

# **Using a Classification Algorithm to Detect Fraud in Healthcare Insurance Claims**

**John Kibunyi Mbugua**

**121458**

**ICS**

**Supervisor Name**

**Allan Vikiru**

**A Project Submitted to the School of Computing and Engineering Sciences in partial fulfillment of the requirements for the award of a Degree in Informatics and Computer Science.**

**April 2025**

**School of Computing and Engineering Sciences**

**Strathmore University**

**Nairobi, Kenya**

### Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the project proposal contains no material previously published or written by another person except where a reference is made in the project proposal itself.

© No part of this project proposal may be reproduced without the permission of the author and Strathmore University

John Kibunyi Mbugua

16/04/2025

..........

### Approval

The project proposal of Adm 121458 was reviewed and approved (for examination) by the following: Allan Vikiru.

..........

## **Abstract**

For insurance firms, fraudulent insurance claims present a serious problem because they can result in considerable monetary losses and erode client confidence. Insurance inflation is a major impact of fraud as fraud has cost insurance companies hundreds of millions of dollars over the years. The study sought to find patterns and signs of fraud by examining historical data, including claim specifics, policy specifics, and client profiles done by machine learning techniques. The trained model was integrated with a web-based application claims management system to quickly identify probable fraud and launch an investigation. In addition to addressing legal requirements, the initiative also addressed moral issues including justice and privacy protection. The research employed extreme programming methodology as it deemed flexible and allowed for changes across its phases. The study used four machine learning classifier algorithms namely Logistic Regression, Support Vector Machine, AdaBoost Classifier, XGBoost Classifier. AdaBoost performed better than the rest with an accuracy of 90%, thus was selected for classification. The rest of the models performed as follows; Logistic Regression 62%, Support Vector Machine 73%, XGBoost 89%. The project's output was a reliable and effective fraud detection system that would help insurance firms reduce financial risks, safeguard legitimate clients, and uphold a relationship of trust with policyholders.

## Table of Contents

Declaration .....	ii
Abstract .....	iii
Table of Contents .....	iv
List of Figures .....	viii
List of Tables .....	ix
Chapter 1: Introduction.....	1
1.1 Background .....	1
1.2 Problem Statement .....	2
1.3 Objectives .....	2
1.3.1 General Objective .....	2
1.3.2 Specific Objectives .....	2
1.4 Research Questions .....	2
1.5 Justification .....	3
1.6 Scope and Limitations.....	3
Chapter 2: Literature Review.....	4
2.1 Introduction.....	4
2.2 History of Healthcare Insurance .....	4
2.3 Health Care Insurance Categories.....	4
2.3.1 Public Health Care Insurance.....	4
2.3.2 Private Health Care Insurance .....	5
2.4 How Health Care Insurance Works .....	5
2.4.1 Health Care Insurance Companies.....	5
2.4.2 Health Care Insurance Subscriber.....	5
2.4.3 Health Care Service Provider.....	6
2.4.4 Health Clearing Houses .....	6
2.5 Health Insurance Fraud .....	6

2.5.1	Health care Service Providers Fraud.....	6
2.5.2	Health Subscriber Fraud .....	7
2.5.3	Impact of Health care Fraud .....	7
2.6	Machine Learning algorithms for Classification .....	7
2.6.1	Logistic Regression.....	7
2.6.2	Support Vector Machine .....	8
2.6.3	Adaptive Boosting (AdaBoost) Classifier .....	9
2.6.4	Extreme Gradient Boosting (XGBoost) Classifier .....	9
2.7	Related Works.....	10
2.7.1	A Predictive Modelling for Detecting Fraudulent Automobile Insurance Claims 10	
2.7.2	Health Claim Insurance Prediction using Support Vector Machine with Particle Swarm Optimization .....	10
2.7.3	Fraud Detection: A Study of AdaBoost Classifier and K-Means Clustering ..	10
2.8	Research Gaps.....	11
2.9	Conceptual Framework .....	11
Chapter 3:	Methodology .....	12
3.1	Introduction.....	12
3.2	Applied Development Approach .....	12
3.2.1	Planning .....	13
3.2.2	Design .....	13
3.2.3	Coding.....	13
3.2.4	Testing.....	13
3.2.5	Listening .....	13
3.3	Justification of the methodology.....	13
3.4	System Analysis.....	14
3.4.1	MySQL .....	14
3.4.2	Python .....	14

3.4.3	Google Collaboratory.....	14
3.4.4	Flask Web Application Framework.....	14
3.4.5	Use Case Diagram.....	14
3.4.6	System Sequence Diagram .....	15
3.4.7	Class Diagram.....	15
3.4.8	Entity Relationship Diagram.....	15
3.4.9	Database Schema .....	15
3.5	System Architecture.....	15
3.6	System Deliverables.....	15
Chapter 4:	System Analysis and Design.....	16
4.1	Introduction.....	16
4.2	System Requirements.....	16
4.2.1	Functional Requirements .....	16
4.2.2	Non-functional Requirements.....	16
4.3	System Analysis Diagrams .....	16
4.3.1	Use Case Diagram.....	17
4.3.2	Sequence diagram .....	17
4.4	System Design Diagrams.....	19
4.4.1	Entity Relationship Diagram.....	19
4.4.2	Database Schema .....	19
4.4.3	Class Diagram.....	20
Chapter 5:	System Implementation and Testing.....	22
5.1	Introduction.....	22
5.2	Description of Implementation Environment .....	22
5.2.1	Hardware Specifications .....	22
5.2.2	Software Specifications .....	22
5.3	Description of Dataset.....	22

5.4	Description of Testing.....	23
5.4.1	Data Preparation.....	23
5.4.2	Model Development.....	26
5.4.3	Claims Classification System .....	28
5.4.4	Testing.....	29
5.5	Testing Results.....	30
Chapter 6:	Conclusions, Recommendations and Future Works .....	34
6.1	Conclusions.....	34
6.2	Recommendations.....	34
6.3	Future Works .....	34
Appendix	.....	38
Appendix A:	Gantt Chart .....	38

## List of Figures

Figure 2.1: How Health Care insurance works .....	5
Figure 2.2: Sigmoid Curve.....	8
Figure 2.3: Support Vector Machine .....	9
Figure 2.4: Conceptual Framework .....	11
Figure 3.1: Extreme Programming Life Cycle .....	12
Figure 4.1: Use Case Diagram .....	17
Figure 4.2: System Sequence Diagram .....	18
Figure 4.3: Entity Relationship Diagram .....	19
Figure 4.4: Class Diagram .....	21
Figure 5.1: Dataset Description .....	23
Figure 5.2: Number of Null Records .....	23
Figure 5.3: Check Duplicate Records .....	24
Figure 5.4: Data Clean-up.....	24
Figure 5.5: Data Transformation.....	25
Figure 5.6:Pearson's Correlation Heatmap .....	25
Figure 5.7: Selected Columns .....	26
Figure 5.8: Splitting Dataset .....	26
Figure 5.9: Classifier Algorithms Performance .....	27
Figure 5.10: Web-Based Application .....	28
Figure 5.11: Claim Classification .....	29
Figure 5.12: Login Module .....	31
Figure 5.13: Dashboard.....	31
Figure 5.14: Medical Claim Form .....	32
Figure 5.15: Claims Classification Form .....	33



## **List of Tables**

Table 4.1: Users Table Database Schema.....	20
Table 4.2: Medical Claims Table Database Schema .....	20

## **Chapter 1: Introduction**

### **1.1 Background**

Over the past few decades, information and communication technology (ICT) has consistently established itself as the system architect by tying together markets, businesses, governments, and people. In addition to boosting job administration and servicing, this connection has shortened distances and given the world a multifaceted aspect, allowing for real-time combat against fraud and ethics. The 21st-century ICT revolution not only shapes business trends and organics, but it also predicts and defines social interaction, culture, and conduct. To enhance their ability to interpret information in this conflict and give them the edge in recognising, managing, and reporting fraud-related incidents, organisations such as insurance companies have made large investments in ICT (Njeru, 2022).

Healthcare fraud is a white-collar crime that happens when medical claims are fraudulently filed to make money. Globally, healthcare fraud and corruption have cost numerous organisations a significant amount of money. Healthcare spending is rising quickly each year in several nations. Approximately 10% of the world's gross domestic product goes towards healthcare (Onyango, 2022). False claims cost enormous sums of money every year. In 2012, 143 cases of healthcare insurance fraud were reported, resulting in a loss of Ksh 253.6 million. Only Ksh 5.2 million of that was recovered. The claims loss ratio for Kenya's healthcare sector is 5% greater than that of other insurance types (Mambo, 2019). The National Health-Care Anti-Fraud Association defines healthcare insurance fraud as the submission of false claims to health insurance programs to benefit oneself. Additionally, providers may submit claims for prescriptions that are partially or never filled. To increase profits, certain medical professionals may coerce patients into taking unnecessary prescription drugs. To identify areas that require extra attention, such as inaccurate and inadequate data input, matching claims, and medically non-covered services, methods of processing electronic claims have recently been progressively modelled. In the healthcare insurance industry, fraud detection has shifted from traditional domain expert analysis to rule-based systems (Gupta, 2021). Even though these systems are used to detect fraud, they often have limited discovery capabilities because most of the discovery is dependent on pre-established standards that are described by fraud field specialists. Hence there is an underlying need for more efficiency in detecting insurance fraud in healthcare insurance claims (Matloob, 2019).

Machine Learning (ML) approaches have the potential to effectively detect fraudulent healthcare insurance claims due to their improved performance and prediction accuracy. Machine learning classification is the process of classifying input samples from a problem area using machine learning algorithms (Burri, 2019). Some of the machine learning classification algorithms include Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), Support Vector Machine, Random Forest, Logistic Regression, Decision Tree, Naïve Bayes classifier (Njeru, 2022).

## **1.2 Problem Statement**

Fraud is any act aimed at defrauding another party financially (Gedela, 2022). The insurance industry in Kenya is well-established, (Insurers, 2020), leading the sub-Saharan African market, and expanding at a rapid pace (Organization, 2018). The insurance sector is reluctant to change, particularly when it comes to utilising new technology to address the concerning problem of fraud (Authority, 2020). The problem of healthcare insurance claims fraud is serious and causes insurance firms to suffer large financial losses. Conventional fraud detection techniques, which frequently include manual investigations, are laborious and prone to human mistake. There is a shortage of information on fraud in Kenya. This is caused by several things, such as the absence of a centralized mechanism for reporting fraud, the unwillingness of victims to disclose fraud, and the complexity of conducting fraud investigations.

## **1.3 Objectives**

### **1.3.1 General Objective**

To develop a predictive model using classification algorithm to detect fraud in healthcare insurance claims made to insurance companies.

### **1.3.2 Specific Objectives**

- i. To research on the current forms in which healthcare insurance claims fraud manifests.
- ii. To analyse existing methods of insurance fraud detection in healthcare.
- iii. To develop a machine learning model using classification algorithm.
- iv. To evaluate the developed machine learning model.

## **1.4 Research Questions**

- i. What are the various forms in which healthcare insurance claims fraud manifests?
- ii. What are the existing methods of insurance fraud detection in healthcare insurance claims?

- iii. How will a classification algorithm be developed to detect fraud in healthcare insurance claims?
- iv. How will the developed machine learning model be tested in detecting fraud in healthcare insurance claims?

### **1.5 Justification**

This study aimed to evaluate the most effective approaches that have been previously employed by examining the various approaches that have been applied to identical challenges. Searching, looking into these approaches, and attempting to improve and develop a predictive model that could identify and highlight the dubious claims based on the investigation, testing, and comparison of these models to develop a sufficiently straightforward, fast, and precise model that can identify the dubious claims without putting undue strain on the system it operates on. The study provided value to the healthcare fraud area by providing insights and recommendations on how to design a fraud detection model based on a classification algorithm. This will result in early detection of fraudulent billings in healthcare insurance firms. The model will allow healthcare insurers to detect fraudulent individuals and take legal action against them.

### **1.6 Scope and Limitations**

Real world data may be limiting therefore open-source health care dataset was used for training and testing. This is because of the data privacy concerns in financial institutions sharing confidential information of their clients. There was also a time constraint limiting full implementation of all the classification algorithms, thus only a few were used.

## **Chapter 2: Literature Review**

### **2.1 Introduction**

This chapter reviewed the history of healthcare insurance, the different insurance categories, how health care insurance works as well as how fraud manifests in healthcare insurance claims. It also investigated on existing studies in the context of fraud detection using classification algorithm. Finally, a conceptual framework to depict the overall functionality of the proposed solution.

### **2.2 History of Healthcare Insurance**

In Germany, the year 1883 was a turning point for health insurance. Industrial employers were required to cover their employees' illness and injury insurance once a law requiring sickness insurance was established. Through their pay, employees made contributions to a sickness fund, and employers made an equivalent contribution (Buttice, 2019).

The modern hospital was first established as a facility for medical care, education, and advancement in the middle of the 1920s. In response to these advancements, the profession developed policies and expanded medical specialisation, training, and the division of labour. Physicians raised their fees, and they now base their choices on factors including the time it takes to acquire new skills, the expense of education, specialisations, administrative fees, and heightened competition (Starr, 2018).

Following World War I, there was a huge demand for labour in the United States. Companies increased their compensation offered to recruit workers to compete effectively. After regulating the pay range, the government enacted legislation enabling businesses to provide health insurance. Since the government was involved, health insurance was offered as part of a corporate package, and participants paid into a medical plan (Onyango, 2022).

### **2.3 Health Care Insurance Categories**

There are two main categories of health insurance. The government pays for public health insurance, which is a subsidised insurance plan that occasionally requires subscribers to pay as well. The subscriber pays the whole premium for private health insurance (Onyango, 2022).

#### **2.3.1 Public Health Care Insurance**

In many parts of the world, this type of government-managed health insurance is most usually referred to as national health insurance. The insurance covers all or a portion of the subscribers' medical expenses. Primary healthcare is primarily handled by it. The country and the program

or scheme determine how the insurance is funded. Some are entirely funded by taxes paid by the government, while others require contributions from the public to be eligible for benefits.

### 2.3.2 Private Health Care Insurance

It is operated by private insurance companies. After selling their goods, the insurance company charges a subscriber. Most businesses pay a set sum to the insurance company, and employees pay a portion that is subsidised. Indirect payments from the government can also be made by subsidising the insurance companies' taxes. A subscriber to private medical insurance can receive a variety of benefits, including coverage for long-term illnesses, mental health, dental, and optical care, depending on the plan they select.

## 2.4 How Health Care Insurance Works

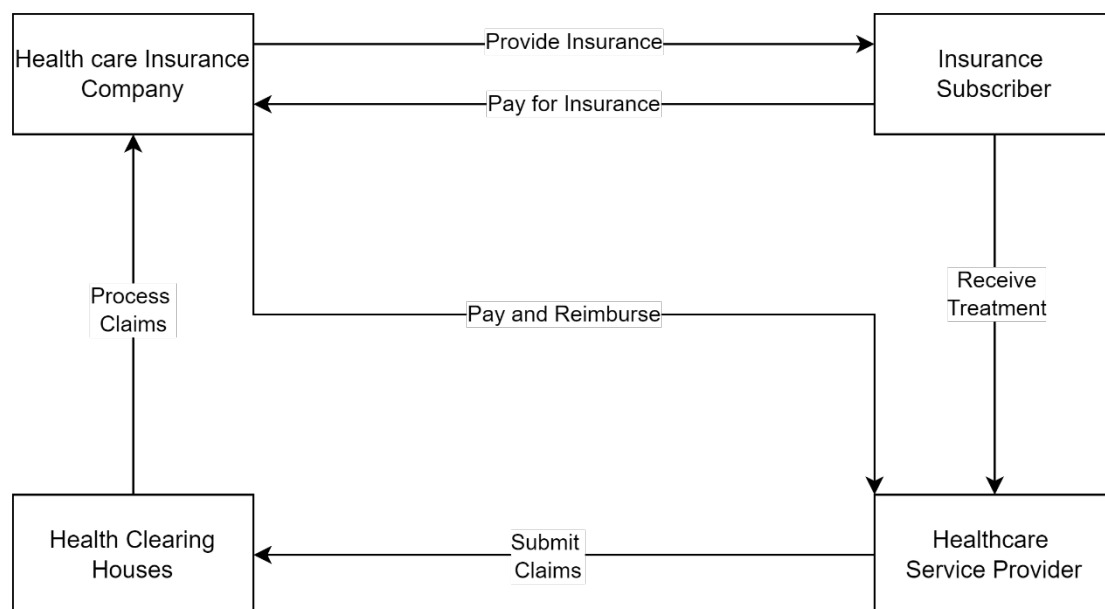


Figure 2.1: How Health Care insurance works

### 2.4.1 Health Care Insurance Companies

An alternative name for them is insurance carriers. These are the companies that offer insurance to consumers. They are compensated by insurance premiums or subscriptions. After their patients are treated, they also reimburse the healthcare providers for their services (Onyango, 2022).

### 2.4.2 Health Care Insurance Subscriber

The individual who pays insurance premiums is known as the subscriber. This may be based on employment; in which case the employer covers the premiums; it may also involve individuals who choose to pay for their own premiums; or it may involve contributions from

both the employer and the employee. Enrolling dependents who might also utilise the insurance is an advantage for subscribers. Subscribers can receive varying degrees of treatment, from basic medical care to full health coverage, depending on their donations (Onyango, 2022).

### **2.4.3 Health Care Service Provider**

A healthcare service provider is an organisation, such as a medical facility or a private practitioner, that is authorised by current legislation to provide healthcare services. Hospitals, physicians, chemists, labs, and ambulance services are among them. They provide the subscriber with services. After that, they submit a claim to the insurance provider, and if it is accepted, they are paid for the services (Onyango, 2022).

### **2.4.4 Health Clearing Houses**

These are the businesses or centres that act as go-betweens for insurance companies and health service providers. Their primary responsibility is to clear the claims' data. A claim is a request for reimbursement for benefits obtained following the provision of a service. To process the claims, they scrub, review, and edit them to remove inaccuracies. This facilitates a quicker payment process (Onyango, 2022).

## **2.5 Health Insurance Fraud**

In the healthcare insurance sector, fraud can take many different forms. Health insurance fraud, waste, and abuse are all included in this. Fraud involving health insurance happens when a subscriber or provider purposefully gives false information to make money. Healthcare abuse happens when necessary best practices are not followed, leading to unnecessary, costly treatment, while healthcare waste happens when health services are used carelessly and recklessly.

### **2.5.1 Health care Service Providers Fraud**

This happens when the service provider takes advantage of the system to make money. Fraud by healthcare service providers can take many various forms, including as impersonation, upcoding, unbundling, unjustified procedures, and manipulating medical records.

When a healthcare professional costs for services that have not been provided, this is known as upcoding. The goal is to inflate a patient's total debt, which is then passed on to the insurance company. The billing of many procedures for a collection of treatments that are billed as a single item is known as unbundling. The service should be invoiced as a single item, but the

provider will make claims for each stage of therapy. The bill amount rises when the payments are broken down.

Manipulating patient medical records to support unnecessary treatments, costly medications, and unwarranted surgeries. This may lead to an incorrect diagnosis and subsequent treatment for the patient.

### **2.5.2 Health Subscriber Fraud**

Some of the ways in which a health subscriber may participate in fraud include falsifying information, impersonation and conspiracy fraud. Intentional collusion between a subscriber and a provider to make a claim for services that were never provided is known as conspiracy fraud. Impersonation happens when someone who is not listed as a dependent or who does not have insurance uses an insured person's information to obtain health care services. This is also termed as identity theft.

### **2.5.3 Impact of Health care Fraud**

Insurance inflation is a major impact of fraud as fraud has cost insurance companies hundreds of millions of dollars over the years (Luther, 2020). Someone must bear the burden of these losses. The cost of government-based insurance is covered by the taxpayers. It's possible that other areas of health lack adequate funding. This causes public hospitals to overwork themselves and lowers the standard of care they give.

Identity theft has targeted health care insurance firms. Medical records contain private information that could be misused if it ends up in the wrong hands.

## **2.6 Machine Learning algorithms for Classification**

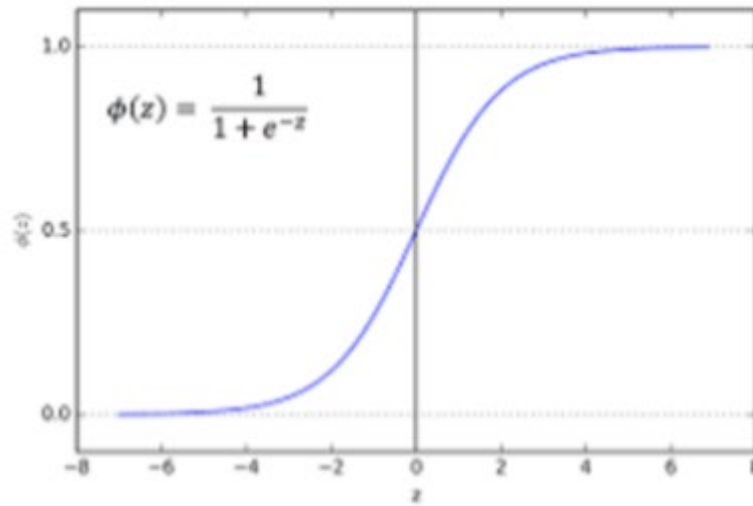
### **2.6.1 Logistic Regression**

It is an algorithm for classification. The relationship between one or more independent variables and other independent variables is modelled. Log odds and linear properties in the independent variables, with little to no multicollinearity among them, form the basis of the procedure. Moreover, independence of observations and errors. The sigmoid function is used in logistic regression to translate predictions into probabilities. The sigmoid function is as represented below:

$$S(x) = \frac{1}{1 + e^{-x}}$$



Where  $S(x)$  is the sigmoid function,  $e$  is Euler's number. The sigmoid curve gives the probability of a class or target prediction which lies in the interval between 0 and 1. The sigmoid curve is as represented below



*Figure 2.2: Sigmoid Curve*

The advantage of logistic regression is that it fails to presume anything about the target variable or class distribution. Additionally, it can be applied to multivariate logistic regression, which makes predictions for several classes. In low-dimensional datasets, the model is less likely to overfit, making predictions more accurate. The major drawback for the algorithm is the assumption of linearity among dependent and independent variables (Onyango, 2022).

### **2.6.2 Support Vector Machine**

Support Vector Machine is a supervised machine learning algorithm used for classification problems as well as regression problems. Hyperplanes are used to classify high-dimensional data. A subspace with one fewer dimension than the feature space is called a hyperplane. It is expected that the input data may be linearly separated in a geometric space, with the value of each feature being the value of the corresponding coordinate, when data points are plotted in an  $n$ -dimensional space (Njeru, 2022). The diagram below shows a support vector machine classifier.

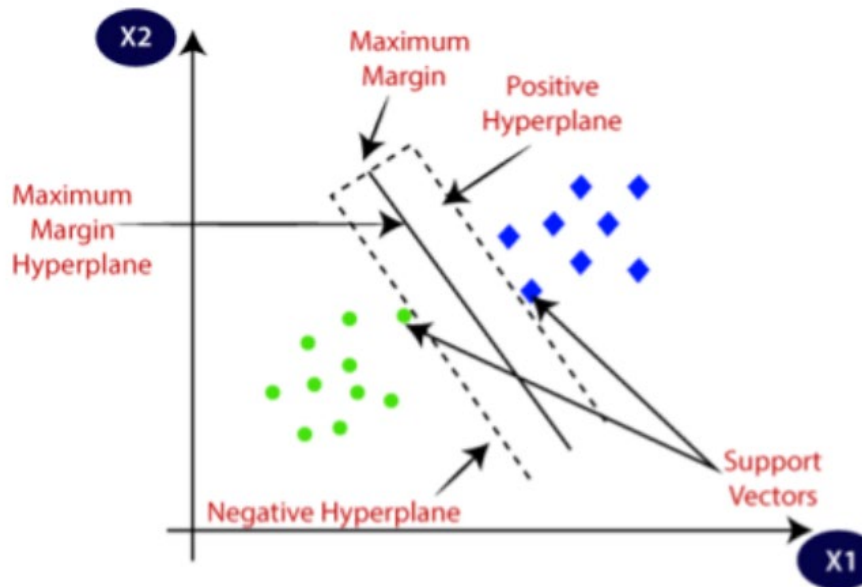


Figure 2.3: Support Vector Machine

### 2.6.3 Adaptive Boosting (AdaBoost) Classifier

This ensemble learning classifier enhances the performance of other classifiers for binary classification when used in combination with them. By passing a coefficient, a weak classifier with high accuracy might be given more weight, reducing training error. Three key factors in the method are the learning rate, base estimator, and number of estimators. The number of estimators there are indicates how many weak learners need to be trained. To train the model, a weak learner called the base estimator is employed. The method favours data samples that are erroneously classified by changing the weights of weak learners according to their learning rate (Njeru, 2022).

### 2.6.4 Extreme Gradient Boosting (XGBoost) Classifier

XGBoost is an ensemble learning technique that uses Gradient Boosted decision trees to increase model speed and performance. It generates forecasts by merging the findings of past learners. To generate decision trees sequentially, all independent variables are weighted before being fed into the decision tree to predict results. The variables are then fed into the second decision tree, which assigns a higher weight to variables predicted incorrectly by the previous tree (Dimitrakopoulos, 2018).

## **2.7 Related Works**

### **2.7.1 A Predictive Modelling for Detecting Fraudulent Automobile Insurance Claims**

Using a logistic regression technique, variables were evaluated and chosen based on a variable relevance ranking process that involved 20 10-fold cross-validation trials. To create the logistic regression model, the learning set was divided into 10 segments, nine of which were utilised as training data. The model was validated using the remaining segment. Each segment was used as the validation set, and the procedure was carried out ten times. To generate 200 logistic regression models, the 10-fold cross validation was carried out 20 times. Stepwise selection was employed in each model to choose its variables. The logistic regression model's probability response was predicated on a 0.5 threshold. Accordingly, a claim was deemed fraudulent if the projected likelihood was more than or equal to 0.5. The accuracy of the model was 87.1% (Moon, 2019).

### **2.7.2 Health Claim Insurance Prediction using Support Vector Machine with Particle Swarm Optimization**

Particle Swarm Optimisation (PSO) was used to search for the parameters of the classification model that was created in this study using Support Vector Machines (SVM). The SVM particles with PSO were identified as candidates based on the SVM's ideal parameters. These particles were produced at random. To find the global optimum location, the study exploited and investigated the search space. The method also uses the fitness function, which was obtained by calculating the F1 score of training SVM on data training. The study's findings demonstrated that SVM outperforms standard SVM when used with PSO. Additionally, it demonstrates that the suggested particle counts of 5, 10, 20, and 50 do not result in any different performance (Syaiful Anam, June 2023).

### **2.7.3 Fraud Detection: A Study of AdaBoost Classifier and K-Means Clustering**

Using data from a big data mining and fraud detection study collaboration between Worldline and the Machine Learning group of the Universite Libre de Bruxelles, Mishra used the algorithm to identify fraud. Random Forest served as the base estimator. Each observation had a distinct number of estimators. For different numbers of estimators with the same learning rate, the model yielded different levels of accuracy. With a learning rate of 0.01 and 200 estimators, the model's accuracy was 92.48%. With 100 estimators and the same learning rate (0.01), the accuracy of the model was 93.09% (Mishra, 2021).

## 2.8 Research Gaps

From the studies discussed in the previous section, adequate comparison analysis of the machine learning algorithms has not been done. Different researchers choose different algorithms, necessitating a comparison of algorithm performance across datasets. Most insurance companies identify fraud via rule-based analysis, which is ineffective and inaccurate. The correctness of the rules is limited to the analysis's effectiveness (zhou, 2020). A small team of auditors manually examines and flags any questionable health insurance claims (Waghade, 2018). The method was time-consuming. Rule-based analysis operates on predefined conditions thus the inability to adapt to new fraud schemes. Rule-based systems also handle moderate volume of claims. The performance degrades when the volume increases as each claim must be evaluated against an extensive list of rules (Agarwal, 2023).

## 2.9 Conceptual Framework

A Web-application was the front end. This was where the claim insurer submitted registered claims for classification as well as get the output report of the categorised claims. The web-application was integrated with a machine learning model which was used to classify the claims, and a database which was used to store the reports.

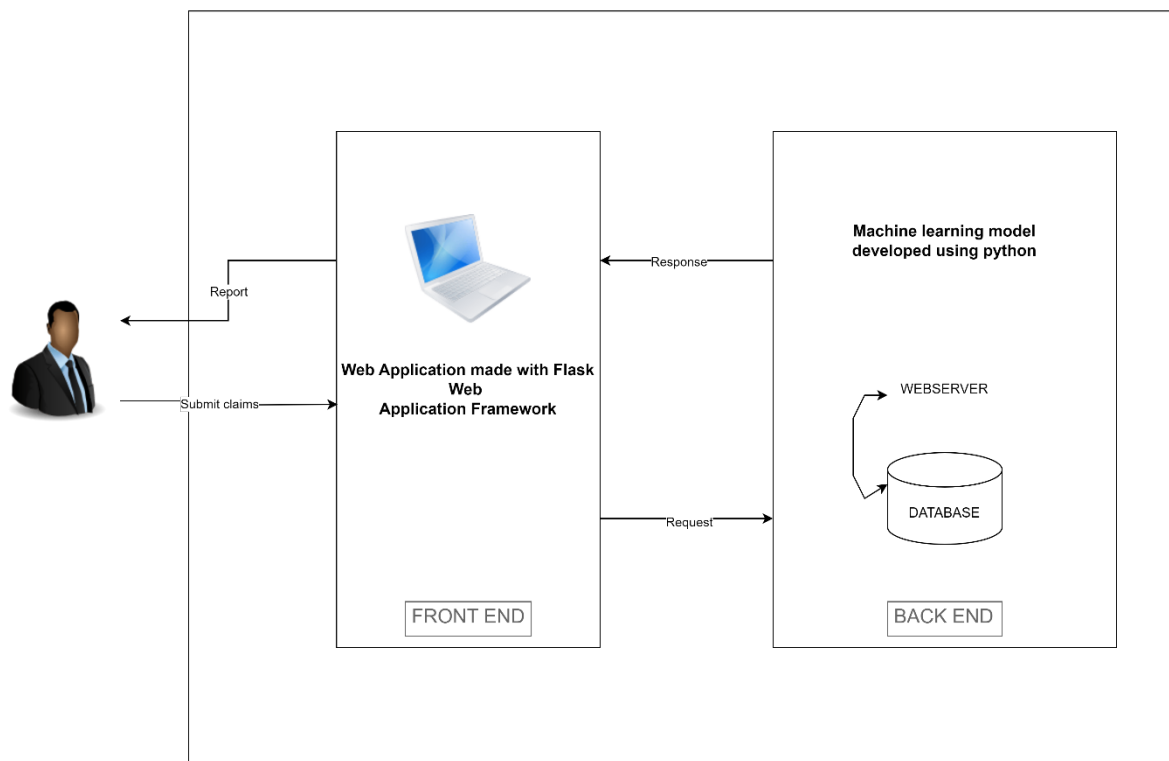


Figure 2.4: Conceptual Framework

## Chapter 3: Methodology

### 3.1 Introduction

The term methodology refers to the systematic, theoretical analysis of the methods that are applied to a given study. The Object-Oriented Analysis and Design Methodology was the preferred approach methodology applied. This was because its techniques are a branch of the software development business that aimed to offer an iterative approach to analysing, designing and implementing complex systems. The methodology processes include planning, analyzing, designing, implementing, and supporting the system (Hammad, 2022).

### 3.2 Applied Development Approach

The system was created utilising an agile-based software development framework called Extreme Programming. It was a model that added flexibility to the project management procedure to overcome some of its drawbacks. It permitted phases to overlap, making it possible to work on several stages at once (Kanade, 2022). It also had a feedback mechanism which allowed revisiting of previous phases if issues were encountered during later stages, facilitating corrections and adjustments without waiting till the end of the project.

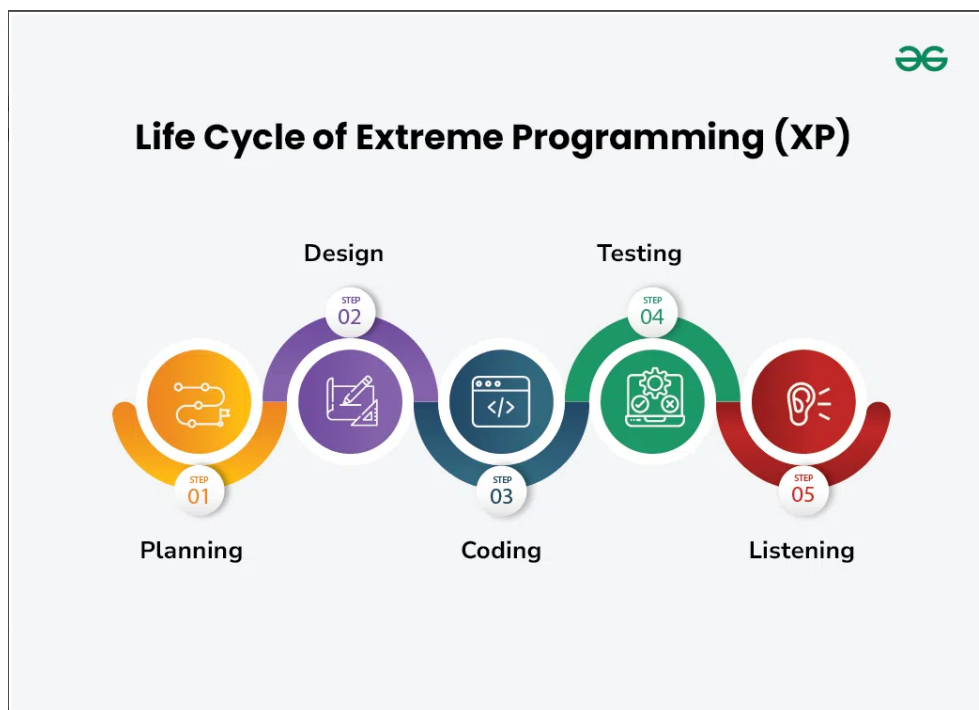


Figure 3.1: Extreme Programming Life Cycle

### **3.2.1 Planning**

This first step entailed gathering all project requirements from user stories and having a plan to guide the work. The Software Requirements were prepared to serve as the foundation for future phases. This phase ensured that all relevant data was acquired before proceeding (Raeburn, 2024).

### **3.2.2 Design**

The system design was developed in accordance with the requirements stated in the preceding phase. This comprised specifying the system architecture, components, modules, interfaces, and data structures. The design step converted the requirements into an architecture for the system (Kramer, 2018).

### **3.2.3 Coding**

In this phase we converted requirements gathering and system analysis to system development, including implementation and coding. The code developed was dependent on the previous phases of the design. It was necessary to employ tools like compilers, interpreters, and debuggers to generate code. It involved pair programming and implementing Test-Driven Development, writing tests before coding to ensure functionality (Kanade, 2022).

### **3.2.4 Testing**

Testing was a continuous activity throughout the Extreme Programming process, ensuring that both unit tests and acceptance tests were performed regularly. These unit tests were automated to check if specific features work correctly. Acceptance tests ensure that overall system met initial requirements (Pal, 2024).

### **3.2.5 Listening**

In this phase regular feedback was done to ensure that the system met the requirements and to make any changes needed in the system.

## **3.3 Justification of the methodology**

Extreme Programming was perfect for projects where specifications change over time since it can adapt to frequent changes in customer needs. Short development cycles and ongoing client feedback enable this agility, enabling teams to make changes fast without incurring large delays (Raeburn, 2024).

### **3.4 System Analysis**

This section contains a list of the tools and diagrams that will be used in this research, along with an explanation of their significance.

#### **3.4.1 MySQL**

Based on Structured Query Language, MySQL is a relational database management system created by Oracle. MySQL was used to store user data and status reports from the model.

#### **3.4.2 Python**

Python is a dynamically typed, dynamically bound, interpreted high-level programming language that has the advantage of data structures. This made it appealing for rapid application development. Python was utilized in the system to build machine learning classification models. It also had the benefit of having numerous libraries already installed, making it simpler to deploy.

#### **3.4.3 Google Collaboratory**

Google Collaboratory is an online Google Platform that enables any Python code to be written and executed and is suitable for education, data analytics, and machine learning. Google Collaboratory was crucial because it enhanced Python during the model's development, guaranteeing the system's seamless operation.

#### **3.4.4 Flask Web Application Framework**

Flask is a compact and lightweight Python web framework that facilitates the development of online applications. Flask's versatility as a framework gives developers the ability to implement the designed solution rapidly (Abdelhadi Dyoury, 2022). Flask was selected for the solution's development because it enabled the implementation of a front-end web application that gave access to the machine learning model to make classification of fraudulent and legitimate claims.

#### **3.4.5 Use Case Diagram**

This is a high-level diagram illustrating the relationship between a system and its external entities. Its purpose was to present a clear and concise overview of a system and its relationships with other systems (Rutkowska, 2023). The use case diagram was employed in this research to demonstrate how the system interacted with its participants.

### **3.4.6 System Sequence Diagram**

This was a visual representation of the flow of data through a system. It focused on the processes that collect, manipulate, store, and distribute data among multiple components and external entities (VanZandt, 2023).

### **3.4.7 Class Diagram**

A class diagram is a fundamental component of the Unified Modeling Language that visually represents the structure of a system by detailing its classes, attributes, methods, and the relationships among them. It served as a blueprint for both the design and implementation of object-oriented systems.

### **3.4.8 Entity Relationship Diagram**

An entity relationship diagram depicts how people, objects, or concepts relate to one another. This diagram used standardised symbols and syntax to depict entities, properties, and their relationships within a system. It also used grammatical structure in its notation, with an entity denoted as a noun and a connection as a verb (Whitfield, 2024).

### **3.4.9 Database Schema**

A schema diagram is an effective visual representation of a database system's structure and organisation. It served as a template for how entities, properties, and connections inside a database are related (Sandoval, 2023).

## **3.5 System Architecture**

The proposed system has a machine learning based classification model, which involved training and testing sample healthcare insurance claims data. The system used four classification algorithms, Logistic Regression, Support Vector Machine, Adaptive Boosting and Extreme Gradient Boosting, for the model's training. The model was then be integrated with a web-based application where it was used to classify fraudulent and legitimate claims.

## **3.6 System Deliverables**

The proposed system has a web application where the claims were submitted for classification as either fraudulent or legitimate claims. The web application was integrated with a machine learning model which had been trained to classify the claims. A system documentation was also provided.



## **Chapter 4: System Analysis and Design**

### **4.1 Introduction**

This chapter examines system design and architecture using diagrams. The diagrams show how the system's many components interact.

### **4.2 System Requirements**

Some of the system requirements reviewed in the project include.

#### **4.2.1 Functional Requirements**

- i. User authentication and authorization – Users logged in securely using the credentials created.
- ii. Data collection and integration – The system collected relevant claim details data and integrated it into claims database.
- iii. User Interface module – A flask web application was created using python and html. The module was integrated with MySQL database and a classification algorithm.
- iv. Classification model – A classification model was created using machine learning in google Collaboratory. The model incorporated machine learning classifiers such as Logistic Regression, Support Vector Machine, AdaBoost and XGBoost.
- v. Database Management – The study used MySQL database to store user credentials as well as claim details. The user interface model was connected to it to POST and GET claim details.

#### **4.2.2 Non-functional Requirements**

- i. System security – These are security measures to protect sensitive client claim details data from unauthorized access. This was achieved by a user authentication and authorization applying role-based access control.
- ii. Data Security – This involves protecting sensitive client information from unauthorized access. This was achieved by implementing role-based access control.

### **4.3 System Analysis Diagrams**

These diagrams effectively represent the system and user needs. The numerous diagrams will show how the system's processes and entities interact.

### 4.3.1 Use Case Diagram

A use case diagram depicts how a user and administrator interact with the system. The actors are the two primary users who interact with the system in this case: the client and the administrator. The use cases represent the system's different functions. The user is dependent on the verification of the password. This represents an include relationship. The admin manages the users as well as the claim classifier and thus the extended relationship. This is shown in figure 4.1.

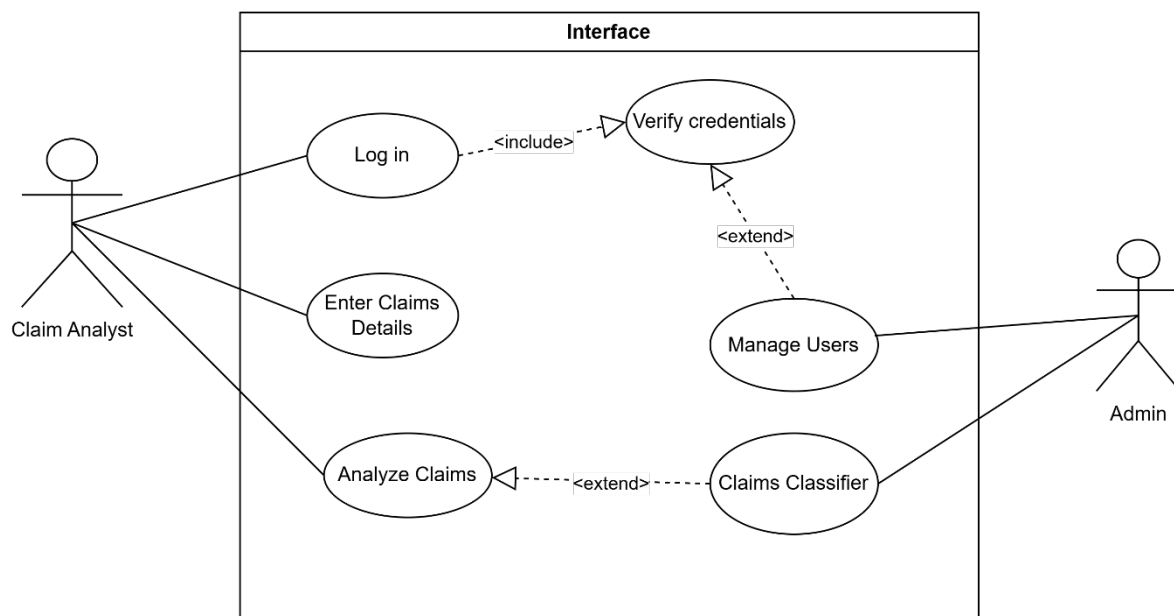


Figure 4.1: Use Case Diagram

### 4.3.2 Sequence diagram

A sequence diagram is a form of Unified Modelling Language interaction diagram that depicts how elements in a system interact with one another in a certain order to complete a task. It focuses on the sequence of message exchanges between objects or components. In this case it shows how a user login into the web application, inputs and submit the claims and lastly the claim classification. This is shown in the figure 4.2.

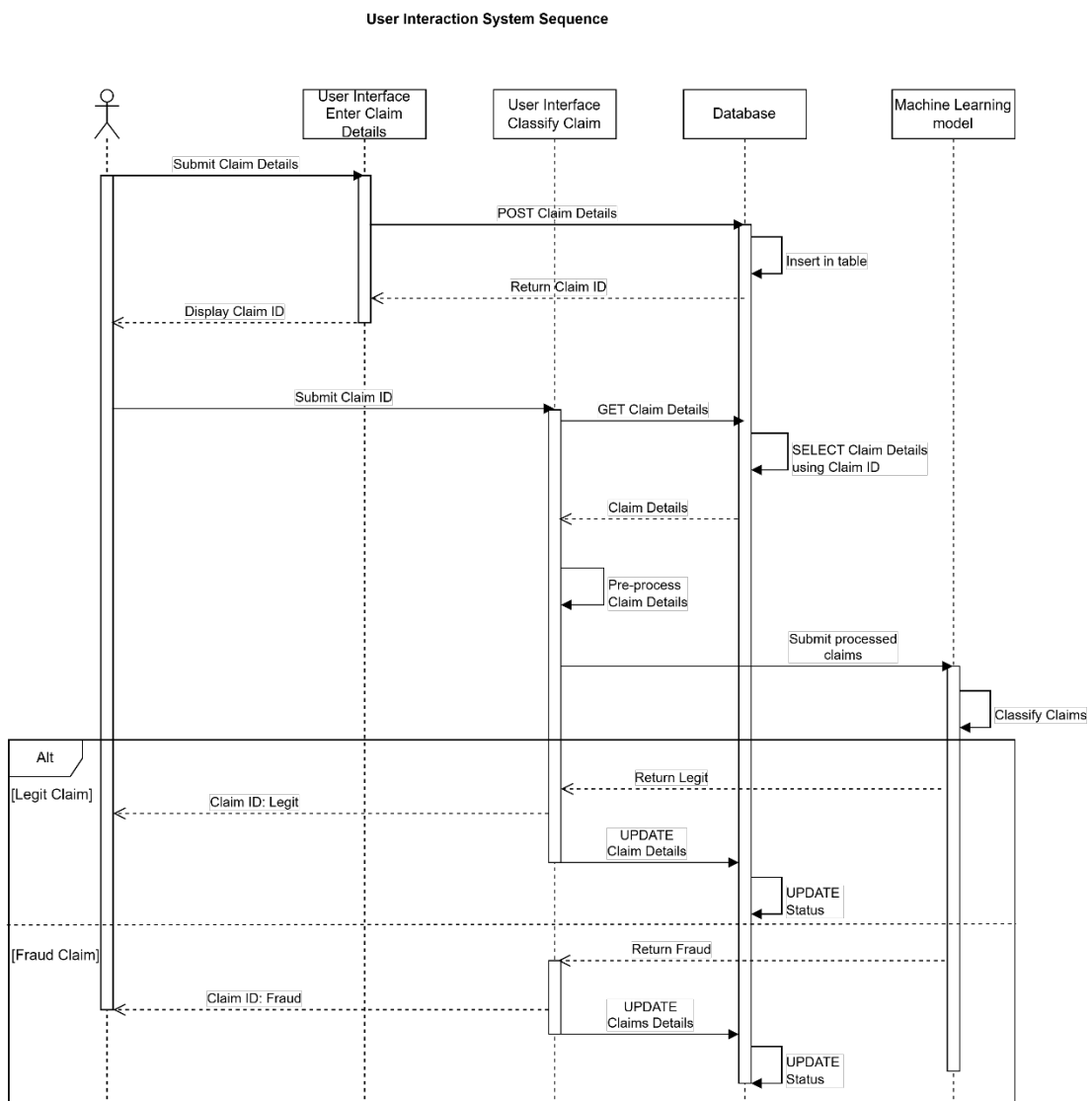
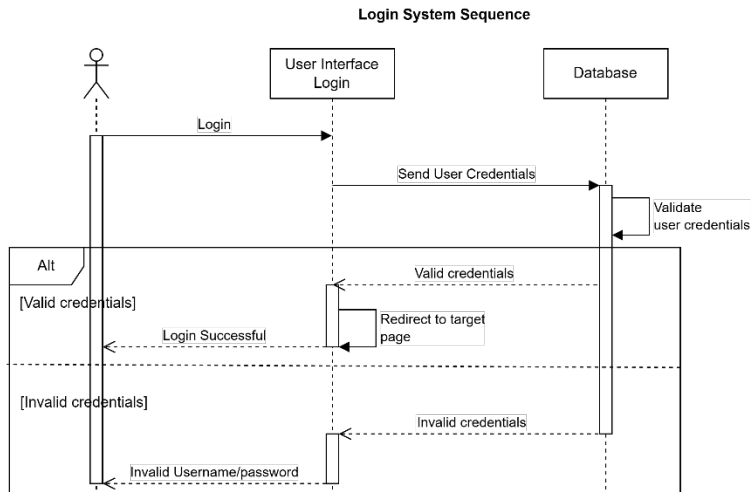


Figure 4.2: System Sequence Diagram

## 4.4 System Design Diagrams

System design diagrams are graphic representations that show the interactions between several parts of a process. These diagrams are crucial for comprehending, organising, and improving systems in a variety of fields.

### 4.4.1 Entity Relationship Diagram

An entity relationship diagram is a visual representation of a relational database's structure. It depicts how entities interact with one another within the system. These are essential for developing and understanding databases because they display entities, properties, and relationships between them.

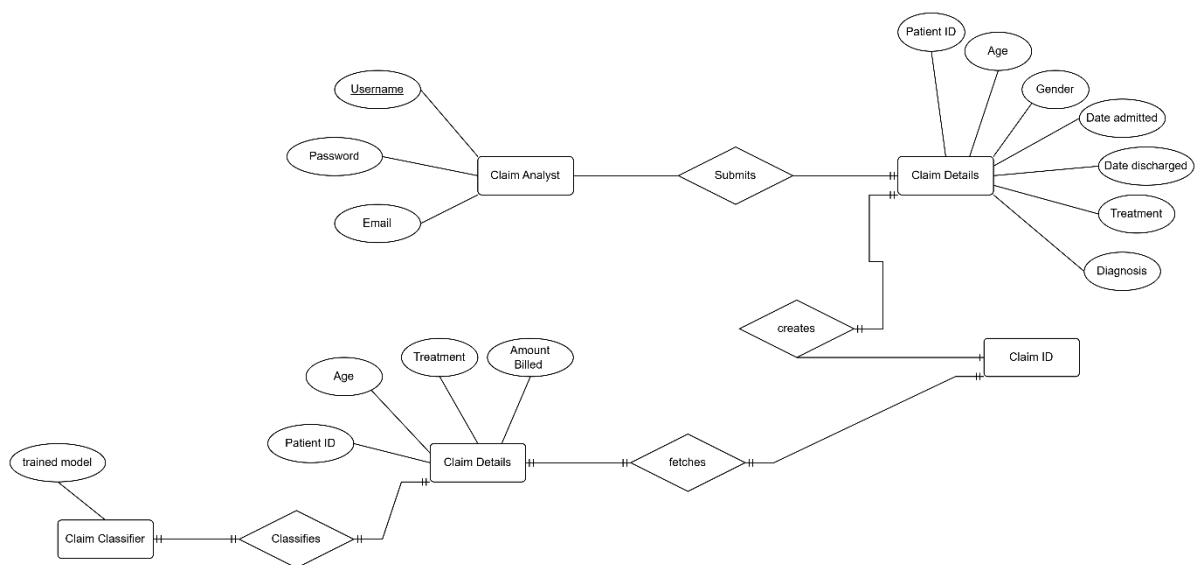


Figure 4.3: Entity Relationship Diagram

### 4.4.2 Database Schema

This is a diagram to show the database schema of the system. The database schema defines how data is organized within a relational database. It shows the various entities in the system inclusive of logical constraints such as, tables and data types.

users Table

Column Name	Data Type	Constraints	Description
id	int	Primary Key	Unique identifier for each user
username	varchar (100)	Unique, Not Null	Login username for user
password	varchar (255)	Not Null	User's password
email	varchar (100)	Unique, Not Null	User's email address

role	varchar (50)	Not Null	User's role (Admin, Claim Analyst)
created_on	datetime	Not Null	Account creation timestamp
last_login	datetime		Last login timestamp

*Table 4.1: Users Table Database Schema*

medical\_claims Table

Column Name	Data Type	Constraints	Description
claim_id	int	Primary Key	Unique identifier for each claim
insured_name	varchar (100)	Not Null	Full name for each client
age	int	Not Null	Client's age
gender	varchar (20)	Not Null	Client's gender
admission_date	date	Not Null	Date of admission
discharge_date	date	Not Null	Date discharged
amount_billed	float (10,2)	Not Null	Amount billed
diagnosis	varchar (20)	Not Null	Client's diagnosis
treatment	varchar (20)	Not Null	Treatment of diagnosis
created_on	Datetime	Not Null	Date claim was created
status	varchar (20)		Status of claim (Fraud, Legit)

*Table 4.2: Medical Claims Table Database Schema*

#### **4.4.3 Class Diagram**

A class diagram is a form of structural diagram in Unified Modelling Language that demonstrates a system's structure by modelling the classes, characteristics, methods, and object relationships. This is shown in figure 4.4.

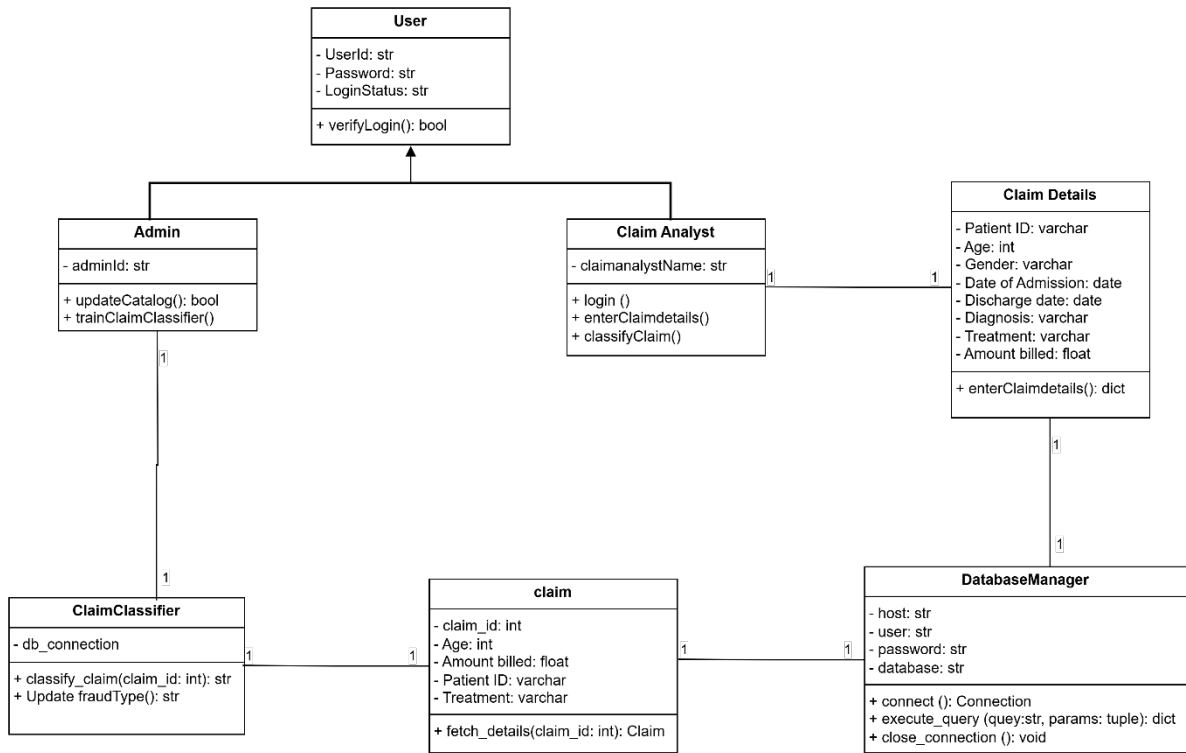


Figure 4.4: Class Diagram

## **Chapter 5: System Implementation and Testing**

### **5.1 Introduction**

The previous chapter covered system analysis, the design process, and feasibility studies. Bugs will very certainly exist in software modules. This is primarily due to complexity rather than irresponsibility on the part of system analysts. Humans have limited ability to manage complexity. This chapter focusses on testing solutions and implementation approaches.

### **5.2 Description of Implementation Environment**

#### **5.2.1 Hardware Specifications**

Computer – The system model is a HP EliteBook 1040 G4 which is a x64-based computer. The systems processor is Intel® Core™ i7-7500U CPU @ 2.70GHz. It has a Random Access Memory of 16 GB and a storage of 500 GB. The system runs on Microsoft Windows 10 Pro.

#### **5.2.2 Software Specifications**

- i. A web browsing service e.g Google Chrome.
- ii. Google Colaboratory which was used to develop the machine learning classification model.
- iii. Sublime Text editor which was used to develop the web-based application.

### **5.3 Description of Dataset**

Real world data was limited therefore an open-source health care dataset was used for training and testing. This is because of the data privacy concerns in financial institutions sharing confidential information of their clients. An open source was obtained from Kaggle (Chosen, 2024), describing the characteristics of fraudulent and genuine insurance claims based on Patient ID, Age, Gender, Date Admitted, Date Discharged, Diagnosis, Treatment, Amount Billed. The dataset has a total number of 1500 rows and 9 columns, as described in figure 5.1.

```
[ ] df.info() # total columns and data types of the dataset

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   Patient ID            1500 non-null  object 
1   Age                   1500 non-null  int64  
2   Gender                1500 non-null  object 
3   Date Admitted         1500 non-null  object 
4   Date Discharged       1500 non-null  object 
5   Diagnosis              1500 non-null  object 
6   Treatment              1500 non-null  object 
7   Amount Billed         1500 non-null  float64 
8   Fraud Type            1500 non-null  object 
dtypes: float64(1), int64(1), object(7)
memory usage: 105.6+ KB
```

Figure 5.1: Dataset Description

There were no null records in the dataset as shown in figure 5.2.

```
df.isna().sum() # number of null records in dataset

0
Patient ID    0
Age           0
Gender        0
Date Admitted 0
Date Discharged 0
Diagnosis     0
Treatment     0
Amount Billed 0
Fraud Type    0
dtype: int64
```

Figure 5.2: Number of Null Records

## 5.4 Description of Testing

### 5.4.1 Data Preparation

This involved removing outliers from the dataset and dealing with missing values. In this case since there were no missing values, we first checked for duplicated data as shown in the figure 5.3.



```
[ ] # check duplicate claims
df.drop_duplicates(inplace=True)
df.shape
```

(1500, 9)

Figure 5.3: Check Duplicate Records

The next step was to remove outliers such as slash (/), hyphen (-), leading and trailing whitespace and finally to convert the data to lowercase.

```
[ ] # Data clean-up
for col in ['Patient ID', 'Gender', 'Date Admitted', 'Date Discharged', 'Diagnosis', 'Treatment', 'Fraud Type']:
    df[col]=df[col].str.replace('/', '') # remove slash (/)
    df[col]=df[col].str.replace('-', '') # remove hyphen (-)
    df[col]=df[col].str.strip() # remove leading and trailing whitespace
    df[col]=df[col].str.replace('\s+', '') # remove extra whitespace
    df[col]=df[col].str.lower() # convert to lowercase
df.head()
```

	Patient ID	Age	Gender	Date Admitted	Date Discharged	Diagnosis	Treatment	Amount Billed	Fraud Type
0	9b4847563a0d47f4ada3fb63ed2d1082	82	male	06022023	06022023	appendectomy	appendectomy	144764.37	no fraud
1	0240f93e8c464c2aa4cb827ecf527d36	29	male	03092022	04092022	cesarean section	fake cesarean section	531434.03	fraud
2	0e0e0476cf0f4b87828ca34bdd780e68	70	male	02102022	17102022	advanced spinal surgery	phantom procedure	128604.41	fraud
3	ad669adf568d4346b60ea26a3374f00d	12	male	15112023	20112023	peptic ulcer	peptic ulcer	304989.18	no fraud
4	9c0cc9c5ff6f442b9c2560ad3f1a4400	72	male	15112022	04122022	appendectomy	appendectomy	277021.33	no fraud

Figure 5.4: Data Clean-up

Columns Patient ID, Diagnosis, Treatment and Fraud Type contain categorical data thus needed to be transformed into formats that machine learning classifiers can understand. The data was transformed using label encoding.

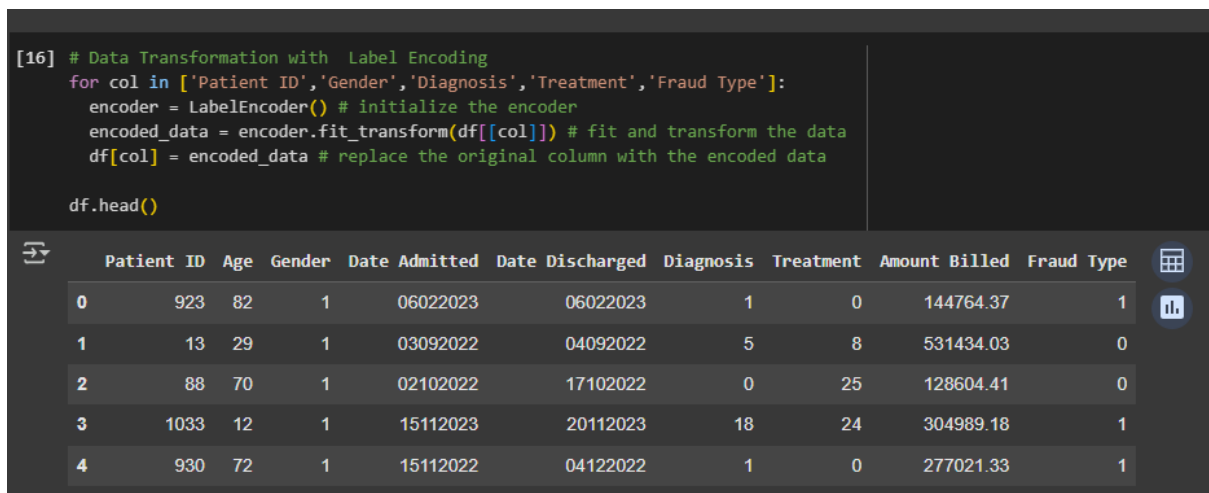


Figure 5.5: Data Transformation

Feature selection is the process of choosing a subset of relevant features to employ in model creation. It is an important stage in machine learning that enhances model performance and interpretability. Feature selection helps to reduce overfitting and improves accuracy. In this study, we used filter methods, Pearson's correlation, for feature selection. Below is the correlation heatmap.

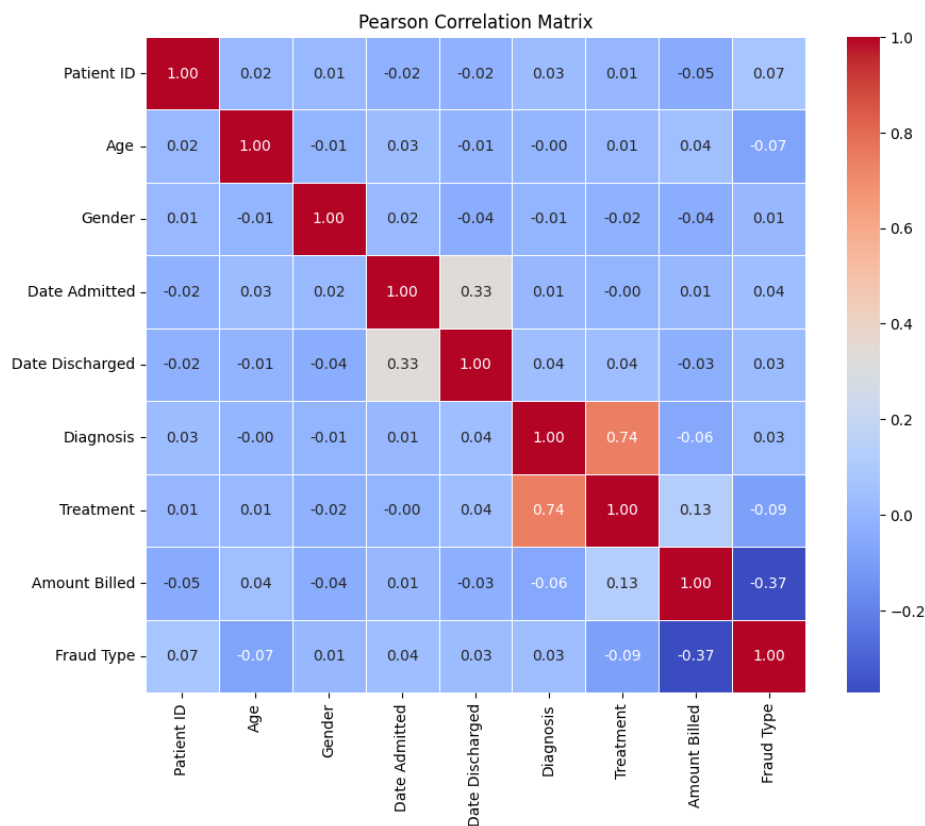
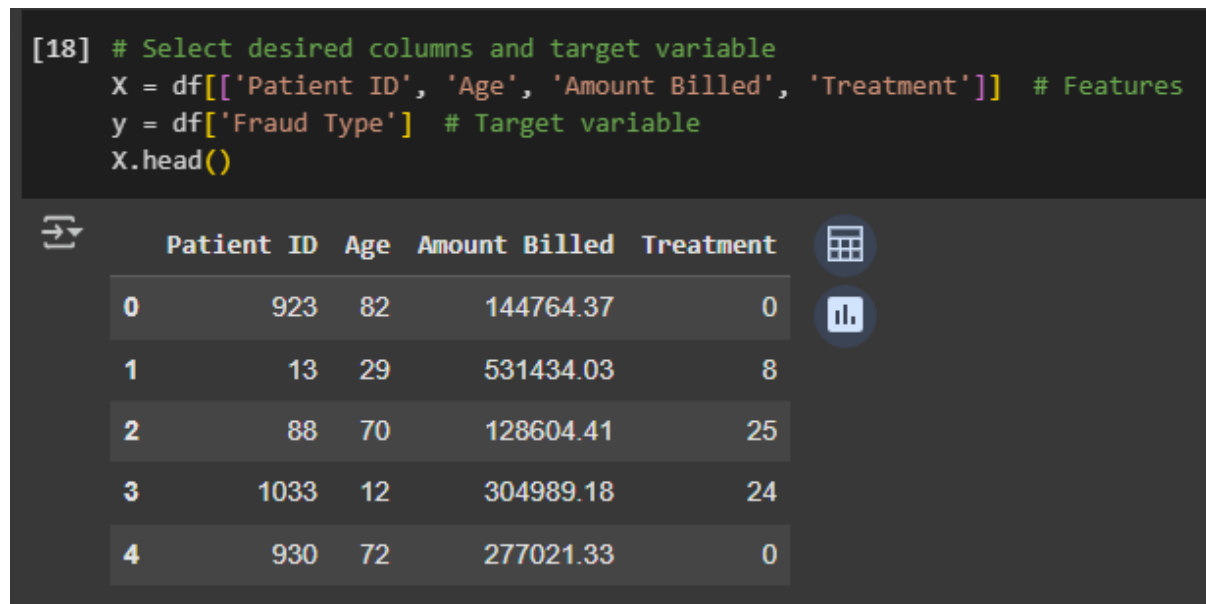


Figure 5.6: Pearson's Correlation Heatmap

Based on the heat map above, the most viable columns selected were, Patient ID, Age, Treatment and Amount Billed since they had a relatively strong correlation with the target variable, Fraud Type.

```
[18] # Select desired columns and target variable
X = df[['Patient ID', 'Age', 'Amount Billed', 'Treatment']] # Features
y = df['Fraud Type'] # Target variable
X.head()
```



The screenshot shows a Jupyter Notebook interface. At the top, there is a code cell with the following Python code: `[18] # Select desired columns and target variable`, `X = df[['Patient ID', 'Age', 'Amount Billed', 'Treatment']] # Features`, `y = df['Fraud Type'] # Target variable`, and `X.head()`. Below the code cell, the output is displayed as a table with 5 columns: Patient ID, Age, Amount Billed, Treatment, and an unlabeled column (likely the index). The table contains 5 rows of data. To the right of the table, there are two circular icons: a calendar icon and a bar chart icon.

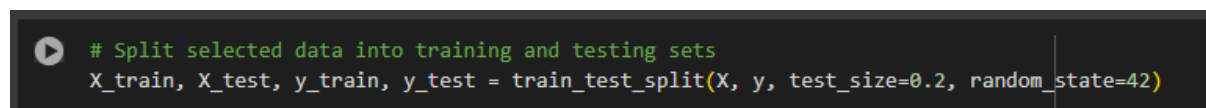
	Patient ID	Age	Amount Billed	Treatment
0	923	82	144764.37	0
1	13	29	531434.03	8
2	88	70	128604.41	25
3	1033	12	304989.18	24
4	930	72	277021.33	0

Figure 5.7: Selected Columns

### 5.4.2 Model Development

The data frame with the selected columns was split into training and testing sets for training with a test size of 0.2 and random state 42.

```
# Split selected data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```



The screenshot shows a Jupyter Notebook interface. At the top, there is a code cell with the following Python code: `# Split selected data into training and testing sets`, `X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)`. Below the code cell, the output is displayed as a table with 5 columns: Patient ID, Age, Amount Billed, Treatment, and an unlabeled column (likely the index). The table contains 5 rows of data. To the right of the table, there are two circular icons: a calendar icon and a bar chart icon.

Figure 5.8: Splitting Dataset

The study used four machine learning classifier algorithms for training namely.

- Logistic Regression -The model was chosen because of its efficiency for faster training and prediction.
- Support Vector Machine – The model was chosen for its robustness to outliers hence expected to maintain a good performance.
- AdaBoost Classifier – For its focus on achieving high accuracy which is crucial in detecting fraudulent claims.
- XGBoost Classifier – For its reputation of excellent performance as well as regularization which ensures model generalizes well to new unseen claims.

The model selected for claim classification was based on the best performance in accuracy which was AdaBoost as demonstrated in figure 5.9. This is due to AdaBoost being an ensemble

learning technique that builds a strong learner by combining several weak learners. These weak learners are trained iteratively, with future iterations assigning greater weight to data points that were incorrectly identified. The performance of the other classification algorithms are presented in figure 5.9.

```

Logistic Regression Metrics:
Accuracy: 0.6166666666666667
Precision: 0.6132810356237793
Recall: 0.6166666666666667
F1 Score: 0.6148060259590838
Confusion Matrix:
[[ 52  60]
 [ 55 133]]
Classification Report:
              precision    recall  f1-score   support

     0       0.49       0.46       0.47       112
     1       0.69       0.71       0.70       188

 accuracy          0.59
 macro avg          0.59
weighted avg          0.61

Support Vector Machine Metrics:
Accuracy: 0.73
Precision: 0.8113011152416356
Recall: 0.73
F1 Score: 0.677459258465823
Confusion Matrix:
[[ 31  81]
 [  0 188]]
Classification Report:
              precision    recall  f1-score   support

     0       1.00       0.28       0.43       112
     1       0.70       1.00       0.82       188

 accuracy          0.73
 macro avg          0.85
weighted avg          0.81

AdaBoost Metrics:
Accuracy: 0.9033333333333333
Precision: 0.9162519201228879
Recall: 0.9033333333333333
F1 Score: 0.8996062044950933
Confusion Matrix:
[[ 83  29]
 [  0 188]]
Classification Report:
              precision    recall  f1-score   support

     0       1.00       0.74       0.85       112
     1       0.87       1.00       0.93       188

 accuracy          0.90
 macro avg          0.93
weighted avg          0.92

XGBoost Metrics:
Accuracy: 0.8833333333333333
Precision: 0.8851519041506204
Recall: 0.8833333333333333
F1 Score: 0.8810659979594597
Confusion Matrix:
[[ 86  26]
 [  9 179]]
Classification Report:
              precision    recall  f1-score   support

     0       0.91       0.77       0.83       112
     1       0.87       0.95       0.91       188

 accuracy          0.89
 macro avg          0.89
weighted avg          0.89

```

Figure 5.9: Classifier Algorithms Performance

### 5.4.3 Claims Classification System

A web-based application was developed using Flask framework. The application was integrated with the machine learning model for classification. This is where claim details were entered and submitted into the database generating a claim ID for each claim. This is shown in the figure 5.10.

The screenshot displays a web application titled "Medical Claims Portal". At the top right, there are "Close" and "Logout" buttons. The main content area features a "Medical Claim Form" with a success message: "Claim submitted successfully! Claim ID: 9". The form contains the following fields:

- Insured Name: Text input field
- Age: Text input field
- Gender: Dropdown menu with "Select" as the current option
- Date of Admission: Date input field with format "dd/mm/yyyy" and a calendar icon
- Discharge Date: Date input field with format "dd/mm/yyyy" and a calendar icon
- Amount Billed: Text input field
- Diagnosis: Dropdown menu with "Select" as the current option
- Treatment: Dropdown menu with "Select" as the current option

A "Submit" button is located at the bottom of the form.

*Figure 5.10: Web-Based Application*

The claim ID was used to fetch details from the database for classification. These details include the Insured Name, Age, Amount Billed and the Treatment. The result of the classification will be displayed on the web-application interface as well as update the status in the database.

The screenshot shows a web application titled "Medical Claims Portal" with a dark blue header. In the top right corner of the header are "Close" and "Logout" buttons. The main content area is light blue and contains a white modal box titled "Claims Classification Form". Inside the modal, it says "Claim ID 9 classified as: **Fraud**". Below this is a label "Claim ID:" followed by an empty text input field. At the bottom of the modal is a dark blue button labeled "Classify Claim".

*Figure 5.11: Claim Classification*

#### **5.4.4 Testing**

In verifying for expected functionality by the system, two approaches were largely applied: black box testing and white box testing. In black box testing, the system was tested with the view of having limited to no knowledge of how it operates. For instance, if a user provides an input and receives a result, the tester does not investigate how the system generates the outcome. The other approach on white box testing focused on the system's internal data structures and algorithms. The tests include the code, branches, pathways, statements, and logic of the code. These assessments demanded programming ability to identify all the paths.

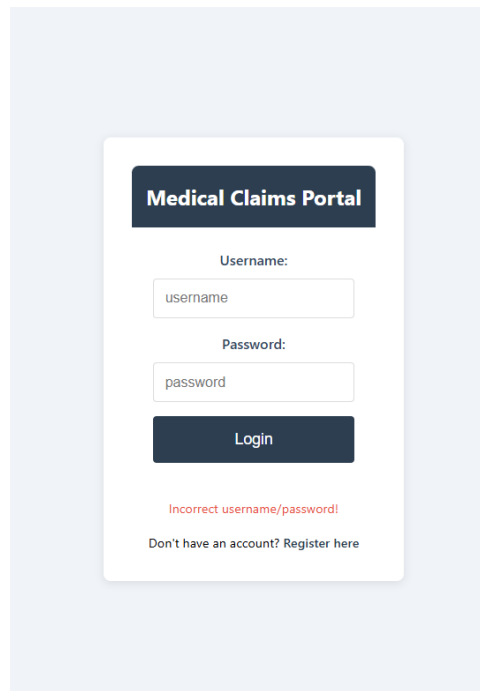
Black box testing was used for the first time to test all functions and procedures. Expected outcomes were tested using various inputs. White box testing was utilised if a function or procedure did not work. As a result, every line of code was put through a series test. The functionalities and methods were combined and tested. Likewise, if it failed to produce the desired results, white box testing was used. The interfaces were tested using the black box technique once the system was finished.

## 5.5 Testing Results

Test Case ID: 1

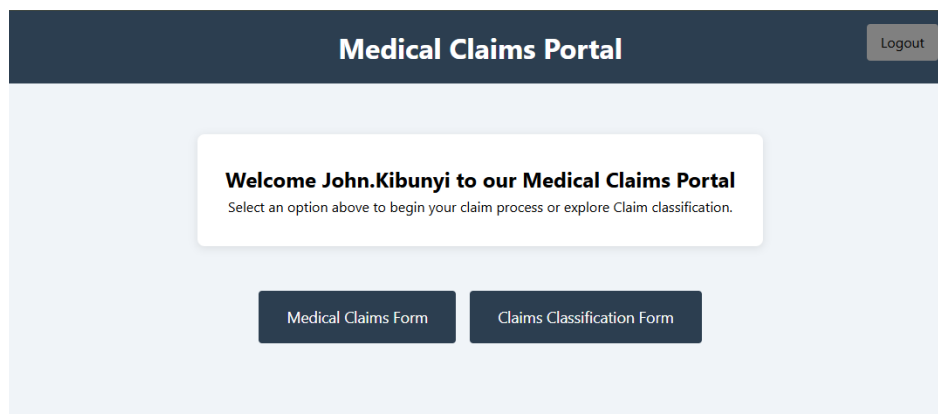
User Interface: User Login

Action	Input	Expected Output	Status
Type Username, password and click log in button (Correct username and password)	Username: John.Kibunyi Password: Password@123	Welcome “Username” to our Medical Claims portal	Pass
Type invalid Username, keep password empty and click login	John	“Please fill out this field” on password	Pass
Type valid User ID and keep password empty then click login	John.Kibunyi	“Please fill out this field” on password	Pass
Keep both User ID and password empty and click login		“Please fill out this field” on password	Pass
Type both User ID and password invalid and click login	Username: John Password: @123	“Incorrect ID/password”	Pass



The image shows a login module for the Medical Claims Portal. It features a dark blue header with the text "Medical Claims Portal". Below the header, there are two input fields: "Username:" and "Password:". The "Username:" field contains the text "username" and the "Password:" field contains the text "password". Below these fields is a dark blue "Login" button. Underneath the button, there is a red error message that reads "Incorrect username/password!". At the bottom of the module, there is a link that says "Don't have an account? Register here".

Figure 5.12: Login Module



The image shows the dashboard of the Medical Claims Portal. It has a dark blue header with the text "Medical Claims Portal" and a "Logout" button. Below the header, there is a white box with the text "Welcome John.Kibunyi to our Medical Claims Portal" and a subtext "Select an option above to begin your claim process or explore Claim classification.". Below this box, there are two dark blue buttons: "Medical Claims Form" and "Claims Classification Form".

Figure 5.13: Dashboard

Test Case ID: 2

User Interface: Medical Claims Form

Action	Input	Expected Outcome	Status
Keep all fields empty and click submit		"Please fill out this field" on Insured Name	Pass
Fill out all the fields with correct values	Fill in details as stipulated	"Claim Submitted Successfully! Claim ID %s"	Pass



Medical Claims Portal

CloseLogout

Medical Claim Form

Insured Name:

Age:

Gender:

Date of Admission:

Please fill in this field.

Select

dd/mm/yyyy

Discharge Date:

Amount Billed:

dd/mm/yyyy

Diagnosis:

Treatment:

Select

Select

Submit

Figure 5.14: Medical Claim Form

Test Case ID: 3

User Interface: Claims Classification Form

Action	Input	Expected Outcome	Status
Type Valid Claim ID	1	“Claim ID %s classified as: Legit/Fraud”	Pass
Type invalid Claim ID	10	“Invalid Claim ID”	Pass
Keep the field empty and click classify Claim		“Please fill in this field”	Pass

Medical Claims Portal

CloseLogout

Claims Classification Form

Claim ID:

Please fill in this field.

*Figure 5.15: Claims Classification Form*

## **Chapter 6: Conclusions, Recommendations and Future Works**

### **6.1 Conclusions**

The study uncovered a variety of types of healthcare insurance fraud, including: Upcoding which refers to billing for more expensive operations than were performed. Unbundling which involves separating procedures that should be billed as a single package. Manipulating Medical Records to support fraudulent claims. Using another person's insurance information to file a claim. Traditional fraud detection techniques mostly rely on rule-based algorithms and manual investigations. Because rule-based systems rely on predetermined criteria developed by subject matter experts, their capacity for discovery is constrained. Four classification techniques were used in the study to successfully create a machine learning model: AdaBoost Classifier (accuracy 90%), XGBoost Classifier (accuracy 89%), Support Vector Machine (73%), and Logistic Regression (62%). AdaBoost was chosen as the best algorithm due to its greater performance at detecting fraudulent claims. The model was integrated into a web-based application for real-time fraud detection and classification.

### **6.2 Recommendations**

The model should incorporate a better labelled dataset to improve on the correlation between features. This would help to avoid overfitting or underfitting of the model. The model should also incorporate other machine learning classifier algorithms such as Decision Tree, Artificial Neural Networks, Naïve Bayes Classifier, K-Nearest Neighbour and Random Forest. The model should also incorporate multi-label classification.

### **6.3 Future Works**

The future work is to find a better labelled dataset to improve the correlation score between features. Moreover, to implement other machine learning classifiers such as Decision Tree, Artificial Neural Networks, Naïve Bayes Classifier, K-Nearest Neighbour and Random Forest. Lastly, to build a model that would support multi-label classification.

## References

- Abdelhadi Dyouri, K. H. (2022, December 21). *How To Make a Web Application Using Flask in Python* 3. Retrieved from DigitalOcean: <https://www.digitalocean.com/community/tutorials/how-to-make-a-web-application-using-flask-in-python-3>
- Agarwal, S. (2023). An Intelligent Machine Learning Approach for Fraud Detection in Medical Claim Insurance: a Comprehensive Study. *Scholars Journal of Engineering and Technology*.
- Authority, I. R. (2020). *Insurance Industry Annual Report* .
- Bhavna, B. &. (2019). Naive Classification Approach for Insurance Fraud Prediction. *Internation Journal of Engineering and Advanced Technology (IJEAT)*, ISSN: 2249-8958 vol 8 Issue 5.
- Burri, R. B. (2019). Insurance Claim Analysis Using Machine Learning Algorithms. *International Journal of Innovative Technology and Exploring Engineering*. Vol, Issue 6, Special Issue 4, 557-582.
- Buttice, C. (2019). *Universal Health Care*. Greenwood Publishing Group.
- Charles, M. A. (2020). Detecting Fraudulent Motor Insurance Claims Using Support Vector Machines with Adaptive Sythentic Sampling Method. *61st International Scientific Conference on Information Technology and Management Science of Riga Technical University(ITMS)*, (pp. 1-5).
- Chosen, B. (2024). Retrieved from Kaggle: <https://www.kaggle.com/datasets/bonifacechosen/nhis-healthcare-claims-and-fraud-dataset>
- Dimitrakopoulos, G. V. (2018). Pathway Analysis Using XGBoost Classification in Biomedical Data. In: *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, (pp. 1-6).
- Gedela, B. &. (2022). Credit Card Fraud Detection using AdaBoost Algorithm in Comparison with Various Machine Learning Algorithms to Measure Accuracy, Sensitivity, Specificity, Precision and F-score. *International Conference on Business Analytics for Technology and Security* , (pp. 1-6).

- Gupta, R. Y. (2021). A Comparative study of using various machine learning and deep learning-based fraud detection models for universal health coverage. *International Journal of Engineering Trends and Technology*, 96-102.
- Hammad, M. (2022, January 19). *Difference between structured and object-oriented analysis*. Retrieved from geekforgeeks: <https://www.geekforgeeks.org/difference-between-structured-and-object-oriented-analysis/>
- Insurers, A. o. (2020). *Insurance Industry Annual Report*. Nairobi.
- Kanade, V. (2022, November 17). *What is Extreme Programming? Meaning, Working, and Principles*. Retrieved from Spiceworks: <https://www.spiceworks.com/tech/devops/articles/what-is-extreme-programming/>
- Kramer, M. (2018). *Best practices in systems development lifecycle: An analysis based on the waterfall model*.
- Luther. (2020, July 23). *Health insurance Fraud and its impact on the healthcare system*. Retrieved from Pacific Prime: <https://www.pacificprime.com/blog/healthcare-system-fraud-impacts.html>
- Mambo, S. R. (2019). *Use of Data Mining to Detect Fraud Health Insurance Claims*. Nairobi: University of Nairobi.
- Matloob, I. &. (2019). *A Framework for fraud detection in government supported national healthcare programs*. Romania: Electronics, computers and Artificial Intelligence, ECAI.
- Miro. (2022, May 18). *What is a Context diagram and how do you use it?* Retrieved from MicroBlog: <https://micro.com/blog/context-diagram/>
- Mishra, A. (2021). Retrieved from Fraud Detection: A study of AdaBoost Classifier and K-Means Clustering: <https://ssrn.com/abstract=3789879>
- Moon, H. P. (2019). A Predictive Modelling for Detecting Fraudulent Automobile Insurance Claims. In *Theoretical Economics Letters* (pp. 1886-1900).
- Njeru, A. M. (2022). *Detection of Fraudulent Vehicle Insurance Claims Using Machine Learning*. Nairobi: Doctoral dissertation, University of Nairobi.

- Onyango, M. N. (2022). *Application of Machine Learning to Detect Fraudulent Maternal Medical Claims*. Nairobi: Doctoral dissertation, University of Nairobi.
- Organization, A. I. (2018). Africa Insurance Barometer 2018; Market Survey. Cameroon, Douala.
- Pal, S. K. (2024, May 29). *What is Extreme Programming (XP)?* Retrieved from geekforgeeks: <https://www.geeksforgeeks.org/software-engineering-extreme-programming-xp/>
- Prashant. (2024). *Waterfall Model in Software Engineering | Modified Waterfall Model*. Retrieved from The Study Genius: <https://radhikaclasses.com/waterfall-model-in-software-engineering/>
- Raeburn, A. (2024, February 13). *Extreme programming(XP) gets results, but is it right for you?* Retrieved from Asana: <https://asana.com/resources/extreme-programming-xp>
- Rutkowska, M. (2023, May 29). *Use Case Diagrams: An Introduction*. Retrieved from Altkom Software: <https://www.altkomsoftware.com/blog/use-case-diagrams-an-introduction/>
- Sandoval, J. (2023, August 15). *What Is a Schema Diagram? A Guide with 10 Examples*. Retrieved from Vertabelo: <https://vertabelo.com/blog/schema-diagram/>
- Starr, P. (2018). *The social transformation of American medicine: The rise of a sovereign profession and the making of a vast industry*. NY:Basic Books.
- Syaiful Anam, M. R. (June 2023). Health Claim Insurance Prediction Using Support Vector Machine with Particle Swarm Optimization. *Journal of Mathematics and Its Applications*, 0797-0806.
- VanZandt, P. (2023, November 03). *What is Sequence Diagram? Definition and Sequence diagrams in UML*. Retrieved from Ideascale: <https://ideascale.com/blog/what-is-sequence-diagram/>
- Waghade, S. S. (2018). A Comprehensive Study of Healthcare Fraud Detection based on machine learning. . Nagpur: *International Journal of Applied Engineering Research*.
- Whitfield, B. (2024, June 03). *What Is an Entity Relationship Diagram (ERD)?* Retrieved from builtin: <https://builtin.com/articles/entity-relationship-diagram>
- zhou, S. H. (2020). *Big data-driven abnormal behavior detection in healthcare based on association rules*. IEEE Access, 129002-129011.

Appendix

Appendix A: Gantt Chart

