International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Understanding Moral Hazard in Auto Insurance Claims with AI-Powered Predictive Models

Rajesh Goyal

GlenMills, PA USA
rajesh.nim@gmail.com

Abstract

Fraudulent auto insurance is a serious problem, as monetary losses happen to insurance companies and the costs increase for policyholders. The goal of this project is to develop a prediction algorithm that is able to correctly guess potential auto insurance fraud claims. Moral hazard issues always challenge the design and regulation of insurance policies. Yet most of these worries are based on theoretical expectations of how rational economic agents will respond to financial incentives. In this study, ML and DL approaches are used in data-driven ways to identify false statements and evaluate behavioral patterns of moral hazard. It involves the preprocessing of the unbalanced Auto Insurance Claims Fraud Detection dataset, training the classification models like XGBoost, Random Forest (RF), Recurrent Neural Networks (RNN), and Multi-Layer Perceptron (MLP) and performing feature engineering with the help of min max normalization as well as one hot encoding. Accuracy, precision, recall, F1 score and AUC ROC are used to evaluate the models. Experimental results show that XGBoost performs better than deep learning models in fraud classification, with the highest accuracy (82.5%) and a balanced trade-off between recall (80.39%) and precision (80.80%). The findings highlight the effectiveness of ML-based ensemble techniques in mitigating moral hazard and enhancing fraud detection strategies.

Keywords: Moral Hazard, Insurance Fraud Detection, Machine Learning, Deep Learning, XGBoost, Automobile, Policy

I. INTRODUCTION

The issue of claims fraud, including illegitimate claims and exaggerated loss amounts (buildup), remains a significant concern for automobile and homeowner's insurance companies. Empirical studies suggest that a large percentage of claims involve fraud or exaggeration[1]. Fraud detection has emerged as a high-priority, technology-driven challenge for insurers, particularly as insurance costs continue to rise globally. This issue is exacerbated in developing regions due to evolving financial regulations and new legislative frameworks[2]. Historically, before the 1980s, underwriting and claims settlement fraud were generally discussed under the broader concept of moral hazard[3]. Moral hazard in insurance arises when policyholders or claimants possess undisclosed information that materially impacts risk exposure or the actual loss incurred. The fundamental premise is that insurance coverage reduces the policyholder's incentive to minimize losses, potentially leading to fraudulent behavior[4].

Hard and soft fraud are the two main categories into which insurance fraud falls. In hard insurance fraud, accidents or damages are purposefully fabricated, while soft insurance fraud occurs when claimants

exaggerate legitimate claims to receive higher compensation[5]. Effective fraud detection and prevention mechanisms enhance customer trust and reduce the financial burden on insurers[6]. In automobile insurance, moral hazard becomes particularly problematic when policyholders benefit financially from an insured loss. For example, if an insurance company replaces a damaged vehicle with a brand-new or almost brand-new one, policyholders may have an incentive to seek loss rather than prevent it[7]. The challenge of fraud detection in auto and homeowner's insurance is compounded by the lack of precise classification rules[8]. ML algorithms are increasingly being used to detect bogus claims; however, class imbalance in fraud datasets often affects model performance, leading to difficulties in accurately predicting fraudulent instances. Addressing moral hazard requires a combination of robust fraud detection methodologies and strategic policy measures to align policyholders' incentives with insurers' risk management objectives.

A. *Significance and Contribution*

The primary aim of this study is to use ML models for fraud detection to examine if moral hazard exists in vehicle insurance claims. The study seeks to develop an effective classification framework that can accurately distinguish between fraudulent and legitimate claims, thereby aiding insurers in minimizing financial losses and improving risk assessment strategies. The main contributions are:

- Using the Auto Insurance Claims Fraud Detection dataset for Auto and Homeowners Insurance Claims.
- The study employs robust data preprocessing techniques, including handling missing values, one-hot encoding, and min-max normalization, to enhance the quality of the dataset.
- The study provides insights into the behavioral patterns associated with fraudulent claims, highlighting key risk factors that contribute to moral hazard in auto insurance.
- Developed machine learning and deep learning XGBoost, Random Forest, Recurrent Neural Networks and Multi-Layer Perceptron.
- Thorough assessment of classification models using performance metrics such as recall, accuracy, precision, and F1-score.

B. *Structure of the paper*

The study is structured as follows: Relevant research on moral hazard in auto and homeowner's insurance claims is presented in Section II, while Section III describes the methods and resources employed. The experimental results of the suggested system are shown in Section IV. Section V wraps up the inquiry and presents a summary of its results.

II. LITERATURE REVIEW

This section discusses some review articles on Auto Insurance Claims using ML techniques. Table I highlights the paper, methods, dataset, key findings, and limitations/future work.

Kowshalya and Nandhini (2018) addition to the insurance business, genuine policyholders are also impacted by the prevalence of false claims. Typically, insurance firms use traditional methods to detect fraudulent claims with the assistance of domain expertise. In recent years, data mining has significantly advanced the field of insurance analysis. In this study, data mining techniques are utilized to estimate insurance premium amounts for various clients based on their financial and personal information, as well as to forecast fraudulent claims[9].

Rahmaniah and Syarif (2018) the dataset was a minority report open collection of German auto insurance data that served as a benchmark. The performance assessment results are then compared to those of previous research projects that employed the same dataset. According to the experiment's findings, the performance measurement acquired using this study's methodology is sometimes better[10].

Subudhi and Panigrahi (2018) the original claim data set's data imbalances are removed using ADASYN in minority class scenarios. The aberrant data is also separated from the regular data using three different classifiers: SVM, DT, and MLP. It uses the 10-fold cross-validation technique to confirm the models' efficacy. An auto insurance data set is used for many tests that show the model's performance[11].

Kareem, Ahmad and Sarlan (2017) present the first findings of their investigation, which seeks to provide a method for recognizing fraudulent health insurance claims by determining a relationship or connection between certain characteristics on the claim papers. This study argues that the successful identification of linked features may effectively resolve the data inconsistencies in fraudulent claims and, therefore, minimize health insurance fraud through the use of a data mining approach using association rules [12].

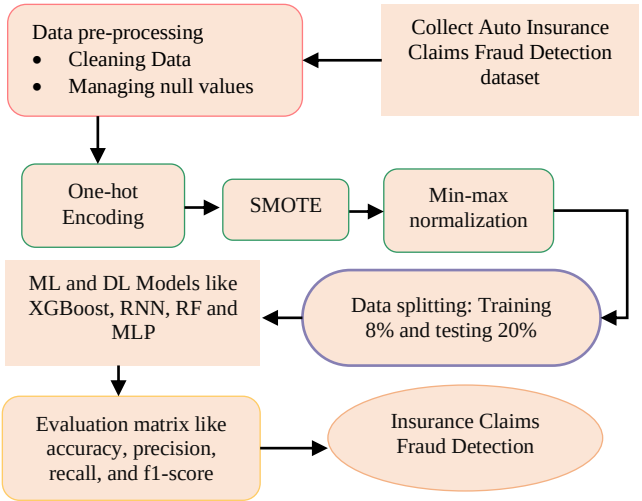
Li et al. (2016) the data was used to filter the index, and the relative importance of each input variable to the output variable was ascertained. The error of the model was analyzed. The method is now supported by empirical evidence. According to real data, the RF-based auto insurance fraud mining technique works better with large, unbalanced data sets than the traditional model. Classifying and forecasting data from vehicle insurance claims and mining fraud rules are better uses for it. Additionally, its precision and robustness are superior[13].

TABLE I. SUMMARY OF RELATED WORK STUDY FOR AUTO INSURANCE CLAIMS USING MACHINE LEARNING

Authors	Methods	Dataset	Key Findings	Limitations / Future Work
Kowshalya and Nandhini (2018)	Predicted false claims and determined premium amounts using data mining methods and subject expertise based on clients' financial and personal information.	Insurance claim data (exact dataset not specified)	Demonstrated that data mining can be effectively applied for fraud detection and premium calculation.	Specific limitations were not discussed; future work could explore the refinement of techniques and dataset expansion.
Badriyah, Rahmaniah and Syarif (2018)	Employed a performance measuring strategy and used the same methodology to compare the results with those of another research.	German car insurance data (benchmark open dataset)	In several instances, the suggested method's performance measurement proved to be better than in earlier studies.	Limitations and further improvements were not explicitly detailed.

Subudhi and Panigrahi (2018)	Addressed class imbalance using ADASYN on minority instances and implemented three classifiers—Support Vector Machine, DT, and MLP—with 10-fold cross-validation.	Auto insurance dataset	The experimental findings demonstrated how well the models distinguished between normal and aberrant (fraudulent) records	Detailed limitations or directions for future enhancements were not provided.
Kareem, Ahmad and Sarlan (2017)	Applied association rule mining to detect correlations among attributes in health insurance claim documents.	Health Insurance Claims Dataset	Identified that determining correlated attributes through association rules can help reduce discrepancies and fraudulent claims.	Limitations and suggestions for future work were not explicitly mentioned.
Li et al. (2016)	Conducted an importance analysis of input variables and error analysis, introducing Random Forest in the fraud mining model; compared against traditional models.	Data from auto insurance claims	The Random Forest-based model outperformed conventional techniques in terms of accuracy and resilience, and it worked well with big and imbalanced datasets.	Specific limitations or recommendations for future work were not outlined.

Fig. 1. Flowchart for Insurance Claims Fraud Detection



III. METHODOLOGY

A data-driven assessment method reveals moral hazard activities in auto insurance claim settlements throughout the investigation. The approach adopts a defined process that begins with obtaining an auto insurance claims fraud detection dataset, which requires preprocessing and value handling along with inconsistency resolution. Several standardization methods, including min-max normalization, data balancing by SMOTE, and one-hot encoding, transform the dataset for standardized analysis. Data partition takes place for testing functions at 20%, and trains function at 80% of the total sample for the development of predictive models. The XGBoost model alongside Random Forest (RF) and Multi-Layer Perceptron (MLP) along with recurrent neural networks (RNN) is used for machine learning to identify false insurance claims with an added capability to identify possible behavioral indications of moral hazard. The system uses F1-score and recall as well as accuracy and precision to evaluate its performance. The findings provide insights into the risk factors associated with moral hazard, aiding insurers in developing strategic measures for fraud prevention and risk mitigation. Figure 1 shows the flowchart of insurance claims fraud detection.

The following steps of the proposed methodology are described in short below:

A. Data Collection with Visualization

The dataset for Auto Insurance Claims Fraud Detection was obtained from Kaggle. This dataset mostly includes data on the different types of motor insurance claims that occurred between 1994 and 1996. The dataset includes one class variable and thirty-one predictor variables. There are 15,420 samples in total; 14,497 of them are not fraudulent, and 923 are. This indicates that 94% of the samples are real, and 6% are fake. Because of this, there is a significant imbalance in the dataset between the percentage of fraudulent and non-fraudulent samples. The data visualization such as pie chart, heatmap, and bar graph, is shown below:

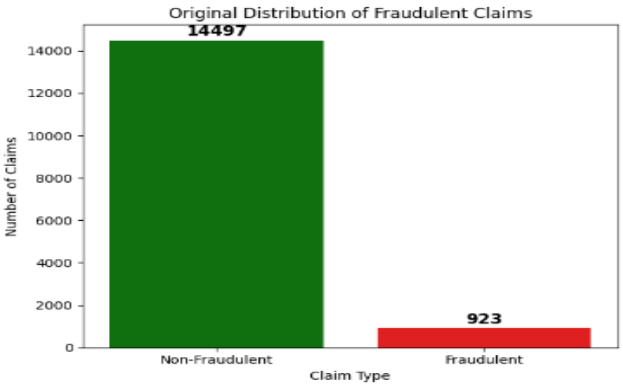


Fig. 2. Distribution of fraudulent claims

Figure 2 illustrates the ratio of false and legitimate claims in the Auto Insurance Claims Fraud Detection dataset, both with and without balancing (SMOTE). The dataset includes 15,420 samples, with a notable disparity between the two classes: only 923 (6%) are fake, while 14,497 (94%) are not. This imbalance highlights the rarity of fraudulent claims, making fraud detection a challenging task for insurers. The distribution emphasizes the need for advanced machine learning techniques to accurately identify fraudulent claims without misclassifying genuine ones.

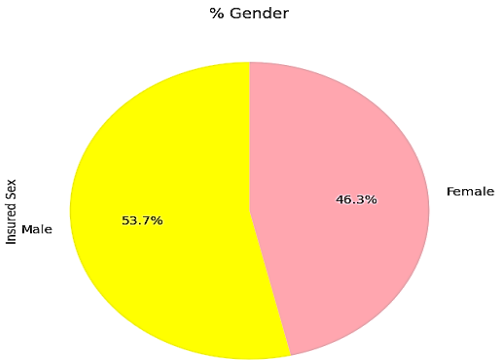


Fig. 3. Distribution of Insured Sex

Figure 3 shows the pie chart illustrating the distribution of insured individuals by gender, showing that 53.7% are male and 46.3% are female, indicating a slightly higher representation of males within the insured population.

Figure 4 illustrates the correlation matrix representing the pairwise relationships between different features in the Auto Insurance Claims Fraud Detection dataset. The heatmap employs a color gradient, where darker shades indicate stronger correlations. Notably, features such as Policy Number and Year exhibit a high correlation (0.94), suggesting redundancy or potential multicollinearity. The fraud indicator variable (FraudFound_P) shows minimal correlation with most features, indicating that fraud detection may require complex feature interactions beyond linear correlations.

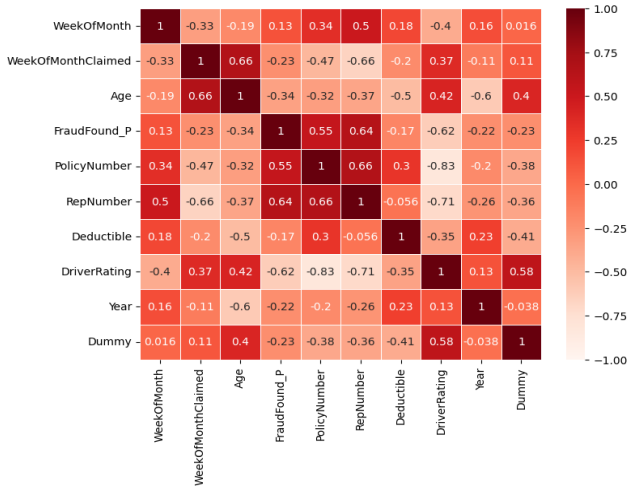


Fig. 4. Correlation Matrix

B. Data Preprocessing

Data Pre-processing in machine learning may be a crucial step that can help enhance the quality of the information and facilitate the extraction of important knowledge from it. After the raw data has been gathered, it is time to organize it so that further processing can use it:

- **Data cleaning:** It is crucial that the data collection be devoid of flaws that can hinder testing or, worse, result in inadequate analysis. These shortcomings or issues brought on by duplicate data, missing values, or dimension loss need to be successfully fixed. Bad data will thus be eliminated in this stage, and missing data will be added.

- **Missing values:** In the initial stage of data cleansing, missing values are handled. Missing values in a record are the absence of information, whether deliberate or not. Finding and encoding missing data is the first stage; dealing with the missing values is the second.

C. *One-Hot Encoding*

Categorical variables are converted into a numerical format that can be comprehended using the appropriate preprocessing techniques, such as one-hot encoding. It is among the most often used techniques, comparing every numerical variable level with a predetermined beginning point.

D. *Synthetic Minority Over-Sampling Technique (SMOTE)*

In order to overcome class imbalance in machine learning, a popular resampling technique is the SMOTE. By interpolating between existing instances instead of just replicating them, it creates synthetic samples for the minority class. Using line segments connecting the original sample and its neighbors, SMOTE generates new synthetic data points after choosing a minority class instance and determining its k-nearest neighbors. This method reduces bias toward the majority class, provides a more balanced dataset, and improves the classifier's capacity to identify uncommon occurrences, such as false claims in insurance datasets, all of which contribute to better model performance.

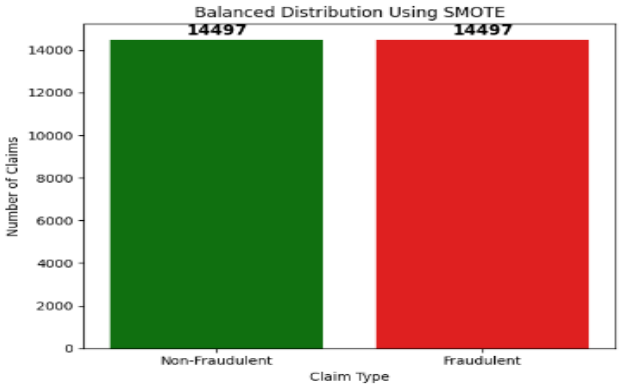


Fig. 5. Balanced Bar Chart of Data Distribution Classes

Figure 5 bar chart demonstrates the balanced dataset after applying SMOTE, which generates synthetic fraudulent claims to match the non-fraudulent claims at 14,497 each. This balancing technique is crucial for training machine learning models because it removes bias towards the majority class and increases the model's ability to identify false claims.

E. *Normalization*

Scaling numerical values in datasets is accomplished by normalization. To resize each feature, Min-Max is chosen to the [0,1] interval. Performing normalization depends on the algorithm that is used. It is calculated in Equation (1) as follows.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \qquad (1)$$

in where $x_{\min}$ is the lowest value and $x_{\max}$ is the highest value of the provided characteristic.

F. *Data Splitting*

In order to compare their proposed fraud detection method to the results of several models, it utilized Python's 80:20 test-train-split module and its machine learning features to calculate classifier scores.

G. Classification with Xgboost Classifier

XGBoost is a gradient-boosting tree library that is widely used to solve supervised learning problems for tabular data. A gradient boosting machine (GBM is commonly used with decision trees as a basic model and is known as a gradient boosting tree[14]. XGBoost is an ML method for classification and regression issues that creates ensemble weak prediction models, which are commonly referred to as decision trees.

Also known as CART, given data set with n examples and m features D = ((xi, yi)) ([D] = ni,xi, e Rm, yi, E R) a tree ensemble model uses K additive function to predict the output. With F = {f(x)=Wq(x)}(q: R → T,WE R) is the space of CART. So, with the Equation (2)

$$\hat{y}_i = \emptyset(x_i) = \sum_{k=1}^{k} f_k(x_i), f_k \in F \qquad (2)$$

One of the new things introduced by XGBoost is the ability to set a default direction on each of its CART nodes in the form of a split search algorithm. Given observations in a node, the algorithm first collects all observations whose values for features are not lost in set 1, then calculates the gain derived from the left and right separations for each observation.

H. Evaluation Metrics

The categorization report contains a number of indicators that are critical to the assessment of any model. The F1-measure, recall, accuracy, and precision are among the metrics that are displayed.

Accuracy: It is the proportion of accurate forecasts to all observations. It is shown in Equation (3)

$$Accuracy = \frac{TP+TN}{TP+Fp+TN+FN} \qquad (3)$$

Precision: The fraction of positive observations that were accurately predicted relative to all positive data. It is given in Equation (4):

$$Precision = \frac{TP}{TP+FP} \qquad (4)$$

Recall: The percentage of correctly predicted positive observations in a class as a percentage of all observations; also called sensitivity in Equation (5):

$$Recall = \frac{TP}{TP+FN} \qquad (5)$$

F1-score: The average of the recall and accuracy scores. It is shown in Equation (6)

$$F1 - Score = \frac{2(Precision*Recall)}{Precision+Recall} \qquad (6)$$

ROC: A performance metric for classification problems is the ROC curve. The True Positive Rate (TPR) is plotted on the y-axis, while the False Positive Rate (FPR) is plotted on the x-axis. The area under the ROC curve, or AUC, or degree of separability, provides a sense of how effectively a model can distinguish between classes. A better model for class prediction has a higher AUC.

where FN, FP, TN, and TP represent false negative, false positive, and true positive, respectively. Multiplication tables feed into algorithms that determine vehicle insurance claim models.

IV. Result Analysis And Discussion

This section delivers the experimental findings regarding ML models used in Auto Insurance Claims. A device utilizing Python operates with Windows 8 processor, CPU, GPU and operating system architecture for execution. The analysis uses F1-score, recall, accuracy and precision to evaluate performance. The XGBoost model achieves these results for insurance claim detection according to Table II.

TABLE II. ML MODEL PERFORMANCE ON THE AUTO INSURANCE CLAIMS FRAUD DETECTION DATASET

Performance Measures	XGBoost
Accuracy	82.5
Precision	80.80
Recall	80.39
F1-score	80.59
AUC	0.81

The performance assessment of ML and DL models on Auto Insurance Claims Fraud Detection data shows results for both Auto and Homeowners Insurance claims through Table II and Figure 6. Accuracy, together with precision and recall and F1-score, constitute essential assessment measures to determine how effectively the model performs fraud classification. XGBoost achieved the highest performance among all models in the dataset with 82.5% accuracy, 80.80% precision, and 80.39% recall, besides 80.59% F1-Score. The data reveals ML models achieve sufficient detection of fraudulent claims through their precise and recall-balanced operations.

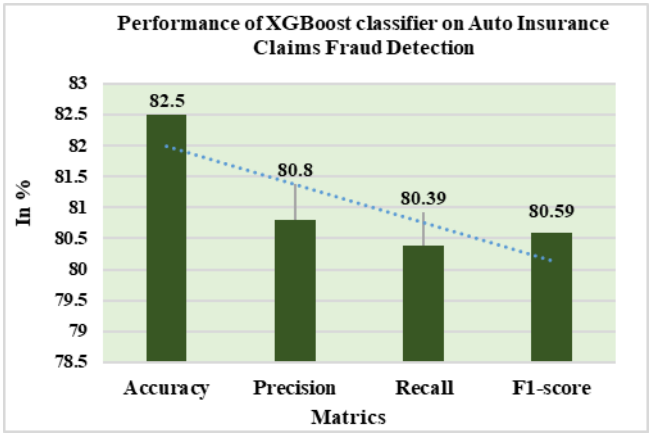


Fig. 6. Bar Graph for XGBoost Model Performance

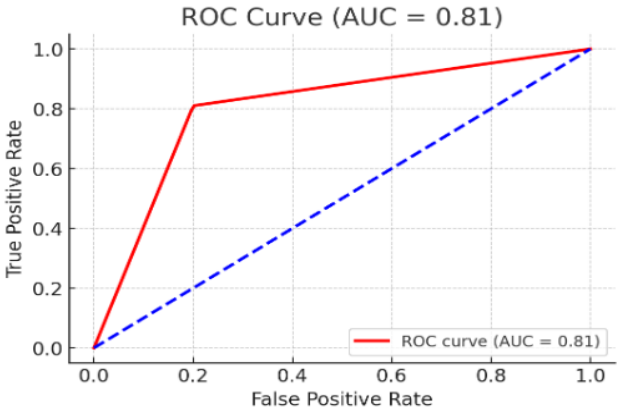


Fig. 7. ROC graph for XGBoost Model

The XGBoost model performance reaches an AUC value of 0.81 according to the ROC curve shown in Figure 7, suggesting high predictive accuracy levels. The red line represents the model, while the blue dashed line shows random performance. The higher curve suggests a strong balance between sensitivity and specificity.

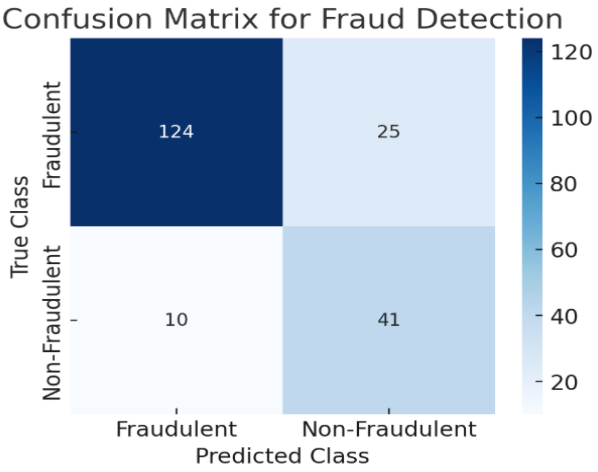


Fig. 8. Confusion Matrix of XGBoost Model

An XGBoost model's confusion matrix is shown in Figure 8, illustrating its performance in classifying fraudulent transactions. The confusion matrix shows that the model correctly classified 124 fraudulent claims (TP) and 41 non-fraudulent claims (TN). However, 25 fraudulent claims were misclassified as non-fraudulent (FN), and 10 legitimate claims were mistakenly reported as fraudulent (FP). Although the model works well, the accuracy of fraud detection might be increased by lowering false positives and erroneous negatives.

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.83	0.88	149
1	0.62	0.80	0.70	51
accuracy			0.82	200
macro avg	0.77	0.82	0.79	200
weighted avg	0.85	0.82	0.83	200

Fig. 9. Classification Report of XGBoost Model

The classification report in Figure 9 presents the performance metrics of an XGBoost model, describing the 200-sample dataset's accuracy, recall, F1-score, and support for two classes (0 and 1). The model demonstrates high precision of 93% and recalls of 83% for class 0, indicating strong performance in correctly identifying this class, while class 1 shows lower precision of 62% and recall of 80%, suggesting a potential imbalance or difficulty in classifying this class accurately. The overall accuracy of the model is 82%, respectively.

A. Comparative Analysis and Discussion

A comparative study of homeowners is presented in this section, as auto insurance claims on the Auto Insurance Claims Fraud Detection Dataset. In Table III, ML and DL models such as XGBoost, RF[15], RNN[16], and MLP[17] are contrasted using performance matrices including AUC-ROC, f1-score, recall, accuracy, and precision.

TABLE III. ML AND DL MODEL'S COMPARISON ON THE AUTO INSURANCE CLAIMS FRAUD DETECTION DATASET

Performance Measures	**XGBoost**	**RF[15]**	**RNN[16]**	**MLP[17]**
Accuracy	82.5	81.2	60.61	72.25
Precision	80.80	80.6	-	-
Recall	80.39	-	90.74	64.58
F1-score	80.59	81.4	58.69	72.25

An analysis of machine learning (ML) and deep learning (DL) models' accuracy, precision, recall, and F1-score using the Auto Insurance Claims Fraud Detection Dataset is shown in Table III. With an F1-score of 80.59%, XGBoost attains the best accuracy (82.5%) while maintaining a balanced precision (80.80%) and recall (80.39%). With an accuracy of 81.2% and a little higher F1-score (81.4%), Random Forest (RF) comes in second, albeit its recall value is not disclosed. Among deep learning models, Recurrent Neural Network (RNN) exhibits the highest recall (90.74%) but suffers from low accuracy (60.61%) and F1-score (58.69%), indicating potential overfitting to positive cases. Multi-Layer Perceptron (MLP) achieves a moderate accuracy (72.25%) with an equivalent F1-score, though its recall (64.58%) is comparatively lower than RNN. Overall, XGBoost outperforms the deep learning model in terms of balanced performance, highlighting its efficacy in fraud detection within auto insurance claims.

V. Conclusion And Future Work

This study uses ML models to identify fraud and employs a data-driven technique to examine moral hazard in auto insurance claims. The research adopts preprocessed data featuring engineered characteristics to run XGBoost models trained for detecting auto insurance fraud. The models received evaluation through multiple metrics, including accuracy as well as precision, recall, F1-score and AUC-ROC scores. XGBoost proved most suitable for insurance claim fraud detection based on experimental data because it achieved 80.80% precision while maintaining 80.39% recall. Ensemble methods developed using ML prove superior to deep learning models for detecting insurance claim fraud, which makes them appropriate for actual insurance fraud detection systems. To improve classification skills, more research should examine mixed models and complex deep learning network architectures. Fraud detection systems can gain interpretability together with robustness through the integration of external data sources in addition to explainable AI techniques.

References

[1] C. O. Omari, S. G. Nyambura, and J. M. W. Mwangi, "Modeling the Frequency and Severity of Auto Insurance Claims Using Statistical Distributions," *J. Math. Financ.*, 2018, doi: 10.4236/jmf.2018.81012.

[2] D. Kenyon and J. H. P. Eloff, "Big data science for predicting insurance claims fraud," in *2017 Information Security for South Africa - Proceedings of the 2017 ISSA Conference*, 2017. doi: 10.1109/ISSA.2017.8251773.

[3] S. Viaene, R. A. Derrig, B. Baesens, and G. Dedene, "A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection," *J. Risk Insur.*, 2002, doi: 10.1111/1539-6975.00023.

[4] J. van Wolferen, Y. Inbar, and M. Zeelenberg, "Moral Hazard in the Insurance Industry," *Netspar Panel Pap.*, 2013.

[5] A. Gogineni, "Novel Scheduling Algorithms For Efficient Deployment Of Mapreduce Applications In Heterogeneous Computing," *Int. Res. J. Eng. Technol.*, vol. 4, no. 11, p. 6, 2017.

[6] R. Roy and K. T. George, "Detecting insurance claims fraud using machine learning techniques," *Proc. IEEE Int. Conf. Circuit, Power Comput. Technol. ICCPCT 2017*, 2017, doi: 10.1109/ICCPCT.2017.8074258.

[7] P. Molk, "Playing with fire? Testing moral hazard in homeowners insurance valued policies," *Utah Law Rev.*, vol. 2, no. 3, pp. 347–409, 2018.

[8] A. H. Anju, "Extreme Gradient Boosting using Squared Logistics Loss function," *Int. J. Sci. Dev. Res.*, vol. 2, no. 8, pp. 54–61, 2017.

[9] G. Kowshalya and M. Nandhini, "Predicting Fraudulent Claims in Automobile Insurance," in *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, 2018. doi: 10.1109/ICICCT.2018.8473034.

[10] T. Badriyah, L. Rahmaniah, and I. Syarif, "Nearest Neighbour and Statistics Method based for Detecting Fraud in Auto Insurance," in *Proceedings of the 2018 International Conference on Applied Engineering, ICAE 2018*, 2018. doi: 10.1109/INCAE.2018.8579155.

[11] S. Subudhi and S. Panigrahi, "Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud," in *Proceedings - 2nd International Conference on Data Science and Business Analytics,*

ICDSBA 2018, 2018. doi: 10.1109/ICDSBA.2018.00104.

[12] S. Kareem, R. B. Ahmad, and A. B. Sarlan, "Framework for the identification of fraudulent health insurance claims using association rule mining," in *2017 IEEE Conference on Big Data and Analytics, ICBDA 2017*, 2017. doi: 10.1109/ICBDAA.2017.8284114.

[13] Y. Li, C. Yan, W. Liu, and M. Li, "Research and application of random forest model in mining automobile insurance fraud," in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2016, pp. 1756–1761. doi: 10.1109/FSKD.2016.7603443.

[14] M. A. Fauzan and H. Murfi, "The accuracy of XGBoost for insurance claim prediction," *Int. J. Adv. Soft Comput. its Appl.*, 2018.

[15] Y. Wang and W. Xu, "Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud," *Decis. Support Syst.*, 2018, doi: 10.1016/j.dss.2017.11.001.

[16] S. P. Sharmila Subudhi, "Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection," *J. King Saud Univ. Comput. Inf. Sci.*, 2017.

[17] G. G. Sundarkumar and V. Ravi, "A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance," *Eng. Appl. Artif. Intell.*, 2015, doi: 10.1016/j.engappai.2014.09.019.