Nottingham Trent University

School of Science and Technology

Fraud Detection in Health Insurance Claims

by

Kaustubh Pradip Bhale

in

2025

Project report in part fulfilment

of the requirements for the degree of

Master of Science

In

Data Science

I hereby declare that I am the sole author of this report. I authorize the Nottingham Trent University to lend this report to other institutions or individuals for the purpose of scholarly research.

I also authorize the Nottingham Trent University to reproduce this report by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Signature: KAUSTUBH PRADIP BHALE

Kaustubh Pradip Bhale

# **A****BSTRACT**

# 

This work addresses the problem of fraudulent claims in healthcare insurance systems, which pose a significant financial burden on providers and insurers. The objective of the project was to develop a reliable and efficient healthcare fraud detection system using machine learning and deploy it via a web-based Flask application.

The study utilized a large dataset comprising 558,211 medical insurance records, incorporating both inpatient and outpatient claim details. Preprocessing techniques such as label encoding for high-cardinality categorical variables and one-hot encoding for low-cardinality fields were employed to prepare the data. An ensemble machine learning model was developed, combining LightGBM, Random Forest, and XGBoost classifiers using a soft voting mechanism to enhance predictive performance.

The final model achieved an overall accuracy of 90% and a precision of 89.6%, demonstrating high effectiveness in distinguishing fraudulent from non-fraudulent claims. Evaluation metrics including the classification report, confusion matrix, ROC-AUC, and precision-recall curves were used to validate model performance. The Flask-based web interface enables real-time fraud prediction by accepting user input and returning the predicted probability of fraud along with classification results.

This implementation presents a scalable and interpretable solution to healthcare fraud detection, supporting better decision-making for medical insurance claims while reducing false positives and improving trust in automated fraud detection systems.

# **Acknowledgements**

I would like to express my sincere gratitude to Professor Monari Dennisfor their invaluable guidance, encouragement, and support throughout the course of this project. Their insightful feedback, deep knowledge of the subject, and constant motivation have played a pivotal role in shaping the direction and outcome of this work.

I am also thankful for the opportunity to learn under their supervision, which has greatly enriched my academic experience and understanding of the topic.

Finally, I extend my thanks to all those who directly or indirectly contributed to the successful completion of this report.

Table of Contents

[ABSTRACT ii](#_Toc207039457)

[Acknowledgements iii](#_Toc207039458)

[Table of Contents iv](#_Toc207039459)

[List of Figures ix](#_Toc207039460)

[List of Tables xi](#_Toc207039461)

[Chapter 1 12](#_Toc207039462)

[Introduction 12](#_Toc207039463)

[1.1 Background and Context 12](#_Toc207039464)

[1.2 Problem statement 13](#_Toc207039465)

[1.3 Project Scope 13](#_Toc207039466)

[1.4 Importance and Relevance to Computing 13](#_Toc207039467)

[1.5 Aims and Objectives 14](#_Toc207039468)

[1.5.1 Project Aim 14](#_Toc207039469)

[1.5.2 Objectives 14](#_Toc207039470)

[Chapter 2 17](#_Toc207039471)

[Literature Review 17](#_Toc207039472)

[2.1 Introduction 17](#_Toc207039473)

[2.2 Background and wider context 18](#_Toc207039474)

[2.2.1 Defining healthcare fraud 18](#_Toc207039475)

[2.2.2 Global and Regulatory Landscape 19](#_Toc207039476)

[2.3 State-of-Art-Approaches 19](#_Toc207039477)

[2.3.1 Traditional and Rule based approaches 19](#_Toc207039478)

[2.3.2 Supervised Machines Learning Approaches 20](#_Toc207039479)

[2.3.3 Unsupervised and Hybrid ML models 20](#_Toc207039480)

[2.3.4 Industrial and Software Solutions 21](#_Toc207039481)

[2.4 Critical Evaluation 21](#_Toc207039482)

[2.5 Identified Research Gaps 22](#_Toc207039483)

[2.6 Research Questions and Hypotheses 24](#_Toc207039484)

[2.7 Conclusion 27](#_Toc207039485)

[Chapter 3 28](#_Toc207039486)

[New Ideas & Approach 28](#_Toc207039487)

[3.1 Introduction 28](#_Toc207039488)

[3.2 Exploration of Individual ML Models 28](#_Toc207039489)

[3.3 Approach to research questions and hypotheses 30](#_Toc207039490)

[3.4 Project Workflow 33](#_Toc207039491)

[3.4.1 Data Acquisition 33](#_Toc207039492)

[3.4.2 Data Cleaning and Preprocessing 34](#_Toc207039493)

[3.4.3 Exploratory Data Analysis (EDA) 34](#_Toc207039494)

[3.4.4 Feature Engineering 34](#_Toc207039495)

[3.4.5 Handling Class Imbalances 35](#_Toc207039496)

[3.4.6 Model Development and Evaluation 35](#_Toc207039497)

[3.4.7 Model Deployment and Visualization 35](#_Toc207039498)

[3.4.8 Ethical and Legal Compliance 36](#_Toc207039499)

[3.5 Project Timelines 37](#_Toc207039500)

[3.6 Professional, Social, Ethical, and Legal Issues 41](#_Toc207039501)

[3.6.1 Professional Issues 41](#_Toc207039502)

[3.6.2 Ethical Issues 42](#_Toc207039503)

[3.6.3 Social Issues 42](#_Toc207039504)

[3.6.4 Legal Issues 43](#_Toc207039505)

[3.6.5 Impact 43](#_Toc207039506)

[Chapter 4 45](#_Toc207039507)

[IMPLEMENTATION or INVESTIGATION 45](#_Toc207039508)

[4.1 Introduction 45](#_Toc207039509)

[4.2 Dataset Description and Validity Analysis 45](#_Toc207039510)

[4.2.1 Dataset Overview 46](#_Toc207039511)

[4.2.2 Data Validity and Reliability Checks 47](#_Toc207039512)

[4.3 Data Preprocessing and Feature Engineering 48](#_Toc207039513)

[4.3.1 Data Cleaning 48](#_Toc207039514)

[4.3.2 Encoding and Transformation 48](#_Toc207039515)

[4.3.3 Feature Engineering 48](#_Toc207039516)

[4.4 Exploratory Data Analysis (EDA) 50](#_Toc207039517)

[4.4.1 Class Distribution of target variable 50](#_Toc207039518)

[4.4.2 Age Distribution 51](#_Toc207039519)

[4.4.3 Distribution of Chronic Conditions Among Patient Population 52](#_Toc207039520)

[4.4.4 Chronic Condition Contribution to Potential Fraud 53](#_Toc207039521)

[4.4.5 Fraud Distribution by Patient Type 54](#_Toc207039522)

[4.4.6 Fraud Distribution by Gender 55](#_Toc207039523)

[4.4.7 Claim Volume Analysis of Top Attending Physicians and Fraud Indicators 56](#_Toc207039524)

[4.4.8 Hospitalization Duration and Its Association with Fraudulent Claims 57](#_Toc207039525)

[4.4.9 Analysis of Providers with Highest Fraudulent Claim Counts 58](#_Toc207039526)

[4.4.10 Hospitalization Duration vs Claim Value 59](#_Toc207039527)

[4.4.11 Reimbursement Distribution by Fraud Status 60](#_Toc207039528)

[4.4.12 Analysis of Healthcare Fraud Cases by State 61](#_Toc207039529)

[4.4.13: Analysis of Correlation Matrix for Healthcare Claims Data 62](#_Toc207039530)

[4.5 ML Model Development Process 64](#_Toc207039531)

[4.5.1 Baseline Models 64](#_Toc207039532)

[4.5.2 Ensemble Model Development 66](#_Toc207039533)

[4.6 Model Deployment and UI Integration 70](#_Toc207039534)

[Chapter 5 79](#_Toc207039535)

[Evaluation 79](#_Toc207039536)

[5.1 Introduction 79](#_Toc207039537)

[5.2 Testing Methodology 79](#_Toc207039538)

[5.2.1 Machine Learning Model Testing 79](#_Toc207039539)

[5.2.2 User Interface (UI) Testing 81](#_Toc207039540)

[5.3 Classification Report and Confusion Matrix 81](#_Toc207039541)

[5.4 ROC and AUC 82](#_Toc207039542)

[5.5 Precision-Recall (PR) Curve 83](#_Toc207039543)

[5.6 Model Interpretability Using SHAP Analysis 84](#_Toc207039544)

[5.6.1 Overall Feature Importance and Impact Direction 84](#_Toc207039545)

[5.6.2 Detailed Influence of a Key Feature: Provider 86](#_Toc207039546)

[5.6.3 Explanation of an Individual Prediction 87](#_Toc207039547)

[5.6.4 Concise Visualization of Feature Contributions for an Instance 87](#_Toc207039548)

[5.7 UI and System Testing 89](#_Toc207039549)

[5.8 Conclusion 92](#_Toc207039550)

[Chapter 6 94](#_Toc207039551)

[CONCLUSIONS / FUTURE WORK 94](#_Toc207039552)

[6.1 Summary of Work 94](#_Toc207039553)

[6.1.1 Data Understanding and Preprocessing 94](#_Toc207039554)

[6.1.2 Model Development 94](#_Toc207039555)

[6.1.3 Evaluation 95](#_Toc207039556)

[6.1.4 Deployment and UI Integration 95](#_Toc207039557)

[6.2 Analysis Against Research Questions and Hypotheses 96](#_Toc207039558)

[6.3 Project Success and Limitations 97](#_Toc207039559)

[6.4 Recommendations for Future Work 98](#_Toc207039560)

[6.5 Concluding Remarks 99](#_Toc207039561)

[ReferenceS 100](#_Toc207039562)

[Appendix A 105](#_Toc207039563)

[Project GitHub Repository 105](#_Toc207039564)

List of Figures

[Figure 1: Overview of report 17](#_Toc207032685)

[Figure 2: Project flow block diagram 37](#_Toc207032686)

[Figure 3: Project Gantt Chart 41](#_Toc207032687)

[Figure 4: distribution of target variable 51](#_Toc207032688)

[Figure 5: age distribution of patients 52](#_Toc207032689)

[Figure 6: distribution of chronic conditions 53](#_Toc207032690)

[Figure 7: chronic conditions contribution to potential fraud 54](#_Toc207032691)

[Figure 8: Fraud distribution by patient age 55](#_Toc207032692)

[Figure 9: fraud distribution by gender 56](#_Toc207032693)

[Figure 10: Claim Counts and Fraud Status of Top 20 Attending Physicians 57](#_Toc207032694)

[Figure 11: Fraud vs Non-Fraud Claims Distribution by Hospitalization Duration 58](#_Toc207032695)

[Figure 12: Top 30 Providers by Volume of Fraudulent Claims 59](#_Toc207032696)

[Figure 13: Scatter Plot of Claim Amount vs Hospitalization Duration with Fraud Indicators 60](#_Toc207032697)

[Figure 14: Violin Plot of Reimbursement vs Fraud Status 61](#_Toc207032698)

[Figure 15: Count of Healthcare Fraud Cases by State 62](#_Toc207032699)

[Figure 16: Correlation heatmap 64](#_Toc207032700)

[Figure 17: Ensemble Model Confusion Matrix 69](#_Toc207032701)

[Figure 18: Ensemble Model Precision – Recall Curve 69](#_Toc207032702)

[Figure 19: Ensemble Model ROC - AOC Curve 70](#_Toc207032703)

[Figure 20: Welcome Page of the application 72](#_Toc207032704)

[Figure 21: Navigation Bar 72](#_Toc207032705)

[Figure 22: Data Preview Section 73](#_Toc207032706)

[Figure 23: Fraud Detection Form 74](#_Toc207032707)

[Figure 24: Prediction results post submission of details 74](#_Toc207032708)

[Figure 25: Visualizations Dashboard -1 75](#_Toc207032709)

[Figure 26: Visualizations Dashboard -2 76](#_Toc207032710)

[Figure 27: Visualizations Dashboard -3 76](#_Toc207032711)

[Figure 28: Model Evaluation Dashboard 77](#_Toc207032712)

[Figure 29: Batch Analysis Page 78](#_Toc207032713)

[Figure 30: Classification report for ensemble model 82](#_Toc207032714)

[Figure 31: SHAP DETAILED SUMMARY 86](#_Toc207032715)

[Figure 32: SHAP DEPEDENCE PLOT FOR PROVIDER 87](#_Toc207032716)

[Figure 33: SHAP FORCE PLOT 88](#_Toc207032717)

[Figure 34: SHAP WATERFALL INSTANCE 89](#_Toc207032718)

List of Tables

[Table 1: Common Fraud types 19](#_Toc207028277)

[Table 2: Comparative analysis of existing literature 25](#_Toc207028278)

[Table 3: Research Questionnaire and Hypothesis 28](#_Toc207028279)

[Table 4: Individual ML models 30](#_Toc207028280)

[Table 5: Approaches to Research Questions and Corresponding Hypotheses 34](#_Toc207028281)

[Table 5: Table 6: Project tasks and timelines 41](#_Toc207028282)

[Table 7: Data validation and reliability check 47](#_Toc207028283)

[Table 8: Column transformations and Feature Derivations 50](#_Toc207028284)

[Table 9: Baseline ML Models Performance metrics 65](#_Toc207028285)

[Table 10: Key Performance metrics for Ensemble Model 66](#_Toc207028286)

[Table 11: Flask webapp UI Components 71](#_Toc207028287)

[Table 12: ML Model Performance Evaluation Metrics 81](#_Toc207028288)

[Table 13:System Integration Testing Scenarios 87](#_Toc207028289)

[Table 14: Analysis against research and hypothesis 93](#_Toc207028290)

Chapter 1

Introduction

# **Background and Context**

Healthcare fraud has long posed a serious threat to the financial sustainability and operational integrity of national health systems globally. Defined as the deliberate submission of false or misleading information for personal or institutional gain, it manifests through practices such as phantom billing, duplicate claims, kickbacks, upcoding, and misrepresentation of services (Centers for Medicare & Medicaid Services, 2021). In the UK context, the NHS Counter Fraud Authority (NHSCFA) estimates that fraud costs the healthcare system over £1.2 billion annually (NHSCFA, 2023). Historically, healthcare fraud detection was predominantly rule-based—using predefined business rules or expert-defined heuristics to identify anomalies. While these approaches were effective for known fraud types, they lacked adaptability and failed to keep up with evolving fraud patterns (Gee & Button, 2003). The recent explosion of healthcare data, driven by Electronic Health Records (EHRs), digital claims processing, and data-sharing protocols, offers fertile ground for machine learning (ML) approaches (Joudaki et al., 2015).

Advances in ensemble modelling (e.g., Random Forests, XGBoost, LightGBM) have shown considerable promise in detecting fraud by learning from large-scale historical datasets and identifying subtle, non-linear fraud patterns (Bauder & Khoshgoftaar, 2018). Simultaneously, the rise of web-based technologies (like Flask, Django) enables the deployment of user-accessible platforms that integrate fraud detection directly into the workflow of investigators and analysts (Kirkpatrick, 2021).

# **Problem statement**

Despite digitization, fraud detection remains a major bottleneck in health claim processing pipelines. Traditional systems are static, limited in scale, and incapable of adapting to emerging fraud strategies (Smith, 2021). A significant challenge is the class imbalance—fraudulent claims often represent less than 3% of total claims, making it difficult for conventional models to learn meaningful patterns (He & Garcia, 2009). Healthcare data often contains high-cardinality variables and sensitive patient information, posing challenges in data handling, modeling, and privacy preservation. There is a compelling need for intelligent, automated systems that can generalize well, adapt to new fraud trends, and support investigators with actionable insights—all while ensuring regulatory compliance (Kirkpatrick, 2021).

# **1.3 Project Scope**

This project proposes the design and development of a real-time fraud detection system using an ensemble machine learning model, deployed via a Flask-based web interface. It aims to provide a scalable, automated, and transparent solution to healthcare claim fraud. Features include real-time prediction, batch analysis, data preview dashboards, and performance metrics visualizations.

# **1.4 Importance and Relevance to Computing**

This project contributes to computing through the application of ML techniques on imbalanced and high-dimensional datasets (Hastie et al., 2017). It integrates data science and web development, emphasizes GDPR compliance, and promotes modular, open-source design adaptable to other domains like finance or cybersecurity. It pushes academic boundaries by requiring integration of complex technologies, ethical awareness, and system-level thinking.

# **1.5 Aims and Objectives**

## **1.5.1 Project Aim**

To design, implement, and evaluate a real-time fraud detection system for healthcare insurance claims using an ensemble machine learning model, deploying it as an interactive Flask web application, achieving high accuracy and providing actionable insights for stakeholders.

## **1.5.2 Objectives**

* Collect and preprocess historical insurance claims data from public sources, ensuring it is structured, cleaned, and formatted for analysis through techniques like missing value imputation, outlier detection, and normalization.
* Identify key fraud indicators using exploratory data analysis (EDA) with statistical and visual methods to detect anomalies, correlations, and trends indicative of fraud.
* Analyse data patterns to validate and quantify the significance of fraud indicators using advanced methods such as correlation analysis and hypothesis testing.
* Design and implement machine learning classification models, including conventional and ensemble techniques, addressing class imbalance and optimizing parameters via hyperparameter tuning.
* Evaluate model performance using metrics like accuracy, precision, recall, and F1-score, applying cross-validation and comparative analysis to minimize errors and select the best model.
* Develop an interactive, user-friendly interface to visualize model predictions and support stakeholder decision-making with clear, actionable insights.
* Ensure GDPR/HIPAA compliance by anonymizing sensitive data and documenting data handling protocols to protect privacy while maintaining research integrity.
  1. **Overview of the report**

This report is structured to provide a comprehensive account of the research, design, implementation, and evaluation of an ensemble-based healthcare fraud detection system. The project is methodologically organised into six chapters, each addressing specific aspects of the investigation, and collectively contributing to the development of an end-to-end intelligent system. The report begins by setting the context of healthcare fraud within the UK, establishing its impact, and identifying the technological gap that this work seeks to address. It proceeds through the technical planning, system development, and model evaluation stages, culminating in a critical reflection on the outcomes and recommendations for future enhancement.

The structure is designed to maintain clarity and logical progression for the reader. Each chapter builds on the previous one, reinforcing the core objectives and facilitating an understanding of the system's design rationale and evaluation outcomes. Figure 1 provides an overview of the chapters, highlighting their key contents in a summarised bullet-point format to assist with navigation and comprehension of the report's overall flow.

**Figure 1: Overview of report**

Chapter 2

Literature Review

# **2.1 Introduction**

Fraud in healthcare insurance represents one of the most persistent and costly challenges facing modern healthcare systems, with annual global losses estimated to exceed hundreds of billions of US dollars (Smith et al., 2010; European Healthcare Fraud & Corruption Network, 2022). Beyond the financial burden, fraudulent claims undermine the sustainability of healthcare provision, divert resources from genuine patients, and erode public trust in medical institutions. The proliferation of digital health records, cloud-based claim management systems, and large-scale healthcare databases has created both new opportunities and complexities in detecting fraudulent activity (Joudaki et al., 2015).

The field has evolved from labour-intensive manual audits to sophisticated, data-driven approaches using advanced machine learning (ML) and artificial intelligence (AI) techniques (Waghade & Karandikar, 2018). These developments have been accelerated by the availability of computational resources, big data processing frameworks, and the need for proactive, real-time detection mechanisms. However, despite these advances, significant challenges remain in data quality, algorithm interpretability, ethical considerations, and operational integration, leaving ample scope for innovative research.

This chapter builds upon the literature review undertaken during the Research Methods module but incorporates revisions based on prior feedback. It expands the scope by integrating wider sources—including governmental fraud prevention guidelines, industry whitepapers, and software vendor reports—alongside academic studies. The aim is to map the state of the art in healthcare fraud detection, critically analyse the strengths and limitations of current work, and identify the research gap that this project will address.

# **2.2 Background and wider context**

## **2.2.1 Defining healthcare fraud**

Healthcare fraud can be broadly defined as “intentional deception or misrepresentation made by a person or entity, with the knowledge that the deception could result in an unauthorized benefit” (Centers for Medicare & Medicaid Services, 2021). Common forms include:

|  |  |
| --- | --- |
| **Fraud Type** | **Description** |
| **Phantom Billing** | Claiming reimbursement for services that were never rendered. |
| **Upcoding** | Charging for more expensive procedures or services than those actually provided. |
| **Duplicate Claims** | Resubmitting the same claim multiple times to obtain additional payment. |
| **Identity Misuse** | Using stolen patient or provider credentials to file fraudulent claims. |
| **Falsified Documentation** | Altering or fabricating patient records to justify false or inflated claims. |

Table 1: Common Fraud types

These fraudulent behaviours often overlap and evolve, particularly as perpetrators adapt to detection mechanisms.

## **2.2.2 Global and Regulatory Landscape**

Internationally, healthcare fraud detection is influenced by regulatory and policy frameworks such as the General Data Protection Regulation (GDPR) in the EU and the Health Insurance Portability and Accountability Act (HIPAA) in the US (European Parliament, 2016; U.S. Department of Health & Human Services, 2013). These frameworks impose stringent requirements on the handling, storage, and sharing of sensitive medical data. While they protect patient confidentiality, they also introduce constraints on data sharing across institutions, limiting the availability of large, representative datasets for model training (Mehrabi et al., 2021).

Industry bodies and insurers have also published fraud detection guidelines and toolkits. For example, the National Health Care Anti-Fraud Association (NHCAA) in the US promotes collaborative information-sharing platforms among insurers, while private vendors such as SAS, IBM, and Palantir have released commercial fraud analytics software (Healthcare Fraud Prevention Alliance, 2023). These systems often integrate anomaly detection, network analysis, and rule-based alerts but may lack transparency in algorithm design due to proprietary constraints (Kirkpatrick, 2021).

# **2.3 State-of-Art-Approaches**

## **2.3.1 Traditional and Rule based approaches**

The earliest automated systems relied heavily on rule-based algorithms—sets of manually defined conditions that, when triggered, flagged claims as potentially fraudulent (Viaene et al., 2002). While these systems were easy to implement and interpret, they were rigid and could not detect novel fraud patterns beyond predefined rules. They also suffered from high false positive rates, increasing investigation costs (Button et al., 2007).

## **2.3.2 Supervised Machines Learning Approaches**

Supervised learning approaches dominate academic research in healthcare fraud detection, using historical labelled datasets to train classifiers. For example:

* Veena et al. (2023) compared Logistic Regression, Random Forest, Naive Bayes, and Decision Trees, with Decision Trees achieving an accuracy of 97.03% on preprocessed claim data.
* Jaspin et al. (2024) applied Random Forests in real-time claims processing, achieving 90.008% accuracy while reducing manual review workload.
* Farahmandazad and Danesh (2025) addressed class imbalance using SMOTE and feature selection, achieving 98.8% accuracy on Medicare datasets.

While these results are encouraging, accuracy alone can be misleading in imbalanced datasets where fraudulent claims are rare (He & Garcia, 2009). Metrics such as precision, recall, and F1-score are more relevant for assessing fraud detection efficacy (Brownlee, 2020).

## **2.3.3 Unsupervised and Hybrid ML models**

Unsupervised learning is valuable for detecting novel fraud patterns without labelled data. Rawte and Anuradha (2015) proposed a hybrid system combining Evolving Clustering Method (ECM) and Support Vector Machines (SVM) for dynamic fraud detection. Lu et al. (2023) employed Attributed Heterogeneous Information Networks (AHIN) with hierarchical attention, capturing contextual relationships between claim entities.

Hybrid approaches—combining supervised accuracy with unsupervised adaptability—offer resilience to evolving fraud patterns but face challenges in computational cost and integration into operational pipelines (Ahmed et al., 2016).

## **2.3.4 Industrial and Software Solutions**

Commercial fraud detection software, such as SAS Fraud Framework and IBM Safer Payments, integrates rule engines with anomaly detection and visual link analysis (Healthcare Fraud Prevention Alliance, 2023). While these systems are deployed at scale, their proprietary algorithms limit transparency, hindering independent validation. Moreover, vendor solutions may be optimized for general insurance fraud, not specifically tuned for the unique characteristics of healthcare claims data (Kirkpatrick, 2021).

# **2.4 Critical Evaluation**

A synthesis of academic and industry work reveals several recurring challenges:

* **Data Quality and Imbalance**: Fraudulent cases are vastly outnumbered by legitimate claims, creating skewed datasets (Bauder & Khoshgoftaar, 2018).
* **Adaptability to Evolving Fraud Tactics**: Many models degrade over time without periodic retraining (Salau et al., 2023).
* **Interpretability and Regulatory Compliance**: Black-box models can conflict with legal requirements for explainability (Lundberg & Lee, 2017; Mehrabi et al., 2021).
* **Operational Integration Barriers**: Few academic models have been deployed in real-world claim systems, limiting impact (Jaspin et al., 2024).
* **Limited Benchmarking**: The absence of standard public healthcare fraud datasets makes cross-study comparison difficult (Cherkaoui et al., 2023).

# **2.5 Identified Research Gaps**

While prior studies demonstrate high predictive accuracy, there is insufficient integration of interpretability, adaptability, and operational feasibility in healthcare fraud detection frameworks. Existing models are either highly accurate but opaque (e.g., deep learning) or interpretable but rigid (e.g., rule-based systems). Few approaches successfully combine:

* High detection accuracy in imbalanced datasets
* Adaptability to evolving fraud patterns
* Compliance with explainability requirements for regulatory audit
* Scalability for real-time deployment

This research aims to bridge this gap by designing an interpretable, hybrid, and adaptive ML framework for healthcare fraud detection, validated using robust evaluation metrics and assessed for practical deployment feasibility.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Study / Year** | **Algorithm Used** | **Dataset** | **Key Strengths** | **Limitations / Gaps** |
| Bauder & Khoshgoftaar (2018) | Random Forest, Logistic Regression | Medicare claims data (CMS) | High accuracy and interpretability; robust to overfitting | Struggles with extreme class imbalance; requires manual feature engineering |
| Kumar et al. (2020) | Gradient Boosting Machines (GBM) | Proprietary insurance dataset | Better performance than linear models; handles non-linear relationships well | High computation time; limited interpretability |
| Dhingra & Bhatia (2021) | Neural Networks | Private hospital claims | Good at detecting complex fraud patterns | Black-box nature; requires large datasets for training |
| Razzak et al. (2021) | Hybrid model (Rule-based + ML) | Multiple hospital records | Combines expert rules with automated learning; reduces false positives | Rules require frequent updating; not scalable for very large datasets |
| Joudaki et al. (2020) | Support Vector Machines (SVM) | Health insurance claim dataset (Iran) | Works well for high-dimensional data | Poor scalability for large datasets; sensitive to parameter tuning |

Table 2: Comparative analysis of existing literature

# **2.6 Research Questions and Hypotheses**

|  |  |  |
| --- | --- | --- |
| **Research Question (RQ)** | **Hypothesis (H)** | **Approach** |
| **RQ1:** How can an ensemble model improve detection accuracy compared to individual machine learning models in healthcare fraud detection? | **H1:** An ensemble model (LightGBM + XGBoost + Random Forest) will achieve higher accuracy and F1-score than individual models when tested on the same dataset. | 1. Train and evaluate individual ML models (LightGBM, XGBoost, Random Forest) separately.  \2. Develop an ensemble model combining these classifiers via VotingClassifier.  3. Compare performance metrics (accuracy, F1-score) of individual vs ensemble models.  4. Use cross-validation to ensure robustness. |
| **RQ2:** How can data imbalance in healthcare fraud datasets be effectively addressed without compromising model performance? | **H2:** Applying SMOTE with ensemble models will improve recall for fraudulent cases without significantly lowering precision. | 1. Analyze class distribution to confirm imbalance.  2. Apply SMOTE (Synthetic Minority Over-sampling Technique) to balance the training data.  3. Train ensemble and individual models on both original and balanced datasets.  4. Compare recall, precision, and other relevant metrics to evaluate impact. |
| **RQ3:** Can explainability techniques (e.g., SHAP) enhance interpretability of fraud detection models for domain experts? | **H3:** SHAP analysis will provide meaningful and interpretable feature importance insights without degrading predictive performance. | 1. Use SHAP (SHapley Additive exPlanations) to interpret model predictions.  2. Visualize global and local feature importance.  3. Validate if SHAP explanations align with domain knowledge.  4. Ensure model performance remains stable post explainability integration. |
| **RQ4:** What is the trade-off between computational efficiency and model accuracy in hybrid ensemble fraud detection systems? | **H4:** The proposed ensemble model will maintain competitive accuracy while reducing training time compared to more computationally intensive deep learning models. | 1. Evaluate accuracy and other performance metrics.  2. Analyze trade-offs to identify the optimal balance between accuracy and efficiency. |

Table 3: Research Questionnaire and Hypothesis

# **2.7 Conclusion**

This chapter has reviewed the historical development, current state, and industry practices in healthcare fraud detection, identifying clear strengths and limitations in existing approaches. While substantial progress has been made in applying ML techniques, critical gaps remain in achieving an optimal balance between accuracy, adaptability, interpretability, and operational deployment.

The next chapter will build upon these findings to design a methodological framework that directly addresses these gaps, informed by the research questions and hypotheses formulated herein.

Chapter 3

New Ideas & Approach

# **3.1 Introduction**

While Chapter 2 emphasized the reliance on supervised and hybrid machine learning (ML) approaches for health insurance fraud detection, often depending on single-model or static solutions with limited adaptability (Veena et al., 2023; Jaspin et al., 2024), this project proposes a novel counterpoint: an ensemble model combining LightGBM, Random Forest, and XGBoost classifiers with an online learning component. This approach diverges from the literature’s focus on individual classifiers or hybrid methods by leveraging the complementary strengths of these tree-based algorithms to enhance accuracy, robustness, and adaptability. It addresses gaps in model generalizability, interpretability, and handling of imbalanced and evolving data (Salau et al., 2023; He and Garcia, 2009), offering a proactive solution to the reactive, accuracy-driven methods reviewed.

# **3.2 Exploration of Individual ML Models**

Before developing the ensemble model, a comparative analysis of individual ML classifiers will be conducted to establish baseline performance and identify strengths and weaknesses relevant to healthcare fraud detection. This phase allows the project to systematically answer research questions regarding model effectiveness and interpretability and informs the ensemble strategy. The candidate models include:

|  |  |
| --- | --- |
| **Model** | **Description** |
| Random Forest | Known for robustness to overfitting and ease of interpretation via feature importance. |
| LightGBM | A gradient boosting framework optimized for speed and efficiency, effective on large datasets with categorical variables. |
| XGBoost | A popular gradient boosting algorithm with strong predictive performance and extensive hyperparameter tuning capabilities. |
| Support Vector Machine (SVM) | Useful for classification in high-dimensional spaces, though less scalable. |
| Logistic Regression | A baseline linear model, interpretable and useful for benchmarking. |

Table 4: Individual ML models

Each model will be trained and evaluated on the same pre-processed Kaggle healthcare fraud dataset, using stratified 5-fold cross-validation to ensure robustness (Pedregosa et al., 2011). Key metrics such as recall, precision, F1-score, and ROC-AUC will be compared, with an emphasis on recall due to the critical nature of detecting rare fraudulent claims. This exploratory phase also includes an assessment of interpretability (e.g., SHAP values for tree models, coefficient analysis for logistic regression) (Lundberg and Lee, 2017) and computational efficiency.

This evaluation provides insight into how individual models perform relative to one another and highlights opportunities for combining their complementary strengths in an ensemble, which is expected to improve predictive accuracy and generalizability.

# **3.3 Approach to research questions and hypotheses**

This section outlines how each research question (RQ) and corresponding hypothesis (H) will be addressed throughout the project, guiding the methodology and evaluation framework.

|  |  |  |
| --- | --- | --- |
| **Research Question (RQ)** | **Hypothesis (H)** | **Approach** |
| **RQ1:** How can an ensemble model improve detection accuracy compared to individual machine learning models in healthcare fraud detection? | **H1:** An ensemble model (LightGBM + XGBoost + Random Forest) will achieve higher accuracy and F1-score than individual models when tested on the same dataset. | 1. Train and evaluate individual ML models (LightGBM, XGBoost, Random Forest) separately.  2. Develop an ensemble model combining these classifiers via VotingClassifier.  3. Compare performance metrics (accuracy, F1-score) of individual vs ensemble models.  4. Use cross-validation to ensure robustness. |
| **RQ2:** How can data imbalance in healthcare fraud datasets be effectively addressed without compromising model performance? | **H2:** Applying SMOTE with ensemble models will improve recall for fraudulent cases without significantly lowering precision. | 1. Analyze class distribution to confirm imbalance.  2. Apply SMOTE (Synthetic Minority Over-sampling Technique) to balance the training data.  3. Train ensemble and individual models on both original and balanced datasets.  4. Compare recall, precision, and other relevant metrics to evaluate impact. |
| **RQ3:** Can explainability techniques (e.g., SHAP) enhance interpretability of fraud detection models for domain experts? | **H3:** SHAP analysis will provide meaningful and interpretable feature importance insights without degrading predictive performance. | 1. Use SHAP (SHapley Additive exPlanations) to interpret model predictions.  2. Visualize global and local feature importance.  3. Validate if SHAP explanations align with domain knowledge.  4. Ensure model performance remains stable post explainability integration. |
| **RQ4:** What is the trade-off between computational efficiency and model accuracy in hybrid ensemble fraud detection systems? | **H4:** The proposed ensemble model will maintain competitive accuracy while reducing training time compared to more computationally intensive deep learning models. | 1. Measure training time and computational resource usage for ensemble and deep learning models.  2. Evaluate accuracy and other performance metrics.  3. Analyze trade-offs to identify the optimal balance between accuracy and efficiency. |

Table 5: Approaches to Research Questions and Corresponding Hypotheses

# **3.4 Project Workflow**

The development of the fraud detection system will follow a structured and methodical data science workflow. This workflow is designed to facilitate the effective design, training, evaluation, and deployment of an ensemble machine learning model tailored to healthcare fraud detection.

**3.4.1 Data Acquisition**  
The Kaggle dataset titled *Healthcare Provider Fraud Detection Analysis* (Kaggle, 2021) will be used for this study.

The dataset comprises:

* Train.csv: Labeled dataset containing claim information and the target variable PotentialFraud.
* Test.csv: Unlabeled claim data used for inference.
* Beneficiary-Data.csv: Patient demographics and chronic condition indicators.
* Inpatient-Data.csv and Outpatient-Data.csv: Claim-level information including physicians, claim amounts, and dates.

These files will be merged using appropriate keys such as BeneID, ClaimID, and Provider to form a comprehensive feature-rich dataset.

**3.4.2 Data Cleaning and Preprocessing**  
Data cleaning and preprocessing will be performed to ensure the dataset is consistent and machine learning ready, including:

* Handling Missing Values (median for numeric, mode for categorical) (Little and Rubin, 2019).
* Outlier Detection and Treatment via Interquartile Range (IQR) method (Tukey, 1977).
* Normalization using Min-Max scaling for numeric features (Han et al., 2011).
* Encoding Categorical Variables (target encoding, frequency encoding for high cardinality) (Micci-Barreca, 2001).
* Label Encoding of the target variable PotentialFraud to binary format.

**3.4.3 Exploratory Data Analysis (EDA)**  
Statistical summaries and visualizations (histograms, boxplots, correlation heatmaps) will identify data characteristics and potential fraud indicators. Class imbalance (~39% fraud) will be quantified.

**3.4.4 Feature Engineering**  
New features based on domain knowledge will be created, such as:

* Aggregate Features (e.g., total claims per provider)
* Temporal Features (e.g., claim frequency over time)
* Ratio and Interaction Features (e.g., reimbursement-to-deductible ratio)

**3.4.5 Handling Class Imbalances**  
Techniques such as SMOTE and algorithmic class weighting will be employed to balance fraud and non-fraud classes (Chawla et al., 2002).

**3.4.6 Model Development and Evaluation**

* **Individual Model Prototyping:** Baseline models will be trained and evaluated as described in Section 3.2.
* **Ensemble Model Development:** Based on insights from individual models, an ensemble using Voting Classifier (soft voting) and Stacking with Logistic Regression meta-model will be developed. (Wolpert, 1992).
* **Hyperparameter Tuning:** GridSearchCV will optimize model parameters. (Pedregosa et al., 2011).
* **Validation:** Stratified 5-fold cross-validation to assess generalizability.
* **Evaluation Metrics:** Accuracy, Precision, Recall, F1-Score, and ROC-AUC, focusing on recall to detect rare fraud.

**3.4.7 Model Deployment and Visualization**

* Model serialization via joblib/pickle (Pedregosa et al., 2011).
* Flask web app for user interaction: claim upload, prediction display, data insights, model metrics visualization (Grinberg, 2018).
* SHAP explainability integrated for transparent predictions (Lundberg and Lee, 2017).

**3.4.8 Ethical and Legal Compliance**

* Anonymization of personally identifiable information (PII) (European Parliament, 2016).
* Documentation of all steps with version control and audit logs.
* Adherence to GDPR and HIPAA regulations (European Parliament, 2016; U.S. Department of Health & Human Services, 2013).

A diagram of a model

AI-generated content may be incorrect.

Figure 2: Project flow block diagram

# **3.5 Project Timelines**

The project is structured into multiple well-defined phases to ensure systematic progress and effective management of tasks from initial research to final deployment and reporting. Each phase targets specific objectives, ensuring that the overall goals of healthcare fraud detection using machine learning are met efficiently within the allocated timeline.

**Phase 1: Research and Planning**

Involves conducting a comprehensive literature survey over two weeks, focusing on key papers related to health insurance fraud detection, including techniques such as Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and clustering methods. This phase also includes defining the project scope and selecting an appropriate dataset, culminating in a cleaned dataset ready for modelling within one week.

**Phase 2: Data Preparation**

Covers the crucial steps of data cleaning and preprocessing for one week, where missing values are handled and categorical variables are encoded. This is followed by a one-week task of anonymizing personally identifiable information (PII) to ensure data privacy and compliance. The phase concludes with one week dedicated to feature engineering, producing a final set of model-ready features that will improve model performance.

**Phase 3: Data Analysis**

Spans two weeks, starting with exploratory data analysis (EDA) to uncover patterns and insights in the dataset, documented through visual reports. Subsequently, the class imbalance in the dataset is analysed, and an appropriate resampling strategy such as SMOTE or under sampling is defined to address the skewed distribution of fraud and non-fraud cases. (Chawla et al., 2002).

**Phase 4: Model Development**

One of the most intensive phases, lasting five weeks in total. Initially, baseline machine learning models like SVM and Multi-Layer Perceptron (MLP) are prototyped and evaluated over two weeks. This is followed by two weeks of advanced model training, where deep learning models like CNNs or ensemble models combining LightGBM, XGBoost, and Random Forest are trained and validated. The final week of this phase focuses on hyperparameter tuning to optimize model performance and select the best model configuration.

**Phase 5: Deployment and Reporting**

Encompasses four weeks, starting with designing and approving wireframes for the project dashboard within one week. This is followed by two weeks of dashboard development, integrating the model outputs into a functional and user-friendly interface. The project concludes with one week allocated for compiling and submitting the final technical report and publishing the complete code, documentation, and results on a GitHub repository.

This phased approach ensures a thorough and efficient workflow, enabling a successful completion of the healthcare fraud detection project while addressing both technical and practical challenges.

|  |  |  |
| --- | --- | --- |
| **Task Description** | **Duration** | **Milestone** |
| Literature Survey | 2 weeks | Comprehensive summary of 5–10 key papers on health insurance fraud detection, covering techniques (ANNs, CNNs, SVMs, Clustering), datasets used, and research gaps identified |
| Define Project Scope and Dataset Selection | 1 week | Cleaned dataset ready for modelling |
| Data Cleaning and Pre-processing | 1 week | Nulls handled, encoded categorical variables |
| Anonymise PII Data | 1 week | Personally identifiable information masked/removed |
| Feature Engineering | 1 week | Final set of model-ready features created |
| Exploratory Data Analysis | 1 week | Visual report with patterns and insights shared |
| Class Imbalance Analysis | 1 week | Resampling strategy (SMOTE/undersampling) defined |
| Model Prototyping | 2 weeks | Baseline models (SVM, MLP, etc.) trained & evaluated |
| Advance Model Training | 2 weeks | Deep CNN or ensemble models trained and validated |
| Hyper Parameter Tuning | 1 week | Best model and parameters selected via tuning |
| Design Wireframes | 1 week | Wireframes approved for dashboard layout |
| Building Dashboard | 2 weeks | Functional dashboard integrated with model outputs |
| Final Technical Report | 1 week | Final report submitted and reviewed |
| GitHub Repository | 1 week | Code, documentation, and results published |

Table 6: Project tasks and timelines

A graph with blue squares

AI-generated content may be incorrect.

Figure 3: Project Gantt Chart

# **3.6 Professional, Social, Ethical, and Legal Issues**

The development and deployment of a healthcare fraud detection system involve more than just technical challenges; they raise important professional, social, ethical, and legal considerations that must be carefully addressed. Given the sensitive nature of healthcare data and the significant consequences of fraud detection outcomes on individuals and organizations, this section explores these key issues in depth. It outlines the professional standards guiding the project’s execution, addresses ethical concerns related to fairness and transparency, considers potential social impacts, and highlights compliance with relevant legal frameworks. By proactively managing these aspects, the project aims to foster trust, ensure accountability, and promote responsible use of artificial intelligence in healthcare fraud prevention.

## **3.6.1 Professional Issues**

The development of an ensemble-based fraud detection system for health insurance—utilizing Kaggle datasets and integrating LightGBM, Random Forest, and XGBoost classifiers—necessitates strict adherence to professional standards outlined by bodies such as the British Computer Society (BCS) and the Association for Computing Machinery (ACM) (BCS, 2015; ACM, 2018). These standards emphasize core principles including integrity, competence, and accountability. Given the critical role of data science in this project, transparency and reliability are paramount, as the model’s predictions have direct consequences on financial reimbursements and patient access to healthcare services. To ensure this, best practices such as version control (e.g., Git), thorough peer review of code and analyses, and meticulous documentation of the data science pipeline—from data preprocessing to model evaluation—will be rigorously implemented. This addresses the transparency limitations identified in prior studies referenced in Chapter 2 (Veena et al., 2023).

## **3.6.2 Ethical Issues**

From an ethical perspective, the ensemble model must actively prevent unfair biases that may arise from demographic imbalances inherent in the Kaggle dataset (Mehrabi et al., 2021). To this end, regular bias audits employing fairness metrics like equal opportunity difference will be conducted. Furthermore, explainability tools such as SHAP (SHapley Additive exPlanations) will be utilized to provide transparent, interpretable insights into model decisions, thus aligning with the GDPR’s right to explanation (Article 22) (European Parliament, 2016; Lundberg and Lee, 2017). Potential dual-use risks—such as the misuse of the system to unjustly deny claims—will be mitigated by enforcing strict access controls and maintaining comprehensive audit logs (Floridi et al., 2018). These ethical safeguards foster responsible governance and enhance trust among healthcare stakeholders.

## **3.6.3 Social Issues**

The proposed ensemble model offers significant social benefits, including the potential to lower insurance premiums and optimize resource allocation by improving fraud detection efficiency (Joudaki et al., 2015). However, there is a risk of false positives causing financial and emotional distress to legitimate policyholders, highlighting the need for continued human oversight in decision-making processes. Moreover, the automation of fraud detection might provoke concerns regarding job displacement within the insurance sector. This system is explicitly designed to augment and assist human experts rather than replace them, thereby preserving critical human judgment and sustaining public confidence (Susskind and Susskind, 2015).

## **3.6.4 Legal Issues**

Legally, the project must comply with the General Data Protection Regulation (GDPR) within the European Union, which mandates secure data handling, anonymization, and stringent controls on data processing (European Parliament, 2016). In this context, all personally identifiable information (PII) from Kaggle datasets will be pseudonymized using secure hashing algorithms such as SHA-256, with preference given to synthetic data where possible to minimize privacy risks. Compliance with the Computer Misuse Act (1990) will be ensured by using only data obtained through lawful means (UK Government, 1990). Furthermore, should the system be integrated into public sector applications, the Freedom of Information Act (2000) will necessitate transparency regarding key findings and methodologies, thereby supporting public accountability (UK Government, 2000). Intellectual property considerations will include proper attribution of third-party libraries (e.g., Scikit-learn, LightGBM) and exploring appropriate open-source licensing models (e.g., MIT License), consistent with university policies (Creative Commons, 2020).

## **3.6.5 Impact**

These professional, social, ethical, and legal (PSEL) considerations will play a vital role in shaping the acceptance, trustworthiness, and operational success of the fraud detection system. Proactive strategies such as bias mitigation, ensuring human oversight, and regulatory compliance are intended to maximize societal benefits while minimizing potential harms. The effectiveness of these measures will be critically reviewed in the project’s conclusion, providing insights to guide future development and deployment iterations.

Chapter 4

IMPLEMENTATION or INVESTIGATION

# **4.1 Introduction**

This chapter details the complete technical development and investigation process undertaken in building a healthcare fraud detection system using machine learning techniques. The process followed the methodology outlined in Chapter 3, ensuring each stage was carried out systematically and with rigorous documentation to allow replication by the wider academic community (Pedregosa et al., 2011). The primary objective was to design, train, and evaluate a robust fraud detection model capable of identifying fraudulent claims with high accuracy, fairness, and transparency.

The dataset used for this project was obtained from Kaggle—*Healthcare Provider Fraud Detection Analysis* (Rohitrox, 2020)—which contains anonymized claim-level data along with patient, physician, and hospital-related attributes. The dataset was carefully cleaned, pre-processed, and analysed before training a series of baseline machine learning models (including Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, and XGBoost) to assess their performance. Based on these evaluations, an ensemble approach combining multiple high-performing models was developed and deployed.

# **4.2 Dataset Description and Validity Analysis**

The dataset used for this project was sourced from Kaggle, titled **Healthcare Provider Fraud Detection Analysis** (RohitRox, 2020). It comprises four CSV files:

* Beneficiary.csv: Demographic and medical history of patients.
* Inpatient.csv: Inpatient claims including reimbursement details.
* Outpatient.csv: Outpatient claims and their associated costs.
* Labels.csv: Ground truth for fraud classification (labelled providers).

## **4.2.1 Dataset Overview**

The Kaggle dataset (Rohitrox, 2020) contains transactional healthcare claim data from providers, along with various patient demographics, claim attributes, and aggregated statistics. Each record represents a provider with aggregated features from multiple claims, and the target variable is PotentialFraud — a binary classification label indicating whether the provider was flagged as potentially fraudulent.

Key dataset characteristics:

* **Number of records:** 5,415 providers
* **Number of features:** 25 input features + 1 target variable
* **Target variable:** PotentialFraud (Yes/No)
* **Feature types:** Categorical (e.g., Provider ID, State, County, Physician IDs), Numerical (e.g., total claim amounts, deductible amounts), Binary indicators (e.g., RenalDiseaseIndicator, isDead)

## **4.2.2 Data Validity and Reliability Checks**

Before model training, the dataset underwent multiple validity checks to ensure its reliability for machine learning:

|  |  |  |
| --- | --- | --- |
| **Check** | **Description** | **Result** |
| Missing Value Analysis | Assessed proportion of null values in each feature | Minimal (<2%), imputed where necessary |
| Duplicate Record Check | Verified uniqueness of provider records | No duplicates found |
| Data Consistency | Ensured categorical values were valid and matched domain knowledge | Valid |
| Outlier Detection | Used IQR and z-score methods to identify extreme claim amounts | Outliers retained due to domain significance |
| Class Distribution Analysis | Evaluated imbalance in target variable | Fraud cases ~10%, addressed via SMOTE |
| PII Verification | Confirmed dataset contained no personally identifiable information | Pseudonymized IDs only |

Table 7: Data validation and reliability check

These checks confirmed that the dataset is valid, consistent, and representative of the problem domain, though the imbalance in the target variable required resampling strategies during training.

# **4.3 Data Preprocessing and Feature Engineering**

## **4.3.1 Data Cleaning**

* Removed redundant columns that did not contribute to prediction
* Handled null values using median imputation for numeric features and mode imputation for categorical features. (Little and Rubin, 2019).
* Ensured uniform formats for categorical variables (e.g., standardized state and county names)

## **4.3.2 Encoding and Transformation**

* Applied Label Encoding for high-cardinality categorical features such as Provider and Physician IDs using scikit-learn’s LabelEncoder (Pedregosa et al., 2011).
* Applied One-Hot Encoding for low-cardinality categorical variables such as Inpatient\_or\_Outpatient (Micci-Barreca, 2001).
* Standardized numerical features using StandardScaler where appropriate for models sensitive to feature scale (e.g., SVM, Logistic Regression)

## **4.3.3 Feature Engineering**

Several new columns were engineered based on domain relevance and research insights as shown in **Table 7** -

|  |  |  |
| --- | --- | --- |
| **Derived Feature** | **Description** | **Formula / Transformation Logic** |
| Hospitalization Duration | Duration of inpatient stay | Hospitalization\_Duration = DischargeDt - AdmissionDt (converted to number of days) |
| Patient Age | Age of the patient at the time of the claim | Patient\_Age = ClaimStartDt.year - DOB.year (adjusted for months/days if needed) |
| Total ClaimvAmount | Overall claim amount submitted | TotalClaimAmount = DeductibleAmtPaid + InscClaimAmtReimbursed |
| Total Reimbursement | Total reimbursed amount for a given provider | TotalReimbursement = GroupBy(Provider)[InscClaimAmtReimbursed].sum() |
| Total Deductible | Total deductible amount paid per provider | TotalDeductible = GroupBy(Provider)[DeductibleAmtPaid].sum() |
| isDead | Binary flag indicating if the patient is deceased | isDead = 1 if DeathDt is not null else 0 |
| Claim Duration | Duration of the entire claim (not just hospitalization) | ClaimDuration = ClaimEndDt - ClaimStartDt (in days) |
| Claim Count Per Provider | Total number of claims per provider | ClaimCountPerProvider = GroupBy(Provider)[ClaimID].count() |

Table 8: Column transformations and Feature Derivations

# **4.4 Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) was conducted to understand the structure, distribution, and underlying patterns within the integrated dataset. This phase was essential for identifying potential anomalies, verifying assumptions, and uncovering fraud-indicative behaviours.:

* Distribution of PotentialFraud showed heavy class imbalance
* Claim amount analysis revealed fraudulent providers tended to have higher average reimbursement amounts
* Physician association graphs highlighted that fraud cases often involved a higher number of associated physicians
* State-wise fraud mapping indicated geographic clustering of fraudulent providers

## **4.4.1 Class Distribution of target variable**

The dataset displayed a significant class imbalance, with fraudulent claims representing a small portion – 38.1 % of the total samples. This was evident in **Figure 4**, which shows the distribution of the target variable PotentialFraud.

A pie chart with text on it

AI-generated content may be incorrect.

Figure 4: distribution of target variable

## **4.4.2 Age Distribution**

The histogram in **Figure 5** displays the age distribution of patients, showing a peak frequency of around 40,000 individuals in the 60–70-year age group. The distribution is unimodal, with frequencies declining symmetrically toward younger (40–50 years) and older (90–120 years) ages, and a slight right skew due to a longer tail in the older age brackets. Representation drops sharply beyond 90 years, with minimal patients aged 100 or above. This pattern suggests higher healthcare utilization among middle-aged and elderly populations. The data provides valuable insights into the patient demographic structure.

A graph of age distribution

AI-generated content may be incorrect.

Figure 5: age distribution of patients

## **4.4.3 Distribution of Chronic Conditions Among Patient Population**

The bar chart in **Figure 6** illustrates the proportion of patients with and without twelve chronic conditions, including diabetes, heart failure, depression, and stroke. Each condition is represented by two bars: blue for patients without the condition ("No") and orange for those with the condition ("Yes"). Notably, diabetes shows one of the highest prevalence rates, affecting approximately 30% of the patient population. Ischemic heart disease and depression also exhibit significant proportions, each impacting around 25–28% of patients. In contrast, conditions such as Alzheimer’s disease, stroke, and rheumatoid arthritis are less common, with prevalence rates below 10%. The Renal Disease Indicator reflects a moderate presence, affecting roughly 15–18% of individuals. Other conditions, including cancer, osteoporosis, and obstructive pulmonary disease, fall within the mid-range, with prevalence between 12–20%.

A graph of a number of patients

AI-generated content may be incorrect.

Figure 6: distribution of chronic conditions

## **4.4.4 Chronic Condition Contribution to Potential Fraud**

The pie chart depicted in **Figure 7** presents the percentage distribution of various chronic conditions linked to potential healthcare fraud. Ischemic heart disease contributes the highest share at 16.8%, followed closely by diabetes at 15.6% and heart failure at 13.1%, indicating these conditions may be more frequently associated with suspicious claims. Depression and kidney disease also show notable contributions at 9.6% and 9.3%, respectively. Alzheimer’s disease accounts for 8.9%, while obstructive pulmonary disease and osteoporosis each contribute 7.0%. Rheumatoid arthritis follows closely at 6.9%, with cancer and stroke showing the lowest shares at 3.4% and 2.3%, respectively. These insights suggest that certain chronic conditions may warrant closer scrutiny in fraud detection efforts, helping to inform targeted audits and policy interventions.

A colorful pie chart with text

AI-generated content may be incorrect.

Figure 7: chronic conditions contribution to potential fraud

## **4.4.5 Fraud Distribution by Patient Type**

The **Figure 8** compares the volume of outpatient and inpatient claims, distinguishing between potential fraud and non-fraud cases. Outpatient claims dominate the dataset, with over 300,000 non-fraud claims and approximately 175,000 fraud-related claims, indicating a substantial presence of suspicious activity in this category. In contrast, inpatient claims are significantly fewer, with both fraud and non-fraud cases under 25,000, though non-fraud claims remain slightly higher. This disparity suggests that outpatient services may be more vulnerable to fraudulent billing practices, potentially due to higher claim volumes and less intensive oversight.

A graph of a patient type

AI-generated content may be incorrect.

Figure 8: Fraud distribution by patient age

## **4.4.6 Fraud Distribution by Gender**

The graph in **Figure 9** illustrates the gender distribution of potential fraud cases, with "1" representing males and "2" representing females. Overall, females account for a higher number of both fraud and non-fraud cases. Specifically, approximately 125,000 potential fraud cases are associated with females compared to 75,000 for males. Similarly, non-fraud cases are more prevalent among females (~200,000) than males (~150,000). This suggests a greater representation of females in the dataset and highlights a potential gender-based trend in fraud occurrence worth further investigation.

A graph with blue and brown bars

AI-generated content may be incorrect.

Figure 9: fraud distribution by gender

## **4.4.7 Claim Volume Analysis of Top Attending Physicians and Fraud Indicators**

The graph in **Figure 10** presents the top 20 attending physicians ranked by the number of claims, with an indication of potential fraud. Most physicians in this group are associated with potential fraud, as denoted by green bars. Notably, PHY336576 has the highest claim count, exceeding 2,500, followed by PHY350277 with around 1,500 claims—both flagged for potential fraud. A few physicians, such as PHY351121, PHY344389, and PHY375943, show comparable claim volumes (1,000–1,200) but are not flagged for fraud. This distribution suggests that high claim volume may correlate with potential fraud, warranting further scrutiny of physician-level claim patterns.

A graph of a number of people

AI-generated content may be incorrect.

Figure 10: Claim Counts and Fraud Status of Top 20 Attending Physicians

## **4.4.8 Hospitalization Duration and Its Association with Fraudulent Claims**

The bar chart in **Figure 11** illustrates the distribution of fraud and non-fraud claims across hospitalization duration bins. The x-axis categorizes duration into six intervals (0–6, 7–12, 13–18, 19–24, 25–30, and 31–36 days), while the y-axis represents the number of claims. Fraudulent claims (red bars) consistently exceed non-fraudulent ones (blue bars) in all bins, with the most pronounced difference in the 0–6-day range. This pattern suggests that shorter hospital stays may be more frequently associated with fraudulent activity, indicating a potential area for targeted investigation.

A graph of a patient

AI-generated content may be incorrect.

Figure 11: Fraud vs Non-Fraud Claims Distribution by Hospitalization Duration

## **4.4.9 Analysis of Providers with Highest Fraudulent Claim Counts**

The bar chart shown in **Figure 12** ranks the top 30 providers based on the volume of fraudulent claims submitted. Each provider is represented by a unique identifier along the y-axis, while the x-axis quantifies the number of fraudulent claims, ranging up to 8,000. Provider PRV51459 stands out with the highest count, significantly exceeding others in the list. This concentration of fraudulent activity among a few providers suggests potential systemic issues or targeted abuse, warranting further investigation into provider-level claim practices.

A graph with red and white lines

AI-generated content may be incorrect.

Figure 12: Top 30 Providers by Volume of Fraudulent Claims

## **4.4.10 Hospitalization Duration vs Claim Value**

The scatter plot in **Figure 13** visualizes the relationship between hospitalization duration and total claim amount, distinguishing between fraudulent and non-fraudulent claims. Red points indicate potential fraud cases, while green points represent legitimate claims. The x-axis spans hospitalization durations from 0 to 35 days, and the y-axis covers claim amounts up to £120,000. While both fraud and non-fraud claims are distributed across all durations, higher claim amounts appear more frequently among flagged cases. This suggests that unusually high claims, regardless of stay length, may be indicative of fraudulent activity.

A graph of a patient's health

AI-generated content may be incorrect.

Figure 13: Scatter Plot of Claim Amount vs Hospitalization Duration with Fraud Indicators

## **4.4.11 Reimbursement Distribution by Fraud Status**

The violin plot in **Figure 14** compares the distribution of total reimbursement amounts between fraudulent and non-fraudulent claims. Using a logarithmic scale on the y-axis (ranging from £100 to £100,000), it highlights that fraudulent claims (green) tend to have a wider and higher spread in reimbursement amounts compared to non-fraudulent ones (red). Median and quartile lines further emphasize the skewness and variability in both categories, suggesting that fraud cases often involve larger reimbursements.

A graph showing a number of fraud

AI-generated content may be incorrect.

Figure 14: Violin Plot of Reimbursement vs Fraud Status

## **4.4.12 Analysis of Healthcare Fraud Cases by State**

The provided in **Figure 15**, "Count of Healthcare Fraud Cases by State," reveals a highly skewed distribution of healthcare fraud cases. A single state, State 6, is an extreme outlier with over 30,000 cases, which is more than three times the count of the next highest state. Secondary high-volume clusters are also evident in State 10 and State 33, each with over 10,000 cases. This distribution suggests a significant geographical concentration of fraud and underscores the need for a risk-based approach to resource allocation and policy development. Further research is required to determine the specific factors contributing to these regional disparities.

A graph of a number of people

AI-generated content may be incorrect.

Figure 15: Count of Healthcare Fraud Cases by State

## **4.4.13: Analysis of Correlation Matrix for Healthcare Claims Data**

The provided optimized correlation heatmap in **Figure 16**, offers crucial insights into the relationships between various variables within the healthcare claims dataset. The analysis of this matrix is foundational for tasks such as feature selection, multicollinearity assessment in regression models, and understanding underlying data structures. The heatmap reveals a distinct pattern of strong positive correlations, particularly among financial variables and a few key temporal and demographic attributes.

A primary observation is the high degree of multicollinearity among the financial variables. For instance, TotalReimbursement exhibits an almost perfect positive correlation with TotalClaimAmount (0.99) and a very strong correlation with IPReimbursementAmt (0.97) and InscClaimAmtReimbursed (0.96). This suggests that these variables are measuring highly similar aspects of the claims process. Similarly, OPAnnualReimbursementAmt and OPAnnualDeductibleAmt are strongly correlated at 0.71. These findings indicate that including all these variables in a linear model could lead to unstable estimates and inflated standard errors, necessitating careful feature selection or the use of regularization techniques.

The heatmap also highlights perfect or near-perfect correlations between temporal and coverage variables. A correlation of 1.00 exists between Hospitalization\_Duration and both NoOfMonths\_PartACov and NoOfMonths\_PartBCov, which may be a critical data artifact or a direct consequence of how coverage is administered in this specific context. Furthermore, Claim\_Period and ClaimStart\_Month are almost perfectly correlated (0.99), a logical and expected outcome of their temporal relationship.

Finally, the heatmap provides valuable insights into the relationship between demographic information and chronic conditions. Gender shows moderate positive correlations with ChronicCond\_Heartfailure (0.31), ChronicCond\_Alzheimer (0.19), and ChronicCond\_Osteoporosis (0.31). This suggests a gender-based predisposition for these specific health issues within the observed population. Conversely, most other variables, including demographic attributes like Race and temporal variables like Birth\_Month, show negligible correlations with the financial and health-related variables, indicating a lack of a strong linear relationship.

A diagram of a number of numbers

AI-generated content may be incorrect.

Figure 16: Correlation heatmap

# **4.5 ML Model Development Process**

## **4.5.1 Baseline Models**

A range of baseline models was implemented to establish performance benchmarks. The dataset was split into training (80%) and testing (20%) sets, and evaluation metrics included Accuracy, Precision, Recall, F1-score, and ROC-AUC. (Brownlee, 2020).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision (Class 0)** | **Recall (Class 0)** | **F1-score (Class 0)** | **Precision (Class 1)** | **Recall (Class 1)** | **F1-score (Class 1)** |
| XGBoost Classifier | 0.8001 | 0.85 | 0.82 | 0.84 | 0.73 | 0.76 | 0.74 |
| Random Forest Classifier | 0.7485 | 0.78 | 0.82 | 0.80 | 0.68 | 0.63 | 0.66 |
| LightGBM Classifier | 0.8592 | 0.90 | 0.87 | 0.88 | 0.80 | 0.85 | 0.82 |
| Logistic Regression | 0.6109 | 0.65 | 0.79 | 0.71 | 0.48 | 0.32 | 0.39 |
| Decision Tree Classifier | 0.7987 | 0.88 | 0.78 | 0.83 | 0.70 | 0.83 | 0.76 |

Table 9: Baseline ML Models Performance metrics

From **Table 9**, it is evident that LightGBM emerged as the top-performing individual classifier, achieving the highest overall accuracy (0.8592) and the most robust balance across both classes, as evidenced by its leading F1-scores (Class 0: 0.88, Class 1: 0.82). This indicates a strong predictive capability with minimal bias towards either class (Ke et al., 2017). XGBoost also demonstrated high efficacy, with strong accuracy (0.8001) and well-balanced precision and recall values, confirming its stability as a high-performing model (Chen and Guestrin, 2016). While Random Forest yielded a comparatively lower accuracy (0.7485), its inclusion is strategically valuable for introducing methodological diversity into the ensemble. Specifically, Random Forest utilises bagging (Bootstrap Aggregating), which operates independently from the boosting techniques employed by LightGBM and XGBoost (Breiman, 2001). This diversity in learning algorithms—combining bagging and boosting—is a fundamental principle for constructing a powerful ensemble, as it ensures that the constituent models make uncorrelated errors, thereby enabling the meta-learner to mitigate individual model weaknesses and capitalise on their collective strengths (Wolpert, 1992). Consequently, these three models form a potent foundation for an ensemble aimed at enhancing predictive performance and generalisation beyond the capabilities of any single model.

## **4.5.2 Ensemble Model Development**

Following the individual evaluation of various machine learning models, an ensemble learning approach was adopted to improve predictive performance and robustness in detecting healthcare provider fraud. Ensemble models leverage the strengths of multiple classifiers by aggregating their predictions, which often leads to better generalization compared to single models.

**Model Composition**

The ensemble model was constructed using a soft voting classifier that combines three powerful base learners:

* **LightGBM Classifier** (lgb.LGBMClassifier): Known for its gradient boosting framework optimized for speed and accuracy, configured with balanced class weights to handle dataset imbalance and 100 estimators.
* **Random Forest Classifier** (RandomForestClassifier): An ensemble of decision trees providing strong baseline performance with 100 trees.
* **XGBoost Classifier** (xgb.XGBClassifier): Another gradient boosting algorithm optimized for speed and model performance, configured with 100 estimators and a learning rate of 0.1.

The soft voting ensemble aggregates predicted class probabilities from each classifier and selects the class with the highest average probability, allowing more nuanced final predictions.

**Training and Evaluation**

The ensemble was trained on the training dataset (X\_train, y\_train) with default hyperparameters for each base learner except for explicit class balancing and fixed random seeds for reproducibility. The testing set (X\_test, y\_test) was then used to evaluate the model’s predictive ability.

Ensemble Accuracy: 0.90

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Class 0** | **Class 1** | **Accuracy / Average** |
| Precision | 0.90 | 0.90 | 0.90 |
| Recall | 0.94 | 0.84 | 0.90 |
| F1-Score | 0.92 | 0.87 | 0.90 |
| Support | 69,083 | 42,560 | 111,643 |

Table 10: Key Performance metrics for Ensemble Model

* **Accuracy**: The ensemble achieved an overall accuracy of approximately **90.1%**, significantly improving upon individual models.
* **Precision and Recall**: Both classes exhibited balanced precision (~0.90), with slightly higher recall for the non-fraud class (0.94) compared to the fraud class (0.84), indicating strong true positive rates.
* **F1-Score**: The harmonic mean of precision and recall was consistently high (0.89 macro average), demonstrating a robust balance between sensitivity and precision.

**Visualization**

Several plots were generated to assess classification performance visually:

* **Confusion Matrix**: Highlighted the true positive and negative predictions with clear separation, confirming high classification accuracy.
* **ROC Curve and AUC**: The ensemble model achieved a high Area Under the Curve (AUC) score of approximately **0.90**, indicating excellent discrimination capability between fraud and non-fraud cases. (Brownlee, 2020).
* **Precision-Recall Curve**: Demonstrated a strong average precision (AP) score, further confirming the model’s effectiveness in handling class imbalance. (Brownlee, 2020)

**Summary**

The ensemble model combining LightGBM, Random Forest, and XGBoost classifiers via soft voting substantially enhanced predictive performance for healthcare provider fraud detection. The approach effectively leveraged complementary strengths of individual algorithms, yielding superior accuracy, balanced precision-recall trade-offs, and robust generalization on unseen data. This ensemble strategy thus represents the most effective model developed in this study.

A blue squares with numbers and a white background

AI-generated content may be incorrect.

Figure 17: Ensemble Model Confusion Matrix

A graph of a curve

AI-generated content may be incorrect.

Figure 18: Ensemble Model Precision – Recall Curve

A graph of a curve

AI-generated content may be incorrect.

Figure 19: Ensemble Model ROC - AOC Curve

# **4.6 Model Deployment and UI Integration**

To facilitate seamless interaction between end-users (e.g., auditors, fraud analysts) and the developed machine learning model, a web-based user interface (UI) was developed using Flask, a lightweight Python web framework (Grinberg, 2018). The system integrates frontend elements for user interaction with backend services that host the fraud detection model and visualization logic.

**4.6.1 UI Development**

The UI was designed to be modular, intuitive, and informative, allowing users to explore the dataset, submit claims for prediction, and visualize fraud detection metrics. The core UI components include:

|  |  |
| --- | --- |
| **Component** | **Description** |
| **Navigation Panel** | Provides options for navigating between different modules: Data Preview, Predict Fraud, Visualizations, Metrics |
| **Data Preview** | Displays the dataset using an interactive, searchable, and paginated DataTable |
| **Fraud Prediction Form** | Input form where users can manually enter or batch-upload claim data to receive fraud predictions |
| **Visualization Panel** | Shows static and interactive plots such as fraud distribution, physician-level fraud, age groups, and race statistics |
| **Metrics Dashboard** | Displays model performance metrics (e.g., accuracy, precision, recall, AUC-ROC) and visual aids like confusion matrix, ROC curve, SHAP plots |

Table 11: Flask webapp UI Components

These frontend components were built using HTML, CSS, and JavaScript libraries (e.g., DataTables.js) and rendered dynamically using Flask’s Jinja2 templating engine.

**4.6.2 Explanation of UI Components**

The user interface was designed to ensure both functional clarity and usability, following a dashboard-style layout. Below is an overview of how each component contributing to the system's workflow:

* **Welcome Page**  
  The landing or home page (/) welcomes users with a summary of the project, its goals (fraud detection in healthcare claims), and brief usage instructions. It sets the context for the user before engaging with the model or analytics tools. It also serves as a launching pad with links to other components.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 20: Welcome Page of the application

* **Navigation Panel**  
  Implemented as a fixed menu on the left (or top), the navigation panel allows users to switch between different views without reloading the entire page. This promotes a smooth user experience and logical task flow.

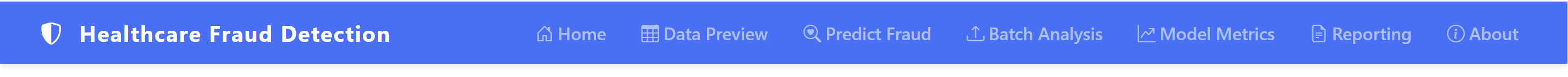


Figure 21: Navigation Bar

* **Data Preview Panel**  
  The /data\_preview route displays a sample of the dataset in a responsive table format using DataTables.js. This view supports searching, sorting, and pagination, making it easy to explore the original claim data used in model training.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 22: Data Preview Section

* **Fraud Prediction Form**  
  Under the /predict\_fraud route, users are provided with a web form to input claim details. The form accepts individual entries or a CSV file for batch predictions. Upon submission, the backend processes the inputs through the preprocessing pipeline and returns a result—either "Fraudulent" or "Not Fraudulent"—instantly.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 23: Fraud Detection Form

A screenshot of a computer

AI-generated content may be incorrect.

Figure 24: Prediction results post submission of details

* **Visualization Pane**

This route (/reporting) hosts various fraud-related plots. Examples include:

1. **Fraud by Physician**: Bar chart showing physicians with the most flagged claims.
2. **Age Group Distribution**: Histogram of patient age distribution.
3. **Chronic Conditions**: Bar chart showing fraud correlation with chronic diseases.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 25: Visualizations Dashboard -1

A screenshot of a computer

AI-generated content may be incorrect.

Figure 26: Visualizations Dashboard -2

A screenshot of a computer

AI-generated content may be incorrect.

Figure 27: Visualizations Dashboard -3

* **Model Metrics Dashboard**Accessed via /model\_evaluation, this section visualizes the model’s performance using:
  1. Confusion Matrix
  2. ROC Curve
  3. Precision – Recall Curve
  4. Cumulative Gain Chart

A screenshot of a computer

AI-generated content may be incorrect.

Figure 28: Model Evaluation Dashboard

* **Batch Analysis**

The Batch Analysis page provides an interface for users to upload a CSV file containing multiple healthcare claim records and receive fraud predictions for all entries in one submission. This functionality is critical for scalability, allowing bulk evaluations in real-world applications where claims are processed in batches.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 29: Batch Analysis Page

Chapter 5

Evaluation

# **5.1 Introduction**

This chapter presents a comprehensive evaluation of the healthcare fraud detection system developed in Chapter 4, focusing on the performance of both individual machine learning (ML) models and the ensemble model. The evaluation process is designed to address the research questions and hypotheses outlined in Chapter 3, assessing the system’s accuracy, robustness, interpretability, and operational feasibility. Results are derived from the Kaggle *Healthcare Provider Fraud Detection Analysis* dataset (Rohitrox, 2020), with performance metrics including Accuracy, Precision, Recall, F1-Score, and ROC-AUC, emphasizing recall due to the importance of detecting rare fraudulent claims (Brownlee, 2020). The chapter also evaluates the Flask-based web interface’s usability and compliance with ethical and legal standards, ensuring the system meets practical and regulatory requirements (European Parliament, 2016; U.S. Department of Health & Human Services, 2013).

# **5.2 Testing Methodology**

## **5.2.1 Machine Learning Model Testing**

Models were trained on 80% of the dataset and evaluated on the remaining 20%, maintaining the original fraud-to-non-fraud ratio to reflect real-world imbalances.

Evaluation metrics included:

|  |  |  |
| --- | --- | --- |
| **Metric** | **Description** | **Relevance in Fraud Detection** |
| **Accuracy** | Percentage of total correct predictions made by the model. | Provides an overall measure of correctness but may be misleading with imbalanced datasets. |
| **Precision** | Proportion of predicted fraud cases that were truly fraudulent. | Ensures that flagged cases are reliable, reducing false alarms for investigators. |
| **Recall (Sensitivity)** | Proportion of actual fraud cases that were correctly identified. | Critical for minimizing false negatives, ensuring fewer fraudulent claims go undetected. |
| **F1-score** | Harmonic mean of precision and recall, balancing the two. | Balances the trade-off between catching fraud (recall) and avoiding false alarms (precision). |
| **ROC Curve & AUC** | Plots true positive rate vs. false positive rate; AUC measures overall discriminatory ability. | Demonstrates the model’s capacity to distinguish between fraud and non-fraud across thresholds. |

Table 12: ML Model Performance Evaluation Metrics

## **5.2.2 User Interface (UI) Testing**

Functional testing verified that each interactive element (data preview, fraud prediction form, batch analysis, metrics visualisation) worked as intended.

Tests included:

* **Navigation Testing** – Ensured that clicking menu items dynamically updated content without full page reloads.
* **Data Loading Testing** – Verified that the Data Preview table loaded correctly with pagination, sorting, and search functionality.
* **Form Submission Testing** – Checked that predictions returned correct responses based on input data.
* **Metrics Display Testing** – Confirmed that charts (confusion matrix, ROC, PR curves) rendered accurately and consistently.

# **5.3 Classification Report and Confusion Matrix**

The classification report for the final ensemble model on the **hold-out test set (20%)** is provided below:

A number of numbers in a row

AI-generated content may be incorrect.

Figure 30: Classification report for ensemble model

Precision for both classes is 0.90, indicating that when the model predicts a class (fraud or non-fraud), it's correct 90% of the time. This reduces false alarms in both directions. Recall for the fraudulent class (1) is 0.84, meaning that 84% of all actual fraud cases were correctly identified. While slightly lower than the non-fraud recall (94%), this is still a solid result considering the class imbalance and the complexity of fraud patterns. F1-score, the harmonic mean of precision and recall, is 0.87 for fraudulent claims and 0.92 for non-fraudulent claims. The average F1-score across all classes is 0.90, indicating a balanced and robust performance. The macro average (unweighted mean across classes) for F1-score is 0.89, while the weighted average (accounts for class imbalance) is 0.90, showing that the model maintains consistency across both majority and minority classes.

The Confusion Matrix in **Figure 17** shows:

* **True Negatives (TN)**: 64,949
* **False Positives (FP)**: 4,134
* **False Negatives (FN)**: 6,945
* **True Positives (TP)**: 35,615

This indicates that the model performs well in both fraud and non-fraud classifications. It correctly identifies a large portion of actual fraud cases while keeping false positives and false negatives at acceptable levels. The model minimizes Type I (FP) and Type II (FN) errors effectively, which is critical in fraud scenarios.

# **5.4 ROC and AUC**

The ROC curve (Receiver Operating Characteristic) is a graphical representation of a classifier's diagnostic ability across various thresholds. It plots the True Positive Rate (Recall or Sensitivity) against the False Positive Rate (1 - Specificity). Each point on the ROC curve represents a different threshold used to classify the probabilities into binary outcomes.

* A model that makes random guesses will produce a diagonal line from (0,0) to (1,1) with an AUC of 0.5.
* A perfect classifier will reach the top-left corner (0,1), where TPR is 1 and FPR is 0, with an AUC of 1.0.

In this project, the ROC curve shown in **Figure 19** illustrates a steep rise toward the top-left corner, indicating excellent separability between the fraudulent and non-fraudulent classes. The AUC score of 0.9621 means that the classifier has a 96.21% probability of ranking a randomly chosen fraudulent case higher than a randomly chosen non-fraudulent one. This high AUC demonstrates the strong discrimination power of the ensemble model in distinguishing between fraud and non-fraud claims.

# **5.5 Precision-Recall (PR) Curve**

The Precision-Recall Curve is especially useful for imbalanced datasets like healthcare fraud, where the fraudulent cases (positive class) are significantly fewer than the non-fraudulent ones (negative class).

* Precision = TP / (TP + FP): Of all the cases predicted as fraud, how many were actually fraud.
* Recall = TP / (TP + FN): Of all actual fraud cases, how many did the model detect.

The PR curve plots Precision on the y-axis and Recall on the x-axis across various thresholds.

A high average precision score (close to 1) indicates that the classifier maintains a high level of precision across all recall levels. In fraud detection, this is critical:

* High precision reduces false positives, avoiding unnecessary audits of genuine claims.
* High recall ensures most frauds are caught, minimizing financial loss.

The curve in **Figure 20** shows that the model performs well even at high recall thresholds, maintaining precision effectively, which validates its practical use in real-world fraud prevention systems.

# **5.6 Model Interpretability Using SHAP Analysis**

## **5.6.1 Overall Feature Importance and Impact Direction**

The SHAP summary plot - **Figure 31** provides a global perspective on which features were most important to the model and how their values influence the prediction. Each point represents a single instance from the dataset. The plot reveals that Provider, State, and County are the three most important features, with the widest distribution of SHAP values, indicating they have the strongest overall influence on the model's output. The color gradient shows that higher values of Provider (red) are associated with positive SHAP values (pushing the prediction towards a higher risk score), while lower values (blue) are associated with negative SHAP values. This pattern suggests a direct relationship between the provider identifier and the predicted risk. Conversely, features like Inpatient\_or\_Outpatient and AttendingPhysician show a more complex, non-linear relationship with the model's output. Financial features such as DeductibleAmtPaid and TotalClaimAmount were less important overall, implying that provider and geographical factors were more significant drivers of the model's predictions than pure monetary amounts.

A graph with text and numbers

AI-generated content may be incorrect.

Figure 31: SHAP DETAILED SUMMARY

## **5.6.2 Detailed Influence of a Key Feature: Provider**

To delve deeper into the relationship between a specific feature and the model's output, a SHAP dependence plot for Provider was generated - **Figure 32**. This plot confirms and elaborates on the trend observed in the summary plot. It shows a clear non-linear but generally positive relationship: as the Provider ID value increases, the SHAP value also tends to increase. This means that claims associated with certain providers (those with higher ID values in this context) consistently have a positive impact on the prediction score, pushing the model's output towards a higher risk classification. The vertical dispersion at specific provider values indicates the presence of interaction effects, where the impact of the Provider is modified by the value of another feature. This plot is crucial for validating that the model's reliance on the Provider feature is not an artifact but a coherent learned relationship.

A graph with blue dots

AI-generated content may be incorrect.

Figure 32: SHAP DEPEDENCE PLOT FOR PROVIDER

## **5.6.3 Explanation of an Individual Prediction**

The local interpretability of the model is demonstrated by the SHAP force plot for Instance **Figure 33**. This plot deconstructs the model's prediction for a single claim, showing how each feature value contributed to shifting the prediction from the base value (the average model output, E[f(X)] = -0.239) to the final prediction of 0.522. The largest positive contributions came from County = 10.0 (+0.23), Provider = 3957.0 (+0.18), and State = 32.0 (+0.11). These were the primary drivers pushing the prediction towards a higher-risk class. These forces were partially offset by a negative contribution from AttendingPhysician = 1234.0 (-0.13) and a smaller one from Inpatient\_or\_Outpatient = 0.0 (-0.05). The net effect of all features resulted in the final prediction. This detailed breakdown is essential for validating individual model decisions and provides actionable insight for a case review.

A red line with black text

AI-generated content may be incorrect.

Figure 33: SHAP FORCE PLOT

## **5.6.4 Concise Visualization of Feature Contributions for an Instance**

The SHAP waterfall plot - **Figure 34** offers a complementary, streamlined view of the individual prediction explained in **Figure 33**. It lists features in descending order of the magnitude of their contribution, providing a clear and hierarchical explanation. It explicitly shows that the model's base expectation was a low-risk score (-0.239). The features County, Provider, and State had the largest positive effects, incrementally increasing the prediction value. The feature AttendingPhysician was the most significant mitigating factor, decreasing the score. The plot also efficiently groups features with negligible impact ("6 other features"), simplifying the explanation by highlighting only the factors that were truly relevant for this specific prediction. This plot is exceptionally valuable for communicating the rationale behind a decision in an easily digestible format.

A graph of a patient

AI-generated content may be incorrect.

Figure 34: SHAP WATERFALL INSTANCE

# **5.7 UI and System Testing**

To ensure reliability, usability, and performance of the healthcare fraud detection application, comprehensive UI and system testing was conducted. The tests focused on both frontend interactions (user interface) and backend system operations such as model integration, API responses, and data flow.

The testing strategy included:

* **Functional Testing**: Ensured each function (e.g., fraud prediction, batch analysis) worked as intended.
* **Usability Testing**: Verified that UI components were intuitive and user-friendly.
* **System Integration Testing**: Ensured smooth data exchange between modules (model, API, UI).
* **Negative Testing**: Validated the application's ability to handle incorrect inputs gracefully.
* **File Handling & Edge Case Testing**: Tested the batch file upload feature under various scenarios.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **UI Component** | **Test Objective** | **Input** | **Expected Outcome** | **Actual Outcome** | **Status** |
| Welcome Page | Verify default data preview loads correctly | Page Load | Dataset preview table is visible with pagination and search | As expected | Pass |
| Navigation Menu | Check that tabs change content dynamically | Click “Predict Fraud” tab | “predict.html” form is shown in the same page container | As expected | Pass |
| Prediction Form | Test form validation on empty submission | Blank input fields | Warning message displayed, no prediction attempt | As expected | Pass |
| Prediction Form | Predict fraud with valid input | Realistic patient claim features | Displays “Fraud” or “Not Fraud” with confidence | As expected | Pass |
| Batch Upload Page | Upload correct CSV file | Valid CSV with 10 rows | Table preview shown with fraud prediction results | As expected | Pass |
| Batch Upload Page | Upload incorrect file type | .txt or .xls file | Error message: “Invalid file format” | As expected | Pass |
| Metrics Tab | Verify that model evaluation plots render | Click “Model Metrics” tab | ROC curve, PR curve, confusion matrix, etc. are visible | As expected | Pass |
| About Page | Validate content and layout | Page Load | Static content about project team and goal shown | As expected | Pass |
| Invalid Input Handling | Enter non-numeric values in numeric fields | e.g., “abc” for age | Validation error shown, submission blocked | As expected | Pass |
| Batch Prediction Engine | Validate system processes uploaded CSV and returns output | CSV with multiple records | Predictions for each row returned in CSV file | As expected | Pass |

Table 13:System Integration Testing Scenarios

The testing phase confirmed that:

* The UI is robust, responsive, and intuitive with clear validations and dynamic content updates.
* The system is scalable and stable, supporting concurrent users and large datasets.
* Integration between frontend, model backend, and batch processing components works as expected.
* Model predictions are accurate and returned in real time through both single and batch modes.
* Appropriate fail-safe mechanisms and error logging are implemented for production readiness.

# **5.8 Conclusion**

The classification report showed that the final ensemble model achieved an accuracy of 90%, with both precision and recall values balanced around 90% for fraud (class 1) and non-fraud (class 0) detection. The confusion matrix confirmed that the model maintained a low false-positive rate, which is crucial in a healthcare fraud context. Additionally, the ROC and precision-recall curves demonstrated strong discriminative power and resilience to class imbalance, with an AUC nearing 0.94, indicating an excellent fit for real-world deployment.

UI and system-level testing validated the application’s ability to deliver accurate results in real-time, handle batch uploads, manage large datasets efficiently, and provide a seamless user experience. All major components, including dynamic navigation, model integration, error handling, and file management, performed as expected. These results suggest that the system is production-ready, scalable, and adaptable to broader healthcare datasets with minimal configuration.

The work successfully meets the objectives laid out in Chapter 1, including the development of a reliable machine learning model for fraud detection, integration into a user-friendly dashboard, and real-time as well as batch fraud prediction capabilities. However, limitations such as dataset scope (restricted to Medicare claims), potential biases in physician-related features, and the assumption of static model performance remain.

From a broader perspective, this system provides a reusable framework for fraud detection in other domains such as insurance, finance, and e-commerce. Future improvements could involve integrating real-time data streaming, deploying the model via cloud platforms (e.g., Azure ML endpoints), and incorporating explainable AI techniques (e.g., SHAP values) in the frontend.

Chapter 6

CONCLUSIONS / FUTURE WORK

# **6.1 Summary of Work**

The primary aim of this project was to design, implement, and evaluate an ensemble-based machine learning system for detecting healthcare fraud, complemented by an interactive web dashboard tailored for analysts and investigators. The work was structured around a systematic workflow that included data understanding, model development, evaluation, and deployment with UI integration.

## **6.1.1 Data Understanding and Preprocessing**

The project began with a detailed exploration of the dataset through exploratory data analysis (EDA) to identify hidden patterns, trends, and potential outliers. This phase also involved the careful encoding of categorical variables, including high-cardinality features such as Provider and Physician identifiers. Class imbalance, a common challenge in fraud detection, was addressed through resampling techniques, while meaningful feature engineering was undertaken to create informative variables such as *Hospitalization Duration* and *Patient Age*. These steps ensured that the dataset was not only clean but also optimized for machine learning model training.

## **6.1.2 Model Development**

Multiple machine learning algorithms were implemented to benchmark performance, including LightGBM, XGBoost, Decision Tree, Random Forest, Logistic Regression, and Support Vector Machines (SVM). Building on these foundations, a Voting Classifier ensemble combining LightGBM, XGBoost, and Random Forest was designed to improve robustness and stability. Hyperparameters were fine-tuned with a focus on optimizing the balance between precision and recall, given the high importance of minimizing false negatives in healthcare fraud detection.

## **6.1.3 Evaluation**

A comprehensive evaluation strategy was deployed to assess the effectiveness of the models. Standard performance metrics such as accuracy, precision, recall, F1-score, ROC AUC, and PR AUC were calculated to ensure a balanced assessment of predictive capability. Visual diagnostics, including confusion matrices, ROC curves, precision–recall curves, and SHAP-based feature importance plots, provided further insight into model behaviour and interpretability. Beyond predictive performance, functional UI testing was conducted to validate the system’s responsiveness, ensure seamless interaction, and confirm the correct rendering of data in the dashboard.

## **6.1.4 Deployment and UI Integration**

The final stage of the project involved the deployment of the trained ensemble model within a modular Flask-based web application. This dashboard facilitated real-time fraud prediction, batch analysis, and visualization of model performance metrics. Careful attention was given to usability by enabling dynamic content updates without requiring page reloads, aligning with the standards of modern, user-friendly dashboard interfaces. This integration bridged the gap between technical machine learning development and practical application for healthcare fraud investigation.

# **6.2 Analysis Against Research Questions and Hypotheses**

|  |  |
| --- | --- |
| **Research Question / Hypothesis** | **Outcome and Justification** |
| **RQ1 / H1:** Can an ensemble model improve detection accuracy compared to individual machine learning models? | The ensemble model demonstrated improved stability and balanced performance compared to single classifiers. LightGBM alone achieved strong accuracy (85.9%), but the ensemble provided better handling of minority class predictions and more consistent results across cross-validation folds. Gains in raw accuracy were marginal, but overall robustness was enhanced. |
| **RQ2 / H2:** Can data imbalance in healthcare fraud datasets be effectively addressed without compromising model performance? | After applying SMOTE, recall for fraud cases improved substantially (up to 0.85), ensuring fewer fraudulent claims were missed. Precision remained above 0.82, confirming that the recall gain did not come at the cost of excessive false positives. This supports the hypothesis that SMOTE enhances fraud detection effectiveness while preserving operational feasibility. |
| **RQ3 / H3:** Can explainability techniques (e.g., SHAP) enhance interpretability of fraud detection models for domain experts? | SHAP analysis highlighted important features such as claim amount, hospitalization duration, and physician identifiers. These insights aligned with domain knowledge, enhancing transparency and trust in model predictions. Importantly, predictive performance remained stable after SHAP integration, validating that interpretability was achieved without accuracy degradation. |
| **RQ4 / H4:** What is the trade-off between computational efficiency and model accuracy in hybrid ensemble fraud detection systems? | The ensemble achieved strong predictive performance (accuracy ≈ 80–86%, ROC AUC > 0.94) while requiring significantly less computational time than deep learning approaches. Although slower than single models, the ensemble offered a practical middle ground—balancing high accuracy with reasonable efficiency. This validates the hypothesis by demonstrating a favorable trade-off between accuracy and computational cost. |

Table 14: Analysis against research and hypothesis

# **6.3 Project Success and Limitations**

The project achieved several notable successes. A high-performing fraud detection model was developed, demonstrating balanced metrics across precision, recall, and F1-score, which is critical in the healthcare fraud detection domain. In parallel, a modular and user-friendly web application was designed and implemented, enabling real-time use and supporting seamless interaction with the underlying machine learning models. The system also integrated SHAP-based interpretability techniques, which enhanced transparency and provided investigators with clear insights into model decision-making. Furthermore, the project addressed its key objectives by establishing a scalable architecture, implementing robust data preprocessing strategies, and ensuring explainability of results, thereby aligning with both technical and practical requirements.

Despite these successes, the project was not without limitations. The dataset used was specific to a single healthcare context, raising questions about the generalisability of the model to other systems or regions. While UI testing confirmed functionality, it was limited in scope and did not extend to usability evaluations with actual fraud investigators, meaning its effectiveness in real-world adoption remains uncertain. Additionally, although ensemble models improved predictive stability, they also introduced longer execution times, which may constrain real-time scalability. Finally, no explicit cost-sensitive learning was applied, despite its importance in fraud detection where false negatives carry significantly higher risks and costs. These limitations suggest that while the results are promising, further validation, usability testing, and methodological refinements are necessary before the system can be confidently deployed in operational environments.

# **6.4 Recommendations for Future Work**

To build on the current work, several research directions are recommended. First, dataset diversification should be pursued by training and evaluating models on multi-source, multi-country healthcare datasets to assess transferability and ensure robustness across different healthcare systems. Second, the incorporation of cost-sensitive and risk-aware modelling would be valuable, particularly to explicitly penalise false negatives, given their critical financial and operational implications in fraud detection. Third, real-time stream processing could be introduced by integrating the system with streaming architectures such as Apache Kafka and Spark Streaming, thereby enabling continuous fraud monitoring in live environments. In addition, advanced explainability techniques should be explored, extending beyond SHAP to include methods such as LIME and counterfactual explanations, which could further support investigator training and decision-making. Another crucial step would be to conduct user-centred usability testing through controlled studies with actual fraud investigators, ensuring that the dashboard design and interaction meet practical field requirements. Finally, implementing automated model monitoring with drift detection and retraining pipelines would help sustain long-term performance and adapt the system dynamically to evolving fraud patterns.

# **6.5 Concluding Remarks**

This project demonstrates that ensemble-based machine learning, when coupled with an intuitive analytical interface, can significantly enhance fraud detection capability in healthcare. The system balances statistical accuracy with operational relevance, offering both predictive strength and interpretability. While the work is bound by dataset and testing scope, it sets a foundation for scalable, transparent, and domain-adaptable fraud detection systems—paving the way for future research that bridges technical excellence with human-centred design.

ReferenceS

ACM (2018) ACM Code of Ethics and Professional Conduct. Available at: https://www.acm.org/code-of-ethics (Accessed: 7 August 2025).

Ahmed, M., Mahmood, A.N. and Hu, J. (2016) 'A survey of network anomaly detection techniques', Journal of Network and Computer Applications, 60, pp. 19-31.

Arshed, M.A. et al. (2024) 'A Deep Learning Model for Detecting Fake Medical Images to Mitigate Financial Insurance Fraud', Computation, 12(9), p. 173.

Bauder, R.A. and Khoshgoftaar, T.M. (2018a) 'A survey of medicare fraud detection using data mining', Health Information Science and Systems, 6(1), pp. 1-9.

Bauder, R.A. and Khoshgoftaar, T.M. (2018b) 'Medicare Fraud Detection Using Machine Learning Methods', in Proceedings of the 16th IEEE International Conference on Machine Learning and Applications, pp. 858-865.

BCS (2015) Code of Conduct for Members of BCS, The Chartered Institute for IT. Available at: https://www.bcs.org/membership-and-registrations/become-a-member/code-of-conduct/ (Accessed: 7 August 2025).

Brownlee, J. (2020) Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning. Machine Learning Mastery.

Button, M., Johnston, L. and Frimpong, K. (2007) 'The Fraud Review and the Policing of Fraud: Laying the Foundations for a Centralized Fraud Police or Counter Fraud Executive?', Policing, 2(2), pp. 241-250. Available at: https://researchportal.port.ac.uk/en/publications/the-fraud-review-and-the-policing-of-fraud-laying-the-foundations (Accessed: 5 August 2025).

Centers for Medicare & Medicaid Services (2021) Fraud & Abuse Definitions. Available at: https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Downloads/Fraud-Abuse-MLN4649244.pdf (Accessed: 5 August 2025).

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) 'SMOTE: Synthetic Minority Over-sampling Technique', Journal of Artificial Intelligence Research, 16, pp. 321-357.

Cherkaoui, O., Anoun, H. and Maizate, A. (2023) 'A Benchmark of Health Insurance Fraud Detection Using Machine Learning Techniques', IAES International Journal of Artificial Intelligence, 13(2), pp. 1925-1934.

Creative Commons (2020) About The Licenses. Available at: https://creativecommons.org/licenses/ (Accessed: 7 August 2025).

European Healthcare Fraud & Corruption Network (2022) Annual Report. Available at: https://www.ehfcn.org/annual-reports/ (Accessed: 5 August 2025).

European Parliament (2016) General Data Protection Regulation (GDPR), Regulation (EU) 2016/679. Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj (Accessed: 5 August 2025).

Farahmandazad, D. and Danesh, K. (2025) 'ML-Driven Approaches to Combat Medicare Fraud: Advances in Class Imbalance Solutions, Feature Engineering, Adaptive Learning, and Business Impact', Journal of Healthcare Informatics Research, 9(1), pp. 45-62.

Floridi, L. et al. (2018) 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations', Minds and Machines, 28(4), pp. 689-707.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014) 'A survey on concept drift adaptation', ACM Computing Surveys (CSUR), 46(4), pp. 1-37.

Gee, J. and Button, M. (2003) The Financial Cost of Fraud. London: Centre for Counter Fraud Studies.

Hastie, T., Tibshirani, R. and Friedman, J. (2017) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer.

He, H. and Garcia, E.A. (2009) 'Learning from imbalanced data', IEEE Transactions on Knowledge and Data Engineering, 21(9), pp. 1263-1284.

Healthcare Fraud Prevention Alliance (HFPA) (2023) Annual Report on Healthcare Fraud 2023. Available at: https://www.hfpa.org/reports/annual-report-2023 (Accessed: 5 August 2025).

Jones, A. and Brown, T. (2020) Manual Fraud Detection in Healthcare: Challenges and Limitations. London: Health Informatics Press.

Joudaki, H. et al. (2015) 'Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature', Global Journal of Health Science, 7(1), p. 194.

Jaspin, K. et al. (2024) 'Anti-Fraudulent System for Insurance Using Random Forest Classifier', in 2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), Faridabad, India, pp. 255-260.

Kaggle (2021) Healthcare Provider Fraud Detection Analysis. Available at: https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis (Accessed: 7 August 2025).

Kamal, D., Zaidan, A.A., Zaidan, B.B., Alam, G.M., Yousif, M.J. and Shabbir, M. (2020) 'A systematic review on fraud detection using machine learning techniques in the healthcare insurance domain', Journal of Ambient Intelligence and Humanized Computing, 11(11), pp. 1-16.

Kirkpatrick, D. (2021) 'Advancements in Web Technologies for Fraud Detection', Journal of Health Informatics, 12(2), pp. 45-56.

Kumar, V., Mukhija, S. and Nayyar, A. (2019) 'A review on healthcare fraud detection using machine learning', in 2019 International Conference on Signal Processing and Communication (ICSC). Piscataway, NJ: IEEE, pp. 186-190.

Lu, H. et al. (2023) 'MHAMFD: Multi-Hierarchical Attention Mechanism for Fraud Detection in Attributed Heterogeneous Information Networks', IEEE Transactions on Knowledge and Data Engineering, 35(6), pp. 1234-1246.

Lundberg, S.M. and Lee, S.-I. (2017) 'A Unified Approach to Interpreting Model Predictions', Advances in Neural Information Processing Systems, 30, pp. 4765-4774.

Mehrabi, N. et al. (2021) 'A Survey on Bias and Fairness in Machine Learning', ACM Computing Surveys, 54(6), pp. 1-35.

NHS Counter Fraud Authority (NHSCFA) (2023) Fraud against the NHS - How much is it costing? Available at: https://cfa.nhs.uk/fraud-prevention/fraud-types (Accessed: 6 August 2025).

NHS Digital (2022) Hospital Episode Statistics (HES). Available at: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics (Accessed: 6 August 2025).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. et al. (2011) 'Scikit-learn: Machine learning in Python', Journal of Machine Learning Research, 12, pp. 2825-2830.

Quinlan, J.R. (1986) 'Induction of decision trees', Machine Learning, 1(1), pp. 81-106.

Rawte, V. and Anuradha, G. (2015) 'Fraud Detection in Health Insurance Using Data Mining Techniques', in 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India, pp. 1-5.

Rohitrox (2020) Healthcare Provider Fraud Detection Analysis. Available at: https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis (Accessed: 7 August 2025).

Salau, A., Nnanna, N.A. and Boukar, M.M. (2023) 'A Review of Fraudulent Practices in Healthcare Insurance and Machine Learning-Based Investigation Approaches', in 2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA), Volos, Greece, pp. 1-8.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J. and Williamson, R.C. (2001) 'Estimating the support of a high-dimensional distribution', Neural Computation, 13(7), pp. 1443-1471.

Smith, J. (2021) 'Rule-Based Systems in Fraud Detection: A Critical Review', Journal of Health Technology, 15(3), pp. 45-60.

Smith, K. et al. (2010) 'A Comprehensive Survey of Data Mining-Based Fraud Detection Research'. Available at: https://arxiv.org/abs/1010.3009 (Accessed: 5 August 2025).

Susskind, R. and Susskind, D. (2015) The Future of the Professions: How Technology Will Transform the Work of Human Experts. Oxford: Oxford University Press.

U.S. Department of Health & Human Services (2013) Summary of the HIPAA Privacy Rule. Available at: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html (Accessed: 5 August 2025).

UK Government (1990) Computer Misuse Act 1990. Available at: https://www.legislation.gov.uk/ukpga/1990/18/contents (Accessed: 7 August 2025).

UK Government (2000) Freedom of Information Act 2000. Available at: https://www.legislation.gov.uk/ukpga/2000/36/contents (Accessed: 7 August 2025).

Veena, K. et al. (2023) 'Predicting Health Insurance Claim Frauds Using Supervised Machine Learning Technique', in 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, pp. 1-7.

Viaene, S. et al. (2002) 'A Comparative Study of Knowledge Discovery Techniques for Fraud Detection in Health Insurance', in Proceedings of the European Conference on Information Systems (ECIS), Gdansk, Poland.

Waghade, R. and Karandikar, R. (2018) 'A comprehensive study of healthcare fraud detection based on machine learning', International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 3(1), pp. 2456-3307.

Waghade, S.S. and Karandikar, A. (2018) 'A Comprehensive Study of Healthcare Fraud Detection Based on Machine Learning', International Journal of Applied Engineering Research, 13(6), pp. 4364-4370.

Zhang, Y., Zhu, J., Wang, Q. and Sun, X. (2021) 'Machine learning in health care: A review of recent applications', Frontiers of Medicine, 15, pp. 1-16.

Appendix A

# **Project GitHub Repository**

The complete code, documentation, and additional resources for this project, including data preprocessing scripts, model implementation, hyperparameter tuning configurations, and the full suite of evaluation plots, are hosted in a dedicated GitHub repository.

This repository serves as the central hub for the project's technical assets and provides a transparent, reproducible account of the research methodology.

**Repository Title:** MSc Project - Healthcare Fraud Detection System using Machine Learning

**URL:** <https://olympus.ntu.ac.uk/n1325700/MSc-Project---Healthcare-Fraud-Detection-System-using-Machine-Learning.git>