



**Cape Peninsula  
University of Technology**

**A machine learning approach for master patient index record linkage and deduplication**

**by**

**Dane Giovanni Hollenbach**

**Thesis submitted in partial fulfilment of the requirements for the degree**

**Master of Information and Communication Technology**

**in the Faculty of Informatics and Design**

**at the Cape Peninsula University of Technology**

**Supervisor: Prof Justine Olawande Daramola**

**Cape Town**

**December 2024**

**CPUT copyright information**

The dissertation/thesis may not be published either in part (in scholarly, scientific or technical journals), or as a whole (as a monograph), unless permission has been obtained from the University

## DECLARATION

I, **Dane Hollenbach**, declare that the contents of this dissertation/thesis represent my own unaided work and that the dissertation/thesis has not previously been submitted for academic examination towards any qualification. Furthermore, it expresses my own opinions and not necessarily those of the Cape Peninsula University of Technology.



11/12/2024

---

**Signed**

**Date**

## ABSTRACT

The research emphasised the vital role of a Master Patient Index (MPI) solution in addressing the challenges public healthcare facilities face in eliminating duplicate patient records and improving record linkage. The study recognised that traditional MPI systems may have limitations in terms of efficiency and accuracy. To address this, the study focused on utilising machine learning techniques to enhance the effectiveness of MPI systems, aiming to support the growing record linkage healthcare ecosystem.

It was essential to highlight that integrating machine learning into MPI systems is crucial for optimising their capabilities. The study aimed to improve data linking and deduplication processes within MPI systems by leveraging machine learning techniques. This emphasis on machine learning represented a significant shift towards more sophisticated and intelligent healthcare technologies. Ultimately, the goal was to ensure safe and efficient patient care, benefiting individuals and the broader healthcare industry.

This research investigated the performance of five machine learning classification algorithms (random forests, extreme gradient boosting, logistic regression, stacking ensemble, and deep multilayer perceptron) for data linkage and deduplication on four datasets. These techniques improved data linking and deduplication for use in an MPI system.

The findings demonstrate the applicability of machine learning models for effective data linkage and deduplication of electronic health records. The random forest algorithm achieved the best performance (identifying duplicates correctly) based on accuracy, F1-Score, and AUC-score for three datasets (Electronic Practice-Based Research Network (ePBRN): Acc = 99.83%, F1-score = 81.09%, AUC = 99.98%; Freely Extensible Biomedical Record Linkage (FEBRL) 3: Acc = 99.55%, F1-score = 96.29%, AUC = 99.77%; Custom-synthetic: Acc = 99.98%, F1-score = 99.18%, AUC = 99.99%). In contrast, the experimentation on the FEBRL4 dataset revealed that the Multi-Layer Perceptron Artificial Neural Network (MLP-ANN) and logistic regression algorithms outperformed the random forest algorithm. The performance results for the MLP-ANN were (FEBRL4: Acc = 99.93%, F1-score = 96.95%, AUC = 99.97%). For the logistic regression algorithm, the results were (FEBRL4: Acc = 99.99%, F1 = 96.91%, AUC = 99.97%).

In conclusion, the results of this research have significant implications for the healthcare industry, as they are expected to enhance the utilisation of MPI systems and improve their effectiveness in the record linkage healthcare ecosystem. By improving patient record linking and deduplication, healthcare providers can ensure safer and more efficient care, ultimately benefiting patients and the industry.

**Keywords:** Machine learning, master patient index, record linkage and deduplication, supervised learning, electronic health records

## **ACKNOWLEDGEMENTS**

I wish to thank:

- I am grateful to Professor Justine Olawande Daramola for his guidance, support, and direction during my coursework and for agreeing to supervise my work. Thank you for the growth opportunity.
- I would like to thank my parents, wife, and kids for their sacrifices, understanding and encouragement throughout my degree process. Thank you for always being there and always supporting my goals and dreams.

This thesis has been a long, exciting, and rewarding journey. It coincided with my family and me taking on a life-changing journey. Therefore, I am genuinely grateful for their support and would like to acknowledge that I could not have done this without it.

## **DEDICATION**

I dedicate this thesis to my mom Natalie, wife Thamson, and kids Vanya, Duncan and Hannah.  
I am truly grateful for your love, prayers, encouragement, and unwavering support.

## **PUBLICATION FROM THESIS**

### **Publications**

- I. Dane Hollenbach, Olawande Daramola (2025). Performing record linkage and deduplication in master patient index using machine learning classifiers, SAI Computing Conference, 2025, London, United Kingdom (In Press)
- II. Dane Hollenbach, Olawande Daramola (2025). Machine learning classifiers for record linkage and deduplication in master patient index. (Journal article in preparation)

## TABLE OF CONTENTS

DECLARATION .....	2
ABSTRACT.....	3
ACKNOWLEDGEMENTS .....	5
DEDICATION.....	6
PUBLICATION FROM THESIS.....	7
TABLE OF CONTENTS .....	8
GLOSSARY .....	12
CHAPTER ONE .....	13
INTRODUCTION .....	13
1.1 Overview of the Study .....	13
1.2 Background .....	14
1.3 Research Problem.....	16
1.4 Aim, Objectives and Research Questions.....	17
1.4.1 Research Aim .....	17
1.4.2 Research Objectives.....	17
1.4.3 Research Questions .....	17
1.5 Delineation of the Study .....	18
1.6 Significance of the Study .....	18
1.7 Thesis Outline .....	19
1.8 Chapter Summary .....	19
CHAPTER TWO .....	20
LITERATURE REVIEW .....	20
2.1 Health Information Exchange (HIE) .....	20
2.2 Patient Record Identification.....	20
2.2.1 Consequences of Poor Patient Identification .....	21
2.2.2 Patient Identification as a Non-Technology Problem.....	21
2.2.3 Patient Identification as a Technology Problem .....	22
2.3 Master Patient Index.....	22
2.4 Record Linkage and Deduplication .....	23
2.5 Record Linkage Algorithms.....	24



2.5.1	Probabilistic Record Linking.....	26
2.5.2	Machine Learning .....	27
2.5.2.1	Random Forests .....	28
2.5.2.2	Extreme Gradient Boosting (XGBoost) .....	29
2.5.2.3	Logistic Regression .....	29
2.5.2.4	Stacking Ensemble .....	30
2.5.2.5	Artificial Neural Networks – Deep Multi-Layer Perceptron.....	31
2.6	Related Work.....	32
2.7	Chapter Summary .....	38
CHAPTER THREE.....		39
RESEARCH METHODOLOGY.....		39
3.1	Research Philosophy.....	39
3.2	Research Approach.....	39
3.3	Research Methodology.....	39
3.4	Research Strategy.....	40
3.4.1	Research Design .....	40
3.4.1.1	Data Sampling and Collection.....	41
3.5	Ethical Considerations.....	42
3.6	Chapter Summary .....	42
CHAPTER FOUR .....		43
EXPERIMENTATION .....		43
4.1	System Architecture .....	43
4.2	Experimentation Workflow .....	43
4.2.1	Software, Tools, and Frameworks .....	44
4.3	Datasets .....	45
4.4	Description of the Dataset .....	46
4.4.1	FEBRL Datasets.....	47
4.4.2	ePBRN .....	47
4.4.3	Custom Synthetic Dataset .....	47
4.5	Data Cleaning.....	47

4.6	Feature Selection .....	48
4.7	Model Training and Testing .....	49
4.8	Chapter Summary .....	51
CHAPTER FIVE.....		52
EVALUATION.....		52
5.1	Model Performance Evaluation.....	52
5.2	Discussion.....	57
5.3	Chapter Summary .....	58
CHAPTER SIX.....		60
SUMMARY, CONCLUSION, AND RECOMMENDATIONS.....		60
6.1	Summary.....	60
6.2	Contributions of the Study .....	62
6.2.1	Theoretical Contribution.....	62
6.2.2	Practical Contribution.....	63
6.3	Limitations of the Study and Potential Impact on the Results.....	64
6.4	Conclusion .....	64
6.5	Recommendations and Future Work .....	64
REFERENCES .....		66
APPENDICES.....		71

## LIST OF FIGURES

Figure 2.1: Architecture of random forests (Khan et al., 2021) .....	28
Figure 2.2: Architecture of XGBoost (Deng et al., 2021) .....	29
Figure 2.3: Architecture of the logistic regression model (Coleman et al., 2023) .....	30
Figure 2.4: Architecture of a stacking ensemble model (Habib & Rahman, 2021) .....	31
Figure 2.5: Architecture of multi-layer perceptron (Naskath et al., 2023) .....	32
Figure 3.1: An overview of the adopted experimental research design .....	41
Figure 4.1: An overview of the proposed machine learning-based record linkage and deduplication process .....	44
Figure 5.1: The ePBRN dataset in which the Random Forest model achieved the best performance .....	54
Figure 5.2: The FEBRL3 dataset in which the Random Forest model performed best. ....	55
Figure 5.3: The FEBRL4 dataset in which the MLP-ANN model performed best. ....	56
Figure 5.4: The Custom Dataset in which the Random Forest model achieved the best performance .....	57

## LIST OF TABLES

Table 2.1 Summary of Related Work .....	35
Table 4.1 Workstation Configuration .....	43
Table 4.2 Libraries Used .....	45
Table 4.3 Understanding duplicates in the Datasets .....	46
Table 4.4 Dataset breakdown .....	46
Table 4.5 Feature selection, description, and justification .....	48
Table 4.5 Machine Learning Model Hyperparameter Configuration and Tuning .....	50
Table 5.1 Model Performance on Datasets .....	53
Table 5.2 Summary of the best-performing models for each of the datasets .....	58

## APPENDICES

APPENDIX A: ETHICS CERTIFICATE .....	71
APPENDIX B: USE OF SYNTHETIC DATA ACKNOWLEDGEMENT .....	72
APPENDIX C: ePBRN Dataset .....	73
APPENDIX D: FEBRL3 Dataset .....	73
APPENDIX E: FEBRL4 Dataset .....	73
APPENDIX F: Custom Dataset .....	74
APPENDIX G: Machine Learning Model Results Per Dataset .....	74

## GLOSSARY

Abbreviation/Acronym	Definition
MPI	Master Patient Index
HIE	Health Information Exchange
MRN	Medical Record Number
CHIME	College of Healthcare Information Management Executives
LR	Logistic Regression
ANN	Artificial Neural Networks
MLP	Multi-layer Perceptron
FEBRL	Freely Extensible Biomedical Linkage
ePBRN	Electronic Practice-Based Research Network
UNSW	University of New South Wales
EHR	Electronic Health Records
RF	Random Forests
XGBoost	Extreme Gradient Boosting
LR	Logistic Regression
SE	Stacking Ensemble
MLP-ANN	Multi-Layer Perceptron Artificial Neural Network
NN	Neural Network
IDE	Integrated Development Environment
TP	True Positives/Actual Duplicates
FN	False Negatives/Actual Not Duplicates
FP	False Positives/Predicted Duplicates
TN	True Negatives/Predicted Not Duplicates

# CHAPTER ONE

## INTRODUCTION

### 1.1 Overview of the Study

Public healthcare services are becoming more disparate, and patients are using multiple healthcare facilities and interacting with various source systems within the same healthcare facility (Beth et al., 2016; Fernandes & O'Connor, 2015; Duggal et al., 2015). These systems often include admissions, dispensaries, and others, all of which may use their medical record number (MRN). Healthcare providers are also becoming more interoperable but still lack a standardised approach to sharing information (Duggal et al., 2015a). Data is stored in different formats, and policies and procedures for capturing and storing information are not consistently implemented (Riplinger et al., 2020; Fernandes & O'Connor, 2015).

Moreover, public healthcare facilities must identify and prevent duplicate records from multiple source systems in their health information exchange (HIE) (Menachemi et al., 2018; Duggal et al., 2015; Harron et al., 2017; Riplinger et al., 2020; Thorell et al., 2019). The HIE aims to enable interoperability between various source systems by following nationally recognised standards (Menachemi et al., 2018). The primary purpose of an HIE is to allow doctors, nurses, pharmacists, and other healthcare providers to access and securely share a patient's electronic medical record, improving speed, quality, safety, and cost of patient care (Anon, 2021). Furthermore, healthcare providers associate the ability to provide efficient quality healthcare services with giving a patient a longitudinal patient record dependent on accurately identifying a patient in an HIE (Morris et al., 2014; Riplinger et al., 2020; Thorell et al., 2019).

However, it is estimated that approximately 1.1 billion people, mainly in Africa and Asia, cannot prove who they say they are. Many of these people are women and children residing in poor rural areas. Their inability to identify themselves is also a barrier to accessing quality healthcare services (Thorell et al., 2019).

Additionally, in many countries, including Africa, a standardised way to capture and store patient information does not exist (Thorell et al., 2019; Morris et al., 2014). In most cases, a minimum required amount of geographical information is common but lacks a standardised mechanism for capturing and storing it (Morris et al., 2014). It is essential to accurately identify and capture a patient's details at registration or healthcare access. The lack thereof results in creating patient records in multiple source systems with potentially multiple source identifiers that are not interoperable but belong to the same individual. These issues are often addressed by using a master patient index (MPI). An MPI system stores a directory of all patients in an

HIE ecosystem and securely exposes patient demographic and clinical data. It has the core role of creating and maintaining a unique identifier for all patients by combining geographic patient characteristics in record linkage (Chouffani, 2017; Nelson et al., 2023).

Record linkage is the science of finding duplicates or matches in records from different source systems using nonunique identifiers, including first name, last name, date of birth, address, telephone number and other similar characteristics (Winkler, 2002). Although the characteristics are not unique, when combined, they produce an accurate individual identity, allowing the system to determine if two or more records are a potential or a complete match (Winkler, 2009). This is an essential function in an HIE for healthcare providers to provide patients with the most efficient and safe care. Being able to link patient records accurately and thereby eliminate duplicate records allows a healthcare facility to give a patient a longitudinal health record.

This study assessed the use of five machine learning algorithms (random forests, gradient-boosted trees, logistic regression, stacking ensemble and artificial neural network) for patient record linkage and deduplication and using the proposed models in the context of a master patient index (MPI) system. This is important because an MPI system serves as a central repository for patient-level information, which will be further explained in this research. Additionally, the primary function of this system is to address the challenges of record linkage and deduplication.

## **1.2 Background**

Hayler (2018) found that 17% of United States (U.S.) healthcare CIOs from 55 hospitals shared an adverse patient-related incident at their medical facility due to patient data duplication. Furthermore, medical record linking is becoming increasingly important as clinical data is more distributed across multiple source systems (Grannis et al., 2004; Fernandes & O'Connor, 2015; Riplinger et al., 2020; Morris et al., 2014).

Little research has been published on the comparative behaviour and output of software programs dedicated to record linkage and patient matching (Karr et al., 2019). Karr et al. (2019) used actual identifiers to evaluate probabilistic approaches comparing two systems, "Link Plus" and "Link King", without explicitly looking at how certain variables affected weights and matches. Furthermore, Beth et al. (2016) identify three major categories for record linkage, which include basic, intermediate, and advanced algorithms. Building on this understanding, the next section will explore probabilistic algorithms and their significance in record linkage and patient matching.

The Fellegi-Sunter theory is the foundation for many probabilistic algorithms (Morris et al., 2014; Goldstein et al., 2017). In addition, the idea is also a complex statistical examination of a collection or string of patient data variables that, when regarded collectively, determine whether there is an automatic match, no match, or manual review required. Probabilistic algorithms commonly use Soundex, edit distance calculations, frequency indexing, and other tools to correct data-entering errors (Kousthubha & Raghuveer, 2018). These approaches can also be combined to form a hybrid matching method. The industry has refrained from recommending a standard matching method or algorithm because these are frequently tailored and fine-tuned for individual data sets, contemplating how demographic factors vary by neighbourhood and ethnicity (Duggal et al., 2015; Chouffani, 2017; Morris et al., 2014; Sauleau et al., 2005; McCoy et al., 2012; Beth et al., 2016). Furthermore, there is a lack of substantiated research on the variability of success with different matching methods, particularly with real-world data sets (Christen & Pudjijono, 2009; Peter, 2005; Vo et al., 2019; Christen, 2008; Nelson et al., 2023). However, one small study of sample data found that simple deterministic methods did not perform as well as probabilistic or hybrid methods because simple deterministic methods struggle to handle data quality issues and have an inability to handle variations in data (Morris et al., 2014).

Record linkage can be approached as a classification problem because the process involves determining whether pairs of records from different datasets are either a match or a non-match (Kousthubha & Raghuveer, 2018). To link and deduplicate data, machine learning algorithms such as decision trees, and support vector machines can be effectively applied (Kousthubha & Raghuveer, 2018). Existing methods require training data to train the machine learning models used for classifying records into matched or unmatched groups (Goldstein et al., 2017). However, when no training data is available, it may be feasible to generate a set utilising comparable data with known matching status or a subset of current data subjected to meticulous hand-matching (Vo et al., 2019; Nelson et al., 2023; Harron et al., 2017; Goldstein et al., 2017).

In conclusion, it is essential to address the implications of duplicate records for healthcare systems and to find effective solutions to alleviate the resulting challenges. This involves enhancing collaboration among healthcare workers in patient admission areas and implementing consistent policies and processes to improve the overall efficiency of the healthcare system. These measures may include refining search protocols, standardising data capture, and developing specific questioning techniques for registrars to identify prior patient visits to the healthcare centre (Beth et al., 2016).

### 1.3 Research Problem

It is a challenge to link and deduplicate patient records within the healthcare industry due to the proliferation of data sources and the existence of redundant data across various databases (Kousthubha & Raghuveer, 2018; Goldstein et al., 2017). Additionally, the United States Centers for Disease Control and Prevention (CDC) also estimates that 85% of all patient health data exist digitally (Centers for Disease Control, 2018). Furthermore, the proliferation of data sources across heterogeneous healthcare information systems always leads to redundant data scattered across various databases (Sauleau et al., 2005; Vo et al., 2019; Nelson et al., 2023; McCoy et al., 2012; Kousthubha & Raghuveer, 2018; Chouffani, 2017). Delivering continuous care and conducting health research involves identifying patients across numerous care facilities or services, a complex problem (Vo et al., 2019; Kousthubha & Raghuveer, 2018).

Patient linking and deduplicating records have long been a challenge for many countries around the globe (Fernandes & O'Connor, 2015). However, poor record-linking processes and duplicate patient records still plague the healthcare industry (Liang et al., 2018; Morris et al., 2014; Fernandes & O'Connor, 2015; Nelson et al., 2023; Vo et al., 2019). In addition, a master patient index (MPI) system is essential in linking and deduplicating patient records (Liang et al., 2018).

A unique identifier that could solve the problem of record linkage and deduplication often does not exist or is not favoured when dealing with many disparate systems (Morris et al., 2014). The use of High Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) as a standards-based interoperability layer for use by the MPI system is thus required to act as an intermediary between healthcare systems with the sole purpose of linking, deduplicating and assigning identifiers to patients (Fernandes & O'Connor, 2015; Saripalle et al., 2019).

Furthermore, some studies have found that the leading reasons for duplicate records include missing data, misspelt record information, the use of multiple healthcare systems and the fact that many countries around the world experience a higher number of the population that share exact name and birthdates (Liang et al., 2018; Morris et al., 2014).

Consequently, not addressing the issue of record linkage and deduplication can cause harm to patients, become costly to healthcare providers and negate the benefits of digital health record systems (Liang et al., 2018; Morris et al., 2014; Fernandes & O'Connor, 2015; Vo et al., 2019; Sauleau et al., 2005). Furthermore, ineffective linking and deduplicating patient records



within the healthcare industry leads to fragmented patient information and potential adverse effects on patient care and healthcare costs (Kousthubha & Raghuv eer, 2018).

## **1.4 Aim, Objectives and Research Questions**

### **1.4.1 Research Aim**

This study aimed to explore machine learning algorithms for patient record linkage and deduplication within a master patient index.

### **1.4.2 Research Objectives**

To accomplish the goal of this research, the following research objectives were established to:

1. Determine the criteria for identifying duplicate records within a data source.
2. Formulate record linkage and deduplication as a machine learning classification task.
3. Apply selected machine learning algorithms for record linkage and deduplication.
4. Evaluate the performance of the selected machine learning algorithms for record linkage and deduplication.

### **1.4.3 Research Questions**

This study's primary research question is: how can machine learning algorithms be applied to patient record linkage and deduplication within a master patient index?

The sub-research questions are the following:

1. What is the basis for identifying duplicate records within a data source?
2. What parameters should be considered when representing record linkage and deduplication as a machine learning classification task?
3. How can machine learning classification algorithms be applied for record linkage and deduplication?
4. What is the performance of selected machine learning algorithms in the context of record linkage and deduplication?

## **1.5 Delineation of the Study**

Given the global issue of record linkage and deduplication of patient records and the limited time scope for this study, open-source datasets were utilised for experiments on patient records for deduplication and record linkage (Peter, 2005; Christen & Pudjijono, 2009; Nelson et al., 2023). The datasets used are described in sections 4.4.1 FEBRL Datasets, 4.4.2 ePBRN, and 4.4.3 Custom Synthetic Dataset, respectively. The experimentation in this study was limited to five machine learning algorithms, described in sections 2.5.2.1 Random Forests, 2.5.2.2 Extreme Gradient Boosting (XGBoost), 2.5.2.3 Logistic Regression, 2.5.2.4 Stacking Ensemble, and 2.5.2.5 Artificial Neural Networks – Deep multi-Layer Perceptron. This study's experimentation used the combination of datasets and machine learning algorithms mentioned above. It is important to note that this study did not consider other algorithms.

## **1.6 Significance of the Study**

The findings of this study are intended to benefit electronic healthcare systems in various ways. According to recent studies, there is a consistent increase in the use of health information exchanges, which increases the need for uniquely identifying a patient across various care settings (Centers for Disease Control, 2018; Fernandes & O'Connor, 2015; McCoy et al., 2012; Morris et al., 2014). Furthermore, this study will directly contribute to enhancing the use of MPI systems and making them more effective in performing their role in this growing connected ecosystem. MPI systems improve health information exchanges and enhance the effectiveness of identifying patients across different care settings (Kousthubha & Raghuveer, 2018; Duggal et al., 2015; Morris et al., 2014; McCoy et al., 2012; Beth et al., 2016). In conclusion, by improving healthcare systems and data accuracy, healthcare providers can ensure better coordination and accuracy of patient information, ultimately leading to improved quality of care and patient safety in the expanding connected healthcare ecosystem.

## **1.7 Thesis Outline**

This thesis is structured as follows: Chapter 2 includes the literature review, covering the background and related work. Chapter 3 outlines the methodology utilised to accomplish the research objectives. Chapter 4 presents the results of the experiments based on the experimental research (ER) design workflow. Chapter 5 delves into a deeper discussion of the results. Finally, Chapter 6 provides a summary, conclusion, and recommendations for further research work.

## **1.8 Chapter Summary**

The chapter discusses the challenges and importance of accurate patient record linkage in the context of public healthcare services. It highlights the lack of standardised approaches for capturing and storing patient information, the need for interoperability between various healthcare systems, and the significance of Health Information Exchange (HIE) in improving patient care. The chapter also touches upon the issue of patient identification, especially in underprivileged areas, and emphasises the role of a master patient index (MPI) system in addressing these challenges. Furthermore, it outlines the significance of record linkage in an HIE ecosystem and proposes using machine learning algorithms within the MPI framework to improve record linkage accuracy. Additionally, it mentions the impact of patient data duplication on healthcare facilities and the potential of machine learning in enhancing record linkage accuracy compared to probabilistic record linkage methods.

## **CHAPTER TWO**

### **LITERATURE REVIEW**

This chapter provides an overview of the literature. The chapter is divided into six sections. It discusses the significance of making a patient's electronic medical record accessible to healthcare providers via Health Information Exchange (HIE). It emphasises the importance of patient record identification, explores using a master patient index (MPI) as a tool within an HIE to solve the issue of record linkage and deduplication, discusses record linkage algorithms, and provides a summary of related work.

#### **2.1 Health Information Exchange (HIE)**

Healthcare providers associate the ability to provide efficient, quality healthcare services with giving a patient a longitudinal patient record, which depends on accurately identifying a patient in a health information exchange (HIE) (Thorell et al., 2019; Beth et al., 2016; Kousthubha & Raghuveer, 2018). In addition, healthcare systems are becoming more disparate, making it more important to have a single unified view of a patient's health record. The HIE aims to enable interoperability between various source systems by following nationally recognised standards (Menachemi et al., 2018).

The primary purpose of an HIE is to allow doctors, nurses, pharmacists and other healthcare providers to access and securely share a patient's electronic medical record to improve the speed, quality, safety and cost of patient care (Anon, 2021). Additionally, using an HIE is often a national objective motivated by the fundamental promise of improved patient care, efficiencies, and reduced healthcare-related costs (Menachemi et al., 2018).

#### **2.2 Patient Record Identification**

The issue of record linkage has gained attention in recent years as the adoption of technology increases and the need for interoperability amongst systems becomes a reality (Fernandes & O'Connor, 2015). To fully understand the issue of record linkage, it must be considered at a high level how data is received from multiple sources which may include hospitals, physicians' offices, clinics, rehabilitation services, long-term care, acute care, and others – and to couple these data sources with the additional factors such as varying data formats and varying level of completeness or lack thereof, as well as the problem of inaccurate data (Fernandes & O'Connor, 2015).

Furthermore, Fernandes and O'Connor (2015) conclude that patient identification and linking have long been a challenge in healthcare worldwide. In addition, Duggal et al. (2015) share that one of the significant challenges in healthcare is the inability to consolidate disparate patient data into one view. This is because patient data resides in multiple source systems such as clinical, pharmacy, billing, laboratory and claim systems.

### **2.2.1 Consequences of Poor Patient Identification**

The inability to accurately identify patients poses a risk of losing the ability to provide a longitudinal health record. This can lead to an increased risk of creating fragmented or duplicate health records (Thornton & Shannon, 2005). Improving the quality of treatment, making care more accessible, and managing rising healthcare expenditures are all global issues. To accomplish the goals of higher quality and lower costs, a unified picture of patient data across care settings must be produced, notwithstanding the variety in data capture, technologies, and standards (or lack thereof). It is critical to have accurate and complete information to provide appropriate, high-quality, cost-effective treatment (Fernandes & O'Connor, 2015).

Furthermore, because health information is kept and transmitted electronically, accurate identification and matching of patient records is critical for assuring patient safety. For example, one-fifth of CIOs polled by the College of Healthcare Information Management Executives (CHIME) said that at least one patient had an adverse incident owing to mismatched information in the past year (Morris et al., 2014; Hayler, 2018).

### **2.2.2 Patient Identification as a Non-Technology Problem**

Researchers have found that resolving the issue of patient record linkage is not a technology problem alone. Several reasons contribute to the existence of duplicate records in healthcare systems. Some of these reasons include various methods for matching patient records; departmental silos; lack of standardisation; lack of policies, procedures, and data ownership; frequently changing demographic data; multiple data points required for record matching; and default and null values in required identifying fields are some of the reasons that duplicate records continue to plague healthcare systems (Beth et al., 2016). The impact of duplicate records can be mitigated by collaborating with colleagues in the patient admissions areas and implementing standard policies and processes, such as improved searching techniques, data recording standards, and questions that registrars can ask the patient to ascertain if the patient has already visited the hospital (Beth et al., 2016). In

addition, staff should be coached and tested on identification data and the repercussions of faulty or incomplete data capture on duplicate record creation regularly (Beth et al., 2016).

### **2.2.3 Patient Identification as a Technology Problem**

Health record sharing and the need for a unified view of a patient's medical record are becoming commonplace (Duggal et al., 2015; Beth et al., 2016). To improve patient linking, the use of more sophisticated technologies, such as biometrics systems, card readers, and machine learning algorithms, needed to be increased. (Beth et al., 2016).

Few research studies have investigated the behaviour and output of linkage software and algorithms' effectiveness (Karr et al., 2019). In addition, this study focused on evaluating the performance and quality of various methods and machine learning algorithms utilised for patient record linking, which can be implemented in a master patient index (MPI) system (Nelson et al., 2023).

## **2.3 Master Patient Index**

An MPI system stores a directory of all patients in an HIE ecosystem and securely exposes patient demographic and clinical data. It has the core role of creating and maintaining a unique identifier for all patients by combining geographic patient characteristics in record linkage (Chouffani, 2017; Nelson et al., 2023; Beth et al., 2016).

The MPI plays a critical role in an HIE system by enabling interoperability between many different systems that often use different medical record numbers (MRNs) for the same patient (Beth et al., 2016). As a result, the MPI solution is a crucial component of an HIE and it must function effectively in an automated setting while complementing the work done by human experts.

To comprehend the real necessity for a Master Patient Index (MPI) solution, it is important to note that medical errors cause one-third of deaths in the United States (Beth et al., 2016). To put this into context, 400,000 patients die yearly, equating to more than 1000 people daily (Beth et al., 2016). In addition, a similar study has found that an estimated 195,000 deaths occur because of medical errors, where 10 of 17 are due to patients being incorrectly identified (Beth et al., 2016). Furthermore, it is estimated that the third most significant preventable cause of death in the United States alone is due to medical errors (Hayler, 2018).

Moreover, in many countries, including countries in Africa, a standardised way to capture and store patient information does not exist (Morris et al., 2014; Thorell et al., 2019). In most cases, a minimum required amount of geographical information is common but lacks a standardised mechanism for capturing and storing it (Morris et al., 2014). This results in duplicate patient records in multiple source systems with potentially multiple source identifiers that are not interoperable but belong to the same individual (Thorell et al., 2019). These issues are often addressed using an MPI system that forms part of a HIE architecture.

In addition, the requirement for more systematic approaches for coordinating, integrating, and managing linked records is becoming more apparent in increasingly distributed healthcare systems. One way to deal with this increasingly complex and distributed landscape is to adopt MPI solutions into an HIE ecosystem (Toth et al., 2014; Nelson et al., 2023; Beth et al., 2016; Thorell et al., 2019).

## **2.4 Record Linkage and Deduplication**

Record linkage is the science of finding duplicates or matches in records from different source systems using nonunique identifiers, including first name, last name, date of birth, address, telephone number and other similar characteristics (Winkler, 2002; Winkler, 2009; Kousthubha & Raghuveer, 2018; Goldstein et al., 2017). Although the characteristics are not unique, when combined, they produce an accurate individual identity, allowing the system to determine if two or more records are a potential or a complete match (Winkler, 2009). For delivering high-quality, high-value healthcare, conducting valid and generalisable research, and evaluating healthcare policy, record linkage among medical databases such as electronic health records (EHRs), health insurer claims, and patient-generated data is becoming increasingly important (Karr et al., 2019). Furthermore, record linkage is essential in an HIE for healthcare providers to provide patients with the most efficient and safe care. Being able to link patient records accurately and thereby eliminate duplicate records allows a healthcare facility to give a patient a longitudinal health record.

In theory, this study deduced that all record linkage algorithms work similarly. They must identify a collection of connecting variables shared by both datasets; these shared variables serve as a comparison point. In this study, a numerical weight gets generated for each attribute pair compared, which is then interpreted as the degree of confidence that the paired data reflect the same person or object. (Karr et al., 2019). Furthermore, at an implementation level, the foundation of record linkage is based on performing string comparisons, weight determination, and match determination.

Medical record linking is becoming increasingly important as clinical data is distributed across multiple source systems (Grannis et al., 2004; Chouffani, 2017; Duggal et al., 2015; Morris et al., 2014; Sauleau et al., 2005; McCoy et al., 2012; Beth et al., 2016). This study explored five machine-learning algorithms to enhance record linkage accuracy and tested them on four synthetic datasets. The results prove that machine learning is an effective record linkage and deduplication tool.

## **2.5 Record Linkage Algorithms**

Record linkage is the process of identifying similar records within the same or different datasets (Kousthubha & Raghuvier, 2018). Record linkage algorithms are utilised to link records across different systems and/or to detect potential duplicate records (Beth et al., 2016). This study established a strong connection between record linkage algorithms, which are responsible for identifying and removing duplicate records, and their usage within an MPI system (Nelson et al., 2023; Beth et al., 2016; Vo et al., 2019).

Little research has been published on the comparative behaviour and output of software programs dedicated to record linkage and patient matching (Nelson et al., 2023). One study used actual identifiers to evaluate probabilistic approaches comparing two systems (Link Plus and Link King) without explicitly looking at how certain variables affected weights and matches (Karr et al., 2019).

Furthermore, we have ascertained the need for an MPI system, which performs a core role in record linkage and related functions. This functionality exists because of the need to provide a patient with a longitudinal health record (Beth et al., 2016; Nelson et al., 2023). Additionally, we have gained insights into how healthcare systems are becoming more complex and distributed. They must adapt and provide new ways of supporting interoperability when disparate systems may have their MRNs. We now focus on how an MPI accomplishes this functionality by focusing on its methods and, more specifically, probabilistic matching and exploring the need for a machine-learning approach.

Record linkage algorithms perform record linkage across disparate systems and provide healthcare workers functionality in identifying duplicate records from a backend perspective (Beth et al., 2016). Furthermore, Beth et al. (2016) define record-linking algorithms as follows:



1. Basic algorithm – Also called deterministic matching: This is the most fundamental method of matching records and uses deterministic matching. This algorithm compares selected elements within specified fields to identify one-to-one character matches, including phonetic matches and wildcards (Beth et al., 2016).
  - a. A data item must be an exact or partial match for a record to be evaluated as a match.
  - b. Comparisons are typically made by name, date of birth, social security number (SSN), and sometimes gender.
2. Intermediate algorithm – Includes fuzzy logic: It uses more powerful programmed techniques than basic algorithms to compare records. These algorithms account for frequently transposed digits or other typographical errors (Beth et al., 2016). This technique uses weights that are assigned to each field. For example, the last names “Hollenbach” and “Hollenbacht” have similar weights.
  - a. To counter misspelt names and nicknames, phonetic encoding schemes and, in some cases, similar name lists are utilised. Field match weights are assigned arbitrarily/subjectively to important patient-identifier features such as first name, last name, date of birth, and SSN, contributing to an overall weight score.
  - b. Additionally, this method may use programmes specifically developed to remedy transpositions, digit rotations, and typographical errors.
3. Advanced algorithm: This category of linkage algorithms depends on mathematical and statistical theories and employs the most advanced tools for matching records. An example of such an algorithm used in combination with machine learning is the Fellegi-Sunter method (Asher et al., 2020).
  - a. One of the essential aspects is the application of probabilistic theory and mathematical or statistical models to determine the likelihood of a match based on specified data qualities.
  - b. This method also incorporates machine learning and neural networks, which use artificial intelligence to replicate human problem-solving.

This study aimed to evaluate and provide better insights into the “advanced algorithm” category, specifically on performance from an algorithm accuracy point of view and a raw performance standpoint, to determine the best implementation given specific machine learning models and datasets.

This study aimed to develop machine learning models for record linkage and deduplication that could be utilised in popular Master Patient Index (MPI) systems, such as SanteMPI and OpenCR. The machine learning models produced by this research are designed to be

compatible with these well-known MPI systems, as demonstrated in a previous study by Nelson et al. (2023). SanteMPI and OpenCR mainly focus on probabilistic algorithms. This study produced machine learning models, which can be used in the same systems for record linkage and deduplication (Nelson et al., 2023). Both systems perform the function of an MPI system and act as a client registry CR. This means these systems are designed to be integrated within an HIE and will serve as promising tools for evaluating the advanced algorithm category.

Lastly, this study intends to make the use of these advanced methods of record linkage more accessible to healthcare facilities, specifically in low-resource settings and countries. We have already determined earlier that record linkage and the ability to reduce duplicate records accurately play a critical role in providing efficient and quality healthcare. This research aims to provide insights into how these tools can be adapted to work well with existing processes and procedures and complement skilled human workers.

### **2.5.1 Probabilistic Record Linking**

Probabilistic algorithms are predominantly based on the Fellegi-Sunter theory (Morris et al., 2014; Goldstein et al., 2017; Tromp et al., 2011; Thornton & Shannon, 2005). The theory is a complex statistical examination of a collection or string of patient data variables that, when taken together, indicate whether an automatic match, no match, or manual review is required. Probabilistic algorithms commonly use Soundex, edit distance calculations, frequency indexing, and other tools to correct data-entering errors (Beth et al., 2016). Both strategies can be used to create a hybrid matching method. The industry has refrained from recommending a standard matching method or algorithm because these are frequently tailored and fine-tuned for specific data sets, considering the differences in demographic variables between communities and ethnicities. Furthermore, there is a lack of research on the variability of success with various matching methods, particularly with real-world data sets, even though one short analysis of sample data indicated that straightforward deterministic methods did not perform as well as probabilistic or hybrid techniques (Morris et al., 2014).

The probabilistic-based approach to record linkage takes trained data to compute a maximum likelihood estimate of whether a record pair is a match (Kousthubha & Raghuveer, 2018). Each record pair is scored independently, e.g., the first name as a source record is compared to a potential target record. These comparisons are made with one of many string comparison algorithms that ultimately contribute to an overall threshold score, determining whether a record matches. Some of the algorithms used include:

1. Jaccard coefficient: This algorithm uses statistics to determine the similarity or divergence of two sequences.
2. Soundex algorithm: This algorithm focuses on indexing the sounds of names as pronounced in English.
3. Levenshtein: It is a string metric for measuring the differences between two sequences. Additionally, this algorithm determines the distance between two words and the number of edits it would take to make them the same.
4. Jaro-Winkler: This algorithm is like Levenshtein distance, determining the distance metric between two sequences.

Furthermore, the goal of probabilistic record linkage methods is to find a set of weights, or scores, for the set  $C$  that will allow elements of  $C$  to be classified as "matches," "non-matches," or indecisive matches, based on the weights assigned (Goldstein et al., 2017). These weights and thresholds are currently not standardised and still require the input of skilled human labour. In addition, making use of skilled human labour still requires extensive testing and manual intervention that could potentially be further optimised with newer and more modern record linkage approaches which follow.

### **2.5.2 Machine Learning**

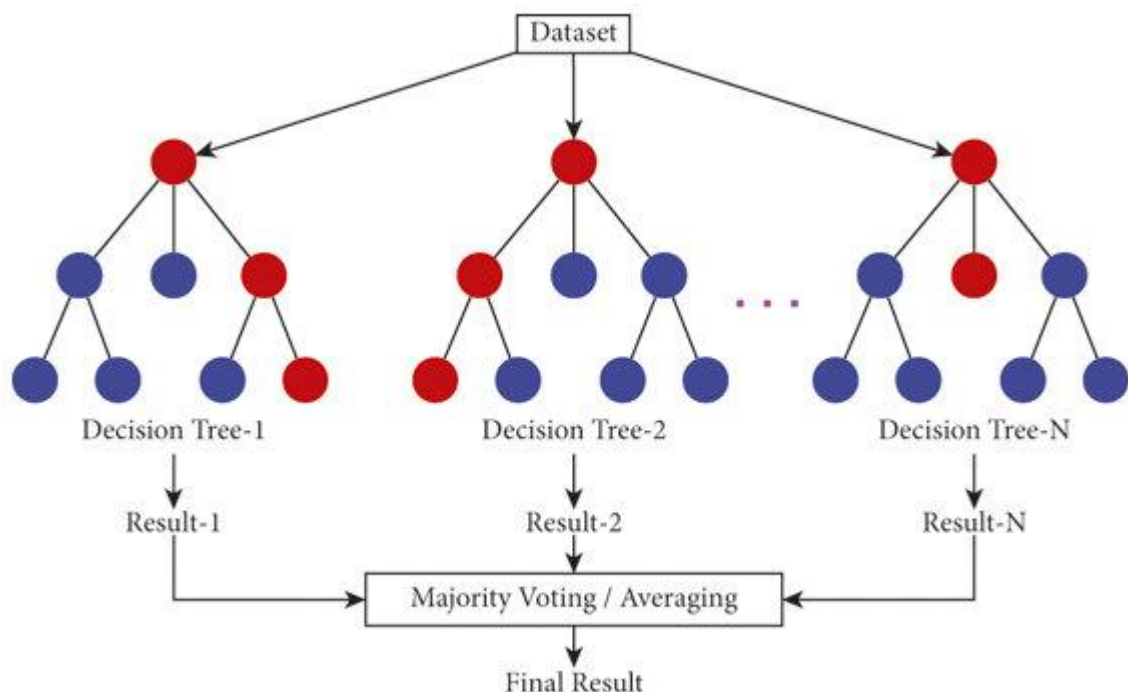
Machine learning can be defined as a computer program that is said to learn from experience when presented with a particular task to solve (Nelson et al., 2023). Many studies often use a single machine-learning algorithm or model to link medical records (Vo et al., 2019). This study used five machine learning algorithms, including random forests, gradient-boosted trees, logistic regression, stacking ensemble and artificial neural networks with four synthetic datasets during experimentation.

Record linkage and deduplication can also be classified as a classification problem (Acheson et al., 2020). Linking patient data and the process of deduplication can utilise machine learning methods such as clustering, decision trees and support vector machines (Kousthubha & Raghuv eer, 2018; Goldstein et al., 2017). In addition, training is used to classify records into matched or unmatched groups. However, when we do not have training data, it may be possible to generate a set utilising similar data with known matching status or a subset of current data subjected to meticulous hand matching (Goldstein et al., 2017). This study leveraged a considerable amount of research in the field of synthetic data generation to create a custom synthetic dataset and utilise datasets used by similar studies (Vo et al., 2019; Nelson et al., 2023; Peter, 2005; Christen & Pudjijono, 2009).

Based on the varied views of which machine learning approach to adopt for record linkage and deduplication, this study aims to explore the most appropriate approach, concluding with utilising five machine learning models and four synthetic datasets to give a more holistic view of approaches (Kousthubha & Raghuveer, 2018; Liang et al., 2018; Pavneet, 2020). The following section will cover the five machine-learning algorithms that were utilised in this study.

### 2.5.2.1 Random Forests

Random forests are non-linear, nonparametric classifiers that are essentially regularised by ensembles and do not tend to overfit. A vital parameter selection in a random forest is the number of trees used in the ensemble. The higher the number, the less likely it is to overfit the model (Acheson et al., 2020). Random forests outperform single decision trees in performance and robustness, yielding more accurate results. Furthermore, they are less prone to overfitting and can handle hundreds of input variables without the need for variable elimination. Random forests may be biased to qualities with a more significant number of levels in categorical data with more than one level (Pita et al., 2017). Figure 2.1 illustrates the architecture of random forests (Khan et al., 2021).



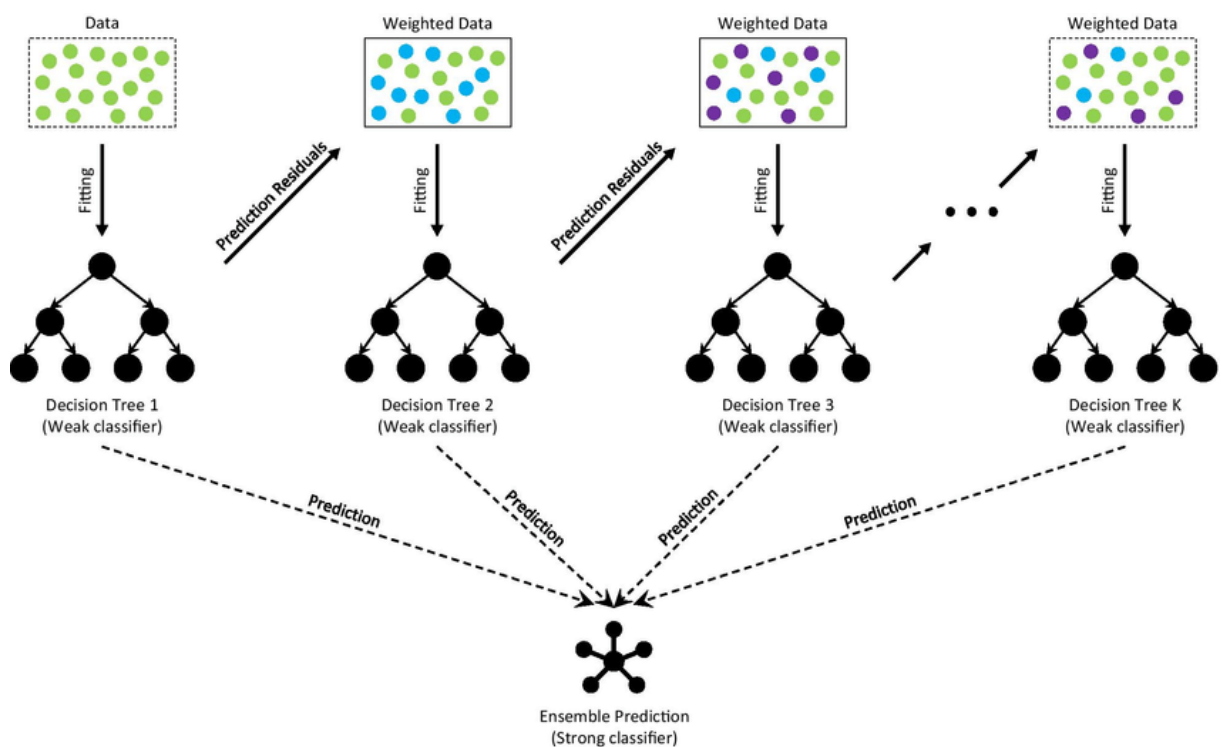
**Figure 2.1: Architecture of random forests** (Khan et al., 2021)

### 2.5.2.2 Extreme Gradient Boosting (XGBoost)

Gradient-boosted trees usually perform well but take longer to learn because they are formed sequentially. They are more prone to overfitting; thus, it is critical to exercise caution during the pre-processing stage (Pita et al., 2017).

Boosting refers to a technique for reducing errors in predictive data analysis. Data scientists train machine learning software (machine learning models) on labelled data to generate educated judgments about unlabelled data (Amazon, 2022).

Furthermore, gradient boosting is an intriguing sequential training strategy because it does not give wrongly identified objects extra weight. It tries to accurately predict target variables by merging estimates from simpler and weaker models (Amazon, 2022). The architecture of gradient-boosted trees is illustrated in Figure 2.2 (Deng et al., 2021).



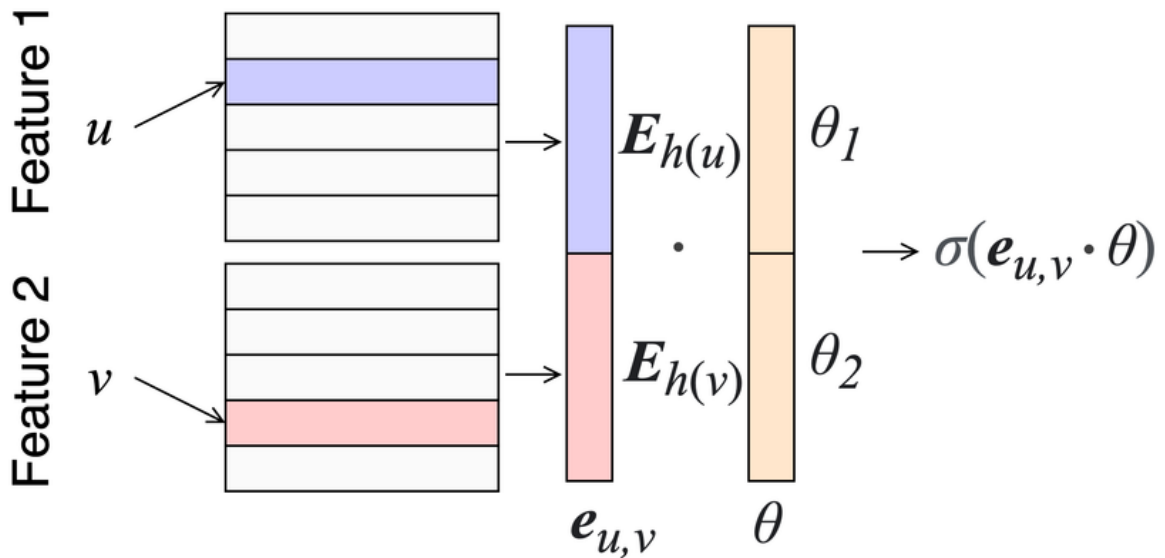
**Figure 2.2: Architecture of XGBoost** (Deng et al., 2021)

### 2.5.2.3 Logistic Regression

Logistic regression is a statistical method to model the relationship between a categorical dependent variable and one or more independent variables. It is commonly employed in medical research to predict binary outcomes, such as disease presence or absence, based on various factors (Boateng & Abaye, 2019). Unlike linear regression, logistic regression does

not assume a linear relationship between the independent and dependent variables; instead, it uses the logit of the outcome to establish the relationship. It is important to note that logistic regression has specific assumptions regarding the data, requires large sample sizes for accurate results, and is suitable for situations where the predicted variable takes two categories.

Additionally, logistic regression can accommodate both categorical and continuous independent variables. However, large sample sizes are required to provide accurate results, and the number of predictor variables should be limited relative to the number of outcome events. Furthermore, the technique assumes that the dependent variable is categorical, the independent variables need not be interval, and the categories must be mutually exclusive and exhaustive (Boateng & Abaye, 2019). Despite its flexibility compared to traditional regression techniques, logistic regression necessitates careful attention to its assumptions and sample size requirements for reliable and meaningful results. The architecture for logistic regression is illustrated in Figure 2.3 (Coleman et al., 2023).



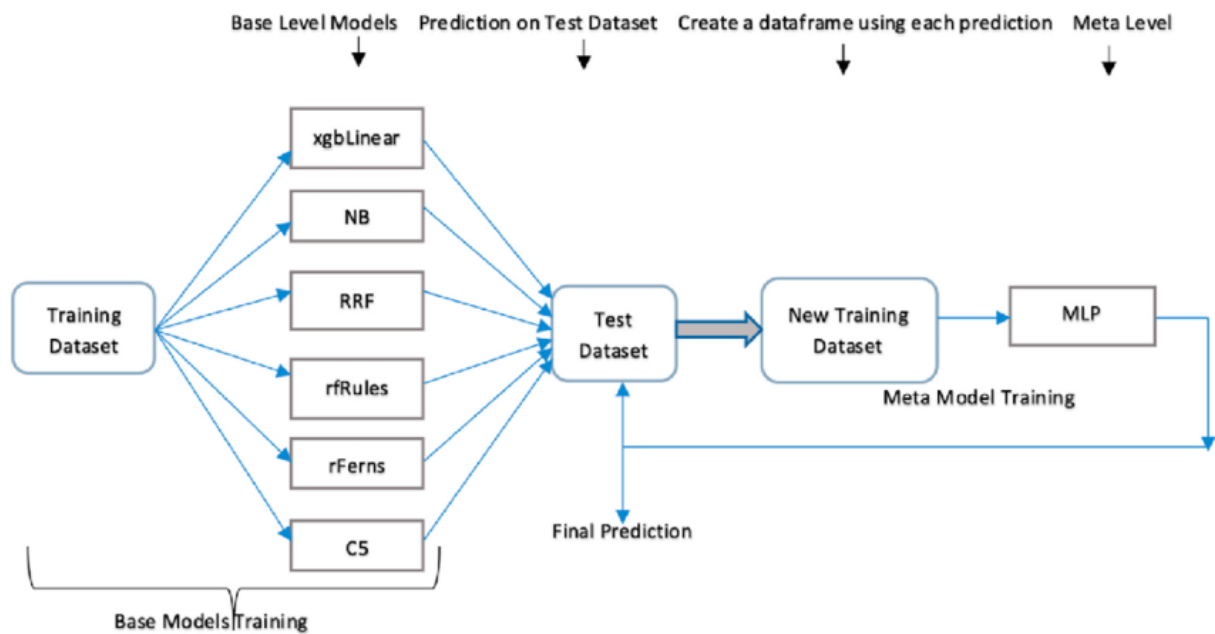
**Figure 2.3: Architecture of the logistic regression model** (Coleman et al., 2023)

#### 2.5.2.4 Stacking Ensemble

Stacking ensemble learning is a method that leverages the complementary strengths of base models to enhance performance and improve generalisation ability (Lu et al., 2023). The process involves two main phases: the first phase includes training the base models using k-fold cross-validation on the original data. In contrast, the second phase entails reassembling the predictions from the base models to create a new training set for a meta-model (Lu et al.,

2023). This meta-model is then trained based on the new dataset, combining predictions from the base models' testing set to obtain the meta-model's testing set.

In summary, stacking ensemble learning involves training base models using k-fold cross-validation, reassembling their predictions to create a new training set for a meta-model, and then training the meta-model based on this new dataset (Lu et al., 2023). This approach capitalises on the strengths of multiple models to enhance overall performance and generalisation ability. Figure 2.4 illustrates the architecture of a stacking ensemble model (Habib & Rahman, 2021).



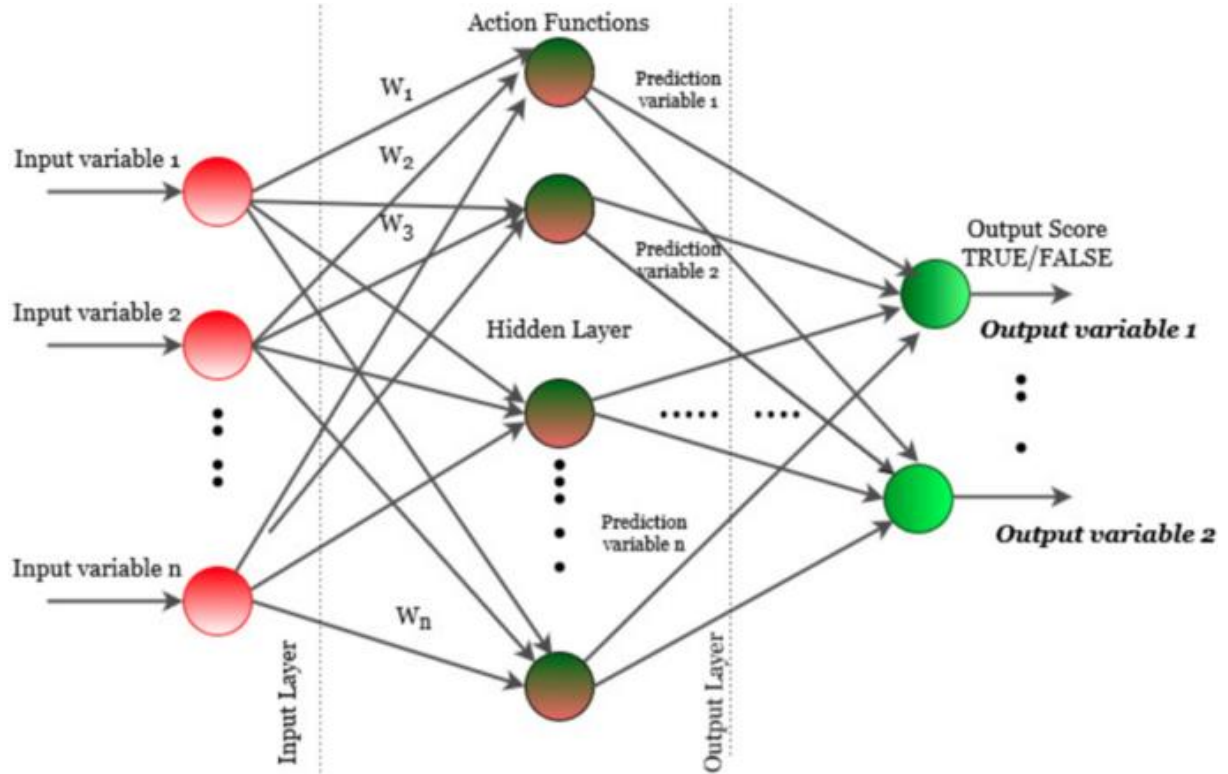
**Figure 2.4: Architecture of a stacking ensemble model** (Habib & Rahman, 2021)

### 2.5.2.5 Artificial Neural Networks – Deep Multi-Layer Perceptron

Deep learning is a highly popular area within machine learning that utilises multiple layers to learn data representations at different levels. This is achieved through nonlinear layers, enabling the creation of complex data abstractions. The approach has led to significant advancements in information processing, mainly through the development of generative and discriminative models, as well as model transfer techniques (Naskath et al., 2023). In addition, deep learning algorithms such as Multi-layer Perceptron, Self-organizing Map, and deep belief networks have found applications across various fields, including wireless networks, speech recognition, medical applications, natural language processing, and remote sensing.

Multi-layer Perceptron (MLP) is a widely used neural network based on supervised learning, characterised by a one-directional flow of information without loops. The primary goal of MLP

is to determine the optimised function  $f()$  that maps input to the desired output while learning the optimised bias value ( $\theta$ ) for it (Naskath et al., 2023). Backpropagation is employed in MLP to adjust the connection weights when there is a discrepancy between the expected and actual output. The main applications of MLP are in solving optimisation problems in diverse domains such as finance, transportation, fitness, and energy. Figure 2.5 illustrates the architecture of artificial neural networks – deep multi-layer perceptron (Naskath et al., 2023).



**Figure 2.5: Architecture of multi-layer perceptron** (Naskath et al., 2023)

## 2.6 Related Work

Much research has been dedicated to improving patient record linkage and deduplication in the last few years (Beth et al., 2016; Centers for Disease Control, 2018; Duggal et al., 2015; McCoy et al., 2012; Kousthubha & Raghuveer, 2018; Tromp et al., 2011; Chouffani, 2017; Vo et al., 2019; Nelson et al., 2023). The authors focused explicitly on framing the issues surrounding patient linking and deduplication. They further focused on why patient matching is an issue and the implications of duplicate records (Beth et al., 2016; McCoy et al., 2012; Kousthubha & Raghuveer, 2018).



Minimal work has been found addressing the specific problem of record linkage and deduplication in public healthcare, providing a practical solution for applying the algorithms to an MPI solution (Nelson et al., 2023). This study produced quantitative metrics based on the performance of machine learning algorithms. Also, it provided insights into the potential use of these algorithms in MPI systems for future work. (Nelson et al., 2023).

As mentioned earlier, the literature review shows no single solution is available. A recent study focused on patient matching and deduplication found that using a two-step Long Short-Term Memory (LSTM) network model produced an accuracy of 99.82% with high performance (Liang et al., 2018). The study further highlights the importance of using newer algorithms within an MPI system (Liang et al., 2018). It is important to note that the use of machine learning was combined with other existing string-based comparison and classification algorithms like named entity recognition (NER), Levenshtein, Jaro-Wingler and Word2vec to find the most optimal model for record linkage and deduplication. Many of these string comparison methods are also adopted by probabilistic approaches.

A study by Vo et al. (2019) focused on improving record linkage performance by employing ensemble strategies. The study showed that this method outperformed traditional approaches focused on a single dataset. Another study by Nelson et al. (2023) focused on using machine learning for record linkage and deduplication but building a practical solution that is generic enough to incorporate into any MPI system. This study showed that using machine learning algorithms in the context of an MPI solution can significantly improve performance over existing record linkage methods.

The industry has concluded that a unified effort is required to understand the root cause of the problem and that only then can we identify a more inclusive and holistic solution. No one solution currently exists that can accurately match patient records 100 per cent of the time and, so doing, eliminate false positives and false negatives (Grannis et al., 2004; Beth et al., 2016; McCoy et al., 2012). The industry at large acknowledges that policies and procedures should be adopted to help resolve the issue of patient matching and that one such solution might be the introduction of A universal/global patient identifier (Morris et al., 2014; Fernandes & O'Connor, 2015; Thorell et al., 2019). However, the industry still recognises that the currently existing methods will still be required for record linkage. Emerging techniques such as machine learning introduce new solutions to problems such as accuracy, scalability and adoption standards (Verschuuren et al., 2020; Pita et al., 2017; Pavneet, 2020; Christen, 2008; Nelson et al., 2023; Brignone et al., 2018)

In addition, it can be concluded that record linkage is not solely a statistical or technological problem. The issue around data integrity and processes for registering patients and validating information at point-of-care scenarios are some of the critical areas that still need to be improved in addition to the enhancement of tools which are there to assist with this growing issue as healthcare facilities deal with more heterogenous datasets. MPI systems are, therefore, an essential part of the puzzle in the healthcare ecosystem that can only work well when fully adopted by healthcare professionals. No technology can solve a problem without being fully acknowledged and adopted by the industry for which it is developed.

Furthermore, the literature reveals that numerous research papers have been published on using machine learning algorithms to solve the problem of record linkage and deduplication; minimal work has been done on implementing this technology into an MPI system, which is one of the main tools for record linkage and deduplication (Pita et al., 2017; Verschuuren et al., 2020; Acheson et al., 2020). This research aims to explore and develop a machine-learning approach for accurate record linking and deduplication in the context of an MPI system.

**Table 2.1 Summary of Related Work**

Author/s	Title	Context	Methodology / Approach	Key Findings
Centers for Disease Control (2018)	Bridging Public Health and Health Care	Patient Linking and Deduplication	Literature review which examined electronic health data.	85% of all patient data exists electronically.
Beth et al. (2016)	Why Patient Matching Is a Challenge: Research on Master Patient Index (MPI) Data Discrepancies in Key Identifying Fields	Patient Linking and Deduplication within an MPI	Experiment: Examined the underlying causes of duplicate records	Increasing the use of more sophisticated technologies is critical to improving patient matching.
Fernandes & O'Connor (2015)	Accurate Patient Identification - A Global Challenge	Patient Linking and Deduplication in different countries.	Literature Review: Focus on record linkage and deduplication.	Patient matching is a global challenge.
McCoy et al. (2012)	Matching identifiers in Electronic Health Records: implications for Duplicate Records and patient safety	Patient Linking and Deduplication and the implications on patients	Experiment: Use personal attributes like name and surname to find duplicate records.	The percentage of records that match patient identifiers is high in several organisations, indicating that the rate of duplicate records or records may also increase.
Sauleau et al. (2005)	Medical record linkage in health information	Duplication of data	Experimental: Finding duplicates	Duplicate-free databases with relevant indexes and similarity values allow

	systems by approximate string matching and clustering	across heterogeneous healthcare information systems	using probabilistic methods relying on the Porter-Jaro-Winkler algorithm.	immediate (i.e., real-time) proximity detection when inserting a new identity.
Thornton & Shannon, (2005)	Reducing Duplicate Patient Creation Using a Probabilistic Matching Algorithm in an Open-access Community Data Sharing Environment	Patient Linking and Deduplication within an MPI	Experimental: Use of probabilistic matching algorithms for matching.	The probabilistic matching algorithm facilitates the management of duplicate patient creation and positions IHC to tune further and refine the EMPI processes.
Nelson et al. (2023)	Optimising Patient Record Linkage in a Master Patient Index Using Machine Learning: Algorithm Development and Validation	Machine Learning: Patient Linking and Deduplication within an MPI	Experimental: Machine learning algorithms within an MPI are used for record linkage.	Developing and evaluating a machine learning-based record linkage and deduplication-based tool using synthetic data.
Verschuuren et al. (2020)	Supervised machine-learning techniques for data matching based on similarity metrics	Machine Learning and Record Linkage for identifying the same	Experimental: Machine learning algorithms are used to record linkage.	Developing and evaluating a machine learning-based record linkage and deduplication-based tool using synthetic data.

		entity in data sources		
Pavneet (2020)	A comparison of machine learning classifiers for use on historical record linkage	Machine Learning and Record Linkage for identifying the same entity in data sources	Experimental: Algorithms used include support vector machine and random forests.	The experimental results show that the Random Forest classifier implemented using the additional attributes produced the highest linkage rate.
Acheson, Volpi & Purves (2020)	Machine learning for cross-gazetteer matching of natural features	Machine Learning and Record Linkage for identifying the same entity in data sources	Experimental: Algorithms include rule-based matching and machine learning (random forests)	Machine learning using random forests offered better performance and greater flexibility, obviating the need to manually align feature types and tune thresholds.
Christen (2008)	Automatic Record Linkage using Seeded Nearest Neighbour and Support Vector Machine Classification	Machine Learning and Record Linkage for identifying the same entity in data sources	Experimental: Algorithms include seeded nearest neighbour and support vector machine classification	The author discovered that utilising nearest-neighbour-based and iterative refinement of SVM led to improved pair classification results compared to other methods.

## **2.7 Chapter Summary**

This chapter highlights the importance of Health Information Exchange (HIE) in improving patient care by enabling healthcare providers to access and share electronic medical records. It also discusses challenges with patient identification and emphasises the need for collaboration and technological advancements to address these issues. The chapter underscores the significance of a unified view of a patient's health record using an MPI system. It suggests methods for mitigating the impact of duplicate records in healthcare systems. Furthermore, this chapter explores the use of machine learning to address the challenges of record linkage and deduplication in the healthcare industry.

## **CHAPTER THREE**

### **RESEARCH METHODOLOGY**

This chapter outlines the research methodology for the study, including the research philosophy, approach, methodology, ethics, and strategy employed. It also emphasises the ethical considerations.

#### **3.1 Research Philosophy**

Research philosophy is a cornerstone that determines the most appropriate approach to conducting research. The researcher considered different research approaches to arrive at a worldview. Furthermore, research philosophy is a belief and assumption for expanding knowledge (Saunders et al., 2009). Consequently, this study's philosophical approach is positivist and forms part of the outer layer of the research onion. This is because positivism focuses on a method that is entirely scientific and empiricist in nature and aims to produce pure data and facts unaffected by human interpretation or bias (Saunders et al., 2009). Furthermore, this worldview promises to create detailed and accurate knowledge. Therefore, this study's approach concentrates solely on a practical problem to solve.

#### **3.2 Research Approach**

This study started by leveraging the deductive approach, whereby the researcher began with an abstract problem and various claims and speculations. Furthermore, this was a good starting point for understanding the research topic better and narrowing it down to a more specific problem and aim (Ho, 2006). In addition, the deductive approach was used in this exploratory study after consulting various academic literature (Saunders et al., 2009). This study further used the deductive approach to aid the researcher in exploring whether using a machine learning approach to master patient index record linkage and deduplication with artificially generated data is feasible in an MPI system.

#### **3.3 Research Methodology**

Quantitative research focuses on producing numeric data for statistical analysis that can be used to derive critical conclusions from collected data (Sabine & Holland, 2009). With the help of the deductive research approach, this study uses the quantitative method.

### **3.4 Research Strategy**

The research strategy can be viewed as a comprehensive plan introducing the research topic area and focus (Anon, 2021). Furthermore, since the research philosophy, approach, and methodology have been defined, the research strategy aims to test the theory (Saunders et al., 2009). This study employed the experimental research (ER) strategy. The ER strategy involved the meticulous execution of the following steps in the Entity Resolution (ER) design:

Firstly, the study investigated the current literature concerning record linkage and deduplication challenges to establish a clear understanding of the existing research landscape.

Subsequently, the study's objectives and research questions were formulated. Following the formulation of the study objectives, the ER strategy was used to formulate the research design of the study. This involved performing experiments to realise the research objectives. Subsequent to the experimental phase, the study collected and processed data for comprehensive analysis. The accuracy of the predictions dictated the progression to the final phase of the revision of exploratory study based on new insights.

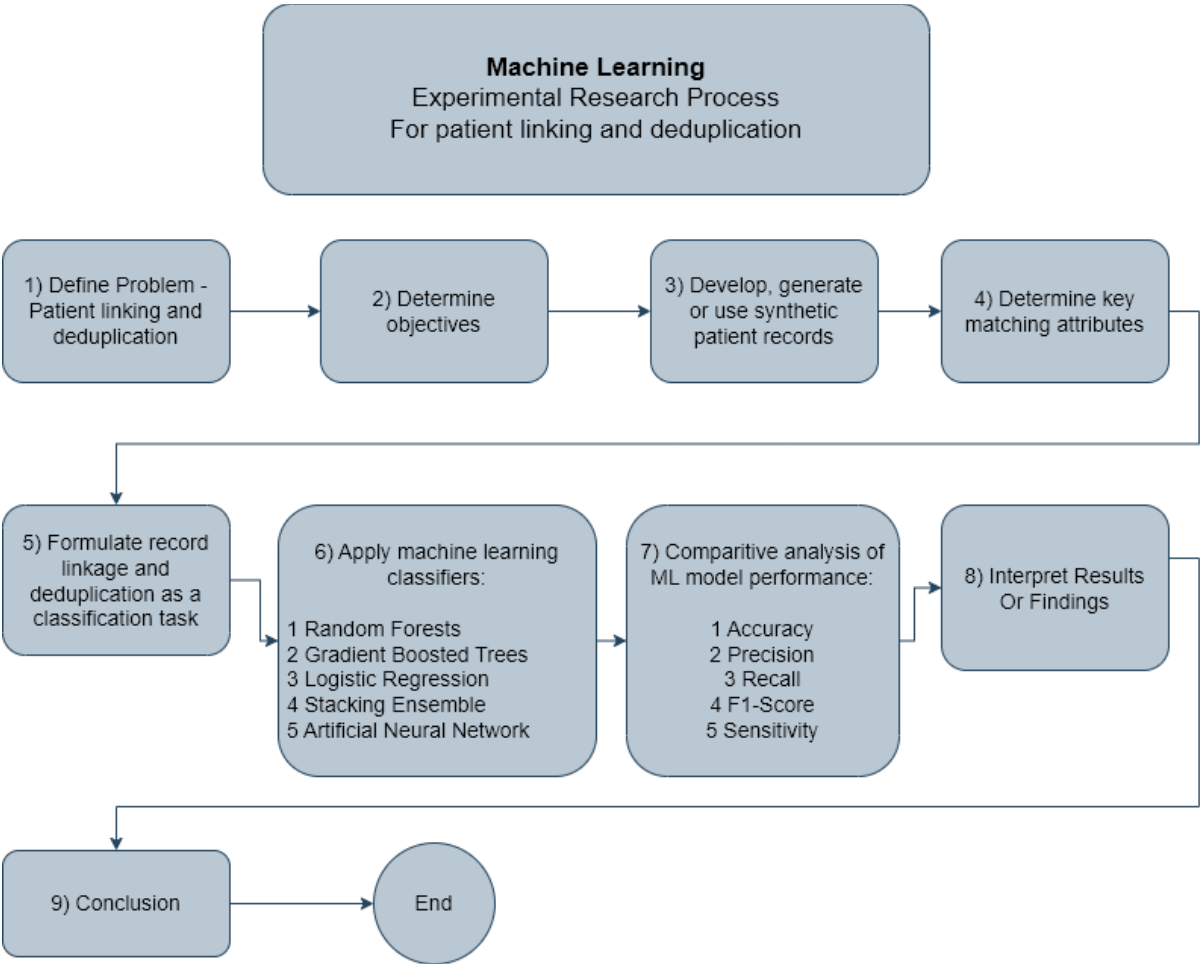
In conclusion, the study drew relevant conclusions and presented the findings to the intended audience in a manner that aligns with academic standards and best practices in research dissemination.

#### **3.4.1 Research Design**

This study used experimental research (ER) to determine if a specific method and treatment applied to a master patient index for record linkage and deduplication influence specific outcomes (Creswell & Cresswell, 2017). Experiments allow the researcher to run tests on synthetic data to collect and validate data, ultimately determining the feasibility and best machine learning approach for record linkage and deduplication. Furthermore, a proper experimental design was used with a pre-test-post-test control group design. This implied splitting our generated patient dataset into known duplicates and unique records. Doing so allowed for checking whether the machine-learning approach could identify duplicate records and flag them as needing to be linked. Furthermore, this information enabled the performance of a post-test validated against the initial dataset for effectiveness and performance.



In Figure 3.1, a diagram of the experimental research process is presented, along with the process for linking and deduplicating patients proposed by this study. Notably, the diagram depicts how the results were verified. The process was repeated for each of the chosen machine learning algorithms, and the process of generating the synthetic data and the attributes used was made available as part of this research.



**Figure 3.1: An overview of the adopted experimental research design**

### 3.4.1.1 Data Sampling and Collection

Synthetic patient data was generated for the custom synthetic dataset in Section 4.4.3 and existing datasets were used in Sections 4.4.1 and 4.4.2 to simulate real-world patient records; the datasets included unique records and known duplicates, following the methods proposed by studies such as Goldstein et al. (2017), Nelson et al. (2023), and Peter (2005). The identified datasets were used to train and evaluate machine learning models for record linkage and deduplication.

### **3.5 Ethical Considerations**

Patient data is inherently sensitive because personal information is included; therefore, a decision was made to utilise a synthetic dataset. Employing a synthetic dataset that contains no identifiable information safeguards the privacy of individuals associated with the data. This approach was chosen in recognition of the importance of protecting patient confidentiality.

This study requires personal information to conduct the research, including name, surname, date of birth, address, phone number, and gender. For this reason, we opted to use the research done by Christen & Pudjijono (2009) to generate a synthetic dataset so that we do not expose confidential information or breach privacy agreements. Additionally, it is a widely adopted approach to use the FEBRL record linkage synthetic datasets to conduct similar experiments that were also used (Nelson et al., 2023; Vo et al., 2019; Heinisch et al., 2019).

1. Data privacy and security: We require personal information to conduct this study, such as name, surname, date of birth, address, phone number and gender. For this reason, we opted to use the research done by Christen & Pudjijono 2009, Peter 2005, and Nelson et al., 2023) to generate a synthetic dataset so that we do not expose confidential information or breach privacy agreements.
2. Open-source licensing: The tools used for evaluation and those developed as part of this research will be made available under the MIT license, a popular open-source license that is permissive and has few limitations.
3. Intellectual Property Rights: The researcher desires that all work produced as part of this research can be used as-is and must adhere to the MIT license with the added condition that attribution must be given to the paper's author.

### **3.6 Chapter Summary**

This chapter provided an overview of the research methodology used, including research methodology, research approach, and ethical considerations. It adopts a positivist philosophical approach and a deductive research approach, using a machine-learning approach with synthetic data. The study utilised a quantitative methodology to produce numeric data for statistical analysis while ensuring data privacy using a synthetic dataset. The research strategy involves experimental research with a pre-test-post-test control group design to ensure valid and reliable results.

## CHAPTER FOUR

### EXPERIMENTATION

This chapter describes the tools, procedures, and systems used to conduct the experiments, outlines the system architecture, and represents the experimentation workflow, model configuration, and datasets.

Additionally, this chapter evaluates the effectiveness of this study's five machine learning models and the manual configuration done. Furthermore, to ascertain the effectiveness of the models, this study employed standard metrics like precision, sensitivity, f1-score, accuracy, and recall. The study will also use plots to visualise the results quickly, making it easier to understand the provided source data.

#### 4.1 System Architecture

Table 4.1 shows the hardware configuration used to experiment with the selected machine-learning models.

**Table 4.1 Workstation Configuration**

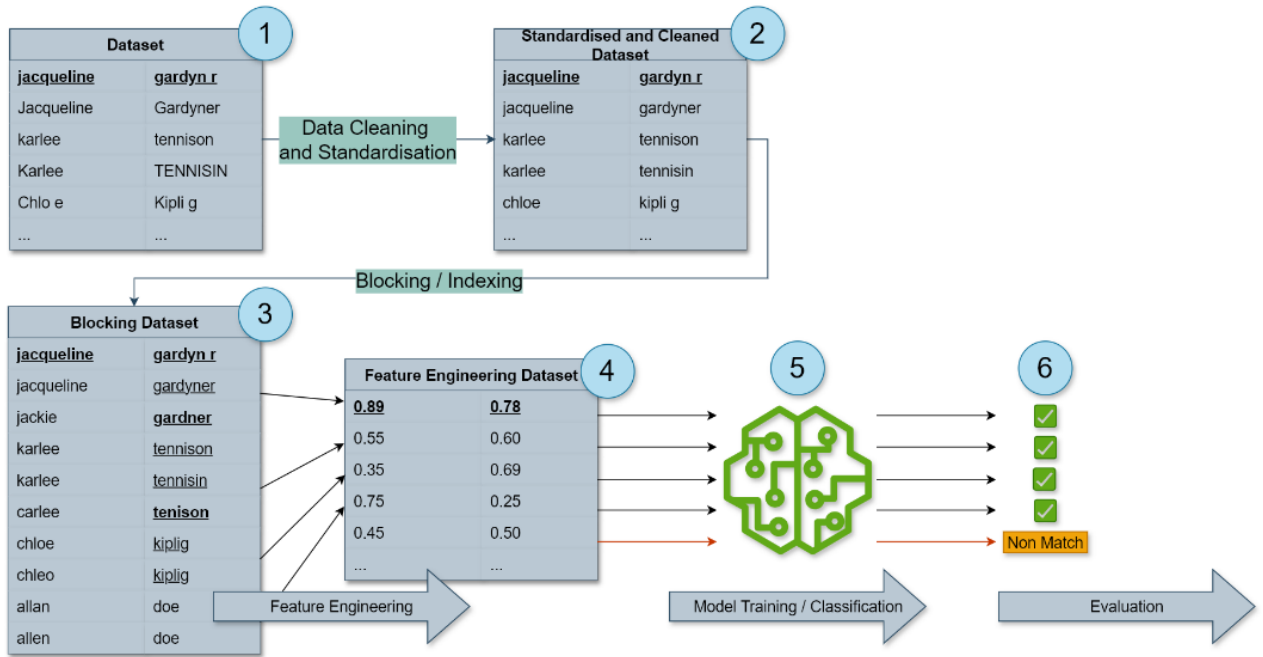
Operating System	Windows 11 Home
CPU	Intel Core Ultra 9 185H 2.3 GHz
GPU	Nvidia RTX 4070 + Intel Arc
RAM	64 GB

#### 4.2 Experimentation Workflow

Figure 4.1 shows the practical machine learning implementation of achieving record linkage and deduplication (Vo et al., 2019; Nelson et al., 2023; Christen, 2008). Records were classified into two major categories: a match or non-match.

Furthermore, all matches were based on the generated synthetic data sources. This study used the fields in Table 4.5 for data cleaning, blocking, and feature engineering.

Additionally, the datasets used for training/classification are described in Table 4.4. The approach followed in this study is deeply rooted in literature (Nelson et al., 2023; Christen, 2008; Christen & Pudjijono, 2009; Christen, 2008a; Christen & Pudjijono, 2009; Morris et al., 2014). The evaluation phase used the five machine-learning models outlined in section 2.5.2.



**Figure 4.1: An overview of the proposed machine learning-based record linkage and deduplication process**

#### 4.2.1 Software, Tools, and Frameworks

Visual Studio Code and Jupyter Notebooks are combined to create an integrated development environment (IDE). This setup was used for experimentation based on the system configuration and the fact that Jupyter Notebooks provide an interactive computational environment for combining code execution, rich markdown text, mathematics, and plots. Python 3.10.11 was used for development and experimentation. The library and frameworks used are discussed below.

**Table 4.2 Libraries Used**

<b>Library</b>	<b>Description</b>	<b>Version</b>
Sklearn	Simple and efficient tools for predictive data analysis. <a href="https://scikit-learn.org/stable/">https://scikit-learn.org/stable/</a>	<b>1.2.2</b>
Recordlinkage	Python record linkage toolkit library. <a href="https://recordlinkage.readthedocs.io/en/latest/">https://recordlinkage.readthedocs.io/en/latest/</a>	<b>0.16</b>
Pandas	An open-source data analysis and manipulation tool, built for Python. <a href="https://pandas.pydata.org/">https://pandas.pydata.org/</a>	<b>2.0.0</b>
NumPy	Bringing computational power of languages like C and Fortran to Python. <a href="https://numpy.org/">https://numpy.org/</a>	<b>1.24.2</b>
Matplotlib	A comprehensive library for creating plots in Python. <a href="https://matplotlib.org/">https://matplotlib.org/</a>	<b>3.9.0</b>
XGBoost	XGBoost is an optimised distributed gradient boosting library for Python. <a href="https://xgboost.readthedocs.io/">https://xgboost.readthedocs.io/</a>	<b>2.1.1</b>

### 4.3 Datasets

This research evaluated the proposed machine learning approach using four synthetic datasets, which include two datasets distributed with the Freely Extensible Biomedical Record Linkage (FEBRL) system (Nelson et al., 2023; Vo et al., 2019), electronic practice-based research network (ePBRN) (Vo et al., 2019), and lastly, a custom-generated synthetic dataset (Goldstein et al., 2017; Peter, 2005). The errors in the datasets are mainly due to human typing mistakes (Peter, 2005; Vo et al., 2019; Nelson et al., 2023). Each dataset has a `rec_id` column that uniquely identifies each record and follows a similar convention to identify a known duplicate. An example of this link and a detailed view of the datasets are shown in Table 4.3.

In addition to understanding how duplicates were generated and identified in the datasets, it was necessary to ensure that duplicates could be easily identified and used with a simplified numeric value. This was achieved by introducing a new `match_id` column in each dataset (Nelson et al., 2023). The `match_id` column was populated with the numeric value from a `rec_id` column and is depicted in Table 4.3.

**Table 4.3 Understanding duplicates in the Datasets**

rec_id	Duplicate rec_id	match_id (applied to original and duplicate)
rec-1070-org	rec-1070-dup-0	1070
rec-4866-org	rec-4866-dup-0	4886
rec-780-org	rec-780-dup-0	780
rec-780-org	rec-780-dup-1	780

For this study, four open-source synthetic datasets were used that are regularly used in the field of record linkage and deduplication (Nelson et al., 2023; Vo et al., 2019; Christen, 2008). This study further selected datasets used in similar studies except for dataset four, which is a custom synthetic dataset used to evaluate algorithm performance against a dataset not usually used (Christen, 2008; Peter, 2005). The datasets used and the record breakdown is explained in Table 4.4.

**Table 4.4 Dataset breakdown**

Dataset Number	Known Duplicates	Total Records	Dataset
1	3000	5000	FEBRL3
2	5000	10000	FEBRL4
3	3192	14078	ePBRN
4	2000	10000	Custom

#### 4.4 Description of the Dataset

As shown in Table 4.4, this study utilised four separate synthetic datasets for experimentation. Furthermore, this study opted to focus on synthetic datasets that are utilised in other studies looking to run similar experiments (Christen, 2008; Christen, 2008; Peter, 2005; Nelson et al., 2023; Vo et al., 2019). Additionally, all datasets are attributed to known duplicates used to train the machine learning models.

#### **4.4.1 FEBRL Datasets**

The freely extensible biomedical record linkage (FEBRL) datasets were used in this study. These datasets were used because they were created specifically because there was a lack of high-quality data for use in medical research split into unique and known duplicate pairs (Christen, 2008). This package contains four separate datasets (Christen, 2008). This study made use of two distinct FEBRL datasets since these packages were developed with error generators in mind (Christen, 2008). This study used the datasets of FEBRL3 (5000 total records) and FEBRL4 (10000 total records). Appendices D and E show the fields provided as part of the FEBRL datasets. It also provides a link to the datasets used during experimentation.

#### **4.4.2 ePBRN**

This study also used the electronic practice-based research network (ePBRN) dataset to conduct additional experiments (Vo et al., 2019). The ePBRN dataset comprises 14078 total records and 3192 known duplicates. Additionally, this dataset exists because the University of New South Wales (UNSW) has been extracting clinical, and administrative data from electronic health records (EHRs) for research improvement purposes (Vo et al., 2019). Appendix C shows the fields provided as part of the ePBRN dataset. It also provides a link to the dataset used during experimentation.

#### **4.4.3 Custom Synthetic Dataset**

This study used the work done by Christen and Pudjijono (2009) to generate an additional synthetic dataset for experimentation. This dataset contains 10000 total records, of which 2000 are known duplicates. Appendix F shows the fields provided as part of the custom dataset. It also provides a link to the dataset used during experimentation.

### **4.5 Data Cleaning**

The data used to perform record linkage will not usually be in a standardised form that can be used for record linkage (Kousthubha & Raghuveer, 2018; Christen, 2008; Vo et al., 2019). In this study, all the datasets used went through a data-cleaning process. The pre-processing process includes removing special characters and spaces, ensuring dates are correctly formatted and parsed into separate fields (day, month, year), and only lowercase is used (Kousthubha & Raghuveer, 2018; Vo et al., 2019).

Furthermore, this study utilised various record-blocking criteria for the FEBRL and ePBRN datasets and did so with the given\_name, surname and postcode fields. For the custom datasets given\_name, surname and phone\_number were used. Blocking is used to limit the number of records considered during the linking process (Vo et al., 2019; Kousthubha & Raghuv eer, 2018).

## 4.6 Feature Selection

The feature selection methods used in this study emulate those employed in previous research and were integrated into the experimentation process to ensure accuracy and provide a solid foundation to build upon (Kousthubha & Raghuv eer, 2018; Vo et al., 2019).

This study used consistent fields during experimentation except for the postcode that was swapped out for a phone number in the custom dataset. The list of fields used in the datasets includes record identifier (rec\_id), first name (given\_name), last name (surname), street number (street\_number), address (address\_1, address\_2, suburb, state, and postcode), phone number (phone\_number), date of birth (day, month, and year), and match identifier (match\_id). The use of match\_id is described in Table 4.3. The match\_id field identified known duplicate records in each dataset. Table 4.5 explains the features used in this study.

**Table 4.5 Feature selection, description, and justification**

Feature Description	Conversion Used & Description	Number of Features
Given Names	Jarowinkler similarity with a threshold of 90%	1
Surnames	Jarowinkler similarity with a threshold of 90%	1
Given Names	Levenshtein with a threshold of 90%	1
Surnames	Levenshtein with a threshold of 90%	1
Exact matches for street number, year, postcode, day, month	Exact Numeric Match	5
Address Line 1	Levenshtein with a threshold of 75%	1
Address Line 2	Levenshtein with a threshold of 75%	1
National Identifier	Numeric comparison	1
Phone Number	Numeric comparison	1



Surname and Given Names	Validate and Check if values are swapped	1
Date Of Birth Day and Month	Validate and Check if Day and Month are swapped	1
Date Of Birth	Validate and check if date of birth was reset to default of (01/01)	1
Surname	Validate and check if surname is joined into a single field	1
Surname and Given Names	Validate if surname and given names are joined with a dash	1
<b>Total Features</b>		<b>18</b>

## 4.7 Model Training and Testing

A blocking process was used to increase the original dataset into a training set of candidate records to train and test the machine learning models used in this study. The blocking process applied considered any two records with a similarity score, based on a text field such as surname, higher than a pre-specified threshold as a suspicious match pair for further classification, significantly reducing the number of pairs to be compared (Vo et al., 2019). The training set was further split based on the current fold, which ranged from 1 to 10. Additionally, the study used the original dataset before blocking was applied to create a test dataset for evaluating the performance of the machine learning models for each hyperparameter and fold used. The training and test split choice was based on research done by Christen (2008) and Vo et al. (2019).

For this study, the researcher opted to tweak the hyperparameters associated with the machine learning model used and further used the tuning values [0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000], and ten folds. The hyperparameter selection was applied to the five machine-learning models used in this study. This study utilised k-fold cross-validation with ten folds to assess the performance of the trained models on new data. This approach involves partitioning the data into k folds, which are then used for training and validation (Daniel Berrar, 2019).

Table 4.5 shows the best hyperparameters used during the model training phase for the five machine learning models used during experimentation. The hyperparameters were selected using grid search, and the tuning values were [0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000] (Vo et al., 2019). The tuning values are

based on a logarithmic scale to more accurately reflect significant changes in machine learning model performance (Vo et al., 2019).

The stacking ensemble model was configured with the logistic regression model as the final estimator, accompanied by additional estimators, including random forest, gradient-boosted trees, and artificial neural networks. The stacking ensemble learning was used to leverage the complementary strengths of base models, enhancing performance and improving generalisation ability (Lu et al., 2023).

**Table 4.5 Machine Learning Model Hyperparameter Configuration and Tuning**

Model	Best Hyperparameter
Random Forests	{ n_estimators: 100, criterion: 'gini', max_depth: 1 }
XGBoost	{ n_estimators: 100, max_depth: 6, learning_rate: 1, gamma: 10, objective: 'binary:hinge' }
Logistic Regression	{ C: 0.001, penalty: 'l2', max_iter: 5000, multi_class: 'ovr' }
Stacking Ensemble	{ stack_method: 'predict', cv: 2, estimators: ['rf', 'xgboost', 'nn'], final_estimator: 'final_estimator' }
Artificial Neural Network	{ alpha: 2000, activation: 'relu' }

The study experiments used code that leveraged existing research, found at <https://github.com/ePBRN/Medical-Record-Linkage-Ensemble>. This strategic choice ensured a robust foundation and allowed for easy modification to accommodate different datasets, features, and machine-learning models. The code utilised included key record linkage, data cleaning, blocking, and evaluation logic based on the work by Vo et al. (2019). It is also important to point out that the work by Vo et al. (2019) made use of key record linkage libraries and features used widely in the field of record linkage (Nelson et al., 2023; Heinisch et al., 2019; Christen, 2008; Kousthubha & Raghuveer, 2018).

## **4.8 Chapter Summary**

This chapter discussed the system architecture used during the experimentation and the workflow and introduced the software, tools, and frameworks used. This chapter also discussed the synthetic datasets used during the experimentation and feature selection processes. In conclusion, the model training process and acknowledgements were addressed.

## CHAPTER FIVE

### EVALUATION

This chapter assessed the effectiveness of five machine learning models and their performance on selected datasets. Standard metrics were used to interpret model performance, including precision, sensitivity, F1-score, accuracy, AUC score, recall, and confusion matrices. The study focused on training machine learning models using synthetic data to evaluate record linkage and deduplication accuracy. The source code for the project implementation is available on GitHub - <https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/README.md>.

#### 5.1 Model Performance Evaluation

This study utilised five machine learning models and various synthetic datasets to determine the best models to include in an MPI solution with higher confidence that new data will also perform well. The best model was selected for each dataset during experimentation. A confusion matrix is provided for each dataset and machine learning model to summarise the models' performance. A confusion matrix is a powerful tool used in classification experiments to assess a system's performance by displaying the number of correctly and incorrectly classified data (Meyer-Baese & Schmid, 2014). This study also calls out the area under the curve (AUC) score for each model against each dataset. The AUC score measures how well a classifier can distinguish between positive and negative classes (Janssens & Martens, 2020).

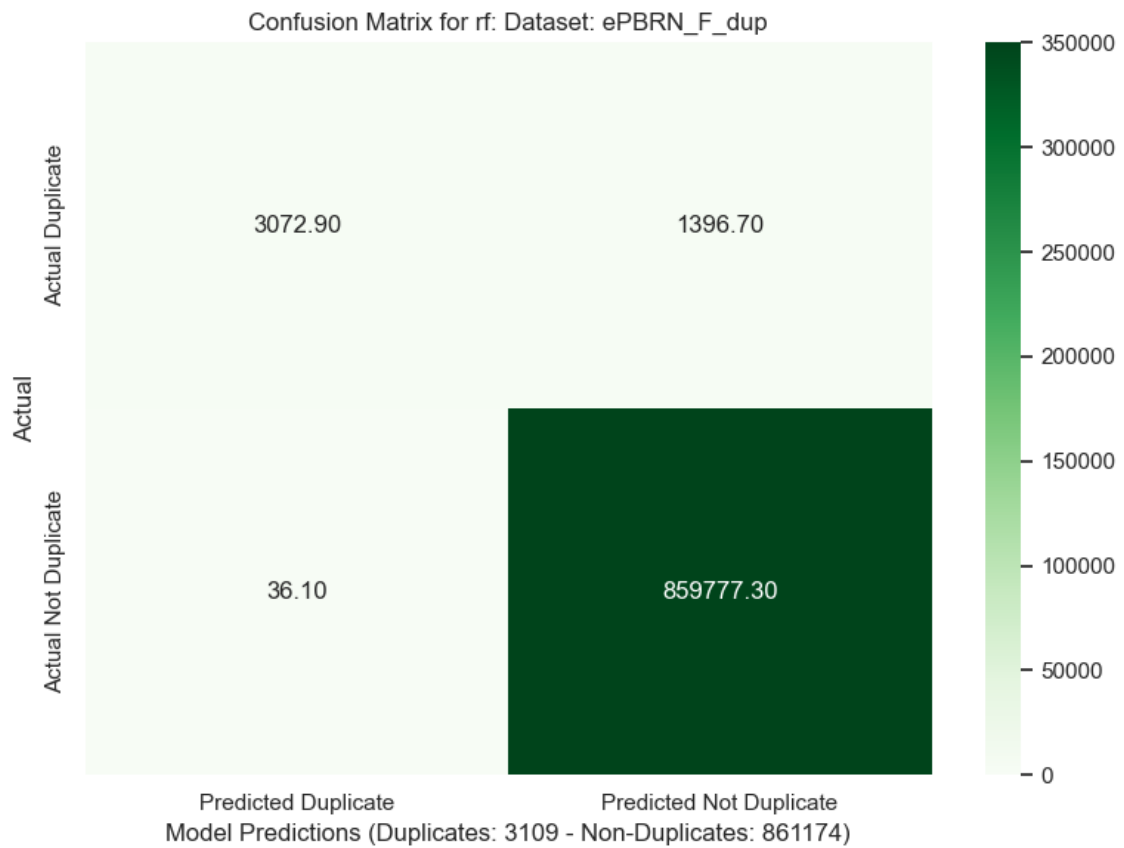
The primary objective of the trained machine learning models was to accurately differentiate between "Duplicate" and "Not Duplicate" records. The effectiveness was assessed using AUC scores and confusion matrices, which provided valuable insights into their performance. These analytical tools enabled informed decision-making by highlighting the strengths and weaknesses of the models employed.

Figures 5.1, 5.2, 5.3, and 5.4 show the confusion matrices for the best-performing machine-learning model based on each dataset. Table 5.1 shows the performance of the ML models on specific datasets.

**Table 5.1 Model Performance on Datasets**

Model	Accuracy	Sensitivity	Precision	Recall	F1-Score	AUC
ePBRN Dataset						
Logistic Regression	99.56%	98.00%	44.70%	98.68%	61.40%	99.93%
Random Forests	99.83%	98.83%	68.75%	98.83%	81.09%	99.98%
XGBoost	86.43%	99.89%	3.01%	99.89%	5.79%	93.14%
MLP-ANN	99.56%	97.79%	44.66%	97.79%	61.32%	99.88%
Stacking Ensemble	99.60%	99.35%	49.10%	99.35%	65.23	99.89%
FEBRL3 Dataset						
Logistic Regression	98.16%	97.95%	77.44%	97.95%	86.50%	99.76%
Random Forests	99.55%	96.98%	95.62%	96.98%	96.29%	99.77%
XGBoost	90.39%	99.88%	43.59%	99.88%	59.01%	94.83%
MLP-ANN	98.22%	97.95%	78.08%	97.95%	86.89%	99.76%
Stacking Ensemble	98.27%	98.26%	78.62%	98.26%	87.30%	99.76%
FEBRL4 Dataset						
Logistic Regression	99.99%	99.52%	94.43%	99.52%	96.91%	99.97%
Random Forests	99.81%	97.47%	87.63%	97.47%	92.29%	99.96%
XGBoost	94.51%	99.92%	25.84%	99.92%	38.65%	97.19%
MLP-ANN	99.93%	99.41%	94.60%	99.41%	96.95%	99.97%
Stacking Ensemble	99.85%	99.49%	88.52%	99.49%	93.68%	99.93%
Custom Synthetically Generated Dataset						
Logistic Regression	99.97%	100%	96.78%	100%	98.36%	99.99%
Random Forests	99.98%	100%	98.40%	100%	99.18%	99.99%
XGBoost	98.58%	99.98%	65.95%	99.98%	76.76%	99.27%
MLP-ANN	99.97%	99.99%	96.77%	99.99%	98.35%	99.99%
Stacking Ensemble	98.70%	100%	79.54%	100%	84.70%	99.99%

### Confusion Matrix for Random Forests: ePBRN Dataset



**Figure 5.1: The ePBRN dataset in which the Random Forest model achieved the best performance**

### Confusion Matrix for Random Forests: FEBRL3 Dataset

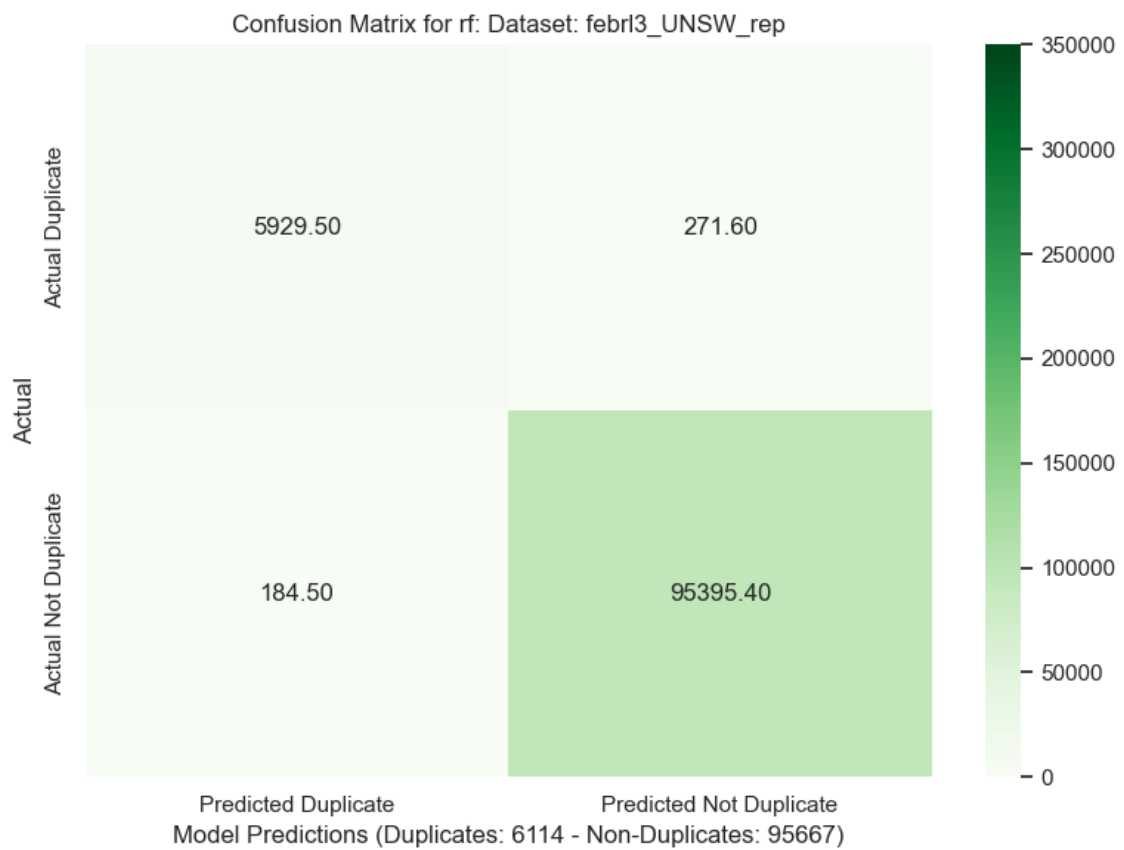
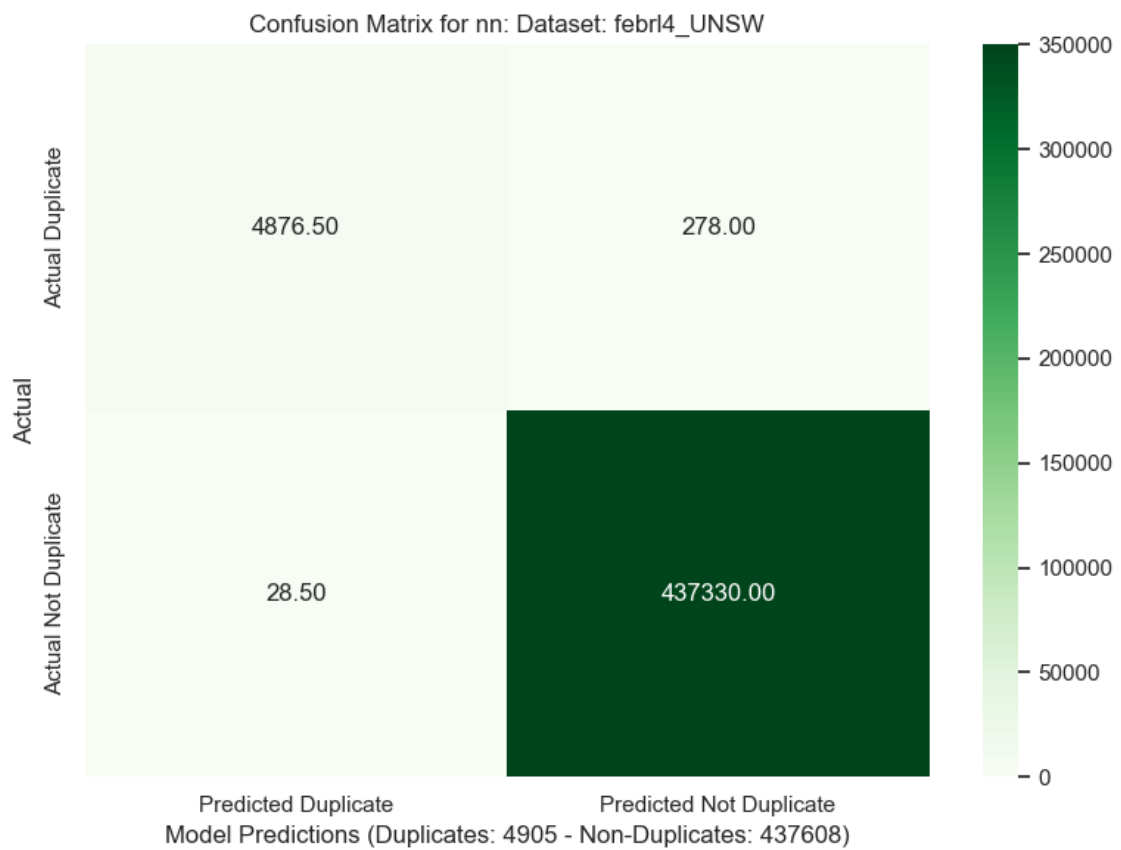


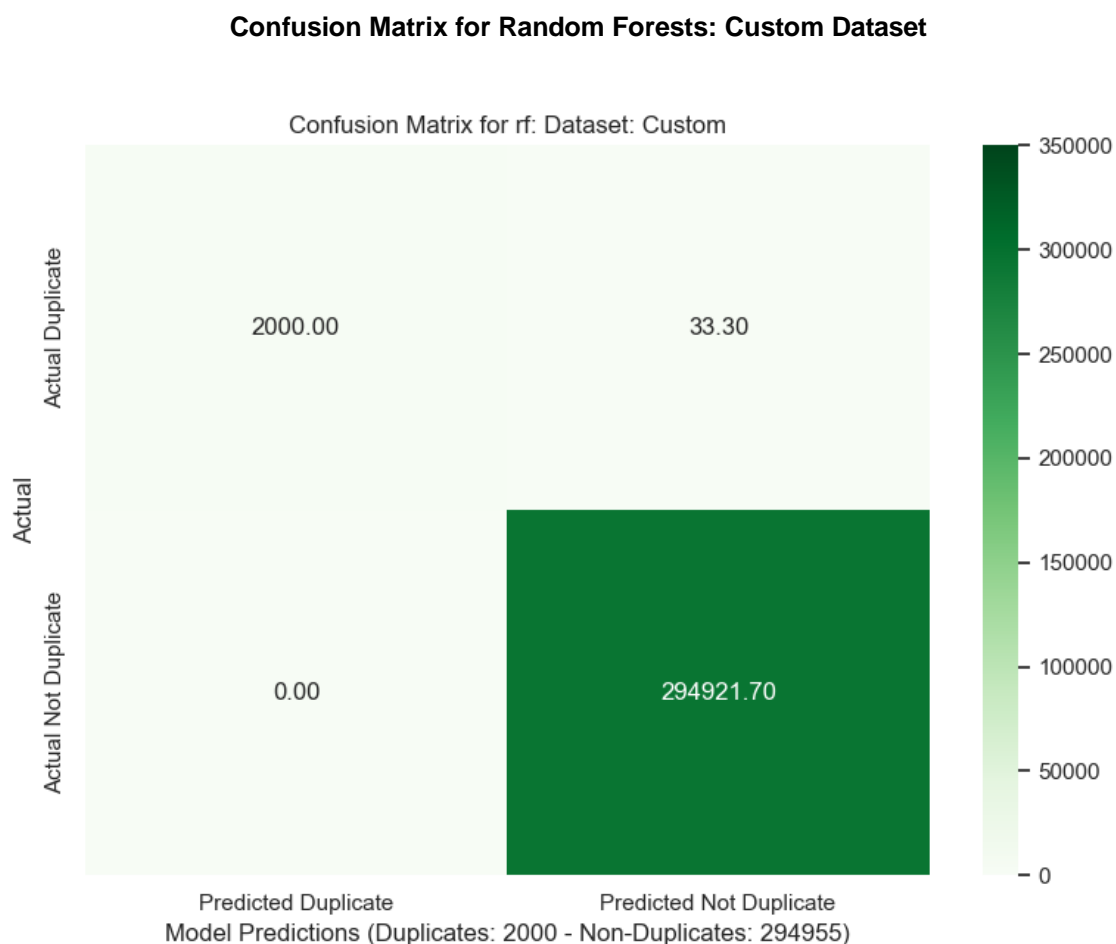
Figure 5.2: The FEBRL3 dataset in which the Random Forest model performed best.

### Confusion Matrix for Neural Network – Multi-Layer Perceptron: FEBRL4 Dataset



**Figure 5.3: The FEBRL4 dataset in which the MLP-ANN model performed best.**





**Figure 5.4: The Custom Dataset in which the Random Forest model achieved the best performance**

## 5.2 Discussion

This study compared the performance of five machine learning models and four synthetic datasets for solving record linkage in the public healthcare sector. The results obtained and the experiments conducted presented record linkage and deduplication as a machine-learning classification task. The datasets provided a good baseline and variance to give the confidence that the code produced by this study can be applied to various datasets. The experiments illustrated that the random forest (RF) model performed best for three (ePBRN, FEBRL3, Custom) of the four datasets. The MLP-ANN model performs best for the FEBRL4 dataset, closely followed by logistic regression (LR). Table 5.2 describes the best-performing models for each dataset and their respective metrics.

**Table 5.2 Summary of the best-performing models for each of the datasets**

Model	Accuracy	Sensitivity	Precision	Recall	F1-Score	AUC
ePBRN Dataset						
Random Forests	99.83%	98.83%	68.75%	98.83%	81.09%	99.98%
FEBRL3 Dataset						
Random Forests	99.55%	96.98%	95.62%	96.98%	96.29%	99.99%
FEBRL4 Dataset						
MLP-ANN	99.93%	99.41%	94.60%	99.41%	96.95%	99.97%
Logistic Regression	99.99%	99.52%	94.43%	99.52%	96.91%	99.97%
Custom Synthetically Generated Dataset						
Random Forests	99.98%	100%	98.40%	100%	99.18%	99.99%

The machine learning models trained on the four synthetic datasets performed well and demonstrated strong potential for enhancing medical record linkage and deduplication. These models effectively identified duplicates and non-duplicates within the datasets, ensuring comprehensive patient record accuracy. By accurately identifying true positives (TP/Actual Duplicates) and minimising false positives (FP/Actual Not Duplicate) and false negatives (FN/Predicted Not Duplicate), the models will support better-informed clinical decisions and improved patient care. Their performance illustrated their capability to improve data accuracy and enhance operational efficiency within the healthcare setting.

In conclusion, the high accuracy demonstrated by these models in distinguishing between "Duplicates" and "Non-Duplicates" reduced the manual effort required for data management, leading to faster, safer, and more efficient processes. These models aimed to assist with automation, preserve patient data integrity, and streamline operations, ultimately contributing to better patient outcomes and more reliable healthcare data management.

### 5.3 Chapter Summary

The experiments run and the results indicate that machine learning can be used for medical record linkage and deduplication. The code produced in this study can be used to implement the functionality into an MPI solution (Nelson et al., 2023). This study utilised five different machine-learning models with varied datasets and hyperparameters tuning. This approach has proven valuable as, even though the models perform well overall, the researcher observed that a model's performance can vary based on the given dataset. This leads the researcher to conclude that no single solution will fit all use cases, and multiple models will need to be tested and evaluated before implementing them within an MPI solution for record linkage and

deduplication. The study found that the random forest (RF) and the logistic regression (LR) models consistently outperformed regardless of the dataset, making them ideal for general-purpose use.

## **CHAPTER SIX**

### **SUMMARY, CONCLUSION, AND RECOMMENDATIONS**

This chapter thoroughly reviews the study's findings, conclusions, and recommendations. It discusses the study's objectives and how they were achieved in the summary section. Additionally, this chapter serves as a conclusion of the study, providing specific insights into its implications. Finally, the chapter discusses recommendations based on the findings and conclusions and provides key insights into future research.

#### **6.1 Summary**

This study aimed to explore machine learning algorithms and their effectiveness when used to solve the problem of record linkage and deduplication. It also aimed to explain how these advanced algorithms can be included in an MPI system to provide better care to patients in an HIE. This study comprised six chapters, each contributing to the overall research.

Chapter 1 provided an overview of the study, including the background, research problem, aim, research questions, and objectives. Chapter 2 conducted a comprehensive discussion of the literature, introducing key concepts, discussing key technologies, explaining record linkage and deduplication, and introducing machine learning as a possible solution. Chapter 3 provided an overview of the research philosophy, approach, methodology, strategy, and ethical considerations. Chapter 4 described the experimentation and evaluation process, providing a comprehensive explanation of the process used during experimentation and the tools and methods used to produce the results. Finally, Chapter 6 presented a summary of the study, contributions, recommendations, and a conclusion.

The following section describes the study's objectives and how each objective was met:

1. To determine the criteria for identifying duplicate records within a data source.

Chapter 2 provided an in-depth review of the current literature on record linkage and deduplication and the use of synthetic datasets. The chapter highlighted research on the essential characteristics required to experiment with record linkage and deduplication. Furthermore, the literature review discussed the issue of record linkage in the healthcare field because no universally unique record identifier exists, and there is no standard way to achieve this.

Given these issues, personally identifiable fields such as first name, surname, date of birth, gender, and address, combined with advanced algorithms, are needed to identify records even when no unique identifier exists. This is how record linkage and deduplication are achieved regardless of the algorithm used.

2. To formulate record linkage and deduplication as a machine learning classification task.

This objective was achieved by doing extensive research on record linkage and deduplication, the consequences of poor patient identification, the technologies currently used in the field, and the application of machine learning to solve this issue. This study highlighted that no single solution exists for solving this very complex issue but that machine learning introduces many efficiencies that would otherwise be lacking. In conclusion, machine learning was very effective when classifying records into a match or a non-match group.

3. To apply selected machine learning algorithms for record linkage and deduplication.

It was essential to consult existing research in record linkage and deduplication to achieve this objective. This was needed due to some limitations when doing a study like this. The limitations are further elaborated on in section 6.3. This process was valuable since the researcher leveraged existing research to use synthetic datasets and generate a custom dataset used for experimentation. Once the datasets were produced and aligned with related studies, the next phase was to select five machine learning models to train on the provided datasets.

The machine learning models used for this study were random forests, extreme gradient boosting, logistic regression, stacking ensemble, and artificial neural networks. Additionally, the datasets used were two FEBRL (3 and 4) datasets, ePBRN, and a custom-generated dataset based on literature.

The trained models all performed well during the evaluation phase of the study. Based on the results achieved, this objective is satisfied.

4. To evaluate the performance of the selected machine learning algorithms for record linkage and deduplication.

This objective was achieved not only based on the findings of this study but also by using some of the datasets used in the industry and implementing some of the logic for

record linkage and deduplication produced by Vo et al. (2019). Using some of the implementation of this research in combination with this study's machine-learning models and four datasets provides a high degree of confidence that machine-learning algorithms perform well when applied to record linkage and deduplication.

Additionally, the study used standard evaluation metrics like accuracy, precision, recall, f-score, sensitivity, AUC score and confusion matrices to assess the trained model's effectiveness in classifying records as either a "Duplicate" or "Non-Duplicate".

False links were captured in the final results using metrics introduced by Vo et al. (2019), which count the overall number of links, but these values are calculated based on the confusion matrix. The false counts (FC) are calculated as  $(FP + FN)$ . Furthermore, the total number of links is calculated as  $(TP + FN)$ . This was done because we knew the number of unique and duplicate records at the time of the experiment. This provides an excellent way to explain the model's effectiveness based on the number of false record links. This was illustrated for each dataset and model using the confusion matrix plots.

In conclusion, using both standard evaluation metrics, five machine learning models, four datasets and a custom metric which measures the number of false links allowed the researcher to gain a comprehensive understanding of the chosen machine learning model's strengths and limitations when applied to the classification task of record linkage and deduplication.

## **6.2 Contributions of the Study**

It is currently known that patient record linking is a big challenge (Fernandes & O'Connor, 2015; Kousthubha & Raghuveer, 2018; Beth et al., 2016). The study aimed to leverage research done in record linking and deduplication and applied it specifically to healthcare data (Winkler, 2002; Beth et al., 2016; Menachemi et al., 2018). This study's theoretical and practical contributions are outlined in sections 6.2.1 and 6.2.2, respectively.

### **6.2.1 Theoretical Contribution**

Machine learning is a viable alternative to existing record linking and deduplication methods, and the researcher applied different techniques to test the four datasets and record the results. Furthermore, many machine-learning algorithms exist that can solve the issue of duplicate record detection (Liang et al., 2018; Almaspoor et al., 2021; Verschuuren et al.,

2020; Christen, 2008; Pavneet, 2020; Nelson et al., 2023). Still, they must be combined with various string cleaning and comparison algorithms to be effective.

The study was one of the first contributions to the public healthcare domain that utilised various datasets and machine learning models for record linkage and deduplication in the context of being applied to an MPI system (Beth et al., 2016; Fernandes & O'Connor, 2015; Duggal et al., 2015). These systems often include admissions, dispensaries, and others, all of which may use their medical record number (MRN). Healthcare providers are also becoming more interoperable but still lack a standardised approach to sharing information (Duggal et al., 2015). Data is stored in different formats, and policies and procedures for capturing and storing information are not consistently implemented (Riplinger et al., 2020a; Fernandes & O'Connor, 2015).

The researcher looked to bring together existing knowledge from various areas and produced a new body of work that can be used to solve current and future record linkage and deduplication issues within the healthcare space. Moreover, public healthcare facilities must identify and prevent duplicate records from multiple source systems to be effective.

Additionally, the researcher used the following machine-learning algorithms:

1. Random Forests
2. Stacking Ensemble
3. Logistic Regression
4. Artificial Neural Network – Deep Multi-Layer Perceptron
5. XGBoost

The algorithms were applied in the context of record linkage and duplication classification to ascertain the performance of each algorithm and, importantly, to visualise the results and how they could be used effectively in hospital care settings to be more efficient and effective.

### **6.2.2 Practical Contribution**

Finding a machine learning approach to record linkage and deduplication produced some key artefacts, including a sample code developed to generate test data and the deduplication and matching machine learning implementation code that can be included in an MPI system.

In addition, this study provides artefacts that can be used as an essential tool in processes (admission, laboratories, and emergencies) where reducing the issue of record linking and duplication related to electronic medical records is necessary. This is substantial since MPI systems are the bridge between various hospital HIE systems.

### **6.3 Limitations of the Study and Potential Impact on the Results**

The main limitation of this study is the use of synthetic data, which has the potential not to reflect real-world data accurately. To that end, this study consulted literature to ensure the synthetic data being used aligns with the standards within the public healthcare space (Vo et al., 2019; Nelson et al., 2023; Christen & Pudjijono, 2009; Christen, 2008a; Peter, 2005; Christen, 2008b; Kousthubha & Raghuveer, 2018). Furthermore, this study will make the synthetic data and source code that can be used freely available for further research.

### **6.4 Conclusion**

This study discovered that no single model consistently produced the best results while evaluating the selected machine learning models (random forests, gradient-boosted trees, logistic regression, stacking ensemble, artificial neural network). This finding, exacerbated by using multiple datasets (ePBRN, FEBRL, Custom), underscores the necessity of conducting experiments before choosing a model for record classification. The study's findings were communicated through visual representations during experimentation, highlighting their importance.

The researcher finds that this study successfully addressed the objectives set out and that valuable insights were provided that will aid further research in this field in both a theoretical and practical sense.

### **6.5 Recommendations and Future Work**

In Section 6.3, the researcher discusses one of the critical limitations of the study, which was the lack of real-world data. The researcher would like to conduct a further study on real-world data that has known duplicates identified or has the human resources to classify data.

Additionally, the researcher would like to incorporate the trained model with the best performance into an MPI system and use it to classify records into duplicate or unique categories. This will capture valuable insights regarding real-world effectiveness and combine technology with human evaluation to enhance model performance for a healthcare setting.



In conclusion, this research has set a good foundation for future work using real data. Utilising real data in a healthcare setting with human input will serve as another key metric for using machine learning in record linkage and deduplication with healthcare tools such as an MPI system.

## REFERENCES

- Acheson, E., Volpi, M. & Purves, R.S. 2020. Machine learning for cross-gazetteer matching of natural features. *International Journal of Geographical Information Science*, 34(4): 708–734.
- Almaspoor, M.H., Safaei, A., Salajegheh, A., Minaei-Bidgoli, B. & Safaei, A.A. 2021. Support Vector Machines in Big Data Classification: A Systematic Literature Review. <https://doi.org/10.21203/rs.3.rs-663359/v1>.
- Asher, J., Resnick, D., Brite, J., Brackbill, R. & Cone, J. 2020. An introduction to probabilistic record linkage with a focus on linkage processing for WTC registries. *International Journal of Environmental Research and Public Health*, 17(18).
- Beth, J.H., Marc, D., Munns, M. & Sandefer, R. 2016a. Why patient matching is a challenge: research on master patient index (MPI) data discrepancies in key identifying fields. *Online Research Journal: Perspectives in Health Information Management*, 13(Spring 2016).
- Boateng, E.Y. & Abaye, D.A. 2019. A review of the logistic regression model with emphasis on medical research. *Journal of Data Analysis and Information Processing*, 07(04): 190–207.
- Brignone, E., Fargo, J.D., Blais, R.K. & Gundlapalli, A. V. 2018. Applying machine learning to linked administrative and clinical data to enhance the detection of homelessness among vulnerable veterans. *AMIA ... Annual Symposium proceedings. AMIA Symposium, 2018*, 305–312.
- Centers for Disease Control. 2018. Bridging public health and health care. <https://www.cdc.gov/surveillance/innovation/sharing-data-digitally.html> [4 May 2022].
- Chouffani, R. 2017. Patient matching algorithms tackle medical record duplication. <https://searchhealthit.techtarget.com/tip/Patient-matching-algorithms-tackle-medical-record-duplication> [4 May 2022].
- Christen, P. 2008a. *Automatic record linkage using seeded nearest neighbour and support vector machine classification*.  
<http://users.cecs.anu.edu.au/~Peter.Christen/publications/kdd2008christen.pdf> (7 June 2022).
- Christen, P. 2008b. Febrl: An open source data cleaning, deduplication and record linkage system with a graphical user interface. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM: 1065–1068.

- Christen, P. & Pudjijono, A. 2009. Accurate synthetic generation of realistic personal information. *Advances in Knowledge Discovery and Data Mining. PAKDD 2009. Lecture Notes in Computer Science*, vol 5476: 507–514.
- Coleman, B., Kang, W.-C., Fahrbach, M., Wang, R., Hong, L., Chi, E. & Cheng, D. 2023. Unified embedding: Battle-tested feature representations for Web-Scale ML systems. *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS 2023)* 56234-56255
- Creswell, J.W. & Cresswell, D.J. 2017. *Research Design: Qualitative, quantitative, and mixed methods approaches*. FIFTH EDITION. Toronto: SAGE Publications, Inc.
- Daniel Berrar. 2019. Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*, 1: 542-545.
- Deng, H., Zhou, Y., Wang, L. & Zhang, C. 2021. Ensemble learning for the early prediction of neonatal jaundice with genetic features. *BMC Medical Informatics and Decision Making*, 21(1): 338.
- Duggal, R., Khatri, S.K. & Shukla, B. 2015. Improving patient matching: single patient view for clinical decision support using big data analytics. In *Noida, India*: IEEE.
- Fernandes, L. & O'Connor, M. 2015. Accurate patient identification - A global challenge. *Perspectives in Health Information Management*: 1–6.
- Goldstein, H., Harron, K. & Cortina-Borja, M. 2017. A scaling approach to record linkage. *Statistics in Medicine*, 36(16): 2514–2521.
- Grannis, S.J., Overhage, J.M. & McDonald, C. 2004. *Real world performance of approximate string comparators for use in patient matching*. Amsterdam: IOS Press.
- Habib, N. & Rahman, M.M. 2021. Diagnosis of corona diseases from associated genes and X-ray images using machine learning algorithms and deep CNN. *Informatics in Medicine Unlocked*, 24: 100621.
- Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimae, M., Barreto, M.L. & Goldstein, H. 2017. Challenges in administrative data linkage for research. *Big Data & Society*, 4(2): 2053951717745678.

Hayler, A. 2018. How data duplication in healthcare is diagnosed. <https://searchdatamanagement.techtarget.com/tip/How-data-duplication-in-healthcare-is-diagnosed> [4 May 2022].

Heinisch, D., Koenig, J. & Otto, A. 2019. The IAB-INCHER project of earned doctorates (IIPED): A supervised machine learning approach to identify doctorate recipients in the German integrated employment biography data. *IAB-Discussion Paper*, 201913.

Ho, Y.C. 2006. Abduction, deduction, and induction: their implications to quantitative methods. *Work*, 480: 812–845.

Janssens, A.C.J.W. & Martens, F.K. 2020. Reflection on modern methods: revisiting the area under the ROC Curve. *International Journal of Epidemiology*, 49(4): 1397–1403.

Karr, A.F., Taylor, M.T., West, S.L., Setoguchi, S., Kou, T.D., Gerhard, T. & Horton, D.B. 2019. Comparing record linkage software programs and algorithms using real-world data. *PLoS ONE*, 14(9).

Khan, M.Y., Qayoom, A., Nizami, M.S., Siddiqui, M.S., Wasi, S. & Raazi, S.M.K.-R. 2021. Automated prediction of Good Dictionary EXamples (GDEX): A comprehensive experiment with distant supervision, machine learning, and word embedding-based deep learning techniques. *Complexity*, 2021: 1–18.

Kousthubha, A.K. & Raghuveer, K. 2018. A survey on record linkage. *The National Institute of Engineering, Mysuru, India*, 3(5): 460–463.

Liang, Z., Huang, J. & Chan, S. 2018. Enterprise master patient index entity recognition by long short-term memory network in electronic health systems. *Electronic Health Systems*. 1-4. <http://dx.doi.org/10.14236/ewic/HCI2018.181>

Lu, M., Hou, Q., Qin, S., Zhou, L., Hua, D., Wang, X. & Cheng, L. 2023. A stacking ensemble model of various machine learning models for daily runoff forecasting. *Water*, 15(7): 1265.

McCoy, A.B., Wright, A., Kahn, M.G., Shapiro, J.S., Bernstam, E.V. & Sittig, D.F. 2012. Matching identifiers in electronic health records: implications for duplicate records and patient safety. *BMJ Journals*, 22(3): 219–224.

Menachemi, N., Rahurkar, S., Harle, C.A. & Vest, J.R. 2018. The benefits of health information exchange: an updated systematic review. *Journal of the American Medical Informatics Association*, 25(9): 1259–1265.

Meyer-Baese, A. & Schmid, V. 2014. Foundations of neural networks. In *Pattern Recognition and Signal Analysis in Medical Imaging*. Elsevier: 197–243.

Morris, G., Farnum, G., Afzal, S., Robinson, C., Greene, J. & Coughlin, C. 2014. Patient identification and matching. *Final Report*. : 93.

Naskath, J., Sivakamasundari, G. & Begum, A.A.S. 2023. A study on different deep learning algorithms used in deep neural nets: MLP SOM and DBN. *Wireless Personal Communications*, 128(4): 2913–2936.

Nelson, W., Khanna, N., Ibrahim, M., Fyfe, J., Geiger, M., Edwards, K. & Petch, J. 2023. Optimizing patient record linkage in a master patient index using machine learning: algorithm development and validation. *JMIR Formative Research*, 7: e44331.

Pavneet, K. 2020. A comparison of machine learning classifiers for use on historical record linkage (Doctoral dissertation). University of Guelph.

Peter, C. 2005. Probabilistic data generation for deduplication and data linkage. *Intelligent Data Engineering and Automated Learning - IDEAL 2005. IDEAL 2005. Lecture Notes in Computer Science*, vol 3578. Springer, Berlin, Heidelberg. [https://doi.org/10.1007/11508069\\_15](https://doi.org/10.1007/11508069_15)

Pita, R., Pinto, C., Barreto, Marcos, Sena, S., Fiaccone, R., Amorim, L. & Barreto, Mauricio. 2017. Design and evaluation of probabilistic record linkage methods supporting the Brazilian 100-million cohort initiative. *International Journal of Population Data Science*, 1(1).

Riplinger, L., Piera-Jiménez, J. & Dooling, J.P. 2020. Patient identification techniques - Approaches, implications, and findings. *Yearbook of Medical Informatics*, 29(1): 81–86.

Sabine, G. & Holland, J. 2009. *Quantitative and qualitative methods in impact evaluation and measuring results*. Birmingham. UK: GSDRC, University of Birmingham.

Saripalle, R., Runyan, C. & Russell, M. 2019. Using HL7 FHIR to achieve interoperability in patient health record. *Journal of Biomedical Informatics*, 94.

- Sauleau, E.A., Paumier, J.P. & Buemi, A. 2005. Medical record linkage in health information systems by approximate string matching and clustering. *BMC Medical Informatics and Decision Making*, 5.
- Saunders, M., Lewis, P. & Thornhill, A. 2009. Understanding research philosophies and approaches. In *Research Methods for Business Students. 8th edition*. Pearson Education: 128–171.
- Thorell, L., Molin, J.D., Fyfe, J., Hone, S. & Lwin, S.M. 2019. Working towards a master patient index and unique identifiers to improve health systems: the example of Myanmar. *South-East Asia Journal of Public Health*, 8(2): 83–86.
- Thornton, S.N. & Shannon, H.K. 2005. Reducing duplicate patient creation using a probabilistic matching algorithm in an open-access community data-sharing environment. *AMIA Annual Symposium*, 2005(2005): 1135–1136.
- Toth, C., Durham, E. & Malin, B. 2014. SOEMPI: A Secure open enterprise master patient index software toolkit for private record linkage. *AMIA Annual Symposium Proceedings*, 2014(2014): 1105–1114.
- Tromp, M., Ravelli, A.C., Bonsel, G.J., Hasman, A. & Reitsma, J.B. 2011. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. *Journal of Clinical Epidemiology*, 64(5): 565–572.
- Verschuuren, P., Palazzo, S., Powell, T., Sutton, S., Pilgrim, A. & Giannelli, M.F. 2020. Supervised machine learning techniques for data matching based on similarity metrics. <http://arxiv.org/abs/2007.04001> [23 May 2022].
- Vo, K., Jonnagaddala, J. & Liaw, S.T. 2019. Statistical supervised meta-ensemble algorithm for medical record linkage. *Journal of Biomedical Informatics*, 95.
- Winkler, W.E. 2009. Handbook of statistics. *ScienceDirect*, 29(Part A): 351–380.
- Winkler, W.E. 2002. Methods for record linkage and Bayesian networks. : 27. Technical report, Statistical Research Division, U. S. Census Bureau.

## APPENDICES

### APPENDIX A: ETHICS CERTIFICATE

**Office of the Research Ethics Committee**  
Faculty of Informatics and Design  
Room 2.09  
80 Roeland Street  
Cape Town  
Tel: 021-469 1012  
Email: [ndedem@cput.ac.za](mailto:ndedem@cput.ac.za)  
Secretary: Mziyanda Ndede

01 February 2023

Mr Dane Hollenbach  
c/o Department of Information Technology  
CPUT

**Reference no:** 209113723/2023/1

**Project title:** A machine learning approach for master patient index record linkage and deduplication

**Approval period:** 1 February 2023 – 31 December 2024

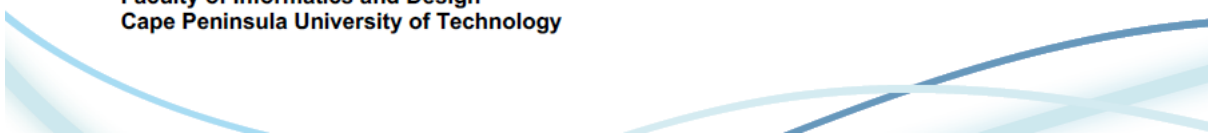
This is to certify that the Faculty of Informatics and Design Research Ethics Committee of the Cape Peninsula University of Technology approved the methodology and ethics of Mr Dane Hollenbach (209113723) for MICT: IT (Magister Technologiae: Information and Communication Technology).

Any amendments, extension or other modifications to the protocol must be submitted to the Research Ethics Committee for approval.

The Committee must be informed of any serious adverse event and/or termination of the study.



**Dr Blessing Makwambeni**  
Acting Chair: Research Ethics Committee  
Faculty of Informatics and Design  
Cape Peninsula University of Technology



## APPENDIX B: USE OF SYNTHETIC DATA ACKNOWLEDGEMENT



PO Box 1908, Bellville, 7535 | Symphony Way, Bellville, Cape Town, South Africa  
+ 27 (0)21 959 6767 | www.facebook.com/cput.ac.za | info@cput.ac.za | www.cput.ac.za

Prof Justine Olawande Daramola  
Associate Professor  
Faculty of Informatics and Design  
Department of Information Technology

29 May 2023

To: Whom it may concern

### Exception from Consent Letter from an Organization

The proposed study by Dane Hollenbach (student Number 209113723) titled **A machine learning approach for master patient index record linkage and deduplication** for the Master of ICT at the Department of Information of the Cape Peninsula University of Technology does not require consent from any person or organization. The study involves experimentation on free and publicly available data sets using machine learning algorithms.

The researcher will use synthetically generated data based on research done by Peter (2005) in his paper titled: *Probabilistic Data Generation for Deduplication and Data Linkage* by using an adapted python script to programmatically generate data needed for the study.

The modality for this was explained in the ethics application that was made and subsequently approved by the FID Ethics Committee (Ref no: 209113723/2023/1).

Thank you.

Prof Justine Olawande Daramola  
Email: daramolaj@cput.ac.za





### APPENDIX C: ePBRN Dataset

Dataset source: [https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/ePBRN\\_F\\_dup.csv](https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/ePBRN_F_dup.csv)

Fields	Rec_id, given_name, surname, street_number, address_1, address_2, suburb, postcode, state, day, month, year, match_id
Total Records	14078
Know Duplicate Records	3192

### APPENDIX D: FEBRL3 Dataset

Dataset source: [https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/febrl3\\_UNSW\\_rep.csv](https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/febrl3_UNSW_rep.csv)

Fields	Rec_id, given_name, surname, street_number, address_1, address_2, suburb, postcode, state, day, month, year, match_id
Total Records	5000
Know Duplicate Records	3000

### APPENDIX E: FEBRL4 Dataset

Dataset source: [https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/febrl4\\_UNSW.csv](https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/febrl4_UNSW.csv)

Fields	Rec_id, given_name, surname, street_number, address_1, address_2, suburb, postcode, state, day, month, year
Total Records	10000
Know Duplicate Records	5000

## APPENDIX F: Custom Dataset

Dataset source: <https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/test5.csv>

Fields	Rec_id, culture, sex, given_name, surname, street_number, address_1, state, date_of_birth, phone_number, national_identifier, blocking_number, address_2
Total Records	10000
Know Duplicate Records	2000

## APPENDIX G: Machine Learning Model Results Per Dataset

The results associated with each dataset follow a similar schema. The results are stored in CSV format, but each row is stored as a JSON object that can be split into unique columns and rows. This was done to capture as much detail as possible during the evaluation phase.

The fields captured during the evaluation phase were fold, model, hyperparameter, precision, sensitivity, f-score, accuracy score (accuracyScore), confusion matrix (confusionMatrix), recall score (recallScore), number of false links (noFalse), and number of links (noLinks).

Dataset	Results
ePBRN	<a href="https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/result_ePBRN_F_dup.csv">https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/result_ePBRN_F_dup.csv</a>
FEBRL3	<a href="https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/result_febrl3_UNSW_rep.csv">https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/result_febrl3_UNSW_rep.csv</a>
FEBRL4	<a href="https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/result_febrl4_UNSW.csv">https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/result_febrl4_UNSW.csv</a>
Custom	<a href="https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/test5.csv">https://github.com/DHollenbach/record-linkage-and-deduplication/blob/main/test/test5.csv</a>