AI-Driven Investment Risk Assessment Using Robotic Process Automation and Multimodal Visual Data Analysis

Akshay Mukhopadhyay
School of Computing, Clemson University, Clemson, SC, USA.
mukhopadhyayakshay@clemson.edu

Abstract

Investment risk assessment increasingly depends on evidence that is heterogeneous, time-sensitive, and partly visual: filings, earnings presentations, charts, dashboard captures, news imagery, and operational documents accompany numerical market data. Conventional portfolio-risk models remain valuable but typically assume that inputs are already structured, synchronized, and complete, leaving a gap between evidence acquisition and risk inference. This paper develops a methodological framework for AI-driven investment risk assessment that combines robotic process automation with multimodal visual data analysis. The proposed architecture separates workflow orchestration from perception and risk modeling: software bots retrieve, validate, timestamp, and version evidence; modality-specific encoders transform numerical, textual, and visual inputs into aligned representations; attention-based fusion produces a compact risk state; and a calibration-aware decision layer translates predicted loss probabilities and uncertainty into escalation or review decisions. The central contribution is a provenance-first formulation in which every automated retrieval action is traceable to the risk estimate it supports, and in which visual evidence is treated as a first-class input rather than an illustration. The paper specifies risk measures, workflow policies, multimodal fusion, temporal aggregation, and uncertainty-sensitive decision rules, then assesses trade-offs involving automation scope, extraction error, distribution shift, visual ambiguity, and human oversight. The framework provides a technically defensible basis for auditable, multimodal investment-risk assessment without relying on unstated empirical assumptions.

Keywords

multimodal risk assessment; robotic process automation; visual document analysis; uncertainty calibration; investment risk; human-in-the-loop decision support

1. Introduction

Investment risk assessment begins with the distribution of portfolio losses and the dependence structure among assets. Mean-variance portfolio theory formalizes risk as variance and establishes diversification as a response to covariance [1], while equilibrium asset-pricing models relate expected returns to systematic risk [2]. These foundations remain essential, but they are not sufficient for settings in which risk evidence arrives as unstructured documents, chart images, news photographs, presentation slides, and portal screenshots. Risk assessment also has an institutional dimension: it must identify hazards, characterize uncertainty, and support decisions under incomplete knowledge [3]. The analytical problem is therefore broader than estimating a covariance matrix. It is the problem of acquiring, aligning, interpreting, and justifying heterogeneous evidence before a risk score is produced.

Machine learning has expanded the representational capacity available for this task. Deep architectures can learn hierarchical features from raw or weakly structured inputs [4], convolutional networks can extract spatial patterns from images [5], residual connections support deeper visual models [6], and attention mechanisms allow models to weight relevant components of an input sequence [7]. Multimodal learning provides a systematic vocabulary for representation, alignment, fusion, and translation across modalities [8]. In parallel, robotic process automation offers a mechanism for executing repetitive, rule-governed digital tasks across interfaces and applications [9]. Automation designs in adjacent domains show that when digital instructions are coupled with sensed outcomes, the system must preserve traceability between intended action and observed result; robotic fabrication workflows illustrate this coupling between computational plans and material execution [10]. The lesson for investment risk is not to imitate physical fabrication, but to make the automated acquisition path explicit and auditable.

RPA research has moved from simple task automation toward programs with governance, exception handling, and integration challenges [11]. Accounting and finance applications demonstrate that digital labor can perform structured, high-volume work while raising questions about controls, evidence, and process redesign [12], and early case evidence shows that automation success depends on process selection and organizational fit [13]. These findings matter for risk assessment because a bot that retrieves the wrong filing version, misreads a chart axis, or omits a timestamp can contaminate the inference chain. The gap addressed here is the absence of a unified formulation in which RPA actions, multimodal visual evidence, risk inference, and decision escalation are designed as one accountable system.

The paper makes one coherent contribution: it specifies a provenance-first architecture for AI-driven investment risk assessment in which RPA orchestrates evidence acquisition, multimodal encoders convert visual, textual, and numerical inputs into aligned risk representations, and a calibrated decision layer determines whether a risk estimate is actionable or requires human review. The framework is methodological and analytical. It defines risk targets, automation policies, fusion operations, temporal aggregation, and uncertainty-sensitive decision rules, and it evaluates design trade-offs rather than reporting experimental outcomes. Section 2 formulates the risk objects and evidence types. Section 3 develops the architecture and its equations. Section 4 assesses technical trade-offs. Section 5 discusses failure modes and sensitivity. Section 6 concludes.

2. Risk Objects, Evidence Types, and Target Measures

An investment risk object is defined over a horizon H and a reference probability space. Let L denote the loss associated with a position, portfolio, or monitored entity over H. For a confidence level α in (0,1), value-at-risk is the smallest loss threshold whose exceedance probability is no greater than 1 minus α. This definition is useful because it yields a single threshold for capital or monitoring purposes, but it does not describe the magnitude of losses beyond that threshold.

$$VaR_\alpha(L) = inf\{\ell \in \mathbb{R} : P(L \leq \ell) \geq \alpha\}$$

Here L is the loss random variable, P is the reference probability measure, α is the confidence level, ℓ is a candidate loss threshold, and inf selects the smallest threshold satisfying the probability condition. The expression clarifies that VaR is a quantile of the loss

distribution. Its limitation is equally important: two portfolios can share the same VaR while having very different tail exposures.

Expected shortfall addresses this limitation by averaging losses beyond the VaR threshold. It is a coherent risk measure under suitable integrability conditions and provides a more informative target for tail-sensitive investment risk assessment.

$$ES_\alpha(L) \;=\; (1/(1-\alpha)) \int_\alpha\text{^}1\, VaR_u(L)\, du$$

In this expression u is a quantile index integrated from α to 1, and the prefactor 1/(1−α) converts the integrated tail into an average. The formulation motivates a model that estimates a conditional loss distribution rather than only a point forecast. In the proposed architecture, the risk model outputs a distribution over L, from which VaR, ES, and threshold probabilities can be derived.

The evidence used to condition that distribution is multimodal. Numerical inputs include returns, volumes, spreads, factor exposures, and accounting variables. Textual inputs include filings, analyst reports, earnings-call transcripts, and news. Visual inputs include chart images, scanned or rendered documents, presentation layouts, dashboard captures, and other pixel-level evidence. Textual sentiment and media content have long been studied as market-relevant signals [14], and dictionary-based textual analysis has shown that domain-specific language matters when interpreting financial disclosures [15]. Risk disclosures contain structured and unstructured information that can be used to identify risk types [16]. Social media provides additional high-frequency signals, including mood indicators [17] and stock opinions transmitted through online networks [18]. These sources are not interchangeable: they differ in latency, reliability, and failure modes. A visual chart may reveal a scale distortion or annotation that text does not mention; a filing may contain a risk-factor revision that no numerical series captures.

Dependence also changes under stress. Equity correlations can be asymmetric, rising in down markets [19], and extreme correlations across international markets can differ from correlations in normal periods [20]. This has a direct implication for multimodal risk assessment: the model must not treat evidence as stationary by default. Visual evidence can signal a regime change before numerical series fully reflect it, but it can also introduce spurious pattern recognition. The risk state must therefore include uncertainty and provenance, not only a point estimate.

3. Methodological Framework

3.1 RPA Orchestration and Provenance

The RPA layer is modeled as a policy over workflow states. A state s_t records the task context, available evidence, prior retrieval outcomes, and pending validation checks. An action a may open a source, authenticate, download a document, capture a dashboard region, run an OCR check, request a human exception, or mark a source as unavailable. The policy selects actions to maximize expected retrieval value net of execution cost, discounted over the remaining workflow horizon.

$$\pi * (s_t) \;=\; argmax_a\, E[\sum_\{k=0\}\text{^}\{T-t\}\, \gamma\text{^}k\, (\kappa\, r_\{t+k\} \;-\; c(a)) \mid s_t]$$

Here π is the chosen policy, s_t is the workflow state at time t, a is an admissible action, T is the terminal step, γ is a discount factor in [0,1], κ converts retrieval value into utility units, r_{t+k} is the relevance or completeness value obtained at step t+k, c(a) is the action cost, and E denotes expectation over workflow outcomes. The expression is a design objective, not

an estimated performance claim. It makes explicit that automation should not maximize document count; it should maximize relevant, validated evidence per unit of execution cost.

Provenance is a constraint on admissible actions. Each retrieved item receives a source identifier, retrieval timestamp, interface state, content hash, extraction method, and confidence score. A visual crop must retain its original page or screen coordinates. A parsed table must retain the bounding region from which cells were extracted. If the bot encounters a changed layout, an expired session, or an ambiguous chart, the policy should route the item to exception handling rather than silently imputing missing evidence. This separates automation reliability from model reliability: a risk model cannot be more trustworthy than the evidence pipeline that feeds it.

3.2 Multimodal Visual and Textual Encoding

Each evidence item is mapped to a modality-specific representation. Let x_{i,m} denote the input of modality m for risk object i, where m ranges over numerical series, text, and visual evidence. A numerical encoder may use temporal convolutions or recurrent summaries; a text encoder may use token embeddings and contextual attention; a visual encoder may use a convolutional backbone for chart or document images and a layout-aware transformer for rendered pages. The fusion layer combines these representations through attention weights that depend on the evidence content.

$$z_i = \sigma\left(\sum_m \beta_m W_m \varphi_m(x_{i,m}) + b\right)$$

Here z_i is the fused risk representation for object i, φ_m is the encoder for modality m, W_m is a modality-specific projection matrix, β_m is a learned attention weight, b is a bias vector, and σ is a nonlinear activation. The term φ_m(x_{i,m}) may be a vector or a sequence of vectors; when it is a sequence, pooling or attention over positions is applied before fusion. The equation allows the model to down-weight a modality that is missing, corrupted, or irrelevant for a particular risk object.

The attention weights are computed from the same modality representations, so the model can adapt fusion to the evidence rather than fixing a single weighting scheme. For modality m, with query vector q, transformation V_m, and bias c_m, the normalized weight is:

$$\beta_m = exp(q \cdot tanh(V_m\, \varphi_m(x_\{i,m\}) + c_m)) \,/\, \sum_k exp(q \cdot tanh(V_k\, \varphi_k(x_\{i,k\}) + c_k))$$

The index k runs over modalities, and the denominator normalizes the weights so that they sum to one. The dot product q·tanh(...) scores how relevant modality m is to the current risk object. This formulation is useful for visual evidence because it can learn that a chart image is highly informative when numerical volatility is unstable but less informative when the chart is a static duplicate of an already parsed table.

3.3 Temporal Aggregation and Risk-State Construction

Investment risk is not a single-document classification problem. Evidence arrives in sequence, and the meaning of a visual disclosure can depend on what was known earlier. A temporal attention layer aggregates representations across a window of prior observations.

$$h_t = \sum_\{k=0\}^\{K\} a_\{t,k\}\, z_\{t-k\}, \quad a_\{t,k\} = softmax(q_t \cdot z_\{t-k\})$$

Here h_t is the aggregated risk state at time t, K is the lookback window, z_{t−k} is the fused representation at lag k, a_{t,k} is the attention weight assigned to that lag, q_t is a time-

specific query vector, and softmax normalizes the weights across k. The equation permits the model to emphasize a recent visual anomaly or a past disclosure that remains relevant. It also makes the temporal dependence explicit: a risk state is a weighted combination of evidence, not an average of document scores.

The resulting risk state feeds a distributional head that estimates parameters of the conditional loss distribution. For example, a Gaussian or Student-t head could produce location and scale parameters, while a quantile head could directly estimate a set of quantiles. The choice should reflect the tail behavior implied by the investment domain. If extreme co-movement is plausible, a heavy-tailed specification or an extreme-value component is more defensible than a Gaussian assumption. The framework does not prescribe one distribution; it requires that the distributional assumption be stated, tested where data permit, and linked to the VaR and ES definitions in Section 2.

3.4 Calibration and Decision Logic

A risk score is only useful if its probability statements are calibrated. A model that outputs a high probability of loss exceedance but is systematically overconfident can trigger unnecessary reviews; an underconfident model can suppress legitimate escalation. Calibration can be assessed with proper scoring rules such as the Brier score or with reliability diagrams, and it can be improved through post-hoc calibration on held-out data when such data are available. The decision layer should use both the calibrated probability and an uncertainty estimate.

Let p_i be the calibrated probability that loss exceeds a threshold for risk object i, and let u_i be an uncertainty measure, such as the dispersion of predictions across an ensemble or the width of a posterior interval. A review rule can be written as:

$$d_i = 1\{p_i \geq \tau + \lambda\, u_i\}$$

Here d_i equals one when the risk object is escalated for human review and zero otherwise, τ is a base escalation threshold, λ is a nonnegative uncertainty penalty, and $1\{\cdot\}$ is the indicator function. The rule states that escalation should require stronger evidence when uncertainty is high. This is preferable to a single threshold because it distinguishes a confident high-risk signal from a noisy one. The parameters τ and λ are policy choices that should be set according to the cost of false escalation, the cost of missed risk, and the capacity of the review function.

4. Analytical Assessment and Design Trade-offs

The first trade-off concerns automation scope. A bot can be designed to retrieve documents, parse tables, capture screenshots, and validate file integrity, or it can be restricted to low-risk retrieval with human confirmation at each step. Broader automation increases throughput and consistency but raises the cost of silent failure. The provenance constraint favors narrow, well-instrumented actions over broad, opaque workflows. For example, a bot that captures a dashboard region should record the dashboard filters, time zone, and screen resolution. Without those fields, a later visual model cannot distinguish a genuine change in risk from a changed display setting.

The second trade-off concerns visual representation. A chart image contains geometry, labels, color, and annotation. A convolutional encoder can learn visual patterns [5], and residual connections can improve optimization in deep visual models [6], but these capacities do not guarantee semantic understanding of axes, units, or logarithmic scales. A layout-aware

model can use text and position together, while a pure image model may overfit to style. The fusion equation mitigates this risk by allowing the model to combine visual and textual evidence, but it does not eliminate the need for extraction checks. A defensible design uses visual encoders for chart-shape and document-layout signals, and uses OCR plus rule-based validation for axis labels, units, and table structures. The visual and textual paths should be complementary, not redundant.

The third trade-off concerns attention fusion. Attention can improve interpretability by exposing modality weights [7], but attention weights are not causal explanations. A high weight on a visual modality indicates that the fusion layer used that representation strongly; it does not prove that the visual evidence caused the risk estimate. For auditability, attention weights should be paired with provenance records and with sensitivity tests that remove or perturb a modality. If removing a document image changes the risk state substantially, the system should record that dependence rather than presenting a single attention score as an explanation.

The fourth trade-off concerns temporal aggregation. A short window is responsive to new evidence but vulnerable to noise; a long window stabilizes estimates but may dilute a recent regime change. The attention formulation in Section 3.3 allows the model to learn lag weights, but the lookback window remains a design choice. In investment risk, the window should be aligned with the horizon H and with the reporting frequency of the evidence. A daily market-risk model may need a short window for numerical returns and a longer window for disclosures. A quarterly credit-risk review may reverse that emphasis. The risk state should therefore be horizon-specific rather than a single universal embedding.

The fifth trade-off concerns calibration and decision thresholds. A highly sensitive escalation rule may catch more risk events but burden reviewers with false positives. A conservative rule may reduce workload but allow ambiguous evidence to pass without scrutiny. The uncertainty-sensitive rule in Section 3.4 makes this trade-off explicit. It also supports tiered review: low-uncertainty, high-probability cases can be routed to standard review; high-uncertainty cases can be routed to specialist review; and cases with missing provenance can be blocked from automated decision-making altogether.

5. Technical Discussion: Failure Modes and Sensitivity

The framework's main failure modes are acquisition error, representation error, fusion error, calibration drift, and decision-policy mismatch. Acquisition error occurs when the RPA layer retrieves an outdated filing, captures a dashboard with the wrong filter, or misses a page because a layout changed. Provenance fields reduce the blast radius of such errors because they allow the risk estimate to be traced to a specific artifact. They do not prevent the error, but they make it detectable and correctable.

Representation error occurs when an encoder misreads a chart, a table, or a document layout. Visual encoders can be sensitive to resolution, compression, color, font, and aspect ratio. A chart image that has been resized may distort line slopes; a scanned document may contain OCR artifacts that change the meaning of a risk factor. Robustness requires preprocessing checks: aspect-ratio validation, OCR confidence thresholds, table-structure validation, and cross-modal consistency tests. If the visual path extracts a revenue trend that contradicts the parsed numerical table, the system should flag the conflict rather than let one modality dominate silently.

Fusion error occurs when attention weights concentrate on a modality that is present but uninformative. For instance, a company logo or a decorative image may dominate a visual representation if the encoder is not constrained to task-relevant regions. Layout-aware cropping, region-level pooling, and modality dropout during training can reduce this risk. Fusion error also arises when modalities are correlated but not redundant; a text encoder and a visual encoder may both capture the same disclosure, causing double counting. The attention layer should be regularized or evaluated for modality redundancy, and the risk model should include a missing-modality path so that a failure in one encoder does not produce an arbitrary score.

Calibration drift occurs when the relationship between evidence and loss outcomes changes. Regulatory changes, market microstructure shifts, and reporting-standard updates can alter the meaning of visual and textual evidence. Because the framework is distributional, drift can be monitored through changes in predicted quantiles, calibration curves, and uncertainty distributions. A rise in ensemble disagreement or posterior width is a useful early warning signal, but it is not a substitute for outcome monitoring. When labels are delayed, as they often are in credit and private-market risk, the system should rely more heavily on provenance completeness and human review until calibration can be re-established.

Decision-policy mismatch occurs when the escalation rule is optimized for a different objective than the one the institution actually cares about. A rule that minimizes review volume may underweight missed risk; a rule that maximizes detection may overwhelm reviewers. The uncertainty penalty in Section 3.4 is a policy parameter, not a model parameter, and it should be set through an explicit cost or constraint analysis. The decision layer should also support abstention: if provenance is incomplete or modality conflict is unresolved, the system should return “review required” rather than a numerical risk score that implies false precision.

Sensitivity analysis should therefore be multimodal and workflow-aware. It should perturb visual regions, text spans, numerical inputs, and RPA retrieval paths. It should ask whether the risk estimate changes when a chart is replaced by its underlying table, when a filing version is rolled back, or when a modality is removed. These tests do not replace empirical validation, but they provide a principled way to identify brittle dependencies before a model is placed in a risk-sensitive workflow.

6. Conclusion

This paper has developed a methodological framework for AI-driven investment risk assessment that integrates robotic process automation with multimodal visual data analysis. The framework treats risk assessment as a provenance-constrained inference problem: RPA orchestrates the acquisition of numerical, textual, and visual evidence; modality-specific encoders transform that evidence into aligned representations; attention-based fusion and temporal aggregation construct a risk state; and a calibration-aware decision layer converts probabilities and uncertainty into review or escalation decisions. The formulation connects established risk measures with modern representation learning while keeping the evidence path auditable. Its central implication is that automation and visual analysis should not be appended to a risk model as separate modules. They should be designed together so that every retrieved artifact, every extracted region, and every fused representation can be traced to the risk estimate it supports. The resulting architecture offers a technically defensible basis for multimodal investment-risk assessment, with explicit trade-offs in automation scope, visual ambiguity, calibration, and human oversight.

References

1. Markowitz, H. (1952). Portfolio selection. The Journal of Finance, 7(1), 77–91.

2. Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. The Journal of Finance, 19(3), 425–442.

3. Aven, T. (2016). Risk assessment and risk management: Review of recent advances on their foundation. European Journal of Operational Research, 253(1), 1–13.

4. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

5. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097–1105.

6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).

7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

8. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

9. van der Aalst, W. M. P., Bichler, M., & Heinzl, A. (2018). Robotic process automation. Business & Information Systems Engineering, 60(4), 269–272.

10. Luo, D., Chen, D., Wang, J., Zhu, G., & Xu, W. (2020). The smart robot crafting approach to computing materials. Construction Robotics, 4(3), 239-249.

11. Syed, R., Suriadi, S., Adams, M., Bandara, W., Leemans, S. J. J., Ouyang, C., ter Hofstede, A. H. M., van de Weerd, I., Wynn, M. T., & Reijers, H. A. (2020). Robotic process automation: Contemporary themes and challenges. Computers in Industry, 115, 103162.

12. Kokina, J., & Blanchette, S. (2019). Early evidence of digital labor in accounting: Innovation with robotic process automation. International Journal of Accounting Information Systems, 35, 100431.

13. Lacity, M. C., & Willcocks, L. P. (2016). Robotic process automation at Telefónica O2. MIS Quarterly Executive, 15(1), 21–35.

14. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139–1168.

15. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65.

16. Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. Management Science, 60(6), 1371–1391.

17. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8.

18. Chen, H., De, P., Hu, Y., & Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. The Review of Financial Studies, 27(5), 1367–1403.

19. Ang, A., & Chen, J. (2002). Asymmetric correlations of equity portfolios. Journal of Financial Economics, 63(3), 443–494.

20. Longin, F., & Solnik, B. (2001). Extreme correlation of international equity markets. The Journal of Finance, 56(2), 649–676.