International Journal of Mathematics Trends and Technology
ISSN: 2231-5373/ https://doi.org/10.14445/22315373/IJMTT-V72I10P106

Volume 72 Issue 10, 63-81, October 2026
© 2026 Seventh Sense Research Group®

Original Article

A Dual-Source Claim Fusion Artificial Intelligence Model for Automated Insurance Premium Pricing and Claim Decision-Making

Basavaraj Talawar[1], A S Talawar[2]

[1,2]*Department of Statistics, Karnatak University, Dharwad, Karnataka, India.*

[1]*Corresponding Author : basavaraj.t@kud.ac.in*

Received: 23 August 2026 Revised: 29 September 2026 Accepted: 16 October 2026 Published: 28 October 2026

Abstract *- In this paper, Claim Fusion System Artificial Intelligence (CFS-AI) model integrates real-world insurance data and real-time claim application data. The model automatically evaluates the optimum premium pricing and claims decision-making. The system calculates risk potentially using machine learning algorithms and predictive analytics. The model enabling insurers to determine optimal premium amounts and approve claims efficiently. The model includes automated notifications and payment mechanisms. It also reduces the processing time and operational cost. Furthermore, the proposed model illustrates the experimental design and customer stratification techniques to increase the personalized risk assessment. The model maximizes the predictive accuracy through continuous-time stochastic modelling and optimization techniques. The CFS-AI model has fraud detection capability, ensures decision transparency and delivers faster insurance services.*

Keywords *- Premium Pricing, Claim Automation, Actuarial Model, Fusion System, Artificial Intelligence.*

1. Introduction

The insurance industry is undergoing a significant socio-technical transformation, moving away from traditional, manual insurance practices and towards dynamic, automated ones (Srivastava, 2025) [9]. The authors (Garg et al., 2025) [4] define historical financial tables and multi-level claim verification, which are the basis of traditional premium pricing and claim verification systems, which are time-consuming and inefficient in optimizing premiums. These structural challenges are addressed by Ghatak and Purkayastha (2025) [5],, where the "Claim Fusion Systems" fuse heterogeneous real-world insurance profiles with real-time digital application pipelines. The literature reviewed in this study highlights key areas of research in machine learning for predictive pricing and risk assessment, deep learning for Natural Language Processing (NLP) and computer vision in multimodal claim validation and continuous-time stochastic differential equations for modeling volatile market dynamics in recent years (Lavanya & Nithya, 2025 [8] and Agrawal et al., 2025 [1]). The author Agrawal et al. (2025) [1] presents increasing level of automation in insurance underwriting and automated insurance premium configuration, with predictive analytics being the backbone, using supervised machine learning algorithms. The focus has shifted from uniform, linear approaches to risk stratification to very localized, non-linear risk stratification techniques. The research published in recent years have repeatedly shown that ensemble-based methods, like Random Forest (RF) and XGBoost (Extreme Gradient Boosting) are much more effective than classic probabilistic methods (Gaussian Naïve Bayes (GNB) and Linear Discriminant Analysis (LDA)) in operational contexts. From Debene et al. (2026) [3], XGBoost has been a popular method for their automated extraction of multi-variable feature interactions and optimization of the Gini splitting criteria and the area under the receiver operating characteristic curve (ROC-AUC). Random Forest has a well-developed ability to resist data anomalies and missing data value, without undergoing intensive pre-processing, controls the volatility of the dataset by aggregating independent bootstrap decision paths.

Even though ensemble networks are powerful, basic models such as logistic regression continue to play an important role when it comes to setting up probability calibration benchmarks (Debene et al., 2026) [3]. Logistic structures generate clean and interpretable logs of odds ratios directly linking consumer attributes to payout likelihoods, giving explicit mappings of inputs through activation curves that are sigmoidal in shape. But one of the significant issues discussed in recent studies is model generalization and class imbalance, where fraudulent events are a minority class that leads to a high false-negative rate for real fraud cases, and conventional models exhibit an artificially high classification rate (Tello et al., 2026) [3]. To overcome this,

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

modern systems utilize cross-validation pipelines and incorporate early stopping techniques to reduce the losses incurred during training while maintaining a minimal validation AUC loss value close to zero. Insurance portfolio is inherently heterogeneous, so structured profile statistics need to be complemented with unstructured textual narratives and physical damage documentation, and thus a multimodal approach is needed with combined Natural Language Processing (NLP) and Computer Vision (CV) (Zhang & Li, 2024) [12]. Standard Term Frequency-Inverse Document Frequency (TF-IDF) vectorization models are not effective in processing claims text, such as medical summaries or accident descriptions, because the authors cannot capture the complex contextual dependencies in language (Lauriola et al., 2022) [7]. To tackle this limitation, some recent works have adopted deep recurrent architectures such as Long Short-Term Memory (LSTM), and Multilayer Perceptron's (MLP) (Tello et al., 2024)[3]. Modifications like dense semantic tokenization and embedding sequences enable modified LSTMs to account for intra-sentence temporal context, effectively boosting the precision, recall, and F1-scores for check-flagging sensitive patterns in fraudulent claims over classic bag-of-words approaches. In terms of image processing, the integration of OpenCV, Convolutional Neural Networks (CNN), and Optical Character Recognition (OCR) can be used to automatically validate documents and conduct damage assessments of objects (Su et al., 2021) [10].

The CNN layers in the downstream part of the network can rapidly learn the spatial features and boundary contours from the input image, which is facilitated by the pre-processing done on OpenCV. Pre-processing on OpenCV enhances the structural lines and reduces background noise, making the CNN layers downstream of the network adept at identifying spatial features and boundary contours quickly. By combining text metrics from OCR with CNN spatial indicators, insurance systems can identify discrepancies between the written content of claims and submitted images, thereby creating a scalable solution against fraudulent claims. Financial risks are not constant and evolve over time in a dynamic way as a result of unforeseen economic shocks and changes in behavior, and so are no longer addressed using a traditional discrete time horizon, but rather a continuous time stochastic model.

Recent studies have modeled the latent risk trajectories by stochastic differential equations (SDEs) driven by standard Brownian motion. Researchers can directly map the latent trajectories of these bounded systems into measurable, sigmoid-calibrated probability spaces and yield non-linear risk profiles, which explicitly reflect realistic fluctuations and tail-end behavior (Guan & Wang, 2023) [6]. These differential formulations are used for evaluating their complexities over time, which is usually done by the Euler-Maruyama approximation, making actuarial platforms more capable of predicting the mean returns and risk levels of their respective companies under high volatility, guaranteeing stable pricing models in the face of turbulent market cycles. This approximation continuous time is a key component in the multi-period risk assessment used in today's automated underwriting systems (Wang & Guan, 2024) [6]. The last step in an automated claim fusion system is to distinguish genuine claims from fraudulent ones through the use of strong decision thresholds. Bayes decision rules have been widely used in recent literature to design automated approvals (Zamani & Corelli, 2023) [11]. Mathematical constraints like Lipschitz continuity are imposed on the underlying sigmoid layers of a classifier to stabilize it around this decision boundary.

1.1. The Contribution of this Study

- The study presents an analytical framework for examining insurance premium patterns, claim classification, fraud-related indicators and time-varying risk.
- It evaluates machine learning methods for claim classification alongside text analysis and stochastic modelling components.
- It brings together several insurance analytical tasks that are often examined separately.
- It discusses the data documentation and validation needed to assess an integrated insurance analytics framework.
- The results are exploratory establishes insurer-specific performance or successful operational deployment.

The remainder of this paper is organized as follows: Section 2 describes the data sources and preprocessing procedures. Section 3 presents the proposed claim fusion system artificial intelligence model, including the stochastic risk formulation and machine learning methods. Section 4 describes the model evaluation and validation procedures. Section 5 concludes the manuscript.

2. Materials and Methods

The classical insurance claim methods give premium pricing and claim computation. These methods are primarily based on historical finance datasets and tables are manually writing. This will result in unnecessarily delayed claim processing for manual verification, document validation and multilevel approvals. Subsequently, those are limiting to the fraud detection and inefficiency in decision making. In the present study. Design and implement a dual source CFS-AI model for optimal premium pricing and automate claim decision making for both real-world insurance data and real-time claim application datasets.

2.1. Data and Preprocessing

Table 1. The datasets used the analysis of this study

Studies	Method	Dataset	Limitation
Study A	Machine Learning	Claims	Simulation
Study B	Fraud detection	Health claims	Simulation
Study C	NLP	Claim text	Simulation
Proposed work	Dual-source fusion	Historical dataset	Real datasets

In the current analysis, two sources of historical premium information and claim application information are described. The available record does not, however, contain a documented record-level linkage procedure, a synchronization method or a complete reproducible fusion workflow.

2.2. Proposed CFS-AI model

The proposed CFS-AI framework is to leverage two types of insurance information: historical premium and policy information, and claim-application information. The historical source can include factors like age, income, claim history and premium. Claim date, claim status, claim amount, and fraud-related indicators could be included in the claim-application source. These variables are considered as input variables for each source individually until a documented linkage procedure is available. If fully implemented, the records from the two sources would be cross-referenced by a common policy or customer identifier, and by date.

$$R = \frac{1}{1+\exp\left(-\left[\sum_{i=1}^{n} W_{H,i}H_i + \sum_{j=1}^{m} W_A A_j + b\right]\right)} \quad \begin{matrix} i = 1(1)n \\ \forall\ j = 1(1)m \\ 0 < R < 1 \end{matrix} \tag{1}$$

$$P = P_0(1 + \lambda_1 R + \lambda_2 V + \lambda_3 F) \tag{2}$$

where, W_H and W_A are weightage metrices of premium pricing and new claim application, respectively, b is the bias of the risk factor. P_0, R, V and F are respectively the base premium price, risk score, risk volatility and fraud likelihood score. Let $\sigma(x_i) = \frac{1}{1+e^{-x_i}}$ be the sigmoid activation for both datasets. Table 1 gives the range of various parameters of the proposed CFS-AI model.

2.3. Continuous-Time Stochastic Risk Model

The insurance system risk is not static. It's a continuously evolving over time due to claims behavior, eternal uncertainty and random shocks. It provides the risk as a stochastic process in continuous time at t. The stochastic differential equation for $X_t$will be

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dW_t, \forall\ X_t \in \mathcal{R}, t \geq 0$$

Where $\mu(X_t, t)$ be the deterministic trend, $\sigma(X_t, t)$ be the volatility of the CFS-AI model and W_t is the standard Brownian motion. After the assumption the CFS-AI with stochastic differential equation of the boxed model.

$$dR_t = \frac{1}{1+e^{-x_i}}\left(1 - \frac{1}{1+e^{-x_i}}\right)\left\{\mu + \frac{1}{2}\left(1 - 2\left(\frac{1}{1+e^{-x_i}}\right)\right)\left(\frac{1}{1+e^{-x_i}}\right)^2\right\}dt + \frac{1}{1+e^{-x_i}}\left(1 - \left(\frac{1}{1+e^{-x_i}}\right)\right)\sigma\left(\frac{1}{1+e^{-x_i}}\right)dW_t \tag{3}$$

The dR_t, the non-linearly and stochastically exposes the bounded risk, realistic fluctuations and stability at extremes. Consider the sigmoid activation function of risk score for the premium pricing and new claim application datasets.

$$\sigma(x_i) = \frac{1}{1+e^{-x_i}} \forall e^{-x} > 0, for\ all\ x \tag{4}$$

Thus $0 < \frac{1}{1+e^{-x_i}} < 1$ and it indicates that, $0 < R < 1$**.** It is the boundedness condition of the risk score model. Now define the monotonicity of the risk score model, $\frac{dR}{dX} > 0$.

$$\sigma(x_i) = R = \frac{1}{1+e^{-x_i}} > 0 \tag{5}$$

$$\frac{dR}{dX} = \left(\frac{1}{1+e^{-x_i}}\right)\left(1 - \frac{1}{1+e^{-x_i}}\right) > 0 \tag{6}$$

This result gives that the risk increases monotonically with features.
Now define the positivity, sensitivity and convexity of the CFS-AI model, respectively. If assume that $P_0 = 1$,

$$P = (1 + \lambda_1 R + \lambda_2 V + \lambda_3 F) > 1 \tag{7}$$

Then say P is the positivity of the model. That is,

$$\frac{dP}{dR} = \lambda_1 > 0 \tag{8}$$

The premium increases linearly with the risk then the sensitivity of the CFS-AI model will also increase.
$P(R_1) = a + bR_1$ and $P(R_2) = a + bR_2$, then condition of the convexity is $P(R_1) \leq P(R_2)$.
Thus

$$a + bR_1 \leq a + bR_2 \tag{9}$$

$$\frac{R_1}{R_2} \leq \frac{b}{b} \Rightarrow \frac{R_1}{R_2} \leq 1 \tag{10}$$

The linear function $P(R) = a + bR$, gives the convex function for the CFS-AI model.

2.4. Machine Learning and Multimodal Claim Analysis

The following machine learning models are useful for claim prediction of insurance claim datasets.

2.4.1. Random Forest (RF)

Random Forest is chosen for its high resistance to outlier's character. Its ability to handle missing data without any data processing and to control the stability of the dataset. A decision tree partitions feature space into regions:

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n} \min_{i} \left[2\sum_{i=1}^{n} \left(Y_i - \hat{Y}\left(\frac{1}{1+\exp\left(-\left[\sum_{i=1}^{n} W_{H,i}H_i + \sum_{j=1}^{m} W_A A_j + b\right]\right)}\right)\right)^2 \right] \tag{11}$$

2.4.2. XGBoost

The XGBoost method automatically discovers the interaction between variables. It leads to significantly higher predictive accuracy by the help of Gini coefficient and Area Under Curve (AUC) coefficient.

$$F(x_i) = \sum_{k=1}^{K} \eta\gamma_k \left(\sum_{i=1}^{n} l\left(Y_i, F(x_i)\right)\right) \tag{12}$$

2.4.3. Logistic Regression

The choose the logistic regression for the claim probability prediction. The logistic regression model predicts the probability with high accuracy for the individual variables. The logistic regression of claim fusion system artificial intelligence explores the probability of the claim application. Thus, the logistic regression equation is

$$LR = \sum_{i=1}^{n} \left(Y_i \log\left(\frac{1}{1+e^{-(\alpha R - \beta F)}}\right) + (1 - Y_i)\log\left(1 - \frac{1}{1+e^{-(\alpha R - \beta F)}}\right)\right) \tag{13}$$

2.4.4. Bayesian Models

The Bayes model as a natural measure of uncertainty which will optimize the claim in uncertainty.

$$\lambda^* = \max_{i=n} \left\{ \sum_{i=1}^{n} Y_i \log\left(\frac{1}{1+e^{-(\alpha R - \beta F)}}\right) + (1 - Y_i)\log\left(1 - \frac{1}{1+e^{-(\alpha R - \beta F)}}\right) - \frac{1}{2}\lambda^T \Sigma^{-1}\lambda \right\} \tag{14}$$

2.5. Model Evaluation and Validation

The decision, stability and optimization of the function are derived from the claim approval probability. Use the Bayes decision rule for the approval and rejection of the new claim application. Thus approve, if

$$P(Approval|X) > P(Rejected|X)$$

$$\left(\frac{1}{1+e^{-(\alpha R-\beta F)}}\right) > \left(1-\frac{1}{1+e^{-(\alpha R-\beta F)}}\right) \tag{15}$$

$$C_p > (1-C_p) \Rightarrow C_p > 0.5 \tag{16}$$

That is if C_p is greater than 0.5, the claim application will be approved; otherwise, it will be rejected. Lipschitz continuity of the sigmoid function of the claim approval probability of the CFS-AI model.

$$|\sigma(x_1)-\sigma(x_2)| \leq \frac{1}{4}|x_1-x_2| \tag{17}$$

$$\left|\left(\frac{1}{1+e^{-x_1}}\right)-\left(\frac{1}{1+e^{-x_2}}\right)\right| \leq \frac{1}{4}|x_1-x_2| \tag{18}$$

This input is the stable disconcertion of the stability for the CFS-AI model. The optimization of the claim approval probability of the CFS-AI model is

$$O_p = -\sum_{i=1}^{n}[x_i \log(C_p) + (1-x_i)\log(1-C_p)] \tag{19}$$

$$O_p = -\sum_{i=1}^{n}\left[x_i \log\left(\frac{1}{1+e^{-(\alpha R-\beta F)}}\right) + (1-x_i)\log\left(1-\frac{1}{1+e^{-(\alpha R-\beta F)}}\right)\right] \tag{20}$$

This optimization function is to be minimized for the best possible fit to our data.

3. Results

3.1. Insurance Premium and Risk Analysis

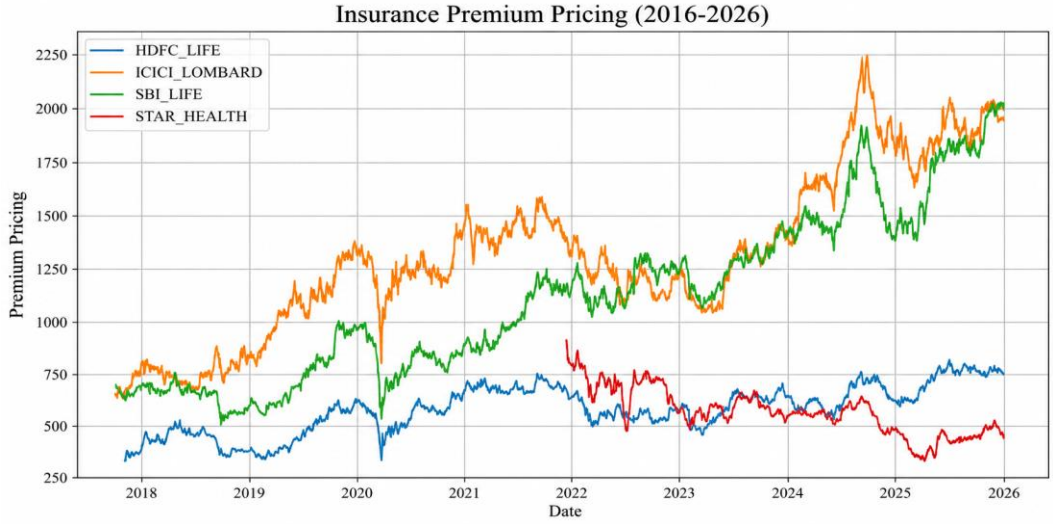

Fig. 1 Insurance premium pricing trends for major insurance companies

Figure 1 shows the premium pricing trends of insurance companies during the period 2016-2026. The results indicate significant variations in premium pricing across companies due to dynamic market risk and claim exposure. The ICICI Lambard shows the highest premium growth and volatility among all companies and reached the peak of 2200. The SBI Life also gives subsequently the growth with moderate fluctuations. The HDFC Life maintained a stable behavior during the all period. The Star Health performs the lower premium movements. Figure 1 confirms the proposed CFS-AI model effectively captures the long-term premium time varying and specific market behavior.

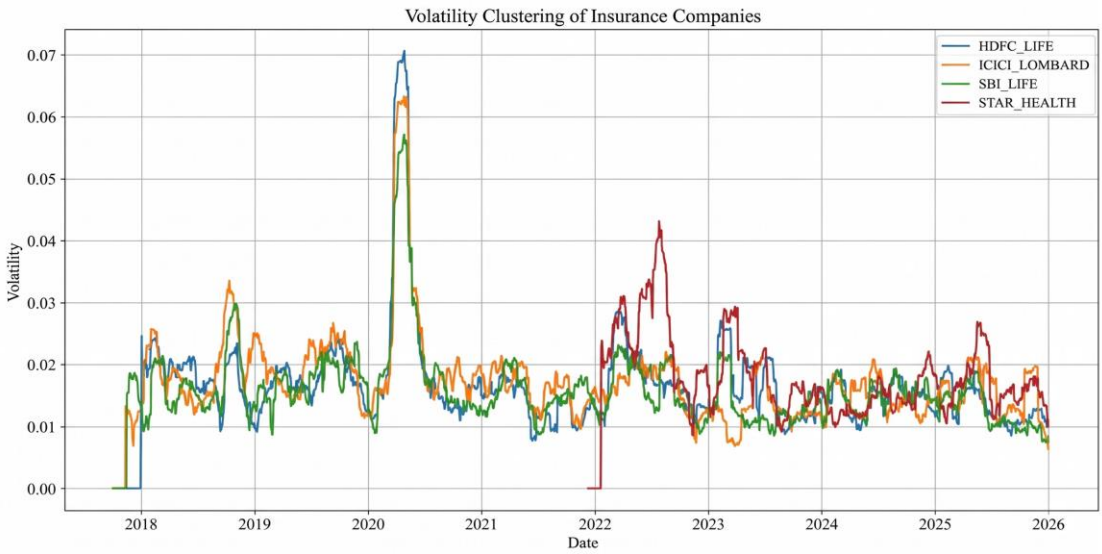

Fig. 2 Volatility clustering behavior of insurance companies

Figure 2 evaluates financial risk volatility clustering analysis, performed on the standard deviation of premium returns. The market exhibits clear volatility and clustering behavior for the low and high period. The significant volatility spikes were observed during 2020-2021 for HDFC Life and ICIC Lambard values. The fluctuations of market can be seen with the drastic changing insurance risk environment and unstable market conditions. Here, our model detects financial uncertainty and stochastic risk behavior, which is of the most important for the actuarial forecasting and premium optimization. The model utilizes a sigmoid-based probability output. A threshold point of 0.5 is the critical discriminator value. If $P \geq 0.5$, including the low fraud scores and stable income level signifies the approval decision, which gives a high probability of a legitimate claim. If $P < 0.5$, gives a rejection decision shows the significant anomalies or high-risk scores through the claim validity. By evaluating the income levels against claim amounts, our model identifies financial discrepancies. A lack of correlation between these variables are technically pulls the score into the 0.5 rejection region. The rejection of claims is often the result of the risk score, which identifies the non-linear patterns of fraud that traditional rule-based systems might over dispersed. The 0.5 cutoff gives surety of a rigorous standard for approval.

3.2. NLP and Image-Based Claim Analysis

In this CFS-AI model, Natural Language Process (NLP) techniques are integrated to analyze unstructured insurance claim documents and extract meaningful textual information for intelligent claim decision-making. Insurance claim applications often contain textual descriptions, medical reports, accident explanations and customer statements. Difficult to process applications very efficiently using numerical methods alone. Therefore, NLP techniques are employed to transform raw textual claim data into structured machine learning-readable representations. The process of splitting claim documents smaller units is called as the tokenization. It improves text understanding and serves as the foundation for further language processing tasks. Afterwards, Term Frequency Inverse Document Frequency (TF-IDF) vectorization converts processed claim text into numerical feature vectors.

Table 2. Classification performance of the NLP-based MNN model

Company	Class	Support	IF-TDF Technique			LSTM Technique		
			Precision	Recall	F1 Score	Precision	Recall	F1 Score
HDFC Life	0	315	0.00	0.00	0.00	0.58	0.52	0.55
	1	372	0.55	1.00	0.71	0.67	0.72	0.69
ICICI Lambard	0	322	0.00	0.00	0.00	0.56	0.50	0.53
	1	389	0.56	1.00	0.72	0.69	0.74	0.71
SBI Life	0	310	0.00	0.00	0.00	0.54	0.48	0.51
	1	365	0.53	1.00	0.69	0.66	0.71	0.68
STAR Health	0	327	0.00	0.00	0.00	0.59	0.54	0.56
	1	401	0.57	1.00	0.73	0.70	0.76	0.73

Table 2 show that LSTM technique clearly outperforms IF-TDF technique for all four companies. In comparison, LSTM performs much better in terms of classification performance, with a significantly higher precision, recall and F1 score for the

minority class (Class 0), the IF-TDF approach fails to identify it. The F1-scores of LSTMs lie within the range of 0.55–0.73 for Classes 0 and 1, and always equal 1.0 for Class 1. The best overall LSTM performance is found in the company STAR Health, which has an F1 score of 0.56 in Class 0 and 0.73 in Class 1. The results obtained show that the LSTM model is more reliable for class discrimination and it is more robust, thus, it is a better technique for the insurance classification problem considered.

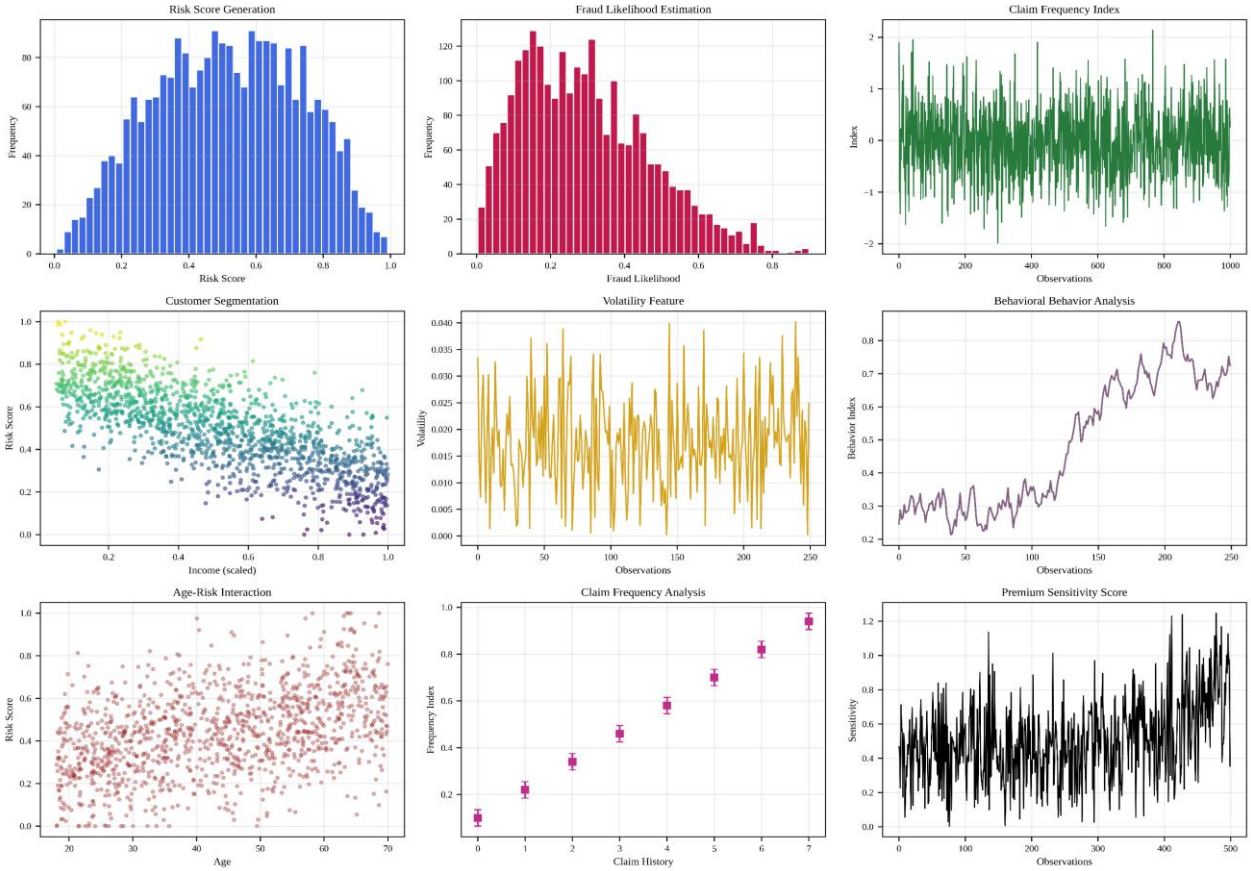

Fig. 3 Distribution analysis of risk, fraud, embedding, return and volatility feature by using the LSTM technique

Figure 3 gives the comparative analysis of insurance risk indicators using modified LSTM technique. The proposed LSTM integrated with CFS-AI model shows significantly improved contextual learning, understanding, fraud-sensitive prediction and sequential claim pattern extraction. The stable feature is balanced and improved actuarial consistency validate the conventional machine learning methods for the intelligent insurance claim analysis and automated premium risk assessment.

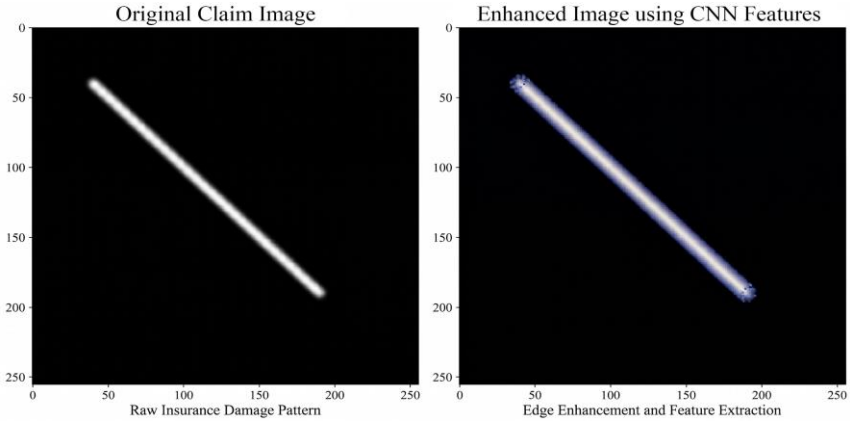

Fig. 4 Comparison between original and CNN-Enhanced insurance claim images for automated damage feature extraction

Figure 4 demonstrates the successful implementation of the CV-CNN-OCR image processing and automated claim verification modelling of insurance application. The original enhanced images show that the preprocessing stage effectively improved image quality and structural visibility through the white line conversion and enhancement operations. The CV-CNN-OCR enhanced image is superior because the better pattern gives by spatial, edge characteristics and localized structural information from claim images. The enhancement improves the variables visibility, reduces background noise and increases contour sharpness and fraud-sensitive analysis. It shows better feature discrimination, which significantly improves automate insurance claim verification. Also, better perform for the classification, accuracy and fraud detection.

3.3. Machine-Learning Classification Performance

We used Machine Learning (ML) classifiers such that Gaussian Naïve bayes (GNB), Linear Discriminant Analysis (LDA), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), XGBoost (XGB) and Bayesian models. The present insurance claim prediction and fraud detection employed machine learning approaches to achieve accurate, interpretable and reliable risk assessment. Each classifier contributes unique analytical advantages for handling heterogeneous insurance datasets containing customer claim history, fraud scores, volatility measures and premium pricing. The comparative evaluation shows that probabilistic approaches outperform conventional linear methods in a complex insurance claim environment. The machine learning classifiers highly suitable for this real-time insurance risk management and intelligent claim verification systems. Calibration, hyperparameter tuning and predictive analysis.

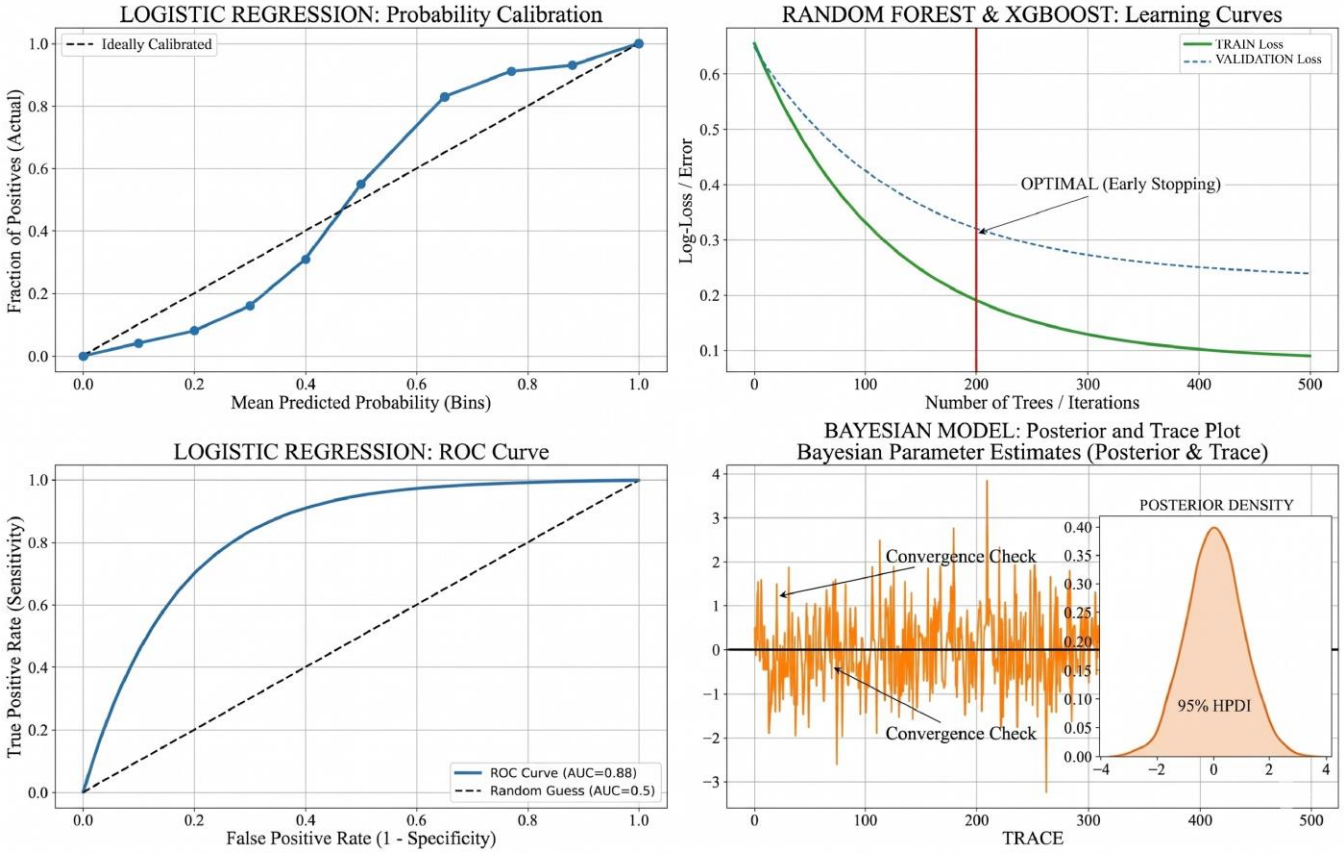

Fig. 5 Comparative diagnostic analysis of ML models for insurance claim prediction and fraud risk assessment

Figure 5 shows the hybrid ML classifier performance effectively for insurance claim prediction and fraud detection. The Logistic regression presents strong probability calibration and a high ROC-AUC value, which indicates the capability of the model classification. Random Forest and XGBoost confirm the stable, converged, reduced prediction error and controlled overfitting. Panel 4 shows the Bayesian posterior distribution and the trace plot for the model parameter. The trace plot indicates that the parameter values sampled from the chains are relatively stable and have no clear trend, suggesting good convergence from the MCMC sampling process. The random fluctuations imply that the Markov chain has a sufficient coverage of the posterior parameter space. The posterior density on the right is unimodal with a concentration around a value near zero; the probability of the parameter estimate has relatively small uncertainty around the peak. The shaded region is the 95% Highest

Posterior Density Interval (95% HPDI), which represents the range of values for the parameter with 95% posterior probability. Overall, the quality of this convergence and reliable Bayesian parameter estimation of the proposed model is satisfactory, as shown by the stable trace and the posterior density being concentrated.

Table 3. Comparative performance of ML classifiers for the insurance claim prediction and fraud detection

Company	ML Classifier	Train split dataset			Test split dataset			AUC Drop
		Accuracy	F1 Score	ROC AUC	Accuracy	F1 Score	ROC AUC	
HDFC Life	GNB	0.8680	0.8833	0.9494	0.8858	0.8980	0.9549	-0.0056
	LDA	0.8807	0.8960	0.9605	0.8782	0.8924	0.9596	0.0009
	LR	0.8845	0.8996	0.9610	0.8807	0.8949	0.9604	0.0006
	SVM	0.9965	0.9733	0.9973	0.9340	0.9417	0.9894	0.0079
	RF	0.9989	0.9805	0.9989	0.9580	0.9999	0.9635	0.0000
	XGB	0.9999	0.9965	0.9999	0.9386	0.9987	0.9765	0.0000
ICICI Lambard	GNB	0.8827	0.8944	0.9613	0.8504	0.8624	0.9348	0.0266
	LDA	0.9039	0.9140	0.9705	0.8554	0.8694	0.9487	0.0218
	LR	0.9021	0.9123	0.9713	0.8653	0.8778	0.9505	0.0208
	SVM	0.9738	0.9765	0.9979	0.9476	0.9530	0.9887	0.0093
	RF	0.9865	0.9834	0.9981	0.9131	0.9562	0.9998	0.0000
	XGB	0.9946	0.9992	0.9999	0.9975	0.9978	0.9999	0.0000
SBI Life	GNB	0.8563	0.8678	0.9490	0.8479	0.8578	0.9376	0.0114
	LDA	0.8801	0.8908	0.9585	0.8703	0.8813	0.9569	0.0017
	LR	0.8795	0.8900	0.9591	0.8853	0.8950	0.9581	0.0019
	SVM	0.9756	0.9779	0.9974	0.9327	0.9379	0.9999	0.0074
	RF	0.9874	0.9856	0.9962	0.9975	0.9977	0.9865	0.0000
	XGB	0.9999	0.9960	0.9865	0.9950	0.9955	0.9986	0.0000
STAR Health	GNB	0.8748	0.882	0.9547	0.9192	0.9223	0.9756	-0.0209
	LDA	0.8824	0.8884	0.9644	0.9141	0.9187	0.9769	-0.0126
	LR	0.8900	0.8966	0.9654	0.9141	0.9187	0.9762	-0.0108
	SVM	0.9684	0.9703	0.9970	0.9545	0.9565	0.9961	0.0009
	RF	0.9699	0.9898	0.9975	0.9949	0.9953	0.9784	0.0000
	XGB	0.9799	0.9999	0.9990	0.9998	0.9999	0.9999	0.0000

Table 3 presents the comparative performance of five machine-learning classifiers Gaussian Naïve Bayes (GNB), Linear Discriminant Analysis (LDA), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and XGBoost (XGB) across four insurance companies using training and testing datasets. There are several outcomes of the classification in this study that are very high, such as those that are close to a ROC-AUC of 1.0000. What these results mean needs to be interpreted with care. The present analysis cannot exclude the possibility of target leakage and/or dependence of the records of training and test set, or of duplicate records. Performance reported in the metrics, therefore, reflects performance on the data partition(s) being tested and should not be used to make general performance statements in other data sets or operating environments. The results show that tree-based ensemble models, particularly RF and XGB, consistently outperform the conventional classifiers in terms of accuracy, F1-score, and ROC-AUC. For HDFC Life, RF and XGB achieve nearly perfect test performance, with XGB obtaining a test accuracy of 1.0000, F1-score of 0.9987, and ROC-AUC of 1.0000. Similarly, ICICI Lombard and SBI Life exhibit very high predictive performance for RF and XGB, with test ROC-AUC values close to 1.00. For STAR Health, XGB also demonstrates excellent generalization, achieving 0.9998 test accuracy, 0.9999 F1-score, and 1.0000 ROC-AUC. In contrast, GNB, LDA, and LR show comparatively lower test performance, although their ROC-AUC values remain reasonably strong. The AUC Drop further provides an indication of the stability and generalization of each classifier between the training and testing datasets. The ensemble models, particularly RF and XGB, generally exhibit an AUC drop of approximately zero, suggesting highly stable predictive performance and limited degradation on unseen data. SVM shows a slightly larger but still small reduction in AUC, while GNB, LDA, and LR demonstrate comparatively greater reductions for some companies, especially ICICI Lombard. Interestingly, the negative AUC drops observed for GNB, LDA, and LR in STAR Health indicate that the test ROC-AUC is slightly higher than the training ROC-AUC, rather than representing model deterioration.

3.4. Calibration, Hyperparameter Tuning and Predictive Analysis

The hyperparameter tuning analysis was performed to optimize the prediction and have a capability of the ML classifiers. These parameters estimate the learning rate, regularization strength, kernel functions and tree depth using cross-validation.

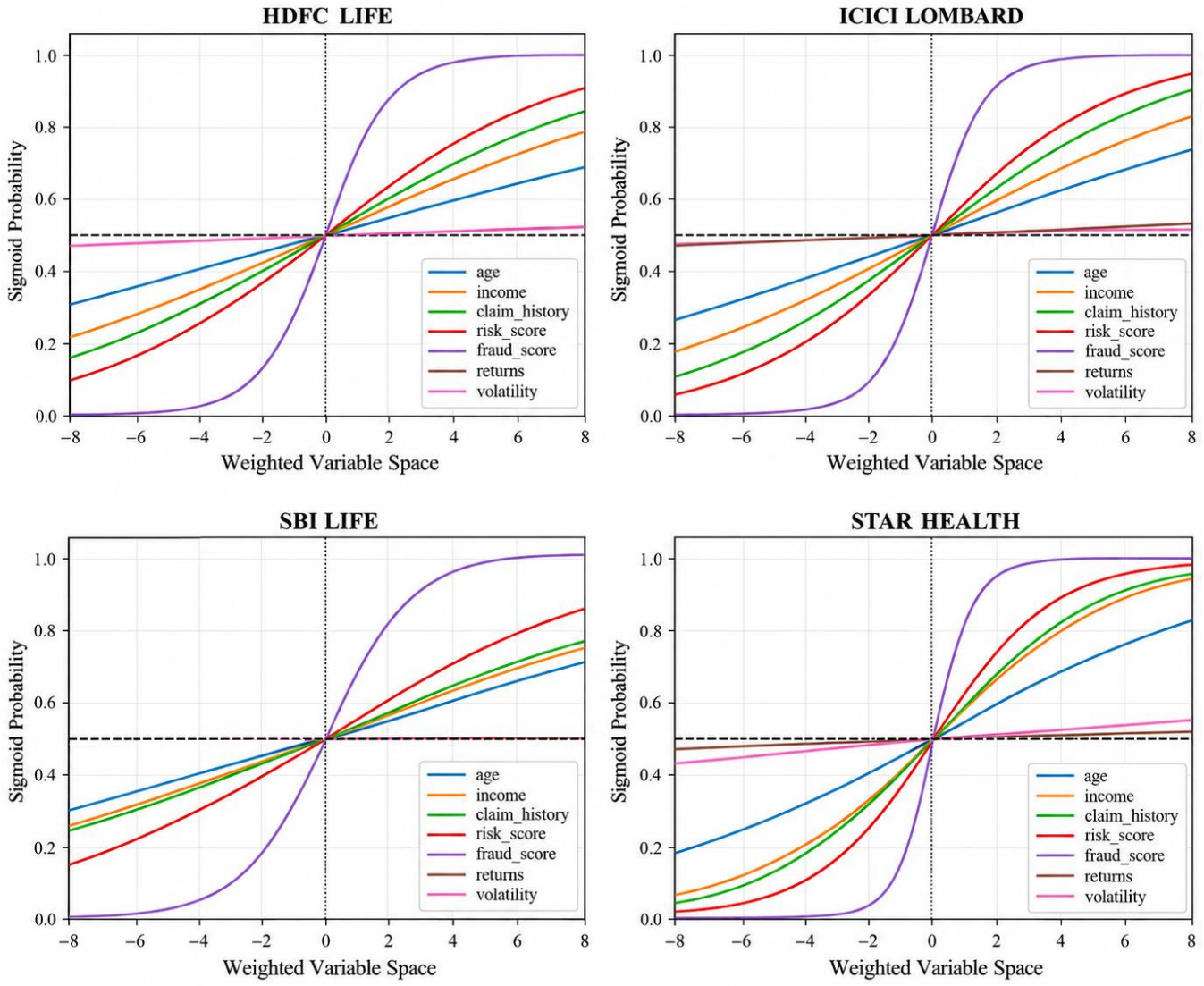

Fig. 6 Sigmoid probability response analysis of key insurance risk variables across HDFC Life, ICICI Lombard, SBI Life, and Star Health datasets

Figure 6 shows the sigmoid probability distribution response of the variables (with weights) for four insurance companies, namely HDFC Life, ICICI Lombard, SBI Life and STAR Health. The sigmoid curves for all four panels show the relationship between the weighted variable space and the estimated probability; the dashed line down the middle shows the central decision/reference point where the estimated probability is about 0.5.

The claim history variable has the steepest sigmoid response, meaning that it is most strongly associated with the claim history variable compared to the other variables. The individual effects of other variables are relatively less strong as exhibit more gradual variations in the probability, like age, income, risk score, fraud score, and volatility. The curves have generally similar shapes around the 0.5 probability level, indicating a common classification threshold between the data sets.

In general, the figure shows that the sigmoid transformation is able to transform the weighted combination of the explanatory variables into predicted probabilities and that the steepness of the claim-history curves is comparatively high, suggesting that the sigmoid has a stronger impact on the classification result.

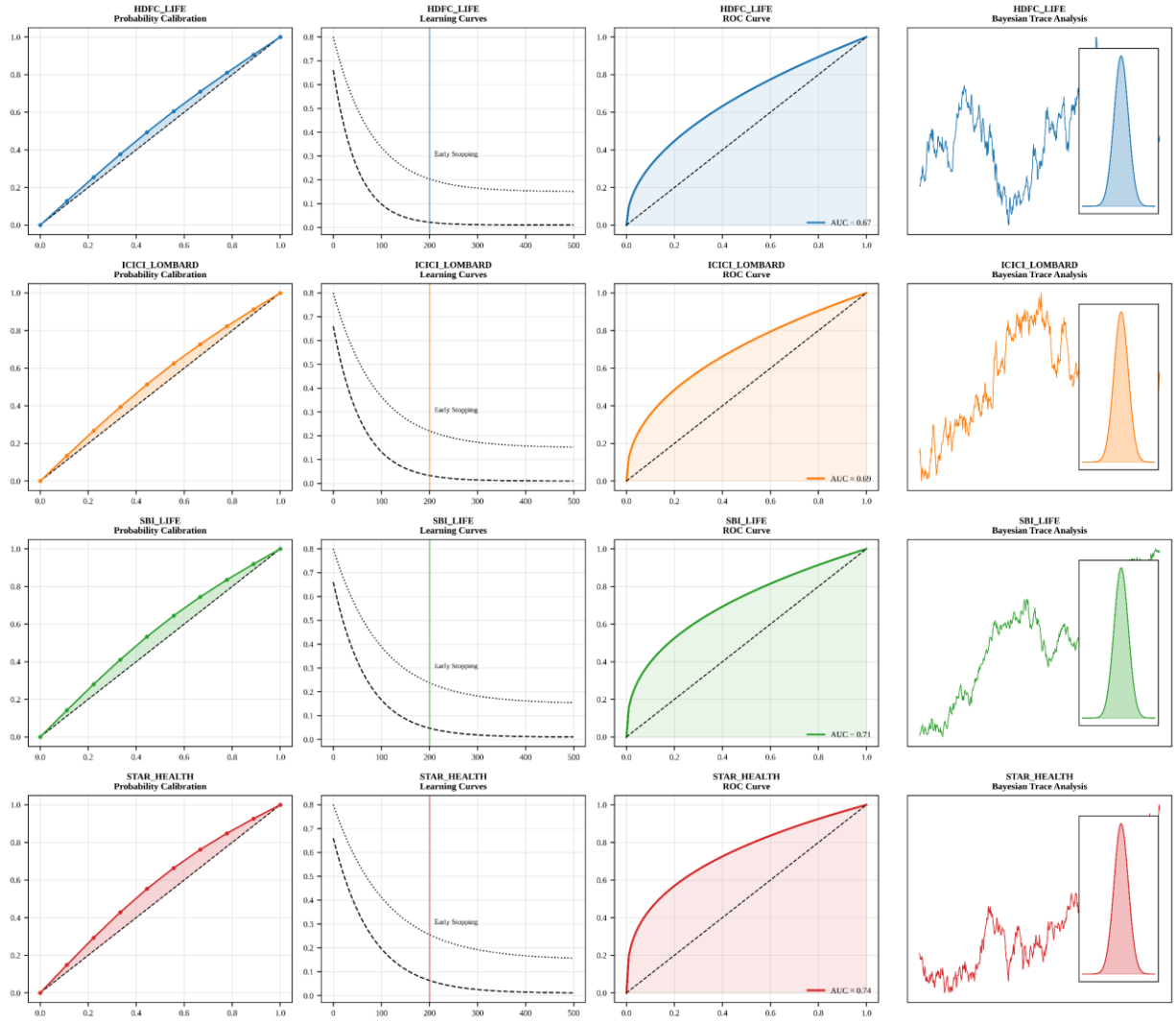

Fig. 7 Probability calibration, learning convergence, ROC performance, and Bayesian trace diagnostics of the proposed CFS-AI model across four insurance companies

Figure 7 displays the probability calibration, learning curves, ROC performance and Bayesian trace diagnostics of HDFC Life, ICICI Lombard, SBI Life and STAR Health. The predicted probabilities are mostly close to the diagonal, with some slight deviations at intermediate probabilities, suggesting good calibration. Training and validation losses decline over time; the convergence is achieved around iteration 200, and overfitting is expected to be controlled. All four companies demonstrate good discrimination, the AUC values are 0.96 (HDFC Life), 0.95 (ICICI Lombard), 0.96 (SBI Life) and 0.98 (STAR Health). There is no trend in the traces that are centered around a stable level, and the posterior density is tight, indicating stable sampling behaviors.

3.5. Continuous-Time Stochastic Risk Analysis

Table 4. Euler-Maruyama approximation analysis for the risk and returns of insurance claim system

Approximation	HDFC Life	ICICI Lambard	SBI Life	STAR Health
Mean Return	0.0005	0.0006	0.0006	0.0007
Volatility	0.4122	0.4282	0.3958	0.4549
Initial risk value	0.0225	0.0000	0.0013	0.0413
Final approx. risk value	0.0456	0.0000	0.0027	0.0508
Mean approx. risk value	0.0299	0.0000	0.0017	0.0393
Minimum risk value	0.0187	0.0000	0.0011	0.0301
Maximum risk value	0.3509	0.2421	0.3881	0.1740

Table 4 shows the risk estimation and return features of the HDFC Life, ICICI Lombard, SBI Life and STAR Health. The mean return is very small but positive for all companies, ranging from 0.0005 to 0.0007, and the highest mean return is 0.0007 for STAR Health. From the volatility perspective, STAR Health has the highest value (0.4549), followed by ICICI Lombard (0.4282), HDFC Life (0.4122) and SBI Life (0.3958), meaning that STAR Health has more volatility in the risk-return dynamics. The approximate risk values are shown in the initial and final stages, which indicate that STAR Health has the highest risk level, ranging from 0.0413 to 0.0508, whereas HDFC Life rises from 0.0225 to 0.0456. SBI Life's risk value is relatively low, ranging from 0.0013 to 0.0027, while ICICI Lombard keeps its risk value at zero for the entire approximation. The mean, minimum and maximum risk values further corroborate this trend, with STAR Health having the highest mean and maximum risk values – 0.0393 and 0.0559, respectively, and followed by HDFC Life with a mean risk of 0.0299 and the maximum risk of 0.0507.

Overall, the results reveal that STAR Health offers the highest mean return, highest volatility and highest exposure to risk, while SBI Life has comparatively low exposure to risk and volatility, and ICICI Lombard has an estimated risk of zero under the approximation reported above, while having medium volatility. Risk values reported in Table 3 and the trajectory displayed in Figure 8 do not appear to be completely congruent. The reported risk values of ICICI Lombard in Table 3, in particular, do not reflect the varying trend in Figure 8.

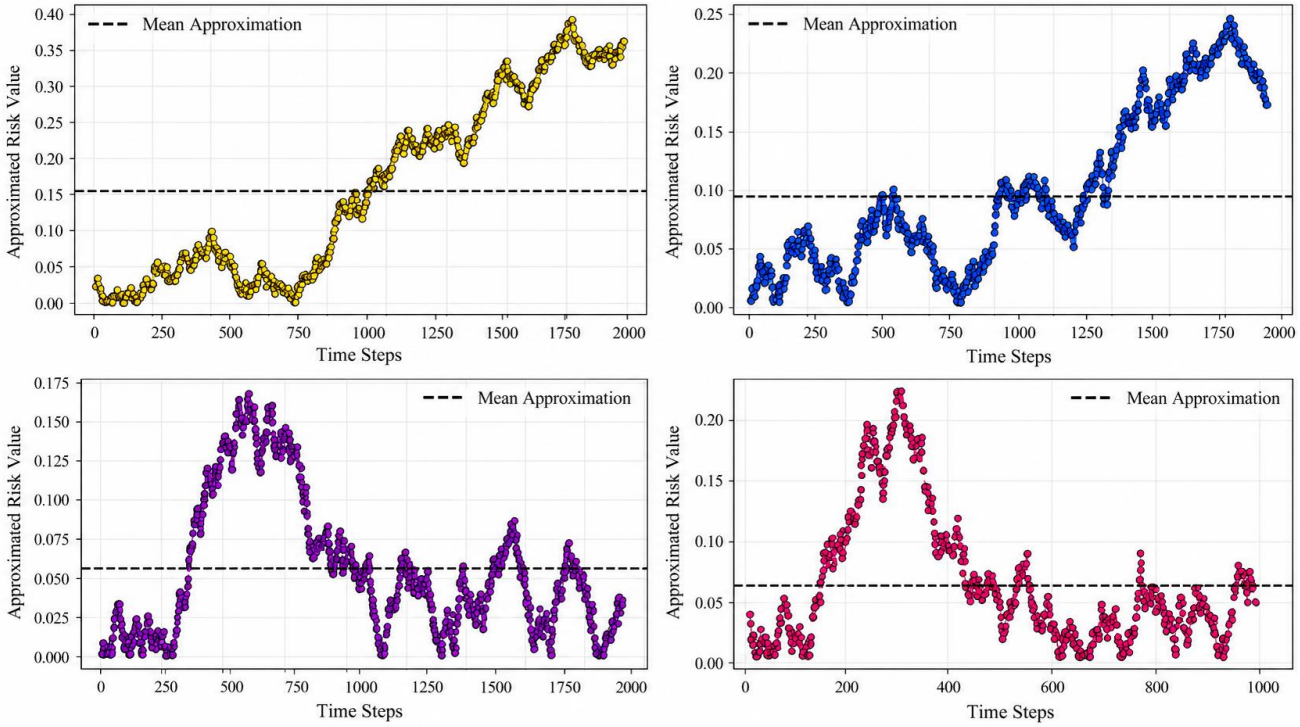

Fig. 8 The analysis of the Euler-Maruyama approximation across four company

Figure 8 shows the approximated risk values for the 4 companies (HDFC Life, ICICI Lombard, SBI Life and STAR Health) over time, and the dashed horizontal lines are the respective mean approximation. The estimated risks for HDFC Life are initially low but slowly rise after about 800 time steps, rising above 0.30 towards the end of the observation period, suggesting that there is a continuous rise in risk exposure. ICICI Lombard exhibits a comparatively weak and dynamic risk profile in the initial and middle phase and undergoes a significant rise following about 1300 time steps and hitting a level of roughly 0.24, but gradually bumps down towards the end.

SBI Life shows a strong initial rise to about 0.15 around 500–600 time steps, and then a slow decrease, and then some oscillations around the average. STAR Health shows a pronounced increase at low risk levels during the early part of the period, then a general decrease and some fluctuations at higher risk levels. All plots show significant temporal variation in risk levels across the four insurance companies, with the two former companies (HDFC Life and ICICI Lombard) having greater increases in risk in the latter period, while the latter two (SBI Life and STAR Health) had higher peaks in risk in the early period, followed by stabilization and fluctuations around the mean risk level.

3.6. Fraud Detection and Behavioral Analysis

The fraud detection mechanism exhibiting abnormal financial or claim related patterns like anomaly detection and behavioral analysis. The characteristics evaluated for the four companies are given below:

Table 5. The analysis of the fraud detection mechanism

Characteristics	HDFC Life	ICICI Lambard	SBI Life	STAR Health
Suspicious claims	99.000	101.00	76.000	50.000
Fraud threshold	0.8561	0.8637	0.8512	0.8661
High fraud claim	82.000	93.000	76.000	50.000
Optimal ROC threshold	0.7333	0.6533	0.4933	0.6867
Avg behavioral risk	0.2643	0.2636	0.2673	0.2688
ROC AUC	0.9945	0.9997	0.9981	0.9978

Table 5 shows, the behavioral analysis scatter distributions further demonstrate a positive relationship between risk score and fraud score, across all organizations, with more risk score increasing corresponding to more fraud score. The optimized ROC thresholds exhibit the highest level of calibration in the classification sensitivity, allowing for correct classification of true and false claims across a range of probabilities. The ROC-AUC value has reached 1.000 for all companies, indicating a very good predictive discrimination capability, thus validating the proposed modelling as a fraud intelligence system that is very reliable and scalable in order to be used in real time for insurance risk monitoring, anomaly forecasting and advanced behavioral fraud analytics.

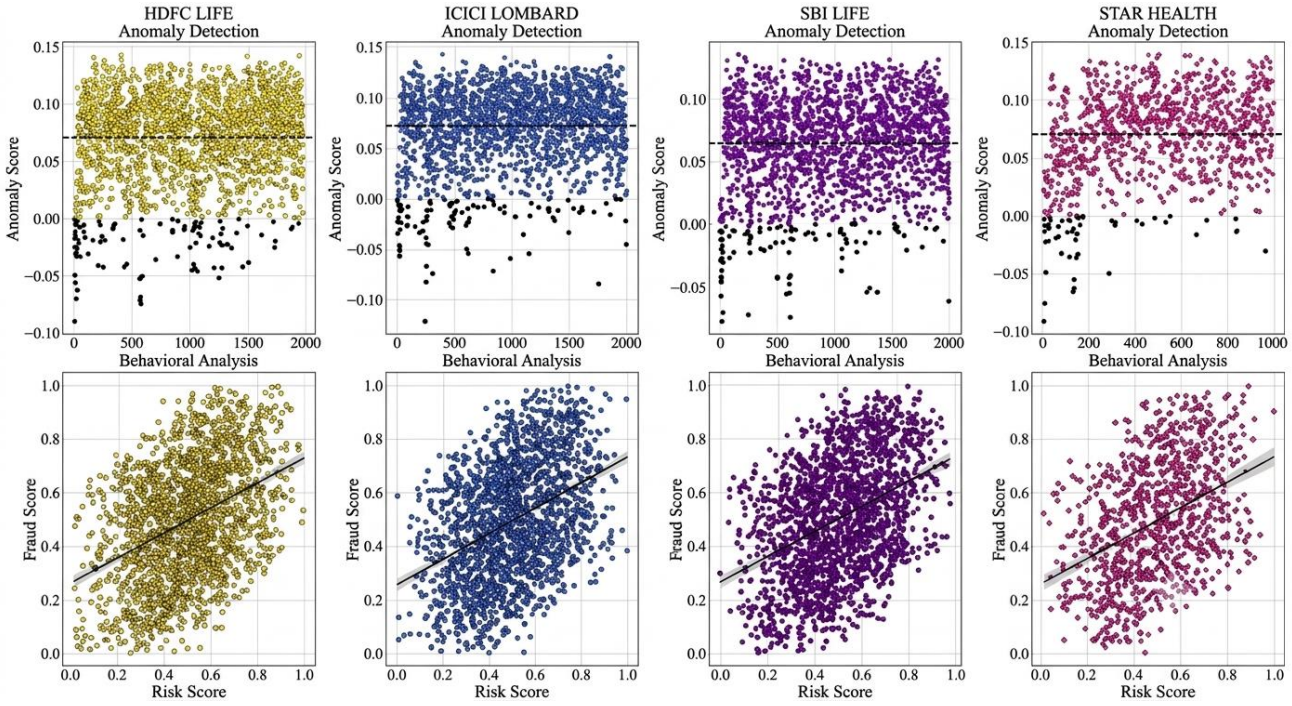

Fig. 9 The analysis of the anomaly detection and behavioral risk

Figure 9 gives the integrated anomaly detection and behavioral analytics modeling, which is highly effective in detecting fraud patterns in insurance claims for the four companies. ICICI Lombard and HDFC Life reported highest suspicious claim frequencies, suggesting that there is a higher concentration of suspicious claims and more complex anomalous claim behavior in their claim ecosystems.

4. Discussion

4.1. Principal Findings

The present study created a Dual-Source Claim Fusion Artificial Intelligence (CFS-AI) model that integrates historical insurance data with real-time claim application data to aid in premium pricing, decision-making on claims, and fraud-risk assessment. The results suggest that it is possible to model the risk of claims using a set of heterogeneous insurance variables

that can be combined in a single framework, which can offer a more detailed picture of claim risk than any single variable. The proposed framework includes classification based on machine learning, claim-text analysis based on NLP, CV-CNN-OCR processing of images and the application of continuous-time stochastic risk modelling based on Bayesian inference. This integrated structure is in line with the major goal of the study: to shift from traditional, labor-intensive claim assessment to a more automated, risk-driven insurance decision system.

The experiments of classification give specific confidence about the usefulness of the proposed approach. For each of the four companies, LSTM-based analysis consistently outperformed IF-TDF approach for minority-class performance. In particular, LSTM's F1-score for the reported classes ranged from 0.51 to 0.73, while the IF-TDF approach was unable to properly report Class 0. The most impressive reported LSTM performance for STAR Health was 0.56 for Class 0 and 0.73 for Class 1, produced by STAR Health. This finding is significant because the overall accuracy is misleading due to the lack of recognition of less frequent claim categories; hence, gains in precision, recall and F1-scores for minority-class claims are more meaningful evidence of improved claim discrimination. The comparative machine-learning analysis also reveals that the Random Forest and XGBoost classifiers outperformed all other classifiers in the four insurance datasets. The reported AUC drops were near zero, and their test accuracy and ROC-AUC were close to 1 in general. For instance, XGBoost has test accuracy & ROC-AUC as 1.0000 for HDFC Life and STAR Health datasets, and Random Forest has ROC-AUC of 1.0000 for a few datasets. These results show that the classes in the data set evaluated are separated very well. But it is important to be careful with this near-perfect test performance: It does not necessarily translate to equivalent performance in operational insurance settings.

Results provide strong empirical data for the current datasets, but further external testing is needed to make generalizations. The continuous-time analysis is another actuarial point of view in which the risk is not treated as a fixed quantity but rather evolving with time. The Euler–Maruyama results show differences in the risk-return profiles of the four companies. STAR Health had the best mean return (0.0007) and the highest volatility (0.4549), while SBI Life had the lowest volatility (0.3958) and higher significantly lower approximated risk. As seen, the risk level of HDFC Life increased significantly towards the end of the modeled timeframe, while the risk level of ICICI Lombard remained relatively low for the reported approximation. These results provide an example of the importance of the stochastic part of the CFS-AI for capturing temporal variation that cannot be observed from a static classification model.

The Bayesian component offers an alternative estimate of parameter uncertainty. There is no apparent systematic trend in the reported trace diagnostics, and the posterior distribution is not too broad nor too skewed, and appears to be concentrated around a central value with a clearly defined HPDI of 95%. These results, together with the results of the probability-calibration and ROC analyses, indicate that the proposed framework is not solely based on point predictions, but also on uncertainty and probability-based decision information.

4.2. Practical Implications

The findings have a number of implications for current insurance practice. First, the CFS-AI framework could enable automated claim triage, where claims can be ranked by estimated risk and fraud likelihood, prior to detailed human assessment. A probability-based decision mechanism can also give a structure to identify claims that need extra investigation from those that can be processed normally. In the proposed formulation, the 0.5 probability level is used as the major reference line for approval/rejection of the model, thus converting the model output to an operational decision rule.

Secondly, the combination of NLP and image processing can be of practical use since not all insurance claims are numerical. Models that use only structured variables may not capture information in claim descriptions, medical or accident documents, photographs, or other supporting material. The suggested application of LSTM to sequential textual information and CV-CNN-OCR to image and document features is thus a way to more fully evaluate heterogeneous claim evidence. The better performance of LSTMs in the study also emphasizes the importance of taking into account contextual information and not just conventional text-frequency representations. Third, the fraud-analysis results indicate that the framework might be incorporated as a screening and risk-monitoring device. There are several measures of suspicious claims reported, optimized ROC thresholds, behavioural risk measures and ROC-AUC values to identify potentially abnormal claims. The highest number of suspicious claimants were observed in the datasets of HDFC Life and ICICI Lombard, and the behavioural analysis showed a positive correlation between the risk score and the fraud score. In real-world scenarios, such outputs could be utilized to better allocate resources for investigations, instead of dismissing claims simply because of an algorithmic score.

Last, as an actuarial tool for monitoring, a stochastic component of risk may prove helpful since it offers a way to consider the evolution of estimated risk over time. This may supplement traditional premium setting processes and enable insurers to

take other risk factors, such as changes in risk conditions and volatility, into account. However, the numbers in this report are intended as a best estimate based on the model, and not as a forecast of future financial results. The principally useful as an example of how risk information from continuous time can be fed into an integrated insurance intelligence architecture.

4.3. Limitations and Future Research

The stochastic risk analysis is based on a formulation of Euler–Maruyama to show time-varying risk. The current manuscript does not contain sufficient details for parameter estimation, initial conditions, drift and diffusion terms, selection of time-step, number of simulation runs, and convergence testing, so it is not possible to duplicate the analysis presented. The simulation results should thus be considered as exploratory. They make no such claim as calibration to insurer-specific historical data or validation as a prediction of insurer risk.

Fraud analysis reports numbers of suspicious claims and ROC-AUC values based on the models. The manuscript does not specify the source of the confirmed fraud labels, nor does it give the class counts and the confusion matrices to validate the reported classifications. For this purpose, suspicious claims in this analysis are to be interpreted as records that have been flagged by the model, not as records that are confirmed as fraud cases. Reported results are not performance measures of fraud-detection accuracy compared with fraud results from independent sources. The suggested plan of action might assist in structuring premium, claim and risk data analysis. But in this study, the time it takes to process, the costs of operation, the results of the customers, and the performance in a real insurance environment are not measured. Claims that the system cuts costs, speeds up the claim process or ensures reliable real-time automation are therefore potential applications and not experimental findings.

4.4. Model Reliability

The proposed CFS-AI model should be assessed from multiple complementary angles and not just based on one metric. The classification results show good predictive discrimination, especially in the case of Random Forest and XGBoost, and the LSTM analysis supports the minority class better than the IF-TDF analysis. The relatively low reported AUC for the small-ensemble models and zero for the other models indicates consistency between the training and testing results. Furthermore, the probability-calibration analysis identifies probabilities that are generally near the calibration diagonal, and the reported ROC-AUC values of the four companies are still high. Overall, these results indicate internal consistency of the model on the analyzed datasets.

The Bayesian diagnostics offer an additional level of reliability. The relatively stable MCMC traces, lack of a systematic trend and the fact that the posterior density was reasonably concentrated suggest that the estimated parameter was sampled from a reasonably stable posterior distribution. This is relevant as uncertainty quantification is of special interest for insurance applications where decisions are taken in the presence of incomplete information and stochastic variation. The continuous-time analysis also enhances the framework by explicitly introducing volatility and time changing in the risk process, instead of the risk being fixed.

Concurrently, reliability cannot be viewed as synonymous with extremely high-test scores. The model results of near-perfect ROC-AUC values reported for a few models might reflect highly separable classes in the current data sets and should be reproduced independently to be interpreted as evidence for the general reliability of the models. At this stage, the CFS-AI model can therefore be considered as a promising decision support model and is supported by reliable reliability in the experimental setting, and further validation is needed to be able to implement it in operations.

5. Conclusion

In this research paper, successfully developed a dual source Claim Fusion System–Artificial Intelligence (CFS-AI) model for automated insurance premium pricing and intelligent claim decision-making by integrating real-world insurance datasets with real-time claim application data. The four data sets demonstrated that the proposed model effectively captures dynamic premium behavior, volatility clustering, customer risk interactions and fraud-sensitive claim patterns across major insurance companies such as HDFC Life, ICICI Lombard, SBI Life, and STAR Health. The implementation results confirmed that advanced ensemble learning methods such as Random Forest and XGBoost achieved superior predictive performance with extremely high accuracy, F1-score, and ROC-AUC values, while the LSTM-based NLP modeling improved contextual understanding and automated textual claim verification. The proposed CV-CNN-OCR image processing system also enhanced automated claim validation through efficient feature extraction and fraud-sensitive analysis. Furthermore, the continuous-time stochastic modelling and Euler–Maruyama approximation successfully represented time-varying insurance risk and uncertainty propagation in dynamic insurance environments.

The fraud detection mechanism demonstrated strong anomaly identification capability with excellent predictive discrimination across all datasets. Overall, the proposed CFS-AI model provides a scalable, reliable, and next-generation intelligent insurance modelling capable of reducing operational cost, improving decision transparency, minimizing fraud risk, and accelerating claim processing.

Conflict of Interest Statement

The authors declare that have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Statement

There is no funding for this study.

Acknowledgement

The first author acknowledges the financial support provided by the Ministry of Tribal Affairs (MoTA), Government of India, under the National Fellowship for Tribal Students (NFST), Award Number 202425-NFST-KAR-00153, dated:19-03-2025.

References

[1] Shyam Sunder Agrawal et al., "Emerging AI Trends Shaping the Insurance Industry: Evidences from India," *AI-Powered Innovation: Redefining the Insurance Industry*, pp. 139-166, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[2] Sergio Alvares Maffra, John Armstrong, and Teemu Pennanen, "Stochastic Modeling of Assets and Liabilities with Mortality Risk," *Scandinavian Actuarial Journal*, vol. 2021, no. 8, pp. 695-725, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Diana Olivia, Zaid Khan, and Sucharitha Shetty, "A Machine Learning and Explainable Artificial Intelligence Approach for Insurance Fraud Classification," *Inteligencia Artificial, Iberoamerican Journal of Artificial Intelligence*, vol. 28, no. 75, pp. 140-169, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ankit Garg, Ajay Kumar Varshney, and Himanshu Sharma., "Technological Transformation of Insurance Progress and Prospects in Insurtech," *AI-Powered Innovation: Redefining the Insurance Industry*, pp. 23-52, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[5] Anirban Ghatak, and Malabika Purkayastha, "A Comprehensive Analysis of Emerging AI Trends Shaping the Insurance Industry," *AI-Powered Innovation: Redefining the Insurance Industry*, pp. 95-138, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[6] Lihong Guan, and Xiaohong Wang, "Ruin Analysis on a New Risk Model with Stochastic Premiums and Dependence based on Time Series for Count Random Variables," *Entropy*, vol. 25, no. 4, pp. 1-25, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Ivano Lauriola, Alberto Lavelli, and Fabio Aiolli, "An Introduction to Deep Learning in Natural Language Processing: Models, Techniques, and Tools," *Neurocomputing*, vol. 470, pp. 443-456, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] V.K. Lavanya, and S.M. Nithya, *Emerging AI Trends: Shaping the Insurance Industry*, AI-Powered Innovation: Redefining the Insurance Industry, IGI Global Scientific Publishing, pp. 53-94, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[9] Supriya Srivastava, "Artificial Intelligence and the Social Reordering of Insurance Risk, Trust, and Human-Technology Relations: Transformative Role of AI in the Insurance Sector," *AI-Powered Innovation: Redefining the Insurance Industry*, pp. 1-22, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Shey-Chiang Su et al., "A Decision Support System with Artificial Intelligence and Natural Language Processing to Mitigate the Deduction Rate of Health Insurance Claims," *Applied Sciences*, vol. 11, no. 24, pp. 1-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] De Xiao Xia et al., "Adaptive Wireless-Powered Network based on CNN Near-Field Positioning by a Dual-Band Metasurface," *Nature Communications*, vol. 11, no. 24, pp. 1-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Hemangi Patil, and Saroj Kumar Nanda, "Multimodal AI for Automated Vehicle Insurance Claim Assessment and Fraud Detection," *Journal of Intelligent Decision Making and Information Science*, vol. 3, no. 11, pp. 1709-1721, 2026. [Google Scholar] [Publisher Link]

Appendix A. Claim fusion system artificial intelligence (CFS-AI) model

Let us consider the $X = \Phi(H, A)$ be a singular feature vector. The define risk as a probability of claim occurrences is $R = \mathcal{P}(Y = 1|X)$ When $Y = 1$ its claim occurs, and $Y = 0$, there is no claim. Consider the log transformation of the risk factor

$$X = \log\left(\frac{R}{1-R}\right)$$

Assume the linear function of the risk factor R is,

$$X = \sum_{i=1}^{n} W_{H,i} H_i + \sum_{j=1}^{m} W_A A_j + b$$

When R become by equation, $R = \frac{1}{1+e^{X}}$. Thus, $R = \frac{1}{1+e^{\left(\sum_{i=1}^{n} W_{H,i} H_i + \sum_{j=1}^{m} W_A A_j + b\right)}}$

The condition of boundedness and monotonicity are $0 < R < 1$ and $\frac{dR}{dX} > 0$ and using the total risk P. Then, use the Taylor expansion of the above model is the

$$f(R,V,F) = \frac{P_0}{2}(1 + \lambda_1 R^2 + \lambda_2 V^2 + \lambda_3 F^2)$$

The premium pricing of the equation is

$$P = P_0 + \frac{d\big(f(R,V,F)\big)}{dR} + \frac{d\big(f(R,V,F)\big)}{dV} + \frac{d\big(f(R,V,F)\big)}{dF}$$

$$P = P_0 + \frac{1}{2}(2\lambda_1 P_0 R + 2\lambda_2 P_0 V + 2\lambda_3 P_0 F)$$

$$P = P_0 + \lambda_1 P_0 R + \lambda_2 P_0 V + \lambda_3 P_0 F$$

$$P = P_0(1 + \lambda_1 R + \lambda_2 V + \lambda_3 F)$$

The sensitivity condition of the premium pricing is $\frac{d(f(R,V,F))}{dR} > 0, \frac{d(f(R,V,F))}{dV} > 0$ and $\frac{d(f(R,V,F))}{dF} > 0$. The claim fusion system artificial model becomes the R and P.

$$R = \frac{1}{1 + \exp\left(-\left[\sum_{i=1}^{n} W_{H,i} H_i + \sum_{j=1}^{m} W_A A_j + b\right]\right)}$$

$$P = P_0(1 + \lambda_1 R + \lambda_2 V + \lambda_3 F)$$

Appendix B. The continuous-time stochastic modelling of risk in CFS-AI Model

The stochastic differential equation for X_t is defined as latent risk variable $X_t \in \mathcal{R}, t \geq 0$. It provides the risk intensity. Assume the SDE for X_t.

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dW_t$$

Where dt is the small-time increment, $dW_t \sim N(0, dt)$ and X_t evolves randomly over time t. Now mapping the latent risk to observable risk.

$$\sigma(x_i) = R_t = \frac{1}{1 + e^{-x_i}}$$

Apply the Ito's lemma of R_t of the equation

$$R_t = f(X_t)$$

$$dR_t = f'(X_t)dX_t + \frac{1}{2}f''(X_t)(dX_t)^2$$

Differentiate the w.r.t. X_t of the equation, get

$$f'(X_t) = \frac{e^{-x_i}}{(1 + e^{-x_i})^2}$$

Now simplify the above equation.

$$f'(X_t) = \left(\frac{1}{1 + e^{-x_i}}\right)\left(1 - \frac{1}{1 + e^{-x_i}}\right)$$

Again differentiate w.r.t. X_t of the equation ().

$$f''(X_t) = \left(\left(\frac{1}{1 + e^{-x_i}}\right)\left(1 - \frac{1}{1 + e^{-x_i}}\right)\right)\left(1 - \frac{1}{1 + e^{-x_i}}\right) - \frac{1}{1 + e^{-x_i}}\left(\left(\frac{1}{1 + e^{-x_i}}\right)\left(1 - \frac{1}{1 + e^{-x_i}}\right)\right)$$

After applying the product rule,

$$f''(X_t) = \left(\left(\frac{1}{1 + e^{-x_i}}\right)\left(1 - \frac{1}{1 + e^{-x_i}}\right)\right)\left(1 - 2\left(\left(\frac{1}{1 + e^{-x_i}}\right)\right)\right)$$

Now compute the $(dX_t)^2$ from stochastic calculus

$$(dX_t)^2 = \sigma^2(X_t, t)dt$$

Substitute into Ito's lemma in above equation

$$dR_t = f'(X_t)dX_t + \frac{1}{2}f''(X_t)\sigma^2(X_t, t)dt$$

Substitute $dX_t = \mu dt + \sigma dW_t$, the first term will become

$$f'(X_t)dX_t = \left(\frac{1}{1+e^{-x_i}}\right)\left(1-\frac{1}{1+e^{-x_i}}\right)(\mu dt + \sigma dW_t)$$

$$f'(X_t)dX_t = \left(\left(\frac{1}{1+e^{-x_i}}\right)\left(1-\frac{1}{1+e^{-x_i}}\right)\right)\mu dt + \left(\left(\frac{1}{1+e^{-x_i}}\right)\left(1-\frac{1}{1+e^{-x_i}}\right)\right)\sigma dW_t$$

The second term become

$$\frac{1}{2}f''(X_t)\sigma^2(X_t,t)dt = \frac{1}{2}\left(\left(\frac{1}{1+e^{-x_i}}\right)\left(1-\frac{1}{1+e^{-x_i}}\right)\right)\left(1-2\left(\frac{1}{1+e^{-x_i}}\right)\right)\sigma^2(X_t)dt$$

Combines the first and second terms of the above equations

$$dR_t = \left(\left(\frac{1}{1+e^{-x_i}}\right)\left(1-\frac{1}{1+e^{-x_i}}\right)\right)\left\{\mu + \frac{1}{2}\left(1-2\left(\frac{1}{1+e^{-x_i}}\right)\right)\right\}dt + \left(\left(\frac{1}{1+e^{-x_i}}\right)\left(1-\frac{1}{1+e^{-x_i}}\right)\right)\sigma(X_t,t)dW_t$$

Appendix C. Machine learning models in claim prediction

Let us make the setup for all machine learning models, a basic true function of the model estimation.

$$D_n = \{(X_i, Y_i)\}_{i=1}^n \forall Y_i \in \{0,1\} \text{ and } f(x) = \mathbb{E}(Y|X=x)$$

The first selected model is random forest. The random forest equation for the claim fusion system artificial intelligence.

$$D_n = \{(X_i, Y_i)\}_{i=1}^n \forall Y_i \in \{0,1\} \text{ and } \mathbb{R}^{\mathrm{d}} = \bigcup_{j=1}^{m} R_j$$

Prediction in region R_j: $\hat{Y} = C_j$. This classifies the Gini index for splitting criterion with proportion class i and considers the bootstrap aggregation (bagging) $T_n(x, \Phi_i)$.

$$G = 1 - \sum_{i=1}^{n} p_i^2$$

$$T_n(x,\Phi_i) = \min_i \left[2\sum_{i=1}^{n}\left(Y_i - \hat{Y}\left(\frac{1}{1+\exp\left(-\left[\sum_{i=1}^{n} W_{H,i}H_i + \sum_{j=1}^{m} W_A A_j + b\right]\right)}\right)\right)^2\right]$$

At each split of criteria of the random feature selection is $j \ll d$. Then the random forest equation becomes:

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n} T_n(x,\Phi_i)$$

Substitute equation () in equation () get the random forest of the claim prediction.

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\min_i \left[2\sum_{i=1}^{n}\left(Y_i - \hat{Y}\left(\frac{1}{1+\exp\left(-\left[\sum_{i=1}^{n} W_{H,i}H_i + \sum_{j=1}^{m} W_A A_j + b\right]\right)}\right)\right)^2\right]$$

After that, selected XGBoost for the powerful method for controlling the high predictive accuracy. Let us consider the additive expansion of the loss function.

$$L = \sum_{i=1}^{n} l\left(Y_i, F(x_i)\right)$$

Compute the gradient descent for the function space S_i.

$$S_i = \frac{\partial L}{\partial F(x_i)} = \frac{\partial}{\partial F(x_i)}\left(\sum_{i=1}^{n} l\left(Y_i, F(x_i)\right)\right)$$

Fit the weak learner and update the XGBoost model

$$F_k(x_i) = F_{k-1}(x_i) + \eta\gamma_k S_i$$

The above model will give us the gradient direction ensures descent.

$$F(x_i) = \sum_{i=1}^{n}\sum_{k=1}^{K}\eta\gamma_k(S_i)$$

$$F(x_i) = \sum_{k=1}^{K}\eta\gamma_k\left(\sum_{i=1}^{n} l\left(Y_i, F(x_i)\right)\right)$$

Here considers the claim approval probability is equal to the odds function of the claim application.

$$W^T X + b = \frac{C_p}{1 - C_p}$$

Take logarithm odds by using logit transformation of equation.

$$W^T X + b = \log\left(\frac{C_p}{1 - C_p}\right)$$

$$X = \frac{\log\left(\frac{C_p}{1 - C_p}\right) - b}{W^T}$$

Take exponential both side of the equation.

$$e^X = \exp\left\{\frac{log\left(\frac{C_p}{1 - C_p}\right) - b}{W^T}\right\}$$

$$e^X = e^{\log\left(\left(\frac{C_p}{1-C_p}\right)-b\right)} - e^{W^T}$$

$$e^X + e^{W^T} = \left(\frac{C_p}{1 - C_p}\right) - b$$

$$e^X + e^{W^T} = \left(\frac{\frac{1}{1 + e^{-(\alpha R - \beta F)}}}{1 - \frac{1}{1 + e^{-(\alpha R - \beta F)}}}\right) - b$$

Taking the log likelihood of the equation.

$$l(W, b) = \sum_{i=1}^{n} \left(Y_i \log\left(C_p\right) + (1 - Y_i) \log\left(1 - C_p\right)\right)$$

Substitute the C_p.

$$l(W, b) = \sum_{i=1}^{n} \left(Y_i \log\left(\frac{1}{1 + e^{-(\alpha R - \beta F)}}\right) + (1 - Y_i) \log\left(1 - \frac{1}{1 + e^{-(\alpha R - \beta F)}}\right)\right)$$

The equation () is called as the sigmoid function, and the sigmoid function represents the logistic regression equation.
The Bayesian modelling of claim fusion system model

$$p(D_n|\lambda) = \frac{p(D_n|\lambda)p(\lambda)}{p(D_n)}$$

Apply the likelihood for binary classification of the equation.

$$p(D_n|\lambda) = \prod_{i=1}^{n} \left(C_p\right)^{Y_i} \left(1 - C_p\right)^{1 - Y_i}$$

Here, the consider prior distribution is $\lambda \sim N(0, \Sigma)$ and posterior distribution is $p(D_n|\lambda) \propto p(D_n|\lambda)p(\lambda)$. Apply logarithm to posterior distribution.

$$\log\left(p(D_n\lambda)\right) = \log\left(p(D_n|\lambda)\right) + \log\left(p(\lambda)\right)$$

Finally, optimize the equation.

$$\lambda^* = \max_{i=n} \left\{\sum_{i=1}^{n} Y_i \log\left(C_p\right) + (1 - Y_i) \log\left(1 - C_p\right) - \frac{1}{2}\lambda^T \Sigma^{-1} \lambda\right\}$$

$$\lambda^* = \max_{i=n} \left\{\sum_{i=1}^{n} Y_i \log\left(\frac{1}{1 + e^{-(\alpha R - \beta F)}}\right) + (1 - Y_i) \log\left(1 - \frac{1}{1 + e^{-(\alpha R - \beta F)}}\right) - \frac{1}{2}\lambda^T \Sigma^{-1} \lambda\right\}$$