

# Artificial Intelligence in Healthcare Fraud Detection

Alisha Hunt - Lakshay Arora - Yoseph Kebede - Roman Libkhen

**ENPM808Y:** Fundamental Concepts of AI and ML and their Applications

**Instructor**: Dr. Massoud Moussavi

**TA**: Rahul Karanam

## 

## Table of Contents

[Artificial Intelligence in Healthcare Fraud Detection 1](#_Toc135077579)

[Table of Figures 3](#_Toc135077580)

[Table of Tables 3](#_Toc135077581)

[Abstract 4](#_Toc135077582)

[Introduction 5](#_Toc135077583)

[Background 6](#_Toc135077584)

[Problem Statement 10](#_Toc135077585)

[Objectives 10](#_Toc135077586)

[Methodology 11](#_Toc135077587)

[Data Analysis 11](#_Toc135077588)

[DATA MODELING AND PROCESSING 11](#_Toc135077589)

[Results 18](#_Toc135077590)

[Potential Issues 20](#_Toc135077591)

[Timeline and Collaboration 22](#_Toc135077592)

[Resources 22](#_Toc135077593)

[Conclusion 23](#_Toc135077594)

[Future Scope 23](#_Toc135077595)

[References 24](#_Toc135077596)

## Table of Figures

[Figure 1 - Telemedicine Fraud Scheme and Data (The US Department of Justice, 2021) 5](#_Toc135116157)

[Figure 2 - Analytics Vidhya Max AUC result of models after using SMOTE 7](#_Toc135116158)

[Figure 3 - Performance of LightGBM in correctly predicting fraudulent and non-fraudulent claims (NYC Data Science Academy, 2022). 8](#_Toc135116159)

[Figure 4 - State space example of an ensemble model with n model types 9](#_Toc135116160)

[Figure 5 - Statistical descriptions of the numerical beneficiary data columns 12](file:///C:\Users\swagg\Downloads\Final%20Report.docx#_Toc135116161)

[Figure 6 - Statistical descriptions of the numerical columns in Inpatient Data 12](#_Toc135116162)

[Figure 7 - Statistical description of numerical columns in Outpatient Data 12](#_Toc135116163)

[Figure 8 - Pie-Charts and Histograms of Beneficiary Data 13](file:///C:\Users\swagg\Downloads\Final%20Report.docx#_Toc135116164)

[Figure 9 - Histograms of Numerical Features of Inpatient Data 14](#_Toc135116165)

[Figure 10 - Histograms of Numerical Features in Outpatient Data 14](#_Toc135116166)

[Figure 11 - Histograms of Newly Created Features 14](#_Toc135116167)

[Figure 12 - Model Clusters 15](#_Toc135116168)

[Figure 13 - Weak Learner Algorithms 16](#_Toc135116169)

[Figure 14 - Diagram of Deep Neural Network 17](#_Toc135116170)

[Figure 15 - Table of metrics 18](#_Toc135116171)

[Figure 16 - Receiver Operating Characteristic Curve 18](#_Toc135116172)

[Figure 17 - Training and Validation Loss 19](#_Toc135116173)

[Figure 18 - Screenshot of GUI Dashboard 20](#_Toc135116174)

## Table of Tables

[Table 1 - Comparison of Results 19](#_Toc135116147)

## Abstract

Healthcare is a trillion-dollar industry in the United States and a prominent problem in this sector is fraudulent healthcare claims made by the healthcare providers. These claims can be made with or without the consent of the patients, but ultimately all fraud causes waste and increases in healthcare costs.

This work used a large public database of Medicare claims to train the deep-learning network to flag providers who likely engage in defrauding the government’s Medicare program. The data were perfectly cleaned in a cluster-density, noise, and variance aware way before feeding it to a novel ensemble neural network, consisting of several meta-learners and a neural network in sequence. The neural network also used branching and recursive gradient transfer to optimize the convergence process. An AUC score of 0.82268 was achieved, exceeding some of the previous works. This demonstrates a technique that could help governments and insurance providers bring the cost of healthcare coverage down and provide affordable healthcare to all.

## Introduction

One of the core sectors for any country is Healthcare. The US National Health Expenditure was reported to be $4.3 trillion and accounted for 18.3% of the nation’s GDP in 2021 (US Centers for Medicare & Medicaid Services, 2023). In the same year, the US spent 10% of its budget on Medicare and another 9% on other healthcare programs, such as Medicaid, ADA and CHIP (Cubanski & Neuman, 2023). It is estimated that approximately 10% of the budget consumed by these programs is wasted on fraudulent claims. Recently, fraudulent activities have been on the rise more than ever and the Federal Trade Commission (FTC) reported $5.8 billion loss to fraud in 2021, a 70% rise from 2020, impacting many sectors, including healthcare (Vedova & Technology, 2022). One example, depicted in Figure 1 below, comes from the 2021 press release on national health care fraud made by the Department of Justice, which outlined the egregious types of fraudulent activities recorded that year.

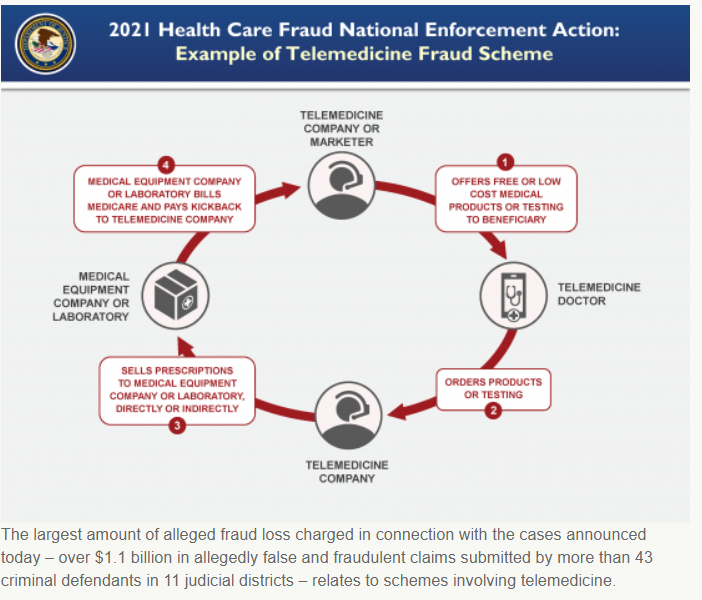


Figure - Telemedicine Fraud Scheme and Data (The US Department of Justice, 2021)

Tackling these illegal activities promises to bring huge savings to both the private sector and government healthcare coverage providers. In order to mitigate this surge in fraud, various entities in academia, insurance industry, and law enforcement - to name a few - have been looking to leverage Artificial Intelligence and Machine Learning tools to process the high volume of transactions in these fields and flag fraud and detect trends, in addition to studying various features (IBM, n.d.). This effort has been supported by the US Federal government, which in 2023 proposed at least $1.6 billion to modernize and improve assets aimed at preventing fraud (The White House, 2023).

Healthcare fraud occurs in various ways - be it double billing, phantom billing (bill for no service), unbundling (separate billing for various sub-steps of the same service), upcoding (billing for more expensive service), identity theft, forgery, and doctor-shopping as defined by the Federal Bureau of Investigations (FBI) - with offenders including medical providers, patients, and others that aim to obtain unlawful payments or benefits (FBI, 2016). The short-term consequences of these types of fraud are financial losses incurred by patients and insurance providers, while in the long run patients and taxpayers are faced with ever-rising coverage costs. On the other hand, manual review of even a small share of the claims and transactions can result in lengthy delays with the high risk of adversely impacting the patients’ health. Given the proven ability of AI-powered software systems to quickly process and categorize a large number of entries, with increasing accuracy, using Artificial Intelligence (AI) and Machine Learning (ML) techniques to screen healthcare claims and records may greatly help in quickly identifying potentially fraudulent claims, flagging medical providers who likely engage in fraudulent activities and detecting new trends, so they can be addressed before they become widespread.

While the private sector keeps its medical records secret for proprietary and legal reasons, data on government-sponsored programs can be found in the public domain and is available to demonstrate the efficiency of AI in mitigating and preventing fraud in healthcare. In this paper, due to availability of data in the public domain and existence of preliminary work connecting to fraudulent claims, the US federal Medicare program has been selected as the target for studying the effectiveness of AI on detecting healthcare fraud.

## Background

The task of applying AI to healthcare fraud is not a new venture. In fact, experts in the field have been and still are experimenting with various approaches to process datasets as quickly as possible and develop models that accurately and precisely identify potential fraud. Private insurance companies have been known to produce and use hundreds of different ML models to this end.

This paper considers a Medicare dataset released with a Creative Commons license in 2019 (Gupta, 2019) and sourced by multiple researchers in the field since then. The dataset contains 100,000s of records of health coverage claims submitted to Medicare in recent years.

Existing work done on this dataset can be largely divided into studies of the pre-processing techniques and model selection methods used during data analysis. To begin with, the dataset has a large number of columns (features) and most of them are missing at least some of the entries, even after careful feature selection, combining and addition of new ones. The work presented by Analytics Vidhya, a Data Science professionals’ community, which used the same dataset, oversamples the data using Synthetic Minority Oversampling Technique (SMOTE), after splitting the dataset into 80:20 train/test groups respectively. The training group was further subdivided into four oversampled groups with different majority:minority sets, which were then fed to Logic Regression (LR), Decision Tree (DT), Support Vector Classifier (SVC), and Naive Bayes (NB) models. Finally, the maximum area under the curve (AUC) metric was applied to each set with the results shown in Figure 2 below.

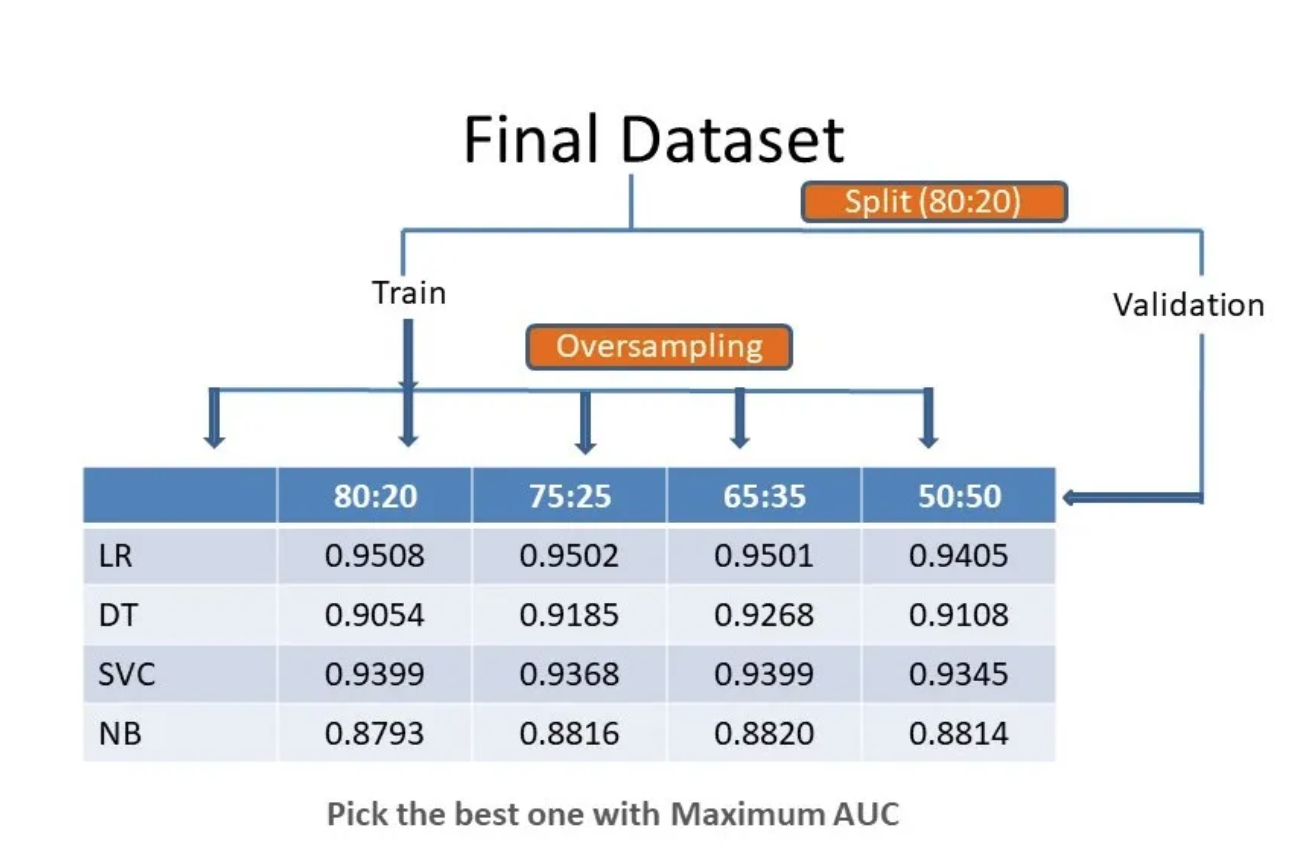


Figure - Analytics Vidhya Max AUC result of models after using SMOTE

Although the claimed maximum AUC score appears high, it is important to note that this approach did not attempt to remove the NaN entries besides the oversampling task and did not consider precision and recall metrics which are important evaluators, as classifying providers wrongly as fraudulent and vice versa is costly for the insurance, the patient, and the medical provider in question. Thus, the metrics may be overly simplistic and mislead the reader.

Another approach to the same dataset was described by Ms. Acharya, a data scientist from the nationally accredited NYC Data Science Academy, who found the LightGBM model to classify fraudulent vs. non-fraudulent claims with high precision and recall (NYC Data Science Academy, 2022). LightGBM model is a decision tree based gradient boosting framework composed of Gradient-Based One Side Sampling (GOSS) as well as Exclusive Feature Bundling (EFB) (Geeks For Geeks, 2021). It employs a histogram-based algorithm to sort continuous features into discrete bins, which shortens training times and reduces demand on memory (LightGBM, 2022). This model optimizes convergence time by following a leaf-wise method with a fixed max depth for node expansion but may be prone to overfitting. Moreover, to mitigate some of the drawbacks of using decision-tree algorithms in classification applications, rather than using one-hot encoders, LightGBM partitions categories into subsets and then focuses on obtaining best features for each subset according to training objectives - in this case, detecting potentially fraudulent providers, as shown in Figure 3 below. Ultimately, the findings from NYC Data Science Academy’s machine learning model demonstrates that the LightGBM model identified key features that enabled it to predict fraudulent cases more accurately than other basic models tried, such as logistic regression.

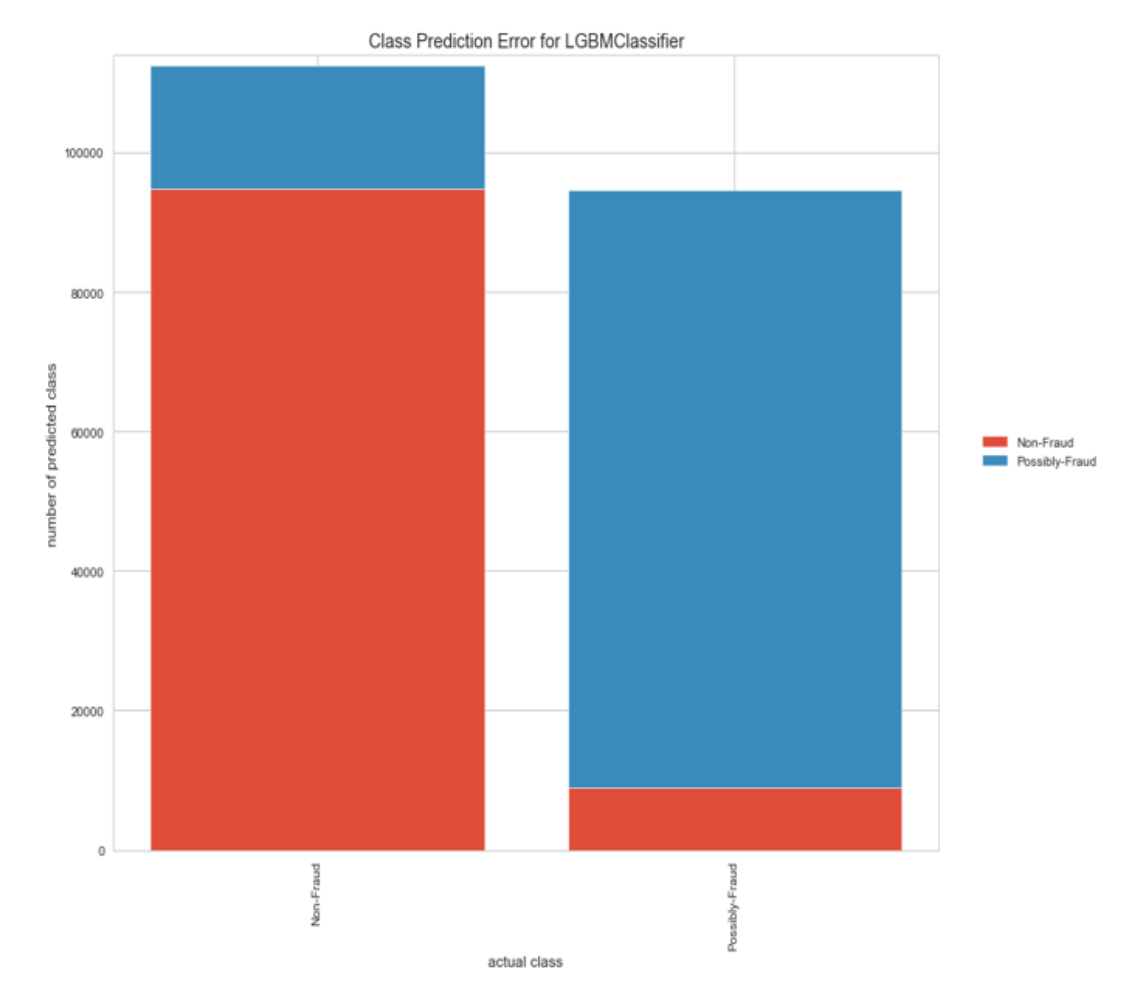


Figure - Performance of LightGBM in correctly predicting fraudulent and non-fraudulent claims (NYC Data Science Academy, 2022).

Nonetheless, although LightGBM appears to be a promising approach for this dataset, it still has key disadvantages that might limit its performing capacity to make accurate and precise classification. For instance, LightGBM uses high correlation to select features, when this does not guarantee selection of features that are most important (Khadka, 2023). Additionally, the model uses several different hyperparameters to optimize accuracy, which makes it tricky and time-consuming to pick the optimal combination. LightGBM is also prone to overfitting if it finds non-representative patterns in training data and uses them to classify and may require significant time and memory resources to work with large, complex datasets (which the Medicare claims dataset is), as it must build a large tree in memory.

In order to address the above hindrances from the models and preprocessing methods followed above, AI researchers and professionals have used the Ensemble ML model to arrange multiple models in a structured sequential order to leverage the combined effect of the models and increase performance with the assumption that weaknesses in one model would be covered for by the others (Manik, 2021).

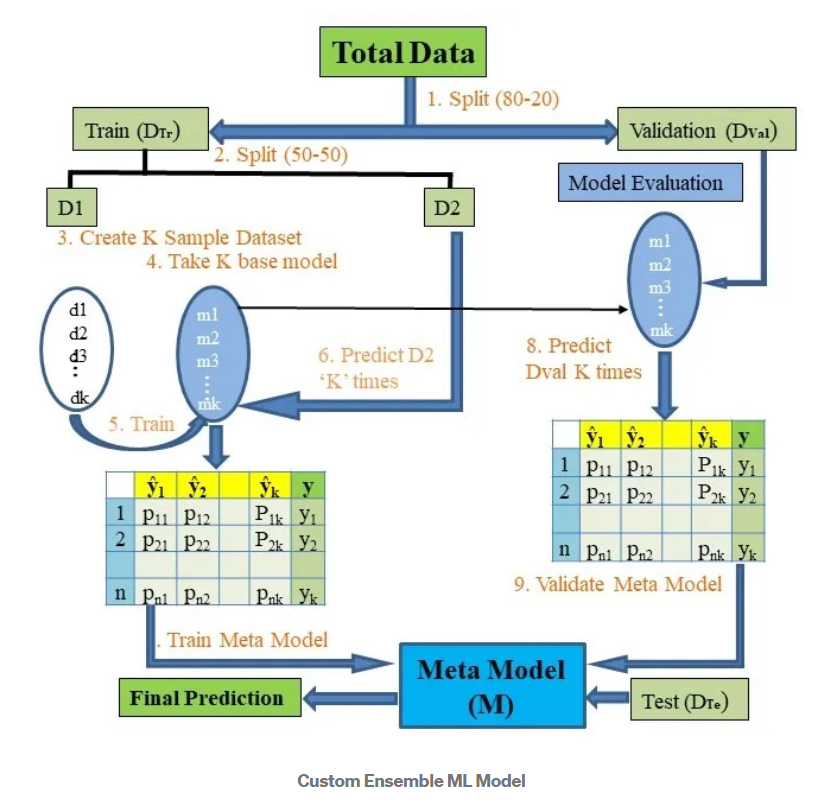


Figure - State space example of an ensemble model with n model types

Outside of this dataset, some researchers have also been working on a theoretical approach, i.e. unsupervised multi-view anomaly detection approach, to solve fraud detection challenges. Proponents of this path, Shekhar et all, argue that the unsupervised method offers interpretability and improved performance when measuring a suspiciousness score (rather than a fraud vs. non-fraud classifier) which ranks likely fraudulent providers at the top (Shekhar, Leder-Luis, & Akoglu, 2023).

This paper explains how a deep learning ensemble framework can be applied to the large Medicare dataset selected.

## Problem Statement

Fraudulent claims against the Medicare program cost the US federal government an estimated 1% of its entire annual budget (around $70B) (Cubanski & Neuman, 2023). To mitigate, the US government must try to filter out as many fraudulent claims as possible and take appropriate measures against high-volume offenders. However, as the number of claims grows, it requires more and more effort (and cost) to evaluate them and correctly identify the improper ones. On the other hand, the costs of mistakes or delays in processing of valid claims are bound to be high - from failure to provide timely healthcare to Medicare recipients, who are already likely to be in poor health due to age or chronic illness, to discouraging certain types of treatments or even diagnoses, if they are known to be likely to result in denials or delays in processing, to erosion of public trust in the government’s programs and discouraging medical providers from accepting Medicare at all.

**Problem: develop a minimally expensive, maximally accurate, tunable, scalable, quick, and efficient way of evaluating Medicare claims databases and identifying the healthcare providers whose claim practices warrant closer scrutiny.**

The target program is Medicare, but the solution may prove to be valid and relevant for other medical insurance programs, both public and private, and may even be applicable to foreign-based programs.

## Objectives

Our first objective is to develop a working neural-network ensemble model to address the problem stated above.

Our second objective is to evaluate the performance and the utility of our model and methodology against existing ones, including traditional, non-AI, non-ML ones, and generate some conclusions on applicability of certain types of models and / or methods to problems, such as this.

Our overarching general objective is to learn and to generate new knowledge. It may be knowledge specific to the issue of Medicare fraud or insurance fraud in general, data science knowledge, technical knowledge about the tools and methods used in the project or a combination thereof.

## Methodology

We began by researching the problem, identifying an appropriate dataset, and exploring the work completed by others in this problem space. Once the dataset was identified, the next step was to conduct exploratory data analysis (EDA) and cleaning of the data. This was done in parallel with exploration of other works for time efficiency.

## Data Analysis

An initial look at the data revealed the complexity of the dataset (56 columns in the combined raw database) and a high percentage of “NaN” or empty entries in many of the columns, particularly procedure and diagnostic codes.

### DATA MODELING AND PROCESSING

Data exploration and modeling is an integral part of Machine Learning development lifecycle. In these steps, the data are carefully explored, curated, cleaned, and transformed while preserving the maximum amount of variance in the data.

#### A. Data Gathering

The data were taken from an open-source Medicare records dataset, publicly available on the Kaggle platform. The dataset contained columnar data in 56 total columns. The dataset was split into 3 separate comma-separated files: inpatient records, outpatient records, and beneficiary data.

#### B. Data Exploration and Cleaning

Data exploration had these primary steps:

1) Statistical Analysis. Statistical analysis inspects all the numerical columns and finds common metrics, such as mean, median, mode, 25th percentile, maximum, standard deviation, null/nan, internal variance analysis, etc. In Python this can be easily done for numerical data types with the df.describe command. The results are shown for each section of data in the figures below, along with some quick analysis.

Figure - Statistical descriptions of the numerical beneficiary data columns

A picture containing text, screenshot

Description automatically generated

A picture containing text, screenshot, line

Description automatically generated

The statistical descriptions of beneficiary data immediately reveal that gender and all of the chronic condition columns have only binary values. The majority of the beneficiaries are race 1.

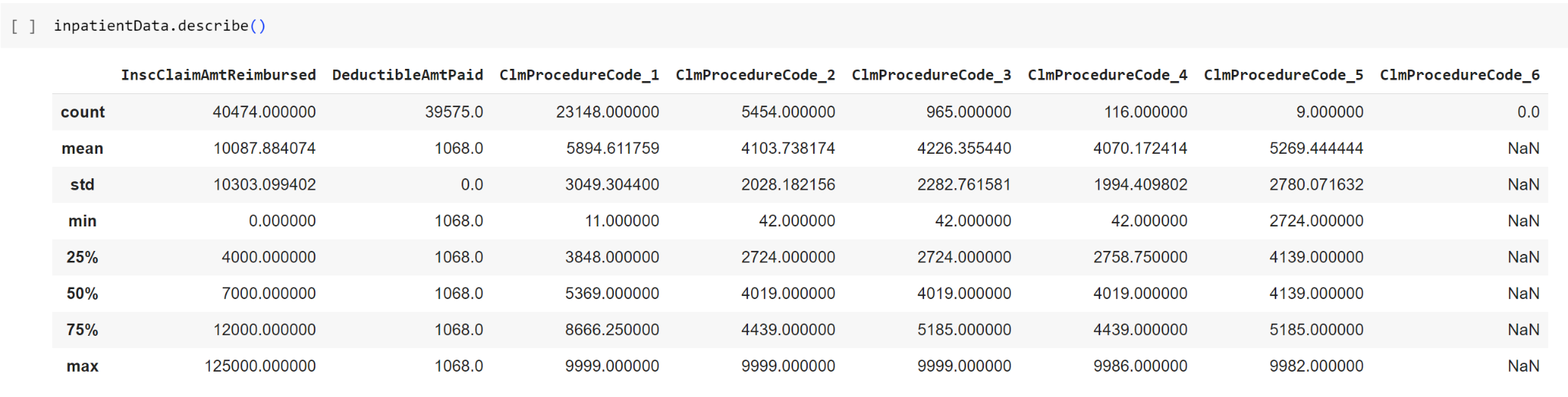


Figure - Statistical descriptions of the numerical columns in Inpatient Data

For inpatient data, it is clear that the deductible amount paid is always the same. This may be peculiar to the Medicare program and another program’s dataset may not exhibit the same finding. Claim procedure codes 2 and 3 have very similar distributions, though there are many more claims with procedure claim code 2. Claim procedure code 6 is never used.

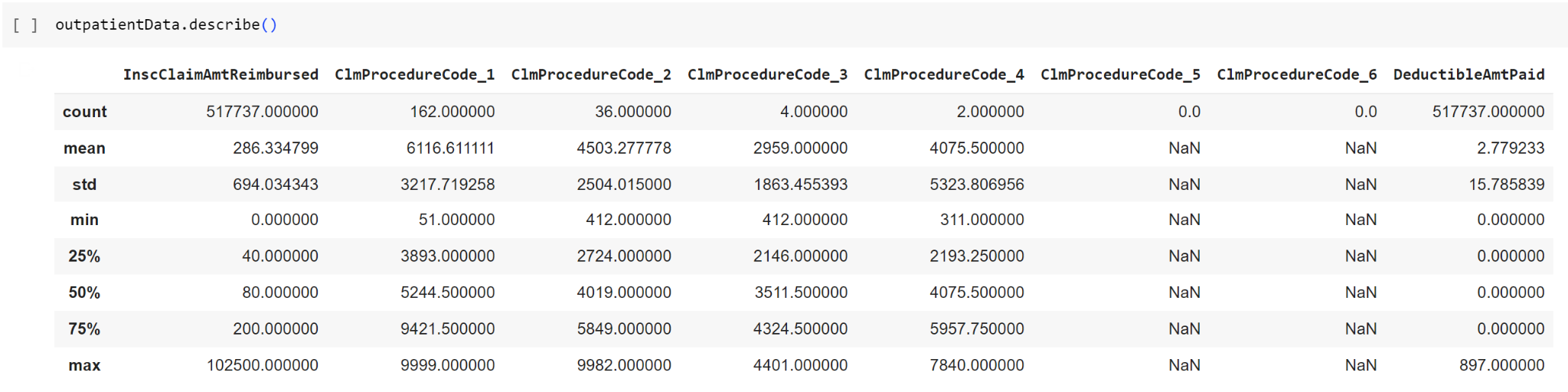


Figure - Statistical description of numerical columns in Outpatient Data

In the outpatient data the deductible amount paid is variable. Procedure codes 5 and 6 are always “NaN”. The amount that is reimbursed is usually low, with an average of $286, but there are some extreme outliers with the maximum reimbursement at $102,500.

2) Visualizations. Several visualizations such as histograms, pairplots, heatmaps, and pie charts were plotted to see the visual inferences from the data. Those are highlighted below with some brief analysis for each section of data.

For beneficiaries, much of the data was segmented into binaries. Race, state, and county were not binary, but had a discrete number of values, so they lent themselves easily to pie plots and histograms as shown in the figure below.

Figure - Pie-Charts and Histograms of Beneficiary Data

A picture containing colorfulness, circle, diagram, design

Description automatically generated

A picture containing text, diagram, line, parallel

Description automatically generated

Inpatient and outpatient data have several procedure or diagnosis code columns. The columns for both data frames are the same, but the values vary between the two. For instance, in inpatient data the deductible amount paid is always the full amount, which is the same for each claim. In outpatient, the full amount may not be paid, so the deductible amount did vary, albeit slightly.

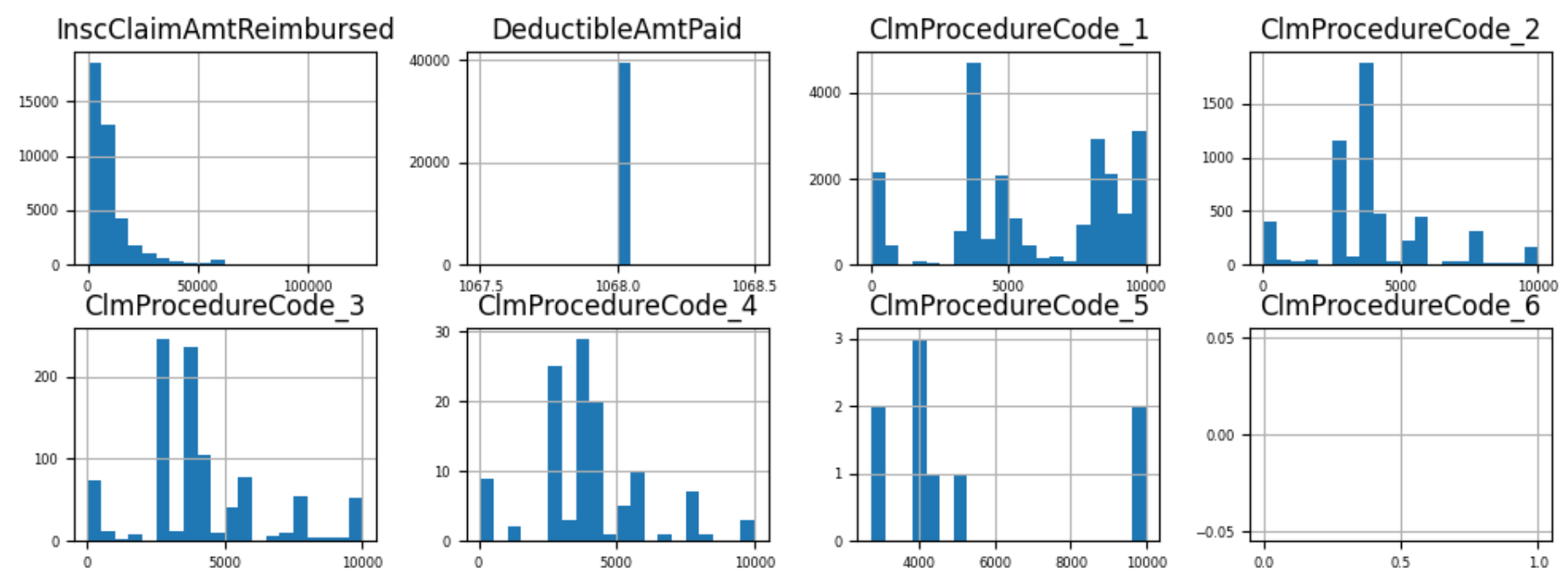


Figure - Histograms of Numerical Features of Inpatient Data

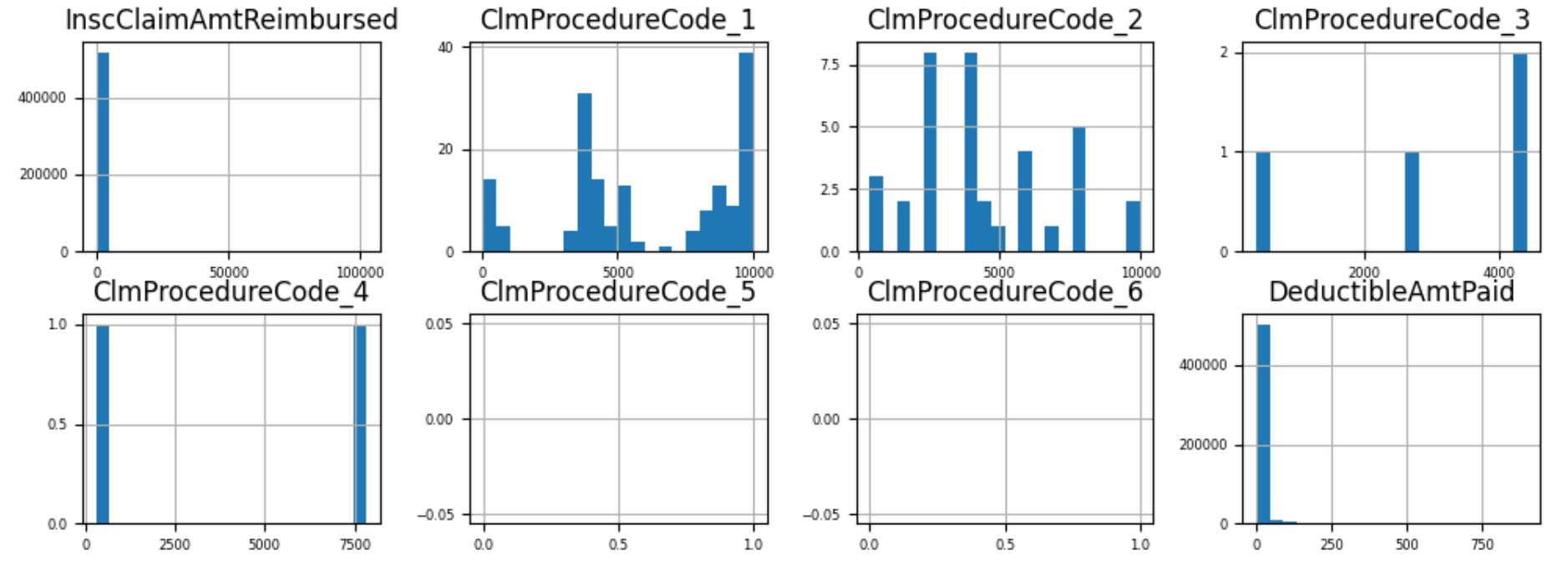


Figure - Histograms of Numerical Features in Outpatient Data

It can also be noted that claim procedure codes 5 and 6 were never used for outpatient claims. Only claim 6 was unused in the inpatient claims data.

3) Feature Engineering. Several new features were created from the existing features or by combining several features by application of several arithmetic operations: addition, mean, etc. For example, the column “claim\_time” was created from admission and discharge time, and “chronic\_count” was created by accumulating the count over several chronic diseases (8 columns) for each patient.

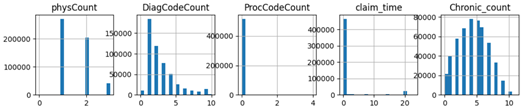


Figure - Histograms of Newly Created Features

4) Data Cleaning. The AutoML library Dabl was used to segregate data into different types: continuous, float, low and high cardinality integer, categorical, etc. According to the results found in the null and internal variance analysis, the columns with small variance and high number of nulls were dropped.

5) Data Transformation. Several features were normalized and transformed into numerical values. Binarization was applied to the categorical columns that had just 2 unique values. Frequency encoder was applied to the numerical columns with high cardinality. Label Encoder was used on categorical columns. Some type conversion was also applied on columns such as DeductibleAmount and claim\_time. Finally, MinMax Scaling and Standard scaler were applied to normalize the data accordingly.

6) Data Sampling. The initial unbalanced dataset was balanced using the Adaptive Synthetic Sampling (ADASYN) (Herland, Khoshgoftaar, & Bauder, 2018) method. This method is beneficial over the typically used sampling methods, such as Synthetic Minority Oversampling Technique (SMOTE) or its variants: SMOTE-Tomeks or SMOTE-ENN (Bauder & Khoshgoftaar, 2017), because of its ability to balance the synthetic samples according to the cluster densities amongst the minority label.

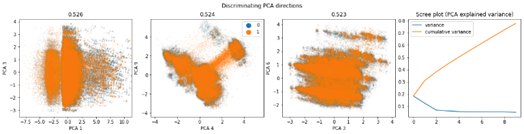


Figure - Model Clusters

7) Feature Selection. Features were finally selected using the Feature Importance score computed using the RandomForest method. The finally selected columns were concatenated with the ensemble predictions.

#### C. Ensemble Predictions

The Ensemble model was used to learn predictions from the weak learners. Several weak learner algorithms: CatBoost, XGBoost, AdaBoost, GaussianNB were used to make preliminary predictions and then pass them to the neural network. These predictions were connected to the feature columns using the column\_stack function in the numpy library.

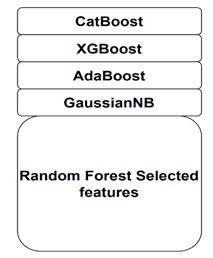


Figure - Weak Learner Algorithms

#### D. Branched Neural Network

The cleaned and transformed features were fed to a novel artificial deep neural network. The neural network utilized the concepts of branching layers and recursive gradient transfer to optimize the convergence time. As shown in Figure 13, the gradients from several dense layers were merged into the concatenation layer. Regular application of the dropout layer prevented the chances of overfitting and BatchNormalization layer tried to normalize the transfer gradients, so that the convergence of the model to optimum happens through a smooth process. Inbuilt regularization (L1) was also used in the dense layers to penalize the wrong results and apply an internal feature selection to the model by reducing coefficients of the less important variables to almost zero.

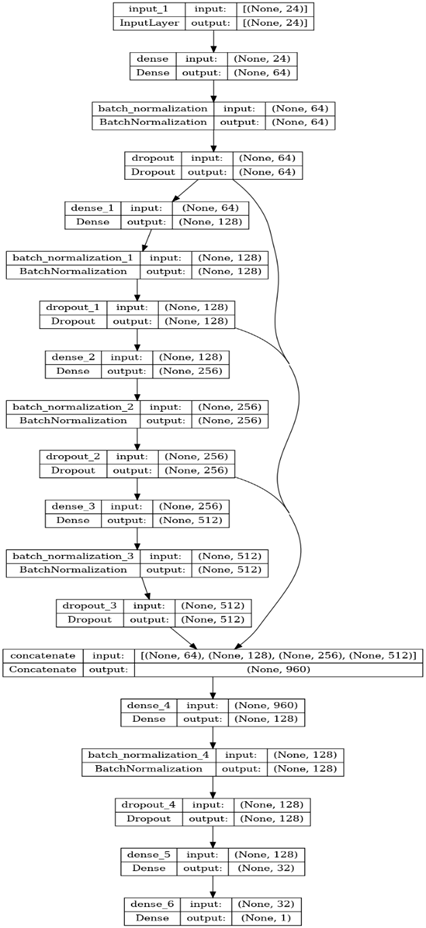


Figure - Diagram of Deep Neural Network

Callbacks, such as CSVLogger, ModelCheckpoint, ReduceLRonPlateau (Matschak, Prinz, Rampold, & Trang, 2022) and EarlyStopping were used in the compilation process to keep a check on validation accuracy and loss. These callbacks help to stop the learning cycle whenever the model starts to overfit. ReduceLROnPlateau reduces the learning rate on hitting a plateau.

## Results

After careful training and evaluation process, the resultant validation accuracy of 82.268% was achieved on the balanced dataset. A perfect balance of precision and recall was maintained over all the labels. As shown in Figure 15, the precision score for the possible fraud and non-fraud label were 0.79 and 0.86 respectively. The F-1 score, which is the harmonic mean of precision and recall was also computed for separate classes to have more visibility into the results.

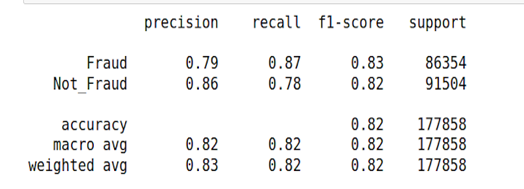


Figure - Table of metrics

As shown in Figure 16, the model attains a Receiver Operating Characteristic (ROC) score of 0.82. The ROC score is insensitive to class imbalances and plots the relation between true and false positive rates.

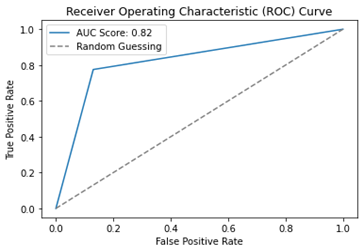


Figure - Receiver Operating Characteristic Curve

Figure 17 shows the loss curve for the first 20 epochs. A total of 148 epochs of training were required for the model to properly converge.

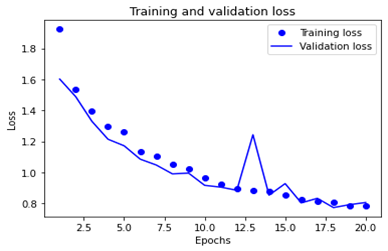


Figure - Training and Validation Loss

Table 1 shows the comparison of the implemented model to the previously proposed solutions in the research world. Other researchers used supervised and unsupervised Machine Learning and Neural Networks to combat the problem of fraudulent healthcare claims. These approaches mainly relied on limited feature engineering and SMOTE for sampling the data. Our implemented solution (“ClaimSafe Health”) defeats all the mentioned research works by a considerable margin.

Table - Comparison of Results

|  |  |  |
| --- | --- | --- |
| **Author** | **Proposed Solution** | **Best Metrics** |
| Matthew[12] | Machine Learning | Acc. 0.816 |
| Richard[13] | Machine Learning | Acc. 0.59 |
| Tizian[14] | Conv. Neural Network | AUC: 0.70 |
| Richard[15] | Unsupervised Machine Learning | AUC: 0.63 |
| Our Solution  (ClaimSafe Health) | Ensemble-Neural Network | Acc. 0.82268  AUC: 0.8225 |

Apart from the AI-powered novel solution, the final deliverables also include a power BI dashboard that would provide a holistic and comprehensive overview of the fraudulent cases spread along the tangents of State, County, and Provider. This would enhance a non-technical user's visibility into the problem and help them make judicious decisions. Armed with accurate data that is presented in a clear and easy-to-understand format, people can avoid visiting providers that have a high number of fraudulent cases in their state or county. Also, metrics, such as average processing time in their states or counties could help patients choose providers based on their efficiency. Having an exact count of patients with chronic diseases (such as cancer, diabetes, etc.), would help policymakers stay in touch with reality on the ground and enact optimal policies.

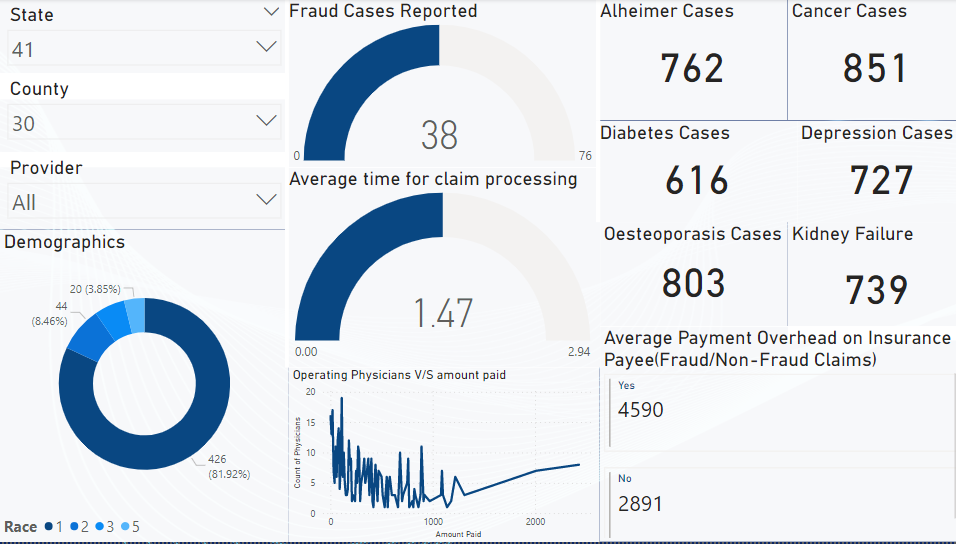


Figure - Screenshot of GUI Dashboard

## Potential Issues

Before beginning the project, there were a host of potential technical challenges. It was possible that the available data would be faulty, biased or difficult to use. The Kaggle dataset (Gupta, 2019) found for this problem has been used broadly, so this was unlikely to be an issue with this dataset. Interestingly, the test portion of the dataset was not labeled, so the amount of usable data was reduced. The training portion of the data included over 711,000 claims of data, which was more than sufficient to test and train the model. Overfitting, underfitting, poor precision, and poor recall are always among the concerns to steer clear of in projects such as this. The model was created and tuned with these potential issues in mind, which allowed for their mitigation.

There are additional challenges which come along in the application and usefulness of fraud detection. Assuming that the fraud detection is implemented by an insurance provider, they may wish to catch every possible fraudulent claim, which would mean tuning the model to minimize false negatives at the expense of including more false positives. This could harm honest providers (and their patients), burdening them with additional paperwork and delays in verification of claims. Changes in policies or even just the implementation of fraud detection could affect the incoming data and drive down the accuracy of models trained on historical data. This would require additional data and model training to correct.

## Timeline and Collaboration

This project was conducted over the course of four weeks during the Spring 2023 semester. The team had the time from submission of the proposal the week of April 11th, until the presentation on May 2nd to complete the project. Given there was an exam on the 25th, this left 3 weeks of time during which the project could be worked. Zoom meetings were held a couple times a week, on nights and weekends, to share the findings. The team collaborated utilizing Google Documents and colab notebooks, which both allow for simultaneous viewing and editing. The first week each person was tasked with finding four references on the subject of healthcare fraud. These were compiled into a Google Doc and discussed on Zoom. The exploratory analysis was also initialized. The 2nd week was an off week, during which the team was studying for a class exam. After the exam, exploratory data analysis was finalized. Model building and adjustments went from the 3rd week into the 4th and final week before the presentation. Team members were able to work the presentation slides in parallel, adding the final pieces of results once the model runs were complete.

## Resources

As discussed in the previous section, the team utilized Google Drive, Google Documents and Google Colaboratory (Colab Notebooks) to coordinate. The dataset was sourced from Kaggle. Coding was completed in Python using standard libraries: numpy, pandas, and matplotlib, and a few more specialized libraries, including Seaborn, SciKitLearn, Imbalanced Learn, and LightGBM.

## 

## Conclusion

Presented is a novel ensemble neural network approach. The research combines the modern concept of recursive gradient transfer, branched learning to the healthcare domain. The EDA and data selection done in the research work carefully analyzed the impact of variance, cluster densities and impact of nulls before making any assumptions (about the model). Data were analyzed by observing several graphs and statistical analysis reports, which led to the final neural network model, one that attains an accuracy of 0.822, beating several state-of-the-art solutions present on the market. Proper training pipelines ensured that optimal convergence occurs in the network with the least impact of noise in the data and overfitting. Results were measured using several metrics, including precision, recall and F-1 score to reduce the impact of any internal biases in the dataset. The metrics were noted down with help of callbacks that automated most of our training pipeline work (check overfitting, converge properly, adjust learning rate, etc.)

### Future Scope

The following items can be made in the future iterations of this solution:

* More time could be spent on the hyper-parameter tuning. Infinite search space can be explored to find the right combination of features, with the limits in computational power in mind.
* Manifold learning visualizations can be plotted to see the cluster organization in the data. Algorithms such as T-distributed Stochastic Neighbor Embedding (T-SNE) can plot out non-linear dependencies in the features.
* Continuing to grow and update the dataset to maintain accuracy and currency of the model. This can be done periodically, as larger or supplementary datasets appear in the future.

## References

Bauder, R. A., & Khoshgoftaar, T. (2017). Medicare Fraud Detection Using Machine Learning Methods. *16th IEEE Conference on Machine Learning and Applications (ICMLA)* (pp. 858-865). Cancun, Mexico: IEEE. doi:10.1109/ICMLA.2017.00-4

Bauder, R., Da Rosa, R., & Khoshgoftaar, T. (2018). Identifying Medicare Provider Fraud with Unsupervised Machine Learning. *2018 IEEE International Conference on Information Reuse and Integration (IRI)* (pp. 285-292). Salt Lake City, Utah: IEEE. doi:10.1109/IR

Cubanski, J., & Neuman, T. (2023, January 19). *What to Know about Medicare Spending and Financing*. Retrieved from KFF: https://www.kff.org/medicare/issue-brief/what-to-know-about-medicare-spending-and-financing/

FBI. (2016, June 1). *Health Care Fraud*. Retrieved from FBI.gov: https://www.fbi.gov/investigate/white-collar-crime/health-care-fraud

Geeks For Geeks. (2021, December 22). *LightGBM (Light Gradient Boosting Machine)*. Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/

Gupta, R. A. (2019, May 9). *Healthcare Provider Fraud Detection Analysis*. Retrieved from Kaggle: https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis?select=Test\_Outpatientdata-1542969243754.csv

Herland, M., Khoshgoftaar, T., & Bauder, R. (2018, May 29). *Big Data Fraud Detection using Medicare Data Sources*. Retrieved from https://doi.org/10.1186/s40537-018-0138-3

IBM. (n.d.). *AI Improves Fraud Detection, Prediction, and Prevention*. Retrieved from IBM.com: https://www.ibm.com/analytics/fraud-prediction

Khadka, N. (2023, April 5). *The Key to Winning Machine Learning Competitions*. Retrieved from DataAspirant.com: https://dataaspirant.com/lightgbm-algorithm/#t-1679668681670

LightGBM. (2022). *Features - LightGBM 3.3.2 documentation*. Retrieved from LightGBM: https://lightgbm.readthedocs.io/en/v3.3.2/Features.html

Manik, A. (2021, June 23). *Healthcare Provider Fraud Detection Analysis using Machine Learning*. Retrieved from Medium.com: https://medium.com/analytics-vidhya/healthcare-provider-fraud-detection-analysis-using-machine-learning-81ebf09ed955

Matschak, T., Prinz, C., Rampold, F., & Trang, S. (2022). *Show Me Your Claims and I'll Tell You Your Offenses: Machine Learning-Based Decision Support for Fraud Detection on Medical Claim Data*. doi:10.24251/HICSS.2022.454.

NYC Data Science Academy. (2022, July 5). *Healthcare Fraud Detection*. Retrieved from Data Science Blog: https://nycdatascience.com/blog/student-works/machine-learning/healthcare-fraud-detection-2/

Shekhar, S., Leder-Luis, J., & Akoglu, L. (2023, February 13). *Unsupervised Machine Learning for Explainable Health Care Fraud Detection*. Retrieved from NBER: https://www.nber.org/papers/w30946

The US Department of Justice. (2021, September 20). *2021 National Health Care Fraud Enforcement Action*. Retrieved from US Department of Justice: https://www.justice.gov/criminal-fraud/2021-national-health-care-fraud-enforcement-action

The White House. (2023, March 2). *Fact sheet: President Biden’s sweeping pandemic Anti-Fraud Proposal: Going after systemic fraud, taking on identity theft, helping victims*. Retrieved from WhiteHouse.gov: http://www.whitehouse.gov/briefing-room/statements-releases/2023/03/02/fact-sheet-president-bidens-sweeping-pandemic-anti-fraud-proposal-going-after-systemic-fraud-taking-on-identity-theft-helping-victims/

US Centers for Medicare & Medicaid Services. (2023, February 17). *NHE Fact Sheet*. Retrieved from CMS: https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet

Vedova, H., & Technology, &. (2022, February 22). *New data shows FTC received 2.8 million fraud reports from consumers in 2021*. Retrieved from Federal Trade Commission: https://www.ftc.gov/news-events/news/press-releases/2022/02/new-data-shows-ftc-received-28-million-fraud-reports-consumers-2021-0