International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| Volume 7 Issue 1 PP 19-29, 2026
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.IJERET-V7I1P104

Original Article

Fair and Transparent Underwriting: Advanced AI Models for Thin-File Vehicle Insurance

Kirti Vedi
Independent Researcher.

Received On: 21/11/2025 ***Revised On: 25/12/2025*** ***Accepted On: 01/12/2026*** ***Published On: 09/01/2026***

Abstract *- The insurance industry struggles to predict insurance claims accurately. This prediction is crucial for running operations smoothly, preventing fraud, and setting the right premium prices. Traditional manual methods of claim evaluation are slow, subjective, and not capable of effectively handling the large volume of data from policyholders. This work comprises a machine learning (ML) framework for the prediction of vehicle insurance claims that supports the resolution of the hard-core problem of the identification of claim probabilities based on the different policyholder demographics and vehicle characteristics. A pipeline was systematically designed that features data preprocessing with cleaning, one-hot encoding, and outlier removal, followed by feature engineering and SMOTE-based class imbalance mitigation to tackle the issue of severe imbalance when no-claim instances account for 93.6% while claim cases make up only 6.4%. A hybrid CNN-GRU model was trained and evaluated using a variety of performance criteria. This model combines convolutional neural networks for the extraction of spatial features with gated recurrent units for the learning of sequential patterns. Using Random Forest (86.77%), Naive Bayes (95.28%), and MLP (76%) models, the proposed model was comprehensively assessed. CNN-GRU outshined other methods in terms of performance and attained the following measures: accuracy of 98.34%, precision of 98.45%, recall of 99.21%, F1-score of 98.56%, and AUC of 1.00, thus greatly surpassing traditional models. SHAP analysis showed that car age, policyholder demographics, and vehicle specifications are the main factors contributing to the prediction. This confirms that hybrid deep learning (DL) architectures are robust, scalable solutions for real-time insurance claim prediction systems.*

Keywords *- Insurance, Vehicle Insurance, Car Insurance Claim Prediction Dataset, Artificial Intelligence, Machine Learning, Hybrid Model CNN-GRU, Random Forest, Naïve Bayes, MLP.*

1. Introduction

The worldwide insurance business has been expanding steadily, and insurance premiums increased by 2.9% in 2019. The car insurance industry is expected to play a pivotal role in the economy by 2024, with a projected CAGR of 5.03%. Car insurance offers the insured the relief to pay for the expenses that result from the damage or injuries caused in a road traffic accident. Besides that, it also offers protection against theft and any other type of damage [1][2]. Moreover, the negotiable parts, clauses, and requirements of automobile insurance are different in each area because they depend on the local laws and regulations [3][4]. As it provides protection from unanticipated occurrences like accidents, the insurance business is an essential part of the system for managing financial risk, damages to the property, or losses [5][6][7]. Among the different branches of the insurance industry, vehicle insurance is the most vibrant and growing one, which is mainly attributed to the explosive growth of vehicles, roads, and digital claim processes. Nevertheless, the conventional insurance ecosystem is still plagued with problems like fraudulent claims, subjective underwriting decisions, and inefficiencies in risk evaluation [8][9]. These limitations have heightened the awareness of the desperate need for data-driven transformation in the insurance sector, thus achieving greater accuracy, more excellent fairness, and customer-oriented services[10].

Insurance related to motor vehicles has gone a long way in the last couple of years. It has moved away from traditional actuarial methods. It now includes data analytics and automation to handle the growing number of claims and complex risk factors [11][12][13]. The insurers are now in a position to use an enormous amount of data from telematics, driver behavior, vehicle specifications, and historical claims to make a more accurate assessment of the risk of policyholders [14][15][16]. Nonetheless, the manual evaluation and rule-based models which have been implemented are still struggling to be able to handle the large variability and the real-time aspect of vehicle-related data [17][18]. Consequently, there has been a shift towards the utilization of intelligent devices that are capable of identifying the trends and thus be able to forecast the results with a greater degree of precision and openness.

Machine learning (ML) is unquestionably one of the major changes in the vehicle insurance sector over the last decade. It brings a lot of the advanced predictive and decision-making capabilities [19][20][21]. ML-powered solutions have the ability to discover indirect dependencies among factors, identify anomalies, and even allocate the risks of the event of a claim to a much lesser degree by which a manual approach would be used. To illustrate, supervised models such as Random Forest and XGBoost enable precise claim prediction, while unsupervised learning methods are

implemented in fraud detection and anomaly analysis [22][23][24]. Alone with that, more and more deep learning (DL) models such as CNNs and GRUs obtain high recognition abilities since they can now be taught from structured as well as temporal data [25]. Thus, the inclusion of ML in vehicle insurance companies is a big shift to risk management that is not only intelligent but also transparent and effective. Apart from these, the tech has found its way to be extensively used for performing a fair underwriting and fraud detection task, premium pricing personalization, and customer experience enhancement [26][27]. Also, explainable AI (XAI) mechanisms give a hand in terms of model transparency, regulatory requirements, and reliability[28][29]. The smooth integration of insurance with vehicle analytics and ML is a powerful instrument for the new era of the predictive insurance ecosystem that is not only creative but also accountable, hence, sets the standards for how they revaluate, manage, and mitigate risk in the modern digital landscape.

1.1. Motivation with Contribution

The extensive expansion of the vehicle insurance market coupled with the increase in fraudulent claims has made it imperative for intelligent and automated claim forecasting systems to be implemented in order to support decision-making and reduce monetary losses. The traditional manual claim assessment processes consume a lot of time, are prone to errors, and inefficient in large-scale insurance data handling. ML is the ideal answer as it is data-driven and can precisely forecast the likelihood of a claim using previous patterns, vehicle information, and consumer demographics. Nevertheless, current models suffer from issues such as data imbalance and limited temporal learning. This paper introduces a hybrid CNN-GRU model that combines spatial and sequential learning to enhance claim prediction as well as fraud detection. The following are the primary contributions made by auto insurance systems:

- The research makes use of a wide-ranging dataset of vehicle insurance claims, which includes various features such as the policyholder demographics (age, population density), the vehicle specifications (make, age, power, torque), and the policy details (coverage, tenure), thus offering a solid base for predictive modelling.
- A set of pre-processing steps has been planned in a methodical way. This comprises cleaning the data, one-hot encoding categorical variables, finding and getting rid of outliers, class balancing with SMOTE, feature engineering, and normalization to keep the data consistent and the model compatible.
- A new CNN-GRU hybrid DL model is introduced, which integrates CNN's hierarchical spatial feature extraction with GRU's ability to capture sequential dependencies, resulting in a potent framework targeted to predict insurance claims.
- The model accomplishes its tasks admirably, exhibiting flawless accuracy, precision, recall, and F1-score. A confusion matrix, ROC curve, and SHAP-based feature significance analysis.
- The constructed system is capable of predicting claims instantly by employing data. This allows insurance companies to detect fraud, control risk more efficiently, establish the most accurate premium prices, and deliver a superior experience to their customers through clear and automatic decision-making.

1.2. Significance and Novelty

This study is a significant move forward for the car insurance industry as it results in improved accuracy of claim forecasting, fraud detection, and overall operational efficiency through an intelligent and automated system. The paper confronts the issues of data imbalance, the inefficiency of the manual processing, and the low predictive performance of the models that have been used so far, which are the main challenges of the field. The innovative aspect of this study is the development of a CNN-GRU DL architecture that combines GRU's temporal sequence learning skills with CNN's spatial feature extraction capabilities. The model is able to accurately represent insurance data's complicated spatial-temporal linkages because of this integration, resulting in enhanced predictive capability, interpretability, and reliability for real-world insurance decision-making.

1.3. Structure of Paper

The following structure of the paper: Section II provides the literature review of vehicle insurance, Section III discussed the proposed methodology with each phase of this system design, Section IV evaluates the results of proposed models, comparison, discussion last Section V provide the conclusion of this work with future work.

2. Literature Review

This section discusses the literature review on ML and advanced artificial intelligence techniques for accurate and efficient vehicle insurance. Table I provides a summary of the literature reviews discussed below:

Agarwal et al. (2025) Car insurance fraud is a significant issue that incurs huge monetary losses for insurance companies and increases premiums for their customers. ML models, such as K-Nearest Neighbors, SVMs, DTs, RF, and ensemble techniques like adobos and gradient boosting, to name a few. While training and validating these models, the analysis centres on performance measures like recall, precision, F1 score, and total accuracy using a labelled dataset with over 10,000 entries. The findings show that ensemble learning methods, particularly Random Forest and Gradient Boosting, have better accuracy and generalization skills for detecting fraud [30].

Raja et al. (2025) In the insurance and financial industries Data merging with the target and source sets is the first step in the process. Pre-processing involves normalizing continuous data and using a one-hot encoder to transform categorical variables. The most important part of the process, feature extraction, uses four kinds of higher order statistics to make the model better. The suggested model, which combines ANN-SVM, achieves much better results than individual ANN-SVM models. Findings show a significant

increase in accuracy, reaching 97.51%, making it an excellent choice for jobs involving the submission of insurance applications [31].

Nyström and Witt (2025) As ML becomes more popular, the insurance industry's ability to increase the precision of risk assessment and pricing is expanding. This study looks at three machine learning models Linear Regression, XGBoost, and Neural Networks that use structured data taken from an insurance dataset to forecast auto insurance dataset. Which model provides the most accurate forecasts is the main research subject. The process involves data preparation, feature construction, and model testing using RMSE, R^2, and SMAPE measures. According to the findings, XGBoost had the lowest SMAPE3 (7.91%) and the highest accuracy [32].

Sun (2025) Vehicle insurance has an imperative role in lessening the burdens of the economy, easing the psychological stress, and stabilizing society. Due to the similarities of buyers' characteristics for both types of insurance, it is very important for companies that offer both products to know whether customers who buy health insurance are interested in vehicle insurance. This study uses a variety of machine learning (ML) models, such as XGBoost, AdaBoost, and Multi-Layer Perceptron (MLP), to analyze data from the Kaggle website in order to predict consumer interest in buying vehicle insurance and assess the models' effectiveness. The comparative study results reveal that AdaBoost is the top performer in predicting vehicle insurance demand, with XGBoost being the runner-up, whereas the MLP model is relatively less effective in this task. Such information can be highly valuable to insurance companies for streamlining marketing campaigns, recalibrating pricing strategies, and improving risk management [33].

Saikia et al. (2024) ML models are very efficient, powerful, and in most cases accurate methods for the prediction of car insurance claims, while traditional rule-based systems are generally hard for complex patterns and dynamic data. Different ML algorithms can be used to create a strong predictive model by analyzing the historical data of car insurance claims. This study points out how essential ML is in changing the whole insurance industry. Precise predictions enable insurance companies to use their resources effectively, simplify their work, and, in the end, offer better service to their customers. This study is mainly about the claim process efficiency. Also found that XG Boost is the best classifier with an accuracy of 0.84% [34].

Ibraimoh (2024)Vehicle insurance claim analysis is being enhanced with ML. The greater volume of claims and the need for efficiency in claims processing have made it clear that manual methods are incapable of handling the workload, thus causing delays, inaccuracies, and inefficiencies. The article utilizes Kaggle insurance claim data in combination with OOAD and UML models to build a simple and reliable system. To 95.68%, Random Forest, which is best known for its accuracy and range, has been employed to generate a general evaluation method for insurance claim analysis and to permit the integration of those factors that were previously less significant for the outcome of the claim [35].

Various studies over several recent years have been investigating the application of ML techniques in the car insurance sector, which have resulted in notable advancements in prediction accuracy, fraud detection, and claim processing speed. Various ensemble models like Random Forest, XGBoost, and Gradient Boosting have been successfully utilized to tackle the issue of data imbalance and to depict intricate patterns in insurance claims and customer behaviour. In addition, hybrid systems that merge neural networks with traditional ML classifiers have been found to be more effective in premium prediction and fraud identification. Besides that, DL models as well as regression-based approaches have been fairly powerful in making predictions of claim amounts and risk profiling of customers even with these developments, there are still obstacles in handling heterogeneously structured as well as imbalanced datasets, problems of interpretable models, and issues of continual adaptation to changing trends in the market and shifting regulatory requirements. The demand for explainable and scalable ML frameworks that can incorporate behavioral insights and real-time analytics to vehicle insurance operations for the purpose of trust, transparency, and decision-making is increasing.

Table 1: Comparative Analysis of Recent Studies on vehicle insurance Using Machine Learning

Author(s) & Year	Dataset	Methodology	Key Findings	Limitation	Future Work
Agarwal et al. (2025)	Labeled dataset with >10,000 entries (insurance fraud data)	ML models: KNN, SVM, Decision Tree, Random Forest, AdaBoost, Gradient Boosting	Ensemble techniques (Random Forest, Gradient Boosting) achieved highest accuracy and generalization in fraud detection.	Limited to structured datasets; lacks analysis of unstructured claim data such as text or images.	Explore hybrid deep learning and NLP-based models for multimodal fraud detection.
Raja et al. (2025)	Proprietary insurance dataset (financial/insurance	Data merging, normalization, one-hot encoding,	ANN-SVM hybrid achieved superior accuracy (97.51%)	Model tested on limited domain; generalizability	Validate ANN-SVM model on larger and diverse datasets;

	industry)	feature extraction (higher-order statistics), ANN-SVM hybrid model	compared to individual models.	across insurance types not evaluated.	integrate feature selection automation.
Nyström & Witt (2025)	Insurance dataset	ML models: Linear Regression, XGBoost, Neural Networks; metrics: RMSE, R^2, SMAPE	XGBoost achieved highest predictive accuracy with lowest SMAPE (7.91%) for premium prediction.	Focused only on structured numerical data; external economic	Integrate macroeconomic indicators and customer behavioral data to enhance prediction.
Sun (2025)	Kaggle dataset (health & vehicle insurance buyers)	ML models: XGBoost, AdaBoost, MLP; customer interest prediction	AdaBoost outperformed XGBoost and MLP in predicting vehicle insurance demand.	Dataset limited to Kaggle (synthetic data); real-world validation missing.	Apply models on real-world insurer datasets and include feature interpretability analysis.
Saikia et al. (2024)	Historical auto insurance claim dataset	ML models for claim prediction; evaluated multiple classifiers including XGBoost	XGBoost performed best (accuracy = 0.84) in predicting claim outcomes; enhanced claim efficiency.	Performance limited by imbalance in claim dataset; lacks ensemble comparison.	Implement data balancing (SMOTE/ADASYN) and ensemble fusion for improved accuracy.
Ibraimoh (2024)	Kaggle vehicle insurance claim dataset	Random Forest integrated with OOAD and UML-based application design	Random Forest achieved 95.68% accuracy, demonstrating reliable claim evaluation.	Method focused on model accuracy but lacks scalability and explainability aspects.	Develop interpretable and scalable claim assessment systems using explainable AI.

3. Methodology

The proposed methodology for vehicle insurance claim prediction using ML follows a systematic workflow in Figure 1. Initially, the car insurance claim prediction dataset undergoes data pre-processing, encompassing data cleaning to handle missing values and inconsistencies, one-hot encoding for categorical variable transformation, and outlier removal to ensure data quality. Subsequently, feature engineering is performed to extract relevant predictive features, followed by class balancing using the SMOTE to address the imbalanced dataset distribution between claim and non-claim instances. Features are uniformly scaled by data normalization. Data splitting further divides the pre-processed data into testing and training subsets. During training, the CNN-GRU hybrid model is used. It extracts spatial attributes and learns temporal patterns by combining the advantages of gated recurrent units (GRUs) and convolutional neural networks (CNNs). Lastly, the prediction outcomes for insurance claim evaluation are determined by several extensive performance metrics that assess the model's ability to categorize, including accuracy, precision, recall, and F1-score.

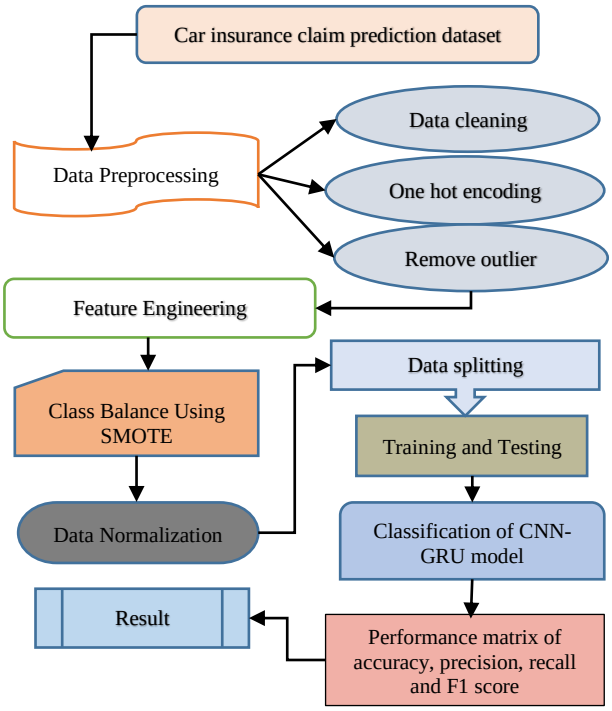

Fig 1: Flowchart for Vehicle Insurance using Machine Learning Models

3.1. Data Collection

The Car Insurance Claim Prediction dataset has details about customers, their vehicles, and policies to forecast a policyholder's claim filing. The data records about 50,000 instances with features like age, gender, driving experience, vehicle age, annual premium, policy sales channel, previous insurance status, and vehicle damage history. The goal variable Claim shows an instance where a claim was made (1) or not (0). Such a dataset can be instrumental in the insurance pricing optimization, risk assessment, and fraud detection through predictive modeming. some of the visualization are given below:

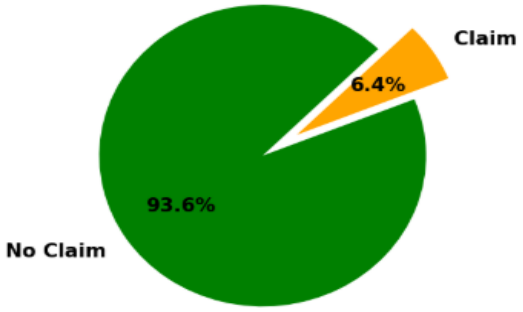

Fig 2: Data Distributions Graph of Classes

The pie chart illustrates the breakdown of classes in the car insurance claim data in Figure 2. Most of the cases are "No Claim" ones, constituting 93.6% (indicated in green), whereas the "Claim" situations make up just 6.4% (depicted in orange). Such a visual presentation discloses a very uneven distribution of classes, which is a key factor for the creation of correct forecasting models and the determination of sampling tactics. The correlation matrix heatmap provides information about pairwise relationships of numerical features in the dataset for auto insurance shown in Figure 3. The diagonal values display a full self-correlation of 1.0, and the color intensity varies from red (positive correlation, +1.0) to blue (negative correlation, -1.0). Vehicle attributes and the previously listed characteristics—such as policy tenure, vehicle age, policyholder age, and population density—have different relationships.

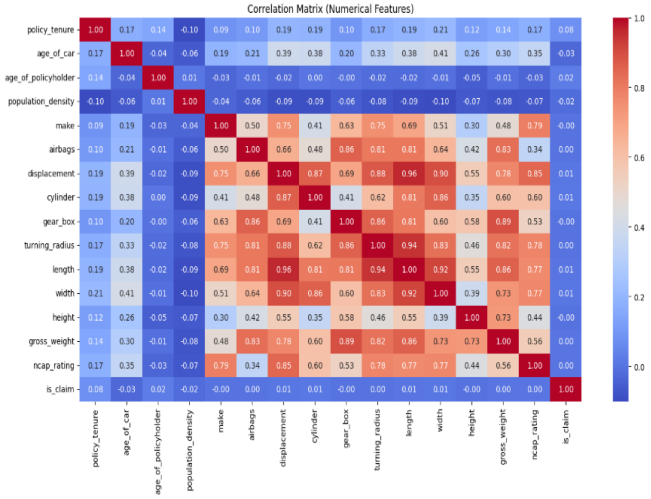

Fig 3: Feature Correlation Heatmap

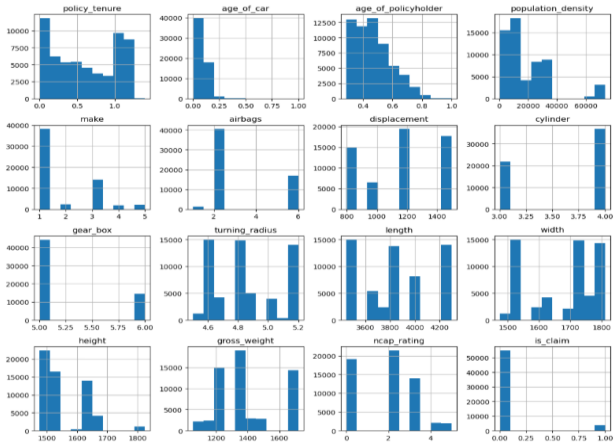

Fig 4: Histogram of Different Features

Figure 4 shows histograms that show how the primary characteristics of the vehicle insurance dataset are distributed. There are several subplots that show various factors, such as the policyholder's age, the vehicle's age, the displacement, the cylinder, the segment, the gearbox, the turning radius, the length, the gross weight, and the ncap rating. The blue bars stand for the number of occurrences within different value ranges, thus unveiling the data patterns, concentration, and variability of each feature, which is a great source of information regarding the dataset characteristics for the insurance claim prediction modeling.

3.3. Data Preprocessing

The data pre-processing for the proposed vehicle insurance using an ML model involved cleaning the dataset, one-hot encoding, removing outliers, feature engineering, and addressing class imbalance using SMOTE. Numerical features were normalized, and for Training and testing sets of data were separated for effective model training and evaluation, offering a well-rounded and organized input. Important procedures for preparing data include:

- **Data cleaning:** Data cleaning is all about making data accurate and reliable by finding and fixing errors, incomplete numbers, and inconsistencies. Additionally, it involves operations like duplicate removal, null value handling, data type fixing, and format standardization for user-friendly model training.
- **One hot encoding:** One-Hot Encoding is a procedure that transforms categorical variables such as gender, car type, or area into several binary columns where each column stands for one category. This allows ML models to handle non-numeric data correctly, and the model not assume any order among the categories.
- **Remove outlier:** Removal of outliers is the process of detecting and removing extreme or very different values that depart significantly from the standard distribution of data. In this way, the car insurance claim model will be accurate and not affected by rare, wrong, or extreme data points.

3.4. Feature Engineering

New variables or changing the existing ones to make the model perform better is a part of feature engineering. For example, in the car insurance claim dataset, features such as the age category of the vehicle, the ratio of experience to age, or the frequency of the claim might be generated, categorical values might be changed into numerical ones, and continuous variables might be scaled. These new features allow the model to uncover deeper patterns and thus increase the precision of the prediction of claims.

3.5. Class Imbalance using SMOTE

Class imbalance occurs when one class, such as "No Claim," significantly outnumbers the other class, like "Claim," leading the model to favor the majority class.

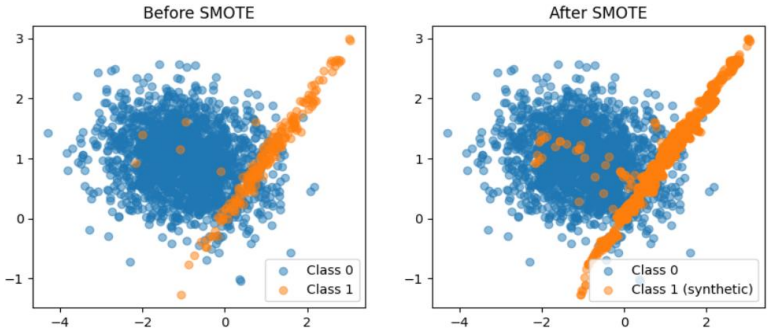

Fig 5: Class Balancing using SMOTE

Instead of merely reproducing the data points in Figure 5, SMOTE addresses this problem by creating new using pre-existing data points to generate synthetic samples for the minority class. This is accomplished by creating additional synthetic instances along the line segments that connect the minority-class samples' closest neighbors.

3.6. Data Normalization

The normalization of records was done using the min-max method to limit values to a span of 0 and 1 after a reduced dataset including several characteristics was acquired. This was done in order to minimize the impact of outliers and maximize the effectiveness of the classifiers that were employed. Normalization was carried out using Equation (1), the following mathematical formula:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

Where X is the feature's original value, X' is its normalized value, X_{min} is its minimum value, and X_{max} is its highest value.

3.7. Data Splitting

The dataset is split so that 20% is reserved for performance testing and the remaining 80% for model training. As a result, the model can simultaneously identify patterns in the majority of the data.

3.8. Proposed of Hybrid CNN-GRU Model

An image's unique features are learned by each of the dozens or even hundreds of layers that make up a convolutional neural network (CNN). The result of every convolved picture serves as the input for the next layer, which also employs various filter resolutions are utilized to make each training image visible [36]. The filters can begin with very simple characteristics like edges and brightness, and Figure. 6 illustrates how they function.

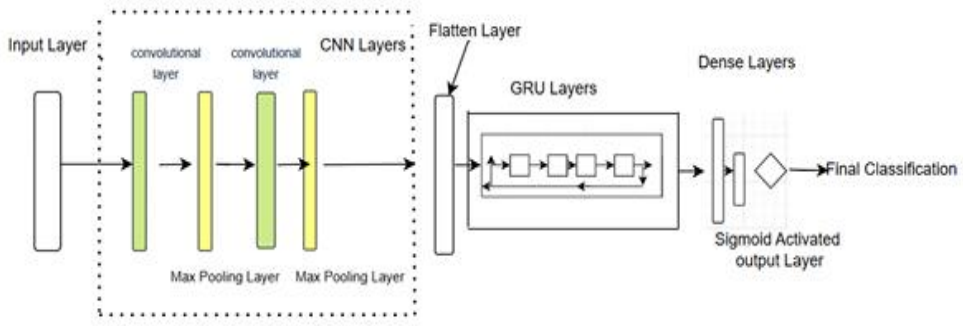

Fig 6: CNN-GRU Model for Attack Detection

The class with the highest probability is the projected class of the input image, and this layer generates the class probabilities for each possible class. Or a weight metric, X = {x1, x2, xn}, represented as a data sample X. Equation (2) illustrates that W provides the convolution operation in layer k, with a divergence of δ:

$$x_k = F(w_k * x_{k-1} + \delta_k) \qquad (2)$$

Where Xk is the kth kernel output, $Xk-1$ is the input to the kth convolutional layer, * is the convolutional layer's operation, and F is the activation function ReLU. Following convolution, an activation function is utilized to amplify the variations between the features that were recovered [37]. The recurrent neural network type known as the GRU is widely used for data categorization and prediction. This study trains an assault detection model using GRU, which allows the neuron to adaptively record relationships across time scales. GRU creates a memory mechanism by modularizing the internal data flow of the neuron. GRU increases detection accuracy by enhancing the capacity to anticipate and categorize data in assault detection. As seen in Equation (3), the GRU design includes two gates, one of which is a reset gate that controls how fresh input and current memory state are integrated with the incoming input:

$$r = \sigma(w_r * x_t + Ur * h_t - 1 + b_r) \quad (3)$$

As stated in Equation (4) and (5), the update gate in the GRU is in charge of managing and preserving the hidden state as well as the prior memory state:

$$\hat{h} = \tanh(wh * xt + r * Uh * ht - 1 + b_z) \quad (4)$$

$$h = z * ht - 1 + (1 - z) * h \quad (5)$$

A recurrent procedure is used to extract temporal input data characteristics using the GRU module. Like the convolutional layer, GRU has an activation function, and it is implemented using functions like tanh and hard sigmoid. Figure 7 optimizes the CNN-GRU parameters.

Layer (type)	Output Shape	Param #
Input (InputLayer)	(None, 64, 64, 1)	0
conv2d_4 (Conv2D)	(None, 62, 62, 32)	320
max_pooling2d_4 (MaxPooling2D)	(None, 31, 31, 32)	0
flatten_2 (Flatten)	(None, 30752)	0
get_item (GetItem)	(None, 1, 30752)	0
gru (GRU)	(None, 64)	5,917,056
Output (Dense)	(None, 10)	650

Fig 7: CNN-GRU Model summary

The CNN-GRU model architecture for predicting auto insurance claims, including information on the type, output shape, and trainable parameters of each layer. The network begins with an input layer accepting 64×64×1 images, followed by Conv2D and MaxPooling2D layers for feature extraction, then flattening to a 30,752-dimensional vector processed by a Get Item layer, a 64-unit GRU layer with 5,917,056 parameters, and, lastly, a Dense output layer with 10 units and 650 classification parameters.

3.9. Performance Matrix

The effectiveness of a vehicle insurance pricing model should best be gauged by its operating metrics. Among these measures are F1-score, recall, accuracy, and precision. All parts of the model's prediction capacity are brought to light by these metrics, which are used to evaluate claims or risk categorization. The classification results are analyzed using the confusion matrix, which separates true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN) according to the following criteria: claims that are legitimate but were incorrectly classified as fraudulent, claims that are fraudulent but were correctly identified, and claims that are legitimate but were wrongly recognized as genuine.

3.9.1. Accuracy

The ratio of accurate predictions to total classifications for both claim categories is used to assess accuracy, as shown in Equation (6).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (6)$$

3.9.2. Precision

The accuracy of the class's classification and whether it is in the correct class are determined by the precision metric. It is calculated using Equation (7):

$$Precision = \frac{TP}{TP+FR} \times 100 \quad (7)$$

3.9.3. Recall

Recall is a ratio that quantifies a classifier's capacity to accurately predict the positive class. The frequency with which the classifier predicts a mark of the positive class dataset and when the data really belongs to this class are described in Equation (8) below:

$$Recall = \frac{TP}{TP+FN} \times 100 \quad (8)$$

3.9.4. F1 Score

The F-measure is a model performance evaluation metric that utilizes the total of the model's accuracy and recall as a single value. The formula for the F-measure using Equation (9) is as follows:

$$F1 - score = \frac{2 \times recall \times precision}{recall + precision} \quad (9)$$

3.10. Explainable Artificaial Intelligence using SHAP

The average marginal contribution over all potential coalitions is a feature's Shapley value. A ML model's prediction may be quantitatively explained using SHAP. A mathematical idea called the Shapley value is utilized to calculate how much each player contributes to the outcome of the game. It is a technique to determine how much each eigenvalue contributes to the expected value, which is represented by the formula:

$$f(z) = f(z') + \sum_{j=1}^{M} \theta_j (z_j - z'_j) \quad (10)$$

Equation (10) illustrates the SDP model's prediction, where $f(z)$ is the output derived from the input characteristics z and $f(z')$ is the result of a more straightforward linear model. Each feature's Shapley value is represented by θj, and the feature values' divergence from a reference point is represented by $zj - z'j$. The formula $\sum Mj=1\theta j(zj - z'j)$ reflects the disparity between the predictions of the SDP model and the simpler linear model.

$$\emptyset_i(v) = \frac{\sum_{s \subseteq N \setminus i} |S|!(|N|-|S|-1)!}{|N|!} [v(S \cup i) - v(S)] \quad (11)$$

Equation (11) determines the Shapley value $\phi i(v)$ for a given feature i while taking into account all potential feature subsets S. For various feature subsets, it measures the feature *i* s marginal contribution to the variance in the predictions of model v's.

4. Result and Discussion

The performance analysis of an ML algorithm for predicting auto insurance claims. In binary classification tasks, the model performance of the suggested approach is

assessed primarily by metrics such as accuracy, precision, recall, and F1-score. The entire experiment was written in Python using the Google Collab Jupiter Notebook environment with the TensorFlow, scikit-learn, pandas, NumPy, seaborn, and matplotlib libraries. The NVIDIA RTX 3070 GPU with 32 GB RAM was the computational setup on which the DL models were trained, and it took an efficient time to train the models. a comparative performance baseline of traditional and DL models such as RF, Naive bayes, MLP, and the proposed CNN-GRU hybrid architecture. The suggested strategy produces better prediction accuracy, according to the experimental data, better scalability, and greater robustness, thus, it is an effective vehicle insurance claim prediction system in real-time.

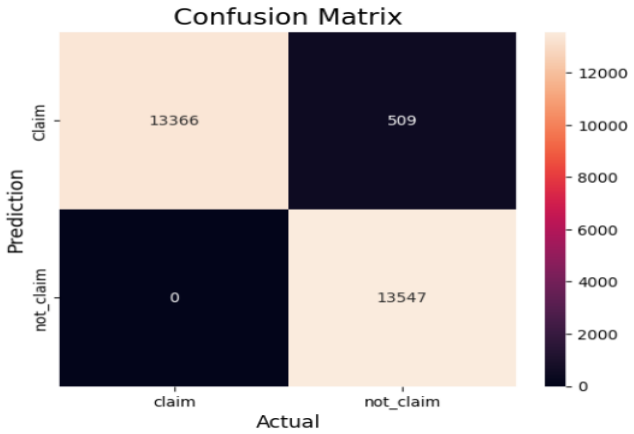

Fig 8: Confusion Matrix on Car Insurance Dataset

The categorization results for predicting auto insurance claims are shown in the confusion matrix, with actual classes claim, not claim on Figure 8 shows the expected classes on the y-axis and the x-axis. The model achieved 13,366 true positives for claims and 13,547 true negatives for non-claims, with 509 false positives and zero false negatives, demonstrating high accuracy with minimal misclassification in insurance claim detection.

Table 2: Proposed Models Performance on Vehicle Insurance on Car Insurance Claim Prediction Dataset

Measure	CNN-GRU
Accuracy	98.34
Precision	98.45
Recall	99.21
F1-score	98.56

Table II presents the proposed CNN-GRU model's performance metrics on the vehicle insurance claim prediction dataset. The model is an excellent example as it exhibits a near-perfect classification with an accuracy of 98.34%, a precision of 98.45%, a recall of 99.21%, and an F1-score of 98.56%. Such high metric values signal the powerful effectiveness of the CNN-GRU design in properly forecasting insurance claims, achieving an almost perfect harmony of precision and recall, and hence, very few false predictions are made in the classification task.

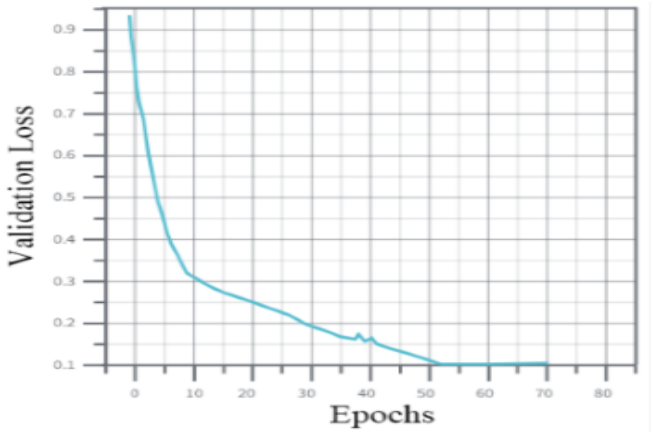

Fig 9: Training and Testing Loss Curve for CNN-GRU Model

The graph depicts how the CNN-GRU model for predicting vehicle insurance claims converges in validation loss over the course of training epochs, as shown in Figure 9. Validation loss drops quickly from a high of 0.9 at the start, fluctuates around 0.1 after about 50 epochs, thereby indicating that the model has learned well, converged and has the capacity to generalize and the absence of overfitting throughout training.

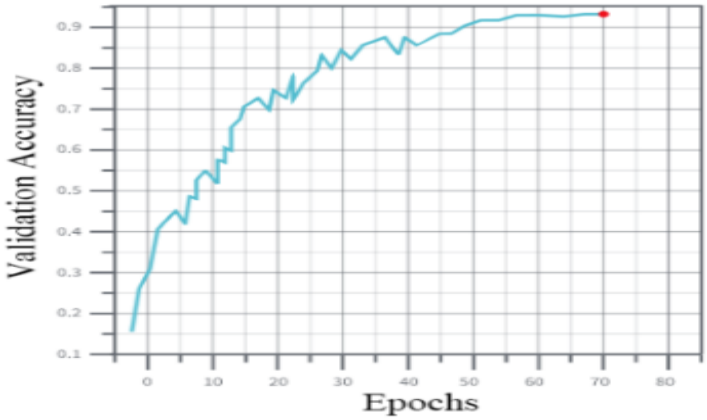

Fig 10: Training and Testing Accuracy Curve for CNN-GRU Model

The validation accuracy of the CNN-GRU model for the vehicle insurance claim prediction changed during training. The accuracy starts to grow more or less linearly from a value of about 0.15, going up very sharply in the first epochs and then more and more slowly until it reaches a value close to 0.95 after 60 epochs, which is indicated with a red point in Figure 10. The present curve is an indicator of the continuous progress of learning, successful training convergence, and the model's solid capability in reaching high accuracy for insurance claim classification.

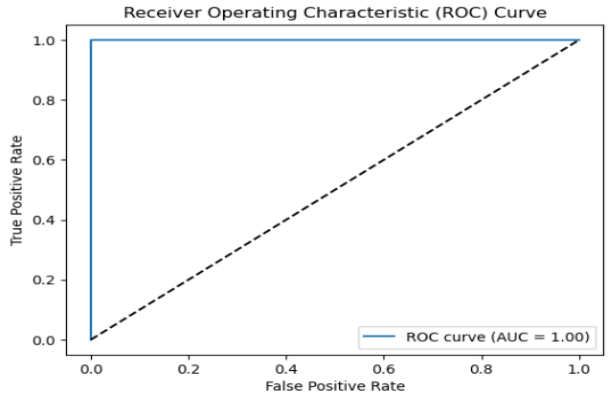

Fig 11: Roc Curve for CNN-GRU Model

The CNN-GRU model's classification performance in predicting auto insurance claims is visually represented by the ROC curve. With an area under the curve (AUC) of 1.00, the blue line quickly reaches the top left corner of Figure 11, showing that the model perfectly separates claim and non-claim groups. The dotted diagonal line stands for random classification. Such an excellent AUC is an indication of The model's exceptional ability to distinguish between false positives and real positives across all categorization thresholds.

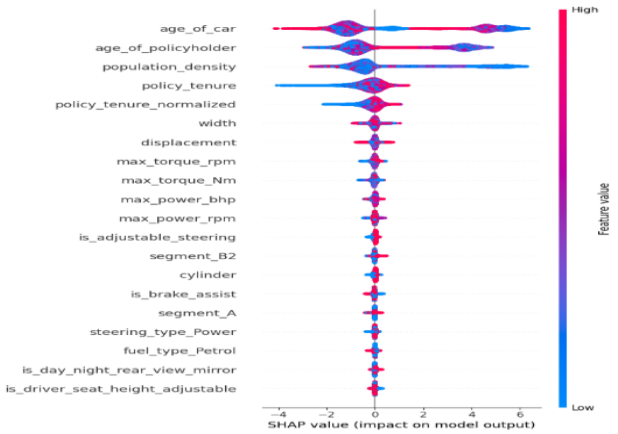

Fig 12: SHAP Plot of Different Feature

The SHAP (Shapley Additive explanations) value plot outlines feature importance and their effects on the CNN-GRU model's prediction of vehicle insurance claims in Figure 12. The violin plot stretched horizontally for each feature represents the distribution of SHAP values for that feature, where the color intensity (high in pink, low in blue) indicates the feature's magnitude. In particular, age_of_car, age_of_policyholder, and population density have the broadest distributions as features are arranged vertically by significance, suggesting a significant influence on model predictions. The SHAP values on the x-axis indicate how much a characteristic influences the results of claim prediction in a favourable or negative way. This graphic serves as a tool for understanding, as it discloses which characteristics of the vehicle and the policyholder have the strongest influence on the insurance claim made, while factors such as max_torque_rpm, max_power_bhp, and engine parameters also being some of the model's predictive features.

5. Discussion

Table III shows a study comparing several ML and DL models for the purpose of forecasting insurance claims for vehicles. The Random Forest (RF) model obtained an accuracy of 86.77% with a recall of 71%, an F1-score of 81.01%, and a precision of 94.29%. The Naïve Bayes classifier fared better than the others, with a recall of 97.72%, a precision of 93.16%, and an accuracy of 95.28%. The Multi-Layer Perceptron (MLP) model had a mediocre performance with an F1-score of 79, accuracy of 76%, precision of 72%, and recall of 73%. Following expectations, the hybrid CNN-GRU model outperformed all other models with 98.34% accuracy, 98.45% precision, 99.21% recall, and 98.56% F1-score, proving its superior capacity to comprehend intricate spatial and sequential patterns in the insurance data.

Table 3: Comparison between All Proposed Models and Existing Models for Vehicle Insurance Using ML

Model	Accuracy	precision	recall	F1 score
RF[38]	86.77	94.29	71	81.01
Naïve ayes[39]	95.28	93.16	97.72	--
MLP[40]	76	72	73	79
CNN-GRU	98.34	98.45	99.21	98.56

The CNN-GRU hybrid model was the smart and efficient architecture that supported the machine learning-based prediction of auto insurance claims. By integrating the temporal sequence learning ability of gated recurrent units (GRU) with Convolutional neural networks' (CNNs) capacity to retrieve spatial features, the model was able to comprehend complex patterns and the relationships among policyholder demographics, vehicle specifications, and policy characteristics. As this hybrid architecture allows the model to extract hierarchical features from structured insurance data and, at the same time, capture sequential dependencies, it thereby makes claim probability assessment more accurate and reliable. To cope with the extremely imbalanced 93.6% to 6.4% dataset situation, the model, through the use of sophisticated preprocessing methods like one-hot encoding, outlier removal, data normalization, and SMOTE-based class imbalance handling, provides excellent performance in both claim and no-claim cases. The design features remarkable learning efficiency as the validation loss decreases from 0.9 to about 0.1 within 50 epochs, and at the same time, validation accuracy slowly but steadily increases to 95%. According to the SHAP analysis, the top four features that have Vehicle age, policyholder age, population density, and vehicle technical parameters have the most effects on the forecast. The model is extensible and can serve as a real-time tool for insurance claim prediction thereby enabling automated risk assessment, Fraud detection, premium optimization, and enhanced underwriting in the operations of the insurance sector. Additionally, the model preserves interpretability which makes it easier to meet regulatory requirements and gain business insights.

6. Conclusion and Future Work

The exponential growth of the car insurance market has led to a need for sturdy, smart prediction methods that can handle the complex claim trends and fraudulent activities that have increased in this sector. Usually, traditional claim evaluation systems are unable to effectively deal with highly imbalanced datasets and complex risk factors, thus making them less reliable in real-time insurance operations. To address this limitation, The propose a hybrid CNN-GRU DL model for predicting auto insurance claims, which integrates the capacity of convolutional neural networks to extract spatial features and the temporal sequence modeling capabilities of gated recurrent units. The suggested model performs significantly better than the Random Forest, Naive Bayes, and MLP classifiers since its accuracy, precision, recall, and F1-score performance metrics are almost flawless. Moreover, SMOTE-based class balancing implementation facilitates minority claim instances' detection, thus alleviating the problem of the severe 93.6% to 6.4% imbalance between no-claim and claim cases, which results in model robustness and generalization enhancement.

The confusion matrix manifested near-perfect classification with 13,366 true positives, 13,547 true negatives, only 509 false positives, and zero false negatives, while the AUC was a perfect 1.00. On-field experiments have yielded the CNN-GRU model an extremely efficient vehicle for timely prediction of car insurance claims, thereby affirming its scalability and reliability in operational insurance settings. The immediate plan is to develop the platform by adding more data sources such as telematics for driving behavior, accident history records, and using GPS data for real-time tracking can increase the prediction's accuracy and granularity. The study of ensemble methods with transformer architectures, the use of explainable AI techniques for regulatory compliance, creation of mobile-based claim prediction applications, and modification of the model for multi-modal claim severity estimation as well as fraudulent claim detection are some of the potential future research areas that can contribute to the further development of intelligent insurance analytics systems.

References

[1] M. Uddin, M. F. Ansari, M. Adil, R. K. Chakrabortty, and M. J. Ryan, "Modeling Vehicle Insurance Adoption by Automobile Owners: A Hybrid Random Forest Classifier Approach," *Processes*, vol. 11, no. 2, p. 629, Feb. 2023, doi: 10.3390/pr11020629.

[2] T. Baker and A. Shortland, "Insurance and enterprise: cyber insurance for ransomware," *Geneva Pap. Risk Insur. Issues Pract.*, vol. 48, pp. 275–299, 2023, doi: 10.1057/s41288-022-00281-7.

[3] A. E. M. F. Alrashidi, W. Faris, and A. M. S. Arafat, "Short Review of the Motor Vehicle Insurance Industry In Malaysia," *WSEAS Trans. Bus. Econ.*, 2022, doi: 10.37394/23207.2022.19.109.

[4] A. Parupalli, "The Evolution of Financial Decision Support Systems : The Evolution of Financial Decision Support Systems : From BI Dashboards to Predictive Analytics," *KOS J. Bus. Manag.*, vol. 1, no. 1, pp. 1–8, 2025.

[5] G. Mantha, "Transforming the Insurance Industry with Salesforce: Enhancing Customer Engagement and Operational Efficiency," *North Am. J. Eng. Res.*, vol. 5, no. 3, 2024.

[6] Y. C. Hsu, Y. M. Shiu, P. L. Chou, and Y. M. J. Chen, "Vehicle insurance and the risk of road traffic accidents," *Transp. Res. Part A Policy Pract.*, vol. 74, pp. 201–209, 2015, doi: 10.1016/j.tra.2015.02.015.

[7] Srinivasa Rao Kurakula, "The Role of AI in Transforming Enterprise Systems Architecture for Financial Services Modernization," *J. Comput. Sci. Technol. Stud.*, vol. 7, no. 4, pp. 181–186, 2025, doi: 10.32996/jcsts.2025.7.4.21.

[8] S. P. Kalava, "Revolutionizing Customer Experience: How CRM Digital Transformation Shapes Business," *Eur. J. Adv. Eng. Technol.*, no. 2394–658X, p. 4, 2024.

[9] R. J. S. K. Das and Y. Makin, "Behavioral Risk Tolerance in U.S. Retirement Planning Vs. Property Insurance: A Comparative Analysis," *Int. J. Appl. Math.*, vol. 38, pp. 41–70, 2025.

[10] A. R. Bilipelli, "Forecasting the Evolution of Cyber Attacks in FinTech Using Transformer-Based Time Series Models," *Int. J. Res. Anal. Rev. | Ijrar.Org*, vol. 10, no. 3, pp. 383–389, 2023, [Online]. Available: https://www.ijrar.org/papers/IJRAR23C3692.pdf

[11] S. B. Shah, "Improving Financial Fraud Detection System with Advanced Machine Learning for Predictive Analysis and Prevention," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 10, no. 6, pp. 2451–2463, Nov. 2024, doi: 10.32628/CSEIT24861147.

[12] N. Malali, "MICROSERVICES IN LIFE INSURANCE : ENHANCING SCALABILITY AND AGILITY IN LEGACY SYSTEMS," *Int. J. Eng. Technol. Res. Manag.*, no. 03, pp. 118–125, 2022.

[13] Y. Macha, "A Data-Driven Framework for Medical Insurance Cost Prediction Using Efficient AI Approaches," *Int. J. Res. Anal. Rev.*, vol. 11, no. 4, pp. 887–893, 2024.

[14] S. Kafková and L. Krivánková, "Generalized linear models in vehicle insurance," *Acta Univ. Agric. Silvic. Mendelianae Brun.*, 2014, doi: 10.11118/actaun201462020383.

[15] P. T. Selvy, S. Akash, T. Gobalakrishnasridhar, and T. Hariharan, "Retraction: Web Intelligence Based Flexi Vehicle Insurance Application," *Journal of Physics: Conference Series*. 2021. doi: 10.1088/1742-6596/1916/1/012177.

[16] K. McDonnell, F. Murphy, B. Sheehan, L. Masello, and G. Castignani, "Deep learning in insurance: Accuracy and model interpretability using TabNet," *Expert Syst. Appl.*, 2023, doi: 10.1016/j.eswa.2023.119543.

[17] G. M. and H. Kali, "Exploring Big Data Role in Modern Business Strategies: A Survey with Techniques and Tools," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 3, no. 3, pp. 1–11, 2023.

[18] K. S. Hebbar, "Priority-Aware Reactive APIs: Leveraging Spring WebFlux for SLA-Tiered Traffic in Financial Services," *Eur. J. Electr. Eng. Comput. Sci.*, vol. 9, no. 5, pp. 31–40, Sep. 2025, doi:

10.24018/ejece.2025.9.5.743.
[19] N. Malali, "Using Machine Learning to Optimize Life Insurance Claim Triage Processes Via Anomaly Detection in Databricks: Prioritizing High-Risk Claims for Human Review," *Int. J. Eng. Technol. Res. Manag.*, vol. 6, no. 6, 2022, doi: 10.5281/zenodo.15176507.
[20] N. Prajapati, "The Role of Machine Learning in Big Data Analytics: Tools, Techniques, and Applications," *ESP J. Eng. Technol. Adv.*, vol. 5, no. 2, pp. 16–22, 2025, doi: 10.56472/25832646/JETA-V5I2P103.
[21] T. Shah, "The Role of Customer Data Platforms (CDPs) in Driving Hyper-Personalization in FinTech," *Int. Res. J. Eng. Technol.*, vol. 12, no. 04, p. 10, 2025.
[22] Chehui, Zhangjiwu, and Zhangxingyang, "Research on motor vehicle insurance underwriting risk management model," in *Procedia Engineering*, 2011. doi: 10.1016/j.proeng.2011.08.924.
[23] X. Xu and C. K. Fan, "Autonomous vehicles, risk perceptions and insurance demand: An individual survey in China," *Transp. Res. Part A Policy Pract.*, vol. 124, pp. 549–556, 2019, doi: 10.1016/j.tra.2018.04.009.
[24] C. Patel, "A Survey of Data-Driven Customer Segmentation Methods for Targeted Marketing Campaigns," *ESP J. Eng. Technol. Adv.*, vol. 3, no. 3, pp. 154–162, 2023, doi: 10.56472/25832646/JETA-V3I7P119.
[25] R. Q. Majumder, "Machine Learning for Predictive Analytics: Trends and Future Directions," *Int. J. Innov. Sci. Res. Technol.*, pp. 3557–3564, May 2025, doi: 10.38124/ijisrt/25apr1899.
[26] R. Q. Majumder, "Assessing the Impact of Audit Committees on Financial Data Reporting Quality and Corporate Accountability," *Int. J. Adv. Res. Sci. Commun. Technol.*, pp. 33–41, 2025, doi: 10.48175/ijarsct-28504.
[27] H. P. Kapadia and K. C. Chittoor, "AI Chatbots for Financial Customer Service: Challenges & Solutions," *J. Adv. Futur. Res.*, vol. 2, no. 2, pp. 1–7, 2024.
[28] R. Palwe, "Adaptive human: AI decision support for high-stakes financial advice," *Int. J. Comput. Artif. Intell.*, vol. 6, no. 2, pp. 385–392, Jul. 2025, doi: 10.33545/27076571.2025.v6.i2e.226.
[29] S. B. Karri, S. Gawali, S. Rayankula, and P. Vankadara, "AI Chatbots in Banking: Transforming Customer Service and Operational Efficiency," in *Advancements in Smart Innovations, Intelligent Systems, and Technologies*, 2025, pp. 61–81. doi: 10.3233/FAIA251498.
[30] R. Agarwal, D. Kalsi, P. Jain, P. Gupta, and R. Goel, "Car Insurance Fraud Detection using Machine Learning Models," in *2025 International Conference on Next Generation Information System Engineering (NGISE)*, 2025, pp. 1–8. doi: 10.1109/NGISE64126.2025.11085234.
[31] S. R. Raja, R. R. Cholla, R. K. Kadu, N. Legapriyadharshini, G. V. Jagatap, and S. Jothilakshmi, "Macroeconomic Modeling for Insurance Applications using the ANN-SVM Method," in *2025 International Conference on Intelligent Computing and Knowledge Extraction (ICICKE)*, 2025, pp. 1–6. doi: 10.1109/ICICKE65317.2025.11136562.
[32] L. Nyström and O. Witt, "Predicting Vehicle Insurance Premiums Using Linear Regression, XGBoost, and Neural Networks A Comparative Study of Predictive Power," 2025.
[33] M. Sun, "Predictive Analysis of Vehicle Insurance Demand Using Machine Learning Techniques," in *Proceedings of the 2024 5th International Conference on Computer Science and Management Technology*, Association for Computing Machinery, 2025, pp. 1193–1197. doi: 10.1145/3708036.3708233.
[34] D. Saikia, R. Barua, M. K. Gourisaria, A. Bandyopadhyay, S. R. Mishra, and S. Bilgaiyan, "Machine Learning Enhancements for Car Insurance Claim Prediction," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1–6. doi: 10.1109/ICCCNT61001.2024.10724028.
[35] R. Ibraimoh, "Using Artificial Intelligence to Improve Insurance Claim Evaluation," *ICONIC Res. Eng. JOURNALS*, vol. 8, no. 2, pp. 749–759, 2024.
[36] B. Cao, C. Li, Y. Song, Y. Qin, and C. Chen, "Network Intrusion Detection Model Based on CNN and GRU," *Appl. Sci.*, 2022, doi: 10.3390/app12094184.
[37] D. Y. Mohammed, "Detection of Vehicle Insurance Claim Fraud: A Fraud Detection Use-Case for the Vehicle Insurance Industry," *Int. J. Progress. Sci.*, vol. 30, no. 1, pp. 504–507, 2021.
[38] M. Hanafy and R. Ming, "Machine Learning Approaches for Auto Insurance Big Data," pp. 1–23, 2021, doi: 10.3390/risks9020042.
[39] G. Mahiyudin, M. Hussain, and D. D. Dewi, "A Comprehensive Study on Predicting the Need for Vehicle Maintenance Using Machine Learning," *Eng. Proc.*, vol. 107, no. 1, 2025, doi: 10.3390/engproc2025107089.
[40] C. Mare, D. Manaţe, G.-M. Mureşan, S. L. Dragoş, C. M. Dragoş, and A.-A. Purcel, "Machine Learning Models for Predicting Romanian Farmers' Purchase of Crop Insurance," *Mathematics*, vol. 10, no. 19, 2022, doi: 10.3390/math10193625.