National College of Ireland

Optimizing Risk Stratification and Premium Allocation for Health Insurance Companies via Supervised Ensemble Learning

MSc Research Project
Data Analytics

Trinita John Peter
Student ID: x23266848

School of Computing
National College of Ireland

Supervisor: Vikas Tomer

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Trinita John Peter
Student ID:	x23266848
Programme:	Data Analytics
Year:	2024-2025
Module:	MSc Research Project
Supervisor:	Vikas Tomer
Submission Due Date:	11/08/2025
Project Title:	Optimizing Risk Stratification and Premium Allocation for Health Insurance Companies via Supervised Ensemble Learning
Word Count:	5928
Page Count:	19

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:	Trinita John Peter
Date:	10/08/2025

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):	

Optimizing Risk Stratification and Premium Allocation for Health Insurance Companies via Supervised Ensemble Learning

Trinita John Peter
x23266848

Abstract

As per the report from 'The Irish Times' published on March 09, 2025, the price revision rolled out by the State's private health insurance companies is heavily impacting both the Irish insurers and policyholders financially. Considering these growing financial pressures, the industry is now seeking innovative ways to predict and manage risks more effectively. This paper explores supervised ensemble machine learning techniques– Gradient Boosting, eXtreme Gradient Boost (XGBoost) and Random Forest that offer a promising pathway towards greater equitable risk assessment and pricing fairness. The proposed framework performs high level feature engineering to flag claims based on upper quartile threshold, bin significant categories and capture claim interaction effects by combining relevant attributes. The results demonstrate an exceptional predictive performance by XGBoost, across all instances with just 1 misclassification, strongly identifying high risk and low risk profiles, and a Reciever Operating Characteristic- Area Under Curve score (ROC) more than 95% confirming the model is well balanced and trustworthy. This engineered model serves as a prototype to streamline solutions on risk stratification, ultimately helping to stabilize insurance premiums for consumers.
Keywords: Patient attributes, health variables, hyper-parameter tuning, risk anticipation, transparent policy pricing.

1 Introduction

Irish Healthcare system offers a tax financed system, ensuring every citizen has access to medical services at a nominal user fee, but just about 50% of the population hold private policyholders. While private insurance companies potentially create inequity among citizens, some of the major benefits they offer include shorter waiting times, additional benefits in dental/optical and enhanced comfort with access to premium amenities. (NDTP; 2024)

The **research** on "Optimizing Risk Stratification and Equitable Allocation for Health Insurance Companies via Supervised Ensemble Learning" seeks to bridge the gap between this asset misallocation, inequity from private insurance and detecting anomalies in claim amount.
The paper aims to address and propose a solution to the following **research question**: "Can the insurance firms better anticipate risks and allocate their assets resourcefully

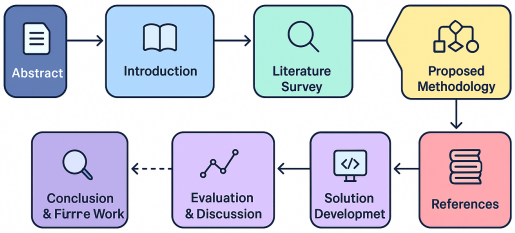


Figure 1: Structure of the report

based on significant patient attributes containing key demographic and health-related variables using supervised ensemble Machine Learning techniques and accurate model evaluation thereby providing fairer and accurate rates for policyholders?". While the analytical approach centers on building a robust and highly reliable risk assessment machine learning framework using supervised ensemble technique (Barbiero et al.; 2024), strong emphasis is placed on addressing class imbalances, executing fine-grained hyperparameter optimization (Abdelshafy et al.; 2024), and facilitating the extraction of empirically driven with statistically valid conclusions.

Objectives and potential outcomes of the work in the field of research are identification of fraudulent or error-prone claims, patterns in high-cost treatments that are often rejected. The trained model may inform in preventative analytics before reimbursement decisions, provider risk profiling, flag patterns of fraud and ultimately stabilize insurance premiums which is a key concern among Irish insurers facing cost escalations. Here is an outline of the structure of report in Figure [1]

2 Related Work

A rigorous literature review of the state of the art is carried out to facilitate selection of evidence-based methodology that compliments the research question, objectives, helps identify key theories with findings and reveal areas that need further investigation.

2.1 Risk Assessment and Underwriting

The structure of the Indian healthcare system follows a dual system, just like in most regions, where public services are free or subsidized while private healthcare is paid out of pocket or via insurance. The journal paper (Abdelshafy et al.; 2024) uses ML model techniques used to predict claim amount by individuals, reduce fraud-related financial losses and predict future claims that help insurance firms better manage risk and allocate resources using GMM. (Alam and Prybutok; 2024) performs an analysis to identify the top contributing factors that increase health insurance claim amounts using PCA, Pearson correlation, multiple regression and SVM.

(Vangala; 2025) This paper aims to create a Machine Learning model to predict the claim amount and out-of-pocket expense of-pocket expense of holders. The conference paper analyses the potential of AI in healthcare insurance claim processing in the US with an aim to determine effective model for predicting claims to save cost for insurance companies. Feature importance is conducted which forecasts customer claim insurance process. 6 machine learning algorithms utilized are (SVM), decision tree (DT), random forest (RF), linear regression (LR), extreme gradient boosting (XGBoost), and k-nearest neighbors (KNN).

Paper (Pragatheeswari et al.; 2025) leverages more sophisticated predictive analytics to generate informed estimations of the most consequential healthcare trends anticipated in the United States throughout the next decade. This Predictive modelling has greatly helped in forecasting future trends in resource utilization and costs. Visualization is done to determine greatest feature impact in (Onwubuariri et al.; 2024). A machine learning based regression model to predict health insurance claim based on different health variables data is widely discussed in where to streamline the expectation interaction of the medical services protection costs in this review, a determination strategy is performed to separate the main elements.

The conference paper (Yang and Wu; 2006) analyses the potential of AI in healthcare insurance claim processing in the US with an aim to determine effective model for predicting claims to save cost for insurance companies. Feature importance is conducted which forecasts customer claim insurance process. 6 machine learning algorithms utilized are (SVM), decision tree (DT), random forest (RF), linear regression (LR), extreme gradient boosting (XGBoost), and k-nearest neighbors (KNN).

Paper (Veena et al.; 2023) leverages more sophisticated predictive analytics to generate informed estimations of the most consequential healthcare trends anticipated in the United States throughout the next decade. This Predictive modelling has greatly helped in forecasting future trends in resource utilization and costs. Visualization is done to determine greatest feature impact in (Paikaray et al.; 2023). A machine learning based regression model to predict health insurance claim based on different health variables data is widely discussed in (Paul; 2024) where to streamline the expectation interaction of the medical services protection costs in this review, a determination strategy is performed to separate the main elements.

The conference paper (Prova; 2024) follows a more traditional approach on deciding whether a person has the right to claim insurance offered by the organization in case of a particular health risk or not. Machine learning handles this particular task in this scenario with various classification algorithms. Fraud detection, risk assessment, claims processing, customer service, disease management are some of the roles played in the market by AI/MLs. Classification algorithms such as logistic regression, naïve bayes, SVM, KNN and decision tree are employed. A practical solution to fraudulent claim detection and risk classification using machine learning concepts like SVM, logistic regression, decision tree and KNN is proposed in paper (Ramani et al.; 2024), and the results prove that decision tree outperforms. The study demonstrates that AI models can improve fraud detection accuracy by up to 30%, significantly reduce operational costs, and provide real-time fraud prevention capabilities. Ensemble models, which combine

the predictions of multiple machine learning algorithms, offer another powerful tool for fraud detection. These models enhance accuracy and reduce the risk of false positives, as they are better at capturing diverse fraud patterns.

2.2 Premium Prediction and Preventive Insights

An approach to build a robust framework to detect and prevent health insurance frauds by leveraging predictive models on a 1300 dataset organized into 7 columns: charges, smoking, region, children, body (BMI), sex, and age is discussed in (Saraswat et al.; 2023). To improve forecast accuracy, add more features, and investigate sophisticated ML approaches like DL models or ensemble methods can be considered. A study to extract distinct information about patient claim history with combined applications of Block chain and Machine Learning is performed by Sandip Vyas in (Sharma and Jeya; 2024). It is an implementation for multi-class classification of insurance claims and Random Forest produces effective outcomes on input data.

Atharv Sharma performs a deep analysis on Forecast insurance costs using the model that exhibits the highest accuracy, in particular XGBoost. In order to improve the functionality of predictive models and make it accessible, a web application is developed (Vyas et al.; 2022).

To address the inaccuracy caused by poor data quality, methods to better handle the unstructured data are shown using SVM that also reduces the number of false negatives, which is discussed in (Barbiero et al.; 2024). The role of AI in shaping the future of automated insurance processes with a focus on the impact of machine learning driven AI models and claim processing is discussed in (Tanwar et al.; 2019). With AI, insurers can now use real-time data to create dynamic pricing models that reflect true risk, leading to fairer premiums which would require significant investment in technology and expertise.

In this paper (Bärtl and Krummaker; 2020), the dataset with 24 features including all relevant attributes needed for prediction of insurance cost was used. The implementation is done using regression algorithms such as Linear Regression, Decision Tree Regression, Lasso Regression, Ridge Regression, Random Forest Regression, ElasticNet Regression, Support Vector Regression, K Nearest Neighbor Regression and Neural Network Regression in R Programming. There were seven metrices applied to measure the performance of the model namely Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), R squared (R2), Adjusted R-Squared (Adj.), and Explained Variance Score (EVS). Random Forest Regression outperformed with 0.9533 as R squared value.

This study (Matloob et al.; 2020) provides a computational intelligence approach for predicting healthcare insurance costs. The proposed research approach uses Linear Regression, Support Vector Regression, Ridge Regressor, Stochastic Gradient Boosting, XGBoost, Decision Tree, Random Forest Regressor, Multiple Linear Regression, and k-Nearest Neighbors A medical insurance cost dataset is acquired from the KAGGLE repository for this purpose, and machine learning methods are used to show how different regression models can forecast insurance costs and to compare the models' accuracy. The results show that the Stochastic Gradient Boosting (SGB) model outperforms the others

with a cross-validation value of 0.0.858 and RMSE value of 0.340 and gives 86% accuracy. The final study (Dhieb et al.; 2020) analyzes data mining strategies in the creation of a predictive model for the detection of auto insurance fraud. The main objective of this research is to build a predictive machine learning model for the detection of fraudulent activity using the most influential features from the chosen dataset. In existing literature, the researchers used various models for detecting the most influential features for fraud detection; instead, we implemented the novel Boruta algorithm to uncover the most significant features for use in fraud detection.

2.3 Summary

In summary, following the literature survey, three ensemble supervised models are adopted-Random Forest, Gradient Boosting and XGBoost that proved to be powerful in tackling complex challenges in health insurance sectors. However, many studies use default ML model settings without rigorous tuning, leading to suboptimal performance, hence suitable feature engineering along with hyperparameter tuning techniques are introduced for optimizing and obtaining a reliable risk assessment model that reduces computational costs and facilitates scalability.
Personal contributions include:

1. Framework threshold tuning, a hyperparameter technique to adjust the decision boundary in classification models when dealing with imbalanced data and optimizing for specific metrics. The threshold value is set to 0.7 in this case, such that the model can effectively flag claim values that are greater than the defined upper quartile range.

2. 'RandomizedSearchCV' is a practical way to automate the different hyperparameter combinations, unlike 'GridSearchCV' optimize model's performance, which has been employed here with a custom value of 0.3-fold cross validation.

3. Additionally, logarithmic transformation on skewed data is applied on the target feature 'ClaimAmount', to normalize right skewed distributions, to reduce the impact of outliers and improve classification tasks.

Table 1: Summary of existing models

Authors & Year	**Datasource**	**Methodology**	**Futurework**
(Abdelshafy et al.; 2024)	3 datasets - general patient data, claim data and hospitalized data.	ML model techniques (regressor - GBR) with 90% performance to predict claim amount by individuals, reduce fraud-related financial losses.	larger dataset to demonstrate anomaly detection
(Alam and Prybutok; 2024)	Insurance Claim Analysis: Demographic and Health	Analysis to identify the top contributing factors that increase health insurance claim amounts.	Include related independent variables to ensure the comprehensiveness of the study
(Barbiero et al.; 2024)	Include related independent variables to ensure the comprehensiveness of the study	Aims to create ML to predict the claim amount and out-of-pocket expense of holders.	Efforts can be directed towards refining machine learning models specifically for premium calculation.
(Bärtl and Krummaker; 2020)	Insurance Claim Analysis: Demographic and Health	Potential of AI in healthcare in US. Aims for predicting claims to save cost for ins comps	Employing more ML models to discover the best AI model to predict claim success
(Dhieb et al.; 2020)	National Public Health statistics	Leverage predictive analytics to generate informed claim estimations in the United States.	Utilize data spanning from 2015 to present
(Paul; 2024)	11,200 items divided into: 'fraudulent and 'non-fraudulent.	In this work, an effort is made to provide a review of healthcare industry frauds and to spot them.	Insurance needs a cloud association to manage vast amount of data and automated claim processing.
(Matloob et al.; 2020)	Medical Cost Personal Datasets	Prediction regarding health ins claim based on diff factors.	Improvement by fine-tuning the parameters in future iterations to further increase its accuracy.
(Onwubuariri et al.; 2024)	Kaggle	This study involves the use of classification algos to evaluate a claim amount is received or not.	Heuristic optimization approaches- machine learning parameters and a deep-learning neural network
(Paikaray et al.; 2023)	Healthcare dataset-40M admission records of 15000 patients	Solution to fraudulent claim detection and risk classification with svm, log reg, dt and knn.	Fraud detection method can be extended to the Adaptive Neuro-Fuzzy Inference System (ANFIS)
(Pragatheeswari et al.; 2025)	Historical Claims data, Customer data and external datasets	This study improves fraud detection accuracy by up to 30	Challenge lies in dealing with missing data and unstructured data, such as social media posts, which calls for NLP
(Prova; 2024)	1300 records in the dataset organised into 7 columns	Robust framework to detect and prevent health insurance frauds by leveraging predictive models.	To improve forecast accuracy add more features, and investigate sophisticated ML approaches like DL models or ensemble methods.
(Ramani et al.; 2024)	Healthcare Providers Data For Anomaly Detection	The study was to extract distinct information about patient claim history.	The system can be implemented for multi-class classification of insurance claims
(Saraswat et al.; 2023)	Kaggle - Medical Cost Personal Datasets	Forecast insurance costs using the model that exhibits the highest accuracy.	Web applications can further improve the accuracy and functionality of predictive models.

Table 2: Con't - Summary of existing models

Authors & Year	**Datasource**	**Methodology**	**Futurework**
(Sharma and Jeya; 2024)	Public Insurance Dataset	Predict insurance cost claims.	Involve better model interpretability, better handling of the unstructured sources of data
(Suneel et al.; 2025)	Insurance Dataset	This article explores the evolution of entity resolution approaches	The implementation of AI-driven entity resolution for a strategic investment.
(Tanwar et al.; 2019)	Auto insurance pricing and claims	Role of AI in shaping the future of auto insurance	Incorporate even more advanced data sources, such as biometric data
(Vangala; 2025)	Kaggle and contributed by ACME, comprises of 1,338	Application of ML in predictive premium pricing.	Advanced hyperparameter tuning approaches like Bayesian Optimization
Veena et al. (2023)	Insurance Dataset	ANN/light BGM for high risk profiling	introduce Advanced hybrid approaches
Vyas et al. (2022)	Insurance dataset of 79,210 rows and 8 columns	underwriting using Random Forest , XGB	apply multi-class classification of insurance claims

3 Justification to Proposed Methodology

The dataset is obtained from Kaggle and a concise statistical summary using the machine learning library 'FastML' is obtained. Advanced Exploratory Data Analysis is conducted to uncover relationships among claim and patient related attributes, enabling deeper insights into patterns and distribution of data. Numerous visualizations are plotted to know the distribution of data across provider specialty, gender, claim type following a Pearson correlation matrix to compute relativities. To improve accuracy, relevant encoding techniques are employed on categorical and numerical features respectively. Although the initial models Gradient Boosting and Random Forest are trained and tested to enhance greater computation in a column transformer with data containing derived features, the metrics report suggests poor performance in classifying high profile claims with the ROC AUC score of only 0.49, implying counter-intuitive predictions and class-imbalances. But next steps will focus on narrowing class disparities, and the addition of new logarithm technique on the significant feature 'ClaimAmount' by training the XGBoost model.

The flow chart in figure [2] illustrates the intricate details involved in performing each step in the process of solving the research question. Data is obtained from the Kaggle platform and ensured is in compliance with the regulations of General Data Protection Regulation GDPR law. There have been several financial fraud detection studies and surveys developed in the past to address similar issues through skilled inspection and auditing. However, due to substantial technological developments, traditional approaches to detect fraud have not been shown to be effective and reliable. A comprehensive survey of existing research in the field has led to the selection of Bagging and Boosting techniques to improve predictive performance, anomaly detection and identify high risk patients. Hence, following the literature survey, three ensemble supervised models are adopted- Random Forest, Gradient Boosting and XGBoost that proved to be powerful in tackling complex challenges in health insurance.

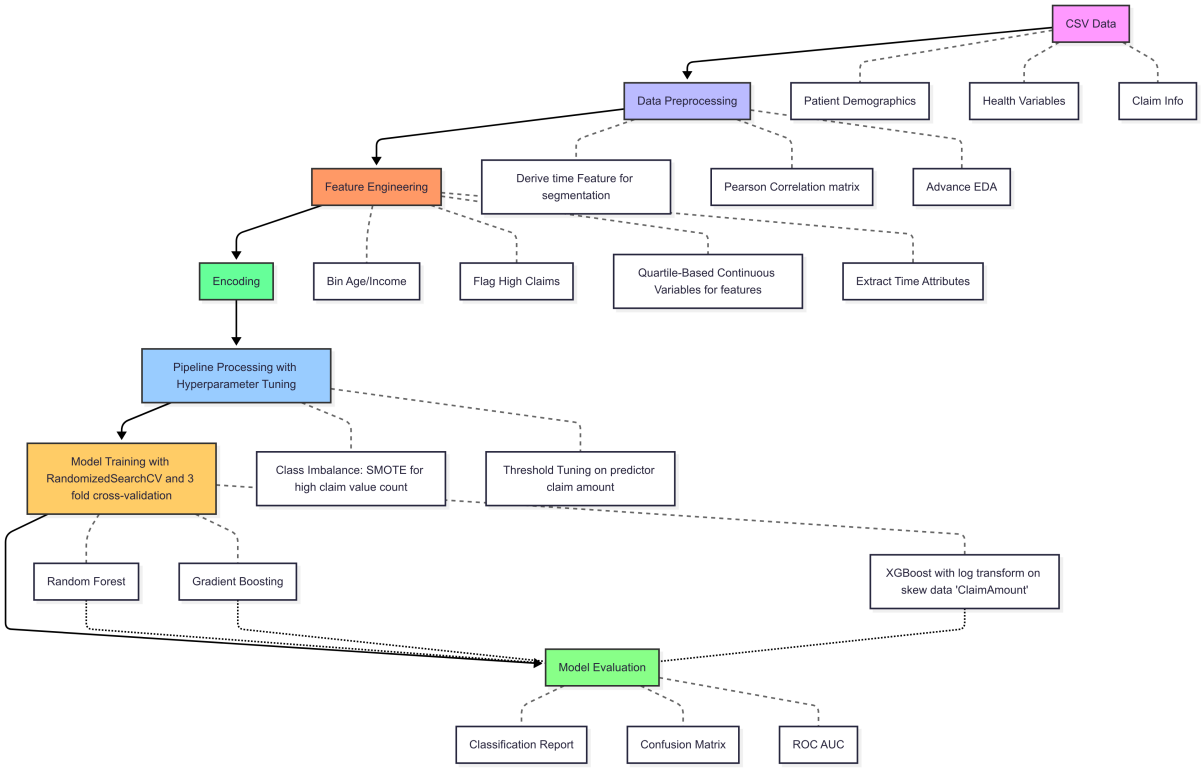


Figure 2: Methodology Flowchart

4 Design Specification

4.1 DataSet

- The US Health Insurance dataset obtained from Kaggle comprises approximately 4500 entries with key attributes such as claim amount, date, patient age, provider specialty, claim status, patient income, marital status, claim type and submission method. A single dataset is employed for this project due to its extensive comprehensiveness and significance containing a multidimensional view of patient population. The variables provided in the dataset are critical to analyze complex patterns in healthcare access, financial burden and claim behaviour. This also ensures consistency.
  Statistical summary present notable findings about highest claim amount recorded, age groups that filed for the reimbursement. The dataset contains ample information on policyholders' income, enabling a more comprehensive analysis.

- Considering and drilling down in these areas play a very crucial role in exposing the underlying trend or correlation if any, that could impact insurance premiums, policy choices, claim behaviours and forecasting risks that will later be taken into consideration. However, efforts are made to best increase the model's capability to understand complex dependencies and provide precise estimations along with trend analysis.

4.2 Data Quality Assessment

- A concise tabular data summary consisting of data type, data type group– categorical/numerical, unique values, sample unique values and percentage of missing values is created using an efficient machine learning library 'FastML' as shown in Table [3].

fields	data_type	num_unique_values	sample_unique_values	num_missing
ClaimID	object	4500	{19044a4f-f7d5-4e16-8216-724f6490ef41,...}	0
PatientID	object	4500	{8532381b-7960-4f16-b190-2b0a8ada0a81, ...}	0
ProviderID	object	4500	{a4cb9c-4863-41cf-8480-2ea17baace88,...}	0
ClaimAmount	float64	4490	3369, 2299, 3940	0
ClaimDate	object	731	{07-06-2024, 30-05-2023, 27-09-2022, 25-05-202...}	0
DiagnosisCode	object	4495	{by0d5, b0052, a2d82, ka21, l2d51, qM187, b2...}	0
ProcedureCode	object	4495	{h5d62, mH831, d9671, k6328, c0658, n05b1, N...}	0
PatientAge	int64	100	23,45,64,89	0
PatientGender	object	2	{M, F}	0

Table 3: Summary of insurance claims data attributes

- Steps to evaluate duplicates, null, missing values and data type format are carried out. Additionally, time feature conversion to the 'datetime' format is carried to ensure uniformity across all date values. Useful features like month and year are derived from 'claimdate' column.

- To summarize key statistics in the data, the describe method of pandas is used on the dataset that returns descriptive information.

4.3 Insights from Pivot tables

Strategic observations obtained from Microsoft Excel's Pivot table are shown below in Figures:

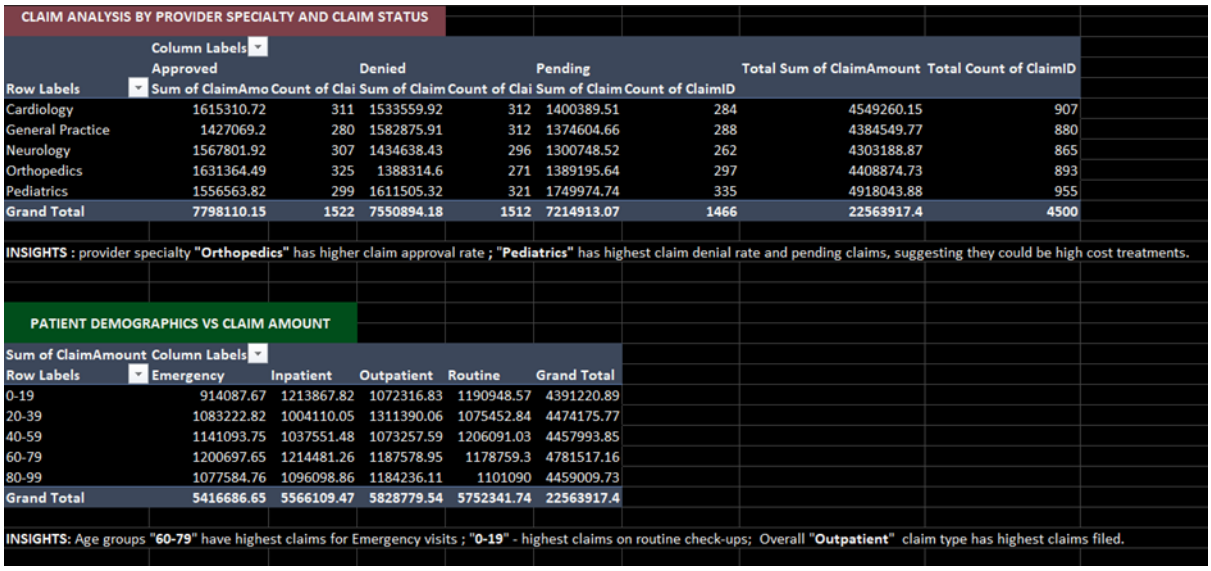


Figure 3: Correlation analysis using Pivot Table with interactive filtering

CLAIM SUBMISSION METHOD EFFICIENCY

	Column Labels							
	Approved		Denied		Pending		Total Average of ClaimDate	Total Count of ClaimID
Row Labels	Average of ClaimD	Count of Clai	Average of Cl	Count of Clai	Average of Cl	Count of ClaimID		
Online	10-06-2023	480	11-07-2023	525	29-06-2023	456	27-06-2023	1461
Paper	06-07-2023	524	03-07-2023	511	11-07-2023	509	06-07-2023	1544
Phone	11-07-2023	518	15-07-2023	476	05-07-2023	501	10-07-2023	1495
Grand Total	**30-06-2023**	**1522**	**09-07-2023**	**1512**	**05-07-2023**	**1466**	**05-07-2023**	**4500**

INSIGHTS: claims filed through **paperwork** have higher approval rates; claims registered via **phone calls** reveal lowest denial rate while they have a likelihood of remaining pending

PATIENT EMPLOYMENT STATUS AND CLAIM ACTIVITY

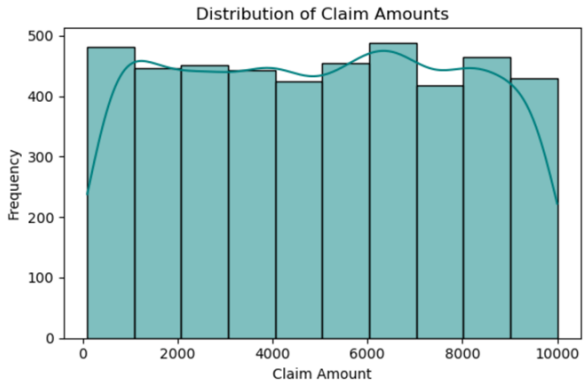
Sum of ClaimAmount	Column Labels				
Row Labels	Divorced	Married	Single	Widowed	Grand Total
Employed	1366617.59	1640361.55	1467012.53	1476604.07	5950595.74
Retired	1267854.25	1283769.96	1318070.6	1438695.11	5308389.92
Student	1415471.61	1478953.27	1276847.96	1445648.54	5616921.38
Unemployed	1558229.57	1467694.77	1340992.23	1321093.79	5688010.36
Grand Total	**5608173.02**	**5870779.55**	**5402923.32**	**5682041.51**	**22563917.4**

INSIGHTS: Unemployed policy-holders have filed for highest claims while **Retired** with less claims

Figure 4: Correlation analysis using Pivot Table for efficient and data-driven decisions

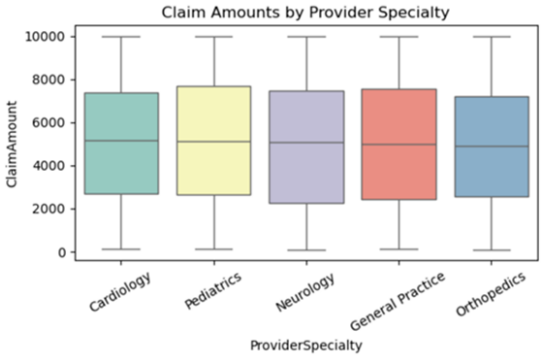
4.4 Data Visualization

- To obtain a clear knowledge of the study and make patterns easier to understand, data visualization tools are widely used. Initially a histogram [5] is plotted to get an idea of the frequency of distribution of claim amounts.



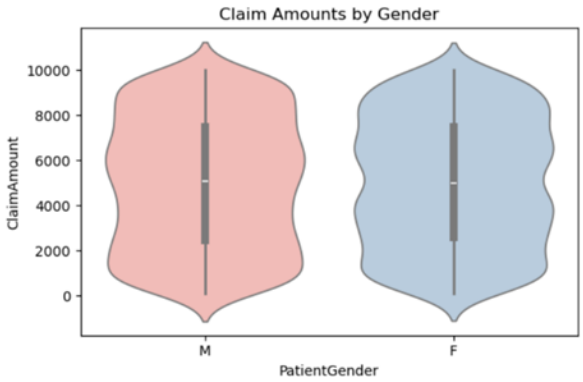
Visualization 5: Histogram distribution of claim amounts

- Outliers reveal data points that stand out because they are unusually high, low or odd compared to the rest- hence a boxplot is plotted to check for anomalies in the data. Below is the corresponding image [6].



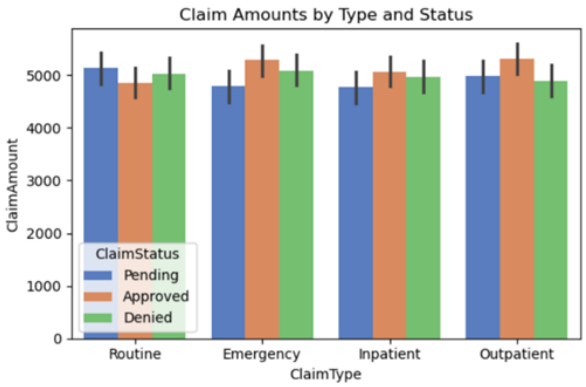
Visualization 6: Boxplot for outlier

- A violin plot [7] is then plotted which is a hybrid of box plot and kernel density plot to get a bigger picture of how the data is distributed as shown below.



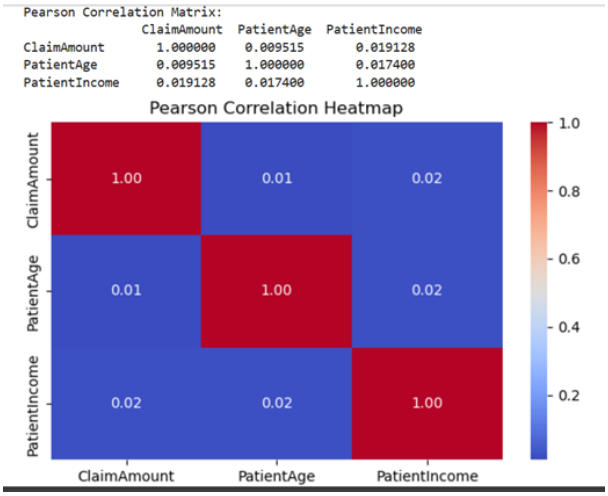
Visualization 7: Violin plot for data density

- A bar plot [8] with hue set to 'ClaimStatus' is plotted to compare subgroups within categories between claim type, claim amount and claim status.



Visualization 8: Barplot for claim type, claim amount and gender

- Finding the correlation between different variables to quantify patterns and predict behavior is done using Pearson's matrix [9]. The results reveal that there is no significant dependency between one another in the data.



Visualization 9: Pearson Correlation

4.5 Feature Engineering

- Feature Engineering is performed to transform raw data into meaningful features that enhance model performance. The features ‘ PatientAge‘ and ‘ PatientIncome‘ are engineered to create a more comprehensive model by binning the categorical and numerical values into quartiles. Income values are converted into categorical bins.

- Subsequently, the claims higher than a threshold of 0.75 are automatically flagged high into the column ‘ IsHighClaim‘. To facilitate good computing in a model, another feature ‘ ClaimInteraction‘ is created by concatenating 2 columns ‘ ClaimType‘ and ‘ ClaimSubmissionMethod‘ for accurate results. These interaction features boost model accuracy and interpretability.

- Encoding of data has proven to make a model with smart features outperform complex algorithms with poor features. Hence, one hot encoding technique is implemented for categorical values.

- Parameters used in hyperparameter tuning are manually chosen to increase the model's behaviour of succeeding.

5 Implementation/ Solution Development

- Random forest is an ensemble method that builds multiple decision trees and combines their output to improve accuracy and reduce overfitting. It uses bagging, a bootstrap aggregation that helps train each tree on a randomized subset of data. This model selects random sample features to reduce correlation between trees. It differs in producing results for classification and regression in a way that majority vote and average predictions are taken into consideration respectively.

- A sequential build of tree, that learns from the error of the previous can be obtained using the Gradient boosting model. It starts by training the weak model and goes on to train and predict the errors of previous trees [10].

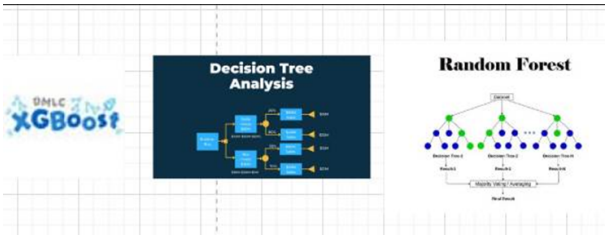


Figure 10: Models Employed

- Data is split into training and testing with a set random state for reproducibility to validate the model. Careful attention is paid to ensure that fraudulent cases don't get disproportionately placed in one set, which would skew model performance. 80% of the data is trained and the rest of 20% is tested with a random state of 0.2 for reproducibility, which is then passed into a column transformer for parallel processing and efficiency. To adjust the decision boundary in a classification model, a popular and efficient threshold tuning which is a popular hyper parameter tuning technique called 'Randomized SearchCV' with 3-fold cross validation is employed.

- The threshold tuning is custom set to 0.3, as it can higher the recall value. Max depth and n estimators are set to 10 and 100 respectively with the idea of making the model perform best. First, 2 models – Gradient Boosting and Random Forest were trained on the extracted data features, which produced fairly moderate results with class 1 being incorrectly predicted in some cases and ROC of 0.47 suggests room for more improvement as shown in the classification report in Figure [11].

```
== Gradient Boosting Evaluation with threshold = 0.3 ==
Confusion Matrix:
 [[387 288]
 [141  84]]
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.57      0.64       675
           1       0.23      0.37      0.28       225

    accuracy                           0.52       900
   macro avg       0.48      0.47      0.46       900
weighted avg       0.61      0.52      0.55       900

ROC AUC Score: 0.4758156378600823
```

Figure 11: Gradient Boosting Report

- Figure 12 shows classification report of Random Forest model performance, when the decision threshold is set to 0.3 – instead of the default 0.5 in an attempt to increase recall metric.

```
== With Custom Threshold (0.3) ==
Confusion Matrix:
 [[387 288]
 [141  84]]
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.57      0.64       675
           1       0.23      0.37      0.28       225

    accuracy                           0.52       900
   macro avg       0.48      0.47      0.46       900
weighted avg       0.61      0.52      0.55       900
```

Figure 12: Random Forest Report

- Class 0 is predicted accurately, and class 1 shows low precision and recall with overall accuracy 52% suggesting, room for improvement.
- Handling the skewing data by applying logarithmic transformations and applying it to the target dependent variable, has shown an increasingly higher accuracy metric, while the rest of the features were retained. A class imbalance was identified which was rectified using the Synthetic Minority Over Sampling (SMOTE) technique. A preprocessing pipeline is established using ColumnTransformer in figure [13] to manage both encoded categorical and numerical features. The resampled data from SMOTE is incorporated into the pipeline, followed by the definition of the XGBClassifier model.

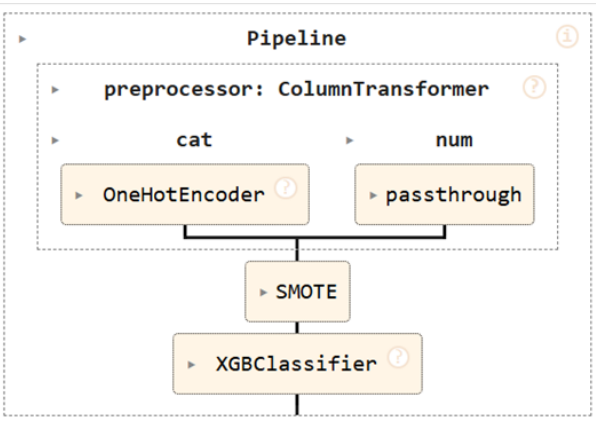


Figure 13: XGBoost Classifier with Smote and log transformed data

- An extreme gradient boosting is a high performance, scalable machine learning algorithm, commonly used for its boosting framework, regularization, parallelization, capacity to handle missing data and built-in importance qualities. This time, the data is trained using an XGBoost model, which tackles complicated classification problems.

6 Evaluation

The confusion matrix shows that the model has correctly identified all the class 0, which is high claim instances with just 1 misclassification in class 1, achieving the best per-

formance overall. This was achieved by combination of SMOTE technique and the hyperparameter tuning 'RandomizedSearchCV' that reduces the samples from unlimited possibilities which is unique to this research and different from the previous ones conducted.

6.1 Case Study 1 - Threshold tuning 'RandomizedSearchCV'

- Of the 3 models built, XGBoost algorithm presents with the best performance, enhancing reliability by profiling high risk patients that helps to assess new applicants in detecting potential fraudulent intentions and improving customer service by anticipating needs. In doing so, the insurers lead to creating fairer pricing policy, ultimately building trust with the policyholders.

- The confusion matrix shows that the model has correctly identified all the class 0, which is high claim instances with just 1 misclassification in class 1, achieving the best performance overall. This was achieved by combination of SMOTE technique and the hyperparameter tuning 'RandomizedSearchCV' that reduces the samples from unlimited possibilities which is unique to this research and different from the previous ones conducted.
  Evaluation metrics equation for classification report is shown below:
  Accuracy = (TP + TN)/ (TP + TN + FP + FN)
  Precision = TP/ (TP + FP)
  Recall = TP/ (TP + FN)
  F1 Score = 2 * [(Precision * Recall)/ (Precision + Recall)]
  where TP - True Positive, TN - True Negative, FP - False Positive, FN - False Negative.

6.2 Case Study 2 - Log transformation of skewed data

- Model reveals that 675 customers are correctly predicted not to file a claim, and 224 customers correctly predicted to file a claim as shown in figure [14] as a result of logarithmic transformation over skewed data.

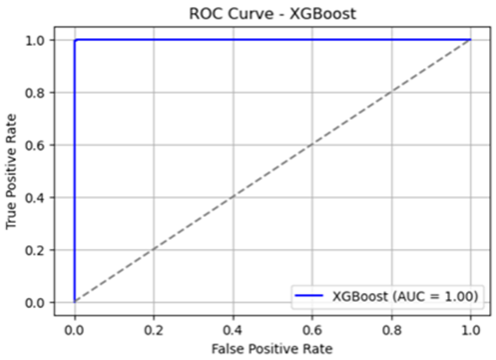
- Such accuracy translates into powerful foresight for anticipating claims and potential liabilities. This capability empowers organizations to allocate resources intelligently- focusing on outreach, or funding to individuals identified as high risk. The model can thus distinguish between claim and no-claim cases as shown below:

```
=== Confusion Matrix ===
[[675   0]
 [  1 224]]

=== Classification Report ===
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       675
           1       1.00      1.00      1.00       225

    accuracy                           1.00       900
   macro avg       1.00      1.00      1.00       900
weighted avg       1.00      1.00      1.00       900

=== ROC AUC Score ===
0.9999736625514404
```

Figure 14: Classification Report of XGBoost

6.3 Case Study 3 - Receiver Operating Characteristic Area Under the Curve

- Scores of precision, recall, f1-score have demonstrated a huge improvement, in contrast with random forest and gradient boosting.
- Overfitting of the model was checked by plotting the accuracy scores between training and testing data, wherein the training score was higher revealing that the model is not overfitting.
- Below image [15] shows the ROC AUC plot high lighting model's effectiveness in distinguishing between positive and negative classes across all thresh olds. The blue line traces in the top left, signs of accurate and reliable classification maximizing True Positives while avoiding False Negatives.



Visualization 15: ROC AUC Curve

- Risk Stratification with Unprecedented Accuracy enables fine grained profiling of policyholders using health demographic attributes, reducing instances of misclassification and adverse selection, leading to more stable risk pools.

ROC Curve plots
TPR = TP/ (TP + FN)

FPR = FP/ (FP + TN)
Where, TPR - True Positive Rate and FPR - False Positive Rate.

6.4 Discussion

In summary, the model XGBoost has worked exceptionally well on data containing demographic insights, claim information, provider location data and gender-based variables. The model also achieved a remarkable predictive performance across this multifaceted dataset, effectively capturing underlying patterns within the data.
Comparison of key factors that influence health insurers and policyholders' decision making with the proposed model are:

1. High Precision risk profiling that enables accurate stratification of patients allowing firms to tailor premiums and coverage was achieved in this current research wherein the previous papers showed much less accuracy scores.

2. Customer trust and retention- transparent pricing builds credibility and fosters long-term policyholders' relationships, which was achieved by introducing threshold tuning - an advanced hyperparameter method, not implemented in the previous papers.

3. Regulatory alignment- this data driven pricing strategy can meet compliance standards while reducing human bias. The dataset used to conduct this research contains of vast demographic data and important attributes of claims, allowing for less biased outcomes.

7 Conclusion and Future Work

The framework applying XGBoost Classifier model performs exceptionally well on high-risk profiling with accuracy rate above 95%. The research findings have addressed the **question**: "Can the insurance firms better anticipate risks and allocate their assets resourcefully based on significant patient attributes containing key demographic and health-related variables using supervised ensemble Machine Learning techniques and accurate model evaluation thereby providing fairer and accurate rates for policyholders?".

The following business objectives are achieved by integrating the proposed model into an existing workflow:
a. Optimized Asset Allocation: Accurate predictions allow insurers to distribute the capital efficiently, minimizing over-reserving.
b. Fair and Transparent Pricing Models: customers receive personalized premium rates that reflect actual risk, boosting trust and satisfaction.
c. Market Leadership and Innovation Branding: Firms adopting advanced ML techniques signal tech savviness and innovation.

While the research demonstrates a strong proof of concept, few areas remain open for exploration such as Model **Generalization**- where future studies should test model robustness across diverse datasets and population, not just pertaining to the US. **Temporal dynamics**- incorporating time-series health data could unlock insights into evolving risk profiles. Domain Knowledge Features can be considered as part of future work to capture fraud-prone signals. **Explainability and Trust- Integrating tools** like Shapley Addictive Explanations (SHAP) values or Local Interpretable Model-agnostic Explanations (LIME) can highly interpret model decisions, ensuring transparency for stakeholders and exercise regulatory laws. **Real-time deployment**- evaluating performance in live settings, that will help bridge the gap between theoretical potential and operational impact.

References

Abdelshafy, N., Abdelshafy, M., Ghazanfar, M. A. and Saglam, R. B. (2024). Machine learning for enhanced underwriting: Predicting premiums in health insurance, *2024 4th International Conference on Robotics, Automation and Artificial Intelligence (RAAI)*, IEEE, pp. 330–339.

Alam, A. and Prybutok, V. R. (2024). Use of responsible artificial intelligence to predict health insurance claims in the usa using machine learning algorithms, *Exploration of Digital Health Technologies* **2**(1): 30–45.

Barbiero, A. et al. (2024). Yet another approximation for the total claims amount using the weibull distribution.

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

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.

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

NDTP, N. D. T. . P. (2024). The irish healthcare system, *Forum of Irish Postgraduate* pp. 1–10.

Onwubuariri, E. R., Adelakun, B. O., Olaiya, O. P. and Ziorklui, J. E. K. (2024). Ai-driven risk assessment: Revolutionizing audit planning and execution, *Finance & Accounting Research Journal* **6**(6): 1069–1090.

Paikaray, B. K., Samantara, T., Rout, D. and Mishra, S. (2023). Machine learning-based regression model to predict health insurance claim, *2023 IEEE 2nd International Conference on Industrial Electronics: Developments & Applications (ICIDeA)*, IEEE, pp. 163–168.

Paul, J. (2024). Financial time series analysis with transformer models, *researchgate*, IEEE, pp. 163–168.

Pragatheeswari, E., Kavitha, M., Saranya, S., Dhanushree, D., Jasmitha, S. and Abiradhie, P. (2025). Prediction of insurance claims for health sector using machine learning techniques, *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, IEEE, pp. 1312–1318.

Prova, N. N. I. (2024). Advanced machine learning techniques for predictive analysis of health insurance, *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, IEEE, pp. 1166–1170.

Ramani, K., Kumar, S. T., Datta, P. P. S., Jamuna, P. and Nithin, K. S. (2024). Predicting health insurance claim amount through machine learning algorithms, *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, IEEE, pp. 1–6.

Saraswat, B. K., Singhal, A., Agarwal, S. and Singh, A. (2023). Insurance claim analysis using traditional machine learning algorithms, *2023 International Conference on Disruptive Technologies (ICDT)*, IEEE, pp. 623–628.

Sharma, A. and Jeya, R. (2024). Prediction of insurance cost through ml structured algorithm, *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Vol. 5, IEEE, pp. 495–500.

Suneel, S., Thatipudi, J. G., Kumari, M. J., Penyameen, K., Sharma, A. et al. (2025). Xannlight: A novel hybrid ensemble model for accurate health insurance prediction and trend forecasting, *2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, IEEE, pp. 1819–1826.

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.

Vangala, B. L. (2025). Data engineering frameworks with ai and ml algorithms: Fraud detection and prevention strategies in insurance sector.

Veena, K., Supriya, S., GM, K. D. et al. (2023). Predicting health insurance claim frauds using supervised machine learning technique, *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, IEEE, pp. 1–7.

Vyas, S., Serasiya, S. and Vyas, A. (2022). Combined approach of ml and blockchain for fraudulent detection in insurance claim, *2022 International Conference on Edge Computing and Applications (ICECAA)*, IEEE, pp. 544–550.

Yang, Q. and Wu, X. (2006). 10 challenging problems in data mining research, *International Journal of Information Technology & Decision Making* **5**(04): 597–604.