Article

A Multimodal Artificial Intelligence Framework for Intelligent Mortgage Risk Assessment Using Underwriting and Loan-Performance Information

Amit Kurchania [1,*]

[1] Digital Solutions, Marlabs LLC, New York, NY 10006, USA.
* Correspondence: Amit.Kurchania@Marlabs.com (A.K.)

Abstract

Mortgage risk models often rely solely on underwriting variables, with unclear targets and evaluation protocols. This study proposes a multimodal artificial intelligence framework using separate encoders for financial and loan-performance features from the same mortgage dataset, adaptive fusion and SHAP explanations. Experiments use Fannie Mae Single-Family Loan Performance Data (2007Q1; 206,914 loans; 4,695 defaults, 2.27%). Default is defined as 90-day delinquency or credit-event exit within 12 months after loan age 12 months, excluding already delinquent loans and non-adverse exits. Evaluation includes stratified 70/15/15 splitting, train-only preprocessing, 10 repeated splits with overlap-corrected inference, loan-disjoint later-age testing and external validation on 377,697 loans from 2018Q1 (0.41% defaults), approximately eleven years later. Compared with Logistic Regression, Random Forest, XGBoost and MLP, the framework achieves ROC-AUC 0.840 ± 0.008 and PR-AUC 0.147 ± 0.011 (random baseline 0.023), with corrected 95% confidence intervals of 0.831–0.850 and 0.133–0.160, respectively. ROC-AUC reaches 0.791 on later-age testing and 0.824 externally, below Random Forest (0.839). Multimodal learning outperforms financial-only (0.797) and loan-performance-only (0.747) models. However, corrected comparisons show adaptive fusion is statistically comparable to concatenation, fixed-weight fusion, Logistic Regression, XGBoost and MLP, outperforming only Random Forest in ROC-AUC on the random split. Class balancing and prevalence shifts require probability recalibration. SHAP highlights FICO score, state, loan purpose and combined LTV. Ethical handling, transparency and human oversight remain essential.

Keywords: Artificial Intelligence, Mortgage Risk Assessment, Multimodal Learning, Loan-Performance Data, Credit Risk, Machine Learning, Explainable AI.

Citation: Kurchania, A. A Multimodal Artificial Intelligence Framework for Intelligent Mortgage Risk Assessment Using Underwriting and Loan-Performance Information. *Impact in Computics* **2026**, *2*, 9. https://doi.org/10.65500/computics-2026-009

Received: 9 September 2026 | Revised: 25 October 2026 | Accepted: 4 November 2026 | Published: 7 November 2026

Copyright: © 2026 by the authors. Licensee Impaxon, Malaysia. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Home mortgages are an integral part of the contemporary monetary system, which makes it possible for households and individuals to buy homes using financial support that is to be paid back over a prolonged duration. Before a lender decides to grant a mortgage, they will have to consider a number of aspects related to the risk of the borrower. Most mortgage risk assessment has been based on financial factors, which include income, employment status, credit history, debt to income ratio, debt obligations, loan repayment history, debt amount and home characteristics. Mortgage markets are becoming increasingly digital, demanding many more transactions that must process a vast amount of information, and adapt to ever shifting financial and economic conditions. In this light, the use of Artificial Intelligence (AI) and Machine

Learning (ML) has proved to be a potential solution to improve the efficiency, uniformity, and forecasting power of mortgage risk assessment [1], [2], [3].

Artificial Intelligence (AI) based mortgage assessment systems can leverage all aspects of the borrower's data and recognize trends and patterns that are associated with repayment issues, mortgage default or mortgage delinquency [4]. Credit scoring and financial risk prediction has been a field where machine learning algorithms like Logistic Regression, Random Forest, Gradient Boosting, XGBoost and Artificial Neural Networks have been widely used [5]. Further, more advanced deep learning models can be designed to capture the nonlinear interactions and sequence in financial behavior. Many of the current mortgage risk assessment models, however, continue to rely on a narrow set of financial data, structured in a certain way. Financial variables are basic elements in assessing repayment ability, but cannot necessarily fully explain a borrower's financial behaviour, stability or evolving risk situation. As digital information has flooded in, there is more opportunity to diversify and supplement traditional financial information with other forms of data. [6]. To address this challenge, the use of Multimodal Artificial Intelligence is explored, where learning and integration of complementary representations from multiple sources of information are possible. In a multimodal mortgage risk assessment system, modality-specific financial data can be preprocessed and features can be learned for each of the modalities first. These representations can then be combined using a feature fusion mechanism to build up a more detailed borrower risk profile. This could be a way to capture traditional indicators of financial capacity and extra behavioral signals of financial stability. The integrated representation can then be used to make mortgage risk predictions, such as knowing who has a relatively high likelihood of having a hard time paying back the mortgage or defaulting on it [4].

While AI lending offers numerous benefits, predictive performance is not enough for impactful mortgage decisions. There are privacy, quality, representativeness, algorithmic bias and transparency concerns arising from such uses of alternative and heterogeneous data [2]. There can also be difficulties in comprehending the individual influences of the various factors on a specific risk prediction in complex multimodal models. Thus, any intelligent mortgage assessment system needs to be both predictive and responsible in its use of information, explainable, and suitably supervised by human beings [7].

To address these challenges, this paper develops and rigorously evaluates a multimodal Artificial Intelligence (AI) framework for mortgage risk assessment. The framework learns financial (underwriting) information and loan-performance (non-underwriting behavioural and contextual) information, both taken from the same mortgage dataset, in separate encoders and combines them with an instance-wise adaptive fusion mechanism. It is compared with conventional machine learning, neural and single-modality baselines under a leakage-controlled protocol, and its explainability is demonstrated with SHAP and an analysis of the learned fusion weights.

Multimodal learning and information fusion have already been explored for credit risk, so the contribution of a new fusion model cannot be claimed on novelty of fusion alone. The studies reviewed in Section 2 differ from mortgage default assessment in three respects. First, most of them fuse numerical data with textual, vocal or social-media signals, or combine the outputs of several classifiers, whereas public loan-level mortgage data contain structured underwriting and loan-performance information only. Second, they address corporate, small-business, consumer or market-risk tasks rather than mortgage default. Third, none of the studies in Table 1 combines underwriting and behavioural loan-performance information on public mortgage data under an explicit prediction-date, leakage-controlled and externally validated protocol. This study addresses that gap, and it tests, instead of assuming, whether adaptive fusion adds value over simple concatenation and fixed-weight fusion.

The contributions of this study are therefore as follows. (i) A mortgage-specific multimodal framework in which underwriting and non-underwriting behavioural/contextual information are encoded separately and combined by a learned, instance-wise fusion network; the architecture, dimensions, scoring network and training procedure are fully specified (Section 3). The individual components (feed-forward encoders, softmax-weighted fusion, SHAP) are established techniques; the contribution lies in their composition for mortgage default and in the controlled test of whether adaptive fusion is beneficial. (ii) A leakage-controlled, reproducible evaluation protocol for loan-level default prediction with an explicit prediction date, observation and label windows, train-only preprocessing, repeated random splits with overlap-corrected confidence intervals, a loan-disjoint later-age snapshot test and a chronologically later external test on a different vintage (Section 4). (iii) An empirical analysis of modality complementarity, class-

imbalance handling, probability calibration and cost-sensitive decisions, together with SHAP and fusion-weight explanations (Section 4). (iv) A transparent account of limitations, including the finding that adaptive fusion matches, but does not exceed, strong baselines on the random split and is below Random Forest and the MLP in external ROC-AUC. The rest of the paper is structured as follows: Section 2 reviews related work on mortgage risk assessment, alternative data and multimodal learning; Section 3 introduces the methodology; Section 4 describes the data, protocol, results and limitations; and Section 5 concludes the study.

2. Related Work

Several studies have been conducted in this research area, exploring different methodologies and approaches to address the existing challenges. Sriram [8] explored the application of AI and deep learning to consumer credit risk assessment, monitoring, and fraud prevention, with a focus on the promise of GANs and reinforcement learning, and regulatory and ethical issues. The potential of AI and ML in predicting mortgage loan defaults was explored by Hung and Lo [9] who found that the techniques are highly useful in enhancing the risk assessment and lending decisions of the borrowers. The study by Konovalova et al. [10] highlighted that, apart from the conventional expert-based and statistical credit assessment methods, AI and ML have improved efficiency, objectivity and predictive power. Ramakrishnan et al. [11] spoke about how AI and ML are being used to measure a borrower's creditworthiness, uncover patterns and automate lending decisions. Ensemble models are also flexible as shown by El Sayad et al. [12] who used them to predict mortgage risk along with three other models: ANN, CNN and Random Forest. Last but not least, Hasan [13] showed the potential application of AI systems in financial information to enhance credit risk prediction, efficiency in financial credit underwriting, and real-time monitoring of risks.

Raji et al. [14] conducted a review of the methods of AI in the field of credit scoring, noting that it can analyze data that are unstructured and non-traditional, and enhance financial inclusion. Chitturi [15] spoke about other data sources that can be used for creating better borrower risk profiles, such as mobile payments, utility records, social media, and digital transactions. In addition to traditional information, Zhang [16] highlighted the importance of non-traditional information to complement conventional financial information and enhance credit assessment through information. Di Maggio and Ratnadiwakara [17] showed that alternative data such as education and employment data can enhance the creditability of the borrower and still be effective at assessing the risk. Barykin et al. [18] created a hybrid model for the risk of the mortgage portfolio that combines traditional risk factors (PD, LGD, EAD) with digital behavioral indicators, showing that the model performs better than traditional risk assessment models in prioritizing risks. Last, but not least, Xu et al. [19] demonstrated that the inclusion of rent, utility, and telecommunication payment data had a positive impact on the availability of credit services for credit invisible borrowers without a significant increase in default risk.

To address the personal credit risk assessment with uncertainty, Wang et al. [20] proposed the MIFCA framework, which combines multiple machine learning algorithms by multi-source information fusion. Gao and Lau [21] showed that deep learning models using multimodal social media signals greatly outperformed the traditional quantitative features in credit risk prediction. Lu et al. [22] proposed a multimodal credit risk prediction approach based on an extended deep belief network that integrates numerical, categorical and textual data for corporate credit risk, and reported that it outperformed benchmark models.

2.1. Critical comparison with multimodal credit-risk studies

The multimodal studies above are close in spirit to the present work, but they differ from it in data, task and fusion level, which limits their direct transfer to mortgage default prediction. Wang et al. [20] fuse the outputs of six base classifiers with Dempster-Shafer evidence theory on a commercial-bank personal-credit data set; this is decision-level fusion of classifiers trained on a single tabular source, so it neither learns modality-specific representations nor quantifies how much each information source contributes. Gao and Lau [21] depend on social-media signals, which are not part of public mortgage records. Lu et al. [22] and Fang [23] use textual information (BERT-based or financial-text encoders) next to numerical data, and Shen et al. [24] build on a large language model with textual SME data; these designs cannot be reproduced on public loan-level mortgage performance data, which contain no free text, and the language-model design adds considerable computational and governance cost. Cao et al. [25] forecast financial risk from textual, vocal and time-series information, which is a different target from borrower-level default. In addition, the comparisons reported in these studies rest on their own data sets and risk

definitions, so they cannot be used to rank methods for mortgage default.

2.2. Closest related work and research gap

The closest study in protocol is the recent Fannie Mae analysis of Hu et al. [26], which stresses leakage control, strict temporal splits and downsampling of the non-default class, but compares single-modality tabular models (logistic regression, random forest, gradient boosting and AutoML) without modality-specific fusion; its test AUROC values range from about 0.72 to 0.82 (about 0.81 for XGBoost and 0.82 for AutoGluon; data from 2023Q1 to 2024Q4) and give a useful reference range for Section 4. Taken together, the literature leaves open how structured underwriting information and behavioural loan-performance information can be combined in a learned latent space with transparent modality weights, and how such a model behaves under a leakage-controlled, time-aware and externally validated evaluation. Table 1 summarises the most relevant studies with their data, methods, reported evaluation and limitations relative to this study.

Table 1. Most relevant related studies: data, methods, reported evaluation, and limitations relative to this study

Ref.	Task	Data / modalities	Method	Reported evaluation	Limitation / gap relative to this study
El Sayad et al. [12]	Mortgage risk prediction	Mortgage loan data (study set in the Canadian housing-market context)	ANN, CNN, RF, XGBoost comparison	CNN validation accuracy 0.912-0.914; ensemble models (RF, XGBoost) retained a slight edge; XGBoost adaptable under market volatility	Single-modality tabular models; no fusion of behavioural information
Barykin et al. [18]	Mortgage portfolio risk prioritisation	Traditional risk factors (PD, LGD, EAD) and digital behavioural indicators	Hybrid composite model	Better risk prioritisation than traditional assessment	Portfolio-level prioritisation rather than loan-level default probability
Wang et al. [20]	Personal credit risk (multi-class)	Real data set of a Chinese commercial bank (sample size not stated in the abstract)	Dempster-Shafer fusion of six base classifiers	Higher multi-class accuracy than the single base classifiers (values not stated in the abstract)	Decision-level fusion of one tabular source; not mortgage; no learned modality weights
Gao and Lau [21]	Microlending credit risk	Real-world client data of a listed microlending firm; financial data and multimodal social-media signals	Deep learning	Accuracy 26.07% higher than the same model with traditional quantitative features alone	Requires social-media data, unavailable in mortgage records
Lu et al. [22]	Corporate credit risk	Numerical, categorical and textual data	Extended deep belief network (ResNet, embeddings, BERT)	Outperformed benchmark models	Text-dependent; corporate rather than mortgage default
Fang [23]	Credit risk forecasting	Financial text, numerical and time-series data	Cross-modal attention fusion	Improved risk identification and early warning	Text-dependent system design; not mortgage
Shen et al. [24]	SME credit risk	Financial, non-financial and textual data	LLM with LoRA	F1-score of 0.829	Text and LLM required; SME rather than mortgage; high compute and governance cost
Cao et al. [25]	Financial risk forecasting	Text, vocal and time-series data	LLM-based multimodal	Multi-source data improved risk forecasting	Market/corporate risk rather than borrower default; audio/text inputs

Ref.	Task	Data / modalities	Method	Reported evaluation	Limitation / gap relative to this study
			framework (RiskLabs)		
Hu et al. [26]	Mortgage default (Fannie Mae, 2023-2024)	Fannie Mae loan-level tabular data, 2023Q1-2024Q4	LR, RF, XGBoost, LightGBM, AutoGluon	Test AUROC about 0.72-0.82 under a strict temporal split (XGBoost about 0.81; AutoGluon 0.823 at 1:2 downsampling)	Tabular models only; no modality-specific fusion
This study	Mortgage default (Fannie Mae 2007Q1: 206,914 loans, 2.27% defaults; external 2018Q1: 377,697 loans, 0.41% defaults)	Underwriting variables plus loan-performance and contextual variables (same dataset)	Multimodal network with adaptive fusion; SHAP	ROC-AUC 0.840, PR-AUC 0.147 (random split); 0.791 (later-age snapshot); 0.824 (external)	Single-vintage training; no independently acquired alternative data; adaptive fusion not better than strong baselines

Table 1 lists only the studies most relevant to the present work; the broader literature on AI-based credit assessment and alternative data is discussed in the paragraphs above. Dataset information and quantitative results are quoted as reported by the cited studies; they cannot be compared across studies because data, targets and metrics differ.

3. Proposed Methodology

The framework (Figure 1) learns modality-specific representations of underwriting information and of non-underwriting behavioural and contextual information, fuses them with an instance-wise adaptive mechanism and predicts the probability of a mortgage default event. This section defines the data modalities, the architecture and its mathematical formulation, the training objective and the algorithm; the experimental protocol is given in Section 4.

3.1. Proposed Multimodal AI Framework

The proposed Multimodal Artificial Intelligence (MMAI) framework combines two modalities. The financial modality X_f contains underwriting variables known at origination: the borrower and co-borrower credit scores, debt-to-income (DTI) ratio, original and combined loan-to-value ratios (LTV, CLTV), original unpaid principal balance (log-transformed), original interest rate, loan term, mortgage-insurance percentage, number of borrowers and units, first-time-buyer flag, loan purpose, property type and occupancy status (30 input columns after encoding). The loan-performance modality X_a (labelled "alternative" in the figure legends) contains non-underwriting behavioural and contextual information from the same mortgage dataset: the payment behaviour observed in the months up to the prediction date (number of observed months, maximum delinquency status, number of delinquent months, months since the last delinquency, modification flag), the loan age and unpaid-principal paydown ratio at the prediction date, and contextual attributes (origination channel, seller, servicer and property state), giving 79 input columns after encoding.

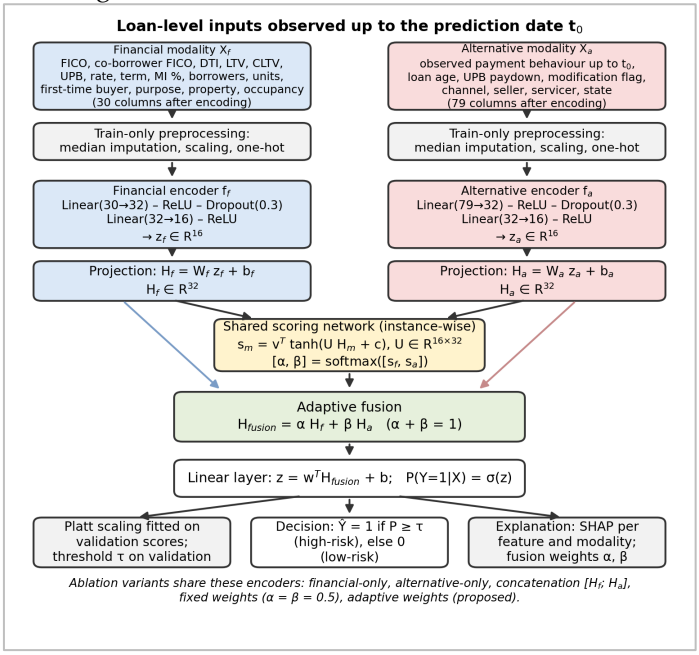


Figure 1. Architecture of the multimodal framework comprising modality-specific encoders (32 hidden and 16 representation units), 32-dimensional projections, a shared scoring network (16 hidden units), instance-wise softmax fusion, sigmoid-based linear prediction, and post-hoc calibration, thresholding, and explanation. Ablation variants retain the same encoders. "Alternative" denotes the loan-performance modality.

Public loan-level mortgage data contain no independently acquired alternative sources such as rent, utility or transaction records. Both modalities are therefore derived from the same Fannie Mae dataset: the second modality consists of loan-performance and contextual information outside the underwriting decision (it is labelled "alternative" in the figure legends), and a loan's own payment history is used only up to the prediction date, never beyond it. We do not claim that this modality is independent of the financial one; its added value is tested empirically in the ablation of Section 4.4. Features such as income, employment status or savings behaviour mentioned in the generic mortgage literature are not part of the public data and are not used.

Each modality passes through its own preprocessing pipeline (median imputation with missing-value indicators, standardisation of numerical features and one-hot encoding of categorical features, with levels below 0.5% grouped as "infrequent"), which is fitted on the training data only. The processed financial and loan-performance data are fed to two separate encoders (Eqs. 2-3). Their outputs are projected to a common dimension (Eq. 4), scored by a shared scoring network (Eq. 5) and fused with learned, instance-wise weights (Eqs. 6-8). The fused representation is passed to a linear layer with a sigmoid output (Eqs. 9-10), and the decision threshold is chosen on validation data (Eq. 11). Adaptive weighting lets the model learn how much each modality contributes for each borrower, which is useful only if the reliability of the two modalities differs between borrowers; Section 4.4 tests this empirically. Figure 1 shows the single architecture that was implemented and evaluated.

3.2. Mathematical Formulation

Let the input information for a loan at the prediction date be represented as:

$$X = \{X_f, X_a\} \tag{1}$$

where X_f is the financial data and X_a is the loan-performance data. The two modalities are processed independently to preserve their individual characteristics. Each encoder is a feed-forward network with one hidden layer:

$$z_f = f_f(X_f; \theta_f) \tag{2}$$

$$z_a = f_a(X_a; \theta_a) \tag{3}$$

where z_f and z_a are the learned representations, $f_f(\cdot)$ and $f_a(\cdot)$ are the encoders, and θ_f and θ_a their trainable parameters. Each encoder is Linear($d_{in} \rightarrow h$) – ReLU – Dropout(p) – Linear($h \rightarrow h/2$) – ReLU, with d_{in} = 30 for X_f and 79 for X_a; the selected values for the proposed model are h = 32, p = 0.3, learning rate 0.0021 and weight decay 1e-03 (Section 4.2). Because the two encoder outputs may have different dimensions in general, each is first mapped to a common space of dimension D = 32:

$$H_m = W_m^p z_m + b_m^p, \qquad m \in \{f, a\} \tag{4}$$

where $W_m^p \in R^{D \times h/2}$ and $b_m^p \in R^D$ are learned projection parameters and $H_m \in R^D$ is the modality representation. The importance score of each modality is produced by a scoring network that is shared by both modalities:

$$s_m = v^\top \tanh(U H_m + c), \qquad m \in \{f, a\} \tag{5}$$

where $U \in R^{16 \times D}$, $c \in R^{16}$ and $v \in R^{16}$ are learned parameters (16 hidden units) and s_m is a scalar score computed separately for every borrower. The fusion weights are obtained by softmax normalisation:

$$\alpha = \frac{\exp(s_f)}{\exp(s_f) + \exp(s_a)} \tag{6}$$

$$\beta = \frac{\exp(s_a)}{\exp(s_f) + \exp(s_a)} = 1 - \alpha \tag{7}$$

so that α, $\beta > 0$ and $\alpha + \beta = 1$ for every borrower. The fused representation is the weighted combination:

$$H_{fusion} = \alpha H_f + \beta H_a \tag{8}$$

Two reference fusion rules are used in the ablation: simple concatenation, $H_{fusion} = [H_f; H_a]$, and fixed-weight fusion, $\alpha = \beta = 0.5$. The prediction layer computes the logit and the risk probability:

$$z = w^\top H_{fusion} + b \tag{9}$$

$$p_{\text{raw}} = \sigma(z) = \frac{1}{1 + \exp(-z)} \tag{10}$$

where $w \in R^D$ and b are the learnable weight vector and bias, $\sigma(\cdot)$ is the sigmoid function of the logit z (not of the input X), and p_{raw} is the raw sigmoid output of the network (a risk score). It is not a calibrated probability, because the training set is class-balanced (Section 4.2). The calibrated probability P_{cal} is obtained from p_{raw} by Platt scaling, a one-dimensional logistic regression fitted on the raw scores of the validation set, so that it corresponds to the natural default rate; P_{cal} is the predicted probability that the loan reaches the adverse outcome defined in Section 4.1. Platt scaling is monotone and does not change the ranking of loans. The predicted risk class is obtained by applying a decision threshold τ to P_{cal}:

$$\hat{Y} = \begin{cases} 1, & P_{\text{cal}} \geq \tau \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where $\hat{Y} = 1$ denotes the high-risk class and $\hat{Y} = 0$ the low-risk class. The threshold τ is selected on the validation set after Platt scaling, either to maximise F1 or to minimise an expected misclassification cost (Section 4.2); it is never selected on test data. Eq. 11 is thus applied to the Platt-

calibrated probability P_{cal}, never to the raw sigmoid output p_{raw}. The parameters are learned by minimising the weighted binary cross-entropy:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[\omega\, y_i \log p_i + (1-y_i)\log(1-p_i)]\,, \qquad \omega = \left(\frac{n_0}{n_1}\right)^{0.5} \tag{12}$$

where N is the number of training loans, $p_i = \sigma(z_i)$ is the raw sigmoid output (Eq. 10) for loan i, and n_0 and n_1 are the numbers of non-default and default loans in the training set. The weight ω corrects for class imbalance; it equals one when the training set is balanced 1:1 (Section 4.2).

In the primary experimental configuration, class imbalance was addressed through random undersampling of the non-default class to obtain a balanced 1:1 training distribution. Consequently, the class-weighting coefficient (ω) was set to 1, and no additional class weighting was applied during model optimization. The primary model is therefore trained with 1:1 undersampling and unit class weights, and Eq. 12 reduces to the standard binary cross-entropy; undersampling and class weighting are not combined. The square-root weighting $\omega = (n_0/n_1)^{0.5}$ and the full inverse-frequency weighting $\omega = n_0/n_1$ were used only in the imbalance comparison of Section 4.7, on the full training set without undersampling. The validation and test datasets retained their original class distributions to ensure unbiased performance evaluation under the natural class prevalence.

3.3. Algorithm

Algorithm 1 gives the complete procedure. The order of its steps is essential to avoid information leakage: the loan-level data set is constructed with an explicit prediction date, the data are split first, and all preprocessing transformations are fitted on the training portion only and then applied to the validation and test portions.

Algorithm 1. **Leakage-controlled multimodal mortgage risk assessment**

Input: Loan-month records R; history K = 12; horizon H = 12; delinquency threshold d = 3 (90+ days); cost ratio r.

Output: Risk class $\hat{Y}$; calibrated probability P_{cal}; explanation E.

1. Construct loan-level dataset

For each loan, set t_0 = first reporting month + K. Use only records at $t \le t_0$. Set y = 1 for delinquency ≥ d or credit-event exit within $(t_0, t_0 + H]$; otherwise y = 0. Exclude missing t_0 records, loans already delinquent at t_0, and censored non-adverse exits.

2. Split data

Stratify loans into train/validation/test (70/15/15) before preprocessing.

3. Preprocess

Fit imputation, standardisation and encoding on training data only for X_f and X_a; apply fitted transformations to validation/test data.

4. Balance

Randomly undersample training non-defaults to 1:1; retain natural prevalence in validation/test. Hence $\omega = 1$ in Eq. 12.

5. Tune

Evaluate six hyperparameter configurations per model using validation PR-AUC of run 0 only; reuse the selected configuration unchanged in all other runs.

6. Train multimodal model

For each mini-batch:

- $z_f \leftarrow \text{FinancialEncoder}(X_f)$
- $z_a \leftarrow \text{LoanPerformanceEncoder}(X_a)$
- $H_m \leftarrow \text{Projection}(z_m)$, $m \in \{f, a\}$
- $s_m \leftarrow \text{ScoringNetwork}(H_m)$
- $(\alpha, \beta) \leftarrow \text{Softmax}(s_f, s_a)$
- $H_{fusion} \leftarrow \alpha H_f + \beta H_a$
- $p_{raw} \leftarrow \text{Sigmoid}(w^T H_{fusion} + b)$

Compute weighted cross-entropy ($\omega = 1$ for the 1:1 balanced training set)

Update parameters using AdamW

End loop. Apply early stopping after eight epochs without validation PR-AUC improvement.

7. Calibrate and select threshold

Fit Platt scaling on validation scores to map p_{raw} to P_{cal}. Select τ by maximum validation F1 or minimum expected cost under ratio r.

8. Predict

For each test loan, calculate P_{cal} from p_{raw}.

If $P_{cal} \ge \tau$, assign $\hat{Y}$ = High-Risk; otherwise $\hat{Y}$ = Low-Risk.

9. Evaluate

Compute ROC-AUC, PR-AUC, MCC, balanced accuracy, precision, recall, F1, confusion matrix, Brier score, ECE and cost per 1,000 loans. Repeat across 10 random splits with new initialisations. Report mean, SD and overlap-corrected 95% CIs using the corrected resampled t-test and Holm adjustment.

Test 2007Q1-trained models on a loan-disjoint later-age snapshot and strictly later 2018Q1 loans.

10. Explain

Compute SHAP values; aggregate by feature and modality. Analyse fusion weights α and β by risk group.

11. Return $\hat{Y}$, P_{cal} and E.

4. Results and Discussion

4.1. Data, Target Definition and Cohort

4.1.1. Data source

The experiments use the Fannie Mae Single-Family Loan Performance Data [27], a public data set of historical U.S. residential mortgages (30-year and shorter, fully amortising, full-documentation, conventional fixed-rate loans acquired by Fannie Mae). Access requires free registration with Fannie Mae. The combined acquisition and performance file of the 2007Q1 acquisition quarter was used for training, validation and testing: 16,588,215 loan-month records of 253,279 loans first reported between January and March 2007. The file of the 2018Q1 acquisition quarter (23,563,399 records, 458,830 loans) was used only as an external test set and was never used for training,

tuning or threshold selection. After the cohort exclusions described in Section 4.1.3 (18,908 loans without a valid record at the prediction date, 4,293 already delinquent and 57,932 censored), 377,697 eligible 2018Q1 loans remain. Both files were accessed in October 2026 and are the combined acquisition-and-performance files of the respective acquisition quarters (one row per loan-month; the 2007Q1 file was used in Parquet format with 110 columns, the 2018Q1 file as a pipe-separated text file). Each record contains the loan identifier, reporting period, credit score, DTI, LTV, unpaid principal balance, interest rate, term, property and geographic attributes, delinquency status and zero-balance code.

4.1.2. Target definition

Each loan is observed at a prediction date t_0, defined as 12 months after its first reporting month (January to March 2008 for the 2007Q1 loans). The observation window contains all loan-month records up to and including t_0. The prediction horizon is the following 12 months. A loan is labelled as adverse (y = 1) if its delinquency status reaches 3 or more (90 or more days past due) in any month of the horizon, or if it exits through a credit event (short sale, third-party sale, real-estate-owned or note sale); otherwise it is labelled y = 0. Loans without a record at t_0 (16,057 loans that left the portfolio before 12 months), loans already delinquent at t_0 (5,717) and loans that left the portfolio within the horizon without an event, for example through prepayment (24,591) are excluded, so prepayment is treated as censoring (the implications of these exclusions for generalisability are discussed in the next paragraph and in Section 4.10). The resulting cohort contains 206,914 loans: 4,695 defaults (2.27%) and 202,219 non-defaults, a ratio of about 43 to 1. Prediction months are January (62,966 loans), February (71,702) and March 2008 (72,246).

4.1.3. Implications of the cohort exclusions

The cohort definition removes three groups of loans: loans without a record at the prediction date (16,057 of 253,279, 6.3%), which left the portfolio within 12 months; loans already delinquent at the prediction date (5,717, 2.3%); and loans that prepaid or otherwise left the portfolio within the 12-month window without an adverse event (24,591, 9.7%). In total 18.3% of the 2007Q1 loans are excluded (for 2018Q1, the corresponding three exclusions total 81,133 of 458,830 loans, 17.7%, as detailed below). For the 2018Q1 external cohort (458,830 loans), the same three exclusions as for 2007Q1 were applied: 18,908 loans lacked a valid record at the prediction date (t_0), which prevented the construction of the loan-level feature representation; 4,293 loans were already delinquent at t_0; and 57,932 loans were censored, that is, they left the portfolio within the 12-month horizon without an adverse event. Hence 458,830 − 18,908 − 4,293 − 57,932 = 377,697 eligible loans (1,542 defaults, 0.41%) were retained, and they form the external test set.

The first two exclusions define the target population: the model scores performing loans that are still observed at loan age 12 months, and it is not intended for loans that are already delinquent, for loans that exit early, or for scoring at origination. The third exclusion is different in kind, because whether a loan prepays inside the window is known only after the prediction date, so a lender could not apply this filter at scoring time. Prepaying borrowers are typically lower-risk, so removing them raises the default rate of the evaluation cohort (2.27%) above the rate that would result if they were retained as non-events, and it can change calibration and, to a smaller extent, ranking metrics. The absolute probabilities, PR-AUC values and cost figures reported here are therefore conditional on this cohort and should not be read as portfolio-level estimates; ROC-AUC values and the comparisons between models are likely to be less affected, but this was not verified, and a sensitivity analysis that retains prepaid loans as non-events is left to future work (Section 4.10).

4.1.4. Leakage control

Feature rows are restricted to $t \leq t_0$ and label rows to $t > t_0$; this separation is asserted in the code for every loan. Delinquency information observed after t_0 is never used as a feature. All preprocessing is fitted after splitting and on the training data only. As a sanity check, a model trained on randomly shuffled labels reached a test AUC of 0.471, as expected without leakage. The borrower payment history in the loan-performance modality therefore describes behaviour before the prediction date and is not an independent data source; Table 2 lists all features by modality.

Table 2 shows that the loan-performance (previously "alternative") modality contains behavioural and contextual information outside the underwriting decision. Median imputation with missing-value indicators handles missing values; categorical levels below 0.5% of the training data are grouped as "infrequent" and unseen levels are mapped to the same group.

Table 2. Dataset characteristics, feature categories and modality assignment (missing rates refer to the training set of run 0)

Modality	Category	Features	Type	Role
Financial	Borrower credit and capacity	Credit score (0.2% missing), co-borrower credit score (56.9% missing, mainly single-borrower loans), DTI (2.6% missing), number of borrowers, first-time-buyer flag	Numerical / categorical	Credit-risk assessment
Financial	Loan terms	Original UPB (log), interest rate, term, mortgage-insurance % (86.8% missing)	Numerical	Mortgage exposure
Financial	Collateral and purpose	LTV, CLTV, property type, number of units, occupancy, loan purpose	Mixed	Collateral risk
Loan-performance	Behavioural (up to t_0)	Observed months, maximum delinquency status, number of delinquent months, months since last delinquency, modification flag, loan age, UPB paydown ratio	Numerical	Payment behaviour before prediction date
Loan-performance	Contextual	Origination channel, seller, servicer, property state	Categorical	Distribution channel and geography
Target	Adverse outcome	90+ days delinquent or credit-event exit within 12 months after t_0	Binary	Default prediction

4.2. Experimental Setup

The cohort was split at loan level into stratified 70% training, 15% validation and 15% test sets, which avoids the ambiguity of a 70%-30% split followed by a further 50% validation split. Preprocessing was fitted on the training set and applied to the other sets. Table 3 reports the exact sizes. Ten repeated random splits were performed: each run draws a new stratified split of the same cohort and a new model initialisation (seed equal to the run index), so the reported standard deviations include both sources of variation. The runs are not independent samples, because the test sets of different runs overlap (each loan is in the test set of about 15% of the runs); the consequences for confidence intervals and significance tests are described under Metrics and calibration. Two additional protocols test generalisation. In the later-age snapshot protocol (called out-of-time in earlier versions), the models are trained on loans at age 12 months (prediction dates January to March 2008) and tested on different loans at age 24 months (prediction dates January to March 2009). The age-24 snapshot contains 159,537 eligible loans (8,438 defaults, 5.29%); in each run a random 30% of these loans, disjoint from the training and validation loans, forms the test set (47,885 loans with 2,495 defaults in run 0, Table 3). Because the test loans have 24 instead of 12 months of history, and the training outcome windows of some loans extend beyond the prediction dates of some test loans, this protocol combines a shift in loan age with a partial shift in calendar time and is not a strictly chronological split; Table 4 gives the exact dates. In the *external* protocol the models trained on 2007Q1 are tested on all 377,697 eligible 2018Q1 loans (1,542 defaults, 0.41%; 18,908 loans excluded for lacking a valid record at t_0, 4,293 excluded as delinquent at t_0 and 57,932 as censored, out of 458,830 loans). This protocol is strictly chronological: the last training outcome (March 2009) precedes the first external prediction date (January 2019), and the loan age at the prediction date is the same as in training (12 months).

Table 3. Cohort and split sizes in run 0 (number of loans, with number of defaults in parentheses)

Protocol	Training	Validation	Test
Random 70/15/15	144,839 (3,286)	31,037 (704)	31,038 (705)
Later-age snapshot (age 24 months)	123,153 (2,785)	21,734 (491)	47,885 (2,495)
External (2018Q1)	175,876 (3,991)	31,038 (704)	377,697 (1,542)

Note: The external test set comprises the 377,697 eligible 2018Q1 loans that remain after the cohort exclusions (458,830 − 18,908 − 4,293 − 57,932 = 377,697; Section 4.1.3). The external training and validation sets are drawn from the 2007Q1 cohort (175,876 + 31,038 = 206,914 loans).

Table 4 gives the prediction dates and outcome windows of each protocol. Within every loan, feature rows precede t_0 and label rows follow t_0, so no loan contributes future information to its own features (asserted in the code

for every loan). Between loans, the three protocols differ. In the random split, training and test loans share the same prediction dates, so it measures generalisation to unseen loans of the same cohort, not to a later time. In the later-age snapshot protocol, test loans are different from training loans, but the training outcome windows end between January and March 2009 and the test prediction dates begin in January 2009, so some training labels were observed after the prediction date of some test loans; this partial calendar overlap, together with the longer history of the age-24 loans, means that the protocol is not a strictly chronological evaluation. In the external protocol, the last training outcome (March 2009) precedes the first test prediction date (January 2019) and the loan age is the same, so only this protocol is strictly chronological.

Table 4. Chronology of training and test prediction dates (t_0) and 12-month outcome windows for each protocol

Protocol	Training: prediction dates (t_0) and outcome windows	Test: prediction dates (t_0) and outcome windows	Chronological relationship
Random 70/15/15	t_0 = Jan-Mar 2008 (loan age 12 months); outcomes observed through Jan-Mar 2009	Same prediction dates as training (random 15% of the same loans); outcomes through Jan-Mar 2009	None: random loan-level split of one cohort
Later-age snapshot (age 24 months)	t_0 = Jan-Mar 2008 (age 12 months); outcomes observed through Jan-Mar 2009	Different loans; t_0 = Jan-Mar 2009 (age 24 months); outcomes Feb 2009-Mar 2010	Test t_0 is 12 months after training t_0, but training outcome windows end Jan-Mar 2009 and so extend beyond the t_0 of some test loans (Jan 2009): partial calendar overlap
External (2018Q1)	t_0 = Jan-Mar 2008 (2007Q1 loans, age 12 months); outcomes observed through Jan-Mar 2009	2018Q1 loans; t_0 = Jan-Mar 2019 (age 12 months); outcomes Feb 2019-Mar 2020	Strictly chronological: last training outcome (Mar 2009) precedes first test t_0 (Jan 2019); same loan age

4.2.1. Class imbalance

Defaults are rare (2.27%). The training set is balanced 1:1 by random undersampling of non-defaults, applied after the split and to the training data only (for example, 3,286 defaults and 3,286 randomly chosen non-defaults in run 0). Validation and test sets keep the natural class ratio. No synthetic oversampling is used. Because the balanced training set has $n_0 = n_1$, the class weight ω in Eq. 12 equals 1 in the primary configuration: the proposed model is trained with 1:1 undersampling and unit class weights, and undersampling and class weighting are not combined. Section 4.7 compares this treatment with no weighting, square-root and full inverse-frequency weighting ($\omega = 1$, $\omega = (n_0/n_1)^{0.5}$ and $\omega = n_0/n_1$, respectively, each on the full training set without undersampling).

4.2.2. Baselines and tuning

The proposed model is compared with Logistic Regression (LR), Random Forest (RF), XGBoost [28] and a multilayer perceptron (MLP) on the concatenation of all features, and with the ablation variants of Section 4.4. All models use the same features, splits and seeds, and the same tuning budget: six random hyper-parameter configurations per model (the first is the default), selected on validation PR-AUC of run 0 only: the search was carried out once, on the validation set of the run-0 split, and the selected configuration of each model was then reused unchanged in all ten runs, without re-tuning on the other splits. Because the splits overlap across runs, validation loans of run 0 reappear in the test sets of other runs, so the repeated-split scores may be slightly optimistic (Section 4.10.5), and the standard deviations over runs do not include the variability of hyper-parameter selection. Selected values were: LR C = 1.0; RF 200 trees, maximum depth 10, minimum leaf size 20, max_features = sqrt; XGBoost 200 trees, depth 4, learning rate 0.139, subsample 0.9, column sample 0.9, minimum child weight 5, L2 regularisation 5, early stopping on validation PR-AUC; MLP hidden size 128, dropout 0.4, learning rate 0.0031, weight decay 1e-05. The proposed model uses hidden size 32, dropout 0.3, learning rate 0.0021 and weight decay 1e-03.

4.2.3. Training configuration

Neural models are trained with AdamW [29], a batch size of 1,024, at most 60 epochs and early stopping with patience 8 on validation PR-AUC, using the weighted cross-entropy of Eq. 12 ($\omega = 1$ in the primary configuration). Experiments ran in Python 3.11.8 with PyTorch 2.7.0+cu118, scikit-learn 1.6.1, XGBoost 3.0.5, NumPy 2.1.3 and pandas 2.2.3 on a Windows machine with a CUDA GPU. Random seeds are set for splits and initialisation, but multi-threaded and GPU training is not bit-wise

deterministic, so repeated executions can differ slightly; all reported values come from one complete execution.

4.2.4. Metrics and calibration

Because the outcome is rare, ROC-AUC and PR-AUC (random-guess PR-AUC equals the default rate) are the primary measures [30]. Matthews correlation coefficient (MCC) [31], balanced accuracy, precision, recall and F1 are computed at a threshold chosen on validation data after Platt scaling [32]; calibration is measured by the Brier score and the expected calibration error (ECE) [33]; and the cost of false approvals and false rejections is evaluated for cost ratios of 5, 10, 20 and 50 [34]. Here "raw" denotes the sigmoid output p_{raw} of Eq. 10 and "calibrated" the Platt-scaled probability P_{cal} of Eq. 11: the raw ECE is computed from p_{raw}, whereas the Brier score, the calibrated ECE, the decision threshold, the class labels, the confusion matrices and the costs are computed from P_{cal}. ROC-AUC and PR-AUC depend only on the ranking of loans and are therefore the same for p_{raw} and P_{cal}, because Platt scaling is monotone. Results are reported as mean ± standard deviation over 10 repeated random splits. Because these runs share one cohort and their test sets overlap, the scores of different runs are positively correlated, and the ordinary Student t interval and paired t-test over runs understate the uncertainty. The 95% confidence intervals for ROC-AUC and PR-AUC and the paired comparisons therefore use the corrected resampled t-test of Nadeau and Bengio [35], in which the variance of the mean over k = 10 runs is multiplied by (1/k + n_test/n_train) / (1/k); this inflates the standard error by a factor of 1.77 for the random split (n_test/n_train = 31,038/144,839) and 2.21 for the later-age snapshot (47,885/123,153; the test loans are redrawn in every run). Holm's step-down adjustment [36] is applied across the 12 comparisons (six comparison models, two metrics). Wilcoxon signed-rank tests are not corrected for the overlap and are not used for inference. In the external test the test set is identical in all runs, so the intervals and tests there reflect training randomness only and do not include test-sampling uncertainty, which is addressed by a paired bootstrap of the test set (supplementary files).

4.3. Results and Baseline Comparison

Table 5 compares the proposed framework with the baselines on the held-out test sets of 10 repeated random splits, and Figure 2 shows the main metrics. Accuracy is not a meaningful measure here: predicting "low-risk" for every loan gives an accuracy of 97.7%, which is higher than that of any model (94.9% to 95.5%). The models should be judged by ranking and detection measures.

Table 5. Performance on the random-split test set (mean ± SD over 10 repeated splits, with overlap-corrected 95% confidence intervals for ROC-AUC and PR-AUC; threshold chosen on validation data)

Model	Accuracy	Precision	Recall	F1	MCC	ROC-AUC	PR-AUC	ROC-AUC 95% CI	PR-AUC 95% CI
Majority class (all low-risk)	0.977±.000	0.000	0.000	0.000	0.000	0.500	0.023	n/a	n/a
Logistic Regression	0.954±.004	0.170±.015	0.265±.023	0.207±.012	0.189±.012	0.839±.007	0.138±.007	0.830–0.848	0.129–0.147
Random Forest	0.952±.009	0.165±.024	0.265±.049	0.199±.011	0.184±.011	0.831±.006	0.130±.010	0.823–0.838	0.117–0.143
XGBoost	0.955±.007	0.182±.021	0.268±.042	0.213±.009	0.197±.010	0.842±.007	0.141±.007	0.833–0.852	0.132–0.149
MLP	0.952±.004	0.175±.011	0.294±.023	0.218±.007	0.203±.007	0.843±.007	0.148±.009	0.834–0.853	0.137–0.159
Proposed (adaptive fusion)	0.949±.006	0.165±.008	0.301±.039	0.212±.007	0.198±.010	0.840±.008	0.147±.011	0.831–0.850	0.133–0.160

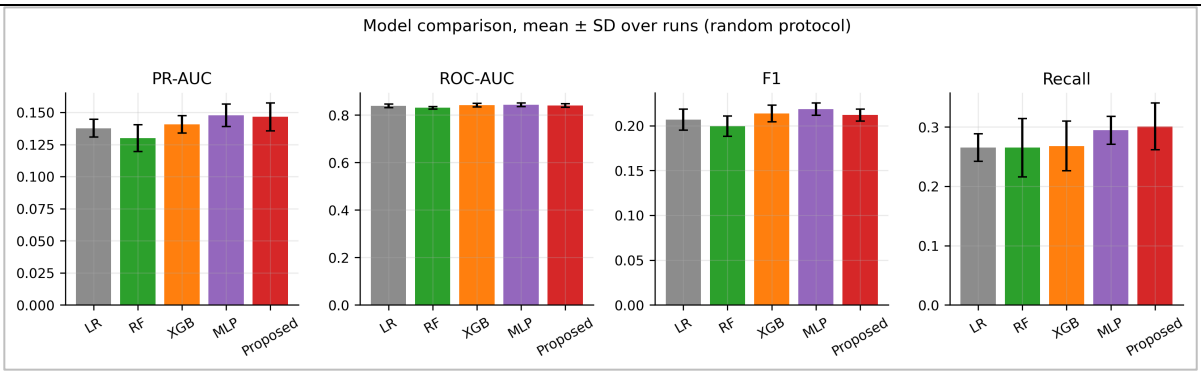


Figure 2. PR-AUC, ROC-AUC, F1 and recall of the baselines and the proposed model (mean ± SD over 10 runs).

The proposed model reaches a ROC-AUC of 0.840 ± 0.008 (corrected 95% CI 0.831-0.850) and a PR-AUC of 0.147 ± 0.011 (0.133-0.160), about 6 times the PR-AUC of random guessing (0.023), and its top-scored 5% of loans contain 33.7% of all defaults (lift 6.7). The MLP (ROC-AUC 0.843 ± 0.007) and XGBoost (0.842 ± 0.007) rank loans equally well within experimental noise, Logistic Regression is slightly lower (0.839 ± 0.007) and Random Forest is lowest (0.831 ± 0.006). Table 6 reports paired comparisons across the 10 repeated splits with the corrected resampled t-test and Holm adjustment for the 12 comparisons (six comparison models, two metrics). Only one difference remains significant: the proposed model has a higher ROC-AUC than Random Forest (+0.0099, corrected p = 0.0006, Holm-adjusted p = 0.007). The PR-AUC advantages over Logistic Regression (+0.0088) and Random Forest (+0.0166) are nominal (corrected p = 0.025 and 0.033; Holm-adjusted p = 0.25 and 0.29), and the lower ROC-AUC compared with the MLP (-0.0030) is not significant after adjustment (corrected p = 0.009, Holm-adjusted p = 0.099). The proposed model is therefore statistically indistinguishable from Logistic Regression, XGBoost, the MLP, concatenation and fixed-weight fusion, and the proposed framework does not outperform the baselines.

Table 6. Paired comparisons of the proposed model with each comparison model across 10 repeated random splits (corrected resampled t-test; Holm adjustment over 12 comparisons; difference = proposed minus comparison model)

Compared model	ΔROC-AUC	p (corrected)	p (Holm)	ΔPR-AUC	p (corrected)	p (Holm)
Logistic Regression	+0.0015	0.321	1.000	+0.0088	0.025	0.251
Random Forest	+0.0099	0.0006	0.007	+0.0166	0.033	0.294
XGBoost	-0.0019	0.140	1.000	+0.0058	0.344	1.000
MLP	-0.0030	0.009	0.099	-0.0013	0.587	1.000
Concatenation fusion	+0.0006	0.278	1.000	+0.0009	0.356	1.000
Fixed-weight fusion	+0.0005	0.369	1.000	+0.0004	0.715	1.000

p (corrected) is the p-value of the corrected resampled t-test and p (Holm) its Holm-adjusted value over the 12 comparisons; the uncorrected paired t-test and Wilcoxon p-values are given in the supplementary files table_stat_tests_across_runs and table_stat_tests_across_runs_external. A paired bootstrap on the test set of run 0 (supplementary file table_stat_tests_bootstrap_run0) quantifies the test-sampling uncertainty of a single split.

4.4. Modality and Fusion Ablation

To test whether both modalities are needed, and whether adaptive fusion is justified over simpler fusion, five configurations were trained and evaluated under identical splits, seeds, tuning budget and training conditions (Table 7): financial-only, loan-performance-only, concatenation of the two representations, fixed-weight fusion ($\alpha = \beta = 0.5$) and the proposed adaptive fusion. Figure 3 shows the ROC and precision-recall curves of run 0 for these five configurations, generated from the same predictions as Table 7.

Using both modalities clearly helps: ROC-AUC rises from 0.797 ± 0.009 (financial-only) and 0.747 ± 0.006 (loan-performance-only) to 0.840 ± 0.008, and PR-AUC from 0.075 ± 0.005 and 0.114 ± 0.009 to 0.147 ± 0.011. The two modalities are complementary: the financial one has the higher ROC-AUC, while the behavioural/contextual one has the higher PR-AUC and precision, i.e. it separates the highest-risk loans better. Adaptive fusion, however, is statistically indistinguishable from concatenation and fixed-weight fusion (ΔPR-AUC +0.0009, corrected p = 0.356, and +0.0004, corrected p = 0.715; ΔROC-AUC below 0.001, corrected p = 0.278 and 0.369). The justification of adaptive fusion in this study is therefore not accuracy: it is the interpretability that instance-wise modality weights provide at no loss of accuracy (Section 4.8). Whether adaptive weighting would pay off when the reliability of the modalities differs strongly between borrowers cannot be concluded from this data set.

Table 7. Modality and fusion ablation on the random-split test set (mean ± SD over 10 repeated splits, with overlap-corrected 95% confidence intervals)

Configuration	Accuracy	Precision	Recall	F1	MCC	ROC-AUC	PR-AUC	ROC-AUC 95% CI	PR-AUC 95% CI
Financial-only	0.918±.020	0.086±.007	0.268±.075	0.128±.009	0.116±.016	0.797±.009	0.075±.005	0.786–0.808	0.069–0.081
Loan-performance-only	0.964±.006	0.203±.032	0.183±.032	0.188±.012	0.173±.011	0.747±.006	0.114±.009	0.739–0.754	0.103–0.125
Concatenation	0.954±.008	0.179±.015	0.277±.054	0.214±.008	0.199±.012	0.840±.008	0.146±.010	0.830–0.849	0.133–0.158
Fixed-weight fusion	0.954±.007	0.175±.012	0.274±.048	0.211±.010	0.195±.012	0.840±.007	0.146±.010	0.831–0.849	0.133–0.159
Proposed (adaptive fusion)	**0.949±.006**	**0.165±.008**	**0.301±.039**	**0.212±.007**	**0.198±.010**	**0.840±.008**	**0.147±.011**	**0.831–0.850**	**0.133–0.160**

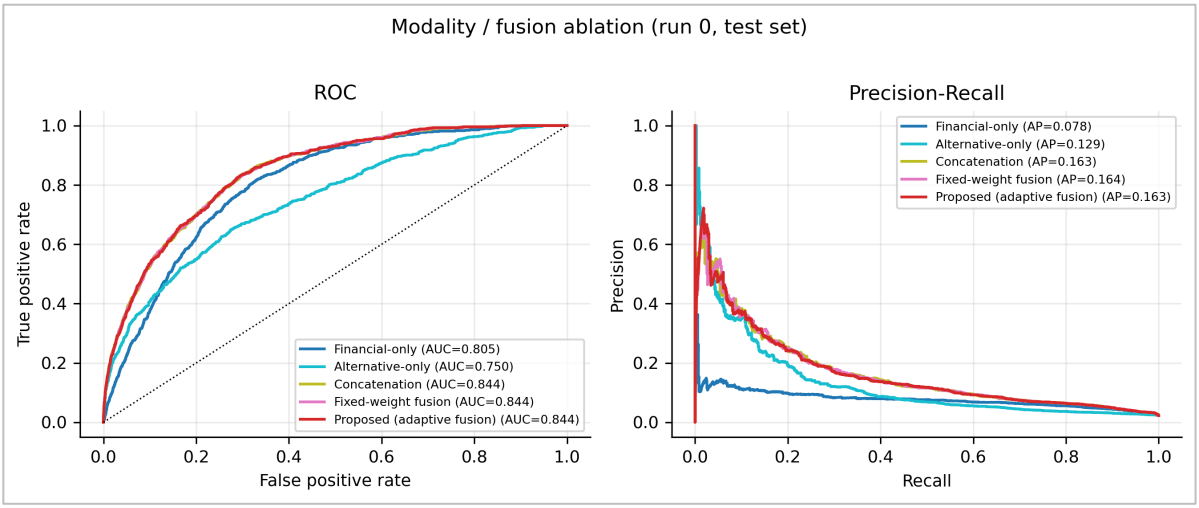


Figure 3. ROC (left) and precision-recall (right) curves of the ablation configurations on the test set of run 0. The legend label "Alternative-only" denotes the loan-performance-only configuration.

4.5. Later-Age Snapshot and External Validation

Generalisation was tested in two ways. In the later-age snapshot protocol (Table 8), models trained on loans at age 12 months are tested on different loans at age 24 months; this combines a shift in loan age with a partial shift in calendar time (Table 4), and the default rate rises from 2.27% to 5.29%, so it is also a test under prevalence shift. In the external protocol (Table 9, Figure 4), models trained only on 2007Q1 are tested on 377,697 loans of the 2018Q1 vintage, an ordinary-era cohort whose default rate is only 0.41%.

Table 8. Later-age snapshot test (age 24 months, unseen loans; mean ± SD over 10 repeated splits, with overlap-corrected 95% confidence intervals)

Model	ROC-AUC	PR-AUC	F1	MCC	Recall in top 5%	Lift (top 5%)	ROC-AUC 95% CI	PR-AUC 95% CI
Logistic Regression	0.799±.002	0.208±.006	0.249±.010	0.216±.009	24.9%	5.0	0.795–0.803	0.198–0.217
Random Forest	0.774±.003	0.190±.007	0.238±.010	0.202±.007	23.8%	4.8	0.770–0.779	0.180–0.201
XGBoost	0.793±.004	0.207±.008	0.245±.014	0.214±.010	25.2%	5.0	0.786–0.799	0.194–0.220
MLP	0.795±.003	0.204±.006	0.253±.013	0.218±.008	25.0%	5.0	0.790–0.800	0.195–0.213

Model	ROC-AUC	PR-AUC	F1	MCC	Recall in top 5%	Lift (top 5%)	ROC-AUC 95% CI	PR-AUC 95% CI
Financial-only	0.742±.004	0.126±.004	0.173±.024	0.125±.019	15.4%	3.1	0.736–0.748	0.119–0.133
Loan-performance-only	0.737±.004	0.175±.006	0.217±.045	0.191±.020	23.6%	4.7	0.731–0.743	0.166–0.185
Concatenation	0.791±.006	0.201±.009	0.238±.020	0.208±.010	24.6%	4.9	0.781–0.800	0.188–0.215
Fixed-weight fusion	0.790±.005	0.202±.007	0.249±.013	0.215±.009	24.5%	4.9	0.782–0.798	0.190–0.213
Proposed (adaptive fusion)	**0.791±.007**	**0.203±.007**	**0.251±.013**	**0.219±.007**	**24.6%**	**4.9**	**0.780–0.802**	**0.192–0.214**

Table 9. External test on 2018Q1 loans (models trained on 2007Q1 only; mean ± SD over 10 runs, with 95% confidence intervals reflecting training variability only)

Model	ROC-AUC	PR-AUC	F1	MCC	Recall in top 5%	Lift (top 5%)	ROC-AUC 95% CI	PR-AUC 95% CI
Logistic Regression	0.816±.004	0.047±.003	0.091±.025	0.103±.017	35.0%	7.0	0.813–0.818	0.046–0.049
Random Forest	0.839±.003	0.057±.004	0.137±.004	0.143±.006	40.1%	8.0	0.837–0.841	0.054–0.059
XGBoost	0.828±.005	0.053±.006	0.127±.016	0.126±.014	35.8%	7.2	0.824–0.831	0.049–0.057
MLP	0.832±.005	0.055±.003	0.129±.009	0.129±.008	36.9%	7.4	0.829–0.835	0.052–0.057
Financial-only	0.803±.007	0.017±.000	0.003±.002	0.008±.004	25.9%	5.2	0.798–0.808	0.016–0.017
Loan-performance-only	0.693±.012	0.049±.002	0.127±.007	0.129±.005	32.1%	6.4	0.684–0.702	0.047–0.050
Concatenation	0.824±.005	0.054±.003	0.110±.020	0.123±.012	37.3%	7.5	0.820–0.827	0.051–0.056
Fixed-weight fusion	0.824±.006	0.053±.003	0.098±.015	0.117±.010	37.7%	7.5	0.820–0.829	0.051–0.056
Proposed (adaptive fusion)	**0.824±.005**	**0.056±.003**	**0.124±.012**	**0.132±.007**	**37.9%**	**7.6**	**0.821–0.828**	**0.054–0.058**

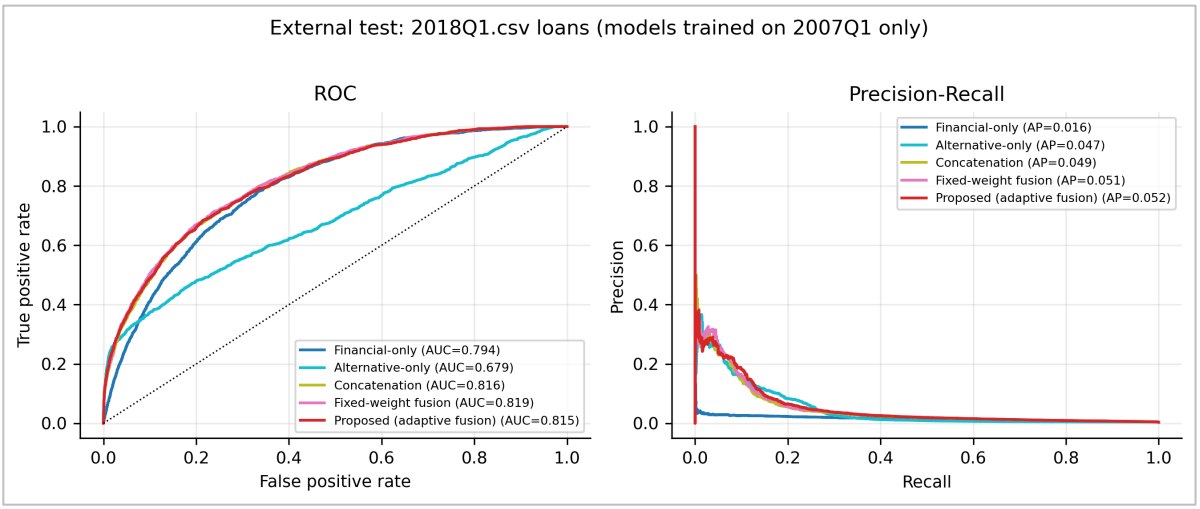


Figure 4. ROC (left) and precision-recall (right) curves of the ablation configurations on the external 2018Q1 test set (run 0). The legend label "Alternative-only" denotes the loan-performance-only configuration.

On the later-age snapshot, the proposed model reaches a ROC-AUC of 0.791 ± 0.007 and a PR-AUC of 0.203 ± 0.007 (random guess 0.053); all baseline and fusion models lie between 0.774 and 0.799, with Logistic

Regression nominally best (0.799) and Random Forest lowest (0.774); after correction for overlapping test sets and Holm adjustment, no difference between the proposed model and any other model is significant on this protocol (smallest Holm-adjusted p = 0.21, against Random Forest on ROC-AUC), so the ranking is nominal. Its F1 (0.251 ± 0.013) and MCC (0.219 ± 0.007) are similar to those of the best baselines (F1 0.238 to 0.253); thresholds chosen on validation data transfer imperfectly when the default rate doubles, and accuracy is less stable across runs (0.893 ± 0.057). On the external 2018Q1 test the proposed model reaches a ROC-AUC of 0.824 ± 0.005 and a PR-AUC of 0.056 ± 0.003, about 14 times the default rate of 0.0041, and its top-scored 5% of loans contain 37.9% of all defaults. Random Forest (ROC-AUC 0.839), the MLP (0.832) and XGBoost (0.828) are higher than the proposed model on ROC-AUC (differences -0.0145, -0.0076 and -0.0034 for the proposed model; paired-test $p < 0.001$, 0.001 and 0.110 across training runs, Holm-adjusted 0.0003, 0.012 and 0.55, so the gap is clear for Random Forest and the MLP but not for XGBoost; these p-values reflect training randomness only, because the external test set is identical in all runs); Logistic Regression is lower (0.816; Holm-adjusted p = 0.002). The proposed model therefore transfers across eras with a small loss in ranking quality (ROC-AUC 0.840 on the random split, 0.824 externally), even though the 2007Q1 training vintage was a crisis cohort and the default rate of 2018Q1 is much lower; because the external test is strictly chronological and at the same loan age as the training data, it is the more reliable evidence of temporal generalisation than the later-age snapshot. However, the proposed model does not lead the external comparison: Random Forest, the weakest model on the random split, achieves the highest external ROC-AUC (0.839) and, nominally, the highest external PR-AUC (0.057) and top-5% recall (40.1%), so the ranking of the models is not stable across protocols and the external results do not show an advantage of adaptive multimodal fusion over the baselines. In the external test, financial-only features keep a ROC-AUC of 0.803 but a PR-AUC of only 0.017, and loan-performance-only features drop to 0.693, so the behavioural and contextual information transfers less well across eras than underwriting variables; the multimodal advantage over financial-only features is visible mainly in PR-AUC (0.056 against 0.017).

4.6. Class-wise Performance, Calibration and Cost-sensitive Evaluation

Table 10 gives the exact test-set size, the class counts, the confusion matrix and class-wise precision, recall and F1 for run 0 (thresholds chosen on validation data). Of the 31,038 test loans, 705 are defaults. The proposed model detects 233 of the 705 defaults (recall 0.330) at a precision of 0.158 for the default class, flags 1,240 non-defaults, and misses 472 defaults; the baselines show the same trade-off between detected defaults and false alarms. Calibration and cost-sensitive results are reported in Table 11 and Figure 5.

Because the training set is balanced, raw sigmoid outputs (p_{raw}) are not calibrated probabilities and are strongly inflated (mean raw ECE 0.301 for the proposed model), but Platt scaling fitted on the natural-rate validation set restores calibration (ECE 0.002; Figure 5). Brier scores are 0.021 for all models, only slightly below the value of a constant predictor at this default rate (0.022), as expected for a rare outcome. In the cost-sensitive evaluation a false approval is assumed to cost r times a false rejection. At r = 10 the expected cost per 1,000 loans is 193.1 for the proposed model, 193.2 for XGBoost, 192.2 for the MLP and 196.6 for Logistic Regression, i.e. the differences are about 2% and not meaningful, whereas financial-only features cost 223.4. The ratio is an assumption that a lender must set; Table 11 shows the expected cost for ratios of 5, 10, 20 and 50.

4.7. Influence of Class Imbalance

To examine whether class imbalance limits performance, the proposed model and XGBoost were retrained under four treatments on the same splits (three runs): no weighting, square-root weighting, full inverse-frequency weighting, and 1:1 undersampling of the training set (the primary setting, with $\omega = 1$; the other three use the full training set) (Table 12). Ranking performance is practically unchanged by the treatment (ROC-AUC 0.838 to 0.843; PR-AUC 0.139 to 0.156). What changes is the threshold behaviour and the calibration: without weighting, raw sigmoid outputs are already well calibrated (ECE about 0.002), while every rebalancing inflates them (ECE 0.08 to 0.31) and requires recalibration. The proposed model identifies defaults reliably rather than optimising overall accuracy: its top-scored 5% of loans contain about 34% of all defaults under every treatment, and detecting more defaults at a fixed threshold costs additional false alarms. Imbalance handling is thus a matter of operating point and calibration here, not of model quality.

Table 10. Confusion matrix and class-wise metrics in run 0 (class 0 = no default, class 1 = default; 31,038 test loans, 705 defaults)

Model	TN	FP	FN	TP	P (0)	R (0)	F1 (0)	P (1)	R (1)	F1 (1)
Logistic Regression	29,382	951	507	198	0.983	0.969	0.976	0.172	0.281	0.214
Random Forest	28,978	1,355	484	221	0.984	0.955	0.969	0.140	0.313	0.194
XGBoost	29,117	1,216	472	233	0.984	0.960	0.972	0.161	0.330	0.216
MLP	29,577	756	525	180	0.983	0.975	0.979	0.192	0.255	0.219
Financial-only	28,722	1,611	538	167	0.982	0.947	0.964	0.094	0.237	0.135
Loan-performance-only	29,945	388	593	112	0.981	0.987	0.984	0.224	0.159	0.186
Concatenation	29,052	1,281	461	244	0.984	0.958	0.971	0.160	0.346	0.219
Fixed-weight fusion	29,356	977	495	210	0.983	0.968	0.976	0.177	0.298	0.222
Proposed (adaptive fusion)	29,093	1,240	472	233	0.984	0.959	0.971	0.158	0.330	0.214

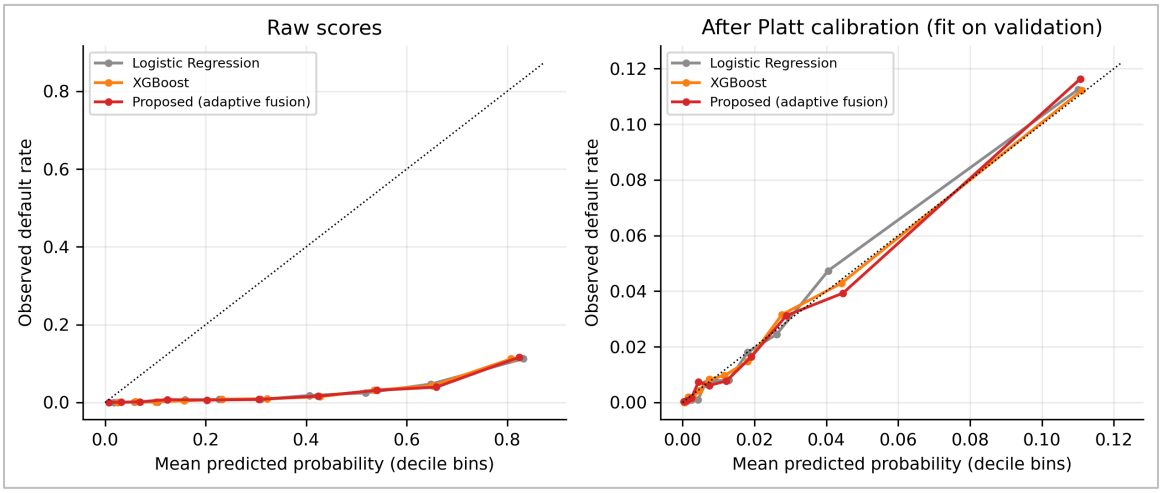


Figure 5. Reliability curves (decile bins) of three models before (left) and after (right) Platt scaling fitted on validation data (run 0).

Table 11. Calibration (Brier score, ECE before and after Platt scaling) and expected cost per 1,000 loans for false-approval to false-rejection cost ratios of 5, 10, 20 and 50 (random split; mean over 10 runs)

Model	Brier (calibrated)	ECE (raw)	ECE (Platt)	Cost per 1,000, ratio 5	Cost per 1,000, ratio 10	Cost per 1,000, ratio 20	Cost per 1,000, ratio 50
Logistic Regression	0.021	0.307	0.002	107.6	196.6	320.3	499.6
Random Forest	0.021	0.367	0.002	108.8	198.0	329.1	528.1
XGBoost	0.021	0.308	0.002	107.1	193.2	317.9	497.9
MLP	0.021	0.293	0.002	106.3	192.2	317.6	495.2
Proposed (adaptive fusion)	0.021	0.301	0.002	106.5	193.1	320.6	499.5

Table 12. Effect of the class-imbalance treatment (mean over 3 runs; random split)

Model	Treatment	ROC-AUC	PR-AUC	Recall	Precision	F1	Recall in top 5%	ECE (raw)
Proposed (adaptive fusion)	full inverse-frequency	0.841	0.152	0.272	0.192	0.225	34.1%	0.281
Proposed (adaptive fusion)	no weighting	0.840	0.153	0.283	0.179	0.218	34.1%	0.002
Proposed (adaptive fusion)	sqrt weighting	0.840	0.153	0.300	0.173	0.219	33.9%	0.083
Proposed (adaptive fusion)	undersampling 1:1 (training only)	0.838	0.154	0.304	0.169	0.216	33.7%	0.301
XGBoost	full inverse-frequency	0.842	0.151	0.293	0.173	0.216	34.2%	0.290
XGBoost	no weighting	0.843	0.156	0.262	0.200	0.223	33.9%	0.002
XGBoost	sqrt weighting	0.842	0.156	0.266	0.193	0.218	34.3%	0.083
XGBoost	undersampling 1:1 (training only)	0.838	0.139	0.297	0.166	0.212	33.1%	0.312

4.8. Explainability

Explainability was implemented with SHAP [37], using the gradient-based (expected-gradients) estimator [38] for the trained model of run 0 on 1,000 test loans (200 defaults and 800 non-defaults) with a background of 200 training loans. Attributions of one-hot columns and missing indicators were summed to the original feature, and attributions were also summed by modality. Figure 6 shows the global importance. The five most important features are the credit score (16.8% of total attribution), the property state (11.5%), the loan purpose (8.2%), the combined LTV (7.7%) and the number of borrowers (6.0%); financial features account for 70% and loan-performance features for 30% of the total attribution. Within the loan-performance modality, the state and the unpaid-principal paydown ratio dominate. The prominence of the state most likely reflects regional house-price declines during the 2007-2009 crisis and should not be read as a causal effect; it also illustrates why geographic variables require fairness scrutiny (Section 4.10). SHAP describes the behaviour of the model, not causal relationships.

Figure 7 gives borrower-level explanations for three correctly flagged defaults, one missed default and one low-risk loan. Each bar shows how much a feature (with its value) pushes the loan's log-odds of default up (red) or down (blue). Such per-loan explanations let a credit officer see which features drive an individual score.

The learned fusion weights were analysed by risk group (Figure 8). The average weight on the financial modality is 0.505 for defaults and 0.530 for non-defaults. Across the five predicted-risk quintiles it falls from 0.559 (lowest risk) to 0.505 (highest risk), and the financial share of SHAP attributions falls from 73% to 60% in the highest-risk quintile, so the model leans slightly more on behavioural and contextual information for the riskiest loans. This pattern is weak and not stable: across the 10 runs the weight for the highest-risk quintile is lower than for the lowest-risk quintile in only 7 runs, and the overall weight varies between 0.46 and 0.54 (supplementary file fusion_weights_stability_across_runs). The weights stay close to 0.5, so adaptive fusion behaves almost like fixed-weight fusion, which is consistent with the ablation. The weights are best viewed as a descriptive diagnostic, not as evidence of when adaptive fusion helps.

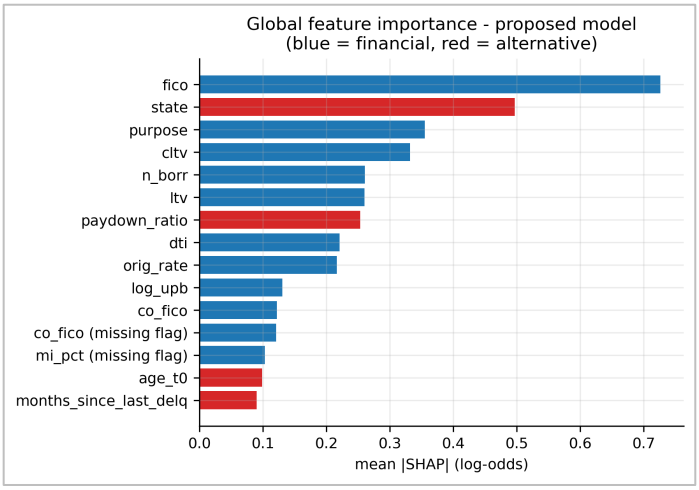


Figure 6. Global feature importance (mean absolute SHAP value, log-odds) of the proposed model; blue = financial, red = loan-performance (labelled "alternative" in the figure).

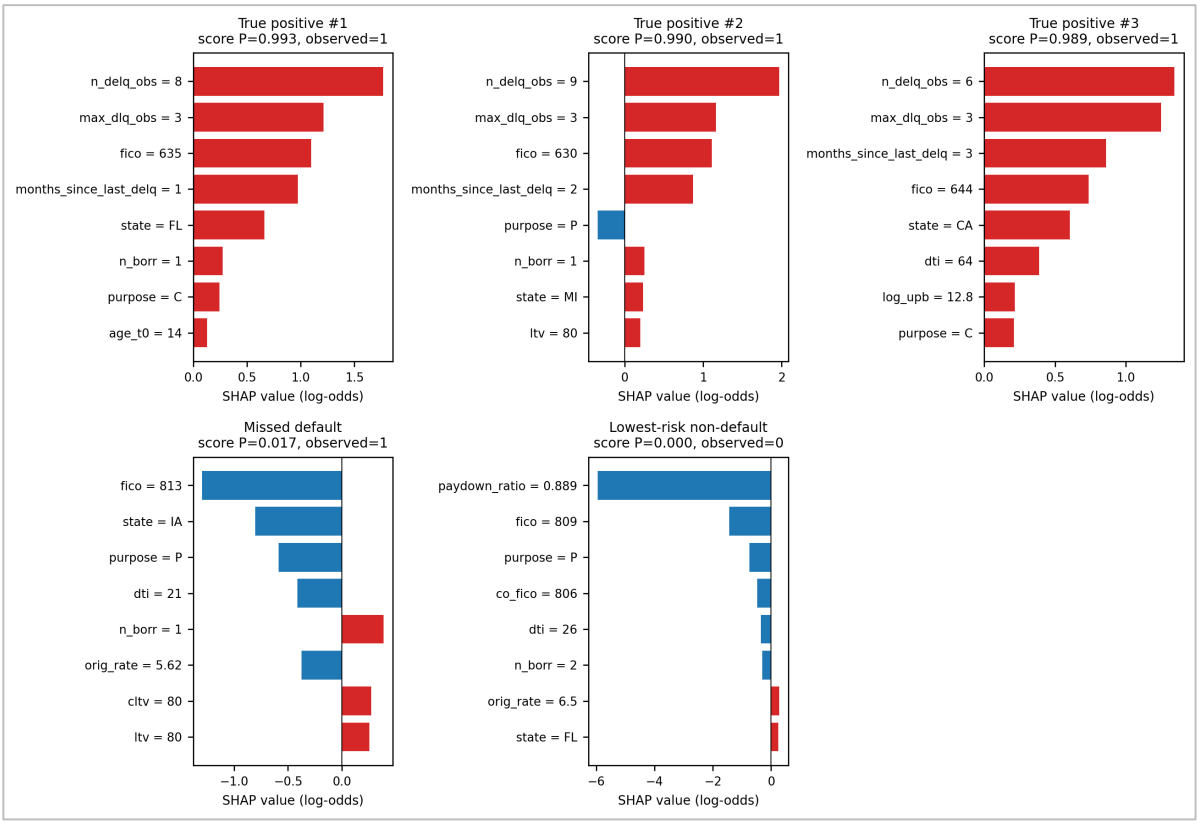


Figure 7. Borrower-level SHAP explanations for five test loans: top eight contributing features, with predicted probability and observed outcome.

4.9. Comparison with Related Studies

Table 13 compares the study qualitatively with related work. Numerical comparison would be misleading because the studies differ in data, risk definitions, splits and metrics. The closest reference is the leakage-controlled Fannie Mae study of Hu et al. [26], whose test AUROC values range from about 0.72 to 0.82 under a strict temporal split; the ROC-AUC values of this study (0.840 random split, 0.791 later-age snapshot, 0.824 external) lie in the same range or above it, though on different cohorts and targets. Several multimodal studies [21], [22], [24] report large gains from text and social-media modalities that are not available in public mortgage data, and our results support the general message that complementary

information helps (Section 4.4), while showing that a learned adaptive fusion adds little over simple fusion on structured mortgage data.

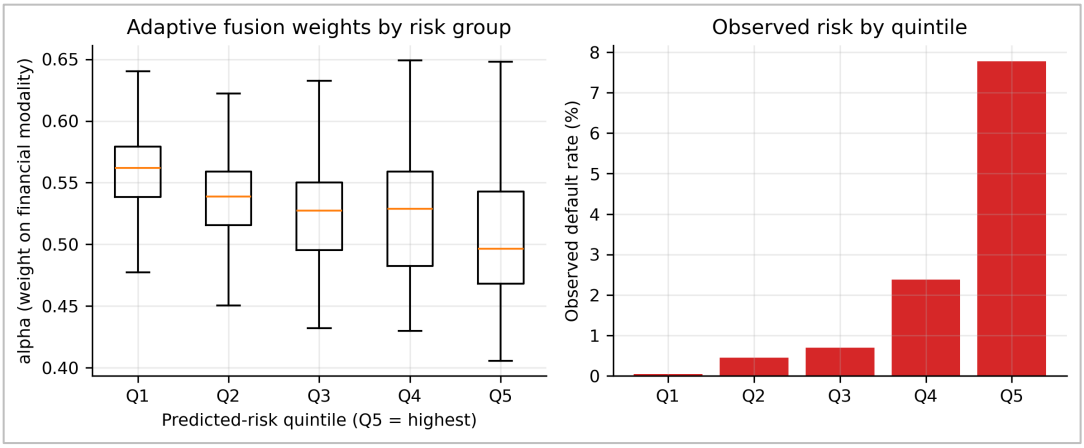


Figure 8. Adaptive fusion weight on the financial modality (left) and observed default rate (right) by predicted-risk quintile (run 0).

Table 13. Qualitative comparison of the present study with related work

Study	Approach	Data	Key result as reported
El Sayad et al. [12]	ANN, CNN, RF, XGBoost	Mortgage and financial data	Ensemble models adaptable under market volatility
Barykin et al. [18]	Hybrid composite model	Financial and digital indicators	Improved mortgage risk prioritisation
Gao and Lau [21]	Deep learning	Financial and social-media signals	26.07% improvement in prediction accuracy
Lu et al. [22]	Extended deep belief network	Numerical, categorical and textual data	Outperformed benchmark models
Shen et al. [24]	LLM with LoRA	Financial, non-financial and textual data	F1-score of 0.829
Hu et al. [26]	LR, RF, XGBoost, LightGBM, AutoGluon	Fannie Mae loan-level data, strict temporal split	Test AUROC about 0.72-0.82
This study	Multimodal network with adaptive fusion	Fannie Mae 2007Q1; external 2018Q1	ROC-AUC 0.840, PR-AUC 0.147 (random split); 0.824 external ROC-AUC

4.10. Limitations and Threats to Validity

4.10.1. Data representativeness

Training and internal testing use one acquisition quarter (2007Q1), a financial-crisis cohort, restricted to fixed-rate, full-documentation, conventional loans acquired by Fannie Mae; external validation covers one ordinary-era quarter (2018Q1). The results do not describe current lending, other loan types or other countries, and the model should be re-validated and recalibrated before any use. The default rate differs strongly between cohorts (2.27%, 5.29% and 0.41%), so probabilities calibrated on one cohort are not valid for another without recalibration. Temporal validation is limited: the later-age snapshot is not a strictly chronological split (Section 4.2, Table 4), and only the external 2018Q1 test follows the training data in calendar time.

4.10.2. Target and censoring

The target is a 90+ day delinquency or credit-event exit within 12 months of loan age 12 months; other horizons or definitions may give different results. Prepayment is

treated as censoring, which ignores the competing risk of prepayment and uses information from after the prediction date to define the cohort (Section 4.1, Implications of the cohort exclusions); a sensitivity analysis that retains prepaid loans as non-events was not performed.

4.10.3. Terminology and alternative data

The public data contain no independently acquired alternative information. The second modality (labelled "alternative" in earlier versions and in the figure legends) consists of loan-performance and contextual information drawn from the same source as the underwriting variables, so the study does not test the value of external alternative data such as rent, utility or transaction records, and it should not be used to argue for or against them. The adaptive fusion did not outperform simpler fusion.

4.10.4. Bias and fairness

Variables such as state, seller, servicer and origination channel can act as proxies for protected characteristics, which are not available in this file and could not be audited directly. Table 14 shows a subgroup audit of the proposed model: ranking quality is similar across channels, first-time-buyer status, occupancy and purpose (ROC-AUC 0.82 to 0.88), but the share of flagged loans varies with borrower and loan type, from 1.8% for first-time buyers to 6.4% for purpose-C loans (6.3% for broker-originated loans) and, across credit-score bands, from 0.2% to 14.1%, largely following observed default rates, and the false-positive rate is 12.6% for credit scores below 680 against 0.2% above 760. Such disparities need review before deployment, together with transparent adverse-action explanations and human oversight; these safeguards were not experimentally evaluated here.

Table 14. Subgroup audit of the proposed model on the run-0 test set (threshold chosen on validation data)

Variable	Group	Loans	Default rate	Flagged	Recall	False-positive rate	ROC-AUC
Channel	B	4,477	3.4%	6.3%	0.375	5.2%	0.822
Channel	C	14,774	2.3%	4.7%	0.342	4.1%	0.849
Channel	R	11,787	1.9%	4.2%	0.282	3.7%	0.844
First-time buyer	N	28,145	2.4%	5.0%	0.331	4.4%	0.839
First-time buyer	Y	2,868	1.5%	1.8%	0.310	1.4%	0.880
Occupancy	I	2,226	2.2%	4.2%	0.340	3.5%	0.859
Occupancy	P	27,438	2.3%	4.8%	0.333	4.2%	0.841
Occupancy	S	1,374	2.0%	3.7%	0.259	3.3%	0.881
Purpose	C	14,397	2.8%	6.4%	0.344	5.6%	0.822
Purpose	P	10,515	1.4%	2.4%	0.297	2.0%	0.873
Purpose	R	6,126	2.4%	4.8%	0.327	4.1%	0.824
Credit score	<680	7,857	5.1%	14.1%	0.409	12.6%	0.749
Credit score	680-719	6,201	2.7%	4.1%	0.295	3.4%	0.815
Credit score	720-759	6,412	1.5%	1.4%	0.172	1.1%	0.788
Credit score	760+	10,506	0.4%	0.2%	0.109	0.2%	0.846

Note: The subgroup analyses were performed using the available observations for each grouping variable. Accordingly, 25 observations (31,038 − 31,013) were excluded from the first-time-buyer analysis and 62 observations (31,038 − 30,976) from the credit-score analysis because of missing or unclassified values. The channel, occupancy and purpose groups each sum to the complete test set of 31,038 loans.

4.10.5. Explanations and reproducibility

SHAP attributions describe the model, not causal effects, and are computed on a sample of test loans. The training of neural models is not bit-wise deterministic, so repeated executions can give slightly different values; the reported standard deviations and the overlap-corrected confidence intervals quantify the variation due to splits and initialisation but not to this non-determinism. Hyper-parameters were tuned once, on the validation set of the run-0 split, and the selected configuration was used in all runs; validation loans of run 0 can therefore appear in the test sets of other runs, which may make the reported scores slightly optimistic. This applies to all models alike, because all received the same tuning procedure, but its size was not quantified. Because the configuration was not re-selected within each split, the standard deviations and overlap-corrected confidence intervals over the 10 repeated splits reflect variation in splits and initialisation only, not variability in hyper-parameter selection. The external 2018Q1 test is not affected, because its loans were never used for tuning. The fusion weights are not stable across runs and should not be over-interpreted.

5. Conclusion

This study developed and evaluated a multimodal AI framework for mortgage default prediction that encodes underwriting and loan-performance (behavioural/contextual) information from the same dataset separately and combines them by instance-wise adaptive fusion. Under a leakage-controlled protocol on Fannie Mae data (206,914 loans, 2.27% defaults), the framework ranks default risk well: ROC-AUC 0.840 ± 0.008 and PR-AUC 0.147 ± 0.011 on the held-out test set, 0.791 on a later-age snapshot and 0.824 on a strictly later external 2018Q1 test, on which Random Forest (0.839), the MLP (0.832) and XGBoost (0.828) reach higher ROC-AUC. Using both modalities improves on either alone, which supports the complementarity of underwriting and behavioural information. After correction for overlapping test sets, adaptive fusion is statistically indistinguishable from concatenation, fixed-weight fusion, Logistic Regression, XGBoost and the MLP, and it is better only than Random Forest on ROC-AUC in the random split, whereas Random Forest is the best model in external ROC-AUC, so its value in this study lies in interpretability, not in accuracy. Accuracy is not informative at this default rate. Raw sigmoid outputs are not calibrated probabilities after class balancing and require Platt recalibration, and calibrated probabilities require recalibration when the default rate changes, and the explanations identify credit score, state, loan purpose and combined LTV as the main drivers. The limitations stated in Section 4.10 apply: a single crisis-era training vintage, a cohort restricted to performing loans that remain observed at loan age 12 months, no independently acquired alternative data, proxy variables that need fairness review, and unstable fusion weights. Privacy, fairness, transparency and human oversight are necessary conditions for any use of such models in lending; this study discusses them but evaluates only part of them (explanations and a subgroup audit). Future work should validate the framework on further vintages and loan types, add independently acquired alternative data where it is legally and ethically available, and test fusion mechanisms when modality reliability varies more strongly.

Funding: This work did not receive any external funding.

Author contributions: A.K. was responsible for conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing—original draft, and writing—review and editing. The author has read and approved the final version of the manuscript.

Data Availability Statement: The study uses the publicly available Fannie Mae Single-Family Loan Performance Data [27] (acquisition quarters 2007Q1 and 2018Q1), available after free registration at https://capitalmarkets.fanniemae.com. The loan-level cohorts can be rebuilt from these files with the supplied code; the cohort counts and exclusions are listed in Section 4.1.

Code availability: The complete pipeline (data preparation, leakage-controlled splits, all models, statistics, figures and tables) is available in a public repository (https://github.com/amitkurchania05-hub/Impaxon-Code.git) together with the stored results of the reported run (feature list, hyper-parameter tuning logs, run configuration, per-run metrics, statistical test outputs and figures), the script for the corrected statistics, a requirements file with the software versions and a README that maps every table and figure of this paper to its output file.

Reproducibility: All hyper-parameters, seeds, splits and software versions are given in Section 4.2; every table and figure is generated by the script from the same stored results. The run reported in this paper is stored unchanged in the repository; running the pipeline again can change the

values slightly because GPU training is not bit-wise deterministic.

Acknowledgments: None.

Conflicts of Interest: The author declares no conflicts of interest.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Ethical Considerations: Not Applicable.

AI Use Declaration: Generative AI tools and Grammarly were used solely for language editing and proofreading in the complete manuscript. The author reviewed and approved all content and takes full responsibility for the manuscript.

References:

1. Chan, L., Hogaboam, L., Cao, R. (2022). Artificial Intelligence in Credit, Lending, and Mortgage. In: *Applied Artificial Intelligence in Business. Applied Innovation and Technology Management*. Springer, Cham. https://doi.org/10.1007/978-3-031-05740-3_13
2. Garcia-Teruel, Rosa M., (2022). Integrating Artificial Intelligence into the Mortgage Credit Market (December 16, 2022). Available at SSRN: http://dx.doi.org/10.2139/ssrn.4304716
3. Bi, S., & Bao, W. (2024). Innovative application of artificial intelligence technology in bank credit risk management. *arXiv preprint arXiv:2404.18183*. https://doi.org/10.48550/arXiv.2404.18183
4. Anderson, J. T., Freybote, J., Lucus, D., Seiler, M. J., & Simon, L. (2022). Using Artificial Intelligence to Identify Strategic Mortgage Default Attitudes. *Journal of Real Estate Research*, *44*(3), 429–445. https://doi.org/10.1080/08965803.2021.2009621
5. Zhou, M. (2022). Credit risk assessment modeling method based on fuzzy integral and SVM. *Mobile Information Systems*, 2022(1), 3950210. https://doi.org/10.1155/2022/3950210
6. Rasel, I. H., Ibrahim, M., Pritty, A. A., Fahim, A. S. M., & Jahan, N. (2023). Beyond FICO: Enhancing mortgage default forecasting and inclusive lending via multimodal graph neural networks and urban mobility analytics. *Frontiers in Computer Science and Artificial Intelligence*, *2*(2), 62-81. http://doi.org/10.32996/fcsai.2023.2.2.5
7. Yohannes, K. T. F. (2022). Next-Generation Mortgage Platforms: Merging Explainable AI, Sign Language Intelligence, and Sustainable IT for Risk-Aware Software Solutions. *International Journal of Computer Technology and Electronics Communication*, 5(3), 5117-5121. https://doi.org/10.15680/vhjvc153
8. Sriram, H. K. (2022). AI Neural Networks In Credit Risk Assessment: Redefining Consumer Credit Monitoring And Fraud Protection Through Generative AI Techniques. *Migration Letters*, 19(6), 1017-1032. https://migrationletters.com/index.php/ml/article/view/11619
9. Hung, M. T., & Lo, H. C. (2024). Risk analysis of mortgage loan default for bank customers and AI machine learning. *Journal of Applied Finance & Banking*, 14(6), 33-50. https://doi.org/10.47260/jafb/1463
10. Konovalova, N., Saksonova, S., Velinov, E. (2026). Improvement of Credit Risk Assessment Using AI. In: Kabashkin, I., Yatskiv, I., Prentkovskis, O. (eds) Reliability and Statistics in Transportation and Communication: Ecosystems for Smart Connectivity and Intelligent Mobility. RelStat 2025. *Lecture Notes in Networks and Systems*, vol 1822. Springer, Cham. https://doi.org/10.1007/978-3-032-19245-5_48
11. Ramakrishnan, R., Rohella, P., Mimani, S., Jiwani, N., & Logeshwaran, J. (2024, March). Employing AI and ML in risk assessment for lending for assessing credit worthiness. *In 2024 2nd International Conference on Disruptive Technologies (ICDT)* (pp. 561-566). IEEE. https://doi.org/10.1109/ICDT61202.2024.10489313
12. El Sayad, I., Chong, F. H. L., & Gosal, B. (2026). Risk assessment in mortgages: a comparative study of AI models. *PeerJ Computer Science*, 12, e3494. https://doi.org/10.7717/peerj-cs.3494
13. Hasan, M. M. (2026). AI-Enabled Financial Information Systems for Credit Risk Forecasting to Support Small Business Growth. American *Journal of Interdisciplinary Studies*, 7(01), 270-314. https://doi.org/10.63125/fhzazv80
14. Raji, A. A. H., Alabdoon, A. H. F., & Almagtome, A. (2024, April). AI in credit scoring and risk assessment: enhancing lending practices and financial inclusion. *In 2024 International conference on knowledge engineering and communication systems (ICKECS)* (Vol. 1, pp. 1-7). IEEE. https://doi.org/10.1109/ICKECS61492.2024.10616493
15. Chitturi, N. (2025). AI-driven credit scoring and alternative data: Expanding financial inclusion and access to credit for underserved populations. *IPHO-Journal of Advance Research in Social Science and Humanities*, 3(09), 06-17. https://doi.org/10.5281/zenodo.17231820
16. Zhang, X. (2025, February). The application of alternative data in credit scoring. *In International Workshop on Navigating the Digital Business Frontier for Sustainable Financial Innovation (ICDEBA 2024)* (pp. 789-794). Atlantis Press. https://doi.org/10.2991/978-94-6463-652-9_84
17. Di Maggio, M., & Ratnadiwakara, D. (2026). Invisible primes: Fintech lending with alternative data. *Management Science*. https://doi.org/10.1287/mnsc.2024.07854
18. Barykin, S., Shulga, A., Dinets, D., & Mikhaylov, A. (2026). Digital twin-enhanced credit risk prioritization in mortgage portfolios: a hybrid model approach. *Quality & Quantity*, 60(2), 7701-7723. https://doi.org/10.1007/s11135-026-02579-3
19. Xu, T., Zhu, W., & Zhang, J. (2026, May). A Study on the Application of Alternative Data in Credit Assessment for the Unbanked Population. In *Proceedings of the 2026 2nd International Conference on Sustainable Development and Digital Transformation* (pp. 379-385) https://dx.doi.org/10.2139/ssrn.6973241
20. Wang, T., Liu, R., & Qi, G. (2022). Multi-classification assessment of bank personal credit risk based on multi-source information fusion. *Expert systems with applications*, 191, 116236. https://doi.org/10.1016/j.eswa.2021.116236
21. Gao, T., & Lau, R. Y. (2024, August). Leveraging Deep Learning and Multimodal Signals from Social Media to Enhance Credit Risk Prediction. In *2024 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICSPCC62635.2024.10770419
22. Lu, S., Zhang, X., Su, Y., Liu, X., & Yu, L. (2025). Efficient multimodal learning for corporate credit risk prediction with an extended deep belief network. *Annals of Operations Research*, **361**, 309–346. https://doi.org/10.1007/s10479-025-06612-w
23. Fang, Y. (2025, September). Design of a Multimodal Credit Risk Forecasting System for Financial Text Analysis. In *Proceedings of the*

2nd International Symposium on Integrated Circuit Design and Integrated Systems (pp. 191-196). https://doi.org/10.1145/3772326.3774731

24. Shen, C., Pan, W., & Shen, X. (2026). Credit risk prediction and heterogeneity analysis for SMEs based on large language models and multimodal data fusion. *Complex & Intelligent Systems*, 12, 82. https://doi.org/10.1007/s40747-025-02192-1
25. Cao, Y., Chen, Z., Kumar, P., Pei, Q., Yu, Y., Li, H., ... & Ndiaye, P. M. (2025, November). RiskLabs: Predicting financial risk using large language model based on multimodal and multi-sources data. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 901-909). IEEE. https://doi.org/10.48550/arXiv.2404.07452
26. Hu, X., Xu, T., Chen, Y., & Wang, S. (2026). Predicting mortgage default with machine learning: AutoML, class imbalance, and leakage control. *arXiv preprint arXiv:2602.00120*. https://doi.org/10.48550/arXiv.2602.00120
27. Fannie Mae. (2026). Single-Family Loan Performance Data [Data set]. Fannie Mae Capital Markets. https://capitalmarkets.fanniemae.com/credit-risk-transfer/single-family-credit-risk-transfer/fannie-mae-single-family-loan-performance-data (accessed October 2026).
28. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785
29. Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *In International Conference on Learning Representations (ICLR 2019).* https://doi.org/10.48550/arXiv.1711.05101
30. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432
31. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 6. https://doi.org/10.1186/s12864-019-6413-7
32. Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Scholkopf, & D. Schuurmans (Eds.), Advances in Large Margin Classifiers (pp. 61-74). *MIT Press*. https://home.cs.colorado.edu/~mozer/Teaching/syllabi/6622/papers/Platt1999.pdf
33. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *In Proceedings of the 34th International Conference on Machine Learning* (PMLR 70, pp. 1321-1330). https://doi.org/10.48550/arXiv.1706.04599
34. Elkan, C. (2001). The foundations of cost-sensitive learning. *In Proceedings of the 17th International Joint Conference on Artificial Intelligence* (IJCAI 2001) (pp. 973-978). https://www.researchgate.net/publication/2365611_The_Foundations_of_Cost-Sensitive_Learning
35. Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, 52(3), 239-281. https://doi.org/10.1023/A:1024068626366
36. Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65-70.
37. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *In Advances in Neural Information Processing Systems* 30 (pp. 4765-4774). https://doi.org/10.48550/arXiv.1705.07874
38. Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., & Lee, S.-I. (2021). Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3, 620-631. https://doi.org/10.1038/s42256-021-00343-w

Disclaimer: All views, interpretations, and data presented in Impaxon publications are the sole responsibility of the respective authors. These do not necessarily reflect the opinions of Impaxon or its editorial team. Impaxon and its editors assume no liability for any harm or loss arising from the use of information, procedures, or materials discussed in the published content.

Publisher's Note: Impaxon remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.