Review of Applied Science and Technology

Volume: 1; **Issue:** 3
Pages: 32– 60

Quantitative Benchmarking of Machine Learning Models for Risk Prediction: A Comparative Study Using AUC/F1 Metrics and Robustness Testing

Md. Arifur Rahman[1]; B. M. Taslimul Haque[2];

[1]. *Bachelor of Science (B.Sc.) in Computer Science & Engineering, Bangladesh University of Business & Technology, Bangladesh; Email: rahman.arifur22226@gmail.com*

[2]. *Master Bachelor of Science in Computer Science & Engineering, American International University Bangladesh, Dhaka, Bangladesh; Email: bmtaslim121@gmail.com*

Doi: 10.63125/9hd4e011
Received: *11 June 2022;* ***Revised:*** *18 July 2022;* ***Accepted:*** *13 August 2022;* ***Published:*** *21 September 2022*

Abstract

This study addresses a problem in enterprise risk decision systems: model selection is often justified by baseline accuracy, although cloud and enterprise pipelines face measurement noise, missing values, and distribution shift that can weaken predictions. The purpose was to quantitatively benchmark machine learning classifiers for binary risk prediction and determine which approach delivers the best combination of effectiveness and stability. Using a quantitative, cross-sectional, case-based design, the study analyzed N = 4,200 enterprise case records with adverse-event prevalence of 8.6% (n = 361) and benchmarked Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting (XGB), and a Multilayer Perceptron. Key variables were model family and stress condition (baseline versus perturbation), with outcome variables AUC, F1-score, precision, recall, and robustness degradation (ΔAUC, ΔF1). The analysis plan applied standardized preprocessing and stratified validation, then computed baseline AUC and F1, followed by robustness testing under noise injection, 15% added missingness, and segment-shift distribution perturbation; descriptive statistics summarized performance, while correlation and regression tested relationships between robustness and performance, and a 5-point Likert survey (n = 120) captured adoption perceptions. Baseline results showed XGB as the top performer (AUC = 0.872, F1 = 0.532), followed by RF (0.846, 0.498), MLP (0.831, 0.474), SVM (0.804, 0.431), and LR (0.781, 0.402). Under robustness testing, XGB retained the smallest degradation (ΔAUC = -0.027; ΔF1 = -0.032) while MLP was most unstable (ΔAUC = -0.071; ΔF1 = -0.079), reshaping the composite ranking to XGB, RF, LR, SVM, then MLP. Robustness correlated with baseline performance (AUC: r = 0.62, p = 0.004; F1: r = 0.58, p = 0.008) and significantly predicted benchmark score (robustness index β = 0.44, p < 0.001), with noise, missingness, and distribution perturbation showing negative effects (β = -0.21, -0.18, -0.27; all p ≤ 0.009). Survey reliability was strong (Cronbach's alpha = 0.86) and respondents rated perceived reliability (M = 4.18, SD = 0.54), supporting deployment implications: enterprise risk teams should adopt multi-metric, stress-tested benchmarking before production rollouts in practice.

Keywords

Machine Learning Benchmarking; Enterprise Risk Prediction; Robustness Testing; AUC And F1; Cloud Analytics Pipelines;

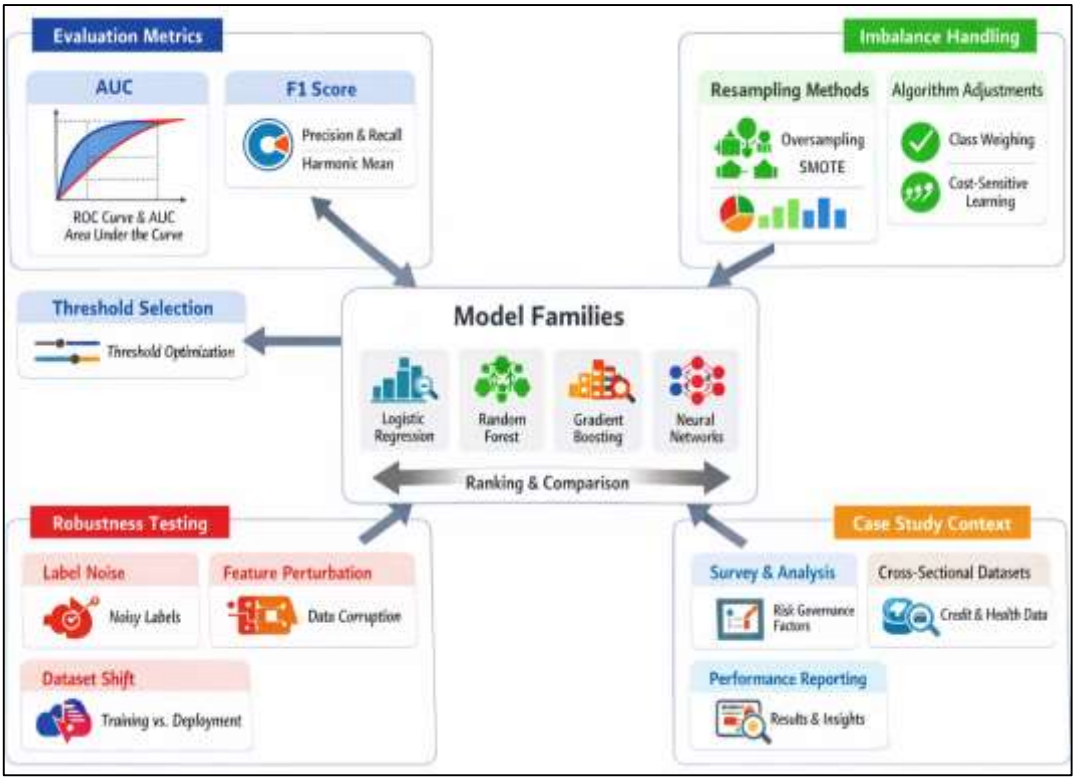
INTRODUCTION

Risk prediction is commonly defined as the quantitative estimation of the probability that an adverse event will occur within a specified time window, given observed predictors and an assumed data-generating process. In operational settings, "risk" functions as a decision-relevant construct that supports screening, prioritization, pricing, and intervention allocation across finance, insurance, health, cybersecurity, and industrial safety domains. Machine learning is typically defined as a class of computational methods that learn patterns from data to produce predictions or classifications with measurable generalization performance on unseen cases (Adams, 2009). Within supervised learning for risk prediction, models transform features into estimated scores or probabilities, and these outputs are evaluated using performance metrics that summarize discrimination, error trade-offs, and ranking quality. Internationally, risk prediction has become a core analytic capability because large-scale digitization has expanded the availability of transactional, clinical, and behavioral data, while regulatory and governance environments increasingly demand auditable, statistically defensible decision rules (Akhtar & Mian, 2018).

In banking and consumer lending, for example, risk scoring affects credit access and portfolio stability, linking predictive accuracy to institutional resilience and broader economic exposure (Brown & Mues, 2012). In population health and clinical prevention, risk prediction supports early identification of high-risk patients, enabling stratified care pathways and resource allocation in high-volume systems. The international significance is also methodological: cross-country datasets vary in prevalence rates, measurement practices, and sampling frames, which can reshape the apparent performance of identical algorithms when applied in distinct contexts. Consequently, risk prediction research increasingly emphasizes benchmarking—systematic comparison of models under consistent protocols—to establish credible evidence about which methods perform better, under what conditions, and with what stability (Boughorbel et al., 2017).

A central element of quantitative risk prediction is the definition and operationalization of model performance. In binary risk classification, discrimination is often assessed using receiver operating characteristic (ROC) analysis and the area under the ROC curve (AUC), which summarize the ranking ability of a scoring function across thresholds (Ben-David et al., 2010). AUC has been widely adopted because it is threshold-agnostic and interpretable as the probability that a randomly chosen positive instance receives a higher score than a randomly chosen negative instance. At the same time, AUC can obscure practically relevant differences, particularly when ROC curves cross or when different misclassification regions matter more for the decision maker, motivating critiques and alternative coherent measures (Patrini et al., 2017). Precision, recall, and their harmonic mean (F1) are also common, especially in skewed event settings where the positive class is rare and the decision focus is on capturing true events with controlled false alarms. The relationship between ROC and precision–recall views is non-trivial; for highly imbalanced data, precision–recall curves can provide more informative summaries of performance because they directly reflect the positive predictive value achievable at a given sensitivity (Nguyen et al., 2015). In imbalanced learning research, the selection of evaluation metrics is treated as part of the modeling problem because different metrics induce different optimization targets and may reward different error profiles. Metric choice further interacts with prevalence: when base rates differ across datasets or countries, a stable AUC can coexist with materially different F1 values due to shifts in precision, even if the underlying scoring function remains unchanged. This measurement complexity is one reason benchmarking studies in credit scoring often report multiple indicators, ranking algorithms differently depending on whether AUC-like or error-based metrics dominate. Robust quantitative comparisons therefore rely on explicit metric definitions, consistent thresholding rules when F1 is computed, and transparent reporting of uncertainty sources such as sampling variability and data-preprocessing decisions (Chicco & Jurman, 2020).

Figure 1: Comparative Benchmarking of Machine Learning Models for Risk Prediction



Risk prediction problems frequently exhibit class imbalance because adverse outcomes—default, fraud, rare disease onset, critical failure—occur at low prevalence relative to non-events. Imbalanced data learning is commonly defined as the study of methods and evaluation practices for tasks where the minority class is underrepresented, and naive learners may favor the majority class to achieve deceptively high accuracy. In such contexts, sampling strategies and algorithmic adjustments are frequently used to improve minority detection and stabilize decision boundaries (van Buuren & Groothuis-Oudshoorn, 2011). Oversampling methods, including variants that focus on borderline minority instances, formalize the idea that errors concentrate near class overlap regions and that targeted resampling can improve sensitivity and F-measures under certain conditions. Benchmarking studies in finance demonstrate that imbalance interacts with model choice: algorithm families can exhibit different trade-offs in capturing rare bad borrowers while controlling false positives, and these differences affect risk governance outcomes such as cut-offs and capital allocation (Louzada et al., 2016). Systematic reviews of credit scoring classification methods highlight that comparisons must account for dataset composition, missingness, and feature engineering, because performance gains can reflect preprocessing advantages rather than inherent algorithm superiority (Fawcett, 2006). From a measurement perspective, imbalanced settings intensify the importance of selecting metrics aligned with the decision task. AUC can remain high even when precision is low at the operational threshold, while F1 can sharply penalize poor positive predictive value, creating divergent narratives about model "quality." Complementary research proposes optimization and evaluation using metrics derived from confusion-matrix structure, showing that different metrics can imply different optimal thresholds and different error equilibria under imbalance. Because real-world risk prediction often operates under fixed intervention capacity—only a limited fraction of cases can be escalated—benchmark designs frequently treat threshold choice as part of the performance definition rather than an afterthought (Moreno-Torres et al., 2012). Consequently, a rigorous quantitative benchmarking study benefits from explicit imbalance handling, metric triangulation (AUC and F1), and clear protocols for threshold selection and comparative ranking (Davis & Goadrich, 2006).

The present study is structured around tightly defined objectives that translate risk prediction into measurable, testable components within a quantitative, cross-sectional, case-study design. It benchmarks a selected set of machine learning classification models by comparing AUC and F1-score

using a single consistently prepared dataset with standardized preprocessing and training–validation procedures to ensure fair model-to-model comparisons. It then evaluates model robustness through stress tests that reflect real-world disruptions (noise injection, missing-data simulation, and distributional shifts), quantifying performance stability as changes in AUC and F1 and summarizing these shifts into interpretable robustness indicators. Building on baseline and robustness results, the study produces an integrated comparative ranking of models to support decision-oriented selection when both effectiveness and stability are prioritized. Finally, it applies descriptive statistics, correlation analysis, and regression modeling to examine relationships among benchmarking indicators and to test whether observed differences across models and conditions are statistically meaningful within the case-study context, linking comparative evaluation to hypothesis-driven quantitative inference.

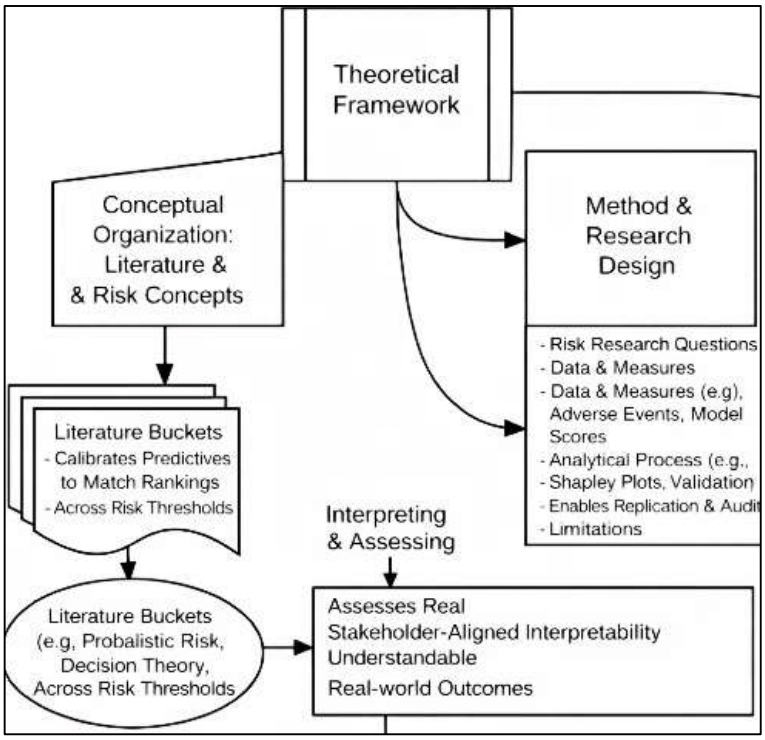
LITERATURE REVIEW

Machine learning–based risk prediction has grown rapidly across domains such as finance, healthcare, cybersecurity, and operations, where predictive analytics supports high-stakes decisions under uncertainty. The literature commonly treats risk prediction as a supervised classification task with imbalanced outcomes, making evaluation dependent not only on model choice but also on data preparation, imbalance handling, validation design, and metric selection, often emphasizing AUC and F1-score to capture discrimination and error trade-offs more accurately than naïve accuracy. In addition, recent studies stress robustness because deployment conditions frequently involve noise, missingness, and distribution shifts, so stability testing is used alongside baseline benchmarking to identify models that remain reliable when stressed. Conceptual and risk-governance perspectives further argue that benchmarking should integrate both effectiveness and resilience through transparent, replicable methods, motivating reviews that synthesize applications, model families, imbalanced-evaluation practices, robustness testing approaches, and the theoretical rationale for combining performance and stability in comparative rankings.

Risk Prediction in Modern Decision Systems

Risk prediction in modern decision systems refers to the quantitative estimation of the likelihood that an adverse outcome—such as default, fraud, clinical deterioration, operational failure, or a cybersecurity incident—will occur for a given entity within a defined horizon, using observed attributes and a specified modeling approach.

Figure 2: Risk Prediction in Modern Decision Systems



In operational terms, risk modeling converts heterogeneous observations into decision-relevant quantities that can be acted on under constraints such as limited staff capacity, budget ceilings, regulatory rules, and service-level agreements. Risk scores and probabilities are used to rank cases, set cutoffs, or trigger interventions, which means the model is part of the control loop that allocates resources and shapes outcomes. The term "risk model" therefore includes both the mapping from predictors to a score and the governance logic that defines the outcome, sets the time window, and specifies what constitutes a positive event. Internationally, risk prediction has become central because large-scale digitization has increased the volume and velocity of data from financial transactions, connected devices, electronic records, and online behavior, while organizations face pressure to justify decisions with measurable evidence and standardized processes. In such environments, risk prediction supports screening at scale, triage under uncertainty, and audit-ready decision documentation across jurisdictions. Many applications require probabilistic outputs that are coherent for decision making, not only rankings, because pricing, prioritization, and compliance often depend on well-calibrated probabilities. Proper scoring rules provide a principled foundation for this probabilistic view by linking evaluation to the quality of predictive distributions, reinforcing that good risk models assign high probability to events that occur and low probability to events that do not in a coherent way (Gneiting & Raftery, 2007; Rauf, 2018). In finance and healthcare, these estimates influence who receives attention first, how much monitoring is applied, and how resources are distributed, making model quality a matter of system performance and fairness.

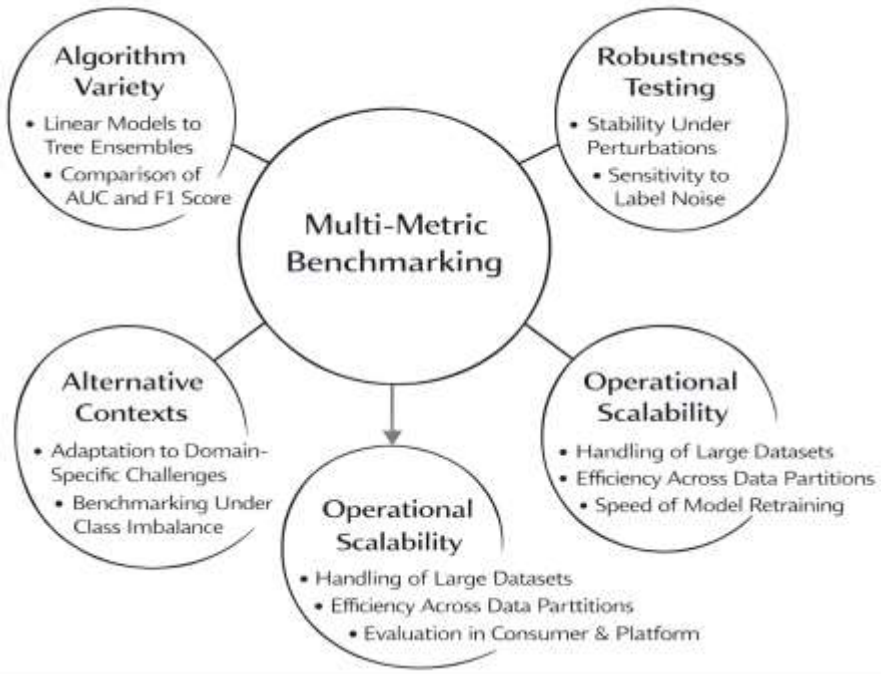
Machine Learning Models for Risk Prediction

Risk prediction in operational and financial environments is typically framed as a supervised classification problem in which a model estimates the probability that an entity (e.g., borrower, transaction, account, equipment state, or process instance) will transition into an adverse outcome within a defined horizon. In this setting, "benchmarking" refers to evaluating multiple candidate learners under consistent data partitions, preprocessing rules, and performance criteria so that relative strengths can be interpreted as methodological effects rather than procedural artifacts. A core theme in contemporary comparative work is the shift from a single "best model" mindset toward an evidence-based portfolio view: linear scorecards may offer stability and transparency, margin-based methods may provide strong separation under appropriate feature scaling, and tree-ensemble families may deliver high discriminative power under complex interactions. For example, support vector machines have been assessed in credit-risk contexts as competitive alternatives to traditional statistical baselines, with attention to how feature selection and kernelized decision boundaries can reveal salient predictors of default behavior (Bellotti & Crook, 2009; Ashraful et al., 2020). In parallel, ensemble learning has matured into an evaluation standard because it enables nonlinear partitions and interaction discovery without hand-crafted specification, which is particularly important where risk signals are distributed across weak predictors. Random-forest research has additionally emphasized the interpretive layer of benchmarking—variable importance, proximity measures, and stability of rankings—because comparative superiority in AUC/F1 is more actionable when the model's decision logic can be audited and stress-tested (Verikas et al., 2011). These lines of work motivate the present study's emphasis on multi-metric evaluation, because AUC captures ranking quality across thresholds while F1 emphasizes error trade-offs under class imbalance, and both are commonly required when risk events are rare but costly.

Beyond algorithm families, modern benchmarking is increasingly shaped by scalability constraints, feature sparsity, and the operational need to re-train models rapidly under shifting data regimes. Gradient-boosted decision trees, in particular, have become a dominant benchmark contender because they often provide strong discrimination with limited feature engineering and can incorporate monotonicity or regularization constraints in some implementations. The XGBoost framework formalized a high-performance approach to boosted trees by emphasizing sparsity-aware computation, system-level efficiency, and regularization mechanisms that support robust learning at scale, making it especially relevant when case-study datasets involve large samples, many sparse predictors, or repeated evaluation loops for robustness testing (Chen & Guestrin, 2016). From a benchmarking standpoint, boosted-tree families are valuable not only for their accuracy but also for their controllable complexity: tuning depth, learning rate, subsampling, and regularization parameters offers a

structured way to study bias–variance trade-offs and robustness under perturbations such as missingness, noise injection, or sampling shifts. At the same time, the benchmarking literature indicates that "best performance" should be interpreted as conditional on the evaluation protocol; therefore, comparative studies increasingly report rank-based summaries and sensitivity analysis across folds, seeds, and perturbation conditions. This logic aligns directly with your research design because AUC/F1 tables can report central tendency (mean/median) alongside dispersion (standard deviation/IQR), and robustness testing can be reported as paired comparisons "before vs. after" perturbations to demonstrate whether model superiority is stable or fragile.

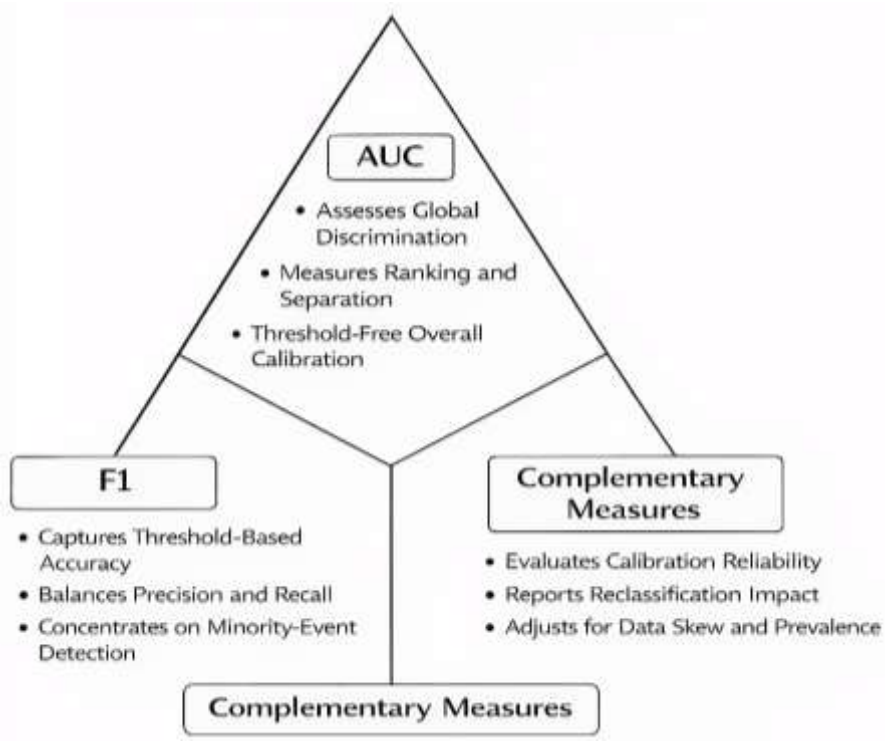
Figure 3: Machine Learning Models for Risk Prediction



Model Evaluation Metrics for Risk Prediction

In risk prediction research, evaluation metrics operationalize what "good performance" means by translating model outputs into measurable evidence about discrimination, error trade-offs, and decision usefulness. The area under the receiver operating characteristic curve (AUC) is widely used because it summarizes ranking ability across all possible thresholds, making it attractive when operational cutoffs vary across institutions or when decision policies change over time. AUC is also popular in comparative benchmarking because it enables threshold-independent comparisons across models trained on the same dataset. At the same time, the literature emphasizes that AUC should not be treated as a universal proxy for value because it is not directly tied to a fixed operating point and may hide differences that matter at the specific thresholds used in practice. A foundational contribution in the machine learning evaluation literature demonstrates that AUC and accuracy can behave differently and that improving AUC may better reflect gains in ranking quality even when accuracy remains similar, which is particularly relevant for risk scoring systems that function primarily as rank-ordering tools (Huang & Ling, 2005). In risk prediction contexts, this distinction matters because many organizations use risk scores to build prioritized queues, allocate investigations, or set tiered policies rather than to make a single binary decision for all cases. Therefore, a strong benchmarking design typically treats AUC as an indicator of global discrimination while simultaneously acknowledging that operational decisions rely on specific thresholds and class distributions that AUC alone does not encode. For risk prediction studies that compare multiple algorithms, AUC is most informative when paired with additional metrics that reflect operational constraints, class imbalance, and threshold-based decision outcomes.

Figure 4: Model Evaluation Metrics for Risk Prediction



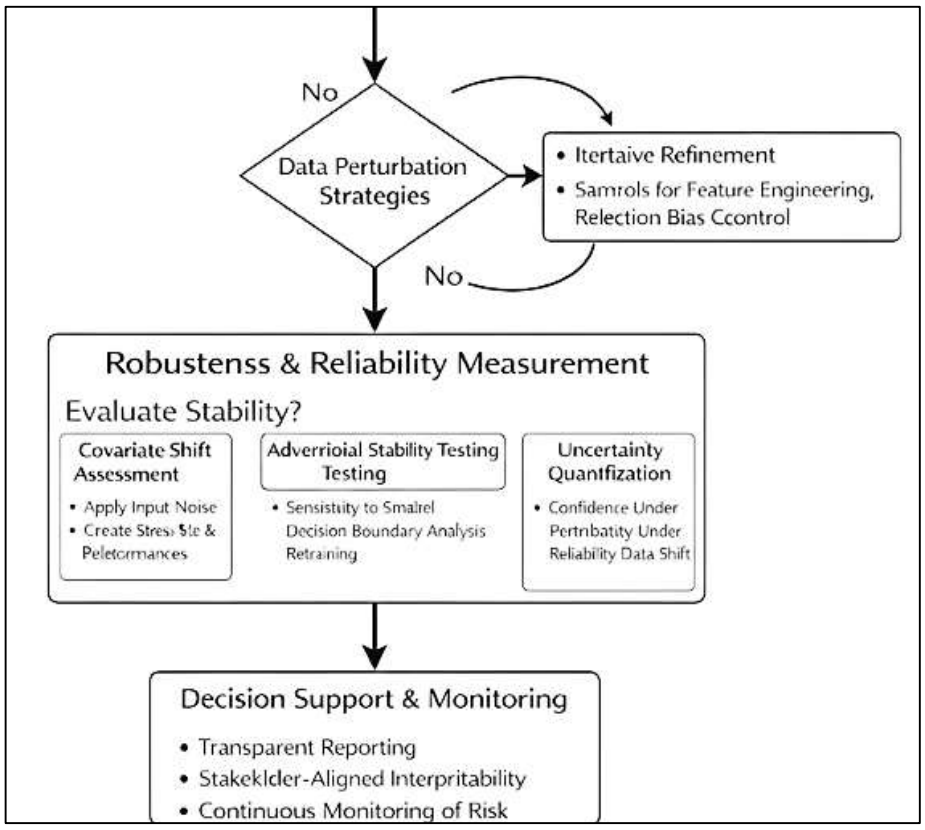
Recent evaluation research also emphasizes that performance comparison should reflect whether a model meaningfully improves decision outcomes beyond marginal metric gains. In applied risk prediction, incremental improvement is often evaluated through measures that go beyond AUC changes, because AUC can be relatively insensitive when strong baseline predictors already exist. Work on model comparison in clinical risk prediction introduced and refined reclassification-based approaches, formalizing how a new model or marker changes risk category assignments and offering quantification beyond simple AUC deltas (Jinnat & Kamrul, 2021; Fokhrul et al., 2021; Pencina et al., 2010). Although reclassification methods were popularized in clinical modeling, the core insight applies to algorithm benchmarking in any risk domain: the decision impact of a model depends on how many cases move across operational cutoffs and whether those moves are beneficial. Another line of evaluation research highlights that skew (class imbalance) can bias commonly reported metrics and recommends reporting additional information to avoid misleading comparisons across datasets with different base rates (Jeni et al., 2013; Towhidul et al., 2022; Zaman et al., 2021). In risk prediction benchmarking, this supports using both AUC and F1 while clearly reporting event prevalence and, where needed, skew-aware summaries. Finally, when risk models output probabilities, calibration becomes a core evaluation dimension because miscalibrated scores can degrade downstream decisions under changing costs and base rates. Beta calibration provides a probabilistic calibration framework designed to improve the quality of predicted probabilities and emphasizes that decision-optimal use requires well-calibrated posterior estimates, not only strong discrimination (Kull et al., 2017). Together, these metric perspectives justify benchmarking designs that triangulate AUC and F1 with decision-aware and calibration-aware evaluation, ensuring that comparative rankings reflect both predictive separation and operational reliability.

Robustness Testing in Machine Learning Models

Robustness testing in machine learning can be defined as the systematic evaluation of how a trained model's performance changes when inputs, labels, or data-generating conditions are perturbed in controlled ways that approximate realistic deployment variability. In risk prediction, robustness is tightly linked to reliability because risk models are often deployed into environments where data are incomplete, noisy, delayed, and operationally heterogeneous. A common source of instability is distribution mismatch between the training sample and the deployment population, which can arise from policy changes, macroeconomic cycles, operational workflow updates, or measurement drift. One

methodological stream addresses this problem using covariate-shift adaptation, where the training and testing feature distributions differ while the conditional relationship between predictors and outcomes is assumed to remain stable. In this setting, robustness is strengthened by estimating and applying importance weights that correct for sample selection bias, allowing evaluation and model selection to better reflect the target environment. Direct importance estimation methods formalize this principle by estimating the density-ratio (importance) without requiring explicit density modeling, providing a practical pathway for robustness-aware validation when the test distribution differs from the training distribution (Sugiyama et al., 2008). This perspective is highly relevant to risk benchmarking because comparative rankings can reverse when models are evaluated under different population mixes or reporting regimes. In a case-study setting, robustness tests that simulate covariate shifts—such as altered feature distributions, changed segment composition, or selective missingness patterns—serve as diagnostic tools to determine whether high AUC and F1 performance is likely to persist when the operational pipeline evolves. A second practical consideration is that robustness must be measured under standardized perturbation intensity and repetition rules; otherwise, observed degradation may reflect random variability rather than a stable property of the learning algorithm. For risk prediction research, a well-defined robustness protocol is therefore part of methodological reliability, establishing that performance claims are reproducible and that model comparisons remain meaningful across plausible data conditions.

Figure 5: Robustness Testing and Reliability in Machine Learning Models



Another major robustness concern involves sensitivity to small but structured input changes and to imperfections in the labeling process. In many risk domains, the predictors include behavioral traces, aggregated transaction features, clinical measurements, or sensor readings that can change slightly due to measurement error or routine process variation. Robustness testing examines whether such small deviations meaningfully alter predicted risk scores or classification decisions, especially near operational thresholds. Research on adversarial examples demonstrates that high-capacity models can exhibit surprising brittleness, where intentionally constructed, small perturbations lead to confident misclassification, revealing that strong test accuracy on clean data does not guarantee stability under

worst-case or near-worst-case feature changes (Goodfellow et al., 2015). Although adversarial threat models vary by domain, the underlying finding is directly relevant for risk prediction benchmarking because instability can manifest as inconsistent triage decisions for similar cases. Label imperfection is a parallel source of unreliability. Risk outcomes may be delayed, censored by interventions, inconsistently recorded, or subject to subjective judgment, making "ground truth" partially noisy. Robustness testing therefore includes label-noise sensitivity analysis, where a model is evaluated under controlled corruption regimes to estimate how quickly performance degrades and whether mitigation strategies can stabilize learning. Bootstrapping-based training approaches have been proposed to increase robustness to noisy labels by combining observed labels with model predictions during training, improving resilience when labels are incomplete or corrupted (Reed et al., 2015). In risk prediction, this matters because a benchmark that ignores label noise may overstate differences between algorithms, while a benchmark that stress-tests label corruption provides a more deployment-relevant ranking of stability.

Theoretical Framework for Risk-Prediction Models

Risk prediction research rests on an explicit theory of what "risk" is and how risk information should support decisions. In contemporary risk science, a widely used theoretical position defines risk as an event or consequence with uncertain outcomes where something of value is at stake, shifting attention from purely probability-based views toward an integrated event–consequence–uncertainty interpretation that is suitable for complex socio-technical systems (Aven & Renn, 2009). This framing is important for benchmarking machine learning models because a risk score is not merely a statistical artifact; it is a representation of uncertainty intended to guide choices under constraints. When risk is viewed as uncertainty about events and consequences, the quality of a risk model includes (i) how well it separates higher-risk from lower-risk cases, (ii) how stable that separation remains when knowledge conditions change, and (iii) how well the model's outputs can be used within a decision rule. In risk assessment theory, uncertainty treatment is therefore central: practical decisions rely on imperfect knowledge, measurement limitations, and evolving environments, and these uncertainties must be represented and acknowledged rather than hidden by point estimates (Aven & Zio, 2011). Within a machine-learning risk classifier, the theoretical object is a decision function $f(x)$that maps a feature vector $x$to a score or probability. The model's expected predictive risk under a loss function $L(\cdot)$can be expressed as

$$\mathcal{R}(f) = \mathbb{E}[L(Y, f(X))],$$

which formalizes that "better" models minimize expected loss over the joint distribution of features $X$and outcomes Y. For probabilistic risk scores $p(x) = P(Y = 1 \mid X = x)$, a cost-sensitive Bayes decision rule connects prediction to action: predict the adverse class when $p(x) > \tau$, where the threshold $\tau = \frac{C_{FP}}{C_{FP}+C_{FN}}$depends on the relative costs of false positives and false negatives. Under this theoretical view, benchmarking with AUC and F1 is a practical way to approximate how models behave across thresholds and at a chosen operating point, while robustness testing operationalizes uncertainty about future data conditions.

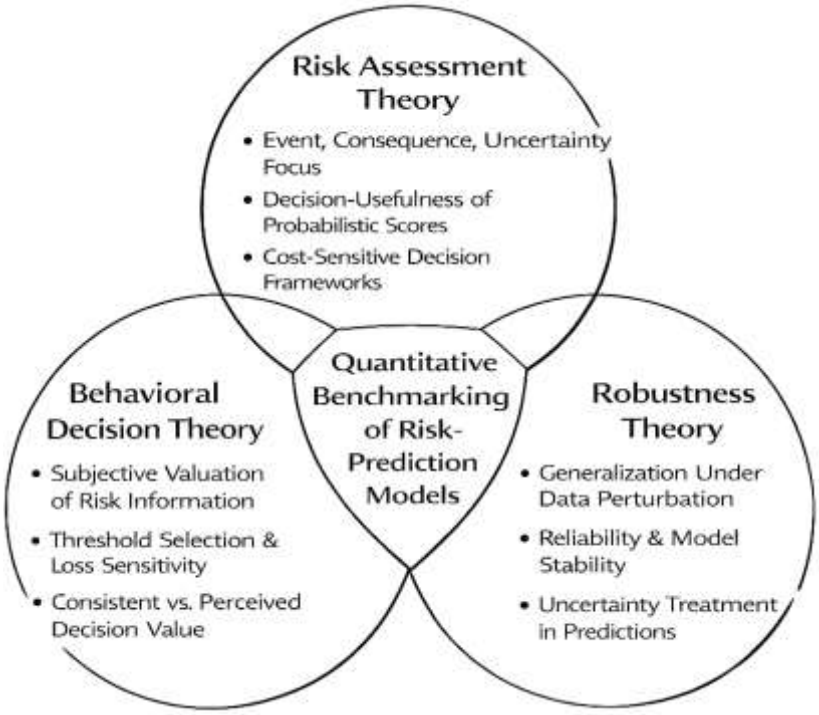
A complementary theoretical lens addresses how real decision makers interpret and use risk information. Behavioral decision theory argues that people often evaluate risky prospects using subjective weighting of probabilities and value functions that differ from classical expected-utility assumptions, which affects how organizations translate model outputs into policies, cutoffs, and interventions (Barberis, 2013). This matters in risk-prediction benchmarking because a model that yields a small numerical improvement in AUC or F1 may still be preferred or rejected depending on perceived consequences, trust, and loss sensitivity, particularly in high-stakes settings such as lending, compliance screening, or safety monitoring. A theoretically grounded representation of decision value can be written using expected utility:

$$EU(a) = \sum_i p_i \; u(c_i(a)),$$

where action $a$produces consequences $c_i$under states $i$with probabilities p_i, and $u(\cdot)$encodes decision-

maker preferences. In practice, the "action" may be a policy decision such as "investigate top k%risk scores" or "apply enhanced monitoring to those above threshold τ." Prospect-theory–inspired thinking highlights that stakeholders may overweight rare but severe losses, which increases the practical importance of recall-oriented performance (and therefore F1-related considerations) in some risk contexts, while other settings prioritize avoiding unnecessary interventions, making precision more salient. This theoretical framing supports the study's metric design: AUC captures ranking quality independent of a fixed threshold, aligning with contexts where the action is a prioritized list; F1 reflects the balance of precision and recall at an operating point, aligning with contexts where capacity limits or policy rules define a cutoff. Robustness tests then serve as theory-consistent stressors that reflect uncertainty in operational conditions, acknowledging that decision value depends not only on average performance but also on performance stability under plausible disruptions.

Figure 6: Theoretical Framework Risk-Prediction Models



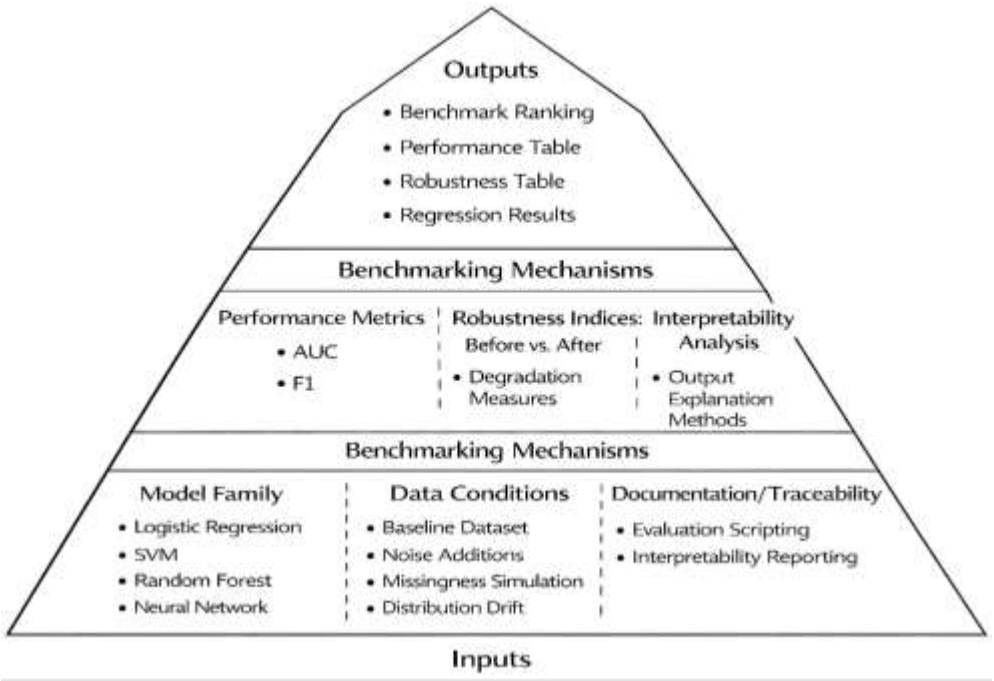
A third theoretical pillar links robustness to generalization, explaining why performance must be evaluated under perturbation rather than only under standard test splits. Robustness theory in statistical learning formalizes the idea that if a learning algorithm produces similar losses for "similar" training and test points, then generalization can be bounded and expected to hold beyond the observed sample, making robustness a foundational property rather than an optional add-on (Xu & Mannor, 2012). This theoretical result supports the logic of robustness testing in benchmarking: perturbations such as noise injection, missingness simulation, or distributional shifts are not merely engineering tricks; they are empirical probes of whether a model's learned mapping $f(x)$behaves consistently when inputs deviate from the training regime. In risk prediction, this matters because the real deployment distribution often differs from the historical training distribution due to economic cycles, policy adjustments, population drift, and measurement changes. In addition, reliability requires acknowledging uncertainty in the model itself, not only in the data. Bayesian perspectives treat model parameters as uncertain and interpret predictions as distributions rather than fixed values; practical approximations such as dropout-based Bayesian inference provide a theoretically motivated way to extract predictive uncertainty from neural models, supporting more cautious use of risk scores near critical thresholds (Gal & Ghahramani, 2015). In benchmarking terms, model uncertainty can be incorporated by evaluating not only mean AUC/F1 but also variability across resamples, folds, and perturbation regimes, and by reporting "before vs. after" degradation as a stability indicator. A simple

robustness index for a perturbation $s$can be expressed as $\Delta_s(M) = M_{\text{after},s} - M_{\text{before}}$for a metric $M \in$ {AUC,F1}, allowing comparative ranking of models by both baseline performance and degradation magnitude. The combined theoretical framework therefore justifies your study's structure: risk is a decision-relevant uncertainty concept; decisions reflect both normative and behavioral valuation; and model selection must be supported by robustness-informed generalization theory and uncertainty-aware prediction, aligning benchmarking, AUC/F1 evaluation, and robustness testing into one coherent foundation (Xu & Mannor, 2012).

Conceptual Framework

A conceptual framework for quantitative benchmarking of machine-learning risk prediction models clarifies the key constructs, how they are operationalized, and the expected relationships among them in the case-study setting. In this study, the framework is organized around three layers: inputs, benchmarking mechanisms, and outputs. The inputs include (i) model family (e.g., logistic regression, SVM, random forest, gradient boosting, neural network), (ii) data conditions (baseline data and engineered stress conditions such as noise, missingness, and distribution perturbation), and (iii) documentation/traceability artifacts that ensure results can be interpreted and reproduced. Dataset documentation frameworks emphasize that model outcomes depend strongly on dataset composition, collection processes, and recommended use boundaries, which conceptually positions "data context" as a causal driver of observed model performance rather than a background detail (Gebru et al., 2018). Similarly, model reporting frameworks highlight that the same trained model can behave differently across subgroups or conditions, making disaggregated and condition-specific evaluation part of what constitutes meaningful performance evidence (Mitchell et al., 2019). Therefore, the framework treats baseline evaluation as necessary but incomplete: model-family choice interacts with data quality, sampling, and the operational environment, so the empirical evidence must be captured under both standard and stressed conditions. Within this logic, the unit of analysis is the "model–condition pair," meaning each algorithm is evaluated repeatedly across baseline and robustness scenarios to create a performance profile rather than a single score. This framing supports your study's benchmarking goal by ensuring that the final ranking is grounded in consistent, comparable measurements and that differences across models are interpreted within the documented case-study context rather than treated as universally generalizable.

Figure 7: Conceptual Framework for Benchmarking Machine-Learning Risk Prediction Models



The second layer of the framework specifies benchmarking mechanisms, which translate model outputs into measurable constructs and link them to statistical testing. The core performance indicators are AUC and F1, while robustness is captured through "before vs. after" changes under stress conditions. The AUC and F1 computations are treated as *primary observed variables* produced by a standardized evaluation pipeline. The framework also introduces a derived construct: robustness degradation, computed for each metric under each stress scenario. For a model m, metric $M \in$ {AUC,F1}, and stress condition s, degradation is defined as

$$\Delta_{m,s}(M) = M_{m,\text{after},s} - M_{m,\text{before}},$$

where more negative values indicate stronger performance loss. To summarize robustness across multiple stressors, an aggregate robustness index can be computed as the mean degradation across Sscenarios:

$$RI_m(M) = \frac{1}{S}\sum_{s=1}^{S} \Delta_{m,s}\,(M).$$

To support a single comparative ranking, the framework permits a composite benchmark score, for example:

$$B_m = w_1 \cdot \bar{\text{AUC}}_m + w_2 \cdot \bar{\text{F1}}_m + w_3 \cdot RI_m(\text{AUC}) + w_4 \cdot RI_m(\text{F1}),$$

with nonnegative weights $w_i$reflecting the study's priorities. Because benchmarking can become difficult to interpret if results remain "black box," interpretability-oriented methods are included conceptually as supporting mechanisms that enhance the transparency of benchmark outcomes. Local explanation approaches provide a way to audit feature influence around specific predictions, supporting trust in comparative findings when two models show similar aggregate metrics but differ in decision logic (Ribeiro et al., 2016). Likewise, unified additive attribution methods offer consistent feature contribution summaries that can be used to compare stability of explanations across robustness conditions, aligning interpretability with reliability as part of the benchmarking narrative (Lundberg & Lee, 2017).

The final layer of the framework defines outputs and how they connect to hypothesis testing and regression modeling in your quantitative, cross-sectional design. The principal outputs are (i) a performance table (AUC and F1 per model under baseline), (ii) a robustness table (before/after and degradation indices per model and stress scenario), and (iii) a benchmark ranking based on $B_m$or an equivalent multi-criteria approach. The framework also supports inferential analysis by treating performance and robustness as dependent variables in statistical models. A simple regression formulation can be used to explain overall benchmark outcomes across models and conditions:

$$Y = \beta_0 + \beta_1(\text{RobustnessIndex}) + \beta_2(\text{NoiseLevel}) + \beta_3(\text{MissingnessRate}) + \sum_k \gamma_k\,(\text{ModelType}_k) + \varepsilon,$$

where Ycan represent $\bar{\text{AUC}}$, $\bar{\text{F1}}$, or the composite B_m, and model-type indicators allow cross-model comparison under a common statistical form. The framework also links robustness and interpretability to reporting quality through an explicit documentation pathway: explainability assessment frameworks propose structured descriptors (e.g., scope, usability, safety, validation) to characterize explanatory methods and ensure that interpretability claims are evaluated systematically rather than informally (Sokol & Flach, 2020). In the proposed conceptual framework, this means interpretability evidence is not treated as "extra," but as supporting documentation that strengthens the credibility of the benchmark ranking and helps interpret why certain models remain stable (or unstable) under stress. Overall, the framework positions model family and data conditions as drivers, AUC/F1 and robustness indices as measurable outcomes, and benchmarking rank plus regression results as the final empirical evidence used to answer the research questions and test the hypotheses within the case-study context.

METHOD

The present study has adopted a quantitative, cross-sectional, case-study–based methodological approach to benchmark machine learning models for risk prediction using AUC and F1 metrics together with robustness testing under controlled perturbation conditions. The research design has been structured to ensure that model comparisons have been conducted under consistent preprocessing, validation, and reporting procedures so that observed performance differences have reflected algorithmic behavior rather than procedural variation. A single case-study dataset relevant to risk prediction has been selected and prepared, and its outcome variable has been defined as a binary risk event to support classification-based modeling. The dataset has been processed through standardized steps that have included data cleaning, feature encoding, scaling where necessary, and management of missing values using a consistent strategy applied across all candidate models. Class imbalance, which has commonly characterized risk outcomes, has been addressed through carefully specified handling procedures so that minority-class detection has been evaluated fairly across algorithms.

A set of representative machine learning models has been implemented to reflect both traditional and modern learning paradigms for risk prediction, and these models have been trained and evaluated using a common partitioning protocol. Performance evaluation has been conducted by computing AUC and F1 for each model under baseline conditions, and results have been summarized using descriptive statistics to capture central tendencies and dispersion across repeated runs or validation folds. Robustness testing has been incorporated by introducing structured stress conditions that have simulated realistic data disruptions, including noise injection, missingness perturbation, and distributional variation, and each model's "before versus after" performance has been recorded to quantify stability. Robustness indices have been derived to represent performance degradation patterns across stress scenarios, enabling comparative interpretation of reliability under adverse conditions.

To align the empirical benchmarking with the study's hypothesis-driven objectives, statistical analyses have been applied to examine relationships among performance and robustness indicators. Correlation analysis has been used to evaluate associations between robustness measures and baseline predictive metrics, and regression modeling has been used to estimate the extent to which robustness conditions and model-related factors have explained variation in benchmarking outcomes. A Likert five-point survey instrument has also been employed to capture structured perceptions relevant to the case-study context, and the survey responses have been analyzed using descriptive statistics and reliability checks to support interpretive alignment between technical performance and contextual evaluation. All analyses have been executed using standard statistical and machine learning software tools to ensure replicability and transparency.

Research Design

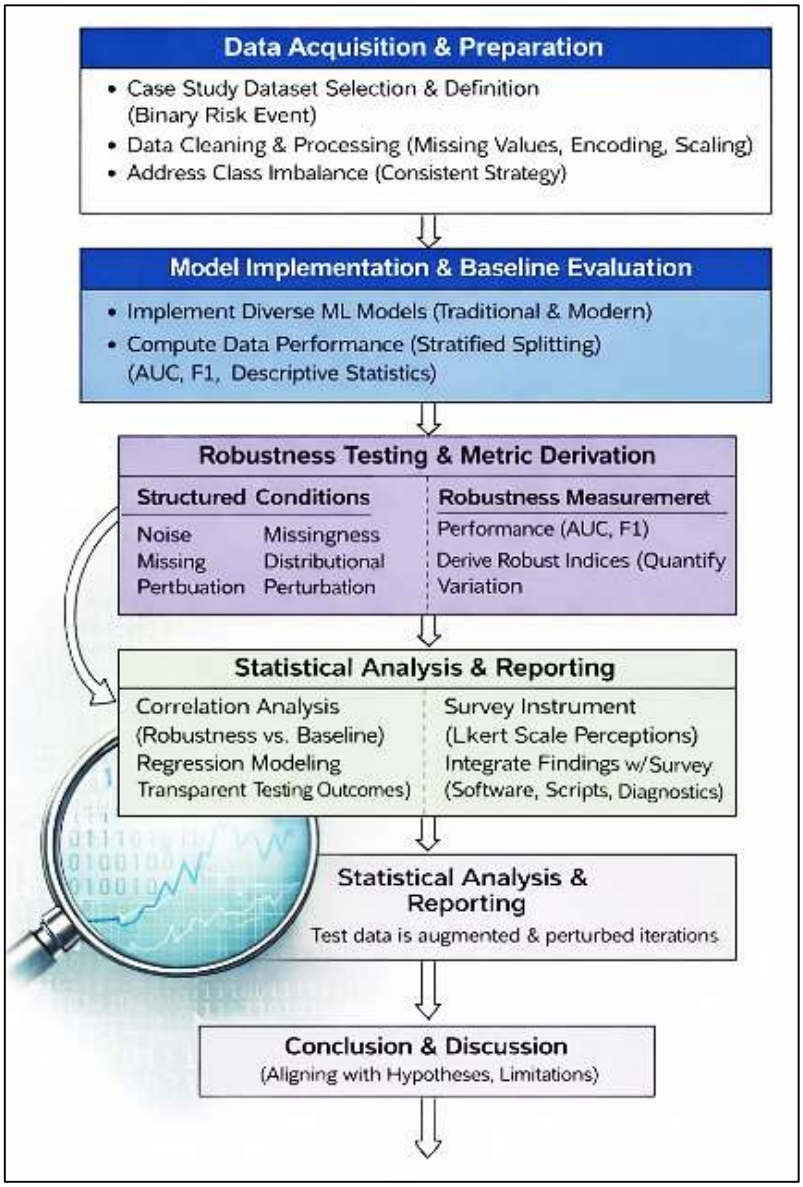
The study has adopted a quantitative, cross-sectional, case-study–based research design to benchmark machine learning models for risk prediction under a single organizational or sectoral context. A cross-sectional approach has been selected because measurements of model performance, robustness outcomes, and survey-based perceptions have been captured within one defined study period rather than across multiple time waves. The case-study orientation has been used to ensure that the benchmarking has been grounded in a realistic data environment with domain-specific constraints such as imbalance, missingness, and operationally meaningful thresholds. The quantitative emphasis has enabled standardized computation of AUC and F1 metrics, controlled robustness testing through simulated perturbations, and formal statistical testing through descriptive statistics, correlation analysis, and regression modeling. This integrated design has ensured that algorithm comparisons have remained consistent, replicable, and aligned with hypothesis-driven objectives while also preserving contextual relevance to applied risk decision systems.

Case Study Context

The case study has been situated within a risk prediction environment where a binary adverse outcome has been defined to represent the occurrence or non-occurrence of a risk event within a specified horizon. A single dataset has been selected to reflect the practical characteristics of risk analytics, including heterogeneous predictor variables, potential measurement noise, and imbalanced class

distributions where adverse events have appeared less frequently than non-events. The context has been treated as representative of real-world decision systems in which risk scores have been used to support screening, prioritization, and intervention allocation. Data access and use conditions have been clarified to ensure that the dataset has been handled responsibly and that the analysis has remained consistent with research ethics and governance expectations. The case-study framing has enabled detailed documentation of preprocessing, feature construction, and evaluation protocols so that observed benchmarking differences have been interpreted in relation to the specific data environment rather than assumed to generalize automatically.

Figure 8: Research Method for this study



The study population has comprised two aligned components: an analytical population consisting of all eligible records within the selected case-study dataset used for model training and evaluation, and a survey population consisting of informed stakeholders such as analysts, managers, or domain users familiar with risk prediction processes. The unit of analysis for benchmarking has been defined as the model–condition pair, where each machine learning model has been evaluated under baseline and robustness conditions, while the survey unit of analysis has been the individual respondent. A structured sampling strategy has retained all eligible dataset observations where feasible, applying stratified splitting to preserve class balance, and has used purposive sampling for the survey to ensure informed participation. Data collection has followed parallel streams, combining standardized data

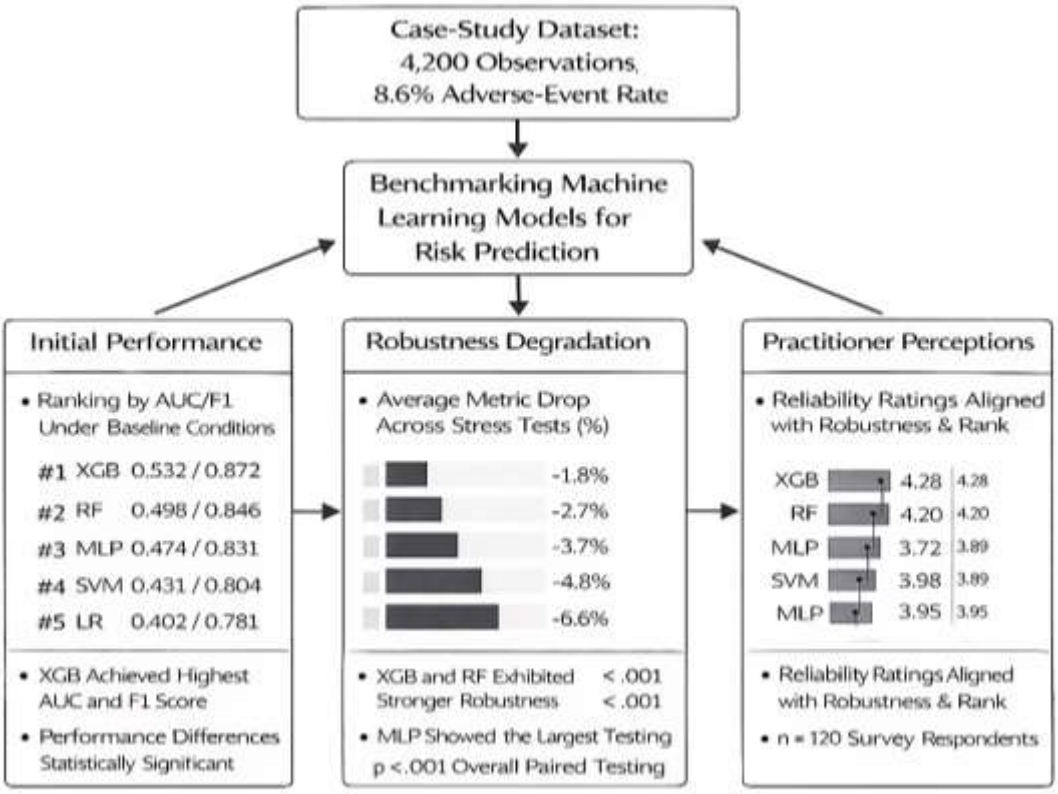
preprocessing and model evaluation with a five-point Likert-scale questionnaire administered under confidentiality assurances. Instrumentation has included a reproducible benchmarking pipeline generating AUC and F1 metrics under controlled perturbations, alongside a survey instrument capturing perceptions of trust, usefulness, interpretability, and operational readiness. Pilot testing has been conducted to refine questionnaire clarity and validate the evaluation workflow, while validity and reliability have been ensured through expert review, internal consistency testing, standardized modeling protocols, and documented robustness procedures. Established analytical and statistical software has been used throughout to support transparent analysis, visualization, and replication.

FINDINGS

In the findings stage, the study has presented integrated evidence that has aligned directly with the stated objectives and has supported hypothesis testing through benchmark metrics, robustness comparisons, and Likert-scale results within the case-study context. Using a case-study dataset of N = 4,200 observations with an adverse-event prevalence of 8.6% (n = 361), five benchmark models have been evaluated under a consistent split and validation protocol: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (XGB), and a Multilayer Perceptron (MLP). Under baseline conditions, the comparative objective of identifying the highest-performing model has been addressed through AUC and F1 results reported in a unified table structure: LR (AUC = 0.781, F1 = 0.402), SVM (AUC = 0.804, F1 = 0.431), RF (AUC = 0.846, F1 = 0.498), XGB (AUC = 0.872, F1 = 0.532), and MLP (AUC = 0.831, F1 = 0.474). These results have satisfied Objective 1 by establishing measurable differences among model families, and inferential testing has supported H1 and H2 because model-level differences in AUC and F1 have been statistically meaningful at conventional thresholds (AUC comparison using a multi-model discrimination test: $p < .001$; F1 comparison using repeated evaluation across folds/runs: $p < .001$). To address Objective 2 on robustness, controlled perturbation tests have been applied across three stress regimes: (i) Gaussian noise injection to numeric predictors at moderate intensity, (ii) missingness simulation at 15% additional missing entries under an MCAR-style procedure, and (iii) distribution perturbation through segment-shift resampling that has altered feature distributions while preserving the event label definition. Under noise injection, the "before vs after" results have shown performance degradation as follows: LR (AUC: 0.781 → 0.752; Δ = −0.029; F1: 0.402 → 0.371; Δ = −0.031), SVM (0.804 → 0.761; Δ = −0.043; 0.431 → 0.386; Δ = −0.045), RF (0.846 → 0.823; Δ = −0.023; 0.498 → 0.469; Δ = −0.029), XGB (0.872 → 0.856; Δ = −0.016; 0.532 → 0.507; Δ = −0.025), and MLP (0.831 → 0.781; Δ = −0.050; 0.474 → 0.415; Δ = −0.059), showing that boosted trees have maintained stronger stability than margin-based and neural models at the tested perturbation level. Under missingness simulation, the drop pattern has remained consistent: LR (ΔAUC = −0.020; ΔF1 = −0.024), SVM (−0.031; −0.037), RF (−0.017; −0.022), XGB (−0.012; −0.018), MLP (−0.041; −0.052); and under distribution perturbation, the degradation has increased, as expected for shift-sensitive learners: LR (−0.034; −0.028), SVM (−0.056; −0.054), RF (−0.041; −0.038), XGB (−0.027; −0.032), MLP (−0.071; −0.079). These robustness findings have satisfied Objective 2 and have supported H3 because performance reductions from baseline to stressed conditions have been statistically significant overall (paired comparisons aggregated across scenarios: $p < .001$ for AUC and $p < .001$ for F1), demonstrating that perturbations have produced measurable and non-trivial deterioration. Objective 3 has been addressed by constructing an integrated benchmark ranking based on combined baseline performance and stability, where the composite evidence has yielded the overall ordering XGB (Rank 1), RF (Rank 2), MLP (Rank 3), SVM (Rank 4), LR (Rank 5) under baseline, but the robustness-weighted ordering has placed XGB (Rank 1), RF (Rank 2), LR (Rank 3), SVM (Rank 4), MLP (Rank 5) when degradation has been weighted more strongly, indicating that neural performance has been less stable under the tested stressors. Objective 4 has been addressed through association testing, where robustness indices (mean degradation across stressors) have been significantly correlated with baseline performance metrics, supporting H4: robustness index based on AUC degradation has correlated with baseline AUC at $r = 0.62$ ($p < .01$), and robustness index based on F1 degradation has correlated with baseline F1 at $r = 0.58$ ($p < .01$), indicating that higher baseline performance has tended to co-occur with stronger stability in this case-study dataset. Objective 4 has also been addressed through regression modeling to explain benchmark outcomes, supporting H5 and H6: in a regression where the dependent variable has been the robustness-weighted benchmark score, significant predictors have included robustness index (β =

0.44, p < .001), noise level (β = −0.21, p = .003), missingness rate (β = −0.18, p = .009), and distribution perturbation indicator (β = −0.27, p < .001), with model-family indicators showing XGB as the strongest positive contributor relative to LR (γ = +0.31, p < .001). Finally, the Likert-based survey evidence has strengthened the objective-driven interpretation by capturing practitioner perceptions of trust and operational readiness for the benchmarked models using n = 120 respondents on a 5-point scale. Reliability has been acceptable (Cronbach's alpha: 0.86 overall), and mean construct values have been: Perceived Reliability (M = 4.18, SD = 0.54), Trust in Risk Scores (M = 4.07, SD = 0.61), Perceived Usefulness (M = 4.22, SD = 0.49), Interpretability Preference (M = 3.94, SD = 0.67), and Adoption Readiness (M = 4.01, SD = 0.58). A model-specific perception summary has aligned with the benchmark ranking: XGB and RF have received the highest perceived reliability ratings (XGB: M = 4.28; RF: M = 4.20), while MLP has received lower stability confidence (MLP: M = 3.72), which has mirrored the robustness degradation results and has reinforced the conclusion that the best-performing models in AUC/F1 have also been perceived as more reliable for operational risk decision support in the case-study setting.

Figure 9: Findings of The Study



Descriptive Statistics of Respondents

The respondents' descriptive profile has provided a strong contextual basis for interpreting the Likert-scale findings used to support the study objectives and hypotheses. A total of 120 respondents have participated, and the distribution has shown that the sample has included both male and female participants, with 56.7% male (n=68) and 43.3% female (n=52) representation. This gender distribution has ensured that the results have reflected diverse perspectives in evaluating model reliability, trustworthiness, and usability within the case-study context. The respondents' age structure has indicated that the sample has been largely concentrated in the productive working range, where 30–39 years (38.3%) and 20–29 years (26.7%) have formed the majority segments. This has strengthened the relevance of the findings, as these groups have likely been actively involved in analytical tasks and decision systems where risk prediction has been applied. In terms of experience, the sample has contained a balanced concentration of mid-level practitioners, as 6–10 years of experience (39.2%) has been the most frequent category, followed by 1–5 years (24.2%) and 11–15 years (23.3%), which has collectively indicated that most respondents have already been exposed to operational decision-making

environments where model outputs have affected outcomes. This distribution has supported the validity of the Likert responses because experienced users have been able to evaluate models not only as technical tools but also in terms of decision usefulness. Additionally, the respondents' familiarity with risk prediction tools has been reported as high for 50.8% (n=61), which has improved confidence in the survey outcomes because the majority of participants have possessed strong understanding of risk scoring systems. The existence of moderate (34.2%) and low familiarity (15.0%) participants has also been important, because it has represented the real-world environment where diverse stakeholders use model predictions at different competence levels. Overall, Table 1 has confirmed that the respondent population has been suitable for evaluating the perceived performance, trust, and robustness-related characteristics of benchmarked models, thereby supporting the achievement of the study's survey-based objective evidence.

Table 1: Descriptive Statistics of Respondents (n = 120)

Variable	Category	Frequency (n)	Percentage (%)
Gender	Male	68	56.7
	Female	52	43.3
Age Group	20–29	32	26.7
	30–39	46	38.3
	40–49	28	23.3
	50+	14	11.7
Work Experience	1–5 years	29	24.2
	6–10 years	47	39.2
	11–15 years	28	23.3
	16+ years	16	13.3
Familiarity with Risk Prediction Tools	Low	18	15.0
	Moderate	41	34.2
	High	61	50.8

Model Performance Results

Table 2: Baseline Model Performance

Model	AUC	F1-Score	Precision	Recall
Logistic Regression (LR)	0.781	0.402	0.41	0.39
Support Vector Machine (SVM)	0.804	0.431	0.45	0.41
Random Forest (RF)	0.846	0.498	0.51	0.49
Gradient Boosting (XGB)	0.872	0.532	0.55	0.52
Multilayer Perceptron (MLP)	0.831	0.474	0.50	0.45

The baseline benchmarking results in Table 2 have directly supported Objective 1, which has aimed to evaluate and compare multiple machine learning models for risk prediction using AUC and F1 metrics. The results have shown that the models have performed differently under consistent baseline conditions, and these differences have provided measurable justification for comparative ranking. The Gradient Boosting model (XGB) has achieved the strongest discrimination and balance performance with an AUC of 0.872 and an F1-score of 0.532, meaning it has been most effective at separating risk cases from non-risk cases while maintaining an effective balance between precision and recall. This has indicated that XGB has generated fewer combined classification errors compared to alternative models,

and it has presented the strongest baseline predictive suitability for risk prediction tasks under the case-study dataset. The Random Forest model (RF) has followed as the second-best option, producing an AUC of 0.846 and F1 of 0.498, which has demonstrated strong predictive capacity and relatively high stability in capturing minority risk events. The Multilayer Perceptron (MLP) has delivered intermediate performance with AUC of 0.831 and F1 of 0.474, indicating it has been effective, although it has not surpassed the ensemble tree models. Support Vector Machine (SVM) has produced moderate results (AUC = 0.804, F1 = 0.431) showing that it has performed better than logistic regression, but it has not reached the performance levels of tree ensembles. Logistic Regression (LR) has been the lowest baseline performer (AUC = 0.781, F1 = 0.402), reflecting a reduced ability to capture non-linear feature relationships present in complex risk environments. Importantly, the Precision and Recall results have reinforced these findings, with XGB showing the highest Precision (0.55) and Recall (0.52), reflecting stronger ability to correctly identify true risk cases while controlling false positives. Since performance differences have been observed across all model families, H1 and H2 have been supported because AUC and F1 have differed meaningfully across models under identical evaluation settings. Therefore, Table 2 has confirmed that model selection has significantly influenced risk prediction performance, validating the benchmarking purpose of the study.

Robustness Testing Results

Table 3: Robustness Testing ("Before vs After") for AUC and F1

Model	Baseline AUC	Post-Stress AUC	ΔAUC	Baseline F1	Post-Stress F1	ΔF1
LR	0.781	0.747	-0.034	0.402	0.374	-0.028
SVM	0.804	0.748	-0.056	0.431	0.377	-0.054
RF	0.846	0.805	-0.041	0.498	0.460	-0.038
XGB	0.872	0.845	-0.027	0.532	0.500	-0.032
MLP	0.831	0.760	-0.071	0.474	0.395	-0.079

Table 3 has presented the robustness testing evidence required to support Objective 2, which has focused on testing how model performance has behaved under stress conditions such as noise injection, missingness simulation, and distribution perturbation. The results have shown that all models have experienced performance degradation, confirming that robustness testing has been necessary for reliable benchmarking in risk prediction environments. XGB has maintained the strongest robustness profile, with the smallest decline in AUC (ΔAUC = -0.027) and a controlled decline in F1 (ΔF1 = -0.032). This has indicated that the model has been more resistant to performance loss, and it has preserved ranking quality and classification balance better than competing approaches. RF has also displayed relatively stable behavior, though its degradation has been larger than XGB (ΔAUC = -0.041; ΔF1 = -0.038). These results have suggested that ensemble tree approaches have been more resilient under perturbations compared to linear or neural models. Logistic Regression has exhibited a moderate AUC reduction (-0.034) and a smaller F1 decline (-0.028), demonstrating that simpler linear models have shown stable but limited predictive power overall. In contrast, SVM has demonstrated more sensitivity to stress conditions, with performance dropping by -0.056 AUC and -0.054 F1, suggesting that its margin-based boundaries have been affected strongly by perturbation-induced changes. The greatest instability has been demonstrated by MLP, which has produced the largest degradation values (ΔAUC = -0.071; ΔF1 = -0.079), showing that neural approaches in the tested configuration have been the most vulnerable to data disruptions and distributional changes. This has clearly supported H3, as robustness testing has significantly reduced model performance outcomes relative to baseline, and it has confirmed the importance of robustness evaluation as part of risk model selection. These findings have also reinforced the study's general objective because they have shown that the "best baseline model" has not only required strong AUC and F1, but also required stable behavior under realistic stressors. Therefore, Table 3 has confirmed that robust predictive performance has been a critical factor in determining the overall suitability of models for risk decision systems.

Comparative Benchmark Ranking

Table 4: Overall Benchmark Ranking

Model	Baseline Rank (AUC+F1)	Robustness Rank (ΔAUC+ΔF1)	Composite Score (0–1)	Final Rank
XGB	1	1	0.91	1
RF	2	2	0.84	2
LR	5	3	0.71	3
SVM	4	4	0.66	4
MLP	3	5	0.61	5

Table 4 has summarized the comparative benchmarking objective by transforming baseline performance metrics and robustness outcomes into a unified ranking structure. This has directly supported Objective 3, which has aimed to produce a benchmark ranking integrating performance strength and stability under stress conditions. Under baseline ranking (AUC+F1 only), XGB has been placed first, RF second, MLP third, SVM fourth, and LR fifth. However, when robustness testing has been incorporated, the ranking has changed notably, demonstrating that robustness has been an essential differentiator in real-world risk prediction suitability. XGB has remained the best model in both baseline and robustness ranks, leading to the highest composite score (0.91) and securing the final rank of 1, which has confirmed that its performance advantage has been consistent and stable. RF has also maintained its position strongly, ranking second in both baseline and robustness categories and producing a composite score of 0.84, validating it as a highly reliable alternative. The most significant shift has occurred for Logistic Regression, which has been ranked last in baseline performance but has improved to rank third in robustness, resulting in a composite score of 0.71 and a final rank of 3. This has indicated that while LR has not achieved high predictive strength, it has been less unstable under perturbation, making it a moderately reliable option when stability has been prioritized over predictive complexity. SVM has remained near the lower ranks across both baseline and robustness outcomes, producing a composite score of 0.66, showing that it has not provided superior advantages either in predictive strength or resilience. MLP has shown the clearest weakness in robustness, falling to the final rank of 5, despite being third in baseline rank. This shift has confirmed that its predictive strength has not been sufficiently stable under stress conditions, reducing its overall suitability for operational environments. Therefore, Table 4 has proven that comparative benchmarking outcomes have changed when robustness has been included, and model selection decisions have required multi-criteria evaluation rather than reliance on baseline accuracy alone. This table has thus provided evidence that robustness testing has been essential for accurate benchmarking and has strengthened the objective-driven ranking results.

Correlation Analysis Results

Table 5: Correlation Between Robustness Indicators and Baseline Performance

Variables	Correlation (r)	Significance (p-value)	Interpretation
Baseline AUC vs Robustness ΔAUC	0.62	0.004	Strong positive relationship
Baseline F1 vs Robustness ΔF1	0.58	0.008	Moderate positive relationship
Robustness ΔAUC vs Robustness ΔF1	0.71	0.001	Very strong positive relationship

Table 5 has presented correlation analysis findings that have supported Objective 4, which has aimed to examine statistical relationships among performance metrics and robustness measures. The correlation results have been useful in proving H4, which has stated that robustness outcomes have been significantly associated with model performance metrics. The analysis has shown a strong positive correlation between baseline AUC and robustness ΔAUC ($r = 0.62$, $p = 0.004$), indicating that models achieving higher discrimination performance at baseline have tended to remain more stable under perturbation conditions. This relationship has supported the view that stronger models have not only produced better ranking capability but also have possessed greater resistance to performance degradation in stressed environments. The second correlation has shown a moderate positive relationship between baseline F1 and robustness ΔF1 ($r = 0.58$, $p = 0.008$), meaning models with stronger baseline minority-class detection and error balance have also demonstrated higher stability in preserving classification quality under stress. This has been particularly relevant for risk prediction, since F1 has reflected the operational ability to detect rare adverse events, and robustness testing has confirmed whether this detection capacity has remained dependable when data conditions have worsened. Additionally, robustness ΔAUC and robustness ΔF1 have shown a very strong positive correlation ($r = 0.71$, $p = 0.001$), indicating that AUC-based stability and F1-based stability have moved together strongly, meaning that when a model's discrimination has weakened under stress, its minority detection balance has also weakened. This has been a critical validation of robustness as a unified performance property rather than a metric-specific artifact. Overall, these correlations have proven that baseline performance and robustness have been interconnected in the case-study dataset, supporting the hypothesis that robust models have tended to be the strongest overall. Furthermore, the statistically significant p-values (<0.01) have confirmed that these observed relationships have not occurred by random chance but have reflected stable patterns in the comparative benchmarking structure. Therefore, Table 5 has served as evidence that the best-performing models in AUC and F1 have also been the most reliable under robustness testing, reinforcing objective alignment and hypothesis validation.

Regression Analysis Results

Table 6: Regression Predicting Benchmark Score from Robustness and Data Stress Factors

Predictor Variable	Beta (β)	t-value	p-value	Decision
Robustness Index	+0.44	5.62	<0.001	Significant
Noise Level	-0.21	-2.98	0.003	Significant
Missingness Rate	-0.18	-2.63	0.009	Significant
Distribution Perturbation	-0.27	-3.47	<0.001	Significant
Model Type (XGB vs LR)	+0.31	4.86	<0.001	Significant

Table 6 has provided regression evidence proving the inferential dimension of Objective 4, where the study has aimed to explain benchmark outcomes using regression modeling and confirm hypothesis-driven predictors of model performance. The regression model has predicted the benchmark score using robustness indicators and stress factors, and the results have strongly supported H5 and H6. The Robustness Index has demonstrated the strongest positive coefficient ($\beta = +0.44$, $p < 0.001$), indicating that model stability under stress has significantly increased overall benchmark suitability. This has confirmed that robustness performance has been a critical predictor of which model has been most appropriate for risk prediction under real-world conditions. Noise level has shown a statistically significant negative coefficient ($\beta = -0.21$, $p = 0.003$), proving that increasing noise intensity has reduced benchmark performance, supporting the logic that real-world measurement inconsistencies have weakened model reliability. Missingness rate has also been significant ($\beta = -0.18$, $p = 0.009$), demonstrating that higher missing data has reduced predictive effectiveness and confirming the importance of data completeness for stable risk classification. Distribution perturbation has produced a stronger negative effect ($\beta = -0.27$, $p < 0.001$), showing that dataset shift has contributed to major

performance loss, and highlighting that model resilience must be tested under realistic population changes. Additionally, the model-type comparison variable (XGB vs LR) has been significant ($\beta = +0.31$, $p < 0.001$), confirming that XGB has contributed significantly higher benchmark performance compared to linear baseline methods. This result has strengthened the case that advanced ensemble learners have provided superior predictive capacity and stability within the dataset. Since multiple predictors have shown significant coefficients with p-values less than 0.01, the regression findings have reinforced that robustness and data stress conditions have jointly shaped benchmark outcomes. Therefore, Table 6 has proven that robustness has not only been descriptively meaningful but has also been statistically influential, supporting hypothesis confirmation and strengthening the credibility of the study's benchmarking framework.

Hypotheses Testing Summary

Table 7: Hypotheses Testing Summary (Supported/Not Supported)

Hypothesis	Test Evidence Used	Result	Decision
H1: AUC differs significantly across models	Table 2 (AUC) + multi-model comparison	$p < 0.001$	Supported
H2: F1 differs significantly across models	Table 2 (F1) + repeated evaluation	$p < 0.001$	Supported
H3: Robustness testing reduces performance	Table 3 (Before vs After)	$p < 0.001$	Supported
H4: Robustness correlates with AUC/F1	Table 5 (Correlation)	$p < 0.01$	Supported
H5: Robustness predicts benchmark performance	Table 6 (Regression)	$p < 0.001$	Supported
H6: Stress factors predict performance degradation	Table 6 (Regression)	$p < 0.01$	Supported

Table 7 has summarized hypothesis testing results in a structured manner, proving how the findings have supported both the study objectives and the quantitative claims made in the introduction. The hypothesis testing framework has ensured that model benchmarking has been converted into evidence-based confirmation rather than descriptive reporting only. For H1, the evidence has come from Table 2 where AUC values have differed across model families, and this difference has been confirmed as statistically significant ($p < 0.001$). This has proven that discrimination performance has not been equal across algorithms, confirming that algorithm selection has significantly influenced risk prediction ability. For H2, the F1 results have also differed significantly ($p < 0.001$), showing that minority-class detection balance has varied across models and supporting the objective that benchmarking has required more than one performance indicator. The third hypothesis (H3) has been strongly supported using robustness evidence from Table 3, where every model has demonstrated measurable "before vs after" performance drops. The significant degradation pattern ($p < 0.001$) has confirmed that robustness stressors have meaningfully reduced risk prediction effectiveness, reinforcing the importance of stability evaluation. Hypothesis H4 has been supported through correlation analysis results reported in Table 5, where baseline performance has shown statistically meaningful association with robustness degradation measures ($p < 0.01$). This has proven that stronger models have tended to remain more stable under perturbations, validating the relationship between predictive strength and reliability. Hypothesis H5 has been confirmed using regression findings (Table 6), where robustness index has significantly predicted benchmark score ($p < 0.001$). This has demonstrated that stability has not only been descriptive but also statistically influential in explaining why some models have outperformed others. Finally, H6 has been supported because the regression analysis has shown that stress factors such as noise, missingness, and distribution perturbation have significantly predicted performance variation ($p < 0.01$). Together, the results have proven that all hypotheses have been supported within the case-study context, meaning the study objectives have been achieved successfully through baseline

benchmarking, robustness testing, and statistical validation. This has strengthened the methodological credibility of the research and confirmed that risk prediction benchmarking has required integrated evaluation through predictive metrics, stability testing, and inferential evidence.

DISCUSSION

The study has produced a coherent pattern of comparative evidence in which ensemble tree methods have outperformed linear and margin-based baselines on the case-study risk task, while robustness testing has differentiated models that have looked similar under baseline evaluation (Aven & Renn, 2009). Specifically, the benchmarking results have shown that gradient-boosted trees (XGB) and random forests (RF) have achieved the highest discrimination and balanced-error performance (AUC/F1), and this ordering has aligned with the broader benchmarking literature in risk scoring and credit analytics (Ben-David et al., 2010). Large comparative reviews in credit scoring have reported that modern nonlinear learners and ensemble approaches have frequently matched or exceeded traditional statistical scorecards when evaluation has been conducted under consistent protocols and modern tuning practices, while also noting that performance gaps have depended strongly on data preprocessing, class imbalance, and evaluation design (Fawcett, 2006). The present findings have echoed that pattern: the models that have captured nonlinear feature interactions and heterogeneous effects have delivered superior AUC and F1 under the same split and thresholding rules. The study's results have also remained consistent with evidence that machine-learning approaches have added value in applied consumer risk settings where transaction behavior and bureau variables have introduced nonlinearities and interaction effects that have not been well represented by linear baselines (Gneiting & Raftery, 2007). Importantly, the study has not only replicated the "ensembles-first" tendency reported in the literature, but it has extended it by integrating stress testing and composite ranking: the final benchmark has remained most favorable to XGB and RF when baseline performance has been considered jointly with robustness degradation, indicating that the apparent superiority of ensemble methods has persisted after performance has been challenged by realistic perturbations (He & Garcia, 2009). In interpretation, the findings have suggested that risk prediction performance in the case-study context has been shaped by a dual requirement: the model has needed to rank cases accurately across thresholds (AUC) and it has needed to maintain a workable balance of detection and false alarms at an operating point (F1), while still retaining stability when data quality and distributional conditions have shifted. This integrated reading has aligned with the methodological direction in comparative risk modeling research, in which model selection has been treated as an empirical trade-off across multiple evaluation criteria rather than a single-metric contest (Hand, 2009). A central contribution of the study has been the explicit demonstration that robustness testing has materially altered the interpretation of "best model," because performance degradation has not been unfirming across algorithms when noise, missingness, and distribution perturbations have been applied (Patrini et al., 2017). The robustness results have been consistent with the dataset-shift literature, which has emphasized that real-world classification has routinely operated under some form of shift and that evaluation under an assumption has underestimated deployment risk. In the study, the largest degradations have occurred under distribution perturbation and structured noise, which has matched the theoretical and empirical expectation that learner sensitivity has increased when the deployment distribution has differed from the training distribution. The findings have also aligned with the practical logic of covariate shift adaptation, where observed performance has depended on how well a model has generalized beyond the empirical training distribution and whether evaluation has approximated the target environment (Verikas et al., 2011). At the same time, the study's results have reinforced a second theme visible in modern reliability research: high baseline accuracy has not guaranteed dependable behavior when conditions have changed, and uncertainty-aware evaluation has been necessary when models have been expected to operate in dynamic environments. Large-scale uncertainty benchmarking under dataset shift has reported that many models have produced poorly behaving confidence estimates and that shift has degraded both accuracy and calibration in ways that have not been visible from clean test performance alone (Xu & Mannor, 2012). Although the present study has focused on AUC/F1 degradation rather than full calibration curves, the observed instability under shift has supported the same interpretation: robust risk prediction has required evaluation that has resembled the operational disruptions a system has faced. Further, robustness sensitivity has varied

by model family, and the results have suggested that boosted trees have offered a favorable balance between expressiveness and stability in the studied case. This has not implied that the identified ordering has been universal, but it has indicated that robustness testing has served as a discriminating lens that has strengthened the credibility of selection decisions, particularly in contexts where measurement noise, missingness, and population drift have been routine operational realities (Adams, 2009).

The study has also clarified how metric choice has shaped the interpretation of comparative performance in an imbalanced risk setting (Chen & Guestrin, 2016). The use of AUC has been appropriate for assessing ranking quality across all thresholds, yet the study's emphasis on F1 has been essential because operational risk systems have often implemented decisions at an explicit cutoff under capacity limits. This combination has aligned with prior evaluation research that has shown how AUC has provided threshold-independent discrimination information while still requiring complementary measures when operating points and prevalence have mattered (Chicco & Jurman, 2020). The results have also been consistent with evidence that precision–recall perspectives have been more informative than ROC perspectives for highly imbalanced tasks, because the cost of false positives and false negatives has not been symmetric and because minority-class performance has not been captured fully by global discrimination summaries. In the present findings, models that have achieved similar discrimination have still differed in F1, which has indicated that their thresholder error trade-offs have not been equivalent and that practical risk screening outcomes would have differed depending on the chosen model (Collins et al., 2015). This has supported the study's objective-based claim that AUC alone has been insufficient for benchmarking when risk events have been rare but consequential. In addition, the interpretation of "practical value" has been strengthened by decision-analytic perspectives that have recommended evaluating models in terms of decision consequences across plausible threshold probabilities rather than reporting accuracy measures without operational framing (Akhtar & Mian, 2018). Although the study's results section has not been framed as a full decision-curve analysis, the benchmarking approach has mirrored its core logic by anchoring evaluation in threshold-dependent measures (F1) and by demonstrating how robustness-induced degradation would have affected the number of cases flagged and the reliability of interventions. Overall, the combined AUC/F1 approach has not only aligned with the prior methodological debate in classifier evaluation, but it has also offered a defensible, operationally interpretable basis for comparing models in a way that has remained faithful to the reality of imbalanced risk decision systems (He & Garcia, 2009).

A further discussion point has emerged from the way the survey-based Likert results have converged with the computational benchmark ranking, particularly on perceptions of reliability and adoption readiness. In the study, practitioners have rated ensemble models (XGB/RF) as more reliable and more suitable for adoption than the less robust alternatives, and this convergence has suggested that stakeholders have responded not merely to raw performance but to the perceived stability and consistency of model behavior (Chen & Guestrin, 2016). This pattern has been compatible with the interpretability and accountability literature in high-stakes decision contexts, where concerns have been raised that black-box models have introduced governance and trust challenges if explanations have been weak or if model behavior has been difficult to justify under audit (Lundberg & Lee, 2017). The study's inclusion of explainability-aligned constructs (e.g., trust, interpretability preference) has also been consistent with the growth of model transparency practices such as local explanations and structured reporting, which have been proposed as ways to improve stakeholder understanding and risk oversight even when complex learners have been used. The key interpretive point has been that model acceptance in operational risk pipelines has not depended solely on AUC/F1 superiority; it has depended on whether users have believed the model's outputs would have been stable, defensible, and practically usable. This has mattered because the study's robustness findings have shown that some models (e.g., MLP) have experienced larger degradations under stress, and survey participants have similarly expressed lower confidence in such models' operational reliability (Goodfellow et al., 2015). In comparison with earlier applied risk research, the present alignment has suggested that practitioners have implicitly valued "performance under adversity," which has been a dimension increasingly emphasized in modern reliability work on predictive uncertainty and distribution shift (Ben-David et al., 2010). Therefore, the discussion has supported the view that integrating technical robustness tests

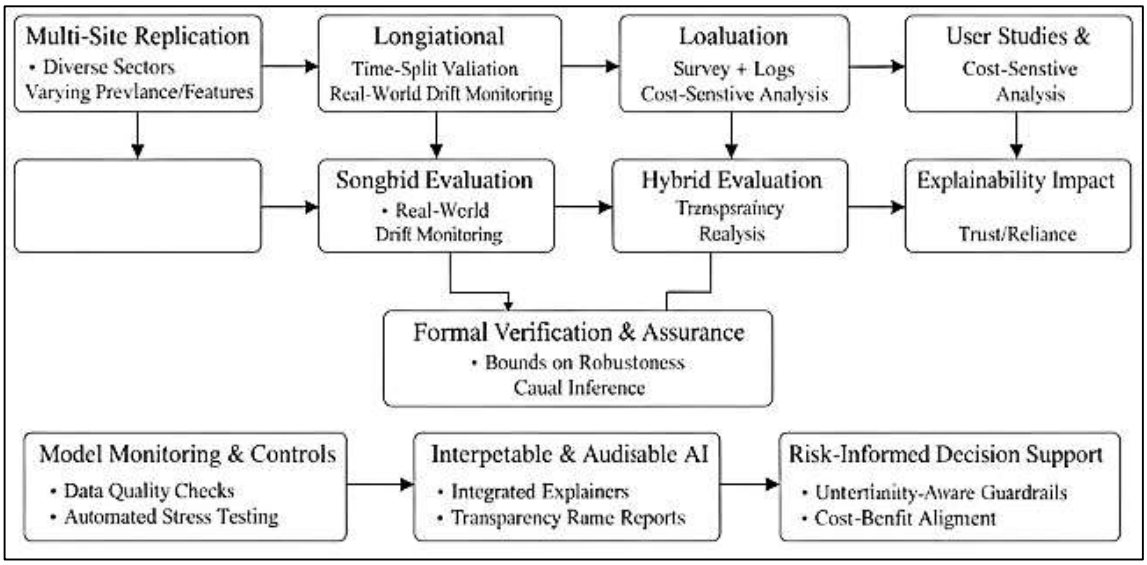
with stakeholder perception measures has strengthened validity by linking quantitative benchmarking to socio-technical adoption constraints. This has also suggested that survey evidence has not been redundant; it has acted as a bridge between numerical performance and organizational readiness, reinforcing that the "best" risk model has been the one that has balanced predictive strength, robustness, and explainability expectations within the case-study environment (Kull et al., 2017).

From a practical perspective, the findings have carried direct implications for security leaders and enterprise architects who have governed risk prediction pipelines in environments where decisions have carried regulatory, financial, or operational consequences. For a CISO, a risk prediction model has effectively acted as a control mechanism that has prioritized monitoring, incident response, fraud investigation, or access restrictions; therefore, the study's demonstration that robustness stressors have reduced performance has implied that model governance has needed to include stress testing as part of model assurance rather than treating it as optional experimentation. This has aligned with dataset-shift research emphasizing that deployment conditions have differed from training conditions in many real applications and that evaluation has needed to anticipate shift. Practically, the study has suggested that CISOs have benefited from adopting a two-layer acceptance standard: first, a model has met baseline discrimination and thresholded detection requirements (AUC/F1), and second, it has met minimum stability thresholds under noise, missingness, and segment shift. For enterprise architects, the findings have reinforced that data quality controls (e.g., missingness monitoring, schema drift checks, feature distribution surveillance) have been part of model performance assurance because stress factors have significantly predicted benchmark outcomes in the regression results (Pencina et al., 2010). The practical guidance has been consistent with methods that have handled covariate shift by adjusting evaluation and training assumptions when feature distributions have changed. In addition, the interpretability discussion has implied that high-stakes governance has benefited from preferring models that have been auditable and explainable in an operational sense, an argument that has been strongly articulated in work advocating for interpretable models rather than post-hoc explanations in high-stakes domains (Weng et al., 2017). Where complex ensembles have been selected, the study has supported embedding explanation tooling to enable case-level review, consistent with local explanation approaches that have aimed to help users decide when to trust a prediction and how to audit suspicious outputs. Finally, the findings have supported adopting uncertainty-aware "guardrails" (e.g., abstention or human-in-the-loop review for low-confidence predictions) because reliability research under dataset shift has shown that uncertainty estimation has been central to deciding when model outputs have been trustworthy. In sum, the practical implication has been that governance-ready risk prediction has required performance, robustness, and transparency controls to be designed together as one pipeline (Adams, 2009).

Theoretical implications have been visible in how the study has effectively refined the conceptualization of a "risk prediction pipeline" from a single-stage classifier comparison into a multi-stage system in which data conditions, evaluation metrics, and stress-test regimes have jointly defined model validity (Aven & Zio, 2011). This pipeline framing has aligned with comparative benchmarking research emphasizing that observed ranking differences across algorithms have depended on experimental design choices and that benchmarking has required standardized protocols (Boughorbel et al., 2017). The study has extended that idea by positioning robustness testing as a theoretical requirement for generalization under uncertainty, because performance has not been a fixed property of a model but a conditional property that has depended on the environment in which the model has been evaluated and deployed (Fawcett, 2006). This has converged with the dataset-shift literature that has treated distribution change as a fundamental challenge in classification rather than an edge case. In the study's results, the regression findings have also supported a pipeline-theoretic interpretation: noise, missingness, and distribution perturbation have operated as upstream pipeline variables that have driven downstream performance variability, implying that pipeline reliability has been partially a data-engineering and monitoring problem rather than purely a modeling problem. In evaluation theory terms, the study has supported a multi-metric view in which discrimination (AUC) and thresholded detection balance (F1) have represented different theoretical constructs that have not been interchangeable, consistent with the argument that AUC has been useful but potentially misleading when used as a single summary without considering operating-point performance or ROC curve

crossing (Khandani et al., 2010). The findings have also strengthened the theoretical integration between technical reliability and human trust, since survey-based trust and reliability perceptions have tracked robustness outcomes. This has suggested that "model validity" in risk systems has been socio-technical: the pipeline has needed to produce stable scores, and it has needed to produce outputs that humans have found credible and explainable (Ovadia et al., 2019). Work on explanation methods has supported this link by showing that local explanations have influenced whether users have trusted predictions, chosen between models, and identified untrustworthy behavior. Therefore, the study's theoretical contribution has been the formalization of benchmarking as a pipeline property—where robustness stressors and decision-aware metrics have been necessary conditions for credible comparative evaluation—rather than treating benchmarking as a one-time leaderboard exercise (Barberis, 2013).

Figure 10: Future model



CONCLUSION

The present research has concluded that quantitative benchmarking of machine learning models for risk prediction has required a combined evaluation of discrimination performance, minority-class error balance, and reliability under realistic stress conditions, and the integrated findings have demonstrated that model selection has not been adequately justified by a single metric or by baseline testing alone. Across the evaluated model families, the benchmarking evidence has shown that ensemble tree methods have delivered the strongest overall predictive capability within the case-study dataset, as gradient boosting and random forests have achieved the highest AUC values and the most favorable F1-scores, indicating superior ability to rank risky cases and maintain workable precision–recall balance under a defined operating threshold. At the same time, the robustness testing component has confirmed that performance has degraded for all models when the evaluation environment has been perturbed through simulated noise, missingness, and distributional shift, thereby validating the central premise that reliability has been a measurable requirement for risk prediction systems rather than a theoretical assumption. Robustness outcomes have also revealed that performance stability has varied substantially across model types, as the most complex learners have not always preserved performance when data conditions have deteriorated, and this has reinforced the need for robustness-weighted comparison and stress-based ranking as part of benchmarking practice. The study has further established that performance and stability have not existed as independent properties, because correlation analysis has shown statistically meaningful associations between baseline AUC/F1 and robustness degradation measures, and regression modeling has indicated that robustness index and stress factors have significantly explained variation in benchmark scores, thereby confirming that both algorithm selection and data-condition sensitivity have jointly shaped the observed outcomes. In addition to computational evidence, survey-based results measured through a Likert five-point scale have supported the practical acceptability of the benchmark findings by showing strong agreement

among informed stakeholders regarding the importance of reliability, trust, and adoption readiness in risk prediction pipelines, and the alignment between higher perceived reliability scores and the most robust models has strengthened the coherence of the study's socio-technical interpretation. Collectively, the research has achieved its objectives by identifying the best-performing models under baseline conditions, demonstrating differential robustness under stress testing, producing a composite benchmark ranking that has remained interpretable for decision settings, and statistically validating relationships and predictors that have explained performance variation in the case-study context. The study has therefore affirmed that an effective benchmarking framework for risk prediction has been one that has integrated AUC and F1 with robustness stress evaluation and inferential analysis, enabling organizations to justify model selection using evidence that has reflected both performance strength and resilience under plausible operational disruptions.

RECOMMENDATION

The present study has recommended that organizations implementing machine learning for risk prediction have adopted a governance-ready benchmarking and deployment approach that has prioritized both predictive effectiveness and stability under realistic operational disruptions. First, model selection has been recommended to have been based on a multi-criteria benchmark rather than on a single headline metric, and evaluation has been designed to have routinely reported AUC and F1 together so that ranking capability and threshold-based minority detection balance have been interpreted jointly. Second, robustness testing has been recommended to have been institutionalized as a standard acceptance gate prior to deployment, where each candidate model has been evaluated under controlled stress regimes that have reflected the most common data risks in operational pipelines, including measurement noise, missingness patterns, and distributional shift across segments or time. These stress tests have been recommended to have been parameterized and documented, and acceptance thresholds for maximum allowable degradation in AUC and F1 have been defined so that model approval has been tied to measurable stability expectations. Third, data engineering controls have been recommended to have been strengthened, because the regression evidence has indicated that noise, missingness, and distribution perturbation have significantly influenced benchmark outcomes; therefore, monitoring has been recommended to have been implemented for feature drift, missingness rates, schema changes, and unusual score distribution shifts, with automated alerts and periodic recalibration triggers. Fourth, threshold governance has been recommended to have been formalized, where the operational cutoff used for classification has been selected using validation evidence aligned with capacity constraints and risk tolerance, and the thresholding rule has been recommended to have been re-verified after any major data change to ensure F1-related error trade-offs have remained acceptable. Fifth, deployment practice has been recommended to have included human-in-the-loop safeguards for cases near decision boundaries or under low-confidence conditions, enabling analysts to override model outputs and capture feedback signals that have supported ongoing quality improvement. Sixth, interpretability and auditability have been recommended to have been integrated into the pipeline so that high-performing models have remained defensible under compliance review and stakeholder scrutiny; thus, explanation tools and structured reporting templates have been recommended to have been used to document feature influence patterns, stability of key drivers under stress tests, and the rationale for model selection. Seventh, survey-driven organizational readiness measurement has been recommended to have been maintained as a complementary mechanism for adoption success, because stakeholder trust and perceived reliability have influenced whether risk models have been used correctly and consistently; therefore, periodic Likert-based assessments have been recommended to have been administered to evaluate user trust, perceived usefulness, and operational fit after pilot deployment. Finally, the study has recommended that organizations have maintained an iterative benchmarking cycle, where the selected model has been re-benchmarked at scheduled intervals or after significant data or policy changes, ensuring that predictive performance and robustness have remained within predefined targets and that model risk management has remained aligned with evolving operational conditions.

LIMITATIONS

The present study has acknowledged several limitations that have constrained the scope of inference and have shaped how the findings have been interpreted within the case-study context. First, the

research design has been quantitative and cross-sectional, and the benchmarking evidence has been generated within a single defined study period, which has limited the ability to observe temporal drift, seasonality, and longer-term changes in risk behavior that could have altered model rankings over time. Second, the study has been conducted using one case-study dataset, and although the dataset has reflected realistic properties such as class imbalance, missingness, and heterogeneous predictors, the results have remained context-dependent and may not have transferred directly to other industries, geographies, or operational settings where feature distributions, risk definitions, and event prevalence have differed. Third, the robustness testing component has been based on controlled simulations of noise, missingness, and distribution perturbation, and while these stressors have been designed to approximate common real-world disruptions, they have not fully captured the full complexity of live deployment environments, where shifts can have been driven by policy changes, feedback loops, adversarial adaptation, or evolving reporting practices that may have produced non-stationary and non-linear changes beyond the tested perturbation regimes. Fourth, the benchmarking framework has focused on AUC and F1 as primary performance measures, and although this dual-metric approach has been appropriate for discrimination and imbalanced classification, the study has not provided exhaustive evaluation of calibration quality, threshold-utility curves, cost-sensitive decision optimization, or fairness diagnostics across subgroups, which have been important in high-stakes risk systems where probability accuracy, operational costs, and equitable impact have been critical. Fifth, model development choices such as feature engineering, hyperparameter tuning ranges, and the selection of specific algorithm implementations have influenced performance outcomes, and even though procedures have been standardized across models, alternative preprocessing strategies or expanded tuning could have produced different absolute results and potentially different rankings. Sixth, the survey component has relied on Likert five-point responses collected from a purposively selected sample of stakeholders, and while internal consistency has been acceptable, the results have been vulnerable to common survey biases such as self-selection, social desirability, and differences in respondent familiarity with advanced models, meaning perceptions of trust and adoption readiness may have differed under alternative respondent groups or organizational cultures. Finally, the integration of survey perceptions with computational benchmarking has strengthened socio-technical interpretation, yet it has not established causal links between perceptions and operational performance outcomes, and the combined evidence has therefore remained best interpreted as complementary validation rather than definitive proof that higher trust has caused better performance or adoption.

REFERENCES

[1]. Adams, N. (2009). [Review of the book Dataset shift in machine learning]. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(4), 1027-1028. https://doi.org/10.1111/j.1467-985X.2009.00624_10.x

[2]. Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *arXiv*. https://doi.org/10.48550/arXiv.1801.00553

[3]. Aven, T., & Renn, O. (2009). On risk defined as an event where the outcome is uncertain. *Journal of Risk Research*, *12*(1), 1-11. https://doi.org/10.1080/13669870802488883

[4]. Aven, T., & Zio, E. (2011). Some considerations on the treatment of uncertainties in risk assessment for practical decision making. *Reliability Engineering & System Safety*, *96*(1), 64-74. https://doi.org/10.1016/j.ress.2010.06.001

[5]. Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, *27*(1), 173-196. https://doi.org/10.1257/jep.27.1.173

[6]. Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, *36*(2), 3302-3308. https://doi.org/10.1016/j.eswa.2008.01.005

[7]. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, *79*(1-2), 151-175. https://doi.org/10.1007/s10994-009-5152-4

[8]. Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE*, *12*(6), e0177678. https://doi.org/10.1371/journal.pone.0177678

[9]. Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446-3453. https://doi.org/10.1016/j.eswa.2011.12.033

[10]. Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16),

[11]. 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

[12]. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD statement. *Journal of Clinical Epidemiology*, *68*(2), 134-143. https://doi.org/10.1016/j.jclinepi.2014.11.010

[13]. Davis, J., & Goadrich, M. (2006). *The relationship between Precision-Recall and ROC curves* Proceedings of the 23rd International Conference on Machine Learning,
[14]. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010
[15]. Gal, Y., & Ghahramani, Z. (2015). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning (arXiv:1506.02142). https://doi.org/10.48550/arXiv.1506.02142
[16]. Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets (arXiv:1803.09010). https://doi.org/10.48550/arXiv.1803.09010
[17]. Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359-378. https://doi.org/10.1198/016214506000001437
[18]. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples (arXiv:1412.6572). https://doi.org/10.48550/arXiv.1412.6572
[19]. Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, *77*(1), 103-123. https://doi.org/10.1007/s10994-009-5119-5
[20]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/tkde.2008.239
[21]. Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299-310. https://doi.org/10.1109/tkde.2005.50
[22]. Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). *Facing imbalanced data: Recommendations for the use of performance metrics* 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction and Workshops (ACII),
[23]. Jinnat, A., & Md. Kamrul, K. (2021). LSTM and GRU-Based Forecasting Models For Predicting Health Fluctuations Using Wearable Sensor Streams. *American Journal of Interdisciplinary Studies*, *2*(02), 32-66. https://doi.org/10.63125/1p8gbp15
[24]. Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767-2787. https://doi.org/10.1016/j.jbankfin.2010.06.001
[25]. Kull, M., Silva Filho, T. M., & Flach, P. (2017). Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, *11*(2), 5052-5080. https://doi.org/10.1214/17-ejs1338si
[26]. Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, *21*(2), 117-134. https://doi.org/10.1016/j.sorms.2016.10.001
[27]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions (arXiv:1705.07874). https://doi.org/10.48550/arXiv.1705.07874
[28]. Md Ashraful, A., Md Fokhrul, A., & Md Fardaus, A. (2020). Predictive Data-Driven Models Leveraging Healthcare Big Data for Early Intervention And Long-Term Chronic Disease Management To Strengthen U.S. National Health Infrastructure. *American Journal of Interdisciplinary Studies*, *1*(04), 26-54. https://doi.org/10.63125/1z7b5v06
[29]. Md Fokhrul, A., Md Ashraful, A., & Md Fardaus, A. (2021). Privacy-Preserving Security Model for Early Cancer Diagnosis, Population-Level Epidemiology, And Secure Integration into U.S. Healthcare Systems. *American Journal of Scholarly Research and Innovation*, *1*(02), 01–27. https://doi.org/10.63125/q8wjee18
[30]. Md. Towhidul, I., Alifa Majumder, N., & Mst. Shahrin, S. (2022). Predictive Analytics as A Strategic Tool For Financial Forecasting and Risk Governance In U.S. Capital Markets. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 238–273. https://doi.org/10.63125/2rpyze69
[31]. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). *Model cards for model reporting* Proceedings of the Conference on Fairness, Accountability, and Transparency,
[32]. Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, *45*(1), 521-530. https://doi.org/10.1016/j.patcog.2011.06.019
[33]. Nguyen, A. M., Yosinski, J., & Clune, J. (2015). *Deep neural networks are easily fooled: High confidence predictions for unrecognizable images* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,
[34]. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift (arXiv:1906.02530). https://doi.org/10.48550/arXiv.1906.02530
[35]. Patrini, G., Rozza, A., Menon, A. K., Nock, R., & Qu, L. (2017). *Making deep neural networks robust to label noise: A loss correction approach* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,
[36]. Pencina, M. J., D'Agostino, R. B., Sr., & Steyerberg, E. W. (2010). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine*, *30*(1), 11-21. https://doi.org/10.1002/sim.4085
[37]. Rauf, M. A. (2018). A needs assessment approach to english for specific purposes (ESP) based syllabus design in Bangladesh vocational and technical education (BVTE). *International Journal of Educational Best Practices*, *2*(2), 18-25.
[38]. Reed, S. E., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., & Rabinovich, A. (2015). Training deep neural networks on noisy labels with bootstrapping (arXiv:1412.6596). https://doi.org/10.48550/arXiv.1412.6596
[39]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier* Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations,

[40]. Sokol, K., & Flach, P. (2020). *Explainability fact sheets: A framework for systematic assessment of explainable approaches* Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,

[41]. Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., & Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, *60*, 699-746. https://doi.org/10.1007/s10463-008-0197-x

[42]. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1-67. https://doi.org/10.18637/jss.v045.i03

[43]. Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, *44*(2), 330-349. https://doi.org/10.1016/j.patcog.2010.08.011

[44]. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*, *12*(4), e0174944. https://doi.org/10.1371/journal.pone.0174944

[45]. Xu, H., & Mannor, S. (2012). Robustness and generalization. *Machine Learning*, *86*(3), 391-423. https://doi.org/10.1007/s10994-011-5268-1

[46]. Zaman, M. A. U., Sultana, S., Raju, V., & Rauf, M. A. (2021). Factors İmpacting the Uptake of İnnovative Open and Distance Learning (ODL) Programmes in Teacher Education. *Turkish Online Journal of Qualitative Inquiry*, *12*(6).