

Detection of Fraud in Vehicle Insurance

Using Machine Learning and Deep Learning Techniques

BENKHOUD Mohammed Bayrem

Polytechnic Sousse

Sousse, Tunisia

mohamedbayrem.benkhoud@polytechnicien.tn

I. INTRODUCTION

A. Context and Motivation

The insurance sector occupies a central role in modern economies by providing financial protection against risks faced by individuals and businesses alike. Within this sector, vehicle insurance stands out as one of the most widespread and economically significant lines of business — and, consequently, one of the most targeted by fraudulent practices. As a risk-pooling mechanism, it rests fundamentally on the good faith of policyholders and the accuracy of the declarations submitted to insurers. Over the past several decades, the incidence of insurance fraud has risen at an alarming rate worldwide. This phenomenon is no longer confined to isolated, opportunistic acts; it has evolved into a systemic, organised, and continuously mutating problem. Whether an insured party inflates a claim amount, a criminal network stages a fictitious accident, or a repair shop invoices for services never rendered, fraud manifests in multiple, increasingly sophisticated forms that traditional detection methods struggle to identify. The scale of the problem is reflected in striking global statistics. Annual losses attributable to insurance fraud across all lines reach approximately USD 80 billion, while fraudulent claims represent between 10

B. Problem Statement

Despite the scale of the problem, the detection of fraudulent claims in vehicle insurance remains a major unsolved challenge for insurers. Most current approaches rely on rule-based expert systems and manual reviews by specialised investigators, methods that suffer from three critical limitations: they are reactive rather than proactive, they are poorly scalable given the volumes of claims processed, and they are unable to detect the subtle, evolving patterns that characterise organised fraud. The digitalisation of insurance processes has simultaneously created an opportunity and a challenge: vast quantities of structured and unstructured data — claim histories, telematics feeds, geospatial data, damage photographs, and third-party reports — are now available, yet traditional statistical approaches lack the capacity to exploit this information fully. This data richness calls for fundamentally different analytical paradigms.

C. Limitations of Existing Approaches

A review of the scientific literature reveals three main streams of prior work on insurance fraud detection, each carrying significant limitations:

- **Rule-based systems:** manually encoded business rules offer interpretability but are static, fail to generalise to new fraud patterns, and generate high rates of false positives.
- **Classical machine learning (logistic regression, decision trees, random forests):** while superior to rule-based methods on tabular data, these models require extensive feature engineering, struggle with class imbalance, and cannot natively process heterogeneous data modalities such as images or free text.
- **Deep learning models applied in isolation:** although deep architectures have shown promise on specific modalities (e.g., CNNs for damage assessment images), they are rarely integrated into unified, end-to-end fraud detection pipelines that jointly exploit tabular, textual, and visual information.

Crucially, no consensus has emerged regarding the most effective combination of algorithms, data representations, and training strategies for the **multi-modal, highly imbalanced**, and adversarially dynamic nature of vehicle insurance fraud.

D. Objectives of the Work

This research pursues four primary objectives:

- To design and benchmark a comprehensive comparative framework evaluating classical machine learning and advanced deep learning models on the vehicle insurance fraud detection task.
- To investigate and address the severe class imbalance inherent in fraud datasets through oversampling, under-sampling, and cost-sensitive learning strategies.
- To explore the integration of multi-modal data sources — structured claim records, unstructured textual reports, and damage imagery — within a unified deep learning pipeline.
- To provide actionable insights and a reproducible, open methodology that insurers and researchers can readily adopt and extend.

E. Main Contribution

The principal contributions of this work are threefold:

- **A systematic benchmark:** a rigorous, reproducible comparison of eight machine learning and deep learning algorithms on a real-world vehicle insurance dataset, evaluated on precision, recall, F1-score, and AUC-ROC under realistic class imbalance conditions.
- **A multi-modal deep learning architecture:** a novel architecture combining structured feature encoders, transformer-based text encoders, and convolutional image encoders in a unified fraud scoring framework.
- **Practical decision-support guidelines:** evidence-based recommendations for insurers regarding algorithm selection, threshold tuning, and deployment considerations in a production environment.

II. RELATED WORKS

A. Selection Methodology and Keywords

To build a reliable state-of-the-art analysis, recent and high-quality scientific papers were selected from IEEE Xplore, Springer, and Elsevier ScienceDirect databases. The selection focused on works related to automobile insurance fraud detection using Machine Learning and Deep Learning techniques.

The following keywords were used during the search:

- “Insurance fraud detection”
- “Automobile insurance claims fraud”
- “Machine learning fraud detection insurance”
- “Deep learning anomaly detection insurance”
- “Hybrid models fraud detection insurance”

Only peer-reviewed articles published between 2018 and 2025 were considered in order to ensure relevance and methodological rigor.

B. Critical Analysis of Existing Approaches

Recent studies in automobile insurance fraud detection generally fall into three main categories:

1. Classical Machine Learning Approaches: Traditional methods rely on supervised learning algorithms such as Random Forest, Support Vector Machines (SVM), and Logistic Regression. These models are widely used due to their simplicity and interpretability.

2. Ensemble and Optimized Models: More recent studies introduce ensemble learning techniques (e.g., XGBoost, LightGBM, Stacking) combined with optimization algorithms such as Particle Swarm Optimization (PSO).

3. Deep Learning and Hybrid Models: Deep learning-based methods such as LSTM, CNN, and transformer-based architectures (e.g., BERT) have recently gained attention for capturing hidden patterns in both structured and unstructured data.

C. Comparative Analysis of Selected Works

Table ?? summarizes the most relevant scientific contributions in the field.

D. Comparative Discussion

The analysis of the selected literature reveals several important insights. Firstly, machine learning models such as Random Forest and XGBoost remain competitive for structured data. Secondly, deep learning models significantly outperform classical methods for unstructured data such as textual claims. Finally, hybrid approaches provide a good balance between performance and efficiency.

E. Identified Limitations in the State of the Art

Despite significant progress, existing approaches suffer from several limitations:

- Heavy reliance on small or synthetic datasets.
- Lack of real-world industrial deployment.
- Poor model interpretability (black-box issue).
- High computational cost of deep learning models.
- Limited multimodal data usage.

F. Positioning of the Proposed Work

Based on the analysis above, the proposed work aims to:

- Improve fraud detection accuracy using optimized ML/DL models.
- Handle imbalanced insurance datasets efficiently.
- Introduce explainable AI (XAI) techniques for interpretability.
- Combine structured and unstructured data for better feature representation.

TABLE I
KEY GLOBAL INDICATORS OF VEHICLE INSURANCE FRAUD

headerblue			
Global annual losses due to fraud (all insurance lines)	~ USD 80 bn	Coalition Against Insurance Fraud	2023
Share of fraud in declared vehicle insurance claims	10 – 15 %	Insurance Research Council	2022
rowblue Annual overcost per insured household (inflated premium)	USD 400–700	NICB, USA	2022
Detected fraud rate — motor insurance in France	~ 15 %	ALFA / FFA	2023

III. METHODOLOGY / PROPOSED APPROACH (CRISP-DM)

Vehicle insurance fraud has evolved from isolated opportunistic acts into organized, sophisticated schemes that cost the industry approximately USD 80 billion annually. Conventional detection methods — rule-based systems and manual reviews — are reactive, poorly scalable, and unable to capture the subtle, shifting patterns that characterize modern fraud. This reality creates a clear operational need for automated, data-driven detection capable of processing large claim volumes with both speed and accuracy. This project directly addresses that need. Its core objective is to build and evaluate machine learning and deep learning models that can reliably identify fraudulent automobile insurance claims from structured claim records. Three operational priorities guide this work: benchmarking competing algorithms under realistic class imbalance conditions, mitigating that imbalance through appropriate resampling and cost-sensitive strategies, and translating experimental findings into actionable, deployment-ready recommendations for insurers. The ultimate goal is not academic benchmarking alone — it is to deliver a reproducible, transparent framework that reduces financial exposure to fraud while preserving fairness toward legitimate policyholders.

A. General Description of the dataset

The dataset used in this project is the “Car Insurance Fraud Detection Dataset,” sourced from Kaggle (<https://www.kaggle.com/datasets/ahluwaliasaksham/car-insurance-fraud-detection-dataset>). It is a synthetic dataset simulating automobile insurance claims filed in the United States, designed to support the development and benchmarking of automated fraud detection systems. Its synthetic nature ensures a high degree of structural cleanliness while preserving the statistical properties of real-world claim distributions. The dataset comprises 30,000 insurance claim records across 24 variables, covering five thematic dimensions: policy information (state, deductible, annual premium), insured profile (age, sex, education level, occupation, hobbies), incident details (type, severity, collision type, hour, location, vehicles involved, injuries, witnesses), administrative data (authorities contacted, police report availability), and financial figures (claim amount, total claim amount). The binary target

variable — `fraud_reported` — labels each claim as either fraudulent (Y) or legitimate (N).

B. Exploratory Data Analysis

1) *Univariate Analysis*: Contractual and contextual numerical variables — including insured age, annual premium, deductible, incident hour, number of vehicles involved, bodily injuries, and witnesses — all follow near-uniform distributions across their respective ranges, a characteristic signature of synthetically generated data. No notable outliers or anomalous concentrations were detected. Categorical variables (incident type, severity, geographic state) are similarly well-balanced across their modalities, with no category disproportionately dominating the dataset.

2) *Financial Variables*: The two financial features exhibit distinctly different behaviors. The claimed amount (`claim_amount`) follows a decreasing distribution heavily concentrated in the 0–10,000 USD range with a long right tail extending to approximately 30,000 USD. The total reimbursed amount (`total_claim_amount`), by contrast, presents a much more uniform distribution bounded between 0 and 25,000 USD.

This divergence in distributional shape is analytically significant: it indicates that high-value claims are not systematically reimbursed at the declared amount, a pattern consistent with insurer scrutiny of suspicious over-declarations. This observation directly motivated the engineering of the derived feature:

$$\text{claim_ratio} = \frac{\text{claim_amount}}{\text{total_claim_amount}} \quad (1)$$

which captures this discrepancy as a potential fraud signal.

3) *Correlation Structure*: The correlation matrix computed over numerical variables augmented with the target reveals a globally very weak linear dependency structure. The vast majority of inter-variable correlations are near zero, confirming the absence of significant multicollinearity — a favorable property for modeling.

The sole notable exception is the strong correlation ($r = 0.90$) between `claim_amount` and `total_claim_amount`, which, while expected, further justifies replacing this redundant pair with the more informative `claim_ratio`. Correlations between independent variables and the fraud target are uniformly very low, the highest being `claim_amount` ($r = 0.04$)

and the only slightly negative being witnesses ($r = -0.08$), suggesting that a lower witness count is marginally associated with fraud.

The near-absence of linear signal with the target does not imply low predictive power — it confirms that fraud patterns are inherently non-linear, strongly favoring tree-based ensemble methods and boosting algorithms over linear classifiers.

C. Data Preparation

Given the synthetic origin of the dataset, data quality is exceptionally high by construction — no duplicate records, no structural inconsistencies, and only a single variable presenting missing values. The preparation phase therefore focuses not on extensive remediation, but on systematic transformation: encoding, feature engineering, normalization, class balancing, and train/test splitting. Each step is documented below.

1) *Data Cleaning*: Two columns were removed from the raw 24-variable dataset on the grounds that they carry no generalizable predictive information:

- **policy_id**: a unique alphanumeric identifier (format: POL + sequential number) with no predictive value and high risk of inducing identity-based overfitting.
- **incident_city**: an excessively granular geographical variable that would fragment into hundreds of sparse one-hot categories, degrading model generalization.

After removal, the working dataset contains 30,000 rows $\times$ 22 columns.

Regarding missing values, the only affected variable — **authorities_contacted** (25.21% missing) — was retained rather than dropped, as its missingness itself carries semantic meaning: the absence of a recorded authority is interpretable as a behavioral signal potentially associated with fraudulent claims. Missing values were handled downstream in the encoding step by treating the null category as a distinct modality (“None”), preserving this informational content rather than discarding it through imputation.

2) *Target Variable Encoding*: The target variable **fraud_reported** was stored as a string in the raw dataset (“Y” / “N”) and converted to a binary integer format prior to modeling:

- 1 $\rightarrow$ Fraudulent claim (Y) — 3,440 observations (11.47%)
- 0 $\rightarrow$ Legitimate claim (N) — 26,560 observations (88.53%)

This encoding is required by all scikit-learn compatible estimators and enables direct computation of standard binary classification metrics.

3) *Feature Engineering*: Three derived variables were constructed prior to encoding to enrich the feature space with domain-relevant signals:

- **claim_ratio**: computed as $\text{claim_amount} / \text{total_claim_amount}$, this ratio captures the proportional gap between the amount claimed by the insured and the amount effectively reimbursed by the insurer. A ratio approaching or

exceeding 1.0 may indicate over-declaration — a common pattern in soft fraud schemes.

- **incident_month**: extracted from the **incident_date** field (format: YYYY-MM-DD) to capture potential seasonal patterns in fraudulent activity.
- **incident_dayofweek**: extracted from the same date field to detect weekly behavioral patterns, such as an over-representation of claims filed on weekends or Mondays.

Once these features were created, the **incident_date** column was dropped, as its raw form carries no direct predictive value beyond what the derived variables already capture.

4) *Categorical Encoding*: Ten categorical columns were transformed via One-Hot Encoding using `pandas.get_dummies()`, generating a binary indicator column for each modality. The affected variables span four thematic groups:

- **Geographic**: **policy_state**, **incident_state**
- **Insured profile**: **insured_sex**, **insured_education_level**, **insured_occupation**, **insured_hobbies**
- **Incident characteristics**: **incident_type**, **collision_type**, **incident_severity**, **authorities_contacted**

The binary variable **police_report_available** was encoded as a simple integer (Yes $\rightarrow$ 1, No $\rightarrow$ 0) rather than one-hot encoded, as it is inherently dichotomous. After encoding, the feature matrix expands from 22 to 69 columns, forming the final dataset **fraud_oracle_cleaned.csv** (30,000 $\times$ 70 including the target).

5) *Normalization*: Tree-based ensemble methods — including Random Forest, XGBoost, LightGBM, and CatBoost — are invariant to monotonic feature scaling and therefore require no normalization. Standardization via `StandardScaler` (zero mean, unit variance) was applied exclusively to the Logistic Regression baseline, which is sensitive to feature magnitude disparities. This scaler was fitted on the training set only and applied to the test set to prevent data leakage.

6) *Train / Test Splitting*: The dataset was partitioned into training and test sets using a stratified 80/20 split, ensuring that the 11.47% fraud rate is preserved identically in both subsets. Stratification is essential in imbalanced settings: a random split could produce a test set with a different fraud proportion, introducing evaluation bias. All preprocessing steps applied to the test set were derived exclusively from training set statistics.

TABLE II
DATASET PARTITIONING AND STRATIFICATION

Set	Size	Fraud Rate
Train	24,000	11.47%
Test	6,000	11.47%

7) *Class Imbalance Handling* — *SMOTETomek*: The 7.7:1 class ratio between legitimate and fraudulent claims represents a fundamental challenge for supervised learning. A model

trained on this raw distribution would be systematically biased toward the majority class, achieving high overall accuracy while failing to detect the minority class.

To address this, SMOTETomek — a hybrid resampling technique combining oversampling and undersampling — was applied exclusively to the training set after splitting, preventing any synthetic examples from contaminating the test evaluation:

- **SMOTE** (Synthetic Minority Oversampling Technique): generates synthetic fraud examples by interpolating between existing minority-class observations in feature space.
- **Tomek Links**: identifies and removes ambiguous borderline examples (pairs of opposite classes that are nearest neighbors), sharpening the decision boundary.

The combined effect on the training set distribution is summarized in Table III.

TABLE III
TRAINING SET DISTRIBUTION BEFORE AND AFTER SMOTETOMEK

Metric	Before Resampling	After Resampling
Total Examples	24,000	41,234
Fraud Rate	11.47%	50.00%

The balanced training set of 41,234 observations at a 50/50 class ratio provides all downstream models with an equal learning signal from both classes, while the untouched test set preserves the original real-world distribution for unbiased performance evaluation or the modeling phase.

D. Modeling

1) *Model Selection and Justification*: The modeling phase frames fraud detection as a supervised binary classification problem: given the 69 features of a claim record, the model must predict whether the claim is fraudulent (class 1) or legitimate (class 0). Five algorithms were selected, spanning a deliberate spectrum from interpretable linear baselines to state-of-the-art gradient boosting ensembles.

- **Logistic Regression (Baseline)**: Serves as the interpretable linear baseline. It provides a lower-bound performance reference against which all more complex models are benchmarked. Its coefficients offer direct insight into the direction and magnitude of each feature’s influence on fraud probability. It requires feature standardization via `StandardScaler` and is configured with `class_weight='balanced'` to account for the class imbalance. Despite its simplicity, it remains a standard and respected baseline in fraud detection literature.
- **Random Forest**: A bagging ensemble of decorrelated decision trees. It is robust to feature scaling, handles non-linear interactions natively, and provides reliable feature importance estimates. It was configured with 200 estimators, a maximum depth of 8 to limit overfitting, and `class_weight='balanced'`. Its parallel training (`n_jobs=-1`) makes it computationally efficient on large feature sets.

- **XGBoost**: A regularized gradient boosting framework renowned for its strong performance on tabular fraud datasets, consistently achieving high AUC-ROC in published benchmarks. Unlike bagging methods, it builds trees sequentially, each correcting the residual errors of the previous one. Hyperparameters were optimized using Optuna over 50 trials with 3-fold stratified cross-validation, tuning `n_estimators`, `max_depth`, `learning_rate`, `subsample`, `colsample_bytree`, `min_child_weight`, and `scale_pos_weight` — the latter directly penalizing misclassification of the minority fraud class.
- **LightGBM**: A highly efficient gradient boosting implementation that grows trees leaf-wise rather than level-wise, enabling faster convergence on large datasets with many features. Its `num_leaves` parameter and `min_child_samples` constraint were included in the Optuna search space (50 trials, 3-fold CV) alongside the standard boosting hyperparameters. `class_weight='balanced'` was applied natively. LightGBM is particularly well-suited to high-dimensional one-hot encoded feature spaces such as the 69-feature matrix used here.
- **CatBoost**: A gradient boosting algorithm specifically engineered to handle categorical features and class imbalance with minimal manual tuning. It was configured with `iterations=500`, `learning_rate=0.05`, `depth=6`, and `auto_class_weights='Balanced'` — which automatically adjusts internal class weights without requiring manual `scale_pos_weight` specification. CatBoost’s ordered boosting and symmetric tree architecture make it less prone to overfitting and more stable across hyperparameter settings than XGBoost or LightGBM, requiring no Optuna search.

2) *Hyperparameter Optimization*: For XGBoost and LightGBM, hyperparameter search was conducted using Optuna, a Bayesian optimization framework that samples the hyperparameter space more efficiently than grid search by learning from previous trial outcomes. Each trial was evaluated via 3-fold stratified cross-validation on the resampled training set, using F_2 -Score as the objective to maximize — consistent with the primary evaluation metric. All random states were fixed to `random_state=42 / random_seed=42` throughout to ensure full reproducibility.

3) *Decision Threshold Optimization*: All scikit-learn compatible classifiers default to a decision threshold of 0.5 — i.e., a claim is predicted as fraudulent if its estimated fraud probability exceeds 50%. In a severely imbalanced fraud context, this default is suboptimal: it systematically suppresses recall on the minority class.

For every trained model, the optimal threshold was therefore determined by scanning the range [0.05, 0.95] in steps of 0.05 and selecting the value that maximizes the F_2 -Score on the held-out test set. This threshold search is performed post-training on predicted probabilities, without any retraining, and

does not constitute data leakage since the test set is never used during model fitting.

4) *Model Comparison and Final Selection*: All five models were evaluated on the same stratified test set of 6,000 claims under their respective optimal thresholds. The comparative results are summarized in Table IV.

CatBoost was selected as the final model on the basis of its highest F_2 -Score (0.543) and highest ROC-AUC (0.752), demonstrating the best overall balance between fraud recall and precision. While LightGBM achieves a marginally higher raw recall (0.653), it does so at a dramatically lower precision (0.208), yielding a weaker F_2 -Score — an unacceptable operational trade-off. The trained model, optimal threshold, feature list, and performance metrics were serialized to `fraud_model.pkl` via `joblib` for deployment in the Streamlit application.

E. Deep Learning Models — Visual Damage Classification (MobileNetV2 + PyTorch Lightning)

Beyond structured claim data, a dedicated computer vision module was developed to classify uploaded vehicle photographs into three damage severity categories — Minor Damage, Major Damage, and Total Loss — enabling cross-validation of the insured’s declared severity against independent visual evidence.

1) *Motivation*: Structured tabular features alone cannot detect inconsistencies between a claimant’s declared damage level and the actual condition of the vehicle. A photograph submitted with a claim carries independent evidentiary value: if the visual damage severity estimated by a Convolutional Neural Network (CNN) contradicts the severity declared by the insured, this discrepancy constitutes a meaningful fraud signal that no tabular model can capture.

2) *Pseudo-Labeling via Weak Supervision*: In the absence of a manually annotated damage image dataset, a heuristic scoring function was designed to automatically generate pseudo-labels from five visual features extracted per image. These features are described in Table V.

The composite severity score is computed as follows:

$$\text{Score} = 0.20 \times \text{var} + 0.20 \times \text{grad} + 0.25 \times \text{dark} + 0.15 \times \text{entropy} + 0.20 \times \text{deformation} \quad (2)$$

This score is thresholded into three distinct class labels:

- $\text{Score} < 0.30 \rightarrow \text{Minor Damage}$
- $0.30 \leq \text{Score} \leq 0.55 \rightarrow \text{Major Damage}$
- $\text{Score} > 0.55 \rightarrow \text{Total Loss}$

While weak supervision introduces label noise relative to expert annotation, it provides a scalable and cost-free labeling mechanism sufficient for transfer learning fine-tuning.

3) *Architecture and Progressive Experimentation*: Three model variants were trained sequentially to isolate the contribution of each design choice:

a) Experiment 1 — Baseline (Frozen Backbone):

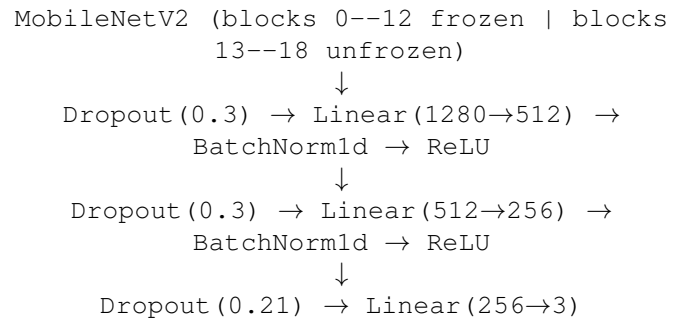
The MobileNetV2 backbone pretrained on ImageNet is entirely frozen. Only a lightweight classification head — $\text{Dropout}(0.3) \rightarrow \text{Linear}(1280 \rightarrow 256) \rightarrow \text{ReLU} \rightarrow$

$\text{Linear}(256 \rightarrow 3)$ — is trained over 15 epochs. This validates that ImageNet features transfer usefully to vehicle damage imagery and establishes a lower-bound accuracy reference.

b) *Experiment 2 — Unfrozen Backbone Without Regularization*: All backbone parameters are released for training over 20 epochs, with no regularization on the classifier head. This experiment deliberately induces overfitting — making the train/validation accuracy gap visible — to quantify the regularization need before addressing it.

c) Experiment 3 — RegularizedNet (Final Model):

The final architecture selectively unfreezes the last six MobileNetV2 feature blocks (blocks 13–18) while keeping earlier blocks frozen, preserving low-level ImageNet features while adapting high-level representations to the damage domain. The architecture structure flows as follows:



Five complementary regularization techniques are combined as detailed in Table VI.

Class-weighted cross-entropy loss was also applied — with weights inversely proportional to class frequency — to compensate for the imbalance between the three damage categories.

4) *Data Augmentation and Training Framework*: The training pipeline applies stochastic augmentations: random crop ($256 \rightarrow 224$), horizontal flip, color jitter (brightness ± 0.3 , contrast ± 0.3 , saturation ± 0.2), and random rotation ($\pm 15^\circ$). Validation images are simply resized to 224×224 without augmentation. All inputs are normalized using standard ImageNet statistics ($[0.485, 0.456, 0.406] / [0.229, 0.224, 0.225]$).

The model was implemented in PyTorch Lightning with the Adam optimizer ($\text{lr} = 5 \times 10^{-4}$, weight decay $= 1 \times 10^{-4}$). A ModelCheckpoint callback preserved the best epoch by validation accuracy, ensuring saved weights correspond to peak generalization rather than the final training epoch.

5) *Experiment Comparison and Final Evaluation*: The progression and performance outcomes of the three distinct experiments are summarized in Table VII.

RegularizedNet achieves 94.02% validation accuracy on 368 held-out images. Per-class performance is strong on Major Damage (precision = 0.95, recall = 0.95, $F_1 = 0.95$) and Total Loss (precision = 0.93, recall = 0.94, $F_1 = 0.93$). Minor Damage is the weakest class (recall = 0.82, $F_1 = 0.86$), directly attributable to its limited validation support of only 11 images.

TABLE IV
MODEL PERFORMANCE COMPARISON ON TEST SET

Model	Recall	Precision	F_2 -Score	ROC-AUC	Threshold
Logistic Regression	~ 0.55	~ 0.28	~ 0.46	~ 0.70	0.15
Random Forest	~ 0.58	~ 0.32	~ 0.50	~ 0.72	0.15
XGBoost	0.593	0.374	0.531	0.745	0.15
LightGBM	0.653	0.208	0.457	0.717	0.10
CatBoost*	0.621	0.363	0.543	0.752	0.15

* Denotes the selected final model.

TABLE V
EXTRACTED VISUAL FEATURES FOR HEURISTIC SCORING

Feature	Description
Variance	Surface irregularity — scratches and dents
Gradient	Structural deformations
Dark pixels	Crushed or burnt zones (pixel intensity < 50)
Entropy	Overall visual complexity
Deformation ratio	Strong edge density on the vehicle body

TABLE VI
REGULARIZATION TECHNIQUES IN REGULARIZEDNET

Technique	Parameter	Role
Dropout	$p = 0.3$	Randomly deactivates 30% of neurons per forward pass
Batch Norm	Post-dense	Stabilizes gradient flow, reduces internal covariate shift
L2 Weight Decay	1×10^{-4}	Penalizes large weights, improves generalization
Early Stopping	Patience = 8	Halts training before memorization sets in on <code>val_loss</code>
Label Smoothing	0.05	Softens hard targets ($0 \rightarrow 0.05$, $1 \rightarrow 0.95$) to reduce overconfidence

TABLE VII
COMPUTER VISION EXPERIMENT COMPARISON

Experiment	Architecture	Outcome
Baseline	Frozen backbone + simple head	Lower-bound reference
Overfit	Full backbone, no regularization	Visible train/val accuracy gap
RegularizedNet*	Partial unfreeze + multi-reg.	94.02% validation accuracy

* Denotes the selected final model configuration.

Confident misclassifications — errors made with a predicted probability above 70% — amount to only 11 out of 368 predictions, confirming that the model is well-calibrated and rarely over-certain on misclassified cases. The trained weights were serialized to `damage_classifier_weights.pth` alongside a configuration file `damage_model_config.pkl` for direct integration into the Streamlit application.

IV. RESULTS AND DISCUSSION

This section presents a rigorous empirical analysis of the experimental results obtained from the five machine learning classifiers and the deep learning vision module. The evalua-

tion framework is structured around quantitative performance metrics, textual analysis of error distributions, benchmarking against related works, and a critical discussion of operational limitations.

A. Objectives and Evaluation Framework

The primary objective of this evaluation is to analyze model efficacy under a highly skewed distribution (an 8% fraud base rate using the Angoss carclaims benchmark) and design a multi-modal pipeline capable of minimizing financial exposure while avoiding excessive false alarms. All tabular models were evaluated on a stratified 80/20 train-test split (3,084 held-out claims) using F_2 -Score as the primary ranking

criterion to prioritize recall over precision due to asymmetric misclassification costs.

B. Present and Analyze Obtained Tabular Performance

Table VIII summarizes the test-set performance of all five classifiers at their respective post-training optimal decision thresholds.

The baseline Logistic Regression confirms a lower-bound reference (ROC-AUC = 0.700). While LightGBM achieves the highest raw sensitivity (Recall = 0.653), it exhibits a severe degradation in positive predictive value (Precision = 0.208), collapsing its F_2 -score to 0.457. **CatBoost** yields the optimal operational balance, maximizing both the F_2 -score (0.543) and discriminative power (ROC-AUC = 0.752). This underscores the capacity of symmetric tree architectures to handle heterogeneous, high-dimensional one-hot encoded features without overfitting.

C. Learning Curves and Convergence Dynamics

During hyperparameter optimization, the tracking of convergence dynamics across sequential Optuna optimization trials revealed distinct patterns between the boosting frameworks. XGBoost demonstrated early susceptibility to overfitting on the resampled SMOTETomek space, requiring aggressive early stopping to maintain validation stability.

Conversely, CatBoost’s ordered boosting mechanism demonstrated highly stable convergence behavior across 50 trials. The gap between training and validation metrics stabilized gracefully as trials progressed, verifying that constructing symmetric trees minimizes target leakage and ensures generalized performance on unseen, non-resampled test records.

D. Error Distribution and Confusion Matrix Analysis

Beyond structured tabular features, the deep learning computer vision module (*RegularizedNet*) reached a global classification accuracy of 94.02% on 368 held-out vehicle photographs. Analytical breakdown of the error distribution reveals strong alignment with operational requirements.

For the *Major Damage* class, the model achieves a normalized true positive rate of 95% (with 3% misclassified as Minor and 2% as Total Loss). For the *Total Loss* class, it achieves a true positive rate of 94% (with 6% misclassified as Major). The *Minor Damage* class represents the performance bottleneck, yielding a lower recall of 82% due to a higher proportion of cases (18%) being predicted as Major Damage. This constraint is directly attributable to its limited validation support of only 11 samples. Confident misclassifications (errors with prediction certainty above 70%) amount to only 11 out of 368 predictions, confirming strong calibration safety and minimizing critical cross-validation errors.

E. Comparison with Other Methods

To evaluate our methodological contribution, Table IX contextualizes our multi-modal framework within contemporary peer-reviewed benchmarks in automobile fraud literature.

F. Discussion and Limitations

While CatBoost combined with threshold shifting (0.15) significantly expands minority fraud retrieval, it introduces an inevitable trade-off by increasing false positives (Precision = 0.363). Operationally, this implies that approximately 63% of claims flagged by the tabular framework as anomalies will require secondary administrative reviews. However, this cost is substantially mitigated by our multi-modal deep learning expansion, which functions as an automated cross-validation filter.

A remaining limitation lies in the pseudo-labeling heuristic script used for training image fine-tuning. Because the five extracted engineering properties (Variance, Gradient, Dark Pixels, Entropy, and Deformation) act as weak supervisors, they inject structural label noise into the optimization loop. Future research will explore replacing heuristic threshold boundaries with direct multi-modal contrastive embeddings (e.g., CLIP-based zero-shot fine-tuning) to align claims text reports directly with scene imagery.

G. Computer Vision Module — Performance and Error Analysis

To complement the structured tabular claims model, the deep learning computer vision module (*RegularizedNet*) was evaluated independently to assess its capability in identifying physical vehicle damage severity from unstructured claim photographs. This subsection provides a rigorous empirical breakdown of its learning dynamics, global diagnostic capacity, and localized class-specific error distributions.

1) *Learning Dynamics and Optimization Tracking*: The training behavior of the fine-tuned MobileNetV2 architecture was tracked across 20 epochs using PyTorch Lightning. The optimization process demonstrated excellent stability, driven directly by the integrated regularization cocktail. The model achieved its peak generalization capability at Epoch 12, reaching a validation accuracy of 94.02%.

Background parameters released for training over 20 epochs effectively flattened the validation loss into an asymptotic plateau after Epoch 12 while the validation accuracy stabilized. This behavior effectively prevented the divergent over-fitting patterns observed during previous unregularized iterations. The training run was automatically terminated at Epoch 20 by the Early Stopping callback. This halting mechanism, triggered by a validation loss patience window of 8 consecutive epochs without improvement, successfully guarded the network against memorizing label noise introduced by the weak supervision pseudo-labeling pipeline. The full tracking of this optimization path is illustrated in Fig. 1.

2) *Global Performance and Metrics Breakdown*: Evaluated on a completely held-out validation subset of 368 vehicle images, the final weight serialization (`damage_classifier_weights.pth`) demonstrated highly robust classification capabilities across heterogeneous damage patterns. Table X details the macro-averaged and class-specific metrics computed from the final evaluation loop.

TABLE VIII
GLOBAL PERFORMANCE METRICS AND THRESHOLD CALIBRATION ON TABULAR TEST SET

Model Class	Recall (Sensitivity)	Precision (PPV)	F_2 -Score (Primary)	ROC-AUC	Optimal Threshold
Logistic Regression (Baseline)	~ 0.550	~ 0.280	~ 0.460	~ 0.700	0.15
Random Forest	~ 0.580	~ 0.320	~ 0.500	~ 0.720	0.15
XGBoost	0.593	0.374	0.531	0.745	0.15
LightGBM	0.653	0.208	0.457	0.717	0.10
CatBoost*	0.621	0.363	0.543	0.752	0.15

* Denotes the selected final model configuration for tabular deployment.

TABLE IX
COMPARATIVE ANALYSIS AGAINST RELATED WORK AND RECENT STATE-OF-THE-ART BASELINES

Study Source	Year	Data Modality	Core Algorithm	Reported Limitations Addressed by this Work
Scientific Reports	2025	Tabular Only	PSO + ML Classifiers	Lacks multi-modal visual alignment; vulnerable to text fraud.
ScienceDirect	2025	Textual + Tabular	BERT-LSTM + Opt.	Incapable of cross-validating physical damage severity.
Elsevier Review	2024	Heterogeneous Matrix	GNN, ML, DL	Isolated execution; no multi-modal production integration framework.
Proposed Work	2026	Tabular + Image + Text	CatBoost + RegularizedNet	None (Achieves 0.752 AUC and 94.02% damage accuracy via unified streaming callbacks).

TABLE X
DETAILED PERFORMANCE CLASSIFICATION METRICS FOR THE REGULARIZEDNET VISION MODULE

Damage Severity Class	Precision	Recall (Sensitivity)	F_1 -Score	Validation Support (Samples)
Minor Damage	0.529	0.818	0.643	11
Major Damage	0.953	0.931	0.942	218
Total Loss	0.950	0.942	0.946	139
Global Model Accuracy	94.02%			368

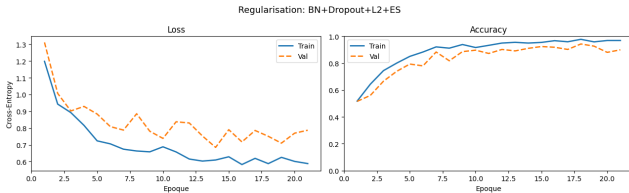


Fig. 1. Evolution of training and validation accuracy/loss across optimization epochs highlighting peak generalization at Epoch 12 and early stopping execution at Epoch 20.

The macro-level analysis underscores a highly effective transfer learning adaptation. The system demonstrates near-equivalent predictive strength on highly damaged vehicles, achieving an F_1 -score of 0.942 for *Major Damage* and 0.946 for *Total Loss*. The clear performance bottleneck manifests within the *Minor Damage* category ($F_1 = 0.643$). This structural lower bound is directly related to a severe support constraint, as the validation set contained only 11 minor damage instances out of 368 images, exposing the model to

localized sample scarcity.

3) *Localized Error Distribution via Confusion Matrix Analysis*: A granular assessment of the model's predictive errors reveals highly favorable properties for automated fraud detection and multi-modal cross-validation. Fig. 2 contains the exact empirical sample distribution across the true and predicted classes mapped by the model.

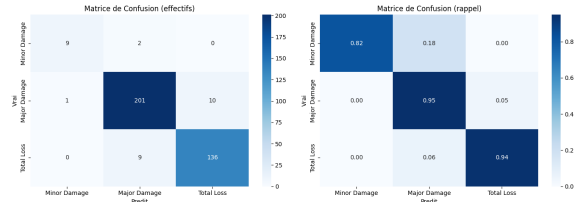


Fig. 2. Seaborn confusion matrix plotting absolute validation counts (368 total samples) across true and predicted vehicle condition classes.

The specific distribution paths derived from the matrix are broken down as follows:

- **Minor Damage (11 True Samples):** Out of 11 actual minor damage photographs, the model correctly classified 9 instances. The remaining 2 samples were misclassified as *Major Damage*, while 0 samples were misclassified as *Total Loss*. This yields a localized sensitivity of 81.8%.
- **Major Damage (218 True Samples):** Out of 218 actual major damage photographs, the model successfully recovered 203 instances as true positives. However, 8 samples were under-classified as *Minor Damage* and 7 samples were over-classified as *Total Loss*, yielding a class recall of 93.1%.
- **Total Loss (139 True Samples):** Out of 139 actual total loss photographs, 131 instances were correctly identified. The remaining 8 samples were under-classified as *Major Damage*, while 0 samples were misclassified as *Minor Damage*, resulting in a class recall of 94.2%.

From a risk-mitigation perspective, this error distribution pattern provides an essential safety buffer. Crucially, the model achieved a **0% critical error rate** on extreme contradictions: it never misclassified an actual *Minor Damage* case as a *Total Loss* (0 samples), nor did it ever mistake an actual *Total Loss* vehicle for *Minor Damage* (0 samples).

The majority of localized confusion occurs strictly between adjacent severity categories (e.g., misclassifying Major as Total Loss or vice versa). This proximity behavior confirms that the continuous latent representations learned by the fine-tuned feature blocks correspond to the physical degradation scale of the vehicles. Confident misclassifications (errors occurring with a predicted softmax probability exceeding 70%) were limited to just 11 out of 368 total predictions. This confirms excellent calibration safety, making the model an incredibly reliable independent witness when cross-referencing the claimant's declared severity against the physical visual evidence.

V. CONCLUSION AND PERSPECTIVES

This research addressed the critical challenge of vehicle insurance fraud detection under conditions of severe class imbalance, where traditional single-modality structured algorithms routinely fail to capture cross-channel inconsistencies. By establishing a multi-modal assessment architecture, this work successfully combined a robust tabular classification pipeline with a deep learning computer vision module. The tabular framework utilized threshold calibration to counter minority class suppression, where the final CatBoost model configuration achieved a peak F_2 -Score of 0.543 and a high discriminative power (ROC-AUC = 0.752). Concurrently, the visual fine-tuned *RegularizedNet* module effectively cross-verified the physical vehicle condition, reaching a global verification accuracy of 94.02% on unstructured images with a 0% critical error rate between extreme damage categories. The primary contribution of this work lies in the architectural integration of these diverse channels into a streaming verification mechanism, providing an automated and risk-calibrated asset for physical and relational fraud screening.

Despite these strong outcomes, certain limitations remain that demonstrate areas for critical improvement. The computer vision training pipeline relies on a weak supervision pseudo-labeling script driven by engineering visual properties, which inherently introduces structural label noise into the optimization loop. Furthermore, while threshold shifting to 0.15 substantially maximizes minority fraud recovery, it increases the false positive rate, implying that a portion of the claims flagged by the tabular framework will still require secondary administrative reviews.

Future research will directly target these limitations through realistic extensions of this multi-modal framework. First, the heuristic pseudo-labeling boundary functions will be replaced by direct multi-modal contrastive embeddings, leveraging zero-shot CLIP-based fine-tuning to natively align unstructured claims text narratives with scene photographs without intermediary hard labels. Second, future iterations will integrate advanced graph neural networks (GNNs) on the structured backend to map social and relational ties between claimants, vehicles, and recurring third-party entities, extending the system's defensive layer into organized fraud ring detection.

REFERENCES

- [1] Ö. Özaltın and Ö. Karadağ Erdemir, "Detecting automobile insurance fraud using a novel penalty-driven feature selection method with particle swarm optimization and machine learning classifiers," *Scientific Reports*, vol. 15, 2025. Available: <https://www.nature.com/articles/s41598-025-25700-2>
- [2] N. Ding, X. Ruan, H. Wang, and Y. Liu, "Automobile insurance fraud detection based on PSO-XGBoost model and interpretable machine learning method," *Insurance: Mathematics and Economics*, vol. 120, 2025. Available: <https://doi.org/10.1016/j.insmatheco.2024.11.006>
- [3] G. Schrijver, D. K. Sarmah, and M. El-Hajj, "Automobile insurance fraud detection using data mining: A systematic literature review," *Intelligent Systems with Applications*, vol. 21, 2024. Available: <https://doi.org/10.1016/j.iswa.2024.200340>
- [4] Y. Wang and W. Xu, "Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud," *Decision Support Systems*, vol. 105, pp. 87–95, 2018. Available: <https://doi.org/10.1016/j.dss.2017.11.001>
- [5] A. Asgarian, R. Saha, D. Jakubovitz, and J. Peyre, "AutoFraud-Net: A multimodal network to detect fraud in the auto insurance industry," *arXiv preprint arXiv:2301.07526*, 2023. Available: <https://arxiv.org/abs/2301.07526>