UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

A FEDERATED LEARNING MODEL FOR THE DETECTION OF INSURANCE CLAIMS FRAUD.

KATIECHI STEPHEN OKENO

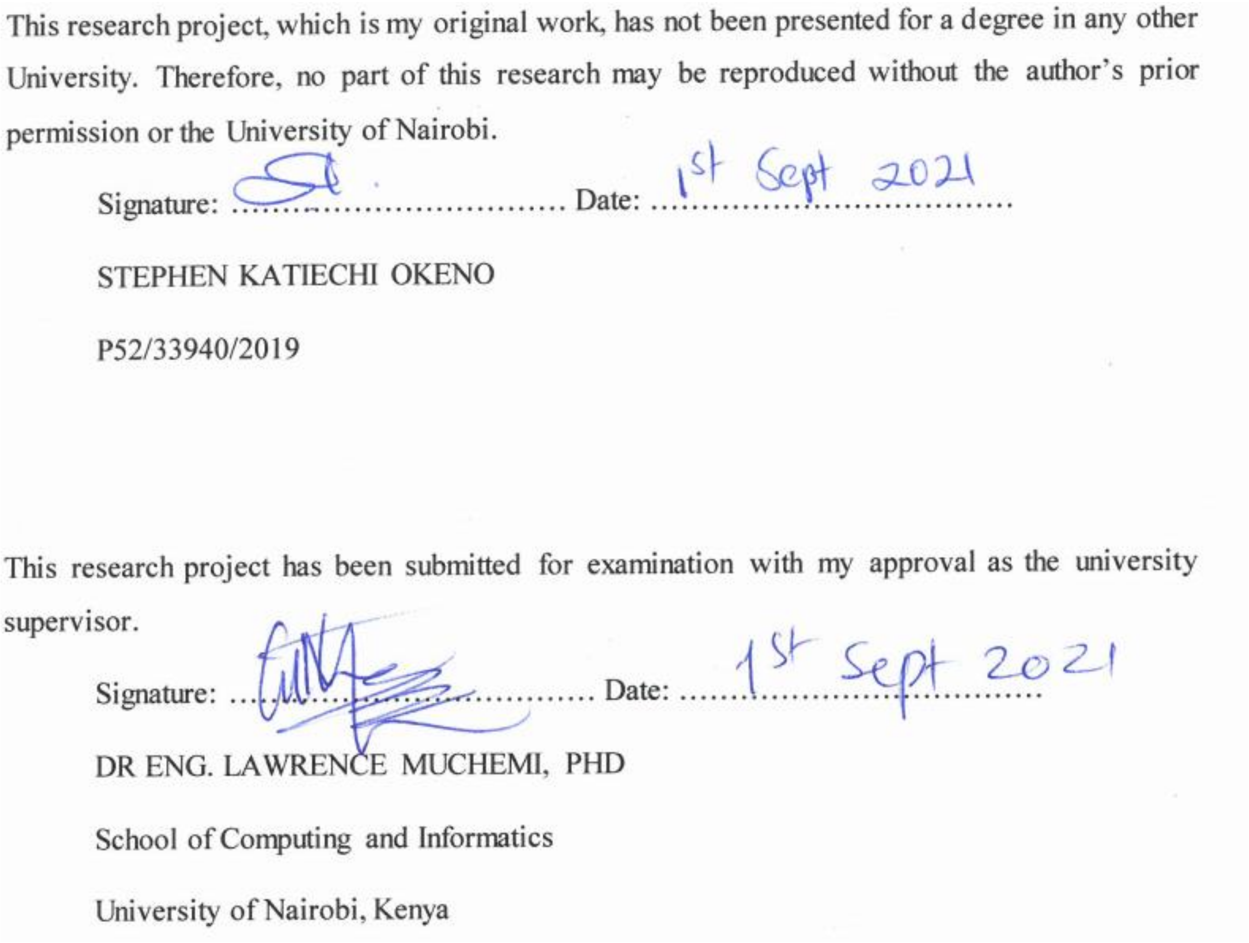
P52/33940/2019

SUPERVISOR:

DR ENG. LAWRENCE MUCHEMI

A Research Project Report Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computational Intelligence, School of Computing and Informatics, University of Nairobi.

DECLARATION



This research project, which is my original work, has not been presented for a degree in any other University. Therefore, no part of this research may be reproduced without the author's prior permission or the University of Nairobi.

Signature: .... Date: 1st Sept 2021

STEPHEN KATIECHI OKENO

P52/33940/2019

This research project has been submitted for examination with my approval as the university supervisor.

Signature: .... Date: 1st Sept 2021

DR ENG. LAWRENCE MUCHEMI, PHD

School of Computing and Informatics

University of Nairobi, Kenya

ACKNOWLEDGEMENT

To the Most High God, the source of all knowledge and understanding, ‘For the LORD gives wisdom; from his mouth come knowledge and understanding’ Proverbs 2:6 NIV.

When writing a supervised report, it is amongst the greatest clichés to say a “big thanks” to your supervisor. As much as I despise starting my report with a stereotype, I do have to express my collective gratitude towards Dr Eng. Lawrence Muchemi, PhD. I thank him for his guidance and patience during this project and his mentoring throughout all my brief encounters in the research and development of the project. Many thanks to Ken Mbuki and Johnstone Mainya whose interest in this study was very beneficial. They helped me in sourcing and verifying insurance data and all their helpful suggestions about the report and presentation towards the end. You served as an inspiration in making this a reality. I also acknowledge the University of Nairobi for allowing me to pursue a Master of Science degree program.

DEDICATION

I dedicate this work to my Mum, Rev. Rose Okeno, without whose caring support, counsel and encouragement it would not have been possible. To my lovely sons Nillan and Ethan, my dear wife Irine, my brother Fanuel, my sisters Loyce and Mary. In loving memory of my Late Dad-Rev. Elishamo Okeno, whose curtains closed before the play could begin, he passed on a love of reading and respect for education. And to scholars interested in federated machine learning.

ABSTRACT

Practical insurance fraud detection solutions require sufficient quality data from insurers to build effective models. However, insurance data is generally proprietary information for specific insurance companies and thus not publicly available. Also, the Insurance datasets are often imbalanced, making it challenging to develop fraud detection models that are not biased. Data privacy and class imbalance are two significant challenges when developing artificial intelligence applications in the insurance setup. In this research study, we tackle these challenges and propose a decentralized and privacy-preserving federated approach using an adjusted random forest model. The method is asynchronous federated learning of the traditional adjusted random forest classifier, i.e., achieving a higher performance and accuracy level than the traditional centralized learning approach. Based on it, we achieved secure collaborative machine learning that allows the training of quality federated fraud detection models from imbalanced data without sharing data. Experiments on Kaggle and Oracle insurance datasets demonstrate that the federated adjusted random forest classifier is more accurate and efficient than the non-federated counterpart. Our model is verified to be practical, efficient and scalable for real-life insurance fraud detection tasks.

Keywords: Fraud Detection, Federated Learning, Adjusted Random Forests, Feature Selection, Ensemble methods.

CONTENTS

DECLARATION .... **I**

ACKNOWLEDGEMENT .... **II**

DEDICATION .... **III**

ABSTRACT .... **IV**

CONTENTS .... **V**

LIST OF FIGURES .... **VII**

LIST OF TABLES .... **VIII**

LIST OF ABBREVIATIONS .... **VIII**

1 INTRODUCTION **1**

1.1 BACKGROUND 1

1.2 PROBLEM STATEMENT 2

1.3 RESEARCH OBJECTIVES 3

1.3.1 General Objective 3

1.3.1 Research Questions 3

1.4 JUSTIFICATIONS 4

1.5 CONTRIBUTIONS OF THE RESEARCH 5

1.6 SCOPE OF THE STUDY 5

2 LITERATURE REVIEW **6**

2.1 INTRODUCTION 6

2.2 INSURANCE FRAUD DETECTION METHODS 6

2.3 PRIVACY-PRESERVING MACHINE LEARNING METHODS 8

2.4 FEATURE ENGINEERING METHODS IN INSURANCE FRAUD DETECTION 10

2.5 RESEARCH GAP 12

2.6 CONCEPTUAL FRAMEWORK 13

3 METHODOLOGY **15**

3.1 INTRODUCTION 15
3.2 STUDY POPULATION 16
3.3 DATA COLLECTION 16
3.4 DATA ANALYSIS 17
3.4.1 Data Cleaning 21
3.4.2 Data Transformation 22
3.5 FEATURE ENGINEERING 22
3.5.1 Feature Engineering 22
3.5.2 Feature Selection 23
3.6 FEDERATED MODEL DESIGNS 24
3.6.1 Feature Alignment 24
3.6.2 Federated Adjusted Random Forests 24
3.6.3 Horizontal Federated Learning 26
3.6.4 Decentralized Architecture Design 27
3.6.5 Decentralized Algorithm Design 28
3.7 IMPLEMENTATION AND PROTOTYPING 29
3.8 MODEL EVALUATION 30
3.9 ETHICAL CONSIDERATIONS 30

4 RESULTS AND DISCUSSIONS 31

4.1 INTRODUCTION 31
4.2 EVALUATION RESULTS 31
4.2.1 Classification Accuracy 31
4.2.2 Confusion Matrix 32
4.2.3 Classification Report 35
4.3 DISCUSSION 37
4.4 MODEL VERDICT 38

5 CONCLUSION AND RECOMMENDATIONS 39

5.1 CONCLUSION 39
5.2 RECOMMENDATION 40

5.3 FUTURE RESEARCH 40

6 REFERENCES 41

7 APPENDICES 43

7.1 IRA STATISTICS 43

7.2 EMAIL CORRESPONDENCE TO REQUEST FOR INSURANCE DATA 44

LIST OF FIGURES

Figure 1 Conceptual Framework 14

Figure 2 Research Process 15

Figure 3 Features of various datasets 17

Figure 4 Features Correlation Heat-Map 18

Figure 5 Features correlating with fraud state 19

Figure 6 Data Analysis 20

Figure 7 Filling Missing Values 21

Figure 8 Data Transformation 22

Figure 9 Kaggle Dataset Selected Features 23

Figure 10 Oracle Dataset Selected Features 23

Figure 11 Feature Alignment 24

Figure 12 Bagging Using Adjusted Random Forest 25

Figure 13 Random Federated Forests 26

Figure 14 Decentralized Design 27

Figure 15 Federated Algorithm 28

Figure 16 Federated Random Forest Before Balancing-Kaggle Dataset 33

Figure 17 Federated Random Forest Before Balancing-Oracle Dataset 33

Figure 18 Balanced Federated Random Forest-Kaggle Dataset 34

Figure 19 Balanced Federated Random Forest- Oracle Dataset 34

Figure 20 Email to Head of Innovations IRA 44

Figure 21 Email to head of Research and Innovation IRA 45

Figure 22 Email to the C.E.O and Commissioner for Insurance IRA 46

LIST OF TABLES

Table 1 Accuracy Score for Kaggle Dataset 32
Table 2 Accuracy for Oracle Dataset 32
Table 3 Classification Report Before Feature Selection-Kaggle Dataset 35
Table 4 Classification Report Before Feature Selection-Oracle Dataset 36
Table 5 Classification Report After Feature Selection-Kaggle Dataset 36
Table 6 Classification Report After Feature Selection-Oracle Dataset 36
Table 7 Models Average Performance 38
Table 10 IRA Statistics 43

LIST OF ABBREVIATIONS

ML	Machine Learning
ANN	Artificial Neural Network
FL	Federated Learning
B2B	Business to business
IRA	Insurance Regulatory Authority
RFM	Recency, frequency, and monetary
HOBA	Homogeneity-oriented behavior analysis
CART	Classification and Regression Trees

CHAPTER ONE

1 INTRODUCTION

1.1 Background

Insurance claims fraud (illegitimate claims), other than tax fraud, is recorded to be the most practised fraud globally. The significant accumulation of liquid financial assets makes insurance companies susceptible to loot schemes and takeovers (Association of Certified Fraud Examiners, 2019). Insurance claims fraud occurs when the insured attempts to gain profits through premiums paid without complying with the insurance agreement terms (Association of Certified Fraud Examiners, 2019). Detecting fraud manually has always been costly for insurance companies. Low incidents that go undetected contribute immensely to the claim ratio. For example, the Industry average Incurred claims ratio (loss ratio) is 64.34%, with motor insurance accounting for 24.6% of the total industry paid claims under the general insurance business (Insurance Regulatory Authority, 2020).

The research community has focused on insurance fraud detection methods that require centralized datasets from specific insurers. There is vast body of literature published on fraud detection methods in the Insurance Industry. These methods, however, use insurance data from specific insurers that might not be representative of the industry fraud problem. Feeble attempts have been made to look at fraud detection methods from an industry perspective. The quality of the data needed to train predictive models is as important as the quantity required. Datasets must be representative and balanced to provide a better picture and avoid bias (Rama Devi Burri, et al., 2019). Recent studies on claim analysis using machine learning recorded data security challenges in implementing machine learning. Vast amounts of data required for machine learning have created additional risk for insurance companies. The increase in data collection and connectivity among applications can lead to data leaks and security breaches. This makes Insurers struggle to provide relevant data for training machine learning models (Rama Devi Burri, et al., 2019).

Significant studies have been conducted to explore the detection and prevention of Insurance fraud. For example, (Phua et al., 2010) have explored holistic and scientific approaches to fraud management. Their respective works observe that studies involving quantitative methods report limitations due to the lack of insurance data (Phua et al., 2010). However, because of the evolving nature of insurance fraud, there still exist challenges due to the lack of sufficient insurance data and a class imbalance problem in claim datasets that have attracted the attention of researchers. Insurance fraud detection problems are often biased because they reduce the overall error rate instead of taking care of minority classes (Johannes & Rajasvaran, 2020). Studies have shown that the lack of primary insurance data and imbalanced datasets is a challenge when developing machine learning models in insurance. Imbalanced datasets often produce biased models that cannot make correct predictions (Johannes & Rajasvaran, 2020). Insurance companies that adopt a centralized approach for insurance claim fraud detection face a class imbalance problem, a case where fraud incidents are less than the total number of claims (R Guha et al., 2017).

1.2 Problem Statement

The research community has focused on insurance fraud detection methods that require centralized datasets to train models. Centralized machine learning methods often produce biased models which are not effective in detecting insurance fraud. The bias in insurance models is primarily attributed to two issues; the class imbalance problem in datasets, where fraud incidents are less than the total number of genuine claims, insufficient insurance data to train the models and. For example, (Johannes & Rajasvaran, 2020) presents a behavioural feature engineering approach for motor insurance fraud detection. However, in the study, they observe that insurance claim data is often imbalanced where at least one class forms a tiny minority of the data. The works by (Burri et al., 2019) provides an in-depth claim analysis using machine learning. The study, however, reports challenges in finding suitable data sources and data security in implementing machine learning (Burri et al., 2019). There have been feeble attempts to look at fraud detection methods that benefit all insurance players instead of individual insurers.

Previous studies present a need for a quality fraud detection system that can tackle the class imbalance problem in the datasets. In this case, all participants who do not have sufficient datasets can collaborate in building a quality model. Studies also present a need to implement privacy-preserving methods that can be used to train machine learning models without sharing data. The methods used in the previous studies suffer drawbacks due to the quality of data used to train fraud detection models; the studies have also presented challenges in accessing quality datasets from insurers (Burri et al., 2019).

1.3 Research Objectives

1.3.1 General Objective

To implement a privacy-preserving federated machine learning framework for the Insurance setup that will be used to train fraud detection machine learning models while preserving the privacy of data. The model will be used to detect insurance claims fraud. We will evaluate the effectiveness of this framework in improving the prediction accuracy of Insurance fraud detection models. The accuracy in the prediction of the model will be assessed against past insurance claims data.

1.3.1 Research Questions

1. What practical technique can be used to build quality insurance fraud detection models while preserving policyholders’ privacy?
2. How can quality fraud detection models be built from imbalanced insurance datasets?
3. What is the prediction performance of federated insurance fraud detection models?
4. Which optimal feature engineering and selection method is used for high dimensional datasets?

1.4 Justifications

This research aims to implement a privacy-preserving machine learning architecture that will be used to train insurance fraud detection models. While machine learning methods such as classification and regression algorithms have been identified and studied in previous research (Burri et al., 2019), the studies do not show that such algorithms can train machine learning models while preserving the privacy of insurance data. In addition, little research has been done to show that decentralized methods can be used with imbalanced datasets to produce quality insurance fraud detection models. The broad topic of insurance fraud detection has received attention, including from insurers and government regulators. Still, decentralized, collaborative and privacy-preserving machine learning methods have not been the focus of that attention. Instead, while acknowledging the challenges in finding quality insurance data and the class imbalance problem in datasets, the research community currently focuses on centralized machine learning methods biased towards individual insurance companies (Dhieb et al., 2019).

The insurance industry, with hundreds of years of history, is characterized by fierce competition. Data has become a valuable resource that Insurers have to protect, hindering the development of solutions that benefit all industry players. Insurers struggle to release relevant data for training AI models (Burri et al., 2019). Brilliant ideas give value to the industry, such as automated underwriting, automated claims fraud detection require privacy-preserving machine learning methods. There is a need to introduce collaborative privacy-preserving approaches to machine learning and data science in insurance. This study will provide insights into insurance claims management practice by exploring privacy-safe methods that can be used to detect insurance claims fraud accurately.

1.5 Contributions of the Research

Research has reported problems in implementing machine learning in the insurance industry, including lack of suitable data sources, data security, and imbalanced datasets (Burri et al., 2019). Balanced datasets give a better picture and avoid bias in prediction. Imbalanced insurance data makes it challenging to produce quality models shared across the industry. This research will draw recommendations on model performance built using imbalanced datasets, improving the prediction accuracy of fraud detection models. The research seeks to demonstrate that using privacy-preserving methods to train a model on decentralized data preserves data integrity, improves prediction accuracy, and, therefore, a practical approach in claims fraud prediction. The study will contribute to the Insurance Claims Management practice and claims fraud detection. In addition, this study will contribute to the knowledge of privacy-preserving machine learning in insurance.

1.6 Scope of the Study

The study will be limited to the General Category of Insurance. This area of the Insurance business is selected because it accounts for the insurance industry’s highest-paid claims. The insurance regulatory authority regulates thirty-six Insurance companies offering a General Category of Insurance. The companies could be used in this research. However, to facilitate this project, we select three major Insurance Companies. The companies understudy would need to be actively engaged in the motor class of insurance.

CHAPTER TWO

2 LITERATURE REVIEW

2.1 Introduction

Fraud detection has blossomed into a high priority and technology-laden problem in the insurance setup. Studies have shown that apart from tax fraud, insurance fraud is reported to be the most practiced fraud globally (Association of Certified Fraud Examiners, 2019). Although fraud detection is seen to be used more in functional fields, it is considered in academia because of the adverse economic effects in insurance pricing and promoting efficiency in the insurance industry. The domain focus of the research community has been on fraud detection methods on centralized datasets. Centralized machine learning often requires extensive data to produce efficient models. Moreover, they suffer a class imbalance problem, a case where fraud incidents are less than genuine claims(R Guha et al., 2017). There have been feeble attempts to look at fraud detection methods from the industry level and provide solutions that cut across the industry instead of individual insurers. Some attempts can be used to address the general insurance fraud problem, and their contributions cannot be ignored even though they address fraud problems in individual insurance companies.

2.2 Insurance Fraud Detection Methods

The works of (Burri et al., 2019) provide a systematic description of machine learning algorithms and their insurance industry application. The study gives an in-depth analysis of predictive models and their application in understanding claim costs. The study also introduces different ways to implement machine learning in the insurance setup; however, the research faces various challenges, including lack of the right data source, data security and higher training requirements. Whereas the study provides a way to analyze claims using machine learning algorithms, it suffers problems because of lack of suitable data sources, data security, and imbalanced insurance datasets (Burri et al., 2019). The study also gives little attention to the challenges of getting suitable data sources and the data privacy problem in applying machine learning in insurance.

Related studies by (R Guha et al., 2017) provides a comparative analysis of machine learning techniques for detecting insurance claims fraud. In the study, the modified random undersampling and the adjusted random forest classifier performed the best on claim datasets, and this is because of their inherent characteristics of aggregating weak classifiers. However, the study reported constraints because of the inability to understand context relationships among features (geography, customer segment, insurance sales process) that might not reflect the typical picture of fraud indicators. The study faces constraints to operate with little known parameters based on heuristic knowledge while having information on other features that can influence decisions (R Guha et al., 2017).

Statistical and regression methods have been studied by (Palacio, 2019). The study focuses on fraud detection using semi-supervised machine learning. The study makes use of mini-batch K-Means and stratified 5-Fold cross-validation. Whereas the method reports good performance, the method faces limitations in that it has to be calibrated each time predictors change. This means the model has to be retrained through dynamic learning (Palacio, 2019). The study fails to show that using random indicators produces a standard fraud detection model shared across the industry. The parameters that are used are extracted from calculations made with data from the industry as a whole. This makes the model irrelevant to insurers who may want to benefit from it.

Previous works by (Phua et al., 2010) reviews data mining techniques from across the industry. The study heavily relies on an extensive literature review of fraud detection literature. The study reports limitations due to the lack of quality and publicly available labelled data. The article discusses fraud detection from both supervised and unsupervised methods. He provides an in-depth analysis of different ways to evaluate the technique’s performance (Phua et al., 2010). The article, however, does not pay attention to the methods of acquiring quality data needed for training machine learning models. The study looks at fraud from an individual insurance company perspective instead of the industry as a whole.

As outlined by (Johannes & Rajasvaran, 2020), research studies present a need for a quality fraud detection system that considers class imbalance problems in the insurance industry. The models trained using standard features can be used by insurers and micro-insurers who do not have sufficient quality datasets to create quality models. Studies have also shown that among the challenges faced in fraud detection include class imbalance, which means that the incidences of fraud are far less than the total number of claims (R Guha et al., 2017). The class imbalance problem is because of the ever-evolving nature of fraudulent claims present in one insurer and not in the other. Methods used in the previous studies suffer drawbacks due to the quality of data used to train fraud detection models; the studies have also presented challenges in accessing quality datasets from insurers (Phua et al., 2010). Studies present a need to implement privacy-preserving systems that can be used to train quality machine learning models from quality imbalanced datasets for different insurers.

2.3 Privacy-Preserving Machine Learning Methods

Privacy and large-scale machine learning are classical problems that have since existed in data-sensitive industries such as insurance. Studies have been conducted to show that it is possible to train fraud detection models while preserving clients' privacy. (Dhieb et al., 2019), in their research, they propose an extreme gradient boosting machine learning algorithm for the detection of fraud. The study uses categorization to transform most attributes of the client's claims into a binary format that is not human readable. The study also uses generalization to replace low-level data with high-level concepts that hide sensitive data such as customer information. Whereas the model presents excellent results with an accuracy of 99.25% compared to the state of the art algorithm, the study uses centralized data with a bias towards individual insurers under investigation. The model does not present an actual picture of fraud incidents in the industry. It cannot be shared for use by other industry players, limiting the possibilities of quality models.

Previous research shows that the adaptation of decentralized training of models in the medical industry is an effective method for disease diagnosis while at the same time preserving the privacy of customer's data. An algorithm to train a shared model using distributed deep neural networks is discussed (Ongati & Lawrence, 2019). The study adopts an approach of sharing the learnt representations (model weights) instead of sharing raw data. The algorithm, however, faces challenges in having a central coordinator who can break the whole network when down. Google researchers have also proven that using Federated Model optimization produces a high-quality centralized model (Konečný et al., 2016). The Federated Averaging protocol, as discussed by (McMahan et al., 2017), assumes that all devices are equally likely to participate and complete each round. This method is primarily limited to small handheld devices as opposed to large and powerful machines and, therefore, a limitation in this case.

Studies were done by (Liu et al., 2019) on the use of Random Forests in a decentralized setup while preserving data privacy. A framework of federated forests was presented based on the CART tree. The study assumes that all domains have the same number of samples and therefore aligns sample IDs across domains. In their research, each tree on a node build by every node working together synchronously. The master node notifies all participating nodes of the features selected. The method, however, hides from the participating nodes of the features that were selected globally. In training, both the master and client nodes get involved in the tree construction. The master node collects all the optimal local features in their algorithm and their corresponding impurity improvements from the client nodes (Liu et al., 2019). The master determines the global best feature from the collected features and notifies the node which provided it. The client node that provided the feature will split the samples and send split information, i.e. IDs that fall in the left and right subtrees. The client then saves the details of the split. If the child tree nodes are created successfully, then the master node doesn't save sample IDs for the subtrees, and when the connection is down, training can be recovered from the breakpoint (Liu et al., 2019).

The studies by (Liu et al., 2019) assumes that domains will have the same number of samples. The assumption is not practical across most industries like insurance which is a limitation in their research. Diverse samples are a strength for machine learning to produce quality models. The study also requires that all participating parties be online during the training; synchronous training is a limitation in decentralized learning. Different entities may encounter different complexities to be online during the training. The study also faces problems of tightly coupled dependencies. All clients depend on the master node that controls all training activities, and in case the master fails, then the entire learning cycle breaks down. The study also faces limitations in prediction due to the partial models stored on client nodes. Multiple rounds of communications are a bottleneck in prediction, especially when one client fails.

Although research has been done to show that using Federated Learning improves prediction accuracy while preserving the accuracy of the data, the proposed methods suffer from abilities to train federated models asynchronously (McMahan et al., 2017) while eliminating a single source of failure (Ongati & Lawrence, 2019). The methods also present limitations due to tightly coupled nodes and dependencies (Liu et al., 2019). Centralized methods that preserve data privacy, like seen in (Dhieb et al., 2019), require centrally aggregated data and, therefore, not effective methods for decentralized machine learning. There is limited research to show the effectiveness of these collaborative model training techniques in the insurance industry. Studies present a need to conduct a study that determines whether using decentralized learning will improve prediction accuracy and prove an effective method of insurance claims prediction.

2.4 Feature Engineering Methods in Insurance Fraud Detection

Feature identification and extraction are essential activities to identify the best predictors in any machine learning application. Models without the right features yield poor results in prediction (R Guha et al., 2017). Data with high dimensionality characterize the insurance industry. Researchers such as (Zhang et al., 2019) have demonstrated that feature engineering significantly influences fraud detection improvements in their respective works. Most researchers approach the feature engineering process by computing features from incoming and past claim entries within a given aggregation period (Johannes & Rajasvaran, 2020). However, the method cannot be used for behaviour analysis in insurance claims. This is because it does not consider the heterogeneity of insurance claims.

In recent times, the study of customer behaviour analysis has been adopted as one of the methods of identifying insurance fraud detection features. For example, (Zhang et al., 2019) proposes the HOBA feature engineering methodology that allows claims to be grouped into homogenous claims while considering different heterogeneity characteristics within the individual groups using an aggregation strategy. The proposed aggregation strategy comprises four steps: determining a set of characteristics worth investigating for fraud detection, then for each customer, a thorough analysis based on recency, frequency, and monetary(RFM) using aggregation statistics in a given aggregation period. The study follows a holistic approach to extracting fraud detection features. However, it suffers limitations as it did not assess the computation costs of the proposed framework. In addition, the feature engineering approach is limited to only one bank instead of an approach that will cut across the industry.

Machine learning applications in insurance often use exploratory analysis techniques to analyze and summarize the datasets' primary information using statistics and visualization methods. To extract the essential features, (Dhieb et al., 2019) have proposed a method that analyzes the frequency of features and calculates the correlation between them. A filtering method is used to filter futile features that are highly uncorrelated to simplify the running time and increase the algorithm's efficiency (Dhieb et al., 2019). While acknowledging the challenges of class imbalance that previous research has faced, the method fails to prove that using the XGBoost algorithm tackles the class imbalance problem in insurance.

Feature selection processes are also documented by (R Guha et al., 2017), the importance of feature engineering in insurance claims fraud detection is discussed in details. Most of the features are engineered based on domain knowledge and datasets attributes. In the study, an in-depth review of different feature selection methods. Forward selection and backward selection are presented and their effects in determining the best features for the model. Dimensionality reduction using Principal Component Analysis is discussed to reduce the number of dimensions and select dimensions that explain most dataset variance. Forward selection performs well in situations where one wants to select the best features that lift the model's performance to a great extend (R Guha et al., 2017).

Backward elimination takes more time compared to forwarding selection. This is because it starts with all features and eliminates features that compromise model performance (R Guha et al., 2017). The method, however, performs better when the model constructed relies heavily on features, for example, decision trees. Dimensionality reduction, for example, the principal component analysis, is helpful, especially when you want to avoid overfitting. Fewer dimensions will produce fewer features that can be iterated over by other feature selection algorithms (R Guha et al., 2017); this results in faster computations. The methods presented in the research provide suitable methods that can be used to select features in insurance. Still, they do not consider the ever-changing features in the insurance industry.

2.5 Research Gap

Insurance fraud detection systems that preserve client’s privacy have been of great interest to researchers as per the literature review. However, the research community has focused on insurance fraud detection methods that require centralized datasets while paying less attention to the sources needed to get quality data required to train the machine learning models. In addition, there have been feeble attempts to look at fraud detection methods that benefit the industry as a whole instead of individual insurers. The community has also focused on fraud predictors limited to individual insurance companies instead of the ever-changing nature of predictors in the industry.

Research studies that have attempted to provide insurance fraud detection solutions have reported limitations due to the lack of quality insurance datasets. As outlined by (Phua et al., 2010) and (R Guha et al., 2017), there is a need for a quality fraud detection system that considers imbalanced claim datasets in the insurance industry. Decentralized training models on imbalanced data will enable insurers and micro-insurers who do not have sufficient quality datasets to create quality models. Studies have also shown that among the challenges faced in fraud detection include class imbalance, which means that the incidences of fraud are far less than the total number of genuine claims (R Guha et al., 2017). The work of (Burri et al., 2019) provides an in-depth claim analysis using machine learning. The study comprehensively compares several classification techniques ranging from simple regression models to complex neural networks. The study, however, reports challenges in finding suitable data sources and data security in implementing machine learning.

Reviews from previous studies (Johannes & Rajasvaran, 2020) present a need for a quality fraud detection system that considers imbalanced data in the insurance industry. There is also a need to implement privacy-preserving methods that can be used to train machine learning models from imbalanced datasets. The methods used in the previous studies suffer drawbacks due to the lack of quality of data and standard features used to train fraud detection models. Limited research has also been done to show the effectiveness of collaborative model training techniques in creating fraud detection models in the insurance industry. There is a need to conduct a study that determines whether using privacy-preserving decentralized methods will improve prediction accuracy and prove an effective method of insurance claims prediction.

2.6 Conceptual Framework

The conceptual framework that aims to resolve issues in the research gap is a practical, federated machine learning model based on adjusted random forests that seeks to improve insurance claims fraud prediction accuracy while considering the security of the data used. First, each node builds a fraud detection model based on standard features shared across participating members while evaluating its accuracy and performance on the training, validation, and test sets. Then, each party uploads the model to a centralized node. The updated model is then used to detect fraud at each node, as shown in figure 1. The architecture requires a central node to manage the decentralized models, get model updates from edge devices, do the ensemble and aggregation work and distribute it back to all participating nodes. In this case, this centralized node is chosen to be a regulatory authority, in this case, the Insurance Regulatory Authority.

We group our input variables into four categories that directly correlate with our target variable fraud status. First, the customer profile has personal and behavior related characteristics of the Insured that are considered during rating and premium calculations. They include sex, marital status, age of the policyholder, years of driving experience. Second, the risk profile has the details of the property, most of which directly affect the end premium. They include vehicle category, the sum insured, car make, age of the car. Third, policy details contain details of the cover. They include cover type, age of the policy, the deductible amount. Finally, claim Details has details of the launched claim; the details include the date on which the incident happened, the incident area, the date a claim was launched, who was at fault if the police report was filed, if witnesses were present during the incident.

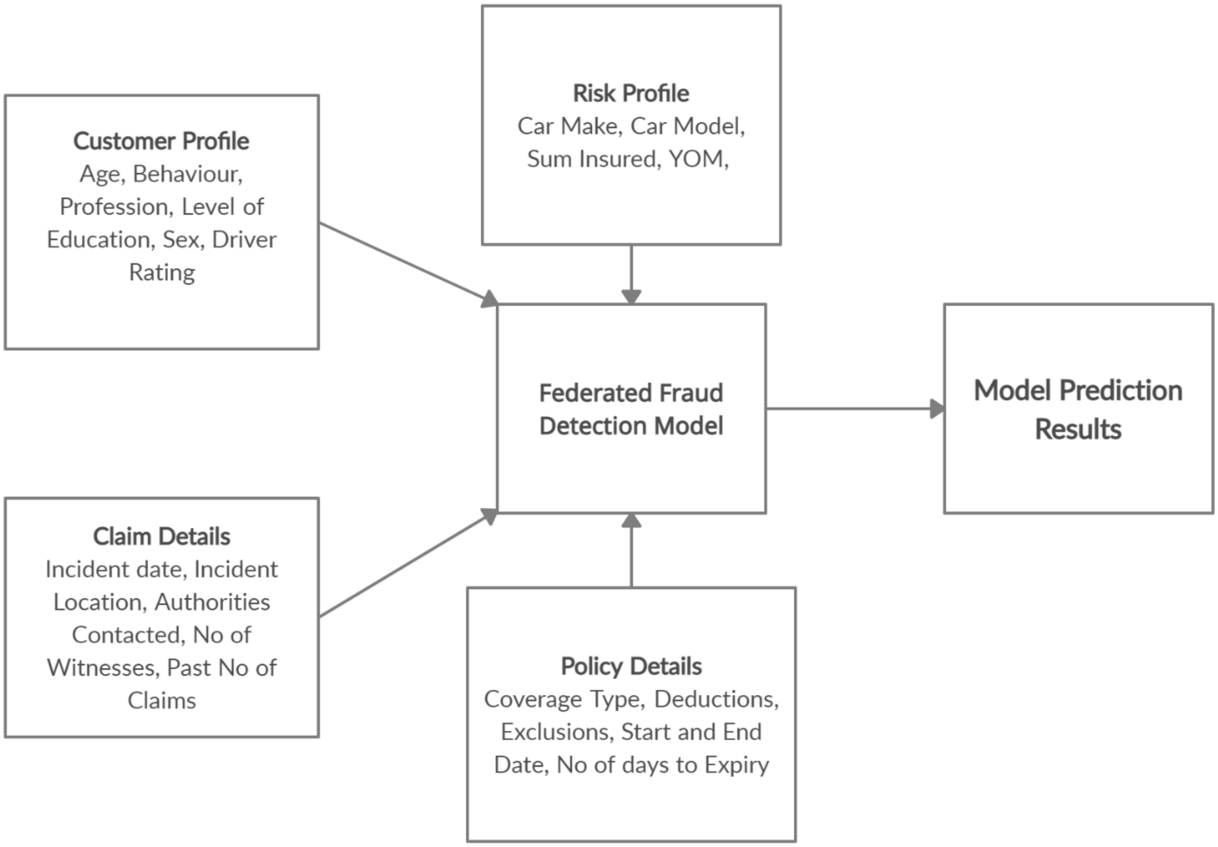

Risk Profile
Car Make, Car Model, Sum Insured, YOM,
Customer Profile
Age, Behaviour, Profession, Level of Education, Sex, Driver Rating
Federated Fraud Detection Model
Model Prediction Results
Claim Details
Incident date, Incident Location, Authorities Contacted, No of Witnesses, Past No of Claims
Policy Details
Coverage Type, Deductions, Exclusions, Start and End Date, No of days to Expiry

Figure 1 Conceptual Framework

CHAPTER THREE

3 METHODOLOGY

3.1 Introduction

The primary focus of the methods used in this research is to answer the questions and objectives and provide a detailed description of how we achieved each objective. First, we conducted theoretical studies of existing literature and previously implemented fraud detection systems to provide insights into how such systems were implemented and the limitations of their implementation; We collected past insurance claims data representing different Insurers for simulations. The collected data was then cleaned and prepared for training. Next, a federated architecture and algorithm were designed; after that, we carried out Simulations and prototyping, the federated model was trained and evaluated. The workflow presented in figure 2 below summarizes the methodology used in the study.

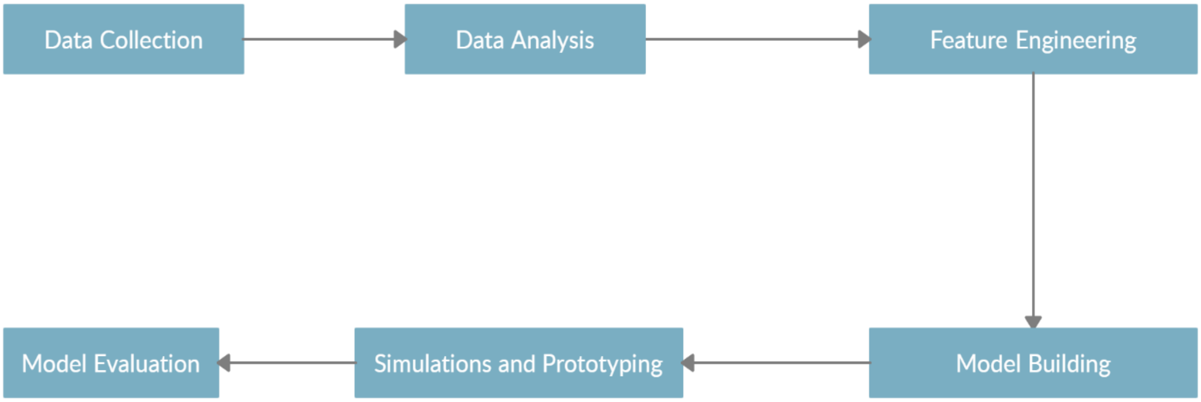


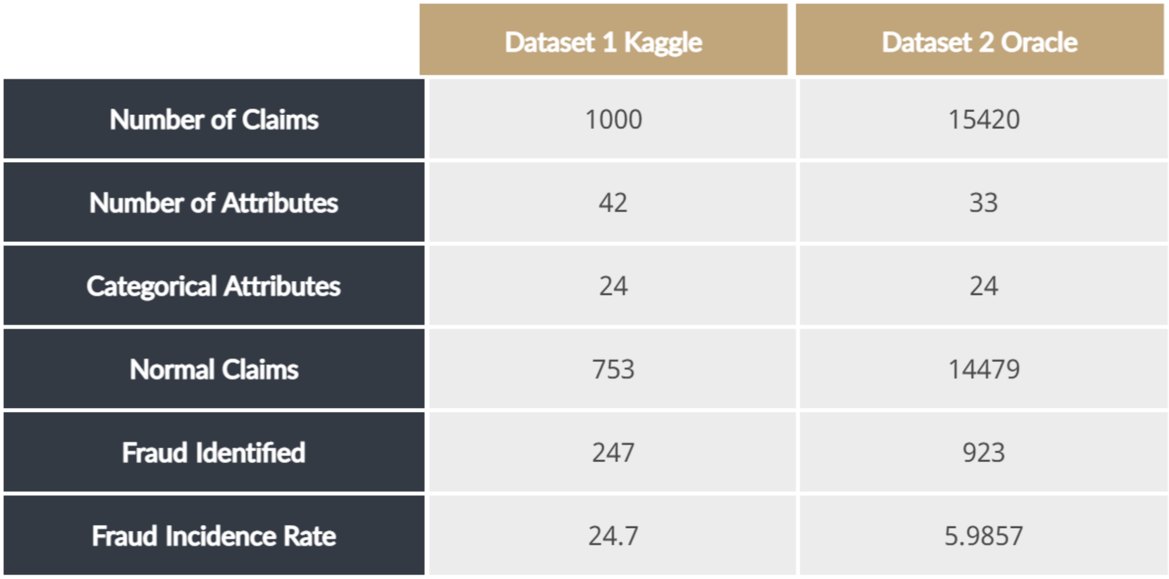
Figure 2 Research Process

3.2 Study Population

The study focuses on the Motor Class of Insurance. This area of the Insurance business is selected because, as reported by IRA (Insurance Regulatory Authority, 2020), it accounts for the insurance industry's highest fraud incidents between 2016 and 2019, as shown in appendix 1. The insurance regulatory authority regulates thirty-six Insurance companies offering a General Category of Insurance in Kenya. Since our project focuses on building a federated model, we work with IRA. This organization regulates the Insurance Industry. Insurance data is proprietary data for specific insurance companies, and regulatory bodies such as IRA do not have such information therefore not publicly available. The datasets used in this study are standard motor claims datasets sourced from a public online platform Kaggle (Roshan,2019).

3.3 Data Collection

The variables used in developing the model for this study were extracted from past motor claims history. Because of the sensitive nature of the insurance data, we used the Kaggle dataset (Roshan, 2019) and the Oracle dataset (Charlie, 2010) to train and test our models. The datasets contain many insurance features and columns used in this study; this includes columns containing the insured personal details. The input variables under study include the customer profile with personal details like name, age, location and sex. We also study the risk profile of the property insured, the claim details and policy details of the insured. They include the type of risk, car make and model, year of manufacture, the number of days to the expiry and the sum insured. For all the datasets, a small portion of claims is identified as frauds while others as normal. Some claims marked as normal might be fraudulent, but the suspicions were not followed through because of time delays, late detection, among other reasons. Table 1 below summarizes our datasets:



	Dataset 1 Kaggle	Dataset 2 Oracle
Number of Claims	1000	15420
Number of Attributes	42	33
Categorical Attributes	24	24
Normal Claims	753	14479
Fraud Identified	247	923
Fraud Incidence Rate	24.7	5.9857

Figure 3 Features of various datasets

3.4 Data Analysis

This research aims to foster collaboration among participating members who work together to improve the model's performance by improving the quality of features presented for machine learning. The study, therefore, presents a classical exploratory data analysis approach that provides the needed intuition about data. Statistical techniques are used in interpreting and selecting quality features relevant to the research to have a global view of the dataset and extract essential features. For example, we analyze the frequency of features and the correlation between them, as shown in figure 1 below.

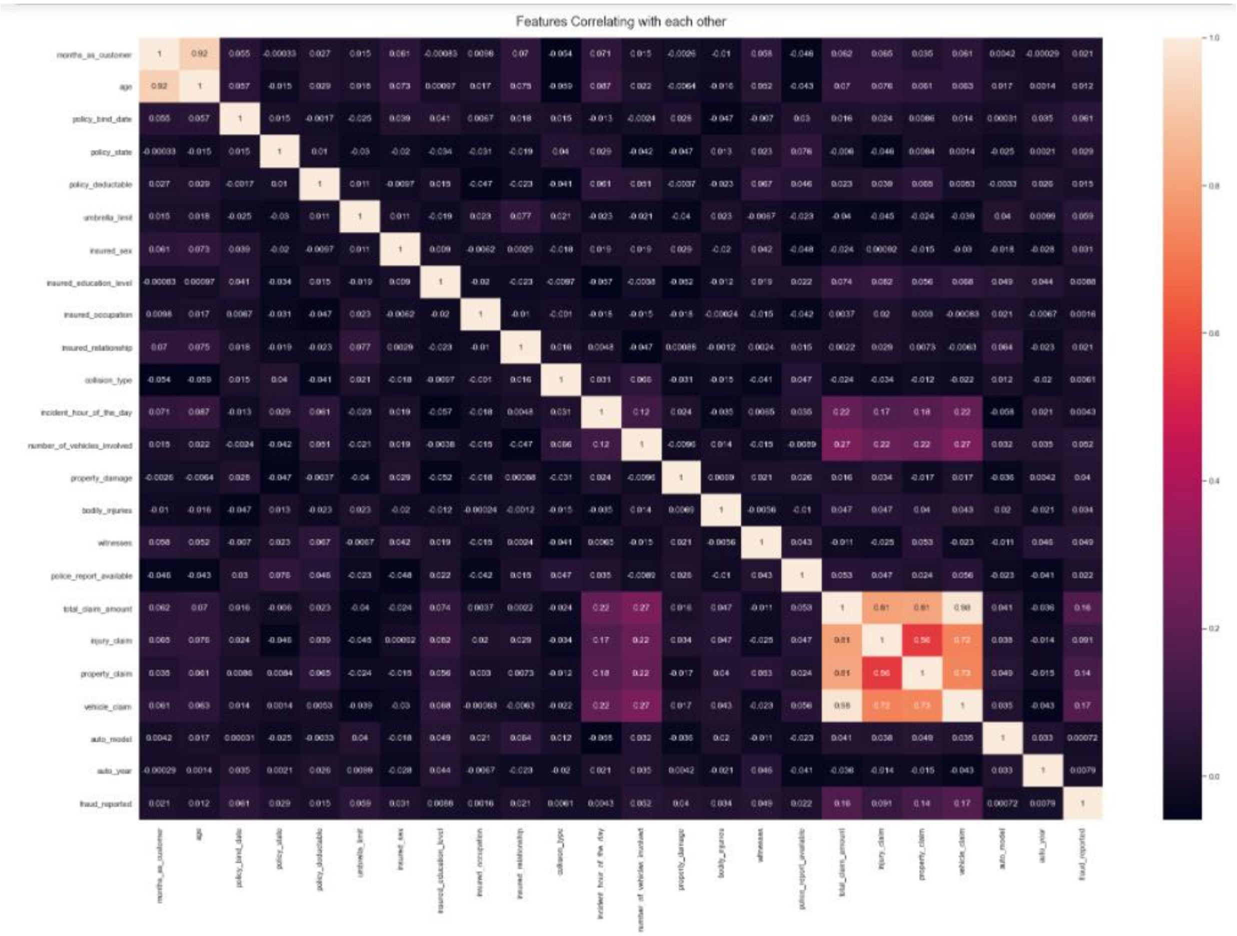


Figure 4 Features Correlation Heat-Map

From the correlation map, we can deduce that the amount claimed for different covers is closely related to fraud. This is because the returns of a fraudulent claim are the motivating factors to the insured committing fraud. The claim amounts include the total claim amount, injury claim, property claim and vehicle claim. To better understand our features, we show the strength in the correlation between each independent variable that we use in our study and the dependent variable, fraud in a claim. We plot a correlation heatmap that spreads the features between -1 and 1, skewed towards one and strongly related to the predicted variable. In contrast, those that skew towards -1 are weak predictors, as shown in the figure below.

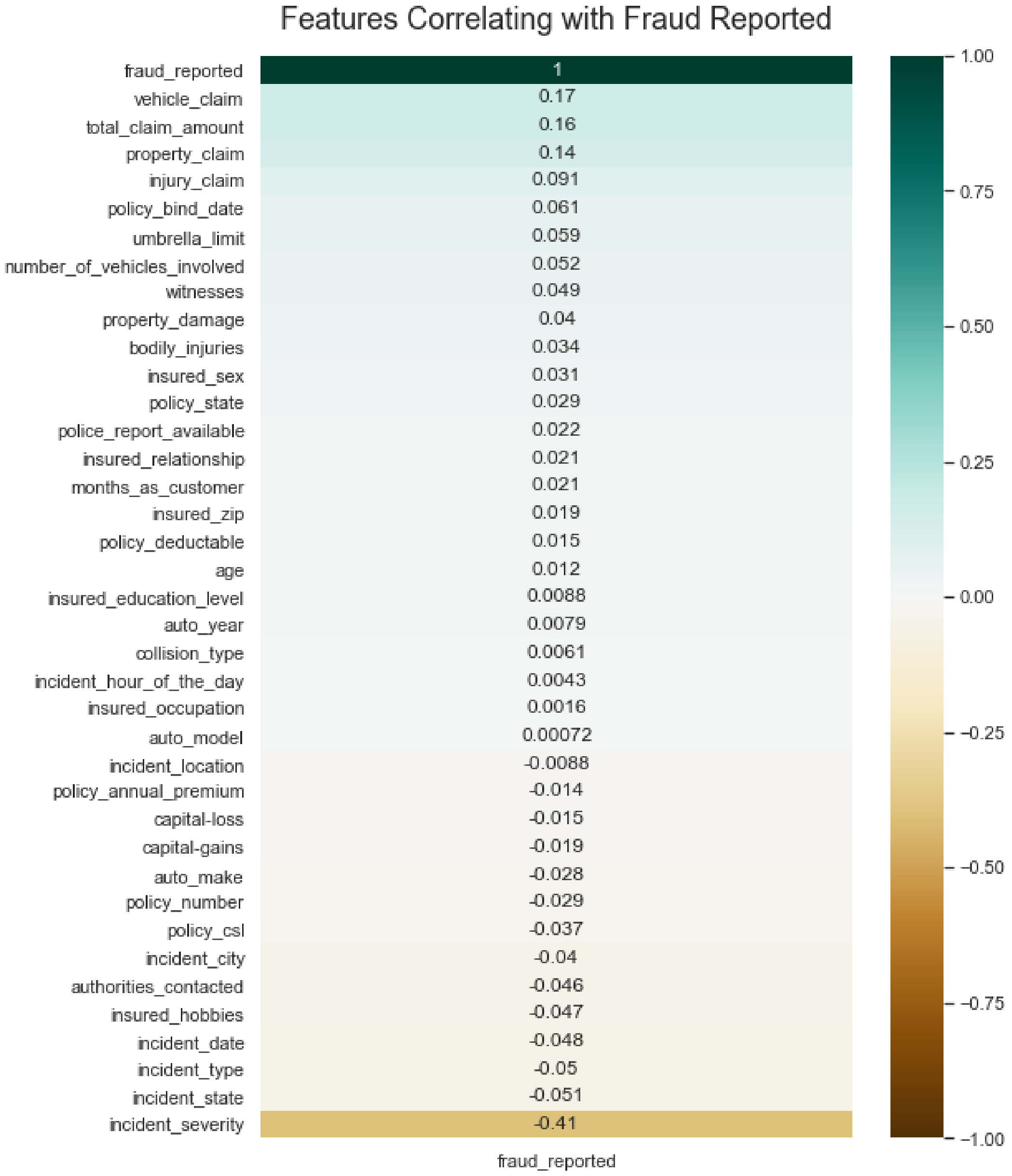


Figure 5 Features correlating with fraud state

We plot graphs to show the count of directly related features to the predicted variable to help us identify features of interest in our study. We note that the incident state, insured sex, and police report availability have minor contributions to fraud. There is minimal difference in the variables makes them weak predictors in our model. We observe the same results in the correlation heatmap presented in figure 4 above. There is an evident variation in the number of witnesses, insured level of education and insured occupation. The variations make the features better predictors for our classification problem. Figure 2 and 3 below provides a graphical representation of different features and how they are related to the predicted variable.

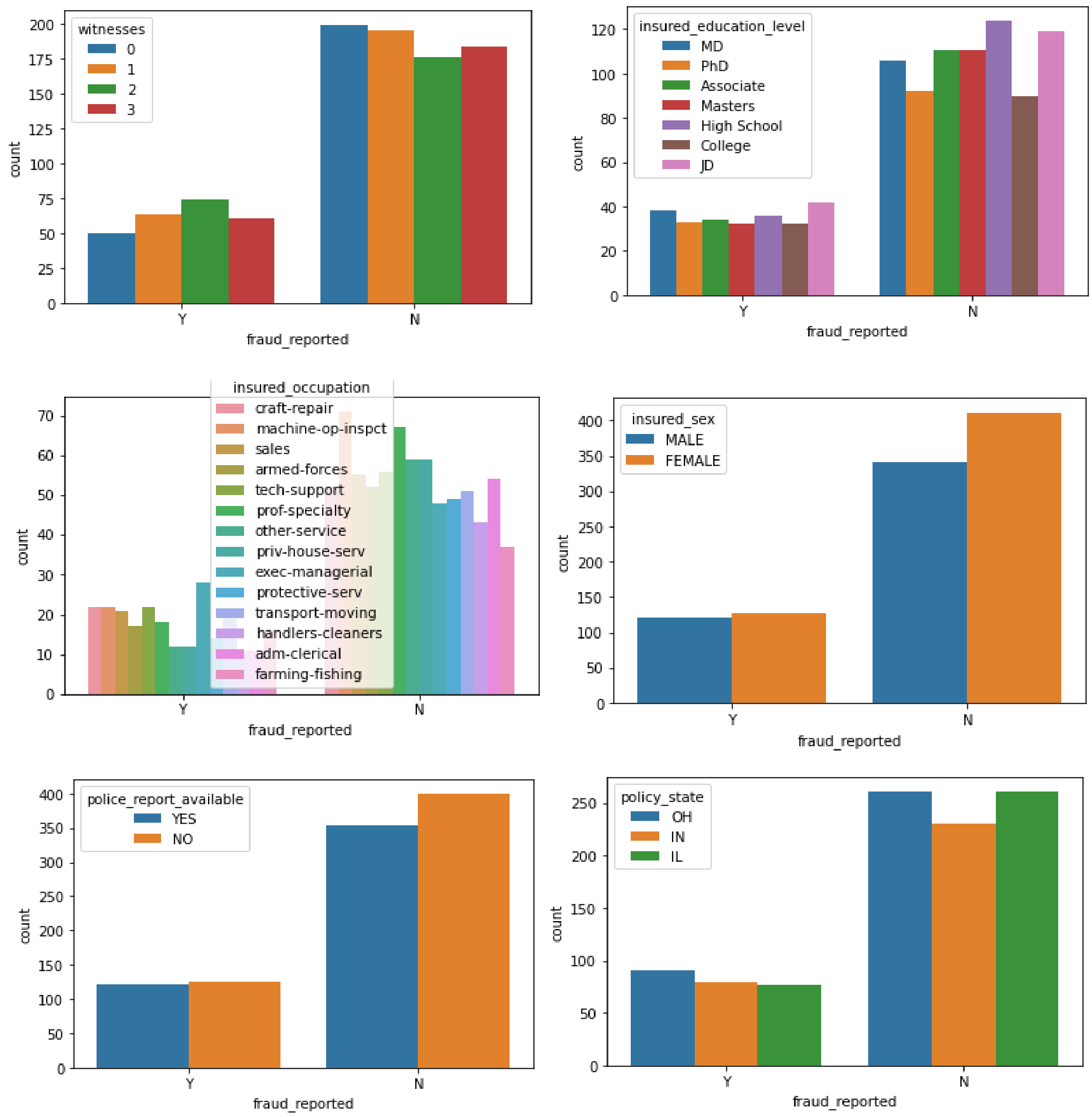
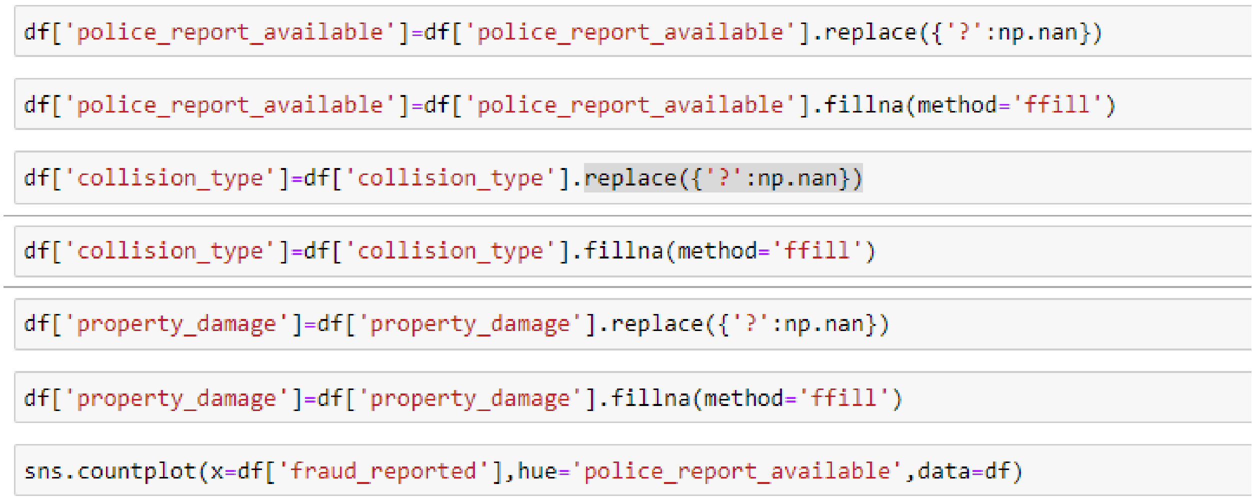


Figure 6 Data Analysis

3.4.1 Data Cleaning

We clean the collected data to remove duplicates, white spaces, and errors. Data cleansing recognizes inaccurate and unfinished parts of the data, filling in missing data and removing data that doesn’t make sense in the study, a process called imputation. The process improves the quality of the dataset and saves training time. This section first drops weak features that we do not intend to use in building the federated model. Features that are removed from the list of columns include: incident_severity, incident_state, incident_type, incident_date, authorities_contacted, incident_city, policy_number, policy_cls, auto_model, insured_hobbies and insured_zip. We then remove all duplicated claims and incomplete claims to remove less significant claims. We fill in missing values by propagating the last valid observation forward to the next valid. We also replace missing values with ‘?’ character with ‘nan’; this is because the python numerical python library ignores the nan values while performing mathematical computations. The operations are shown in figure 5 below.

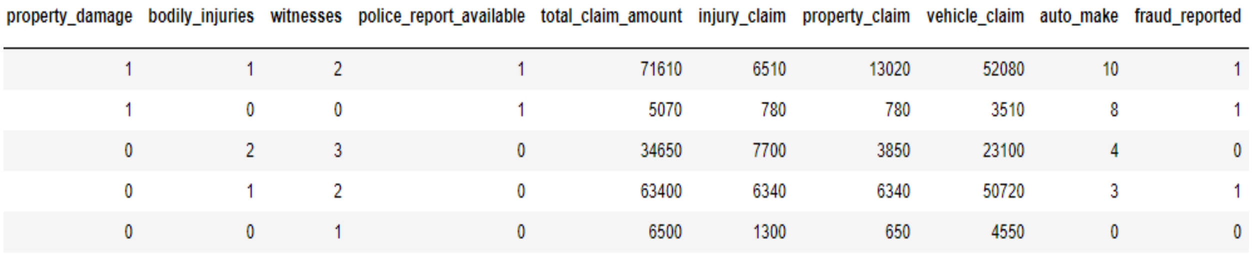


```
df['police_report_available']=df['police_report_available'].replace({'?':np.nan})

df['police_report_available']=df['police_report_available'].fillna(method='ffill')

df['collision_type']=df['collision_type'].replace({'?':np.nan})

df['collision_type']=df['collision_type'].fillna(method='ffill')

df['property_damage']=df['property_damage'].replace({'?':np.nan})

df['property_damage']=df['property_damage'].fillna(method='ffill')

sns.countplot(x=df['fraud_reported'],hue='police_report_available',data=df)
```

Figure 7 Filling Missing Values

3.4.2 Data Transformation

We transform data into formats that machine learning algorithms can understand and model to discover helpful information that will help select the studied features. For example, some machine learning algorithms may not understand text values. We use nominal and ordinal encoding techniques to handle categorical features. Since ML algorithms involve many mathematical calculations, we convert the categorical values into integer or floating-point values that ML algorithms can understand.



property_damage	bodily_injuries	witnesses	police_report_available	total_claim_amount	injury_claim	property_claim	vehicle_claim	auto_make	fraud_reported
1	1	2	1	71610	6510	13020	52080	10	1
1	0	0	1	5070	780	780	3510	8	1
0	2	3	0	34650	7700	3850	23100	4	0
0	1	2	0	63400	6340	6340	50720	3	1
0	0	1	0	6500	1300	650	4550	0	0

Figure 8 Data Transformation

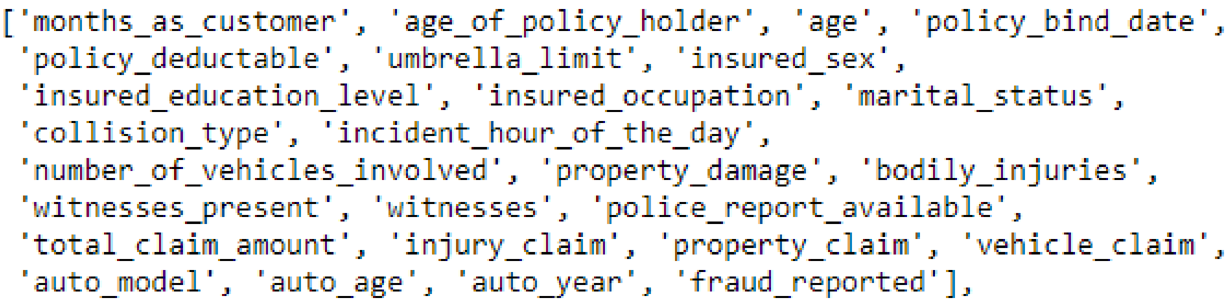
3.5 Feature Engineering

3.5.1 Feature Engineering

Claims datasets often contain different types and shapes of data. Additionally, this data is collected and aggregated from different intervals of time. This data may also be from external sources such as demographic data. One feature for the prediction can be an aggregate of features observed in a dataset in some predefined timeframe. Feature engineering aims to prepare datasets to be compatible with the requirements of a machine learning algorithm and improve the performance of the machine learning model. Research studies have proven that algorithms require features that portray specific characteristics in a domain to work appropriately (Zhang et al., 2019). Different approaches to feature engineering have been studied and proposed for insurance fraud detection. The HOBA approach used by Zhang (Zhang et al., 2019) allows claims to be grouped into homogenous claims while considering different heterogeneity characteristics. The behavioural feature engineering approach by (Johannes & Rajasvaran, 2020) adds policy expiration behaviour in the RFM model using the HOBA methodology. This study uses the classical feature engineering method as outlined in data analysis to select our variables with the forward-selection and backward elimination techniques.

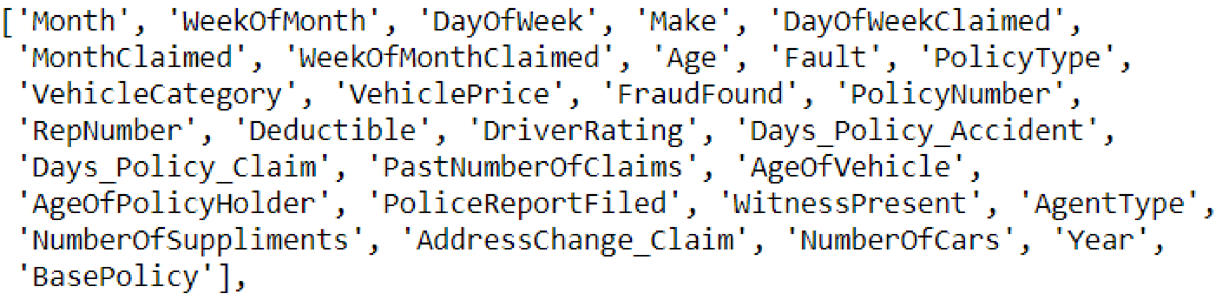
3.5.2 Feature Selection

Insurance claims fraud detection involves many features that can cause increased dimensionality, multicollinearity and overfitting. Studies have shown that as the number of features increases, the classifier’s performance increases to a specific limit of features. Adding more features past the optimal limit degrades the performance of the classifier. The feature selection process selects a subset of features with more information from the original features by removing redundant and irrelevant features without losing information. Because of the many features we have selected to determine whether a claim is fraudulent, as shown in figure 8, we adopted embedded methods (Random Forests) suitable for feature selection in high dimension datasets. Random Forests uses a tree-based strategy that naturally ranks the features by how best they improve the purity of the node. The methods use modelling in their approaches, and hence they account for each feature interaction during training (Johannes & Rajasvaran, 2020).



```
['months_as_customer', 'age_of_policy_holder', 'age', 'policy_bind_date',
 'policy_deductable', 'umbrella_limit', 'insured_sex',
 'insured_education_level', 'insured_occupation', 'marital_status',
 'collision_type', 'incident_hour_of_the_day',
 'number_of_vehicles_involved', 'property_damage', 'bodily_injuries',
 'witnesses_present', 'witnesses', 'police_report_available',
 'total_claim_amount', 'injury_claim', 'property_claim', 'vehicle_claim',
 'auto_model', 'auto_age', 'auto_year', 'fraud_reported'],
```

Figure 9 Kaggle Dataset Selected Features



```
['Month', 'WeekOfMonth', 'DayOfWeek', 'Make', 'DayOfWeekClaimed',
 'MonthClaimed', 'WeekOfMonthClaimed', 'Age', 'Fault', 'PolicyType',
 'VehicleCategory', 'VehiclePrice', 'FraudFound', 'PolicyNumber',
 'RepNumber', 'Deductible', 'DriverRating', 'Days_Policy_Accident',
 'Days_Policy_Claim', 'PastNumberOfClaims', 'AgeOfVehicle',
 'AgeOfPolicyHolder', 'PoliceReportFiled', 'WitnessPresent', 'AgentType',
 'NumberOfSuppliments', 'AddressChange_Claim', 'NumberOfCars', 'Year',
 'BasePolicy'],
```

Figure 10 Oracle Dataset Selected Features

3.6 Federated Model Designs

3.6.1 Feature Alignment

An effective decentralized design was done considering the security concerns in the insurance industry. To align features, clients participating in the training will securely download a list of standard participating features from the master node. Since training is asynchronous, all nodes don't have to be online for training, making the alignment process more practical. In addition, the participants can recommend features that may not be in the list of standard features downloaded by re-uploading the document with an updated list. After approval from participating members, a new list of features is made available for download.

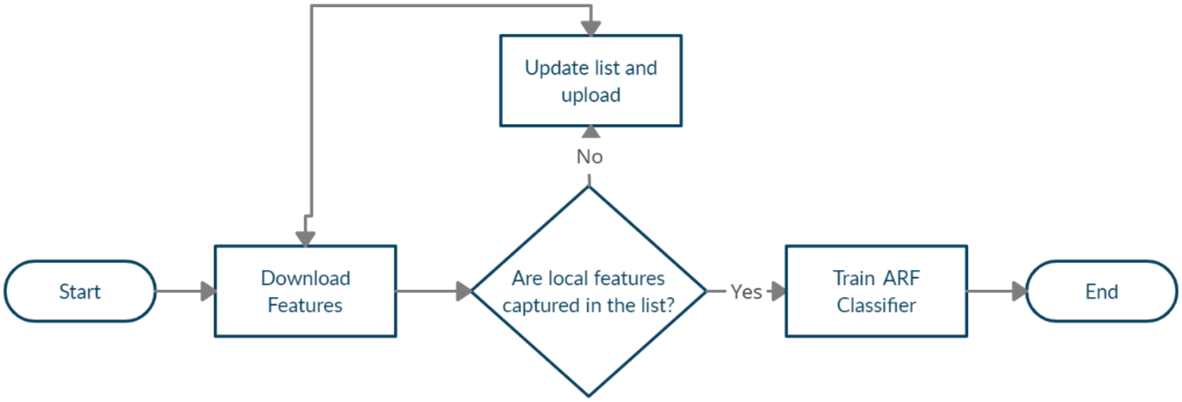


Figure 11 Feature Alignment

3.6.2 Federated Adjusted Random Forests

Fraud detection can be considered a classification problem where machine learning models need to classify input features and predict which class the features fall. Suppose we assume that our predicted variable falls in either two classes (fraud or no fraud). In that case, categorical classification techniques cannot be applied, and binary classification techniques solve this problem. In the study by (Dhieb et al., 2019), an extreme gradient boosting algorithm was used to detect the status of a claim for insurance applications as fraudulent, not fraudulent, and other categories. This study presents a federated adjusted random forests algorithm based on the CART tree to deal with the class imbalance problem and make it a practical approach. The adjusted random forest algorithm works as shown below.

i. For each round, we randomly select a sample of fraudulent claims with a replacement that is from the lower class. We also select a similar number of non-fraudulent claims with replacement.
ii. We create a tree with the sample above to a full depth, i.e. without pruning. A split is made on each node should be based on a set of randomly selected features. This makes the correlation between the trees to be lower.
iii. We repeat steps (i) and (ii) for n times and use bootstrapped aggregation to find the final result we use as the final predictor.

The Adjusted Random Forest classifier doesn’t overfit imbalanced data because it performs tenfold cross-validations at every iteration level. Figure 8 below illustrated the functionality of adjusted random forest.

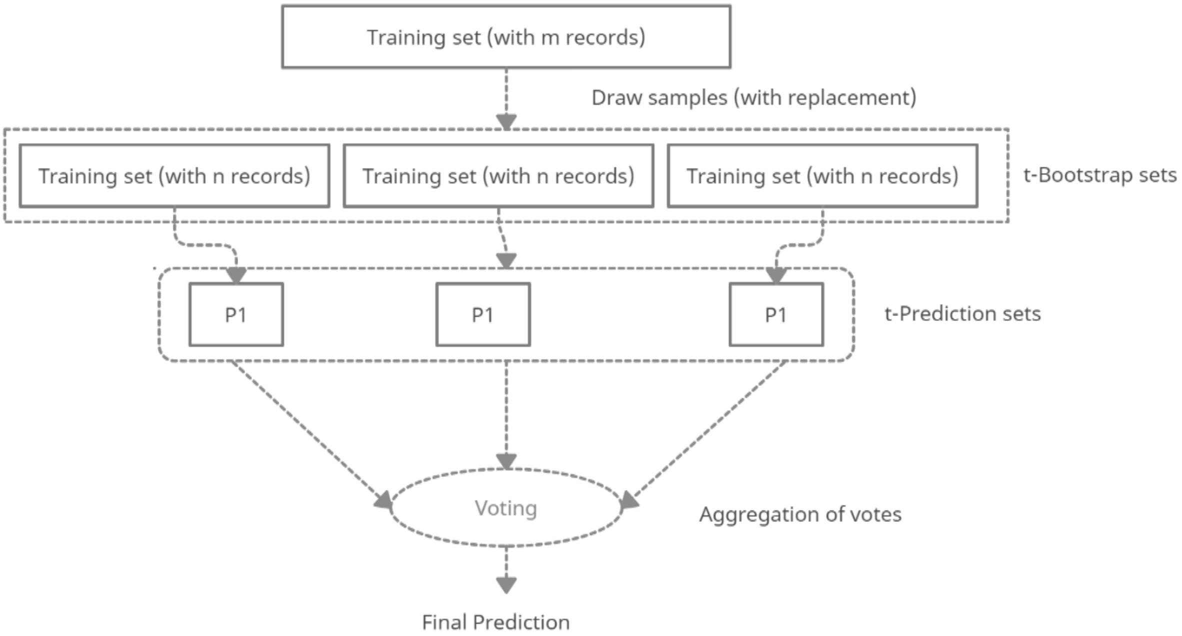

Training set (with m records)
Draw samples (with replacement)
Training set (with n records)
Training set (with n records)
Training set (with n records)
t-Bootstrap sets
P1
P1
P1
t-Prediction sets
Voting
Aggregation of votes
Final Prediction

Figure 12 Bagging Using Adjusted Random Forest

Each client trains an ensemble adjusted random forest classifier on their data. Training on the Individual node happens such that individual models are trained in parallel; a random subset of data trains each model. Thus, the adjusted random federated forest becomes, by default, an ensemble model of bagged federated trees. Figure 9 below shows the framework that is based on the CART tree. The framework can be used for both classification and regression problems.

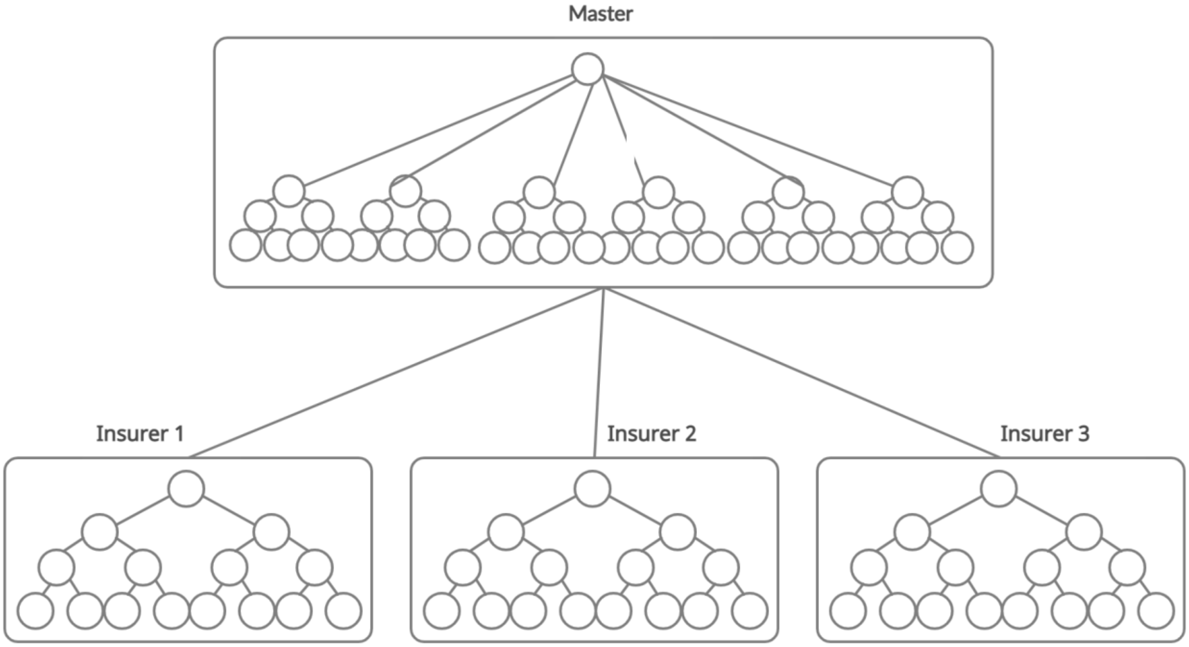


Figure 13 Random Federated Forests

3.6.3 Horizontal Federated Learning

This type of learning makes use of standard features across devices involved in the learning. This, therefore, means that client A and client B have the same set of features. For example, two insurance companies in different regions and their customers come from their respective areas. Suppose customers in the two companies present a frequency in purchasing products in the General Insurance Category. Their mutual intersections are minor, making them belong to the same feature space. Therefore, the two companies can contribute to a joint model (Yang et al., 2019). This study makes use of this type of learning. We assume that all participating Insurers belong to the same feature space. Each member's standard feature space is addressed by downloading standard features before training their models. If an Insurer finds a vital feature missing, they can recommend the feature to participate members who approve it on a solid predictor.

3.6.4 Decentralized Architecture Design

The Insurance Regulatory Authority is a government regulatory agency mandated by law to regulate, supervise and develop the insurance industry (Insurance Regulatory Authority, 2020). In regulating and developing the Industry, IRA is a natural target as a central node and ensures fairness in rewarding participating members. IRA will be a choice for the central node that will manage other insurers and reinsurers. Each primary insurer downloads a list of participating features from the central node. They then train their model with their data on their servers and upload them to the Insurance Regulatory Authority centralized node. IRA ensembles the edge models through aggregation to form a standard centralized model and distributes them back to all the primary insurers. This approach will be a simple version of Federated Learning because the number of insurers is expected to be smaller. In addition, the communication bandwidth between enterprise servers is expected to be more than that of mobile phones. The study uses serialization and encryption of the models during transfer to and from the central node to protect privacy for policyholders, protect the insurer's data, and benefit from the global federated model. Figure 6 below illustrates the federated design that effectively trains a shared model in the Insurance setup.

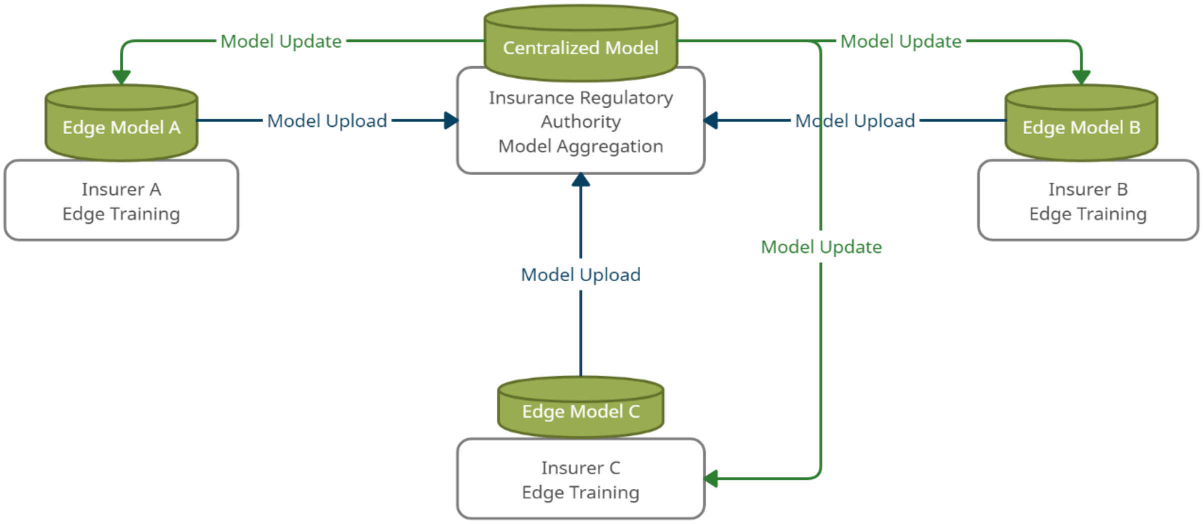
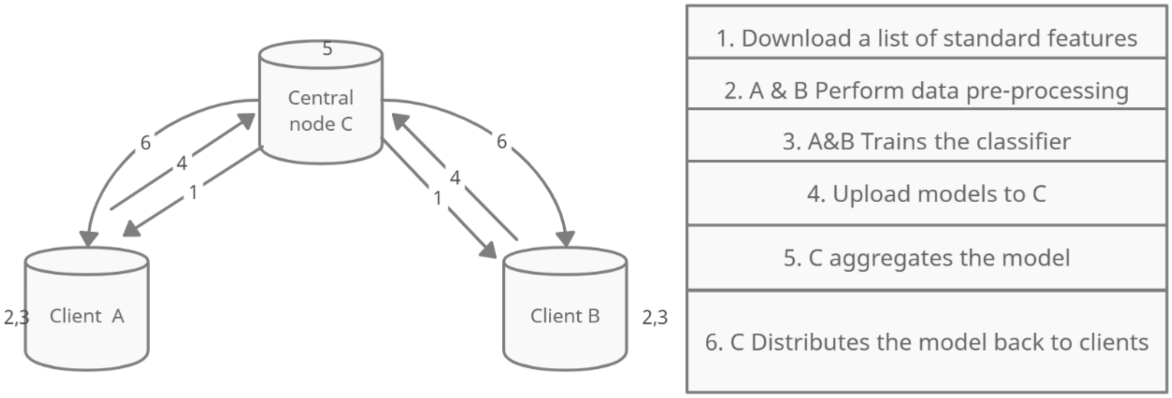


Figure 14 Decentralized Design

3.6.5 Decentralized Algorithm Design

An optimal decentralized algorithm is implemented by improving the asynchronous limitations of the existing federated forest algorithm presented by (Liu et al., 2019). In our work, the forest classifier is built and evaluated for accuracy and efficiency on the node. As opposed to (Liu et al., 2019), all parties don’t have to work together to build the federated model making the process asynchronous and therefore eliminating points of failures and problems of multiple dependencies as reported by (Ongati & Lawrence, 2019). After downloading a list of similar features from the central node, the dataset is ready to train a model. The training process is asynchronous, which makes client nodes ensure the confidentiality of their training process. Using Random federated forests as an example, training the model is divided into four steps, as shown in Figure 12 below.



1. Download a list of standard features
2. A & B Perform data pre-processing
3. A&B Trains the classifier
4. Upload models to C
5. C aggregates the model
6. C Distributes the model back to clients

Figure 15 Federated Algorithm

1. Clients A and B downloads the latest list of standard features.
2. Clients A and B perform data preprocessing and analysis to align data to the downloaded features.
3. Clients A and B independently trains their classifier using Adjusted Random forests until they reach a splitting state.
4. Clients A and B upload their serialized models to the central Node C for aggregation.

5. Central node C de-serializes the uploaded models. The central node also aggregates the models by combining the estimators from individual trees.
6. The central node C then distributes the combined model back to the client's nodes and completes a single training cycle.

We repeat the process above when a single client node uploads an updated model; The respective data of A and B is kept locally and not shared. There is no interaction during training; therefore, no sharing of client data.

3.7 Implementation and Prototyping

This study used the python programming language to develop model architectures. We also used secure shell protocol (SSH) for communication and transfer of model representations and weights. The environment was set up on personal computers, running on Windows 10 Pro operating system with Python 3.6 installed with NumPy, pandas and sklearn python libraries. First, we loaded data from (Roshan, 2019) with 1000 claim records and got it through the data analysis steps as shown in sub-section 3.5. We assigned the target variable "fraud reported" to be our y and dropped it from the x-axis. Next, we shuffled the data with a random state to reproduce it when the code is rerun. Shuffling helps to balance the data to reduce noise.

We then split our training data into 80% training sets and 20% test/validation sets. Studies have shown that assigning more samples (n) to the training set improves accuracy (Dobbin & Simon, 2011). We repeated the same process with the dataset from (Charlie, 2010). The dataset contains 15420 claim instances between January 1994 to December 1996 and an average of 430 claims per month. We used embedded methods (Random Forests) suitable for feature selection in high dimension datasets. The methods use modelling in their approaches, and hence they account for each feature interaction during training (Johannes & Rajasvaran, 2020). For each dataset, we recorded metrics before feature selection and after feature selection.

The model development involved classical algorithms as well as our federated algorithms. On each algorithm, we trained and recorded the metrics required to evaluate its performance. The selected algorithms included the classical Random Forest Classifier, adjusted random forest classifier and Extreme Gradient boosting classifier. We created 20 Random forest classifiers and recorded the evaluation metrics for each forest. We also created 20 adjusted random forest classifiers and recorded the evaluation metrics. The last step involved combining individual bagged random forest models into one giant federated model by aggregating its estimators. We tested both the classical and the federated models on the two datasets. We then calculated the results that are considered as the performance criterion and did a comparison.

3.8 Model Evaluation

After the training of local random forest and adjusted random forest classifiers, the updates are aggregated to the global federated model. The evaluation is composed of metrics computed on classical algorithms and those computed from the federated model. The performance of the federated global model is compared against implementing the same model in a centralized architecture where all the required data is centrally aggregated. We used three main categories of classification metrics: classification accuracy, confusion matrix, and a classification report that includes precision, recall, and F-score. We evaluated the model based on its prediction accuracy on the test data. We used the cross-validation technique that involves partitioning the original dataset into a training dataset used during model training and an evaluation dataset used to evaluate the analysis.

3.9 Ethical Considerations

According to the industry’s laid down procedures, relevant permissions were sought from the Insurance Regulatory Authority and Association of Kenya Insurers before using insurance data. The institutions regulate research and innovation activities in the insurance industry. Data collection was done by sourcing public datasets from online platforms, posing no risk to the policyholder. The study was conducted openly to ensure transparency for all interested parties.

CHAPTER FOUR

4 RESULTS AND DISCUSSIONS

4.1 Introduction

The federated model draws its validity from performance reports measured during experiments and prototyping. Therefore, it is essential to know whether we can live with the model's performance and its ability to improve the performance of fraud detection algorithms. In this chapter, we report the performance of our adjusted federated random forest approach. We also evaluated the federated model by comparing its performance against classical algorithms used in this study. We then discuss the effectiveness of the approach presented in this study to build fraud detection models in the insurance setup.

4.2 Evaluation Results

The results first detail the performance of the classical random forest model and the model's performance when using a balanced random forest classifier. The performance of the other classical algorithms chosen in this study is also measured. We use different evaluation measures to quantify the model's performance. In this study, we used the following classification metrics: Classification accuracy, confusion matrix, and a classification report that includes: precision, recall and F-score. Different evaluation methods were used in this study because of the class imbalance problem in our dataset. This helps us to avoid bias by subjecting the models to multiple evaluations.

4.2.1 Classification Accuracy

This score is used to measure the percentage of labels the model predicted accurately over the total number of predictions. The federated random forest classifier performs better than most algorithms. As shown in the table, the federated adjusted random forest classifier had the lowest accuracy score. However, it should be noted that accuracy is not a great measure of classifiers performance, especially when classes are imbalanced, as is our case. The accuracy of the models for the study is given below.

Model	Accuracy Score
Logistic Regression	0.755
Random Forest Classifier	0.75
Balanced Random Forest Classifier	0.5
Gradient Boosting Classifier	0.695
Extreme Gradient Boosting Classifier	0.72
Federated Random Forest Classifier	0.755
Federated Balanced Random Forest Classifier	0.51

Table 1 Accuracy Score for Kaggle Dataset

Model	Accuracy Score
Logistic Regression	0.94
Random Forest Classifier	0.94
Balanced Random Forest Classifier	0.66
Gradient Boosting Classifier	0.94
Extreme Gradient Boosting Classifier	0.96
Federated Random Forest Classifier	0.94
Federated Balanced Random Forest Classifier	0.63

Table 2 Accuracy for Oracle Dataset

4.2.2 Confusion Matrix

We plotted a confusion matrix to present a notable performance. The confusion matrix shows how many of the classifier’s predictions we correctly classified, and it also tells us where the classifier got confused. We present two confusion matrices: federated Random Forest Classifier before balancing the classifier and after balancing the classifier for the two datasets.

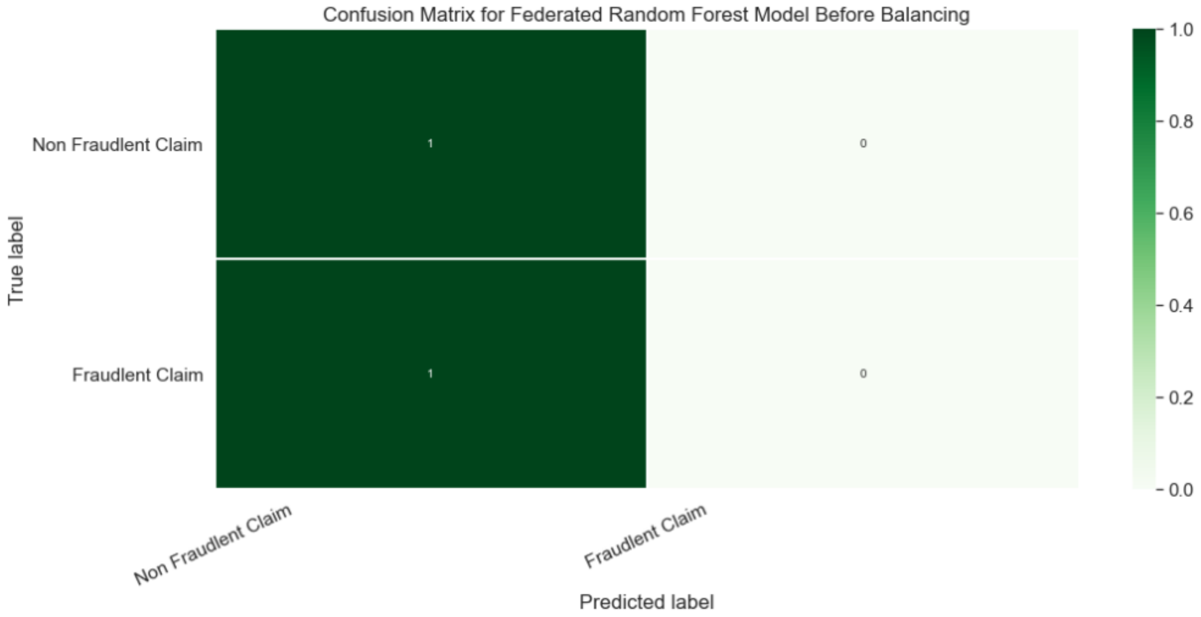


Figure 16 Federated Random Forest Before Balancing-Kaggle Dataset

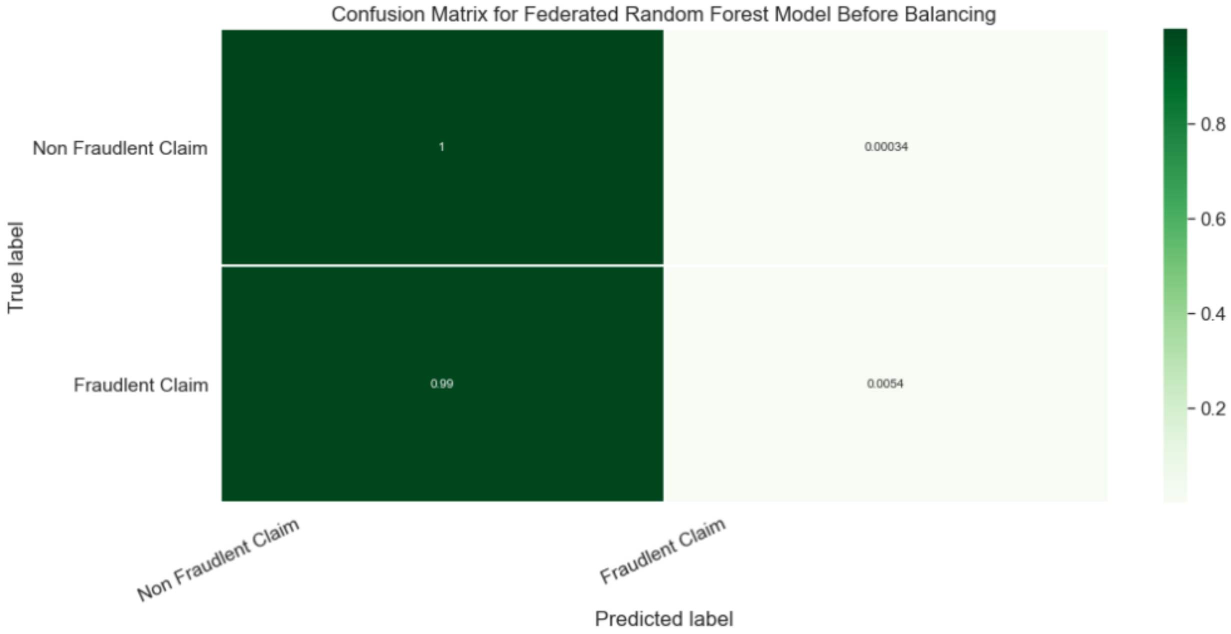


Figure 17 Federated Random Forest Before Balancing-Oracle Dataset

As shown in the confusion matrix figures 16, the unbalanced random forest model found difficulties predicting a fraudulent claim at 0% and predicted 99% of the fraudulent claims as non-fraudulent. However, the model also scored higher at 100% at predicting non-fraudulent claims and predicted 0% of non-fraudulent claims as fraudulent. Figure 17 shows similar results; it records 100% in predicting non-fraudulent claims. The model got confused and predicted 99% Fraudulent claims as Not fraudulent.

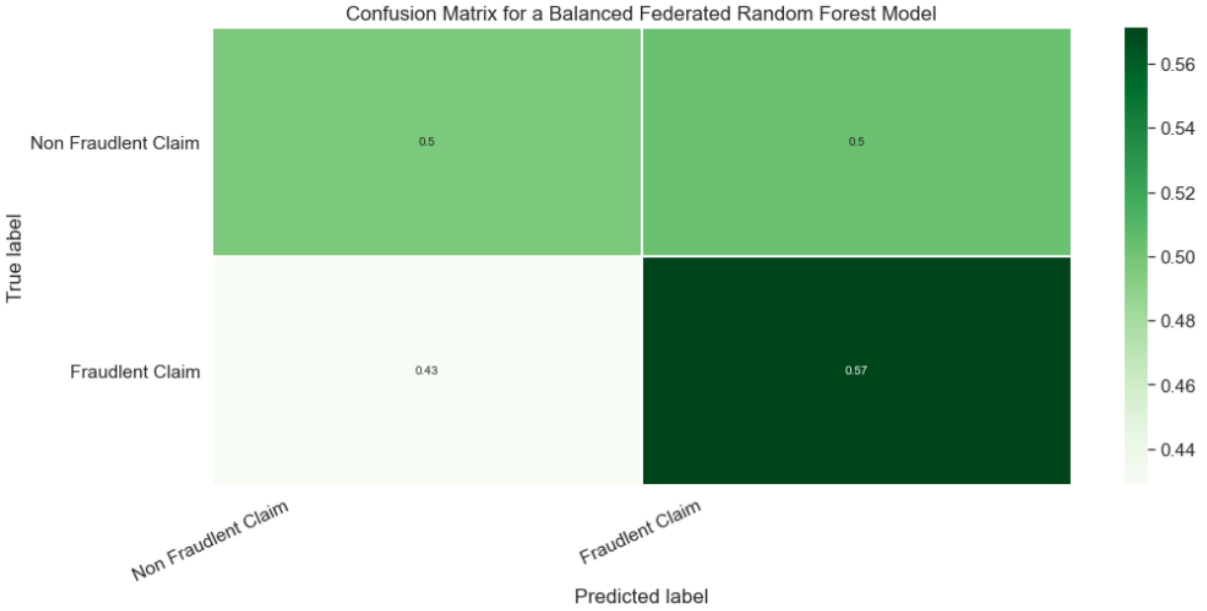


Figure 18 Balanced Federated Random Forest-Kaggle Dataset

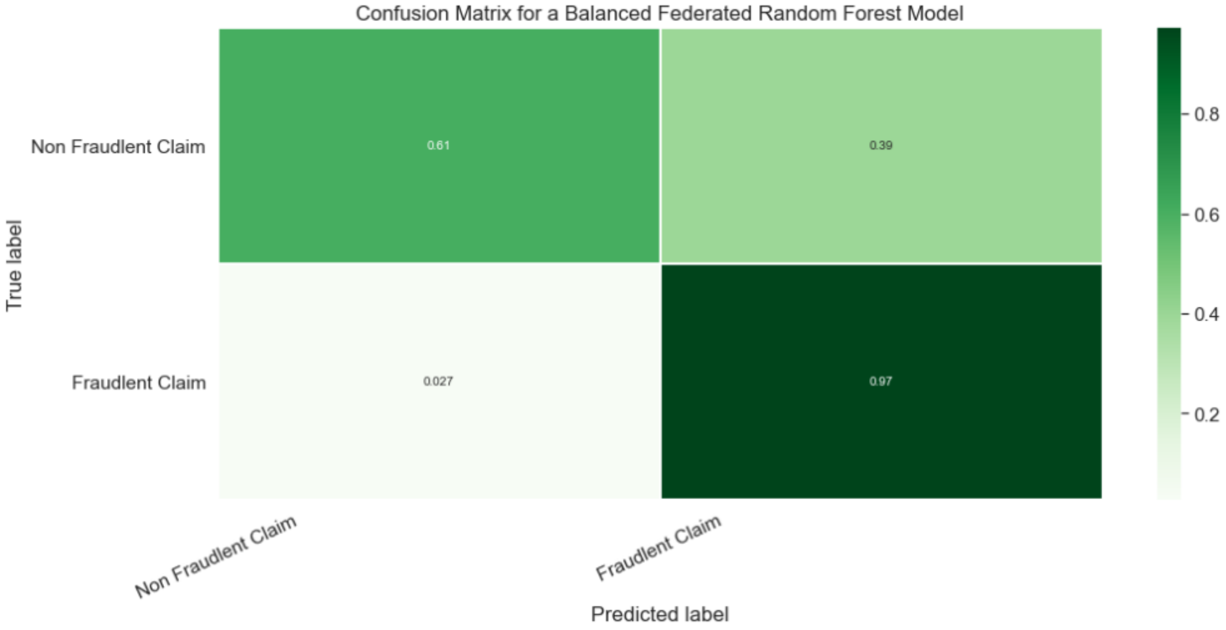


Figure 19 Balanced Federated Random Forest- Oracle Dataset

The confusion matrix in figure 18 shows the result for the federated adjusted random forest classifier on the Kaggle dataset. The classifier gets confused and classifies 50% of the non-fraudulent claims as fraudulent and 43% of the fraudulent claims as non-fraudulent. The model classifies 50% of the non-fraudulent claims correctly. The performance of the model improves and classifies 57% of the fraudulent claims correctly. The confusion matrix in figure 19 shows the results for the balanced federated random forest on the Oracle dataset. The federated model presents notable performance on the dataset. The classifier correctly identifies 61% of non-fraudulent claims and 97% of fraudulent claims. The percentage of incorrectly classified claims significantly reduced. The classifier recorded 27% for the fraudulent claims classified as non-fraudulent and 39% for the non-fraudulent claims classified as fraudulent.

4.2.3 Classification Report

We examine other metrics like precision, recall, and F1 score to get more insight into our model's performance. Precision is a fraction of members of a class that were correctly identified amongst all members that were predicted to belong in a particular class. A model's recall is a fraction of members who were predicted to belong to a class amongst all of the members that belong to the class. F1-Score combines both precision and recall metrics into one metric. If precision and recall are high, F1-score will be high, and if they are low, the F1-Score will be lower. The figures below give classification reports for the algorithms in this study.

Model Name	**AVG Precision**	**AVG Recall**	**Avg F1-Score**	**Accuracy**
Federated Balanced Random Forest Classifier	0.53	0.53	0.49	0.52
Federated Random Forest Classifier	0.38	0.50	0.43	0.76
Balanced Random Forest Classifier	0.55	0.56	0.53	0.57
Random Forest Classifier	0.50	0.50	0.45	0.74

Table 3 Classification Report Before Feature Selection-Kaggle Dataset

Model Name	AVG Precision	AVG Recall	Avg F1-Score	Accuracy
Federated Balanced Random Forest Classifier	0.57	0.79	0.50	0.63
Federated Random Forest Classifier	0.72	0.50	0.49	0.94
Balanced Random Forest Classifier	0.56	0.76	0.51	0.66
Random Forest Classifier	0.80	0.51	0.51	0.94

Table 4 Classification Report Before Feature Selection-Oracle Dataset

Model Name	AVG Precision	AVG Recall	Avg F1-Score	Accuracy
Federated Balanced Random Forest Classifier	0.74	0.78	0.75	0.80
Federated Random Forest Classifier	0.66	0.60	0.61	0.76
Balanced Random Forest Classifier	0.72	0.78	0.73	0.77
Random Forest Classifier	0.73	0.63	0.65	0.79

Table 5 Classification Report After Feature Selection-Kaggle Dataset

Model Name	AVG Precision	AVG Recall	Avg F1-Score	Accuracy
Federated Balanced Random Forest Classifier	0.57	0.79	0.50	0.63
Federated Random Forest Classifier	0.72	0.50	0.49	0.94
Balanced Random Forest Classifier	0.56	0.76	0.51	0.66
Random Forest Classifier	0.83	0.51	0.51	0.94

Table 6 Classification Report After Feature Selection-Oracle Dataset

4.3 Discussion

The federated insurance claims fraud detection approach we adopted in this study was a decentralized approach that solves the class imbalance problem and the data privacy problem. The proposed federated architecture and algorithm also enables insurance players to collaborate in the training of quality models. In this part, we discuss the efficiency of the federated adjusted random forest classifier algorithm against the results of the classical algorithms. We, therefore, look at the effectiveness of the approach on imbalanced datasets and evaluate our model based on its efficiency in securely learning from decentralized datasets. We also evaluate its efficiency and in detecting insurance claims fraud. We tested the model on imbalanced datasets drawn from Kaggle datasets (Roshan, 2019) and Oracle (Charlie, 2010). The interpretation of the results aims to help us validate our federated approach in detecting Insurance fraud.

Many research shows that the quality and quantity of available data significantly impact insurance fraud detection models (Johannes & Rajasvaran, 2020) (R Guha et al., 2017). The federated models performed poorly on the Kaggle dataset (Roshan, 2019). The performance is majorly attributed to the small number of claims (1000) and more significant variance in the classes. The federated models, however, performed well on the oracle dataset (Charlie, 2010). The performance is attributed to a large number of claims (15420). Therefore, the quality and quantity of data at the nodes have significant effects on the overall performance of the federated model. The overall performance of the federated model depends on the performance of the local model.

There is a significant improvement in all algorithms when using the embedded Random Forest feature selection method. The federated adjusted random forest classifier records the highest balance of metrics after feature selection, increasing accuracy to 28% from 52% on the Kaggle dataset. As discussed by (Johannes & Rajasvaran, 2020), Random Forests uses a tree-based strategy that naturally ranks the features by how best they improve the purity of the node, and they, therefore, account for each feature interaction during training. We required additional information from the classification report and confusion matrix to help us understand how well the model performed, how it performed on each class and where it found difficulties to distinguish classes. We note that accuracy is not a great measure of classifiers performance, especially when classes are imbalanced, as is our case.

The confusion matrix shows that the federated random forest classifier performs poorly on imbalanced datasets because it introduces bias. The model is therefore not suitable for insurance claims fraud prediction. Therefore, we cannot use it in a federated setting and collaborative learning. On the other hand, the balanced federated random forest classifier gives an exemplary performance on the imbalanced insurance dataset and outshines its classical counterpart on both datasets. Moreover, the model has a balanced representation of the classification metrics making it an optimal model for collaborative machine learning in detecting insurance claims fraud.

4.4 Model Verdict

We performed experiments to benchmark the balanced federated random forest algorithm and its performance in classifying fraudulent and honest claims from decentralized datasets. As seen in the results, the machine learning models presented varying performance levels on different input datasets. Therefore, we used the average F1-score metric that provided the model's overall performance to rank the models. The F1-score results show that the federated adjusted random forest classifier outshines the classical algorithms. In addition, the method performs well in highly skewed datasets, as is our case. Finally, the results demonstrate that the adjusted random forest ensemble classifier performs well on decentralized datasets. The table below compares federated models against classical models using the average F1-score; the higher the average F1 score, the better the model's performance.

Model Name	Avg F1-Score	Rank
Federated Balanced Random Forest Classifier	0.63	1
Federated Random Forest Classifier	0.55	4
Balanced Random Forest Classifier	0.62	2
Random Forest Classifier (Johannes & Rajasvaran, 2020)	0.58	3

Table 7 Models Average Performance

CHAPTER FIVE

5 CONCLUSION AND RECOMMENDATIONS

The field of federated learning has attracted attention from researchers, and its growth has just begun. This research, offered us a platform to demonstrate that federated learning can be applied to solve real-world challenges in the insurance setup. This research also demonstrates that insurance industry players can collaborate and train a model that benefits all as opposed to individual companies. This research also shows that indeed adjusted random forest models can be trained in a decentralized way. The model helps insurance risk managers make informed decisions on claims by detecting insurance claims fraud while reducing the loss ratio.

5.1 Conclusion

This study proposed a federated ensemble and an adjusted tree-based machine learning approach, which is asynchronous, meaning all nodes don’t have to be online for training to occur. The method, therefore, solves the problem of synchronous learning as shown by (Liu et al., 2019) and the challenges of having a central coordinator to manage all the nodes during training as shown by (Ongati & Lawrence, 2019). Furthermore, the federated adjusted random forest model solves the class imbalance problem on the dataset by performing tenfold cross-validation at every iteration level during training as it builds the forest. As a result, the adjusted random forest classifier was able to identify most of the fraudulent cases with a higher precision, F1-score and a lower false-positive rate compared to the classical counterpart proposed by (Johannes & Rajasvaran, 2020).

The federated model for insurance claims fraud prediction presented in this research was tested and found efficient for the detection of fraudulent claims in the insurance setup. The federated model allows all players in insurance to collaboratively build efficient fraud detection models without sharing data. The federated model for insurance claims fraud prediction allows smaller insurance entities to detect claim fraud without sufficient insurance data. Furthermore, the asynchronous federated machine learning approach achieves data privacy by allowing each node to train an ensemble adjusted random forest classifiers. Generally, this solution is an important asset for insurance entities as it helps them reduce the loss ratio.

5.2 Recommendation

We selected IRA as our central node that aggregates models weights. We however recommend a neutral node applicable in the region to be a central node that aggregates the model weights. Given the difference in insurance fraud predictors at different geographical locations, we recommend that each implementation of this method adopts a feature engineering and selection method that best suits the fraud scenario in the specific regions. It is also essential that participants in a collaborative machine learning model do sufficient tests on their local models before uploading, as poor models will affect the overall performance of the federated model. Quality local models can be built by ensuring that sufficient data is aggregated before training.

5.3 Future Research

Given the inherent characteristics of various datasets on different nodes, it would be impractical to recommend an optical technique or a feature engineering methodology that would best perform the model at the node. Therefore, a future study can present a standard feature engineering methodology used with federated learning in the insurance setup. Future research can also focus on a verification method that will be used to verify the updates before they are aggregated. The method will ensure that poor updates don’t spoil the quality model. The method can verify the model updates by running the update on the test dataset before they are accepted and integrated into the global model.

6 REFERENCES

Association of Certified Fraud Examiners. (2019). *INSURANCE FRAUD HANDBOOK*. Association of Certified Fraud Examiners, Inc.

Burri, R. D., Burri, R., Reddy Bojja, R., & Rao Buruga, S. (2019). Insurance Claim Analysis using Machine Learning Algorithms. *International Journal of Innovative Technology and Exploring Engineering*, *8*(6S4), 577–582. https://doi.org/10.35940/ijitee.F1118.0486S419

Charlie, B. (2010, January 18). *Fraud and Anomaly Detection Made Simple*. https://blogs.oracle.com/machinelearning/fraud-and-anomaly-detection-made-simple

Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2019). Extreme Gradient Boosting Machine Learning Algorithm For Safe Auto Insurance Operations. *2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, 1–5. https://doi.org/10.1109/ICVES.2019.8906396

Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics*, *4*(1), 31. https://doi.org/10.1186/1755-8794-4-31

Insurance Regulatory Authority. (2020). *INSURANCE INDUSTRY ANNUAL REPORT 2019* [ANNUAL REPORT]. Insurance Regulatory Authority.

Johannes, S. K., & Rajasvaran, L. (2020). AUTO-INSURANCE FRAUD DETECTION: A BEHAVIORAL FEATURE ENGINEERING APPROACH. *Journal of Critical Reviews*, *7*(03). https://doi.org/10.31838/jcr.07.03.23

Konečný, J., McMahan, H. B., Ramage, D., & Richtárik, P. (2016). Federated Optimization: Distributed Machine Learning for On-Device Intelligence. *ArXiv:1610.02527 [Cs]*. http://arxiv.org/abs/1610.02527

Liu, Y., Liu, Y., Liu, Z., Zhang, J., Meng, C., & Zheng, Y. (2019). Federated Forest. *ArXiv:1905.10053 [Cs, Stat]*. https://doi.org/10.1109/TBDATA.2020.2992755

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. *ArXiv:1602.05629 [Cs]*. http://arxiv.org/abs/1602.05629

Ongati, F., & Lawrence, M. (2019). *Big Data Intelligence Using Distributed Deep Neural Networks*. 7.

Palacio, S. M. (2019). Abnormal Pattern Prediction: Detecting Fraudulent Insurance Property Claims with Semi-Supervised Machine-Learning. *Data Science Journal*, *18*(1), 35. https://doi.org/10.5334/dsj-2019-035

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). *A Comprehensive Survey of Data Mining-based Fraud Detection Research*. 14.

R Guha, Shreya Manjunath, & Kartheek Palepu. (2017). Comparative Analysis of Machine Learning Techniques for Detecting Insurance Claims Fraud. *Wipro*, 19.

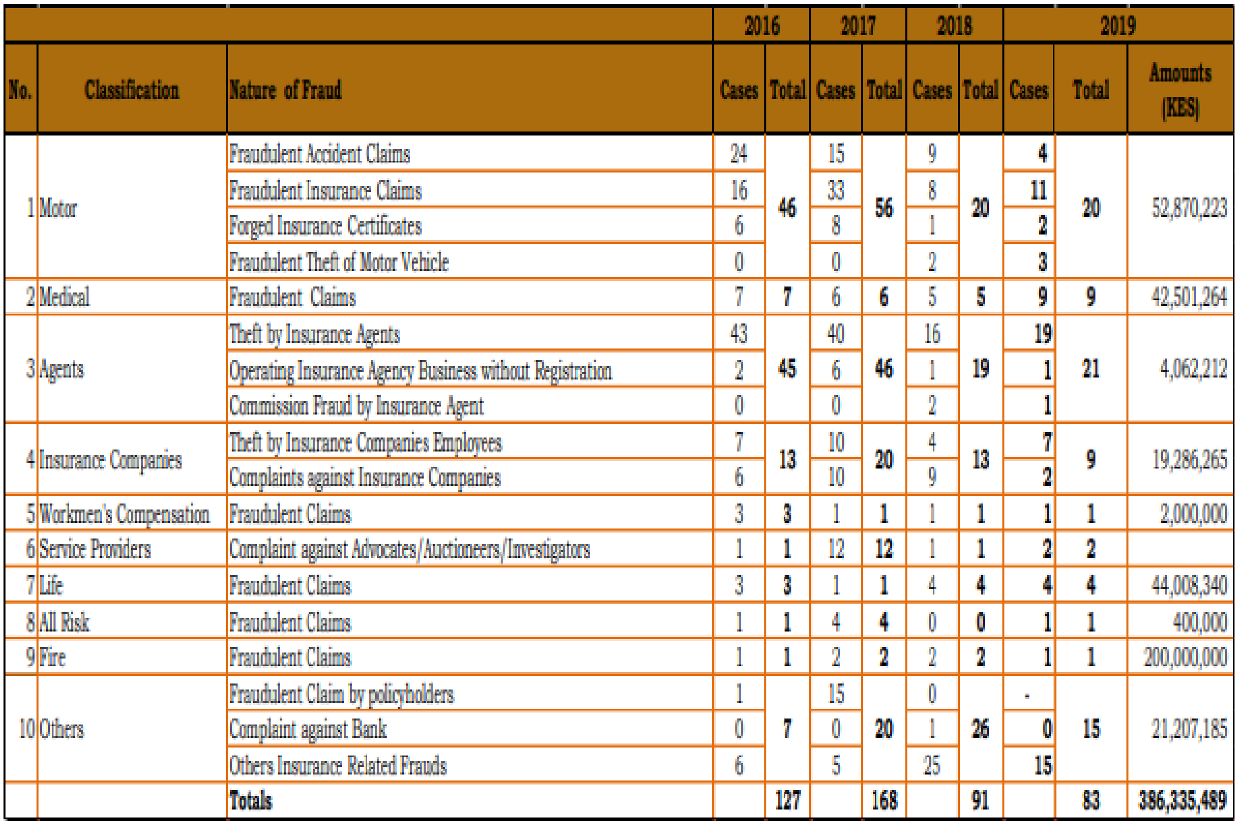
Roshan, S. (2019, March 7). *Insurance Claim*. https://www.kaggle.com/roshansharma/insurance-claim

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Machine Learning: Concept and Applications. *ArXiv:1902.04885 [Cs]*. http://arxiv.org/abs/1902.04885

Zhang, X., Han, Y., Xu, W., & Wang, Q. (2019). HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. *Information Sciences*, *557*, 302–316. https://doi.org/10.1016/j.ins.2019.05.023

7 APPENDICES

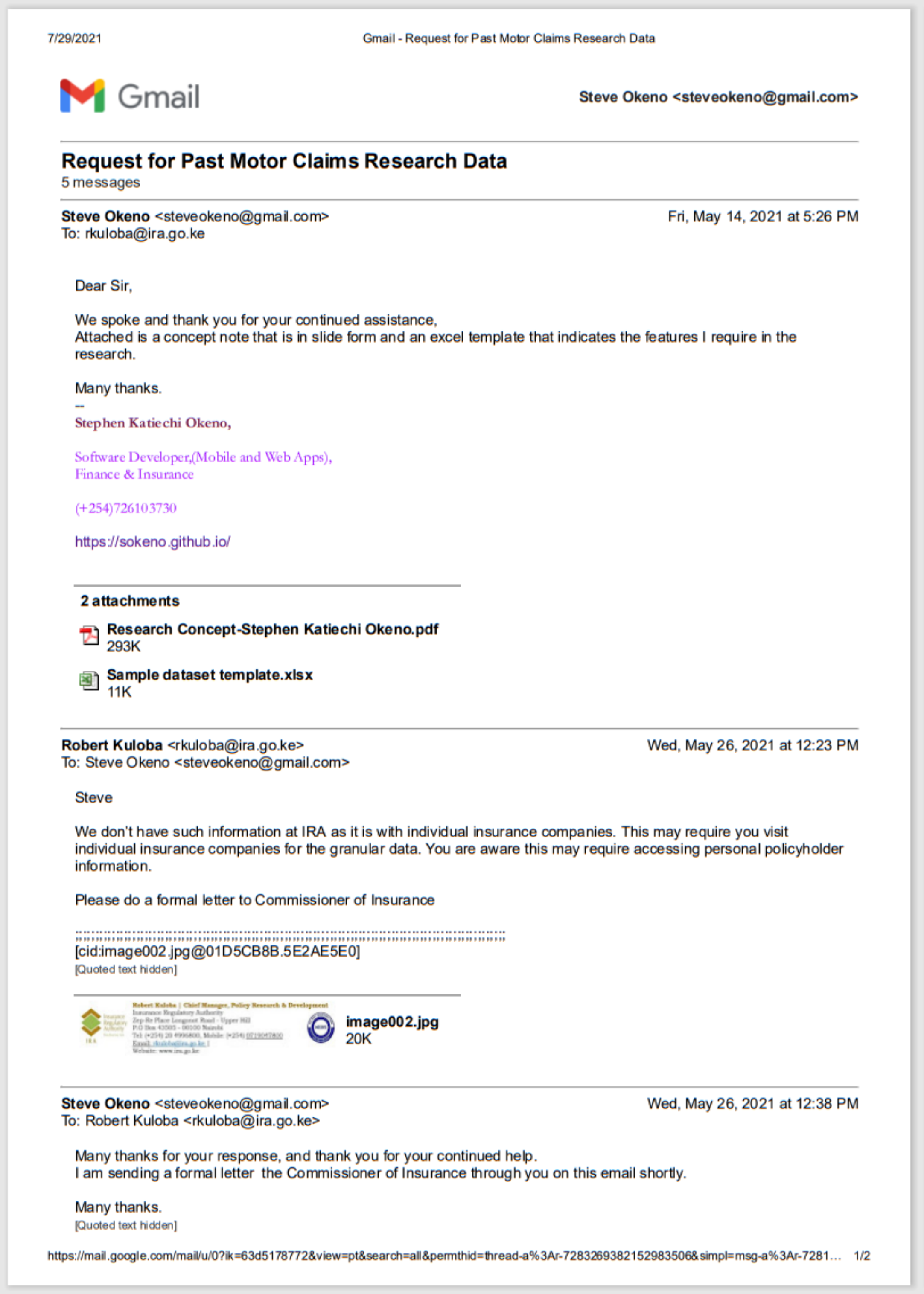
7.1 IRA Statistics



			2016		2017		2018		2019		
No.	Classification	Nature of Fraud	Cases	Total	Cases	Total	Cases	Total	Cases	Total	Amounts (KES)
1	Motor	Fraudulent Accident Claims	24	46	15	56	9	20	4	20	52,870,223
		Fraudulent Insurance Claims	16		33		8		11		
		Forged Insurance Certificates	6		8		1		2		
		Fraudulent Theft of Motor Vehicle	0		0		2		3		
2	Medical	Fraudulent Claims	7	7	6	6	5	5	9	9	42,501,264
3	Agents	Theft by Insurance Agents	43	45	40	46	16	19	19	21	4,062,212
		Operating Insurance Agency Business without Registration	2		6		1		1		
		Commission Fraud by Insurance Agent	0		0		2		1		
4	Insurance Companies	Theft by Insurance Companies Employees	7	13	10	20	4	13	7	9	19,286,265
		Complaints against Insurance Companies	6		10		9		2		
5	Workmen's Compensation	Fraudulent Claims	3	3	1	1	1	1	1	1	2,000,000
6	Service Providers	Complaint against Advocates/Auctioneers/Investigators	1	1	12	12	1	1	2	2	
7	Life	Fraudulent Claims	3	3	1	1	4	4	4	4	44,008,340
8	All Risk	Fraudulent Claims	1	1	4	4	0	0	1	1	400,000
9	Fire	Fraudulent Claims	1	1	2	2	2	2	1	1	200,000,000
10	Others	Fraudulent Claim by policyholders	1	7	15	20	0	26	-	15	21,207,185
		Complaint against Bank	0		0		1		0		
		Others Insurance Related Frauds	6		5		25		15		
		Totals		**127**		**168**		**91**		**83**	**386,335,489**

Table 8 IRA Statistics

7.2 Email Correspondence to Request for Insurance Data



7/29/2021 Gmail - Request for Past Motor Claims Research Data

Steve Okeno <steveokeno@gmail.com>

Request for Past Motor Claims Research Data
5 messages

Steve Okeno <steveokeno@gmail.com> Fri, May 14, 2021 at 5:26 PM
To: rkuloba@ira.go.ke

Dear Sir,

We spoke and thank you for your continued assistance,
Attached is a concept note that is in slide form and an excel template that indicates the features I require in the research.

Many thanks.
--
Stephen Katiechi Okeno,

Software Developer,(Mobile and Web Apps),
Finance & Insurance

(+254)726103730

https://sokeno.github.io/

2 attachments

Research Concept-Stephen Katiechi Okeno.pdf
293K

Sample dataset template.xlsx
11K

Robert Kuloba <rkuloba@ira.go.ke> Wed, May 26, 2021 at 12:23 PM
To: Steve Okeno <steveokeno@gmail.com>

Steve

We don't have such information at IRA as it is with individual insurance companies. This may require you visit individual insurance companies for the granular data. You are aware this may require accessing personal policyholder information.

Please do a formal letter to Commissioner of Insurance

[cid:image002.jpg@01D5CB8B.5E2AE5E0]
[Quoted text hidden]

Robert Kuloba | Chief Manager, Policy Research & Development
Insurance Regulatory Authority
Zep-Re Place Longonot Road - Upper Hill
P.O Box 43505 - 00100 Nairobi
Tel: (+254) 20 4996000, Mobile: (+254) [illegible]
Email: rkuloba@ira.go.ke |
Website: www.ira.go.ke

image002.jpg
20K

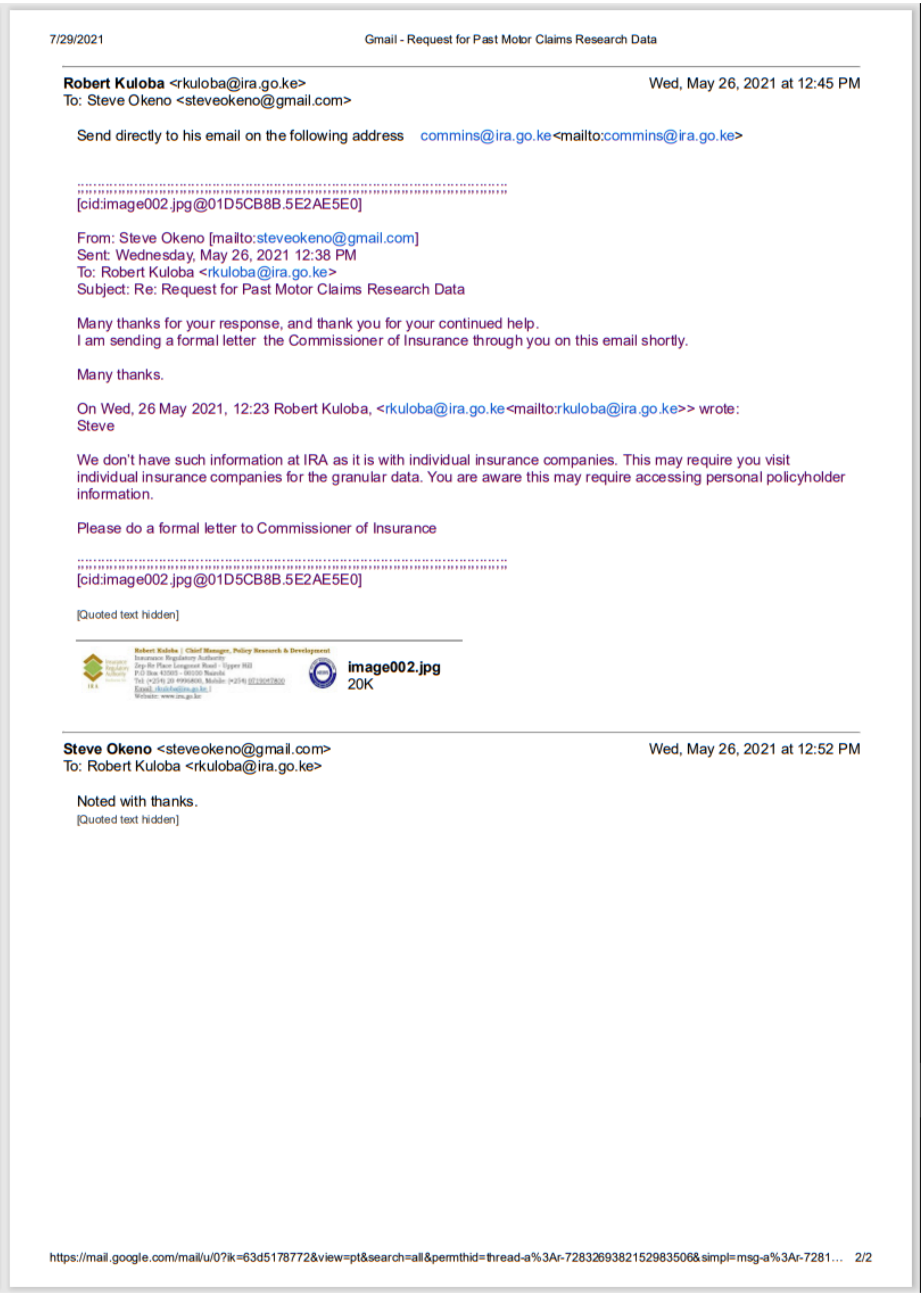
Steve Okeno <steveokeno@gmail.com> Wed, May 26, 2021 at 12:38 PM
To: Robert Kuloba <rkuloba@ira.go.ke>

Many thanks for your response, and thank you for your continued help.
I am sending a formal letter the Commissioner of Insurance through you on this email shortly.

Many thanks.
[Quoted text hidden]

https://mail.google.com/mail/u/0?ik=63d5178772&view=pt&search=all&permthid=thread-a%3Ar-7283269382152983506&simpl=msg-a%3Ar-7281… 1/2

Figure 20 Email to Head of Innovations IRA



7/29/2021 Gmail - Request for Past Motor Claims Research Data

Robert Kuloba <rkuloba@ira.go.ke> Wed, May 26, 2021 at 12:45 PM
To: Steve Okeno <steveokeno@gmail.com>

Send directly to his email on the following address commins@ira.go.ke<mailto:commins@ira.go.ke>

[cid:image002.jpg@01D5CB8B.5E2AE5E0]

From: Steve Okeno [mailto:steveokeno@gmail.com]
Sent: Wednesday, May 26, 2021 12:38 PM
To: Robert Kuloba <rkuloba@ira.go.ke>
Subject: Re: Request for Past Motor Claims Research Data

Many thanks for your response, and thank you for your continued help.
I am sending a formal letter the Commissioner of Insurance through you on this email shortly.

Many thanks.

On Wed, 26 May 2021, 12:23 Robert Kuloba, <rkuloba@ira.go.ke<mailto:rkuloba@ira.go.ke>> wrote:
Steve

We don't have such information at IRA as it is with individual insurance companies. This may require you visit individual insurance companies for the granular data. You are aware this may require accessing personal policyholder information.

Please do a formal letter to Commissioner of Insurance

[cid:image002.jpg@01D5CB8B.5E2AE5E0]

[Quoted text hidden]

Robert Kuloba | Chief Manager, Policy Research & Development
Insurance Regulatory Authority
Zep-Re Place Longonot Road - Upper Hill
P.O Box 43505 - 00100 Nairobi
Tel: (+254) 20 4996000, Mobile: (+254) 0719047800
Email: rkuloba@ira.go.ke |
Website: www.ira.go.ke

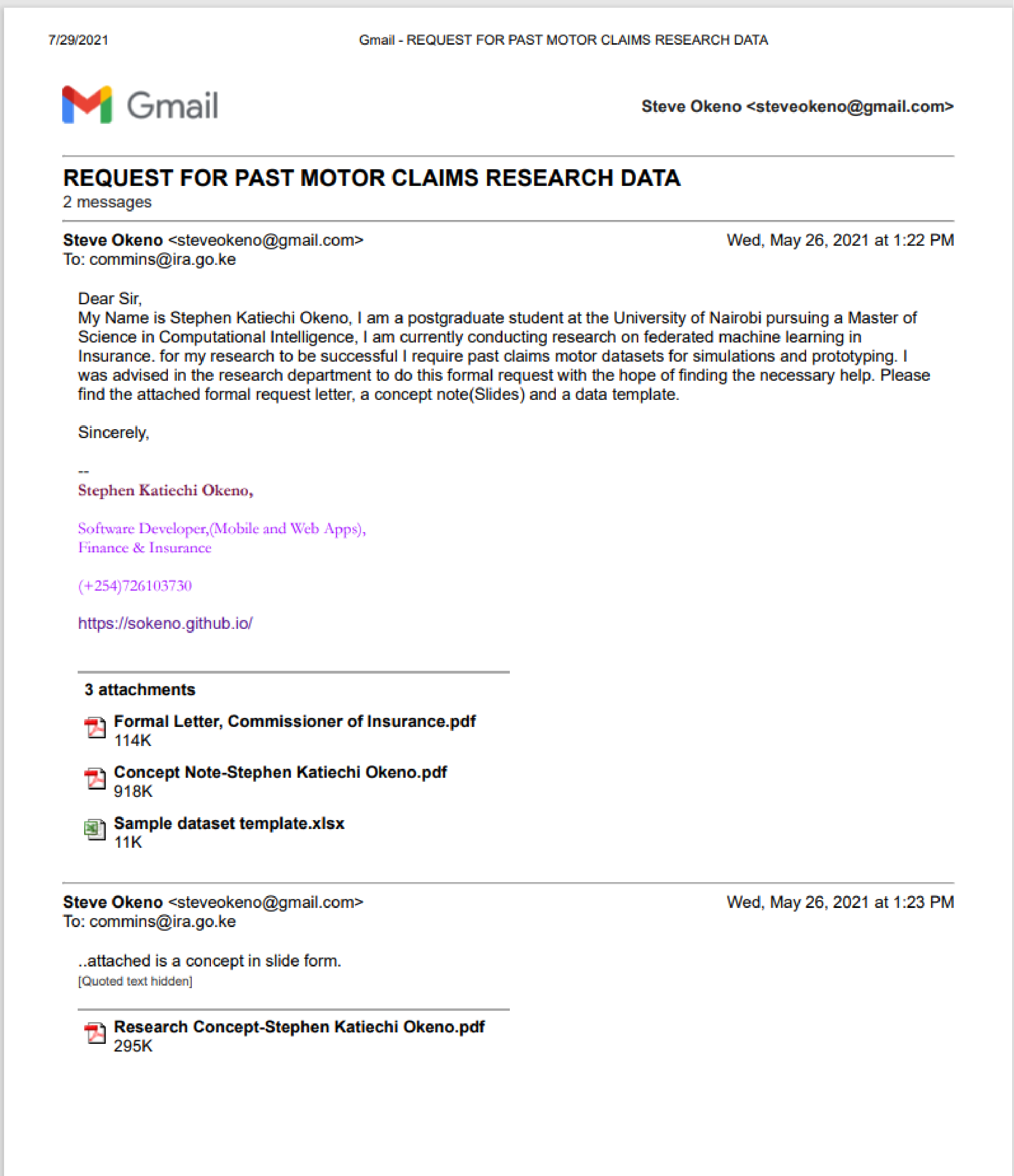
image002.jpg
20K

Steve Okeno <steveokeno@gmail.com> Wed, May 26, 2021 at 12:52 PM
To: Robert Kuloba <rkuloba@ira.go.ke>

Noted with thanks.
[Quoted text hidden]

https://mail.google.com/mail/u/0?ik=63d5178772&view=pt&search=all&permthid=thread-a%3Ar-7283269382152983506&simpl=msg-a%3Ar-7281… 2/2

Figure 21 Email to head of Research and Innovation IRA



7/29/2021 Gmail - REQUEST FOR PAST MOTOR CLAIMS RESEARCH DATA

Steve Okeno <steveokeno@gmail.com>

REQUEST FOR PAST MOTOR CLAIMS RESEARCH DATA

2 messages

Steve Okeno <steveokeno@gmail.com> Wed, May 26, 2021 at 1:22 PM
To: commins@ira.go.ke

Dear Sir,
My Name is Stephen Katiechi Okeno, I am a postgraduate student at the University of Nairobi pursuing a Master of Science in Computational Intelligence, I am currently conducting research on federated machine learning in Insurance. for my research to be successful I require past claims motor datasets for simulations and prototyping. I was advised in the research department to do this formal request with the hope of finding the necessary help. Please find the attached formal request letter, a concept note(Slides) and a data template.

Sincerely,

--
Stephen Katiechi Okeno,

Software Developer,(Mobile and Web Apps),
Finance & Insurance

(+254)726103730

https://sokeno.github.io/

3 attachments

Formal Letter, Commissioner of Insurance.pdf
114K

Concept Note-Stephen Katiechi Okeno.pdf
918K

Sample dataset template.xlsx
11K

Steve Okeno <steveokeno@gmail.com> Wed, May 26, 2021 at 1:23 PM
To: commins@ira.go.ke

..attached is a concept in slide form.
[Quoted text hidden]

Research Concept-Stephen Katiechi Okeno.pdf
295K

Figure 22 Email to the C.E.O and Commissioner for Insurance IRA