019)

[98] Kaiser, M., Klier, M., and Heinrich, B., "How To Measure Data Quality?-a Metric-Based Approach," *ICIS 2007 Proceedings* pp. 108 (2007)

[99] Liu, J., Wang, G., Hu, P., Duan, L-Y, and Kot A. C, "Global Context-Aware Attention LSTM Networks for 3D Action Recognition," *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1647-1656 (2017)

[100]Bahdanau D., Cho, K., and Bengio, Y. S., "Neural machine translation by jointly learning to align and translate," *In ICLR* (2015)

[101]Yang, Z. C., Yang, D. Y., Dyer, C., He, X. D., Smola, A. J., and Hovy, E. H., "Hierarchical Attention Networks for Document Classification," *In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480-1489 (2016)

[102]Ma, F. L., Chitta, R., Zhou, J., You, Q. Z., Sun, T., and Gao, J., "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," *In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *KDD* '17, pp. 1903–1911 (2017)

[103]Haque, A., Alahi, A., and Li, F. F., "Recurrent Attention Models for Depth-Based Person Identification," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1229-1238 (2016)

[104]Pan, P. B., Xu, Z. W, Yang, Y., Wu, F., and Zhuang, Y. T., "Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1029-1038 (2016)

[105]Du, XQ., Yin, SS., Tang, RJ., Zhang, Y., and Li, S., "Cardiac-DeepIED: Automatic

Pixel-Level Deep Segmentation for Cardiac Bi-Ventricle Using Improved End-to-End Encoder-Decoder Network," *IEEE Journal of Transactional Engineering in Health and Medicine*, Vol.7, Art.1900110 (2019)

[106]Chawla, A., Lee, B., Fallon, S., and Jacob, P., "Host based intrusion detection system with combined cnn/rnn model," *In Proceedings of Second International Workshop on AI in Security* (2018)

[107]Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lo, P., and Bengio, Y., "Graph attention networks," *in ICLR* (2018)

[108]Riskstaff. "Top 10 operational risks for 2019," *in Risk.net* (2019)

[109]BIS. "OPE25 - Standardized approach," *The Bank for International Settlements (BIS)* (2019)

[110]Wang, C., Chi, C.H., Zhou, W., and Wong, R., "Coupled Interdependent Attribute Analysis on Mixed Data," *In: Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015)

[111]Zaidi, N. A., Cerquides, J., Carman, M. J., and Webb, G. I, "Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting," *Journal of Machine Learning Research*, pp. 1947-1988 (2013)

[112]Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H., and Chawla, N.V., "A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data," *in AAAI* (2019)

[113]Margarit, H., and Subramaniam, R., "A Batch-Normalized Recurrent Network for Sentiment Classification," *Advances in neural information processing systems* (2016)

[114] Xiao, J., Ye, H., He, X., Zhang, H., Wu, F., and Chua, T-S., "Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks," *in IJCAI*, pp. 3119-3125 (2017)

[115]You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J., "Image Captioning with Semantic Attention," *in CVPR*, pp. 4651-4659 (2016)

[116]Riemer, M., Vempaty, A., Calmon, F. P., Heath III, F.F., Hull, R., and Khabiri, E., "Correcting forecasts with multifactor neural attention," *in ICML*, pp. 3010–3019 (2016)

[117] Dietterich, T. G., "Machine Learning for Sequential Data: A Review," Springer-Verlag Berlin Heidelberg, *SSPR /SPR 2002: Structural, Syntactic, and Statistical Pattern*

Recognition, pp. 15-30 (2002)

[118]Dennis, D., Pabbaraju, C., Simhadri, H.V., Jain, P., "Multiple Instance Learning for Efficient Sequential Data Classification on Resource-constrained Devices," *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada* (2018)

[119]Dachraoui, A., Bondu, A., and Cornu*é*jols, A., "Early Classification of Time Series as a Non Myopic Sequential Decision Making Problem," *ECML PKDD 2015: Machine Learning and Knowledge Discovery in Databases*, pp. 433-447 (2015)

[120]APRA, "Prudential Practice Guide CPG 235 - Managing Data Risk", *Australian Prudential Regulation Authority (APRA)*, pp. 1-13 (2013)

[121]Bakker, B., and Heskes T., "Task clustering and gating for bayesian multitask learning," *Journal of Machine Learning Research* 4(May): p.p. 83–99 (2003)

[122]Mohtarami, M., Baly, R., Glass, J., Nakov, P., Mrquez, L., and Moschitti, A., "Automatic stance detection using end-to-end memory networks," *In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 767-776 (2018)

[123]Taleb, I., El Kassabi, HT., Serhani, MA., Dssouli, R., Bouhaddioui, C., "Big data quality: A quality dimensions evaluation," *In: 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*, IEEE, pp. 759–765 (2016)

[124]Ge, M., Helfert, M., and Jannach, D., "Information quality assessment: Validating measurement dimensions and processes," (2011)

[125]Tesfay, WB., Hofmann, P., Nakamura, T., Kiyomoto S. and Serna, J., "I Read but Don't Agree: Privacy Policy Benchmarking using Machine Learning and the EU GDPR," *Companion Proceedings of the Web Conference*, pp. 163-166 (2018)

[126]Li, H., Shen, Y. Y., and Zhu, Y. M., "Stock Priced Prediction Using Attention-based Multi-Input LSTM," *in Proceedings of Machine Learning Research*, pp. 454-469 (2018)

[127]Naumann, F., Leser, U., and Freytag, JC., "Quality-driven integration of heterogeneous information systems," V*ery Large Data bases* (1999)

[128]Xu, H., Nord, JH., Brown, N., and Nord, GD., "Data quality issues in implementing an

ERP," *Industrial Management & Data Systems*, 102, 47, (2002)

[129]Singh, R., and Singh, K., "A descriptive classification of causes of data quality problems in data warehousing," *International Journal of Computer Science Issues*, 7, pp. 41-50 (2010)

[130]Yang, Y., Wang, X., Guan, T., Shen, J., and Yu, Li., "A Multi-dimensional Image Quality Prediction Model for User-generated Images in Social Networks," *Information Science 281*, pp. 601-610 (2014)

[131]Das, S., and Saha, B., "Data Quality Mining using Genetic Algorithm," *International Journal of Computer Science and Security*, *IJCSS*, Vol (3): Issue 2, pp. 105-112 (2009)

[132]Helfert, M., Foley, O., Ge, M., and Cappiello, C., "Limitations of Weighted Sum Measures for Information Quality," *In: Proceedings of the Fifteenth Americas Conference on Information Systems* (2009)

[133]Jayawardene, V., Sadiq, S., and Indulska, M., "An Analysis of Data Quality Dimensions," ACIS 2013, *24th Australasian Conference on Information Systems*, pp. 1-11 (2013)

[134]Hartig, O., and Zhao, J., "Using Web Data Provenance for Quality Assessment," *CEUR Workshop Proceedings* (2009)

[135]Manzoor, A., Truong, H. L., and Dustdar, S, "Quality of Context: models and applications for context-aware systems in pervasive environments," *The Knowledge Engineering Review*, Vol. 29:2: pp. 154-170. *Cambridge University Press* (2014)

[136]Kaiser, M., "A Conceptual Approach to Unify Completeness, Consistency, and Accuracy as Quality Dimensions of Data Values," *European and Mediterranean Conference on Information Systems* (2010)

[137]Talhofer, V., Hošková, S., and Hofmann, "A. Improvement of Digital Geographic Data Quality," *International Journal of Production Research* 50 (17), pp. 4846-4859 (2012)

[138]Emran, N. A., Embury, S., Missier, P., Mat Isa, M. N., and Muda, A. K, "Measuring Data Completeness for Microbial Genomics Database," *Asian Conference on Intelligent Information and Database Systems*, pp. 186-195 (2013)

[139]Cappiello, C., Daniel F., Matera, M., and Pautasso, C., "Information Quality in Mashups," *Internet Computing* 14(4), pp. 14-22 (2010)

[140]Galárraga, L., Razniewski, S., Amarilli, A., and Suchanek, F. M., "Predicting

Completeness in Knowledge Bases," *International Conference on Web Search and Data Mining*, pp. 375-383 (2017)

[141]Martin, L., Suneel, Sharma., and Koilakuntla, M, "Machine learning in banking risk management: A literature review," *Risks Journal*, Vol 7, No. 1, Multidisciplinary Digital Publishing Institute, pages 29 (2019)

[142]Dey, D., "Growing Importance of Machine Learning in Compliance and Regulatory Reporting," *European Journal of Multidisciplinary Studies*, September-December 2017, Volume 2, Issue 7 (2017)

[143]Venkataraman, S., Yang, ZH., Franklin, M., Recht, B., and Stoica, I., "Ernest: Efficient Performance Prediction for Large-Scale Advanced Analytics," *Proceedings of the 13th USENIX Symposium on Networked Systems Design and Implementation (NSDI '16)*, ISBN 978-1-931971-29-4 (2016)

[144] Gudivada, V., Apon, A., and Ding, J., "Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations," *International Journal on Advances in Software*, Vol. 10, Number 1, pp. 1-20 (2017)

[145] Asheibi, A., Stirling, D., and Soetanto, D., "Analyzing Harmonic Monitoring Data Using Supervised and Unsupervised Learning," *IEEE Transactions on Power Delivery*, Vol. 24, No. 1 (2009)

[146]Baldi, P., "Autoencoders, Unsupervised Learning, and Deep Architectures," *JMLR: Workshop and Conference Proceedings* 27: pp. 37–50 (2012)

[147]Kaelbling, L. P., Littman, M. L., and Moore, A. W., "Reinforcement Learning: A Survey," *Journal of Artificial Intelligence Research 4*, pp. 237-285 (1996)

[148]Dietterich, T. G., "Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition," *Journal of Artificial Intelligence Research*, 13: pp. 227–303 (2000)

[149]Sutton, R. S., Precup, D., and Singh, S. P., "Between MDPs and semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning," *Artificial Intelligence*, 112(1-2): pp. 181– 211 (1999)

[150]Taylor, M. E., and Stone, P., "Transfer Learning for Reinforcement Learning Domains: A Survey," *Journal of Machine Learning Research 10*, pp. 1633-1685 (2009)

[151]Weller, A., "Data Governance: Supporting Datacentric Risk Management," *Journal of*

Securities Operations and Custody, Volume 1, Number 3, pp. 250-262(13) (2008)

[152]Lemieux, V., "Data Governance, Analytics and Life Cycle Management," *Financial Analysis and Risk Management*, Springer, ISBN 978-3-642-32231-0 (2012)

[153]Moghe, P., "Controlling Risk with a Data Governance Framework: How People, Processes, and Technology Must Come Together for an Entity to Properly Use and Manage Data," *Bank Accounting & Finance*, Vol. 22, Issue 2 (2009)

[154]Schilling, R., Aier, S., Winter, R., and Haki, K., "Design Dimensions for Enterprise-Wide Data Management: A Chief Data Officer's Journey," *Proceedings of the 53rd Hawaii International Conference on System Sciences* (2020)

[155]Ngu, H. C. V., and Huh, J. H., "B+-tree construction on massive data with Hadoop", *Clustering Computing 22*, pp. 1011-1021 (2019)

[156]Alexander, L., Das, S. R., Ives, Z, Jagadish, H. V., and Monteleoni, C., "Research Challenges in Financial Data Modeling and Analysis," *Big Data*, Vol. 5, No. 3 (2017)

[157]Chan, K., Marcus, K., Scott, L., Hardy, R., "Quality of information approach to improving source selection in tactical networks," *In: 2015 18th International Conference on Information Fusion (Fusion)*, IEEE, pp. 566–573 (2015)

[158]Good, I. J., "Probability and the Weighing of Evidence," London, Charles Grin (1950)

[159]Zhang, T., "An introduction to Support Vector Machines and Other Kernel-based Learning Methods," *AI Magazine*, Vol. 22, Number 2, pp. 103 (2001)

[160]Archer, K. J., and Kimes, R. V., "Empirical Characterization of Random Forest Variable Importance Measures," *Computational Statistics & Data Analysis*, Vol. 52, Issue 4, pp. 2249-2260 (2008)

[161]Bailey, T. L., and Elkan, C., "Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization," *Machine Learning*, 21, pp. 51-80 (1995)

[162]Liu, J., Li, J., Li, W., and Wu J., "Rethinking Big Data: A Review on the Data Quality and Usage Issues," *ISPRS Journal of Photogrammetry and Remote Sensing* 115: pp. 134–142 (2016)

[163]Vizhi, J.M., and Bhuvaneswari, D.T., "Data quality measurement with threshold using genetic algorithm," *International Journal Eng Res Appl* 2(4): pp. 1197–120 (2012)

[164]Mendes, P.N., M*ü*hleisen, H., and Bizer, C., "Sieve: linked data quality assessment and fusion," *In: Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pp. 116–123 (2012)

[165]Lin, H., Hu, J., Tian, Y., Yang, L., and Xu, L., "Toward better data veracity in mobile cloud computing: A context-aware and incentive-based reputation mechanism," *Information Sciences* 387: pp. 238–253 (2017)

[166]Saha, B., and Srivastava, D., "Data quality: The other face of big data," *In: 2014 IEEE 30th International Conference on Data Engineering*, pp. 1294–1297 (2014)

[167]Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F., Grafberger, A., "Automating largescale data quality verification," *Proceedings of the VLDB Endowment* 11(12): pp. 1781–1794 (2018)

[168]Hunt, L., and Jorgensen, M., "Clustering Mixed Data," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1(4): pp. 352–361 (2011)

[169]Kaiser, M., "A conceptional approach to unify completeness, consistency and accuracy as quality dimensions of data values," *In: European and Mediterranean Conference on Information Systems* (2010)

[170]O'Hagan, A., Murphy, TB., Scrucca, L., and Gormley, IC., "Investigation of parameter uncertainty in clustering using a gaussian mixture model via jackknife, bootstrap and weighted likelihood bootstrap," *Computational Statistics* 34(4): pp. 1779–1813 (2019)

[171]Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B., "Attention-based bidirectional long shortterm memory networks for relation classification," *In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), pp. 207–212 (2016)

[172]Mun, J., Cho, M., and Han, B., "Text-Guided Attention Model for Image Captioning," *in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (AAAI-17) (2017)

[173]Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S., Hitzler, P., "Quality assessment methodologies for linked open data," Submitted to Semantic Web Journal 1:1–5, IOS Press (2013)

[174]Kingston, J., "Using Artificial Intelligence to Support Compliance with the General Data Protection Regulation," *Artificial Intelligence and Law*, Springer (2017)

[175]Even, A., and Shankaranarayanan, G., "Understanding impartial versus utility-driven quality assessment in large datasets," *In: ICIQ*, pp 265–279 (2007)

[176]Singh, K., and Best, P., "Anti-Money Laundering: Using Data Visualization to Identify

Suspicious Activity," *International Journal of Accounting Information Systems*, ACCINF-00418 (2019)

[177]Ye, Y., He, R., and Cheng, X., "Research on Data Quality Management and Data Audit & Monitor in Telecom Industry," *Telecommunications Science* 28 (2), pp. 1-6 (2012)

[178]Emeka-Nwokeji, N. A., "Repositioning Accounting Information Systems Through Effective Data Management A Framework for Reducing Costs and Improving Performance," *International Journal of Scientific & Technology Research,* Vol. 1, Issue 10 (2012)

[179]Chieu, T. C., Singh, M., Tang, C. Q., Viswanathan, M., and Gupta, A., "Automation System for Validation of Configuration and Security Compliance in Managed Cloud Services," *IEEE International Conference on e-Business Engineering* (2012)

[180]Bajaber, W., AIQulaity, M., and Zafar, A., "An Overview of Strategic Information Systems Planning in Banking Sectors: A case study of Riyadh Bank of KS," *International Journal of Computer Applications*, Vol. 144, No. 7 (2016)

[181]Boehm, J., Curcio, N., Merrath, P., Shenton, L., and Stahle, T., "The Risk-Based Appraoch to Cybersecurity, " Mckinsey & Company (2019)

[182]Stollenga, MF., Masci, J., Gomez, F., and Schmidhuber, J., "Deep networks with internal selective attention through feedback connections," *In: Advances in Neural Information Processing Systems*, pp 3545–3553 (2014)

[183]ISACA., "Risk and Compliance – For Better or Worse?," *ISACA Journal Archives* (2013)

[184]Graves, A., Mohamed, Ar., Hinton, G., "Speech recognition with deep recurrent neural networks," *In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 6645– 6649 (2013)

[185]Upadhyay, A., "How Long Does It Take to Complete a Big Data or Data Science Project in Real Time on Large Data Sets?," *Quora*, pp 1 (2017)

[186]Abdel-Nasser, M., and Mahmoud, K., "Accurate Photovoltaic Power Forecasting Models Using Deep LSTM-RNN," *Neural Computing and Applications*, Springer, pp. 1-14 (2017)

[187]Wang, Y., and Zang, J., "Keyword Extraction from Online Product Reviews Based on Bi-directional LSTM Recurrent Neural Network," *2017 IEEE International Conference on*

Industrial Engineering and Engineering Management (IEEM), pp. 2241-2245 (2017)

[188]Ostemyer, J., and Cowell, L., "Machine Learning on Sequential Data Using a Recurrent Weighted Average," *arXiv preprint arXiv*:1703.01253 (2017)

[189]Yildirim, O., "A Novel Wavelet Sequence Based on Deep Bidirectional LSTM Network Model for ECG Signal Classification," *Computers in Biology and Medicine*, 2018. 96: pp. 189-202 (2018)

[190]Liebergen, B V., "Machine learning: a revolution in risk management and compliance?," *Journal of Financial Transformation*, The Capco Institute (2017)

[191]Frühwirth-Schnatter, S., and Kaufmann, S., "How Do Changes in Monetary Policy Affect Bank Lending? An Analysis of Austrian Bank Data," *Journal of Applied Econometrics* 21(3): pp. 275–305 (2006)

[192]Pipino, L. L., Lee, Y. W., and Wang, R. Y., "Data Quality Assessment," *Communications of the ACM*, 45, 4, pp. 211-218 (2002)

[193]Wong, Eric., and Cho, H. H., "A liquidity risk stress-testing framework with interaction between market and credit risks," *Hong Kong Monetary Authority – Research Department*, *Working Paper 06/ 2009*, pp. 1-33 (2009)

[194]Chordia, T., Roll, R., and Subrahmanyam, A., "Commonality in Liquidity," *Journal of Financial Economics*, pp. 3-28 (2000)

[195]Regina, E. T., Edward, Y. B., and Gideon, E. W., "A Machine Learning Approach for Predicting Bank Credit Worthiness," *Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, ISBN: 978-1-4673-9187-0, IEEE (2016)

[196]Siddayao, G., Valdez, S., and Fernandez, P., "Analytic Hierarchy Process (AHP) in Spatial Modeling for Floodplain Risk Assessment," *International Journal of Machine Learning and Computing*, 4(5), pp. 450-457 (2014)

[197]Kaya, M. E., Gurgen, F., Okay, N., "An Analysis of Support Vector Machines for Credit Risk Modeling," In Soares, *Frontiers in Artificial Intelligence and Applications*, IOS Press, pp. 27-35 (2008)

[198]Grolinger, K., Capretz, M. AM., and Seewald, L., "Energy Consumption Prediction with Big Data: Balancing Prediction Accuracy and Computational Resources," *IEEE International Congress on Big Data*, pp. 157-164 (2016)

[199]Ruta, D., "Automated Trading with Machine Learning on Big Data," *IEEE*

International Congress on Big Data, pp. 824-830 (2014)

[200]Serhani, M. A., Kassabi, H. T. E1., Taleb, I., and Nujum, A., "An Hybrid Approach to Quality Evaluation Across Big Data Value Chain," *IEEE International Congress on Big Data*, pp. 418- 425 (2016)

[201]AI-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., and Taha, K., "Efficient Machine Learning for Big Data: A Review," *Big Data Research* 2(3), pp. 87-93 (2015)

[202]Lv, Y. S., Duan, Y. J., Kang, W. W., Li, Z. X., and Wang, F. Y., "Traffic Flow Prediction with Big Data: A Deep Learning Approach," *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, No. 2, pp. 865-873 (2015)

[203] Chen, H. M., Zhao, H., Shen, J., Zhou, R., and Zhou, Q. G., "Supervised Machine Learning Model for High Dimensional Gene Data in Colon Cancer Detection," *IEEE International Congress on Big Data*, pp. 134-141 (2015)

[204]Shevchenko, P. V., and Wüthrich, M. V., "The Structural Modelling of Operational Risk via Bayesian Inference: Combining Loss Data with Expert Opinions," *Journal of Operational Risk* 1(3), pp. 3-26 (2006)

[205]Darema, F., "Dynamic data driven applications systems: A new paradigm for application simulations and measurements," *In: International Conference on Computational Science*, Springer, pp. 662–669 (2004)

[206]Samuel, JC., Sankhulani, E., Qureshi, JS., Baloyi, P., Thupi, C., Lee, CN., Miller, WC., Cairns, BA., Charles, AG., "Under-reporting of road traffic mortality in developing countries: application of a capture-recapture statistical model to refine mortality estimates," *PloS on*e 7(2):e31091 (2012)

[207]Wichern, G., and Lukin, A., "Low-latency Approximation of Bidirectional Recurrent Networks for Speech Denosing," *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, pp. 66-70 (2017)

[208]Allen, L., and Bali, T. G., "Cyclicality in Catastrophic and Operational Risk Measurements," *Journal of Banking & Finance*, Volume 31, Issue 4, pp. 1191-1235 (2007)

[209]Dan, R., and David, S, "Risk Factor Contributions in Portfolio Credit Risk Models," *Journal of Banking & Finance* (34), pp. 336-349. http://dx.doi.org/10.1016/j.jbankfin.2009.08.002 (2010)

[210]Bolt, W., De Haan, L., Hoeberichts, M., Van Oordt, MR., and Swank, J., "Bank profitability during recessions," *Journal of Banking & Finance* 36(9): pp. 2552–2564 (2012)

[211]Song, X., Kanasugi, H., and Shibasaki, R., "Deeptransport: Prediction and Simulation of Human Mobility and Transportation Mode at a Citywide Level," *In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2618–2624 (2016)

[212]Yu, R., Gao, J., Yu, M., Lu, W., Xu, T., Zhao, M., Zhang, J., Zhang, R., and Zhang, Z., "LSTM-EFG for Wind Power Forecasting Based on Sequential Correlation Features," *Future Generation Computer Systems* 93:33–42 (2019)

[213]Zhou, P., Shi, W., Tian, J., Qi, Z. Y., Li, B. C., Hao. H. W., and Xu, B., "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification," *in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Germany, pp. 207-212 (2016)

[214]Hu, T., Zhang, P., Zhang, X., and Dai, H., "Gender Differences in Internet Use: a Logistic Regression Analysis," *AMCIS 2009 Proceedings* pp. 300 (2009)

[215]Cheng, TH., Lan, CW., Wei, CP., and Chang, H., "Cost-sensitive Learning for Recurrence Prediction of Breast Cancer," *In: PACIS*, pp. 118 (2010)

[216]Zhou, C., Bai, J., Song, J., Liu, X., Zhao, Z., Chen, X., and Gao, J., "Atrank: An attention-based user behavior modeling framework for recommendation," *In: Thirty-Second AAAI Conference on Artificial Intelligence* (2018)

[217]Cho, K., Van Merriënboer B, Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y., "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv preprint arXiv*:14061078 (2014)

[218]Sutskever, I., Vinyals, O., and Le, QV., "Sequence to Sequence Learning with Neural Networks," *In: Advances in Neural Information Processing Systems*, pp. 3104–3112 (2014)

[219]Graves, A., and Schmidhuber, J., "Framewise Phoneme Classification with Bidirectional LSTMNetworks," *in Proceedings of International Joint Conference on Neural Network*, Canada, IEEE, Vol. 4, pp. 2047-2051 (2005)

[220]Bengio, Y., Simard, P., and Frasconi, P., "Learning Long-Term Dependencies with

Gradient Descent is Difficult," *IEEE Transactions on Neural Networks* 5(2): pp. 157–166 (1994)

[221]Hochreiter, S., and Schmidhuber, J., "Long Short-Term Memory," *Neural Computation* 9(8): pp. 1735–1780 (1997)

[222]Cross, J., and Huang, L., "Incremental Parsing with Minimal Features Using Bi-directional LSTM," *arXiv preprint arXiv*:160606406 (2016)

[223]Irie, K., Lei, Z., Deng, L., Schlüter, R., and Ney, H., "Investigation on Estimation of Sentence Probability by Combining Forward, Backward and Bidirectional LSTM-RNNs," *In: INTERSPEECH*, pp. 392–395 (2018)

[224]Hwang, S., and Satchell, S. E., "Modelling Emerging Market Risk Premia Using Higher Moments," *International Journal of Finance & Economics* 4(4), pp. 271-296 (1999)

[225]Sun, Q., Lee, S., and Batra, D., "Bidirectional Beam Search: Forward-Backward Inference in Neural Sequence Models for Fill-in-the-Blank Image Captioning," *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6961–6969 (2017)

[226]Werbos, PJ., "Backpropagation Through Time: What It Does and How To Do it," *Proceedings of the IEEE* 78(10): pp. 1550–1560 (1990)

[227]Li, J., Luong, MT., and Jurafsky, D., "A Hierarchical Neural Autoencoder for Paragraphs and Documents," *arXiv preprint arXiv*:150601057 (2015)

[228]Xu, H., and Al-Hakim, L., "Criticality of Factors Affecting Data Quality of Accounting Information Systems," In Wang, R. Y., Pierce, E. M., Madnick. and Fisher C. W. (Eds.), *Information Quality*, pp. 197-214. New York; M.E. Sharpe (2005)

[229]Rai P. K., and Bunkar, R. K., "Study of Security Risk and Vulnerabilities of Cloud Computing," *International Journal of Computer Science and Mobile Computing*, Vol. 3 Issue. 2 (2014)

[230]Popovic, K., and Hocenski, Z., "Cloud Computing Security Issues & Challenges," *33rd International Convention MIRPO*, IEEE, pp. 344-349 (2010)

[231]Kamongi, P., Kotikela, S., Kavi, K., Gomathisankaran, M., and Singhal, A., "Vulcan: Vulnerability Assessment Framework for Cloud Computing," IEEE *International Conference on Software Security and Reliability* (2013)

[232]Peotta, L., Holtz, M. D., David, B. M., Deus, F. G., Sousa, RT. De., "A Formal Classification of Internet Banking Attacks and Vulnerabilities," *International Journal of Computer Science & Information Technology*, Vol 3, No 1 (2011)

[233]Joshi1, K. P., Mistry, N. R., and Dahiya, M. S., "A Comprehensive Study of Vulnerability Assessment Techniques of Existing Banking Apps," *International Journal for Research in Applied Science & Engineering Technology*, Vol 6, Issue IV (2018)

[234]Ward, R., Wu, X., and Bottou, L., "Adagrad Stepsizes: Sharp Convergence over Nonconvex Landscapes, from My Initialization," *arXiv preprint arXiv*: 1806.01811 (2018)

[235]Ruder, S., "An overview of gradient descent optimization algorithms", *arXiv preprint arXiv*:1609.04747 (2017)

[236]Shin, S., and Samanlioglu, F., and Cho, B. R., and Wiecek, M. M., "Computing trade-offs in robust design: perspectives of the mean squared error," *Computers & Industrial Engineering*, Elsevier, Vol 60, No. 2, pp. 248-255 (2011)

[237]Laub, J., and Muller, K-R., "Feature Discovery in Non-Metric Pairwise Data," *Journal of Machine Learning Research*, Vol. 5, pp. 801-818 (2004)

[238]Earp, JB., and Payton, FC., "Information Privacy in the Service Sector: An Exploratory Study of Health Care and Banking Professionals," *Journal of Organizational Computing and Electronic Commerce* 16(2): pp. 105–122 (2006)

[239]De Amicis, F., Barone, D., and Batini, C., "An Analytical Framework to Analyze Dependencies Among Data Quality Dimensions," *In: ICIQ*, pp. 369–383 (2006)

[240]Cai, L., and Zhu, Y., "The challenges of data quality and data quality assessment in the big data era," *Data Science Journal* 14 (2015)

[241]Margarit, H., and Subramaniam, R., "A batchnormalized recurrent network for sentiment classification," *Advances in Neural Information Processing Systems* pp. 2–8 (2016)

[242]Shi, W., Zhu, Y., Zhang, J., Tao, X., Sheng, G., Lian, Y., Wang, G., and Chen, Y., "Improving power grid monitoring data quality: An efficient machine learning framework for missing data prediction," *In: 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*, IEEE, pp. 417–422 (2015)

[243]BIS, “Statistical thinking & methodology: pillars for quality in the big data era, basel committee on banking supervision,” Tech. rep., Bank for International Settlements (BIS) (2016)

[244]Xiao, Y., Xiao, J., Liu, J., and Wang, S., “A multiscale modeling approach incorporating arima and anns for financial market volatility forecasting,” *Journal of Systems Science and Complexity* 27(1): pp. 225–236 (2014)

[245]Rather, AM., Agarwal, A,, and Sastry, V., “Recurrent neural network and a hybrid model for prediction of stock returns,” *Expert Systems with Applications* 42(6): pp. 3234–3241 (2015)

[246]Schelter, S., Lange D., Schmidt P., Celikel M., and Biessmann, F, “Automating large-scale data quality verification,” *International Conference on Very Large Databases* 11(12), pp. 1781-1794 (2018)

[247]Barakat, A., and Hussainey, K., “Bank governance, regulation, supervision, and risk reporting: Evidence from operation risk disclosures in European banks,” *International Review of Financial Analysis*, 30, pp. 254-273 (2013)

[248]Hevner, A., and Chatterjee, S., “Design science research in information systems,” *In: Design Research in Information Systems. Integrated Series in Information Systems*, vol. 22, pp. 9–22. Springer, US (2010)

[249]Wang, R., and Li, J., “Bayes Test of Precision, Recall, and F1 Measure for Comparison of Two Natural Language Processing Models,” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pages 4135–4145 (2019)

[250]Yang, Y.P.O, Shieh, H. M., and Tzeng, G. H., “A VIKOR technique based on DEMATEL and ANP for information security risk control assessment,” *Information Sciences* 232, Elsevier, pp. 482-500 (2013)

[251]Irie, K., Lei, Z., Deng, L., Schl*ü*ter, R., and Ney, H., “Investigation on estimation of sentence probability by combining forward, backward and bidirectional lstm-rnns” *In: INTERSPEECH*, pp 392–395 (2018)

[252]Contissa, G., Docter, K., Lagioa, F., Lippi, M., Micklitz, H. W., Palka, P., Sartor, G., and Torroni, P., “CLAUDETTE meets GDPR Automating the Evaluation of Privacy Policies using Artificial Intelligence,” *European Consumer Organization*, pp. 1-59 (2018)

[253]Westphal, P., Fernandez, JD., Kirrane, S., and Lehmann, J., "SPIRIT: A Semantic Transparency and Compliance Stack," *CEUR Workshop Proceedings* (2018)

[254]Jassal, R. K., and Sehgal, R. K., "Study of Online Banking Security Mechanism in India: Take ICICI Bank as an Example," *IOSR Journal of Computer Engineering*, Vol. 13, Issue 1, pp. 114- 121 (2013)

[255]Abukhzam, M., and Lee, A., "Factors Affecting Bank Staff Attitude Towards e-Banking Adoption in Libya," *Electronic Journal on Information Systems in Developing Countries*, 42, 2, pp. 1-15 (2010)

[256]Farn, K-J., Lin, S-K., Fung, and A. R-W., "A Study on Information Security Management System Evaluation—Assets, Threat and Vulnerability," *Computer Standards & Interfaces*, Vol. 26, Issue 6, pp. 501-513 (2004)

[257]Katkar, A. S., and Kulkarni, R.B., "Web Vulnerability Detection and Security Mechanism," *International Journal of Soft Computing and Engineering*, ISSN: 2231-2307, Vol-2, Issue-4 (2012)

[258]Brooks, T. T., Caicedo, C., and Park, J. S., "Security Vulnerability Analysis in Virtualized Computing Environments," *International Journal of Intelligent Computing Research* (2012)

[259]Grüttner, V., Pinheiro, F., and Itaborahy, A., "IT Governance Implementation – Case of a Brazilian Bank," *AMCIS* (2010)

[260]Nedelcu, B., Stefanet, M-E., Tamasescu, L-R., Tintoiu, S-E., and Vezeanu, A., "Cloud Computing and its Challenges and Benefits in the Bank System," *Database Systems Journal*, Vol. VI, Issue 1 (2015)

[261]Carroll, M., Merwe, A. V. D., and Kotzé, P., "Secure Cloud Computing, Benefits, Risks and Controls," IEEE *Information Security for South Africa* (2011)

[262]Schaad, A., Moffett, J., and Jacob, J., "The Role-Based Access Control System of a European Bank: A Case Study and Discussion," *Proceedings of the Sixth ACM Symposium on Access Control Models and Technologies*, pp. 3-9 (2001)

[263]Venkatraman, S., "Biometrics in Banking Security: a Case Study," *Management and Computer Security* (2008)

[264]Baldwin, L. P., Irani, Z., and Love, PED., "Outsourcing Information Systems: Drawing Lessons from a Banking Case Study," *European Journal of Information Systems*, pp. 15-

24 (2001)

[265]Mbelli, T. M., and Dwolatzky, B., "Cyber Security, a Threat to Cyber Banking in South Africa An Approach to Network and Application Security," IEEE *International Conference on Cyber Security and Cloud Computing* (2016)

[266]Zhu, D., "Security Control in Inter-Bank Fund Transfer," *Journal of Electronic Commerce Research*, Vol. 3, No. 1 (2002)

[267]Zachary, B. Omariba., Nelson, B. Masese., and G. Wanyembi., "Security and Privacy of Electronic Banking," *International Journal of Computer Science Issues*, Vol. 9, Issue 4 (2012)

[268]Chen, H., and Corriveau, J. P., "Security Testing and Compliance for Online Banking in RealWorld," *International MultiConference of Engineers and Computer Scientists* (2009)

[269]Gharibi1, W., and Mirz, A., "Software Vulnerabilities, Banking Threats, Botnets and Malware Self-Protection Technologies," *International Journal of Computer Science Issues*, Vol. 8, Issue 1 (2011)

[270]Balusamy, B., Velu, M., Nandagopal, S., and Mano, S. J., "Achieving Security to Overcome Attacks and Vulnerabilities in Mobile Banking Security," *Online Banking Security Measures and Data Protection* (2017)

[271]Kaur, D., Kaur, P., and Singh, H., "Insecurity Status and Vulnerability Density of Web Applications: A Quantitative Approach," *International Journal of Computer Science and Information Security*, Vol. 15, No. 1 (2017)

[272]Jassal, R. K., and Sehgal, R. K., "Online Banking Security Flaws: A Study," *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol 3, Issue 8 (2013)

[273]Lawati, A. AI., and Ali, S., "Business Perception to Learn the Art of Operating System Auditing: A Case of a Local Bank of Oman," IEEE *GCC Conference and Exhibition* (2015)