Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 13, No. 3, 2025

Predicting Internal Control Deficiencies with Deep Learning: Evidence from IPO Review Records

Dongjie Lin *

School of Public Finance and Taxation, Central University of Finance and Economics, Beijing 102206, China
*** Corresponding author Email:** rayldj@foxmail.com

Abstract: Internal control is a core governance mechanism for improving information quality and reducing information risk. Identification of internal control deficiencies (ICDs) based on annual report disclosures or audit opinions suffers from reporting lags and sparse positives. Leveraging China's issuance regime, this study uses IPO rejection outcomes as proxy labels and builds a financial-ratio–based prediction model. Incorporating proxy information lifts PR-AUC and ROC-AUC over the baseline by 22.78% and 4.52%; a full-parameter search yields gains of 24.05% and 5.32%. Under fixed thresholds, F1, precision, and recall also rise. Interpretability analyses show that ROA, OPM, and PPE retain persistent importance in identifying high-risk firms. The approach advances label construction and methodology, and offers an operational tool to target regulatory inquiries, optimize audit procedures, and support firms' internal control self-assessments, with clear policy and theoretical value.

Keywords: Internal Control Deficiency Prediction; IPO Rejection; Deep Learning.

1. Introduction

As a contractual arrangement to mitigate agency conflicts, internal control aims to enhance information quality and prevent material risks, playing a foundational role in standardizing operations, ensuring the reliability of financial reporting, and sustaining capital market trust. By standardizing processes and allocating responsibilities, internal control improves the production and transmission of information and, in turn, affects resource allocation efficiency and risk pricing. When it fails, firms become more prone to earnings management, heightened information risk, fraud, and stock price crashes. Internal control deficiencies (ICDs) are associated with lower accounting information quality, greater information asymmetry, and higher financing costs and crash risk [1-2]. However, identifying ICDs from post-listing disclosures or external audit opinions entails latency, sparsity, and selective disclosure constraints, limiting forward-looking prediction useful for regulators, auditors, and investors.

To overcome the twin bottlenecks of sparse ICD positives and disclosure lags, this study uses IPO rejection outcomes as proxy labels. Under the registration-based regime and routine regulatory inquiries, IPO examination not only emphasizes the conformity of financial statements and the completeness of information disclosure but also conducts multi-round inquiries and onsite inspections that penetrate the design and execution of internal control. IPO rejection therefore often maps to substantive deficiencies in accounting fundamentals, governance arrangements, and risk control, and can be viewed as a forward-looking regulatory signal that characterizes internal-control fragility. Compared with ICD reports disclosed only after listing, this proxy label offers greater observability, earlier timing, and broader coverage. It substantially expands the number of positive-class observations and mitigates undertraining induced by the scarcity of severe ICD cases.

This study introduces IPO-rejected firms as proxy labels to alleviate positive-class scarcity and builds an ICD prediction model. Out-of-sample evaluation on a common test set shows that incorporating IPO-rejected cases improves performance relative to the baseline: PR-AUC rises by 22.78% and ROC-AUC by 4.52%. Building on this, the optimal model from a full-parameter search further improves PR-AUC by 24.05% and ROC-AUC by 5.32% over the baseline. At fixed thresholds, F1, precision, and recall increase concurrently, implying higher coverage while maintaining precision. Overall, both the IPO-tuned and full-parameter models outperform the baseline in ranking quality and threshold performance, with the full-parameter model exhibiting the best robustness and balance.

The main contribution is a forward-looking ICD prediction framework that marries institutional context with methodological innovation. Relative to prior work—for example, Liu and Lin (2024) [3], which highlights the advantages of deep learning over logit by emphasizing structural specification and predictive accuracy—this study takes an institutional perspective by converting IPO rejection outcomes into proxy labels for ICDs and implementing a "proxy in training, truth in evaluation" design. Through weak labeling and robust learning, this design alleviates positive-class scarcity while controlling potential noise. The study provides an operational quantitative tool to better target regulatory inquiries, inform audit-procedure design, and support firms' internal control self-assessments.

The remainder of the paper is organized as follows. Section 2 reviews the literature. Section 3 presents the institutional background. Section 4 develops the ICD prediction model. Section 5 describes the sample and reports descriptive statistics. Section 6 evaluates model performance. Section 7 concludes.

2. Literature Review

As a core governance mechanism, internal control reshapes the production, transmission, and use of corporate information through process standardization and the allocation of responsibilities. Discussion of its economic consequences hinges on establishing a measure of internal control effectiveness that is operable, comparable, and

verifiable. Against this backdrop, this study first reviews the mechanisms through which internal control operates and then assesses the existing research on effectiveness measurement.

By shaping firms' information production and dissemination, internal control continually improves the information environment. Firms with internal control deficiencies (ICDs) exhibit poorer accrual quality and a stronger propensity for earnings management[4], and they face elevated risks of financial fraud [1]. Changes in the information environment affect analyst coverage and forecast quality and are subsequently reflected in pricing in equity and debt markets. In equity markets, ICDs are positively associated with the cost of equity, indicating that investors demand higher risk premia [5], and they are an important driver of stock price crash risk [6]. In debt markets, Costello and Wittenberg-Moerman (2011) find that when firms report material internal control weaknesses, lenders reduce reliance on financial covenants and performance pricing and place greater weight on collateral and rating terms, revealing a contractual substitution in response to information risk [7]. External governance and market conditions can moderate these effects: institutional investor monitoring, industry competition, and high-quality auditing mitigate the negative impact of ICDs through shareholder action, market discipline, and professional oversight, respectively. In short, internal control quality affects capital and debt markets through the mechanism of reducing information noise, reshaping external oversight, and altering pricing and contracting structures, with pronounced heterogeneity across settings such as the strength of external monitoring, media and analyst attention, and underwriter reputation.

As an embedded organizational system, internal control influences "real-side" decisions in operations, investment, and cost management via process standardization and authority allocation. At the operational level, ICDs heighten internal information noise and agency frictions, undermining process and cash-conversion efficiency. Relative to financial-reporting internal control alone, operations-level internal control has a larger impact on firm performance [8]. Following remediation of material weaknesses, operating efficiency improves significantly, especially in firms with higher information demand, more severe weaknesses, and smaller size [9]. Regarding investment efficiency, weak internal control exacerbates information asymmetry and governance failures, inducing both over- and under-investment [10]. When ICDs relate to capital expenditures, the association between future cash flows and current investment is systematically weakened, indicating control failures over real investment processes.

In evaluating internal control effectiveness, U.S. research commonly relies on internal control over financial reporting assessments and deficiency disclosures under Section 404 of the Sarbanes–Oxley Act (SOX) as an observable measure. Based on this, Doyle et al. (2007) match firms disclosing material weaknesses to peer characteristics and show that such firms tend to be smaller, younger, financially weaker, more complex, or undergoing restructuring—factors that correlate with control weakness [11]. Before internal control audits became mandatory, Ashbaugh-Skaife et al. (2007) examine the discovery and disclosure of weaknesses in the pre-mandate period and confirm that business complexity, organizational change, and accounting risk increase the likelihood of ICDs; disclosure reflects not only actual control conditions but also managerial incentives and audit-supply constraints [5].

Compared with the SOX 404–centered U.S. approach, evaluation of internal control effectiveness in China has evolved from early "defect disclosure" toward index-based composite measures. Lin et al. (2014) construct an outcome-oriented internal control index that refines the five objectives of the Basic Internal Control Norms into basic, operating, and strategic layers and measures effectiveness by goal attainment [12]. Lin et al. (2016) develop a disclosure-based index by mapping disclosure items to the COSO (2013) and COSO-ERM (2016) frameworks, yielding five first-level, thirty-one second-level, and eighty-seven total indicators [13].

While disclosure-based and index-based measures offer an operational foundation for studying the economic consequences of internal control, there remains scope to better capture heterogeneity, nonlinear relations, and the integration of multi-source data. It is therefore valuable to advance machine-learning–based ICD prediction models that jointly emphasize out-of-sample predictability and causal identification, thereby verifying and extending both the measurement of internal control effectiveness and evidence on its underlying mechanisms.

3. Institutional Background

China's capital market has undergone gradual reform from a quota system to an approval-based IPO regime and then to a registration-based regime. The regulatory logic has shifted from administrative gatekeeping at the point of entry to market-based discipline grounded in sufficient, truthful, and understandable information disclosure. The earlier phase emphasized qualification screening and quota control—filtering issuers by entry criteria and quality thresholds—whereas the later phase emphasizes disclosure and inquiry as process constraints, with ongoing oversight exercised by investors and market mechanisms. For the sample period in this study (2010–2022), the institutional setting spans the approval-based regime characterized by gatekeeping by the CSRC Issuance Examination Committee (IEC) and, from 2019 onward, the registration-based pilot launched on selected boards. The comprehensive registration-based reform in 2023 marks a watershed but falls outside our sample window. These institutional transitions provide an operational basis for identifying internal control effectiveness using examination and disclosure information.

Alongside the evolution of the issuance regime, the organization and procedures of examination were restructured. Under the approval-based regime, regulators conducted substantive reviews of compliance, profitability, going-concern status, and the quality of accounting information. Starting from the completeness and conformity of application materials, the IEC exercised comprehensive gatekeeping over financial metrics, governance structures, and the adequacy of disclosure, reflecting strict entry control. Rejection decisions in this phase typically pointed to externally verifiable deficiencies, such as material compliance gaps, insufficient disclosure, inappropriate accounting policy choices, or weak internal oversight. Under the registration-based pilot, the guiding philosophy shifted from administrative admission to disclosure-based review, yet the constraint on key risks did not loosen. Multi-round regulatory inquiries brought forward questions concerning accounting treatments, major transactions, going-concern uncertainties, and governance arrangements, forming a traceable evidence trail through inquiries and responses. Stock exchanges and the CSRC

could still issue decisions of non-registration or non-approval to list where disclosure was untruthful, inadequate, or noncompliant. Formally, the two stages emphasize "substantive review of quality" versus "adequacy review of disclosure." Substantively, the risk content of rejection is highly similar across stages, concentrating on compliance, disclosure quality, financial-reporting discipline, and governance structure.

The grounds for IPO rejection exhibit stable, structured patterns. In the compliance dimension, gaps often involve hard constraints such as taxation, environmental protection, and industry admission. In the disclosure dimension, issues commonly include insufficient or unpersuasive disclosure of key transactions, related-party relationships, and customer or supplier concentration. In the financial-reporting dimension, problems include improper revenue recognition, ambiguous classification of non-recurring items, and repeated audit procedures. In the governance dimension, weaknesses include unclear boundaries between control rights and operating authority, breakdowns in internal oversight, or weak decision procedures for major related-party transactions. These reasons are systematically tracked and evidenced in the regulatory process and, in substance, map to weaknesses in the design and execution of internal control. Even within the registration-based pilot, a rejection constitutes a negative conclusion on the truthfulness, adequacy, and compliance of disclosure, and its risk content is, to a considerable extent, the same as under the approval-based regime. Accordingly, in the coexistence of approval-based and registration-based pilot regimes, IPO-rejected firms, in most cases, reflect insufficient internal control effectiveness and can serve as proxy labels for internal control deficiencies (ICDs).

4. Constructing the Internal Control Deficiency Prediction Model

4.1. Modeling Framework and Learning Objective

Financial fraud prediction and internal control deficiency (ICD) prediction are structurally analogous. Both are classification tasks with low base rates (few positives), high-dimensional and heterogeneous features, and tight regulatory constraints, and they share a common toolbox of predictive methods. The fraud-prediction literature has developed along three main routes. First, interpretable linear models such as logit provide a disciplined baseline and stable ranking [14]. Second, nonlinear classifiers led by kernel methods and ensemble learning—e.g., support vector machines [15-16] and bagging/boosting [17]. Third, deep learning, which leverages deep networks and attention mechanisms to enhance representation and discrimination [18].

In ICD prediction, early studies largely relied on logit models. The linear specification limits the ability to capture complex nonlinear relations and intertemporal dynamics. Bringing deep learning into this setting, Liu and Lin (2024) employ a recurrent neural network (RNN) and organize predictors under the POP (pressure–opportunity–propensity) framework, delivering sizable performance gains. Building on that line of work [3], this study also focuses on deep learning but extends the research design: during training, this study introduces IPO rejection as proxy positives to alleviate rarity and severe class imbalance, thereby enriching the set of learnable risk signals.

The core classifier is an RNN. The inputs are firm-level financial ratios organized in a rolling two-period window. Through hidden units and nonlinear activations, the network performs representation learning that flexibly captures nonlinearities and potential higher-order interactions among indicators. In the statistical–econometric tradition, linear or semiparametric models depend on prespecified functional forms and hand-crafted interaction terms. By contrast, the RNN learns end-to-end which feature combinations, under which conditions, jointly indicate elevated ICD risk. This yields stronger separability and better extrapolation when indicators are correlated and the structure is complex.

Conceptually, the RNN acts as a representation learner that compresses high-dimensional financial signals into salient risk cues. Through state recursion, it can encode temporal dependence, allowing the model to distill risk information from levels, changes, and volatility without manual construction of large polynomial or interaction sets. The result is a unified pipeline that performs feature composition, nonlinear transformation, and a probabilistic output via a logit link with reduced risk of model misspecification.

The learning objective is to estimate $P(\text{ICD}=1|x)$ reliably and to maintain stable out-of-sample ranking and discrimination. Parameters are estimated by minimizing a weighted binary cross-entropy loss, with regularization (early stopping) to curb overfitting and improve generalization. To address the scarcity and lag of true ICD disclosures, training uses IPO rejections as proxy labels to expand the library of risk signals, whereas validation and testing strictly revert to the ground-truth ICD disclosure criterion. This "proxy in training, truth in evaluation" design strengthens risk identification while containing the influence of proxy-label noise through a clear evaluation boundary.

4.2. Feature Engineering and Data Processing

This study uses financial-statement ratios as the primary inputs. These indicators can be validated against standardized disclosures and provide a low-cost, comparable mapping to key facets of internal control effectiveness. In terms of risk formation, funding stress, operating constraints, and governance incentives leave observable traces in financial condition and cash flows. Accordingly, financial ratios that summarize firms' pressure, opportunity, and execution provide a stable information base for forward-looking identification.

The features are grouped by economic meaning as follows: liquidity and short-term solvency (current ratio CR, quick ratio QR, current liabilities to total assets CL); profitability (return on assets ROA, operating profit margin OPM); operating efficiency (total asset turnover TAT, inventory turnover INVTO, accounts receivable turnover ARTO); capital structure and long-term leverage (debt-to-equity DE, long-term debt-to-equity LTDE); cash quality (operating cash flow to total assets CFO_TA, operating cash flow to revenue CFO_REV, net change in cash to total assets NCF_TA); asset structure (intangible assets to total assets INTAN, property, plant and equipment to total assets PPE); and size/scale (log total assets LN_TA, log revenue LN_REV). Variable definitions are reported in Table 1.

Table 1. Variable Definitions

Code	Name	Definition
CR	current ratio	current assets / current liabilities
QR	quick ratio	(current assets − inventory) / current liabilities
CL	current liabilities to total assets	current liabilities / total assets
ROA	return on assets	net income / total assets
OPM	operating profit margin	operating profit / operating revenue
TAT	total asset turnover	operating revenue / total assets
INVTO	inventory turnover	operating revenue / inventory
ARTO	accounts receivable turnover	operating revenue / accounts receivable
DE	debt-to-equity ratio	total liabilities / shareholders' equity
LTDE	long-term debt-to-equity	non-current liabilities / shareholders' equity
CFO_TA	operating cash flow to total assets	net cash flow from operating activities / total assets
CFO_REV	operating cash flow to revenue	net cash flow from operating activities / operating revenue
NCF_TA	net change in cash to total assets	net increase in cash and cash equivalents / total assets
INTAN	intangible assets to total assets	intangible assets / total assets
PPE	property, plant and equipment to total assets	property, plant and equipment / total assets
LN_TA	log total assets	ln(total assets + 1)
LN_REV	log revenue	ln(operating revenue + 1)

No separate category of growth variables is constructed. The research design adopts a rolling two-period window: for the same firm, revenues, assets, and related ratios from adjacent periods enter the model jointly. Information on expansion or contraction is thus implicitly captured by level differences and relative changes. This allows the model to learn the linkage between "levels and changes" directly from the raw financial ratios without injecting explicit growth rates that would raise collinearity. It also avoids look-ahead and timing-mixing risks. In short, growth is not represented as an independent variable; it is absorbed and utilized by the model through the existing ratios and windowed inputs.

Where financial restatements or misstatements exist, the original (pre-restatement) disclosed values are uniformly used for both training and evaluation. This choice serves three purposes. First, it preserves ex-ante information sets and prevents the introduction of post-hoc corrections that would create hindsight bias in the prediction period. Second, restatements are often outcomes of weak internal control; using corrected numbers would artificially "clean" signals left by ICDs and understate true, model-detectable risk. Third, original disclosures better reflect the contemporaneous information environment faced by investors and regulators, aligning predictive assessment with actual decision contexts. Together with consistent treatment conventions, this yields features that are economically meaningful and auditable, providing a reliable baseline for subsequent data splits, validation, and model evaluation.

All inputs used to predict ICDs at t+1 come strictly from information disclosed at t or earlier, ensuring forward-looking alignment. To reduce the influence of extreme observations on estimation and training, winsorization is applied at the 1st and 99th percentiles by year.

Material weaknesses, significant deficiencies, and general deficiencies disclosed in internal control evaluation reports are coded as "any deficiency" (ICD = 1). This broader definition rests on three considerations. First, from regulatory and governance perspectives, these categories lie on a common risk continuum; material weaknesses are tail events that are extremely rare and sensitive to exogenous forces such as inquiry intensity and disclosure conventions. Using only material weaknesses as positives would exacerbate class imbalance (reducing statistical power and stability) and introduce selection bias tied to instances of particularly strong oversight/disclosure. Second, for forward-looking identification, important/general deficiencies often precede material weaknesses; they reflect weaknesses in process execution, authorization, and information systems and therefore carry clear early-warning content. Including them helps the model learn the trajectory from mild to severe risk and improves recall of high-risk firms. Third, to mitigate measurement error and enhance comparability, grading differences are collapsed into a binary indicator of "any deficiency," which preserves economic meaning while reducing noise from heterogeneous grading practices across firms, industries, and years.

4.3. Data Splitting

The unit of observation is the firm–year. The goal is to assess out-of-sample, forward-looking discrimination for the next period's internal control deficiency (ICD). The predictive setup uses financial information at year t as inputs and outputs the probability that an ICD will be disclosed at t+1. To reflect expansion or contraction without introducing separate growth-rate variables, a rolling two-period window is applied on the input side: core ratios from adjacent periods jointly capture the linkage between levels and changes, allowing the model to absorb "growth" information end-to-end within existing variables.

Data are randomly split with stratification at the company level and firms are assigned to training and test sets (approximately 70:30). Within the training set, a further subset of firms (about 15%) forms the validation set. The split follows a company-level, mutually exclusive principle: the same firm never appears in more than one set, blocking any spillover of adjacent-period information for the same entity across sets. All preprocessing parameters are estimated only on the training set and then applied in a frozen manner to validation and test sets; no validation/test information feeds back into training or preprocessing. To avoid hindsight bias and preserve observable traces of control failures, training, validation, and test use pre-restatement original disclosures

whenever subsequent restatements or corrections exist.

Given the scarcity and lag of ground-truth ICD positives, IPO rejection outcomes are incorporated as proxy positives during training in a controlled way to raise the effective positive share and the density of learnable signals, thereby stabilizing decision boundaries in sparse positive regions. Recognizing that rejection does not always stem from weak internal control, robust constraints are imposed on proxy samples in training, including soft labels, separate loss weights, and joint control of the oversampling multiplier and proportion cap.

This training convention follows a "moderate augmentation, bounded influence, truth-anchored evaluation" rule. Proxy samples serve only to increase learnable density; the three constraints bound their influence; and the modeling objective and evaluation criterion remain anchored to ground-truth ICD disclosures. No proxy information is injected at validation or test. Model selection and threshold setting are based on validation performance, and the test set is reserved for final evaluation.

4.4. Evaluation Metrics

This study uses average precision (AP / PR-AUC) and the area under the ROC curve (ROC-AUC) as the primary criteria for model selection and comparison. The two are complementary under class imbalance: AP emphasizes minority-class detection effectiveness, while ROC-AUC captures global separability.

Average precision (AP) equals the area under the precision–recall curve and can be viewed as a recall-weighted average of precision. When recall increases (more potential deficiencies are identified) while precision remains high, AP rises accordingly. Because internal control deficiency (ICD) prediction is highly imbalanced, AP is more sensitive to the scarcity of positives and better reflects improvements in identifying the high-risk tail. AP is therefore set as the core metric.

ROC-AUC summarizes the trade-off between the true positive rate and the false positive rate across all thresholds and is equivalent to the probability that a randomly chosen positive receives a higher score than a randomly chosen negative. Ranging from 0.5 to 1 and being insensitive to class prevalence, ROC-AUC enables robust comparisons across datasets and time. A notable rise in ROC-AUC with only modest movement in AP typically indicates clearer overall ranking, but improvements on rare positives still require confirmation via AP.

For interpretability, continuous scores are also mapped to a hit/coverage view at a preset threshold of 0.5, reporting precision, recall, and $F1 = 2PR/(P+R)$. These indicators clarify the trade-off between false alarms and misses. Because they are threshold-dependent, they are not used for model selection.

5. Sample and Descriptive Statistics

Financial-statement data for listed firms are drawn from CSMAR from 2010 to 2022. Year 2022 serves as the cutoff for inputs so that internal control evaluation reports disclosed in 2023 provide the t+1 labels for forward-looking prediction. Given major differences in reporting structures, firms in finance and insurance are excluded, as are observations with missing key variables, yielding 39,785 firm-year observations.

The IPO rejection sample comes from Choice. Target firms are identified by records such as "not approved/not passed," "non-registration," or "registration terminated," with finance and insurance again excluded, resulting in 180 firms. For these firms, the three fiscal years prior to the IPO are hand-collected and cross-checked from prospectuses and related disclosures, producing a theoretical 180 × 3 = 540 firm-years. Because the research design treats two adjacent years as one record to construct level–change linkages, observations must form consecutive pairs. After removing non-pairable cases and those missing key variables, 524 valid records remain.

Table 2. Descriptive Statistics (Listed Firms)

Variable	Mean	Std. Dev.	P25	Median	P75
CR	2.672	2.957	1.165	1.726	2.937
QR	2.172	2.747	0.784	1.282	2.366
CL	0.337	0.183	0.195	0.316	0.454
ROA	0.035	0.079	0.013	0.038	0.069
OPM	0.065	0.267	0.024	0.080	0.159
TAT	0.604	0.421	0.335	0.509	0.744
INVTO	22.794	95.999	2.893	5.225	9.548
ARTO	36.947	173.106	2.811	5.129	12.124
DE	1.115	1.420	0.323	0.681	1.336
LTDE	0.228	0.393	0.016	0.070	0.263
CFO_TA	0.045	0.074	0.006	0.044	0.086
CFO_REV	0.085	0.201	0.010	0.081	0.171
NCF_TA	0.020	0.112	-0.029	0.006	0.049
INTAN	0.047	0.053	0.017	0.034	0.058
PPE	0.242	0.179	0.101	0.207	0.347
LN_TA	22.112	1.314	21.181	21.932	22.856
LN_REV	21.391	1.506	20.379	21.257	22.262

Table 2 reports descriptive statistics for the listed-firm sample (excluding IPO-rejected firms). The means of log total assets (LN_TA) and log revenue (LN_REV) are 22.112 and 21.391, with medians 21.932 and 21.257, respectively. In asset structure, property, plant and equipment to total assets (PPE) has a mean 0.242 and median 0.207, exceeding

intangible assets to total assets (INTAN) with mean 0.047 and median 0.034. In capital structure, long-term debt-to-equity (LTDE) shows a mean 0.228 and median 0.070, indicating substantial dispersion. Regarding cash quality, operating cash flow to total assets (CFO_TA) averages 0.045 (median 0.044), operating cash flow to revenue (CFO_REV) averages 0.085 (median 0.081), and net change in cash to total assets (NCF_TA) averages 0.020 (median 0.006), reflecting notable year-to-year cash volatility.

Table 3 reports the annual incidence of internal control deficiencies (ICDs) over the sample window. Rows are indexed by fiscal year t; each row aligns financial information at t with ICD disclosures in t+1. The "2010" row (56 cases, 3.51%) therefore corresponds to disclosures made in 2011, and the "2022" rate of 2.78% corresponds to 2023 disclosures. In total, there are 39,785 firm-year observations, of which 2,224 disclose at least one deficiency, yielding an overall incidence of 5.59%. The rate peaks at 8.31% in 2012 (the sample maximum), remains elevated in 2017 (7.85%), and then declines to 2.78% by 2022.

Table 3. Annual Distribution of Internal Control Deficiencies (Listed Firms)

Year	Firm–years	ICD count	ICD rate
2010	1596	56	3.51
2011	2070	139	6.71
2012	2249	187	8.31
2013	2441	153	6.27
2014	2468	163	6.60
2015	2631	152	5.78
2016	2948	177	6.00
2017	3313	260	7.85
2018	3399	231	6.80
2019	3576	211	5.90
2020	3995	166	4.16
2021	4395	198	4.51
2022	4704	131	2.78
Total	39785	2224	5.59

6. Evaluation and Analysis of ICD Prediction Performance

6.1. Benchmark Setup and Out-of-Sample Results

Table 4 compares two models on the same test set. Column (2) reports the baseline model, trained without IPO data under default hyperparameters. Column (3) reports the IPO-tuned model, which mixes IPO rejections as proxy positives during training and selects hyperparameters (label-smoothing strength, proxy-sample weight/proportion, class weights, etc.) by search. Data processing and evaluation are identical for both models, including a company-level mutually exclusive split, forward-looking alignment, and preprocessing parameters estimated on the training set and then frozen. Therefore, performance differences mainly reflect the introduction of proxy information and tuning or regularization choices.

Ranking quality improves jointly. Average precision (AP / PR-AUC) rises from 0.079 to 0.097 (+22.78%), indicating stronger ordering on the scarce positive class (ICDs). Given the overall incidence of 5.59% in Table 3, the AP-to-incidence ratio is about 1.41 for the baseline and 1.74 for the IPO-tuned model, implying that with the same review resources the tuned model surfaces true ICD cases more effectively toward the top of the list and is more sensitive to the high-risk tail. ROC-AUC increases from 0.620 to 0.648 (+4.52%), meaning the probability that a randomly drawn ICD observation receives a higher score than a non-ICD observation rises from 62.0% to 64.8%. The joint gains in these threshold-invariant metrics indicate learning of more discriminative financial-signal combinations rather than threshold-specific artifacts.

At a fixed threshold of 0.5, F1 improves from 0.133 to 0.136 (+2.26%). Precision is essentially unchanged (+1.32%), while recall increases from 0.552 to 0.578 (+4.71%). Thus, at the preset threshold the IPO-tuned model achieves broader coverage of potential deficiencies without materially diluting precision.

Table 4. Out-of-Sample Performance (Baseline Model vs. IPO-Tuned Model)

Metric	Baseline	IPO-tuned	Change
AP (PR-AUC)	0.079	0.097	22.78%
ROC-AUC	0.620	0.648	4.52%
F1 @ 0.5	0.133	0.136	2.26%
Precision @ 0.5	0.076	0.077	1.32%
Recall @ 0.5	0.552	0.578	4.71%

6.2. Full-Parameter Optimal Model

A full-parameter search is conducted under the same train/validation/test split and evaluation convention; the configuration that performs best on the validation set is then evaluated once on the test set. The search spans network size and regularization strength, learning rate and batch size, class weights, and proxy-sample settings (label-smoothing strength, loss weights, and mini-batch proportion). The 'global' optimum attainable under the given data and training protocol is reported in Table 5, column (2).

Relative to the baseline model, core metrics rise in tandem. AP (PR-AUC) increases from 0.079 to 0.098 (+24.05%), and ROC-AUC from 0.620 to 0.653 (+5.32%), indicating improved ability to raise recall while maintaining precision on the scarce positive class and stronger separability across thresholds. Using "AP / baseline incidence" to gauge tail improvement, the optimum also dominates the baseline, implying that with the same review resources more true ICD cases are pushed toward the top of the list. At the fixed threshold of 0.5, F1 rises to 0.146 (+9.77%); precision to 0.083 (+9.21%); and recall to 0.590 (+6.88%), delivering higher coverage and a higher hit rate without changing the threshold.

Compared with the IPO-tuned model, gains are incremental at the ranking level: AP from 0.097 to 0.098 (+1.03%), and ROC-AUC from 0.648 to 0.653 (+0.77%). This suggests diminishing returns after one effective round of tuning, with the full search mainly refining and smoothing the decision boundary. At the fixed threshold of 0.5, however, improvements are more tangible: F1 from 0.136 to 0.146 (+7.35%), precision from 0.077 to 0.083 (+7.79%), and recall from 0.578 to 0.590 (+2.08%), translating ranking gains into actionable review benefits.

Table 5. Out-of-Sample Performance of the Full-Parameter Optimal Model

Metric	Full-parameter optimal model	vs. Baseline	vs. IPO-tuned
AP (PR-AUC)	0.098	24.05%	1.03%
ROC-AUC	0.653	5.32%	0.77%
F1 @ 0.5	0.146	9.77%	7.35%
Precision @ 0.5	0.083	9.21%	7.79%
Recall @ 0.5	0.590	6.88%	2.08%

6.3. Variable Importance and Relative Contribution

Table 6 shows that property, plant and equipment (PPE) and profitability measures (ROA, OPM) consistently rank near the top across all three models. For PPE, the change in average precision (ΔAP) is 0.011 (14.13%) in the baseline model, 0.018 (18.08%) in the IPO-tuned model, and 0.012 (12.38%) in the full-parameter model. ROA exhibits a similar pattern, with relative contributions of 11.24% (baseline), 15.52% (IPO-tuned), and 11.70% (full-parameter), indicating that asset structure and earnings quality provide stable ordering power for identifying ICD risk. OPM also peaks in the IPO-tuned configuration at 0.011 (11.20%), exceeding the baseline's 0.007 (9.22%) and the full-parameter model's 0.007 (6.93%).

Liquidity and leverage display model-specific patterns. QR is broadly comparable across models: 0.007 (8.64%) in the baseline, 0.007 (7.55%) in the IPO-tuned model, and 0.006 (6.35%) in the full-parameter model. LTDE is highest under IPO tuning at 0.007 (6.84%), versus 0.005 (5.91%) in the baseline and 0.005 (5.36%) in the full-parameter model. Size and short-term pressure also diverge by model: LN_TA registers 0.009 (9.39%) in the IPO-tuned model, compared with 0.002 (3.15%) in the baseline and 0.005 (5.29%) in the full-parameter model; CL is more pronounced only in the full-parameter model at 0.003 (2.91%), remaining negligible in the baseline (0.00%) and modest in the IPO-tuned model (0.001, 1.40%).

For cash quality, CFO_TA contributes 0.006 (6.57%) in the IPO-tuned model and 0.005 (5.77%) in the baseline, but attenuates to 0.003 (2.78%) in the full-parameter model, suggesting that the full search diffuses importance away from cash-flow signals toward a broader mix spanning liquidity, size, and leverage. DE similarly rises under IPO tuning, at 0.005 (5.39%), relative to the baseline and full-parameter settings (both 0.002; 2.08% and 2.24%, respectively), while LN_REV remains minor across models (0.00%, 1.24%, and 2.21%).

Overall, the IPO-tuned configuration concentrates weight on PPE, ROA, OPM, and firm size, whereas the full-parameter model distributes contributions more evenly across core features (for example, higher CL and mid-range LN_TA and LTDE), reflecting different emphasis rather than a uniform uplift of any single variable.

Table 6. Variable Importance and Relative Contribution (Top 10)

Variable	Baseline model		IPO-tuned model		Full-parameter optimal model	
	ΔAP	ΔAP/AP	ΔAP	ΔAP/AP	ΔAP	ΔAP/AP
PPE	0.011	14.13%	0.018	18.08%	0.012	12.38%
ROA	0.009	11.24%	0.015	15.52%	0.011	11.70%
OPM	0.007	9.22%	0.011	11.20%	0.007	6.93%
QR	0.007	8.64%	0.007	7.55%	0.006	6.35%
LTDE	0.005	5.91%	0.007	6.84%	0.005	5.36%
LN_TA	0.002	3.15%	0.009	9.39%	0.005	5.29%
CL	0.000	0.00%	0.001	1.40%	0.003	2.91%
CFO_TA	0.005	5.77%	0.006	6.57%	0.003	2.78%
DE	0.002	2.08%	0.005	5.39%	0.002	2.24%
LN_REV	0.000	0.00%	0.001	1.24%	0.002	2.21%

7. Conclusion

Anchored in the IPO issuance regime, this study proposes a "proxy in training, truth in evaluation" framework for predicting ICDs. Empirically, introducing proxy labels yields steady gains over the baseline (+22.78% PR-AUC, +4.52% ROC-AUC), and the full-parameter search achieves further improvements (+24.05% PR-AUC, +5.32% ROC-AUC). Fixed-threshold metrics (F1, precision, recall) also improve, indicating broader coverage of high-risk firms without sacrificing hit rate. These findings can be translated into practice by structuring rejection decisions and inquiry focal points into ICD-indicative indicators for routine monitoring and tiered supervision, aligning early-warning thresholds and review checklists with financial ratios. Standardizing regulatory inquiry corpora, audit opinions, sanctions, and modified audit reports—and linking them to XBRL disclosures—would provide auditable, long-horizon data support for interpretable risk-control models.

Several caveats point to avenues for refinement. The design centers on financial ratios for availability and auditability; unstructured text has not yet been incorporated systematically, suggesting future integration of textual features from prospectuses and annual-report inquiry responses. Moreover, institutional and economic shifts may induce distribution shift and threshold drift; practical deployment therefore calls for rolling retraining, monitoring and alerting, and probability calibration, together with an evaluation of cost–benefit trade-offs under alternative thresholding strategies.

Acknowledgments

This work was supported by the Humanities and Social Sciences Research Youth Fund of the Ministry of Education of China (Project title: "Machine Learning for Internal Control Deficiency Prediction: Model Development and Applications"; Grant No. 19YJC790072).

References

[1] Donelson D C, Ege M, McInnis J M. Internal control weaknesses and financial reporting fraud[J]. Auditing: A Journal of Practice & Theory, 2017, 36(3): 45-69.

[2] Boulhaga M, Bouri A, Elbardan H. The effect of internal control quality on real and accrual-based earnings management: Evidence from France[J]. Journal of Management Control, 2022, 33(4): 545-567.

[3] Liu C, Lin B. Deep-Learning-Based Prediction of Internal Control Deficiencies in Listed Firms: New Theories and Methods[J]. Accounting Research (China), 2024, (06): 119-134.

[4] Doyle J T, Ge W, McVay S E. Accruals quality and internal control over financial reporting[J]. The Accounting Review, 2007, 82(5): 1141-1170.

[5] Ashbaugh-Skaife H, Collins D W, Kinney W R Jr, LaFond R. The effect of SOX internal control deficiencies on firm risk and cost of equity[J]. Journal of Accounting Research, 2009, 47(1): 1-43.

[6] Kim J-B, Song B Y, Zhang L. Internal control weakness and stock price crash risk[J]. Accounting & Finance, 2019, 59(2): 1197-1233.

[7] Costello A M, Wittenberg-Moerman R. The impact of financial reporting quality on debt contracting: Evidence from internal control weakness reports[J]. Journal of Accounting Research, 2011, 49(1): 97-136.

[8] Chen H, Wang S, Yang D, Zhou N. COSO-based internal control and comprehensive enterprise risk management: Institutional background and research evidence from China[J]. Encyclopedia, 2025, 5(3): 106.

[9] Cheng Q, Goh B W, Kim J-B. Internal control and operational efficiency[J]. Contemporary Accounting Research, 2018, 35(2): 1102-1139.

[10] Lee J, Cho E, Choi H. The effect of internal control weakness on investment efficiency[J]. Journal of Applied Business Research, 2016, 32(3): 649-662.

[11] Doyle J T, Ge W, McVay S E. Determinants of weaknesses in internal control over financial reporting[J]. Journal of Accounting and Economics, 2007, 44(1-2): 193-223.

[12] Lin B, Lin D, Hu W, et al. A Goal-Oriented Internal Control Index[J]. Accounting Research (China), 2014, (08): 16-24.

[13] Lin B, Lin D, Xie F, et al. An Information-Disclosure-Based Internal Control Index[J]. Accounting Research (China), 2016, (12): 12-20.

[14] Dechow P M, Ge W, Larson C R, Sloan R G. Predicting material accounting misstatements[J]. Contemporary Accounting Research, 2011, 28(1): 17-82.

[15] Cecchini M, Aytug H, Koehler G J, Pathak P. Detecting management fraud in public companies[J]. Management Science, 2010, 56(7): 1146-1160.

[16] Perols J. Financial statement fraud detection: An analysis of statistical and machine learning algorithms[J]. Auditing: A Journal of Practice & Theory, 2011, 30(2): 19-50.

[17] Bao Y, Ke B, Li B, Yu Y J, Zhang J. Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach[J]. Journal of Accounting Research, 2020, 58(1): 199-235.

[18] Craja P, Kim A, Lessmann S. Deep learning for detecting financial statement fraud[J]. Decision Support Systems, 2020, 139: 113421.