Peer reviewed Journals

SciChronice

RESEARCH

Service Specific Micro Adaptation for Specimen Labeling Discrepancy Detection in Surgical Pathology Accession and Grossing Workflows

Nguyen Minh Kiet,[1] Tran Quoc Bao,[2] and Le Duc Thang[3]

[1]Faculty of Information Technology, Quy Nhon University, 170 An Duong Vuong Street, Nguyen Van Cu Ward, Quy Nhon City 55100, Binh Dinh Province, Vietnam
[2]Department of Computer Science, Thai Binh University, Tan Binh Ward, Thai Binh City 410000, Thai Binh Province, Vietnam
[3]Faculty of Engineering and Technology, Hong Duc University, 565 Quang Trung Street, Dong Ve Ward, Thanh Hoa City 440000, Thanh Hoa Province, Vietnam

Abstract

Surgical pathology depends on the correct linkage between patient identity, specimen container, requisition text, tissue site, laterality, procedure description, and gross-room handling. A discrepancy at accession or grossing can delay diagnosis, require recollection, create amended reports, or in rare circumstances place a patient at risk of an incorrect result being associated with the wrong body site or specimen. Automated safety review is possible because accession systems contain structured fields and free-text requisitions, yet the language and workflow patterns differ across services and hospitals. This paper reports an empirical study of service-specific micro adaptation for detecting specimen labeling discrepancies before final pathology sign-out. We assembled 128,740 surgical pathology cases from three hospitals and manually adjudicated 9,860 discrepant or high-risk cases into five discrepancy categories: patient-identifier mismatch, container-requisition mismatch, site or laterality conflict, specimen-count inconsistency, and unsupported free-text amendment. A general transformer classifier was compared with a micro-adapted model that uses small service-specific adapter modules, accession-field consistency checks, and gross-room event features. The proposed model improved macro F_1 from 0.692 to 0.801, increased recall for site or laterality conflict from 70.5% to 86.9%, and reduced false safety holds from 11.8 to 7.2 per 1,000 cases. Cross-hospital temporal testing showed smaller degradation than a single global model. Error review indicated that micro adaptation helped most in breast, dermatopathology, gastrointestinal, and gynecologic workflows, where terminology and container conventions differed substantially.

1. Introduction

A surgical pathology specimen begins its diagnostic life before a pathologist examines tissue under a microscope. It arrives in a container, with a label, a requisition, a procedure description, and often a sequence of location details that specify organ, side, lesion, margin, or biopsy level. The accessioning and grossing process turns these materials into a case record that will later support microscopic interpretation, diagnostic reporting, billing, and communication with clinicians. The safety of that process depends on many small correspondences. The patient identifier on the container must match the requisition. The number of containers must agree with the specimen list. The description of the tissue site should not contradict the procedure [1]. Laterality should be consistent across label, requisition, and gross description. A separately submitted margin should not be merged with a primary specimen unless that is intended. A free-text correction should not silently override structured fields without review.

Laboratories already use manual checks, barcodes, accession rules, and supervisory review [2]. Even with those controls, discrepancy detection remains difficult because many problems are not visible as a single missing field. A container may say “left breast,” while the requisition says “right breast mass.” A dermatology specimen may be labeled with an anatomic phrase that is common in one clinic and ambiguous in another [3]. A gastrointestinal case may contain multiple jars with similar levels, such as distal, proximal, random, terminal, and rectal, where one swapped descriptor changes the clinical meaning. A gynecologic specimen may use abbreviations that are clear to the submitting clinic but incomplete for accession. These problems are not simply typographic; they are relational.

This paper studies automated discrepancy review

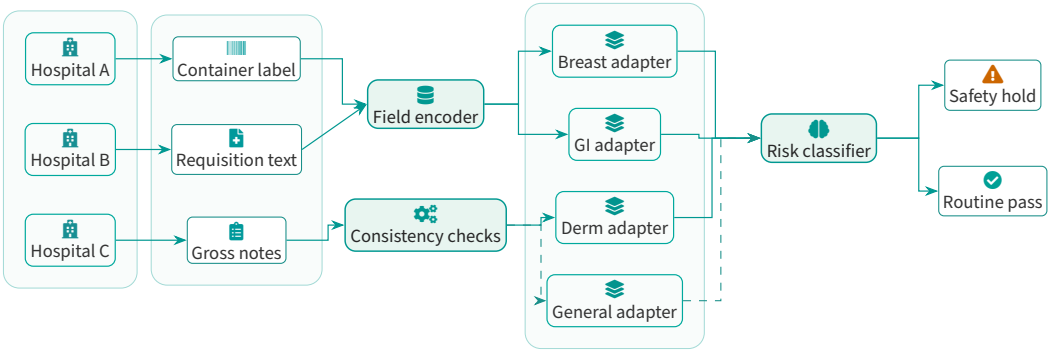

Figure 1. Service-specific micro adaptation links hospital accession streams, structured consistency checks, and compact adapters before issuing pre-sign-out review recommendations.

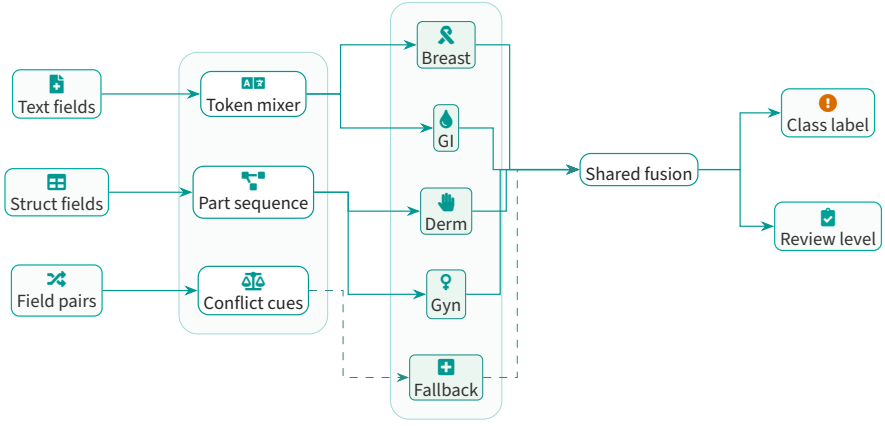

Figure 2. Adapter routing preserves shared accession knowledge while letting high-volume services specialize in local specimen wording, part ordering, and discrepancy cues.

for surgical pathology accession and grossing. The goal is not to replace accession staff, grossing personnel, pathologists, or laboratory policy. It is to identify cases that deserve safety review before the case advances too far into the diagnostic workflow. A useful model should catch discrepancies that are easily missed during high-volume processing, while avoiding excessive false holds that slow routine cases. A model that flags every unusual phrase is not useful. A model that misses laterality conflict because it treats all anatomic text as free-form language is also not useful. The problem requires both structured field comparison and local knowledge of service-specific wording.

The proposed approach is called service-specific micro adaptation. It starts with a shared language and field encoder trained across hospitals. It then adds small adapter modules for major pathology services, such as breast, dermatopathology, gastrointestinal, genitourinary, gynecologic, head and neck, and general surgical pathology. These adapters do not replace the global model. They adjust the representation for local terminology, specimen conventions, and field patterns. A breast adapter learns that "clock face," "margin," "clip," and laterality phrases are often safety-relevant. A dermatopathology adapter learns that skin site text can be long, informal, and full of clinic-specific body-location shorthand. A gastrointestinal adapter learns that specimen count and level sequence matter. The central hypothesis is that small service-specific adaptation improves discrepancy detection more effectively than simply increasing the size of one global classifier.

A methodological precedent for this choice comes from a different medical artificial intelligence setting. Sadanandan and Behzadan [4] reported that targeted LoRA applied to selected layers of a medical vision-language model used only 0.1% of parameters yet achieved the best accuracy-consistency trade-off across three datasets, outperforming a substantially larger MedGemma-27B model on the compared metrics; in the present study, that result is taken only as a design cue that selective adaptation can be more useful than indiscriminate scale when the task depends on narrow clinical behavior.

The study uses de-identified surgical pathology accession data from three hospitals. Cases include structured fields, container labels, requisition text, procedure descriptions, gross-room timestamps, specimen counts, amendment notes, and service routing. A subset of discrepancy-enriched cases was manually reviewed by pathology quality personnel and pathologist adjudi-

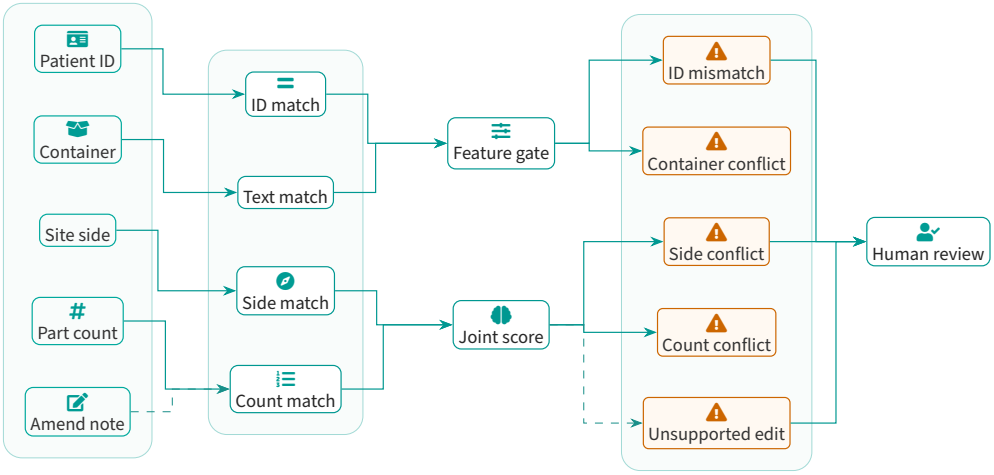

Figure 3. Relational consistency features transform accession fields into discrepancy evidence without allowing deterministic checks to become final decisions.

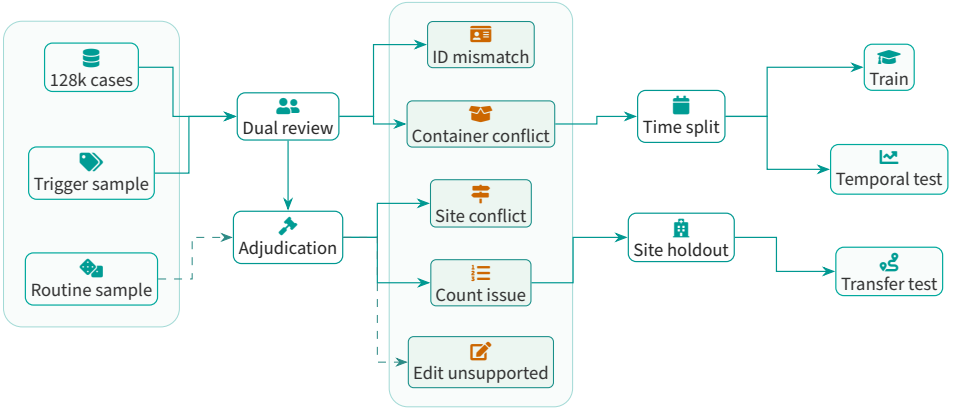

Figure 4. Discrepancy labels are derived from enriched and routine samples, then separated by accession time and hospital to test drift and transfer.

cators. The primary outcome is a five-class discrepancy label, with a no-discrepancy class used for routine cases. The evaluation emphasizes recall for safety-relevant discrepancy types, false safety holds per 1,000 cases, temporal robustness, and cross-hospital transfer.

The paper is deliberately practical. It does not propose a new diagnostic pathology model and does not analyze slides. It stays at the accession and grossing layer, where many preventable workflow errors can be intercepted before interpretation. It also does not assume that every discrepancy is clinically severe. Some discrepancies are clerical and resolved quickly. Others require specimen hold, clinician contact, amended accession, or pathologist review. The model therefore produces a recommended review level rather than a final legal or diagnostic judgment.

The findings show that service-specific micro adaptation improves both detection and workflow efficiency. The global transformer model performed reasonably on common discrepancy patterns but struggled with service-specific language. The micro-adapted model improved macro F_1, laterality conflict recall, specimen-count inconsistency recall, and cross-hospital temporal stability. It also reduced false holds, suggesting that adapters did not merely make the model more aggressive. Manual error review indicated that the model learned useful local patterns but still failed when the source record lacked enough information or when submitting clinics used highly ambiguous shorthand.

2. Data Sources and Discrepancy Labels

The study corpus contained 128,740 surgical pathology cases accessioned across three hospitals over a thirty-month period. The hospitals differed in case mix and accession workflow. Hospital A was an academic center with high subspecialty volume and complex resections. Hospital B was a community hospital with broad general surgical pathology and a high proportion of endoscopy specimens [5]. Hospital C was a specialty referral hospital with large breast, dermatopathology, and gynecologic volumes. All data were de-identified before modeling. Direct patient identifiers, accession

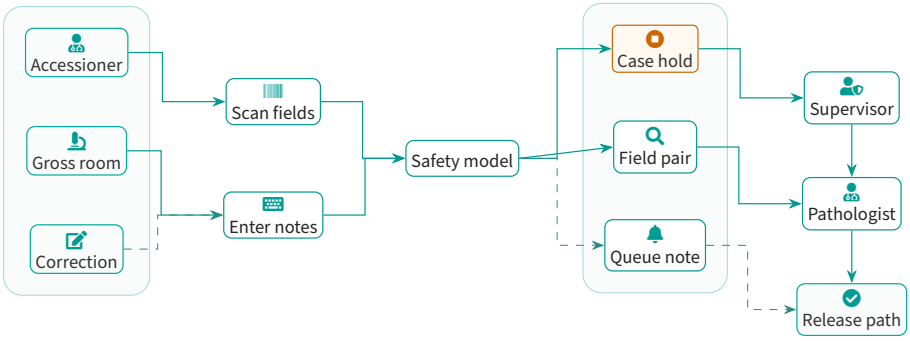

Figure 5. The deployment loop keeps the model in a recommendation role, surfacing review evidence while supervisors and pathologists resolve discrepancies.

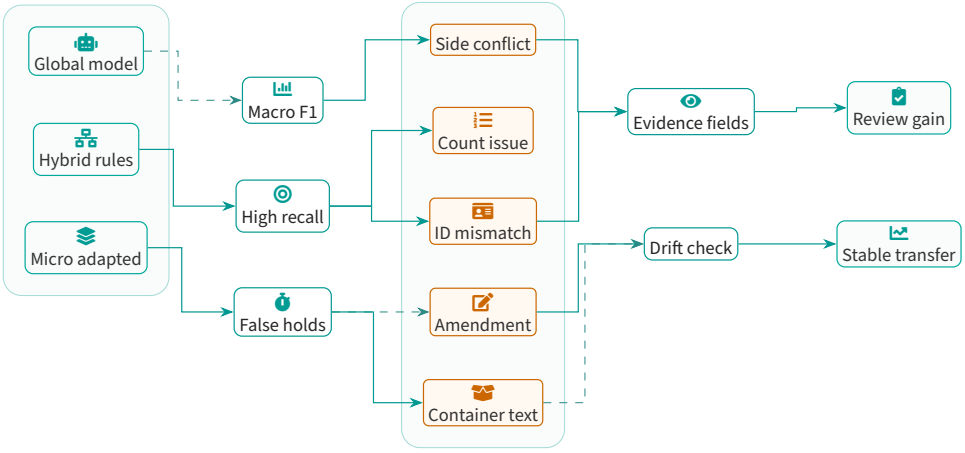

Figure 6. Performance analysis connects model variants with safety-critical labels, explanation fields, false-hold behavior, and temporal transfer stability.

numbers, clinician names, clinic addresses, and exact dates were removed or replaced with typed placeholders.

Each case record contained structured accession fields and text fields. Structured fields included service line, container count, specimen part labels, procedure category, collection location, received time category, grossing status, amendment status, and quality event code when present. Text fields included container label text, requisition diagnosis or indication, procedure description, submitted specimen description, gross-room notes, accession comments, and free-text correction notes. The model did not receive final diagnostic text for the main prediction task because the goal was to detect discrepancy before sign-out. Final diagnostic text was used only during retrospective adjudication when reviewers needed to understand whether a discrepancy had later affected reporting.

The manually adjudicated set contained 9,860 cases. Sampling intentionally enriched likely discrepancies because true discrepancies are uncommon in routine pathology flow. Cases were selected using existing quality-event logs, accession comment triggers, laterality mismatch rules, amendment codes, specimen-count mismatch rules, and random sampling of routine cases. The random routine sample was necessary because rule-triggered cases contain a biased mixture of obvious problems and false alarms. The adjudicated set contained 4,180 no-discrepancy cases and 5,680 cases with one primary discrepancy label. Secondary discrepancy tags were also recorded when a case had more than one issue.

The five discrepancy categories were patient-identifier mismatch, container-requisition mismatch, site or laterality conflict, specimen-count inconsistency, and unsupported free-text amendment. Patient-identifier mismatch included discordance among container label, requisition, and accession record. Container-requisition mismatch occurred when the tissue described on the container differed from the tissue described on the requisition without a documented correction. Site or laterality conflict included left-right discordance, incompatible body site, conflicting margin orientation, or incompatible lesion location. Specimen-count inconsistency included missing containers, extra containers, duplicate part letters, skipped part labels, or mismatch between submitted count and accessioned count [6]. Unsupported free-text amendment included a correc-

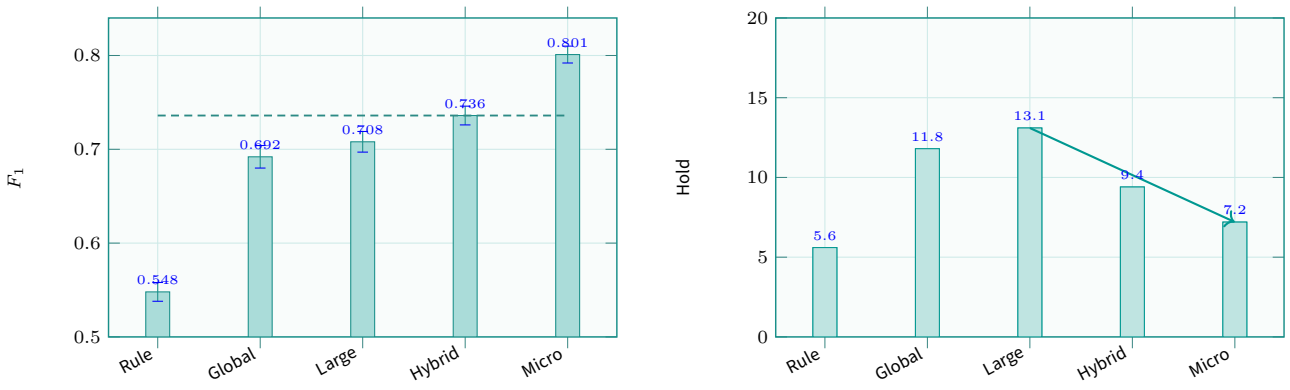

Figure 7. Model comparison.

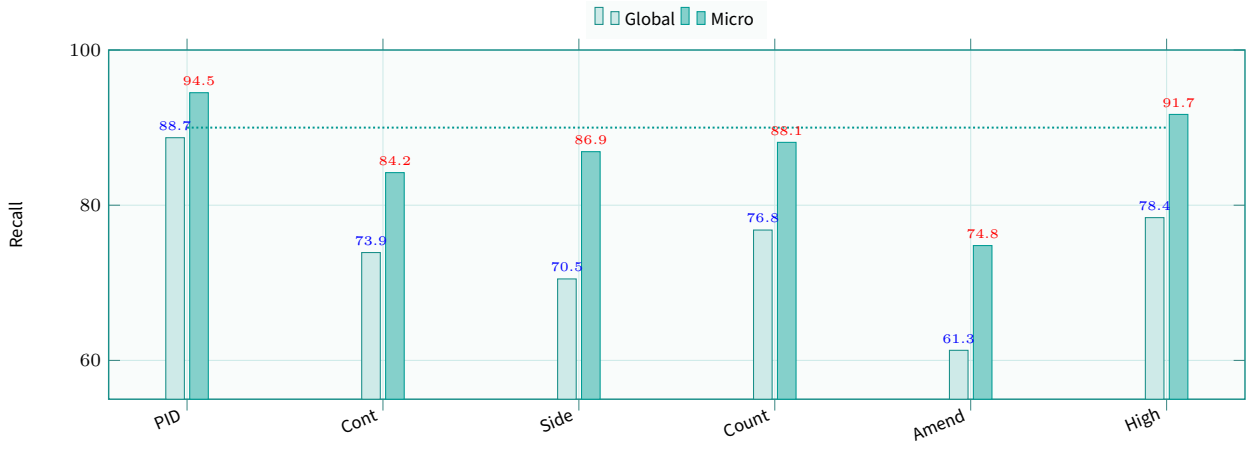

Figure 8. Recall lift.

tion note that changed a safety-relevant field without enough supporting documentation.

The category distribution in the adjudicated set was uneven. No-discrepancy cases made up 42.4%. Container-requisition mismatch made up 16.2%. Site or laterality conflict made up 13.8%. Specimen-count inconsistency made up 12.5%. Unsupported free-text amendment made up 8.9%. Patient-identifier mismatch made up 6.2%. Patient-identifier mismatch was less frequent but treated as high priority because it can affect the entire case identity. Site or laterality conflict was also treated as high priority because it can affect downstream clinical interpretation, procedure correlation, and treatment planning.

Adjudication was performed by a team that included pathology quality specialists, senior accessioning staff, gross-room supervisors, and pathologists. Each case was reviewed by two annotators. Disagreements were resolved by a pathologist or laboratory quality lead. Reviewers examined the de-identified container text, requisition text, accession record, correction comments, quality event code, and available workflow notes. They assigned a primary discrepancy category, a severity level, and a recommended workflow action. Severity levels were informational, moderate, and high. High severity included patient-identity concern, unresolved laterality conflict, missing specimen container, and unsupported correction to a critical field.

Agreement before adjudication was 84.1% for discrepancy versus no discrepancy and 76.8% for the six-way label including no discrepancy. Agreement was highest for patient-identifier mismatch and specimen-count inconsistency. Agreement was lowest for unsupported free-text amendment because reviewers sometimes differed on whether documentation was sufficient. Site or laterality conflict had high agreement when left-right words were explicit, but lower agreement when location was expressed through body-region shorthand or procedure-specific wording. These disagreement patterns motivated the model's use of service-specific adapters and uncertainty reporting.

The full non-adjudicated corpus was used for self-supervised adaptation and field reconstruction tasks. These tasks included predicting masked specimen site phrases, reconstructing structured service line from text, identifying whether two specimen parts belonged to the same container sequence, and distinguishing true part labels from shuffled part labels. The purpose was

Table 1. Dataset Composition Across Hospitals

Hospital	Primary Focus	Approximate Case Pattern	Notable Workflow Traits
Hospital A	Academic center	Complex resections	High subspecialty volume
Hospital B	Community hospital	Broad surgical pathology	Endoscopy-heavy workload
Hospital C	Specialty referral	Breast and gyn emphasis	Specialized accession terminology
Combined Cohort	Multi-site corpus	128,740 cases	30-month collection period
Adjudicated Subset	Enriched review set	9,860 cases	Dual-review adjudication

Table 2. Primary Discrepancy Categories

Discrepancy Type	Proportion (%)	Typical Example
Patient-identifier mismatch	6.2	Container and requisition identity conflict
Container-requisition mismatch	16.2	Tissue description disagreement
Site or laterality conflict	13.8	Left-right inconsistency
Specimen-count inconsistency	12.5	Missing or duplicate container
Unsupported amendment	8.9	Unverified correction note
No discrepancy	42.4	Routine concordant case

to help the encoder learn laboratory language without using unverified discrepancy labels as truth. Historical quality-event codes were not used as final labels because they reflected local reporting habits and did not capture all discrepancies.

Data were split by accession month and hospital. The primary training set included earlier cases from all three hospitals. The validation set contained later cases from the same period but separate patients and accession batches. The internal test set contained 2,420 adjudicated cases. A temporal test set contained 1,380 adjudicated cases from a later six-month period after two hospitals changed accession templates. A cross-hospital test simulated deployment by training without one hospital and evaluating on that hospital's adjudicated cases. This was repeated for each hospital. The same manually adjudicated labels were used for all final scoring.

The case mix differed substantially across services [7]. Breast cases had many laterality and margin-orientation fields. Dermatopathology cases had many body site phrases, sometimes with clinic-specific abbreviations. Gastrointestinal cases had multiple containers and sequential biopsy sites. Gynecologic cases often contained procedure terms with abbreviations and combined specimens. Genitourinary cases contained laterality, level, and core-count information. General surgical pathology included the widest range but fewer repeated conventions. These service differences were the basis for adapter design.

The study did not include slide images, molecular data, or final diagnoses in the predictive input. It also did not include staff identity, clinician identity, patient demographics, or exact clinic names [8]. The intention was to evaluate whether accession and grossing records themselves contained enough signal for discrepancy screening. Including staff or clinic identity might improve apparent performance but could create undesirable monitoring or bias effects. The model is designed to review cases, not people [9].

3. Micro Adapted Discrepancy Detection Model

The model has three information streams. The first stream encodes free text from container labels, requisition text, procedure descriptions, submitted specimen descriptions, accession comments, and gross-room notes. The second stream encodes structured accession fields such as service, container count, part labels, and workflow status. The third stream encodes consistency checks derived from the relationship among fields. These checks include whether laterality terms agree, whether part counts align, whether site terms are compatible, whether amendment text changes a structured field, and whether a container label lacks a corresponding requisition part [10].

The base text encoder was trained on the full de-identified corpus using masked field reconstruction and next-field prediction. It learned laboratory phrases such as "jar," "part," "submitted as," "designated," "margin," "left," "right," "random," "distal," "proximal," "biopsy," "excision," and "received in formalin." It also learned service-specific abbreviations and spelling variants. The encoder did not use final diagnosis as supervision. After this pretraining stage, the encoder was fine-tuned on adjudicated discrepancy labels.

Service-specific micro adapters were placed inside the encoder. Each adapter was a small trainable module activated by the service line [11]. The base encoder remained shared, so common concepts such as patient-

Table 3. Architecture Components of the Proposed Model

Component	Input Sources	Primary Role
Text encoder	Requisitions, labels, notes	Learn pathology language patterns
Structured stream	Counts, status fields, labels	Encode workflow metadata
Consistency checks	Cross-field comparisons	Detect relational conflicts
Service adapters	Service-specific terminology	Capture local workflow semantics
Review-level head	Risk severity features	Generate hold recommendations
Explanation module	Attention and evidence fields	Highlight relevant discrepancies

Table 4. Comparison of Baseline and Proposed Models

Model	Macro F_1	False Holds / 1,000	Key Limitation
Rule system	0.548	5.6	Weak language understanding
Global transformer	0.692	11.8	Service-language confusion
Large global transformer	0.708	13.1	Increased false alarms
Hybrid rule-classifier	0.736	9.4	Limited specialization
Micro-adapted model	0.801	7.2	Sensitive to hybrid services

identifier conflict or generic part-count mismatch could be learned across services. The adapters adjusted the representation for service-local language [12]. For example, the breast adapter was allowed to specialize in laterality, margins, clips, and clock-face location [13]. The gastrointestinal adapter specialized in container sequence and level terms. The dermatopathology adapter specialized in body-site text and biopsy descriptors. The model also had a fallback general adapter for mixed or uncertain service routing.

The structured field stream used embeddings for service, procedure category, container count, part count, amendment status, and grossing status. Numeric counts were represented as categories rather than continuous values because a difference between one and two containers is more important than a difference between 21 and 22 in many workflows. Part labels were encoded as sequences. Duplicate, skipped, or repeated labels were represented explicitly. The structured stream was especially important for specimen-count inconsistency, where text alone often missed the issue.

The consistency-check stream used deterministic comparisons [14]. These comparisons did not make final decisions by themselves [15]. They created features for the model. Laterality checks identified whether left, right, bilateral, upper, lower, medial, lateral, and side-specific abbreviations appeared consistently across fields. Site compatibility checks mapped phrases to broad anatomic groups and then identified conflicts. Count checks compared received containers, accessioned parts, and text mentions of specimen number. Amendment checks detected whether correction notes introduced a new patient identifier, site, side, count, or container description. The model could learn when a check mattered and when it was likely a benign wording difference.

The classifier produced a discrepancy category and a review level. Review levels were no hold, accessioner review, supervisor review, and pathologist or quality lead review. The discrepancy category supported explanation, while the review level supported workflow. A laterality conflict in a breast excision usually required higher review than a minor wording inconsistency in a routine skin biopsy. A patient-identifier mismatch usually required immediate hold. The model was trained to predict both category and review level, but the final evaluation emphasizes discrepancy category and false safety holds.

Training used three kinds of examples. Clear adjudicated examples trained the category classifier directly. Reviewer-disagreement examples trained the model to distribute probability across plausible categories while preserving the adjudicated primary label [16]. Routine cases trained the model to avoid false holds. High-severity cases received larger loss weight, but the weight was capped to prevent the model from flagging every uncommon phrase as dangerous. The validation set was used to choose thresholds that balanced high-severity recall against false holds per 1,000 cases.

The model also produced a local explanation. It identified the fields that contributed most to the predicted discrepancy. For a laterality conflict, it might point to container label and requisition laterality fields. For count inconsistency, it might point to received con-

Table 5. Recall by Discrepancy Category

Category	Global Transformer (%)	Micro-Adapted Model (%)
Patient-identifier mismatch	88.7	94.5
Container-requisition mismatch	73.9	84.2
Site or laterality conflict	70.5	86.9
Specimen-count inconsistency	76.8	88.1
Unsupported amendment	61.3	74.8
High-severity overall recall	78.4	91.7

Table 6. Service-Specific Performance Gains

Service	Main Improvement Area	Baseline Recall (%)	Adapted Recall (%)
Breast pathology	Laterality and margins	72.1	90.4
Dermatopathology	Body-site compatibility	66.8	82.7
Gastrointestinal	Sequence consistency	79.5	91.2
Gynecologic pathology	Amendment validation	68.3	81.5
General surgical pathology	Calibration stability	74.6	80.2

tainer count and accessioned part sequence. For unsupported amendment, it might point to the correction note and the field changed by that note. Explanation quality was evaluated by comparing these highlighted fields with reviewer-marked evidence fields in the adjudicated set. Explanations were not presented as proof. They were intended to help staff review the case efficiently.

A global transformer baseline used the same text and structured fields but no service-specific adapters. A larger global transformer baseline had more parameters but still no adapters. A rule system baseline used deterministic field checks. A hybrid rule-classifier baseline combined deterministic checks with a classifier. These baselines help separate the value of service adaptation from the value of simply using more data or more rules.

The model was designed for a pre-sign-out safety review point. It can run after accession and gross-room entry, and it can run again after a correction note or amendment is entered. In the simulated workflow, high-priority flags created a hold recommendation, moderate flags created accessioner or supervisor review, and low-priority flags created a passive queue note. The model did not alter the case record [17]. It did not release, cancel, or amend a case. It only recommended review.

4. Experimental Protocol

The primary metric was macro F_1 across the six labels, including no discrepancy. Macro F_1 was used because patient-identifier mismatch and unsupported amendment were less common than no-discrepancy and container-requisition mismatch. Overall accuracy was also reported but treated as secondary. A model can achieve high accuracy by correctly passing routine cases while missing rare safety issues. The main safety metrics were high-severity recall, site or laterality conflict recall, patient-identifier mismatch recall, and false safety holds per 1,000 routine cases.

False safety hold was defined as a model recommendation for supervisor or pathologist review in a case adjudicated as no discrepancy or informational-only discrepancy. Not every false hold is equally harmful, but excessive holds slow laboratory workflow and can reduce trust. The study therefore reports both sensitivity and false hold rate. A useful model should reduce missed discrepancies without creating a large number of unnecessary holds.

Temporal robustness was evaluated on a later period after accession templates changed [18]. One hospital changed how container labels were imported into the laboratory information system. Another changed the amendment comment template. These changes altered text patterns without necessarily altering discrepancy risk. A robust model should not fail simply because a field appears in a slightly different format. The temporal test also included a higher share of gastrointestinal cases due to seasonal procedure volume.

Cross-hospital transfer was evaluated by leaving one hospital out of supervised training and testing on that hospital. The model still used self-supervised pretraining on de-identified text from all hospitals in one variant, and a stricter variant excluded the target hospital entirely from pretraining. The main cross-hospital result uses the stricter variant because it better reflects deployment in a new laboratory network. Calibration thresholds were selected on the source hospitals and then applied to the target hospital. A small calibration-only sensitivity analysis used 200 adjudicated cases

Table 7. Temporal and Cross-Hospital Evaluation

Evaluation Setting	Global Transformer	Micro-Adapted Model	Observation
Internal test macro F_1	0.692	0.801	Strongest overall gain
Temporal test macro F_1	0.624	0.768	Lower degradation observed
Hospital A holdout	0.653	0.742	Better transferability
Hospital B holdout	0.668	0.756	Improved robustness
Hospital C holdout	0.662	0.721	Specialty terminology challenge

Table 8. Ablation Study Results

Ablation Variant	Macro F_1	Primary Effect
Full micro-adapted model	0.801	Best balanced performance
Without adapters	0.736	Loss of service specialization
Without consistency checks	0.752	Reduced relational detection
Without structured fields	0.703	Weak count handling
Without disagreement training	0.775	More false holds
Without review-level head	0.789	More high-severity misses

from the target hospital, but that result is reported separately.

Explanation quality was evaluated on 1,940 adjudicated discrepancy cases with reviewer-marked evidence fields. A field-level explanation was considered correct if the model highlighted at least one field pair that reviewers used to justify the label. For example, in a laterality conflict, highlighting the requisition and container label counted as correct if those were the conflicting fields. Highlighting only the service line did not count. Explanation quality was not used to select final models, but it was important for assessing whether the model pointed reviewers to relevant parts of the record.

An ablation study removed major components one at a time. The no-adapter ablation used the shared encoder only. The no-consistency-check ablation removed deterministic comparison features. The no-structured-field ablation used free text only. The no-disagreement-training ablation treated adjudicated labels as one-hot labels even when reviewers disagreed. The no-review-level head ablation predicted only discrepancy category. The larger-global baseline increased model size without adapters. These experiments tested whether performance came from service-specific adaptation, explicit consistency features, or general capacity.

The statistical analysis used bootstrap resampling by case batch. Case batches were used rather than individual cases because accession workflow errors can cluster by day, clinic, procedure type, or template issue. Confidence intervals were computed with 2,000 bootstrap samples. Differences in high-severity recall, false hold rate, macro F_1, and temporal degradation were reported with 95% intervals. The paper focuses on practical effect sizes rather than only statistical significance.

Manual error review was performed on 620 cases. The sample included missed high-severity discrepancies, false holds, cases corrected by the proposed model relative to the global model, and cases where the global model was correct but the proposed model was wrong. Reviewers categorized errors as insufficient source information, ambiguous submitting terminology, service routing error, laterality abbreviation issue, count sequence issue, amendment documentation issue, or model overreaction to unusual wording. This review helped interpret quantitative results.

The simulated workflow analysis estimated case holds and review workload. Historical discrepancy review logs provided approximate review capacity and hold resolution times. The simulation compared routine practice triggers, the global transformer model, and the micro-adapted model. Outcomes included additional cases reviewed per 1,000 accessions, high-severity discrepancies captured before sign-out, median hold duration for true discrepancies, and false hold volume [19]. The simulation is not a prospective trial. It is a retrospective estimate of workflow consequences.

5. Results

The deterministic rule system had high precision for explicit count mismatches and direct left-right conflicts, but it missed many discrepancies expressed through local wording. Its macro F_1 was 0.548 on the internal test set. Patient-identifier mismatch recall was 84.2%, but site or laterality conflict recall was only 51.6%, and unsupported amendment recall was 39.8%. The rule system created 5.6 false safety holds per 1,000 routine cases. It was conservative in volume but incomplete in detection.

Table 9. Common Error Patterns Identified During Review

Error Pattern	Representative Cause	Typical Consequence
Missing source information	Absent container text	Unresolved mismatch detection
Ambiguous terminology	Clinic-specific shorthand	Incorrect site interpretation
Laterality abbreviation issue	Multi-meaning abbreviations	Side confusion errors
Amendment documentation ambiguity	Policy variation	Unsupported amendment flags
Service routing mismatch	Incorrect adapter selection	False positive review hold
Hybrid cross-service specimen	Mixed terminology usage	Adapter misclassification

Table 10. Workflow Simulation Outcomes

Workflow Metric	Routine Practice	Global Transformer	Micro-Adapted Model
Additional true discrepancies detected	–	10.2	14.6
Additional false holds	–	37.4	5.1
Median hold duration	Longer	Moderate	Shorter
Reviewer field guidance	Minimal	Partial	Strong evidence highlighting
Adaptation to template changes	Weak	Poor	Moderate robustness

The global transformer model improved macro F_1 to 0.692. It captured more container-requisition mismatches and site conflicts, but it produced false holds on unusual service-specific language. The larger global transformer improved macro F_1 only slightly to 0.708 and increased false holds to 13.1 per 1,000 cases. The hybrid rule-classifier model reached macro F_1 of 0.736. The proposed service-specific micro-adapted model achieved macro F_1 of 0.801, with overall accuracy of 90.6%. It produced 7.2 false safety holds per 1,000 routine cases.

High-severity recall improved from 78.4% for the global transformer to 91.7% for the micro-adapted model. Site or laterality conflict recall improved from 70.5% to 86.9%. Specimen-count inconsistency recall improved from 76.8% to 88.1%. Patient-identifier mismatch recall improved from 88.7% to 94.5%. Unsupported amendment recall improved from 61.3% to 74.8%, which remained the weakest category. Container-requisition mismatch recall improved from 73.9% to 84.2%. These improvements were accompanied by fewer false holds than the larger global model, indicating that the adapters did not simply increase sensitivity by flagging more cases.

Performance differed by service. Breast cases showed the largest improvement in laterality and margin-related conflicts. The global model often recognized left-right disagreement but missed conflicts involving margin orientation or lesion-location phrasing. The breast adapter improved conflict recall from 72.1% to 90.4%. Dermatopathology improved mainly in body-site compatibility, where recall increased from 66.8% to 82.7%. Gastrointestinal improved in specimen-count and sequence issues, with count inconsistency recall increasing from 79.5% to 91.2%. Gynecologic cases improved in container-requisition mismatch and unsupported correction notes. General surgical pathology improved less, but still showed better calibration and fewer false holds.

The temporal test exposed the weakness of the global model. After template changes, macro F_1 for the global transformer decreased from 0.692 to 0.624. False holds increased from 11.8 to 18.7 per 1,000 cases. The larger global transformer degraded similarly. The micro-adapted model decreased from 0.801 to 0.768 and false holds increased from 7.2 to 9.6 per 1,000 cases. The model still degraded, especially for unsupported amendments, but the drop was smaller. The amendment template change caused many false holds in models that relied heavily on old wording patterns.

Cross-hospital transfer was more challenging. When Hospital A was held out, the micro-adapted model reached macro F_1 of 0.742. When Hospital B was held out, it reached 0.756. When Hospital C was held out, it reached 0.721. The lower performance on Hospital C reflected more specialty-specific breast and dermatopathology terminology. The global model averaged 0.661 across held-out hospitals. Adding 200 adjudicated calibration cases from the target hospital improved the micro-adapted model's average macro F_1 to 0.773, mostly by adjusting review thresholds and adapter selection. This suggests that small local calibration is useful even when the model transfers reasonably.

The ablation study showed that service adapters

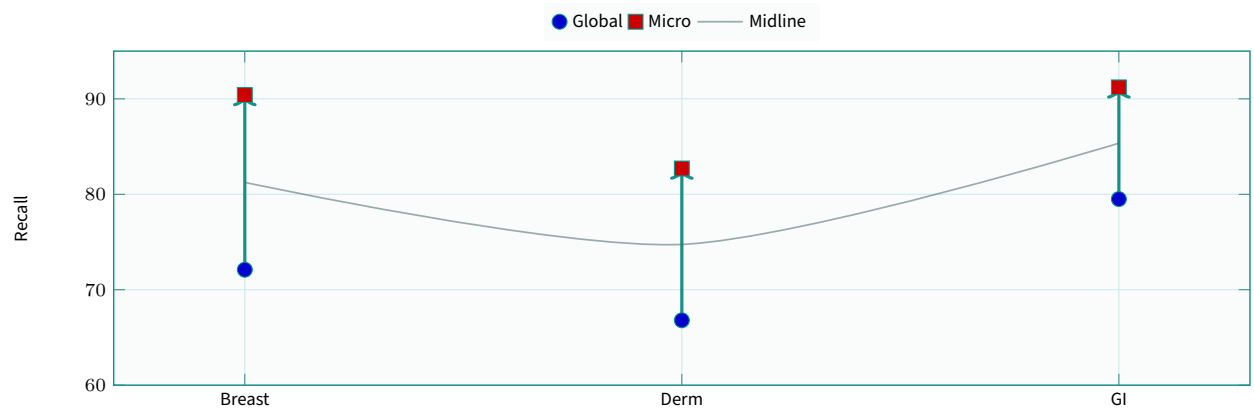

Figure 9. Service gains.

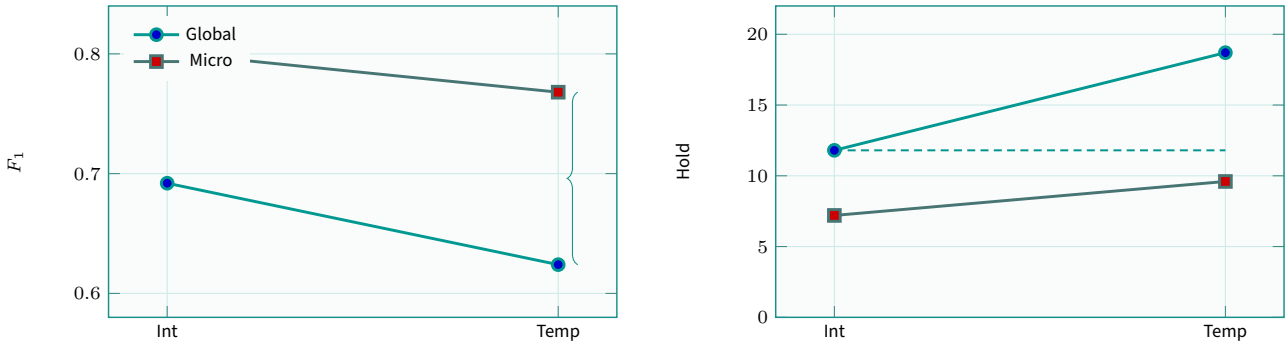

Figure 10. Temporal drift.

were the largest contributor. Removing adapters reduced macro F_1 from 0.801 to 0.736. Removing consistency-check features reduced it to 0.752. Removing structured fields reduced it to 0.703. Removing disagreement-aware training reduced it to 0.775 and increased false holds. Removing the review-level head reduced macro F_1 slightly to 0.789 but increased high-severity misses, suggesting that workflow-level supervision helped the model distinguish minor discrepancy from hold-worthy discrepancy. Increasing global model size did not recover the full adapter benefit.

Explanation quality improved with micro adaptation [20]. The global transformer highlighted reviewer-relevant fields in 68.5% of discrepancy cases. The hybrid rule-classifier reached 73.9%. The micro-adapted model reached 82.6%. Explanation accuracy was highest for specimen-count inconsistency and patient-identifier mismatch. It was lowest for unsupported amendment, where the model sometimes highlighted the correction note but not the field that the correction changed. Reviewers judged explanations as practically useful in 78.3% of true positive high-severity cases. They were less useful in false holds, where the model often highlighted unusual but harmless wording.

Workflow simulation showed that the micro-adapted model would increase review of true discrepancies while limiting unnecessary holds. Compared with routine practice triggers, it identified an estimated 14.6 additional high-severity discrepancies per 10,000 accessions before sign-out. It created 5.1 additional false holds per 10,000 accessions after replacing several existing broad triggers. The global transformer identified 10.2 additional high-severity discrepancies but created 37.4 additional false holds. Median simulated hold duration for true discrepancies decreased because the model highlighted relevant fields, allowing reviewers to locate the issue faster. False holds still consumed staff time, especially in dermatopathology and gastrointestinal cases with unusual but correct site wording.

The model performed best when discrepancies were expressed in multiple fields. For example, a laterality conflict present in both container text and requisition text was easier than a conflict implied only through procedure description. Count inconsistency was easier when structured count and part labels disagreed than when the issue appeared only in a free-text comment. Unsupported amendment was difficult because the model had to judge whether a correction note was supported by surrounding documentation. This category often required local policy knowledge, and even human reviewers disagreed more often.

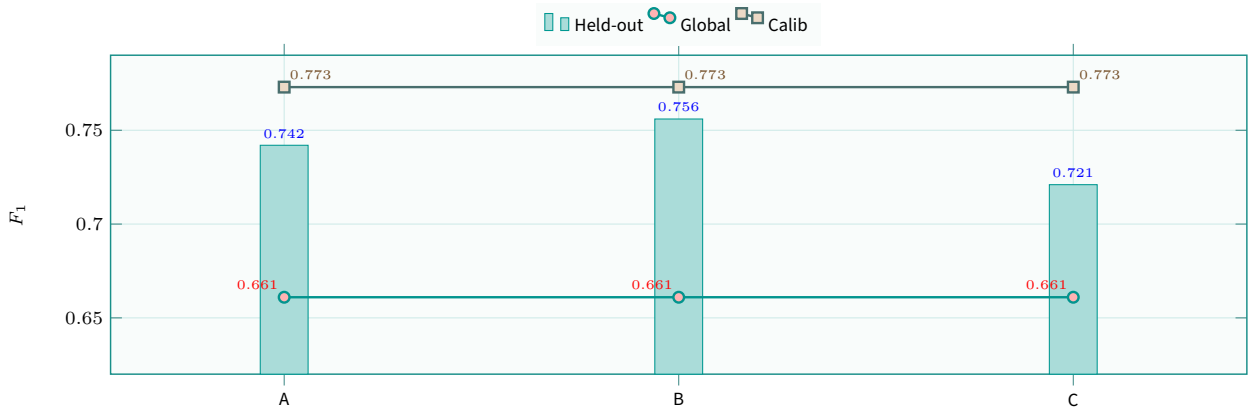

Figure 11. Hospital transfer.

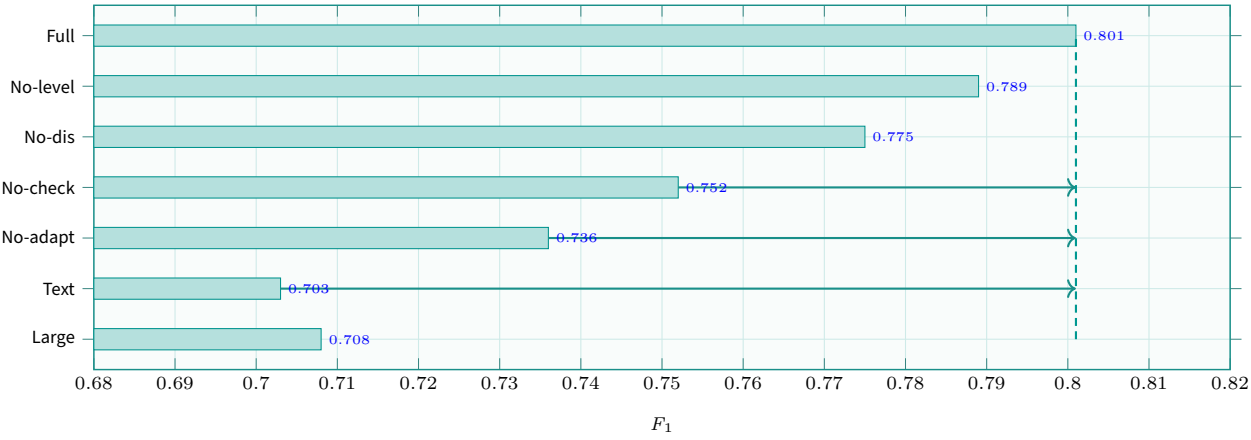

Figure 12. Ablation profile.

6. Error Analysis

Missed high-severity discrepancies fell into several patterns. The first pattern was insufficient source information. Some cases lacked the original container label text or had only a scanned image placeholder. Without the missing field, the model could not compare container and requisition. Human reviewers sometimes inferred the issue from later comments, but the model did not have enough reliable pre-sign-out input. These cases suggest that document availability is a safety dependency. A model cannot recover information that the workflow has not captured.

The second pattern was ambiguous submitting terminology. Some clinics used shorthand body-site descriptions that were locally understood but not clearly mapped to standard anatomic terms. Dermatopathology had many such cases. For example, a site phrase could refer to a nearby region, a lesion nickname, or a procedure-specific location. The adapter improved performance but did not eliminate ambiguity. In some cases, the right response is not better modeling but a better requisition interface that requires structured body-site selection [21].

The third pattern was laterality abbreviation. Certain abbreviations can mean different things depending on service or procedure. A short token may indicate left, lateral, lower, or a named anatomic landmark. The model sometimes interpreted an abbreviation incorrectly, particularly in head and neck and dermatopathology cases. Rule systems also struggled with these abbreviations. Human reviewers used broader context, including procedure description and clinic convention. A future system could incorporate a service-specific abbreviation dictionary maintained by the laboratory.

The fourth pattern was amendment documentation. Unsupported amendment cases often depended on whether a correction note was adequate according to local policy. The model could detect that a correction changed a field, but it could not always determine whether the documentation was sufficient. Some hospitals require explicit source confirmation, while others allow certain accession corrections based on staff verification. This policy variation limited cross-hospital transfer. Local threshold calibration helped, but the category remained challenging.

False holds were also reviewed. Many false holds involved unusual but correct wording. Breast cases with

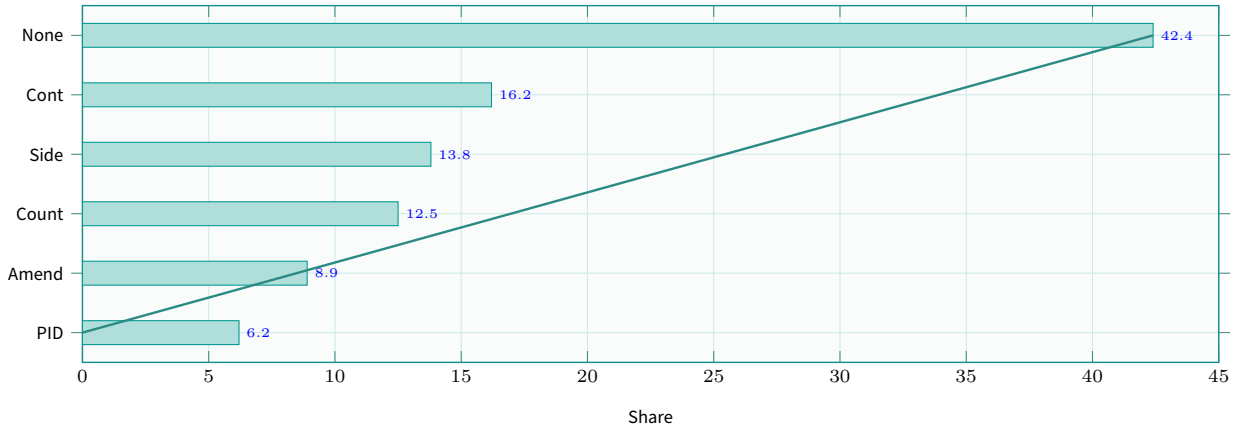

Figure 13. Label distribution.

detailed margin descriptions sometimes looked contradictory because the same side appeared with multiple orientations [22]. Dermatopathology cases sometimes used compound body sites that triggered compatibility checks. Gastrointestinal cases with many jars and repeated biopsy levels sometimes triggered count concerns even when the sequence was correct. These false holds were not random. They occurred where the record was complex and the model's caution was understandable. Still, unnecessary holds can delay reporting, so reducing them matters.

Some false holds came from service routing errors. If a case was assigned to the wrong service adapter, the adapter could misinterpret ordinary terminology. For example, a specimen routed as general surgical pathology rather than gynecologic pathology could lose service-specific interpretation of procedure phrases. The model included a fallback adapter, but service routing remained important [23]. This suggests that adapter selection should be based on both accession service and text-inferred service, with conservative handling when they disagree.

Cases corrected by the micro-adapted model showed the value of local specialization. In breast cases, the model learned that laterality, clip status, margin designation, and lesion location form a tight field group. In gastrointestinal cases, it learned that part sequence and container count interact with site order. In dermatopathology, it learned that body-site phrases can be long and variable but still contain laterality or region cues. These patterns helped the model catch discrepancies that a global representation treated as ordinary text variation.

Cases where the global model was correct and the micro-adapted model was wrong often involved rare cross-service specimens. A specimen routed to one service but described using terminology common in another service confused the adapter. Some large resections also crossed ordinary service boundaries. The global model sometimes handled these because it did not specialize as strongly [24]. This suggests a trade-off: adapters improve common service-specific patterns but can mislead on hybrid cases. The fallback adapter and uncertainty logic reduced but did not remove this issue.

The model's explanations were useful but imperfect. In true positive cases, highlighting conflicting fields helped reviewers confirm the discrepancy. In false holds, explanations sometimes revealed why the model was concerned, even when the concern was not a true discrepancy. Reviewers reported that explanations were most useful when they showed two fields side by side. Single-field explanations were less helpful because discrepancy is relational. A future interface should display the model's field pair and the specific terms that disagree, not merely a risk score.

Temporal errors after template changes were concentrated in amendment notes. The new template inserted standardized phrases that looked similar to discrepancy language. The global model flagged many of these. The micro-adapted model also flagged some, but fewer. After a small set of local calibration examples, false holds decreased. This indicates that model monitoring should be tied to laboratory information system changes. Any template revision, accession interface update, or new clinic submission format can shift the input language.

The cross-hospital results raise a governance issue. A model trained at one laboratory may not be ready for another laboratory without local validation. Specimen terminology, accession policy, amendment documentation, and service routing differ. The proposed model transferred better than a global model, but still benefited from local calibration. A safe implementation would require a silent evaluation period, review of false holds and misses, and adjustment of thresholds before active use. It should also include a way for staff to mark model concerns as helpful, unnecessary, or incorrect.

The study also showed that some discrepancies are

better prevented upstream. Automated review can catch a left-right conflict after submission, but structured requisition design can prevent it. A model can flag missing container count, but barcode scanning and required container reconciliation can reduce the occurrence. A model can detect unsupported amendment language, but policy-aligned correction forms can standardize documentation. The model should therefore be viewed as one layer in a safety system, not the main solution.

7. Conclusion

This paper presented a service-specific micro-adapted model for detecting specimen labeling and accession discrepancies in surgical pathology workflows. The model combined shared text encoding, structured accession fields, deterministic consistency checks, and small service-specific adapters. It improved discrepancy detection compared with rules, a global transformer, a larger global transformer, and a hybrid rule-classifier [25]. The largest gains appeared in site or laterality conflict, specimen-count inconsistency, and service-specific container-requisition mismatch.

The results support a practical claim: specimen safety review benefits from local workflow knowledge. A single global model can learn common discrepancy signals, but pathology services use different language and conventions. Small adapters captured these differences without requiring a separate full model for every service. The approach improved recall while reducing false holds relative to larger global modeling. This matters because laboratories cannot adopt systems that flag too many routine cases.

The study also identified persistent limitations. Unsupported free-text amendment remained difficult because it depends on local policy and documentation sufficiency [26]. Highly ambiguous submitting terminology remained difficult because the source text sometimes lacked enough information. Cross-service hybrid cases could confuse specialized adapters. Template changes caused temporal drift, especially in amendment notes. These limitations show that model performance depends on both algorithm design and laboratory information quality [27].

A discrepancy detection model should not operate as an autonomous authority. It should recommend review, highlight relevant fields, and support existing quality processes. Accession staff, grossing personnel, supervisors, and pathologists remain responsible for resolving discrepancies. The model's role is to catch cases that deserve attention earlier and to reduce the chance that subtle conflicts travel downstream unnoticed. Additional studies should examine smaller laboratories, pediatric pathology, cytology, and molecular specimen workflows. The present findings indicate that service-specific micro adaptation can improve surgical pathology accession safety review, but safe use requires ongoing monitoring, policy alignment, and human oversight.

References

[1] E. Rometsch, P.-O. Thuresson, N. Hurst, E. Eule, I. Bachmeier, B. Blanchard, J.-A. Sahel, and I. Audo, "Patient experience in retinitis pigmentosa and choroideremia- a concept elicitation study in 17 patients based on qualitative interviews.," Orphanet journal of rare diseases, vol. 20, pp. 418–418, 8 2025.

[2] K. Nwanyanwu, J. Andoh, E. Chen, Y. Xu, A. Granados, Y. Deng, M. Desai, and M. Nunez-Smith, "Social determinants associated with diabetic retinopathy awareness: National health and nutrition examination survey (2005-2008).," American journal of ophthalmology, vol. 278, pp. 389–401, 6 2025.

[3] Y. Yang, P. Wang, C. Yu, J. Zhu, and J. Sheng, "Application of artificial intelligence medical imaging aided diagnosis system in the diagnosis of pulmonary nodules.," BMC medical informatics and decision making, vol. 25, pp. 185–185, 5 2025.

[4] B. Sadanandan and V. Behzadan, "Psf-med: Measuring and explaining paraphrase sensitivity in medical vision language models," arXiv preprint arXiv:2602.21428, 2026.

[5] J. Harewood, M. Contreras, K. Huang, S. Johnson, and J. Wang, "Access to myopia care in the united states-a narrative review.," Investigative ophthalmology & visual science, vol. 66, pp. 5–5, 6 2025.

[6] S. A. Merchant, N. Merchant, S. L. Varghese, and M. J. S. Shaikh, "Large language models and large concept models in radiology: Present challenges, future directions, and critical perspectives.," World journal of radiology, vol. 17, pp. 114754–114754, 11 2025.

[7] P. Scheffler, N. Neidert, J. Straehle, D. Erny, M. Prinz, U. Hubbe, R. Rölz, D. H. Heiland, J. Beck, and A. E. Rahal, "Artificial intelligence-based analysis and diagnosis of intradural extramedullary spinal tumors by stimulated raman histology.," Neuro-oncology advances, vol. 7, pp. vdaf211–vdaf211, 1 2025.

[8] X. Qian, H. Wang, X. Wang, T. Wu, H. Han, X. Bu, and F. Teng, "Patients' knowledge, attitude, and practice toward stroke rehabilitation: a web-based cross-sectional study.," Frontiers in public health, vol. 13, pp. 1593802–1593802, 8 2025.

[9] F. Rafsani, D. Sheth, Y. Che, J. Shah, M. M. R. Siddiquee, C. D. Chong, S. Nikolova, K. Ross, G. Dumkrieger, B. Li, T. Wu, and T. J. Schwedt, "Leveraging multi-modal foundation model image encoders to enhance brain mri-based headache classification.," Scientific reports, vol. 15, pp. 33256–33256, 9 2025.

[10] Y. Luo, H. Hooshangnejad, X. Feng, G. Huang, X. Chen, R. Zhang, Q. Chen, W. Ngwa, and K. Ding, "A language vision model approach for automated tumor contouring in radiation oncology.," Bioengineering (Basel, Switzerland), vol. 12, pp. 835–835, 7 2025.

[11] S. D. Landes, B. K. Swenor, J. P. Hall, A. J. Forber-Pratt, N. Vaitsiakhovich, K. Caldwell, M. Kakara, D. Lefkowitz, A. Myers, S. J. Popkin, N. S. Reed, E. F. Rothman, and M. Salinger, "The disability mismatch: the case for a comprehensive disability status measure.," Health affairs scholar, vol. 3, pp. qxaf137–qxaf137, 7 2025.

[12] M.-H. Van, A. N. Carey, and X. Wu, "Influence-based approaches for tumor classification in noisy brain mri with deep learning and vision-language models," International Journal of Data Science and Analytics, vol. 20, pp. 6475–6487, 6 2025.

[13] R. Gifford, A. Reid, S. R. Jhawar, K. VanKoevering, and S. Krening, "Convolutional neural network for classification of oropharynx cancer with video nasopharyngolaryngoscopy.," Journal of otolaryngology - head & neck surgery = Le Journal d'oto-rhino-laryngologie et de chirurgie cervico-faciale, vol. 54, pp. 19160216251326590–19160216251326590, 3 2025.

[14] D. T. Leffa, Y. Zhang, Y. Wang, E. G. Teverovsky, E. Jacobsen, B. Bellaver, P. C. L. Ferreira, K.-H. Fan, M. I. Kamboh, T. K. Karikari, C.-C. H. Chang, B. E. Snitz, B. S. G. Molina, T. A. Pascoal, and M. Ganguli, "Association between the genetic risk for attention-deficit/hyperactivity disorder and cognitive function in older age: The myhat population-based study.," The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry, 10 2025.

[15] S. A. Sebton, C. M. Bridger, P. A. Arias, C. K. Okeibunor, S. R. Spencer, R. E. Willis, S. E. Taylor, and B. J. Alsip, "Comprehensive trauma-informed organizational change at an academic medical center in south texas.," BMC health services research, vol. 25, pp. 1433–1433, 11 2025.

[16] K. M. Chen, K. W. Chen, V. Diaconita, S. Chang, and L. H. Suh, "Ophthoacr (ophthalmology automated chart review): An ai-powered tool for complete automation of ophthalmology chart reviews and cohort data analysis.," Translational vision science & technology, vol. 14, pp. 8–8, 10 2025.

[17] A. Sheikhy, F. D. Firouzabadi, N. Lay, N. Jarrah, P. Y. Anari, and A. Malayeri, "State of the art review of ai in renal imaging.," Abdominal radiology (New York), vol. 50, pp. 5305–5323, 4 2025.

[18] C. L. Ortz, M. S. Duncan, O. Leshi, W. B. Burrows, and B. L. Smalls, "Influence of perceived health provider communication, diabetes duration and age at diagnosis with confidence in diabetes self-care.," BMJ open diabetes research & care, vol. 13, pp. e004645–e004645, 4 2025.

[19] Z.-A. Li, Y. Gao, K. Ji, K.-Y. Zhang, Q. Zhang, J. Wang, L. Han, X.-Y. Zhai, W.-P. Wang, W.-L. Liu, H.-Y. Wei, R.-J. Qin, Y.-D. Li, H.-L. Zhao, R.-F. Yan, and H.-K. Cui, "Ai-driven diagnosis of vulnerable intracranial atherosclerotic plaques using large language models and vision transformers: a multi-center study.," European radiology, vol. 36, pp. 2700–2711, 10 2025.

[20] S. Schmidgall, J. D. Opfermann, J. W. Kim, and A. Krieger, "Will your next surgeon be a robot? autonomy and ai in robotic surgery.," Science robotics, vol. 10, pp. eadt0187–eadt0187, 7 2025.

[21] X. Hu, M. Varkanitsa, E. Kropp, M. Betke, P. Ishwar, and S. Kiran, "Aphasia severity prediction using a multimodal machine learning approach.," NeuroImage, vol. 317, pp. 121300–121300, 6 2025.

[22] S. J. Price, J. G. Hughes, S. Jain, C. Kelly, I. Sederias, F. M. Cozzi, J. Fares, Y. Li, J. C. Kennedy, R. Mayrand, Q. H. W. Wong, Y. Wan, and C. Li, "Precision surgery for glioblastomas.," Journal of personalized medicine, vol. 15, pp. 96–96, 2 2025.

[23] D. Dai, Y. Zhang, Q. Yang, L. Xu, X. Shen, S. Xia, and G. Wang, "Pathologyvlm: a large vision-language model for pathology image understanding," Artificial Intelligence Review, vol. 58, 3 2025.

[24] M. Omar, R. Agbareia, M. E. Naffaa, A. Watad, B. S. Glicksberg, G. N. Nadkarni, and E. Klang, "Applications of artificial intelligence in vasculitides: A systematic review.," ACR open rheumatology, vol. 7, pp. e70016–e70016, 3 2025.

[25] R. Hu, Y. Yang, S. Liu, Z. Li, J. Liu, X. Ding, H. Sun, and L. Ren, "Large language model driven transferable key information extraction mechanism for nonstandardized tables.," Scientific reports, vol. 15, pp. 29802–29802, 8 2025.

[26] H. Lai, Z. Jiang, Q. Yao, R. Wang, Z. He, X. Tao, W. Lv, W. Wei, and S. K. Zhou, "Bridged semantic alignment for zero-shot 3d medical image diagnosis.," IEEE journal of biomedical and health informatics, vol. PP, pp. 1–14, 11 2025.

[27] F. Zhu, Z. Liu, J. Chang, Y. Qin, and L. Wang, "Deep learning for scene understanding in mitochondrial dysregulation and blood cancer diagnosis.," Frontiers in oncology, vol. 15, pp. 1609851–1609851, 10 2025.