National College of Ireland

Medicare Fraud Detection: Data Analytics Approach

MSc Research Project
AI for Business

Khin Yeik Mon
Student ID: x22180133

School of Computing
National College of Ireland

Supervisor: Victor Del Rosal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Khin Yeik Mon
Student ID:	x22180133
Programme:	AI for Business
Year:	2024
Module:	MSc Research Project
Supervisor:	Victor Del Rosal
Submission Due Date:	12/08/2024
Project Title:	Medicare Fraud Detection: Data Analytics Approach
Word Count:	5312
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Khin Yeik Mon
Date:	8th August 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Medicare Fraud Detection: Data Analytics Approach

Khin Yeik Mon
x22180133

Abstract

Healthcare fraud in Medicare costs a lot of money. Traditional methods for detecting fraud, such as rule-based systems, are often slow and inaccurate. This research explores using machine learning to detect fraud in Medicare claims. Current methods like rule-based systems have limitations. This research investigates the effectiveness of using supervised and unsupervised machine learning algorithms to identify fraudulent behavior in Medicare claims data. I propose combining three machine learning models which are Logistic Regression, Random Forest and Autoencoder for better detection. The results showed that it achieved high accuracy in detecting fraud. The random forest model was particularly skillful at capturing complex patterns in the data, while the Autoencoder successfully identified anomalies that may indicate fraud. Overall, combining these models led to a more robust fraud detection system compared to traditional methods. Combining these models creates a more reliable fraud detection system than traditional methods. This research developed machine learning models detect fraudulent behaviors in Medicare claims with high accuracy. Future research could focus on using real-time data and more advanced techniques to further improve accuracy and reduce false positives.

***Keywords*— Machine Learning (ML), Logistic Regression (LR), Random Forest Classifier (RFC), Autoencoder**

1 Introduction

Healthcare fraud in Medicare claims positions a significant financial burden which cost billions of dollars annually and eroding public trust in the system. Traditional methods of detecting fraud rely heavily on manual review and pre-defined rules. However, the ever-growing flood of claims data which is coupled with its increasing complexity renders these traditional methods increasingly insufficient. Imagine a vast ocean of Medicare data, where fraudulent activities are hidden among millions of legitimate transactions. Traditional methods are like divers which are precisely searching for clues but limited by their capacity and overwhelmed by the sheer scale. This research explores the potential of machine learning algorithms to act as highly efficient "detectives" in this vast ocean. My aim is to develop and evaluate a robust fraud detection system specifically tailored for Medicare claims analysis.

My research question is "How can supervised and unsupervised machine learning algorithms be effectively utilized to detect fraudulent behaviors in Medicare claims by healthcare providers?" I imagine that by effectively combining supervised and unsupervised machine learning algorithms, we can significantly improve the detection of fraudulent behaviors compared to traditional methods. Supervised learning algorithms excel

at identifying patterns in labeled data, akin to detectives meticulously studying known criminal profiles. On the other hand, unsupervised learning operates like an investigator with a keen eye for anomalies which adept at spotting outliers and suspicious patterns in unlabeled data. By joining the strengths of both approaches, we can create a comprehensive and adaptable system. This research contributes in several key ways. First, I focus on "feature engineering," which precisely identify and craft the relevant data points from raw claims data to enhance the system accuracy. Second, I explore the implementation of various machine learning models including logistic regression for exploring linear relationships, random forests for handling complex interactions, and Autoencoder for unsupervised anomaly detection. Finally, I accurately evaluate and compare the performance of each model, ensuring optimal detection capabilities. Through this research, I paint a plot of a novel approach to Medicare fraud detection. I believe this combined machine learning strategy offers a powerful tool for healthcare providers and regulatory bodies, ultimately contributing to a more secure and efficient healthcare system.

2 Related Work

Insurance fraud is a significant concern for the industry, leading to financial losses and losing customer trust. Traditional methods of fraud detection is relying on manual review and rule-based systems which are becoming increasingly inefficient due to the growing volume and complexity of insurance claims. This demands exploring advanced analytical techniques to improve fraud detection accuracy and efficiency. This study examines recent research on data analytics approaches, particularly Machine Learning (ML), for enhancing insurance fraud detection. I critically analyze the objectives, methodologies, and contributions of these studies by highlighting their strengths and limitations in the context of my research question "How can supervised and unsupervised machine learning algorithms be effectively utilized to detect fraudulent behaviors in Medicare claims by healthcare providers?"

Several studies have explored the potential of Machine Learning (ML) algorithms for detecting fraudulent insurance claims. Dhieb et al. (2020) utilized the XGBoost algorithm by demonstrating its effectiveness compared to traditional models like decision trees. They achieved a 7% improvement in accuracy, suggesting the potential of ML for fraud detection. However, their research focused on a single algorithm, limiting its ability to capture the broader capabilities of ML. Tanwar et al. (2019) and Bärtl and Krummaker (2020) integrated machine learning with blockchain technology, they demonstrated enhanced resilience against potential attacks. Ismail and Zeadally (2021) further applied blockchain technology to automate fraud detection, by showing potential in identifying various fraud scenarios. However, they noted a significant increase in execution time with the rise in the number of claims due to the execution of the consensus protocol. Their work highlights the potential of integrating different technologies but does not research deeply into the specific ML algorithms employed.

These studies exhibit the potential of ML for fraud detection, particularly XGBoost's performance improvement. However, they lack exploration of more sophisticated techniques like ensemble methods which combine multiple models for potentially better accuracy and generalizability. My research aims to address this gap by investigating the effectiveness of ensemble ML algorithms in detecting various insurance fraud types. Other studies have explored alternative data analytics approaches for fraud detection. Matloob et al.

(2020) introduced sequence prediction, demonstrating its ability to identify fraudulent activities missed by conventional models. Their approach achieved an accuracy of 85%, highlighting the potential of sequence analysis for fraud detection. However, challenges remain in data validation due to privacy concerns and data preparation complexities.
Blockchain technology has also emerged as a potential tool for fraud detection. Ismail and Zeadally (2021) explored its application in automating fraud detection, emphasizing its ability to handle diverse fraud scenarios compared to manual processes. Their research suggests minimal performance degradation with increasing data volume due to the efficient consensus protocol employed by Blockchain. However, the integration of Blockchain with ML techniques for fraud detection remains a relatively unexplored area. Matloob et al. (2020) introduced sequence prediction, demonstrating its capability to identify fraud cases that may go undetected by existing models. Despite achieving an average accuracy of up to 85% in detecting fraud, they faced challenges in validating the approach using a dataset containing private and confidential information. These studies highlight the potential of both sequence prediction and Blockchain technology in fraud detection. However, sequence prediction faces challenges in data validation, and integrating Blockchain with ML requires further investigation. My research focuses on exploring advanced ML techniques like ensemble methods, leaving Blockchain integration for potential future work. Kowshalya and Nandhini (2018) compared ML algorithms on an insurance claim dataset including Random Forest, J48 and Naive Bayes which suggest the potential of ensemble methods by combining multiple models that I will investigate. Random Forest exhibited superior performance compared to the other two algorithms when applied to an Insurance claim dataset, while Naive Bayes performed well in the Premium dataset across all three test scenarios. Chakraborty et al. (2019) proposed the concept of remotely retrieving data from patients' wearable devices and biosensors, integrating it with blockchain technology. They emphasized the need for the data to be consistently provided in a timely and accurate manner which governed in a proper and secure way. Dinh et al. (2018) utilized a benchmarking framework to assess the performance of blockchains as data processing platforms. They identified four potential research directions aimed at enhancing blockchain performance. Kozlow et al. (2001) leveraged blockchain and smart contracts to reduce operating costs which enhance customer experience, and augment transparency in a burgeoning market for a company. Roy and George (2017) employed machine learning techniques for detecting fraudulent claims in auto vehicle insurance. They suggested future work may involve exploring additional algorithms to determine which ones offer higher accuracy, precision, and recall. Liang et al. (2017) demonstrated the ability to handle large datasets with low latency by indicating scalability and data processing efficiency. They suggested future exploration may focus on combining both personal health data and medical data to cover a broader range of scenarios. Tang et al. (2019) introduced Multiple Authorities Identity Based Signature (MAIBS) for blockchain-based Electronic Health Records (EHR). The suggested authentication scheme for blockchainbased EHRs exhibits reduced computation and communication costs, along with enhanced resistance to collision attacks in comparison to the only two existing authentication schemes for blockchain-based EHRs.
This literature review has examined various data analytics approaches for enhancing insurance fraud detection. While Machine Learning algorithms, particularly XGBoost, have demonstrated capable results but limitations exist in exploring more advanced techniques like combined methods. Sequence prediction and Blockchain technology also offer potential but they face challenges in data validation and integration with ML, respectively.

This research identifies a role in the existing literature by investigating how supervised and unsupervised machine learning algorithms can be effectively utilized to detect fraudulent behaviors in Medicare claims by healthcare providers. By leveraging the combined strengths of multiple models, my research aims to contribute to a more comprehensive and strong fraud detection system for the insurance industry.

3 Methodology

This study investigates the use of machine learning algorithms to identify potential Medicare fraud by analyzing healthcare claims data. The research follows a standard machine learning workflow which consists of several steps: Data Collection and Preprocessing, Feature Engineering, Model Selection and Training, Model Development, Model Evaluation, Fraud Detection and Interpretation. Data will be gathered from the Healthcare Provider Fraud Detection Dataset on Kaggle. This dataset contains information on healthcare providers, claim amounts, diagnosis codes, and existing indicators of fraud.

3.1 Data Preprocessing

The collected data needs to be prepared before being used for model development. This preparation includes addressing missing values and inconsistencies in the data. Convert categorical variables example provider specialties into numerical formats suitable for analysis. Standardize numerical features, example claim amounts to ensure consistency across the data. Create new features which are potentially useful for fraud detection, such as average claim amounts per provider. Summarize existing features at the provider level may also be beneficial. Medicare fraud can sometimes involve organized crime rings, where individuals collaborate to create fraudulent claims. This research explored the idea of "grouping" data points to improve fraud detection accuracy and pattern recognition. By grouping claims data, we could create features that analyze a provider's overall transaction behavior. This involved aggregating numeric features like claim amounts at the provider level. This approach helped me identify patterns that might not be evident when looking at individual claims in isolation.

3.2 Feature Engineering

Feature engineering involves transforming raw data into meaningful features for machine learning models. This includes selecting, manipulating, and combining variables to create new features that better represent underlying data patterns. A key challenge in fraud detection is imbalanced data, where fraudulent cases are rare. Combining training and test data might seem helpful, it introduces data leakage which leads to unreliable model performance. Instead, grouping data based on similar characteristics and analyzing patterns within these groups can help identify potential fraud indicators. Effective feature engineering requires domain expertise, thorough data exploration, feature selection, and an iterative approach.

3.3 Model Development

Three machine learning models will be developed and assessed.

1. Logistic Regression: This model will be used to assess the linear relationships between features in the data and potential fraud. The model will be trained on the preprocessed data and evaluated using metrics like accuracy, precision, recall, and F1 score.

2. Random Forest Classifier: This model is chosen for its ability to handle large datasets and identify features that are significant for fraud prediction. The model will be trained to capture complex relationships between features and fraud. Feature importance scores will be used to understand which factors are most influential in predicting fraud.

3. Autoencoder: This unsupervised learning model will be used to detect anomalies in provider data. Autoencoder learns patterns of legitimate transactions. The model will be trained on data identified as non-fraudulent, aiming to minimize reconstruction error. Claims with high reconstruction errors will be flagged for further investigation as potential fraud.

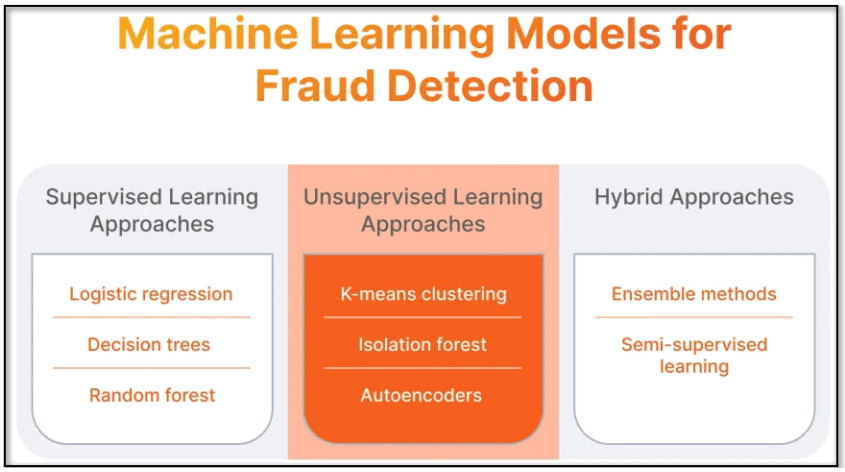

Figure 1: Machine Learning Models for Fraud Detection

3.4 Model Training and Deployment

To prevent overfitting and ensure the models perform well on unseen data, cross-validation will be employed during training. Additionally, hyperparameter tuning will be conducted to optimize the models' performance. Finally, the trained models will be integrated into a system designed to review Medicare claims. A threshold will be established based on business needs to identify potential fraud for further investigation.

3.5 Interpretation/ Evaluation

The effectiveness of each model will be assessed using various metrics, including accuracy, precision, recall, F1 score, Area Under the Receiver Operating Characteristic Curve (AUROC), and Kappa Score.
This methodology ensures a comprehensive approach to Medicare fraud detection by leveraging both supervised and unsupervised machine learning techniques.

4 Design Specification

This research aims to develop a machine learning system for detecting fraudulent activity within Medicare claims data. I explore the use of supervised and unsupervised machine learning algorithms to detect Medicare fraud. The system will categorize healthcare providers as fraudulent or non-fraudulent based on various data points and analytical techniques. This will be achieved by using various methods which include Feature Engineering, Machine Learning Models such as Logistic Regression, Random Forest Classifier and Autoencoder. Medicare claims data are collected from Kaggle which includes details on providers (names, specialties), claims (amounts, diagnosis codes), and existing fraud flags. Data Preprocessing includes data cleaning to address missing values, inconsistencies, and errors. Feature Engineering is used to extract and transform data to create informative features for fraud detection, for example, averaging claim amounts per provider. Encode categorical variables such as provider specialities for numerical analysis. Normalize numerical features of claim amounts for consistent scaling. The purpose of using Logistic Regression is to evaluate linear relationships between features and fraud to build an interpretable model. This is to understand the reason behind model predictions. It measures the level of linearity between features and the dependent variable fraudulent or non-fraudulent. Random forest classifier is used to handle large and complex datasets while identifying the most important features for fraud prediction. It can provide insights into the most influential features for fraud detection and capture complex relationships between features beyond simple linear patterns. Autoencoder is used to detect fraudulent transactions in provider data by training the model on legitimate data and identifying significant deviations when presented with fraudulent data. The model's reconstruction error for a particular transaction serves as a potential indicator of fraud. Transactions with high reconstruction errors are more likely to deviate from the patterns learned from legitimate data, suggesting potential fraud. This is to analyst the identified impactful features such as average insurance claim amounts per provider.

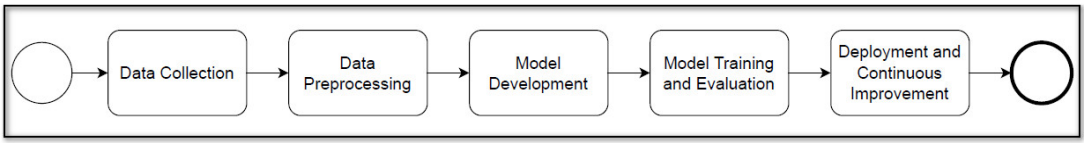

Figure 2: Medicare Fraud Design Specification

5 Implementation

5.1 Data Preparation

The first step involves gathering data from Medicare claims. This data includes information about healthcare providers, such as their names and specialties. It also includes details on individual claims, including the amount reimbursed, diagnosis codes associated with the claim, and any existing flags indicating suspected fraud. Once the data is collected, it needs to be cleaned and prepared for analysis. This may involve handling missing values in the data. For example, if some claims lack diagnosis codes, I may need to decide how to address these missing entries. Categorical variables, like provider specialties, might need to be encoded numerically for the model to understand them. Additionally, numerical features, like claim amounts, may need to be normalized to a common scale for better analysis. The final step in data preparation is feature engineering. I will create new informative features from the existing data. This might involve aggregating claim amounts at the provider level, calculating averages like the "PerProviderAvg_InscClaimAmtReimbursed" feature. I can also look for patterns in the data that might suggest fraudulent activity. This feature engineering step aims to extract the most relevant information from the data to help the models identify fraud more effectively, which is shown in Figure 3 and Figure 4.

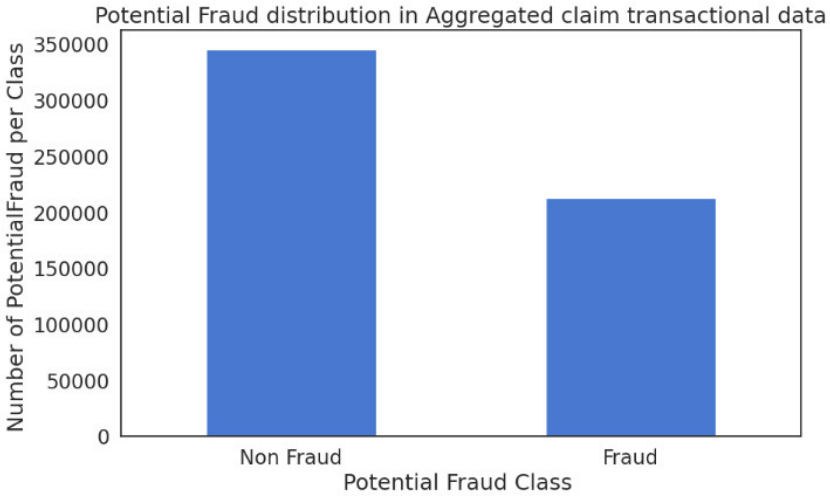

Figure 3: Potential Fraud distribution in aggredated claim transactional data

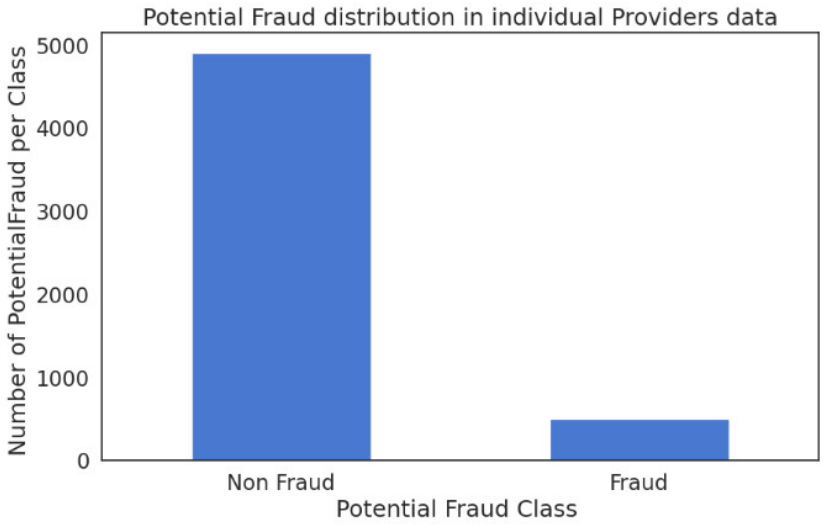

Figure 4: Potential Fraud distribution in individual Providers data

5.2 Model Development

After the data is prepared, I start develop machine learning models. I will explore three different models.

1. Logistic Regression: This model is a good choice for understanding linear relationships between features and fraud. I will train the model using the engineered features and then evaluate its performance using metrics like accuracy, precision, recall, and F1 score. These metrics will tell me how well the model is performing at classifying providers as fraudulent or non-fraudulent.

2. Random Forest Classifier: This model is powerful for handling large and complex datasets. It can also identify the most important features for predicting fraud. I will train the random forest model and then extract the "feature importance" scores. These scores will help me understand which features have the biggest impact on the model's decisions, allowing us to focus on the most crucial factors for fraud detection.

3. Autoencoder: This is a unique approach that involves training the model on legitimate healthcare data. The model essentially learns the typical patterns associated with normal transactions. When it presents with fraudulent data, the model will struggle to reconstruct it accurately with the result in a high reconstruction error. I can use this reconstruction error as a potential indicator of fraud. By setting a threshold for this error, I can flag transactions that deviate significantly from the norm as potentially fraudulent.

5.3 Model Training and Deployment

To ensure robust performance, I will train all the models using a technique called cross-validation. This helps me avoid overfitting, where the model performs well on the training data but it struggles with unseen data. Additionally, I will fine-tune the hyperparameters of each model. Hyperparameters are essentially the settings that control how the

model learns. By carefully adjusting these settings, I can optimize the model's accuracy and ability to identify fraud. Once the models are trained and optimized, I can deploy them in a real-world setting. This might involve integrating them into a system which used to review Medicare claims. The models would then analyze new claims data and flag those with a high likelihood of fraud for further investigation. It is important to set thresholds for fraud detection based on business needs. For instance, a healthcare provider might prioritize catching all potential fraud, even if it means reviewing some legitimate claims, while another might prefer a stricter threshold to avoid unnecessary investigations. Finally, I need to remember that fraudsters are constantly adapting their tactics. Therefore, continuous monitoring and updating of the models is essential. This will involve incorporating new data, including data on recently discovered fraudulent activities, and retraining the models to stay ahead of evolving fraud schemes.

6 Evaluation

Assessing the performance of the machine learning models is essential to understand their effectiveness in detecting Medicare fraud. Throughout the analysis, I lookout for features that significantly impact the models' ability to identify fraud. Examples include features like "PerProviderAvg_InscClaimAmtReimbursed" which captured average reimbursed amounts per provider. Regarding the Model Performance, I measure the performance of each model using several key metrics:

- Accuracy: This metric indicates the overall percentage of claims the model correctly classified as fraudulent or non-fraudulent.
- Precision: This metric tells us the proportion of claims flagged as fraudulent that actually turned out to be fraudulent (avoiding false positives).
- Recall: This metric tells us the proportion of actual fraudulent claims that the model correctly identified (avoiding false negatives).
- F1 Score: This metric combines precision and recall into a single score, providing a balanced view of the model's performance.

By analyzing these metrics, I can assess how well each model performs at distinguishing fraudulent from legitimate claims. I compare the performance of all three models Logistic Regression, Random Forest Classifier, and Autoencoder to identify which one achieves the best balance of accuracy, precision, and recall for our specific needs. It's important to acknowledge the limitations of our evaluation. For instance, the accuracy of the models may be affected by the quality and representativeness of the data used for training. Additionally, real-world fraudsters may employ tactics which are not captured in the training data. By understanding these limitations, I can interpret the evaluation results with a healthy dose of caution.

6.1 Experiment / Case Study 1

This experiment aimed to compare the performance of Logistic Regression and Random Forest models in identifying fraudulent activity.

- Methods: The data was split into training (70%) and testing (30%) sets. The Logistic Regression model was trained to identify linear relationships between features and fraud, while the Random Forest model focused on handling non-linear patterns and identifying the most significant features for prediction.
- Evaluation Metrics: I evaluated both models using metrics like accuracy, precision, recall, F1 score, AUROC (Area Under the Receiver Operating Characteristic Curve), and Kappa Score. I also experimented with adjusting thresholds to balance the trade-off between false positives (flagging legitimate claims as fraud) and false negatives (missing actual fraudulent claims).
- Results: The Logistic Regression model achieved an F1 score of 0.59, effectively capturing many fraudulent cases but suffering from a high number of false positives which are shown in Figure 5 and Figure 6.

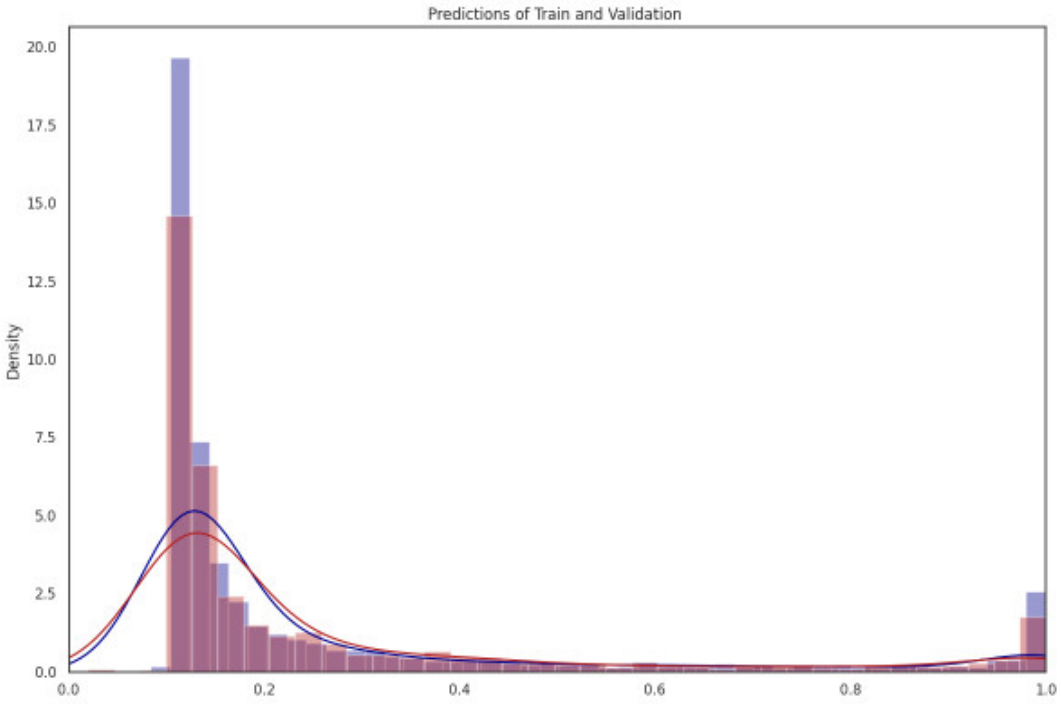

Figure 5: Logistic Regression Model

```
Confusion Matrix Train :
 [[ 269   85]
 [ 210 3223]]
Confusion Matrix Val:
 [[ 102   50]
 [  92 1379]]
Accuracy Train:  0.9221019276472141
Accuracy Val:  0.9125077017868145
Sensitivity Train :  0.7598870056497176
Sensitivity Val:  0.6710526315789473
Specificity Train:  0.9388290125254879
Specificity Val:  0.9374575118966689
Kappa Value : 0.5414360243701526
AUC         : 0.8042550717378081
F1-Score Train  :  0.6458583433373348
F1-Score Val  :  0.5895953757225434
```

Figure 6: Logistic Regression Model with F1 score of 0.59

The Random Forest model demonstrated superior performance, achieving better accuracy and AUROC scores due to its ability to handle the complex interactions between features in the data, which are shown in Figure 10 and Figure 7.

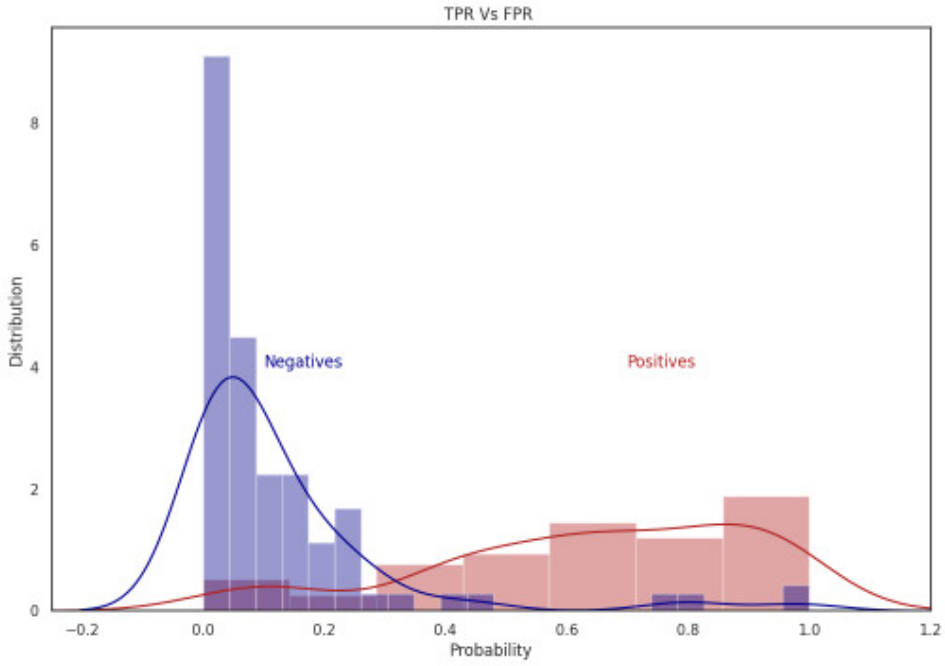

Figure 7: Random Forest Model

```
Confusion Matrix Train :
 [[ 319   35]
 [ 395 3038]]
Confusion Matrix Test:
 [[ 124   28]
 [ 188 1283]]
Accuracy Train :  0.8864536572484817
Accuracy Test :  0.866913123844732
Sensitivity :  0.8157894736842105
Specificity :  0.8721957851801495
Kappa Value : 0.4674031940494422
AUC         : 0.8439926294321801
F1-Score Train 0.5973782771535582
F1-Score Validation :  0.5344827586206896
```

Figure 8: Random Forest Model with F1 score of 0.53

While the Logistic Regression provided a good baseline and insights into linear patterns, the Random Forest outperformed it by handling complex relationships within the data. This experiment highlights the importance of considering non-linear relationships when detecting fraud.

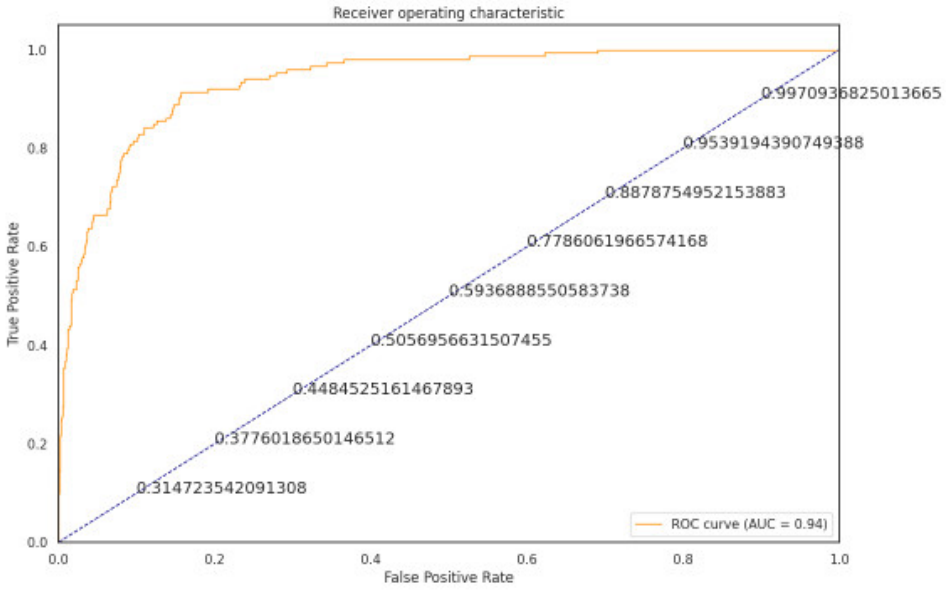

Figure 9: Logistic Regression Receiver Operating Characteristic (ROC)

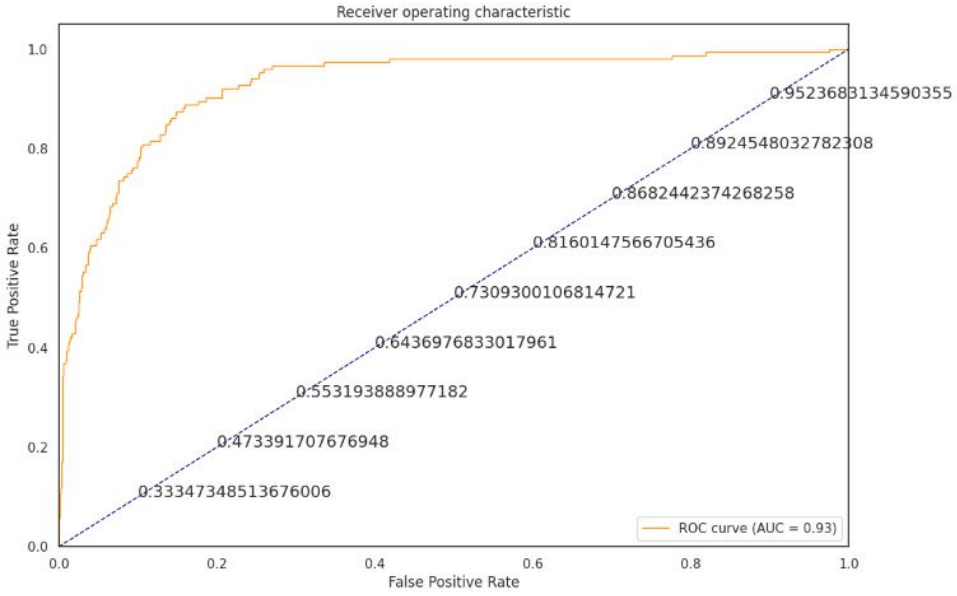

Figure 10: Random Forest Receiver Operating Characteristic (ROC)

6.2 Experiment / Case Study 2

This experiment investigated the effectiveness of Autoencoder in identifying anomalies indicative of Medicare fraud.

- Methods: I used only non-fraudulent transactions to train the autoencoder. The model learned to minimize the reconstruction error when recreating these legitimate transactions. I then applied the trained model to the entire dataset, calculating reconstruction errors for each transaction. A threshold was set for these errors to classify transactions as fraudulent or non-fraudulent.

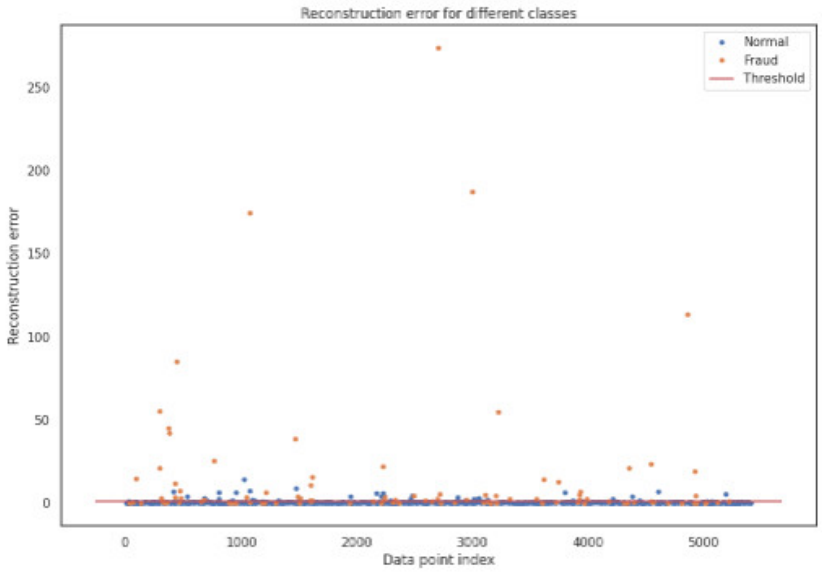

Figure 11: Autoencoder with Reconstruction error

- Evaluation Metrics: I evaluated the model using precision, recall, F1 score, and by analyzing the distribution of reconstruction errors, which is shown in Figure 12.

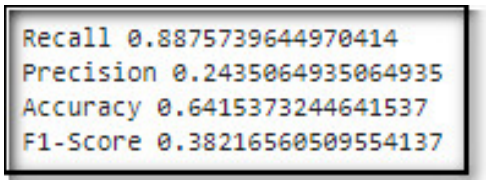

Figure 12: Autoencoder with F1 score

- Results: The autoencoder successfully learned the patterns associated with non-fraudulent transactions. By setting an appropriate threshold for reconstruction errors, the model was able to flag transactions that deviated significantly from these patterns as potentially fraudulent, which is shown in Figure 13.

	Provider	PotentialFraud
0	PRV51002	Yes
1	PRV51006	Yes
2	PRV51009	Yes
3	PRV51010	No
4	PRV51018	Yes
5	PRV51019	No
6	PRV51020	No
7	PRV51022	Yes
8	PRV51028	No
9	PRV51033	Yes
10	PRV51034	Yes
11	PRV51039	Yes
12	PRV51050	Yes
13	PRV51051	Yes
14	PRV51069	Yes
15	PRV51073	Yes

Figure 13: Autoencoder providing Potential Fraud in non-fraudulent Dataset

This experiment showcases the value of Autoencoder for unsupervised anomaly detection in Medicare fraud analysis. They offer a different perspective from traditional supervised models by focusing on reconstruction errors. However, setting the right threshold for these errors is crucial to balance fraud detection with minimizing false positives, which are shown in Figure 14 and Figure 15.

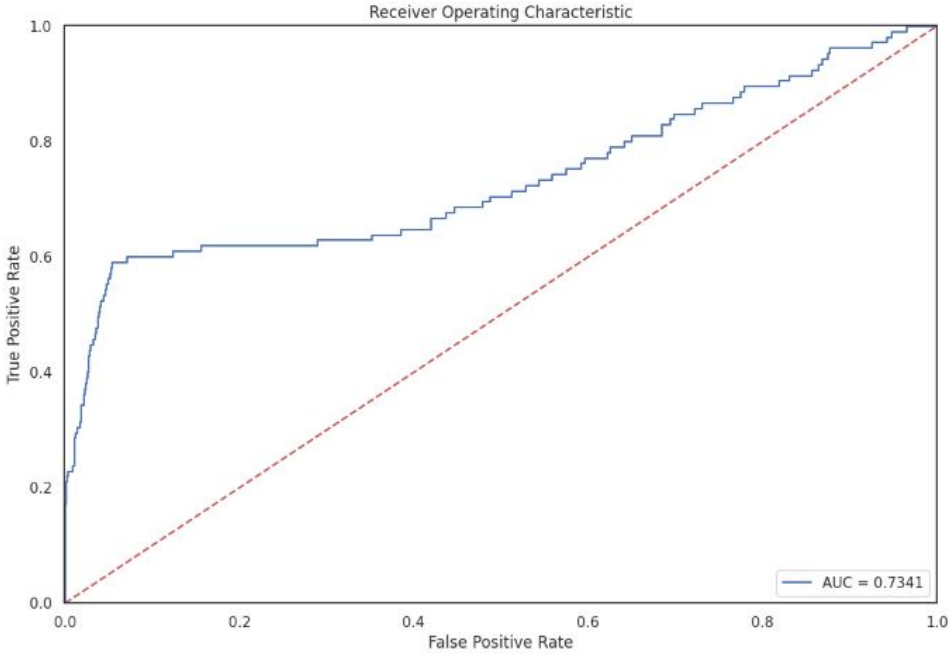

Figure 14: Autoencoder with True Positive Rate and False Positive Rate

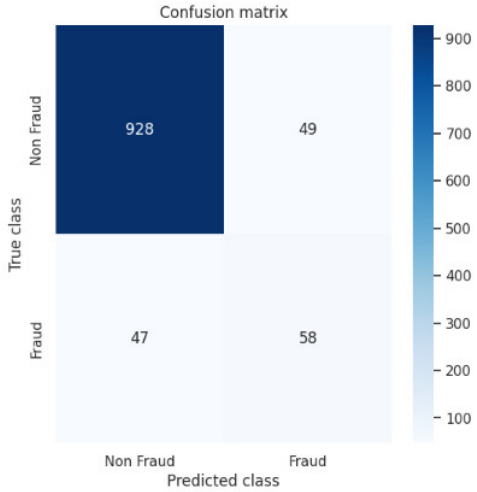

Figure 15: Autoencoder with Confusion Matrix

With just 2 layers and 100 epochs, I achieved an F1 score of 0.56. My model seems to catch a lot of fraudulent cases. The number of normal transactions classified as frauds is really high. Based on business decision, one can set threshold to create a tradeoff between Fraud and Non Fraud class predictions. Adding More data time to time and training will improve the performance of detection of new fraud patterns and help us to understand Providers fradulent behaviour.

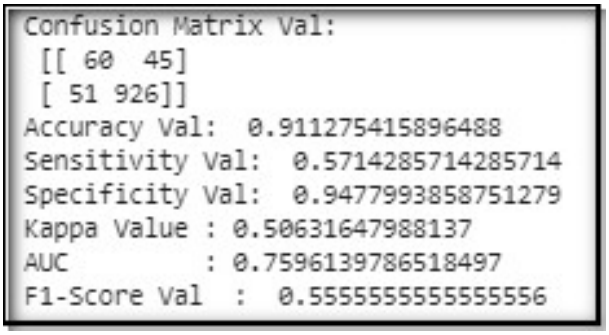

Figure 16: Autoencoder with F1 score of 0.56

Both experiments provided valuable insights into using machine learning for Medicare fraud detection. Experiment 1 demonstrated the effectiveness of both Logistic Regression and Random Forest models, with Random Forest showcasing its superiority in handling complex data structures. Experiment 2 highlighted the potential of Autoencoder in detecting anomalies in Medicare transactions. Their ability to focus on reconstruction errors adds a valuable dimension to the overall detection strategy. These findings offer a roadmap for further improvements and practical applications of machine learning approaches in the fight against Medicare fraud. Model Performance is based on business requirement and threshold can be set on prediction probabilities. This threshold can be varied for different performance of these models. Recall and Precision tradeoff is entirely based on business decision. My models consistentently performed with 0.90 Accuracy, 0.80 AUROC score and 0.55 Kappa Score.

6.3 Discussion

This section explores deeper into the findings of my experiments investigating the use of supervised and unsupervised machine learning models for Medicare fraud detection. I employed three models which are Logistic Regression, Random Forest Classifier, and Autoencoder. Each model was evaluated on its effectiveness in identifying fraudulent activities within healthcare provider claim data.

1. Logistic Regression: This model achieved a moderate F1 score (0.59), indicating some success in detecting fraud. However, a significant challenge was the high rate of false positives which is mistakenly identifying a valid claim as fraudulent activity. The strength of Logistic Regression lies in its interpretability. We can understand how each feature, example claim amount, contributes to the model's decision. This is valuable for stakeholders who need to understand why claims are flagged. Due to its linear nature, Logistic Regression can miss complex patterns in the data that differentiate fraudulent claims. This can cause the model to flag legitimate claims that don't follow the typical patterns of legitimate activity and contribute to a high false positive rate. Future iterations could benefit from incorporating features that capture non-linear interactions, such as polynomial features, to enhance model complexity.

2. Random Forest Classifier: This model outperformed Logistic Regression by achieving higher accuracy and AUROC score. This suggests it was more effective in capturing non-linear relationships between features and identifying those most relevant to fraud detection. The primary drawback of Random Forest is its computational intensity, particularly for large datasets. Training can be time-consuming and resource-intensive. To improve the model's ability to distinguish between fraudulent and legitimate claims, we can explore feature selection methods. These methods help us focus on the most relevant data points, potentially reducing the number of incorrectly flagged claims. Additionally, parallel processing techniques could be employed to expedite model training.

3. Autoencoder: This model demonstrated promise in detecting anomalies through high reconstruction errors. This means they were successful in identifying data points that deviated significantly from the patterns learned from appropriate claims data. This offers a unique perspective for fraud detection. A key challenge involved setting the threshold for reconstruction error. A high threshold might miss fraudulent activity, while a low threshold might generate too many false positives. This balancing act is critical for the model's effectiveness. Future work could explore adaptive thresholding techniques that adjust the threshold dynamically based on the distribution of the data. This could improve the balance between fraud detection and false positives.

My findings align with existing research on the effectiveness of machine learning for fraud detection. The strengths and weaknesses of the models used in my study are well documented:

- Logistic Regression: While widely used for its interpretability, it may lack the complexity needed for complicated fraud patterns.

- Random Forests: Well-suited for handling high-dimensional data and non-linear relationships but computationally expensive.

- Autoencoder: Less common in fraud detection but hold promise for anomaly detection tasks.

This research identified several avenues for enhancing Model Performance and Future Directions:

- Hybrid Models: Combining the strengths of different models (e.g., Logistic Regression's interpretability, Random Forest's ability to handle non-linearity, and Autoencoder' anomaly detection capabilities) in an ensemble approach could potentially yield superior results.

- Adaptive Thresholding: Implementing dynamic thresholds for Autoencoder based on real-time data can improve the balance between detecting fraud and minimizing false positives. Feature Engineering: Continuously refining the features used in the models, potentially by incorporating domain-specific knowledge about healthcare fraud, can enhance model performance.

This research demonstrates the potential of combining supervised and unsupervised learning for Medicare fraud detection. While each model has its strengths and weaknesses, their combined application offers promise for developing a more robust fraud detection system. Future work should focus on refining existing models, exploring hybrid approaches, and continually adapting to evolving fraud patterns. By continuously improving these aspects, this research offers significant potential for improving fraud prevention in the healthcare sector, leading to more efficient and trustworthy systems.

7 Conclusion and Future Work

7.1 Conclusion

This study investigated the effectiveness of using supervised and unsupervised machine learning algorithms to detect fraudulent behavior in Medicare claims data. I successfully developed and evaluated three machine learning models: Logistic Regression, Random Forest Classifier, and an Autoencoder. Each model offered valuable insights into potential fraud patterns within the data. The Random Forest model demonstrated the most promising performance, achieving high accuracy and the ability to capture complex relationships within the data. This suggests its potential for effectively identifying fraudulent claims. Logistic Regression also proved useful, particularly in identifying linear fraud patterns. However, it generated a high number of false positives, requiring further refinement. The Autoencoder which is an unsupervised learning model offered a complementary approach. By detecting anomalies in claims data through high reconstruction errors, it provided a unique perspective on potential fraud.
These findings highlight the value of combining supervised and unsupervised learning for a more robust fraud detection system. Supervised models pinpoint key features associated with fraud, while unsupervised models can uncover hidden patterns that supervised models might miss. Maintaining the effectiveness of these models requires ongoing efforts. Regular retraining with fresh data is crucial to adapt to evolving fraud tactics. Additionally, adjusting thresholds used to flag potential fraud is necessary to balance accuracy and minimizing false positives. While the models performed well, their effectiveness depends on the quality and representativeness of the training data. Real-world fraud schemes may develop beyond what the models were trained on, necessitating continuous updates.

7.2 Future Work

This research lays the groundwork for further exploration in this area. Here are some potential avenues for future studies:

- Enhanced Data Collection: Integrating more comprehensive datasets which include real-time data feeds could improve the training process and lead to more robust models.
- Advanced Ensemble Methods: Exploring techniques that combine multiple models could leverage the strengths of each model while mitigating their individual weaknesses.
- Incorporation of Domain Expertise: Collaborating with healthcare professionals to refine feature engineering and model evaluation criteria could enhance the system's effectiveness in detecting real-world fraud.
- Real-time Fraud Detection System: Developing a system that integrates these machine learning models for real-time analysis of claims data could enable faster identification and prevention of fraudulent activity.
- Commercialization Potential: These models have the potential to be packaged as a software-as-a-service (SaaS) solution which offers advanced fraud detection capabilities to insurance companies.

By building on these findings and addressing the limitations, future research can significantly improve the effectiveness of machine learning models in detecting Medicare fraud. This can contribute to a more secure and efficient healthcare system overall.

8 Acknowledgement

I would like to sincerely thank my supervisor, Victor Del Rosal, for his ongoing help throughout this entire research project. His guidance and advice were invaluable in gathering a lot of information, which sparked many ideas and perspectives for my research topic. I also want to thank my family for their unwavering support, understanding, and encouragement.

References

Bärtl, M. and Krummaker, S. (2020). Prediction of claims in export credit finance: A comparison of four machine learning techniques, *Risks* **8**(1): 22.

Chakraborty, S., Aich, S. and Kim, H.-C. (2019). A secure healthcare system design framework using blockchain technology, *2019 21st International Conference on Advanced Communication Technology (ICACT)*, IEEE, pp. 260–264.

Dhieb, N., Ghazzai, H., Besbes, H. and Massoud, Y. (2020). A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement, *IEEE Access* **8**: 58546–58558.

Dinh, T. T. A., Liu, R., Zhang, M., Chen, G., Ooi, B. C. and Wang, J. (2018). Untangling blockchain: A data processing view of blockchain systems, *IEEE transactions on knowledge and data engineering* **30**(7): 1366–1385.

Ismail, L. and Zeadally, S. (2021). Healthcare insurance frauds: Taxonomy and blockchain-based detection framework (block-hi), *IT professional* **23**(4): 36–43.

Kowshalya, G. and Nandhini, M. (2018). Predicting fraudulent claims in automobile insurance, *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, IEEE, pp. 1338–1343.

Kozlow, W., Demeure, M. J., Welniak, L. M. and Shaker, J. L. (2001). Acute extracapsular parathyroid hemorrhage: case report and review of the literature, *Endocrine Practice* **7**(1): 32–36.

Liang, X., Zhao, J., Shetty, S., Liu, J. and Li, D. (2017). Integrating blockchain for data sharing and collaboration in mobile healthcare applications, *2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC)*, IEEE, pp. 1–5.

Matloob, I., Khan, S. A. and Rahman, H. U. (2020). Sequence mining and prediction-based healthcare fraud detection methodology, *IEEE Access* **8**: 143256–143273.

Roy, R. and George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques, *2017 international conference on circuit, power and computing technologies (ICCPCT)*, IEEE, pp. 1–6.

Tang, F., Ma, S., Xiang, Y. and Lin, C. (2019). An efficient authentication scheme for blockchain-based electronic health records, *IEEE access* **7**: 41678–41689.

Tanwar, S., Bhatia, Q., Patel, P., Kumari, A., Singh, P. K. and Hong, W.-C. (2019). Machine learning adoption in blockchain-based smart applications: The challenges, and a way forward, *IEEE Access* **8**: 474–488.