## Part 1: Introduction

Healthcare insurance is crucial for citizens' access to quality healthcare and national well-being. But it's threatened by healthcare insurance fraud, which involves deceptive activities among medical providers, patients, and insurance companies, posing a significant challenge in the healthcare sector. Insurance companies are frequently on the receiving end of these bad practices, which in turn has caused them to hike the prices of their insurance premiums, making healthcare costs surge periodically.[1] It is pretty evident that patients and their healthcare information can easily be exploited which later can hamper the overall cost.[1] Healthcare fraud has far-reaching consequences. It jeopardizes the healthcare system's stability, erodes patient trust in insurance, and can lead to unnecessary expenses and health risks from unnecessary treatments. It also poses financial risks to insurance companies and damages the reputation of all healthcare providers.

To detect and avoid the fraud, data mining techniques are applied.[2] Data mining aids in uncovering fraud patterns and unusual behavior in extensive medical data, helping the healthcare system spot unnecessary procedures and ensure appropriate care for patients. Predictive models using historical data can also identify potential fraudulent providers, enabling early detection and prevention of medical insurance fraud. This leads to cost reduction for insurance companies and, in turn, lower insurance premiums.

In this report, we will examine academic papers on healthcare technologies and algorithms to improve our approach. Also, we will leverage the dataset we explored in our previous work, containing medical insurance claims data with provider details, billing information, patient records, procedure data, and fraud indicators. Our primary goal is to extract meaningful features from this dataset and create a predictive decision tree model for identifying potential fraudulent providers.

## Part 2: Related Work

In our earlier report, we reviewed several academic papers that explored healthcare fraud prediction from various angles. Incorporating the feedback, we received and reflecting on these papers has led to new insights and understandings. The feedback on our previous work encouraged us to enhance our feature engineering efforts. We realized that careful feature selection and construction are critical in healthcare fraud detection. This insight has reinforced our focus on designing and selecting informative features to enhance the accuracy of our predictive model.

First, the paper titled "Healthcare Provider Summary Data for Fraud Classification"[3] by J. M. Johnson emphasizes the critical role of feature engineering and dataset construction in healthcare fraud detection. The author leverages the latest publicly available data from CMS and introduces two new labeled Medicare Part B datasets for supervised learning. Their research demonstrates that, through careful selection and construction of the SbP feature set, significant improvements can be achieved in the performance of practical healthcare fraud detection models. This is of paramount importance to our project as we need to carefully consider how to design and select the most informative features to enhance the accuracy of our model.

Furthermore, the paper titled "A Comparative Analysis of Fraud Detection in Healthcare using Data Balancing & Machine Learning Technique"[4] by Nikita Agrawal et al. The paper's results indicate the superior performance of two data balancing techniques, namely Class Weighing Scheme (CWS) and Adaptive Synthetic Oversampling (ADASYN), in handling imbalanced datasets. This prompted us to consider the use of data balancing techniques to address the data imbalance issue. This approach is crucial for improving model performance metrics and mitigating the impact of skewed data distribution.

Additionally, the paper "Predicting health insurance claim frauds using supervised machine learning technique"[5] by Veena K et al. The paper emphasizes the high accuracy of the decision tree classifier, surpassing the performance of the other three algorithms it was compared to: logistic regression, random forest, and naive Bayes. We have decided to incorporate this algorithm into our model to leverage its exceptional accuracy in identifying potentially fraudulent activities.

Through the feedback and insights gained from these papers, we have gained insights into the pivotal roles played by feature engineering, data balancing and decision tree classifiers techniques in healthcare fraud prediction. These insights will guide us in formulating sound data preprocessing, feature extraction plans and methodological strategies for our project, ensuring that our project is better aligned with the challenges of healthcare fraud prediction.

## References:

[1] A. Bhardwaj, S. Kumar and A. Naidu, "Predictive analysis and supervised detection for fraudulent cases in healthcare," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2022, pp. 416-421, doi: 10.1109/Confluence52989.2022.9734195.

[2] V. Rawte and G. Anuradha, "Fraud detection in health insurance using data mining techniques," 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India, 2015, pp. 1-5, doi: 10.1109/ICCICT.2015.7045689.

[3] J. M. Johnson and T. M. Khoshgoftaar, "Healthcare Provider Summary Data for Fraud Classification" 2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI), San Diego, CA, USA, 2022, pp. 236-242, doi: 10.1109/IRI54793.2022.00060.

[4] N. Agrawal and S. Panigrahi, "A Comparative Analysis of Fraud Detection in Healthcare using Data Balancing & Machine Learning Techniques" 2023 International Conference on Communication, Circuits, and Systems (IC3S), BHUBANESWAR, India, 2023, pp. 1-4, doi: 10.1109/IC3S57698.2023.10169634.

[5] V. K et al., "Predicting health insurance claim frauds using supervised machine learning technique" 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICONSTEM56934.2023.10142604.