HQAI-TB: A MULTIMODAL HYBRID QUANTUM–CLASSICAL ARTIFICIAL INTELLIGENCE FRAMEWORK FOR TUBERCULOSIS DIAGNOSIS FROM CHEST RADIOGRAPHS AND CLINICAL RECORDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tuberculosis (TB) is one of the most dangerous and lethal infectious external agents. In highly prone countries, the diagnosis of TB is hindered by three interconnected bottlenecks: low-quality X-rays, lack of expert readers, and clinical information locked away in bilingual or multilingual paper records and not in digital versions. There have been many AI solutions to address these bottlenecks and challenges in silos: chest X-ray classifiers that incorporate clinical information from radiology reports as covariates, document digitization pipelines that largely ignore imaging findings, and hybrid quantum-classical related learning systems that have not yet been tested on TB diagnosis with integrated clinical records. In this paper, we propose a multi-modal framework with four clinically deployable modules, QASP-Net, which conditions the images and enhances on the estimated quality descriptors, called HQAI-TB. VQC-ResNet is a combination, and it works. In VQC-ResNet works and combines. In this paper, we propose HQAI-TB, which is a multi-modal framework consisting of four clinically deployable modules, QASP-Net, which conditions the images and improves the estimated quality descriptors. VQC-ResNet: a hybrid which operates on a residual convolutional encoder and a parameterized variational quantum circuit. TBScript-OCR that converts bilingual or multilingual clinical documents into digital data and finally TB-ICD-BERT that extracts entities and makes it easy for ICD diagnostic coding. A diagnostic probability and a structured clinical summary are output by an attention-weighted fusion layer. We describe the formalism of all the components and analyze the pipeline cost. Final clinical validation will determine ultimate utility to patient care, but we describe a pre-specified validation protocol: corpora are fixed (800 public domain frontal views, 138 Montgomery, 662 Shenzhen), reference standards hold, and five ablation configurations are compared with standard statistical tests. We give one initial quantitative result: TBScript-OCR, initialized from an off-the-shelf transformer-based recognizer, and fine-tuned on a 500-form institutional pilot corpus, recognized approximately 82% of characters (18 percent error rate) on the printed bilingual set at one site. We report this as a preliminary finding and note that an 18 percent character error would be unacceptable for downstream diagnostic coding without the ICD ontology filter, we propose here.

Keywords— tuberculosis, hybrid quantum-classical learning, variational quantum circuits, multimodal medical AI, chest radiography, optical character recognition, clinical natural language processing.

1 INTRODUCTION

Tuberculosis (TB), caused by Mycobacterium tuberculosis, remains one of the highest mortality risks from a single infectious agent. The WHO estimates 10.8 million incident cases and 1.25 million deaths in 2023, with an overwhelming majority in low- and middle-income countries (World Health Organization, 2024); Indian authorities report treating one in five global cases. Culture remains

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the diagnostic gold standard but takes weeks to yield results; nucleic acid amplification tests such as Xpert MTB/RIF are rapid and informative on rifampicin resistance but require infrastructural investments unavailable at the periphery (Boehme et al., 2011). Hence, chest radiography is the mainstay of TB triage, where the limiting factor is the availability of readers.

Learned representations have largely replaced hand-designed descriptors (LeCun et al., 2015), and residual encoders enable very deep architectures (He et al., 2016). For TB diagnosis, convolutional networks achieve discriminatory performance close to, if not higher than, expert readers on the Montgomery County and Shenzhen corpora (Lakhani & Sundaram, 2017)–(Jaeger et al., 2014), with multi-site evaluations showing that operating points may not transfer across sites. (Qin et al., 2019). Such studies seldom acknowledge that a single-image classifier lacks the benefit of laboratory results, treatment history, or concurrent clinical notes.

A second literature addresses the diagnostic record itself: in many high-incidence settings, patient records are paper-based and bilingual, requiring manual entry for storage or analysis. Transformer-based optical character recognition (OCR) has improved the recognition of degraded print (Li et al., 2023), and clinical language models can extract relevant entities or support diagnostic coding (Devlin et al., 2019)–(Alsentzer et al., 2019), but these methods are typically evaluated on standalone document tasks rather than in conjunction with imaging.

A separate literature proposes quantum machine learning models to encode high-dimensional interactions with relatively few parameters (Biamonte et al., 2017); given current hardware limitations, realistic proposals involve variational circuits, shallow quantum circuits whose parameters are optimized classically (Cerezo et al., 2021). The value of such approaches for medical data is still to be established. Multimodal radiology imaging systems typically process structured fields alongside images, which fails to address low-resource settings that lack digitized records.

Research gap. Four factors distinguish this work from the existing literature. First, no reported TB diagnosis system closes the loop from paper records through language understanding to a joint decision with radiographic findings. Second, image enhancement uses a fixed recipe regardless of estimated quality, so a correctly exposed and a motion-degraded image are enhanced identically, with potential losses in diagnostic yield. Third, hybrid quantum-classical learning has been proposed for medical data, but never specified, costed, and evaluated in a multimodal TB setting. Fourth, a recurring challenge across all four areas is the lack of explicit hypotheses of practical relevance: the literature typically does not state, up front, what outcome would constitute evidence in favor of the reported advance.

Objective. This paper describes HQAI-TB, a multimodal hybrid quantum-classical framework that processes radiographic and documentary evidence with separate modules and combines their outputs in a fusion layer that uses attention mechanisms, and specifies the evaluation strategy up front. This work makes no claims regarding diagnostic performance, quantum advantage, or ultimate clinical utility.

Contributions. **1)** HQAI-TB, a multimodal architecture that brings chest radiographic and documentary evidence into a joint diagnostic pipeline. **2)** QASP-Net, a preprocessing formulation that selects the enhancement operator based on estimated no-reference image-quality descriptors, rather than using a fixed recipe. **3)** VQC-ResNet, a hybrid quantum-classical predictor: a residual convolutional encoder followed by a parameterized variational quantum circuit, with explicitly described encoding, ansatz, measurements, and gradient estimation. **4)** TBScript-OCR and TB-ICD-BERT, a document-digitization and language-processing pipeline that processes bilingual scanned records and produces structured entities and diagnostic codes. **5)** An attention-weighted fusion layer that specifies recognition probabilities, diagnostic codes, and structured clinical summary in case of missing modalities. **6)** A pre-specified evaluation and ablation protocol, identifying reference corpora, statistical tests, and conditions under which each component can be considered dispensable.

2 Related Work

AI-based TB diagnosis. Automated radiograph interpretation relies on large thoracic x-ray corpora, which also highlight the limits of report-derived labels (Wang et al., 2017); residual and

Table 1: Capability coverage of representative prior work.

Study / approach	Img	OCR	NLP	MM	QC	Principal limitation for programmatic TB use
Qin et al. (2019)	✓	–	–	–	–	Triage only; operating point not site-transferable
Li et al. (2023) (TrOCR)	–	✓	–	–	–	General domain; not clinical or bilingual
Mullenbach et al. (2018)	–	–	✓	–	–	Coding from typed notes; no imaging
Huang et al. (2020)	✓	–	✓	✓	–	Requires structured EHR; not TB-specific
Havlíček et al. (2019)	–	–	–	–	✓	Non-medical benchmark tasks
HQAI-TB (this work)	✓	✓	✓	✓	✓	Specification and protocol; not yet validated

Img = chest imaging; OCR = document digitisation; NLP = clinical language understanding; MM = multimodal fusion; QC = quantum or hybrid quantum-classical component. For HQAI-TB, ✓ denotes coverage in the proposed design, not validated performance.

densely connected encoders are now the norm (He et al., 2016; Huang et al., 2017). For TB, convolutional networks achieve close to expert performance on the Montgomery County and Shenzhen collections (Lakhani & Sundaram, 2017; Pasa et al., 2019; Jaeger et al., 2014), with lighter networks deployable on modest hardware (Pasa et al., 2019). Performance depends strongly on the training corpus: one multi-center evaluation found operating points non-transferable across sites (Qin et al., 2019), and similar limitations likely apply to mixed-scanner deployments (Rahman et al., 2020).

OCR and clinical NLP. Segmentation-based OCR engines are effective on clean print but perform poorly on degraded handwriting, skew, or contrast (Smith, 2007); recent recognizers exploit large contexts but still require grid-like regular structures to handle skew (Li et al., 2023). Biomedical and clinical language models are the default clinical text representation (Devlin et al., 2019; Lee et al., 2020; Alsentzer et al., 2019), typically evaluated on intensive care unit notes from public databases (Johnson et al., 2016). Label attention architectures identify diagnostic labels while highlighting supporting evidence (Mullenbach et al., 2018), but again assume the input text is readily available in digital form—a questionable premise in resource-limited settings.

Multimodal healthcare AI. The literature suggests that leveraging auxiliary context to interpret imaging findings is the primary opportunity in multimodal clinical AI (Acosta et al., 2022; Kline et al., 2022). Combining chest radiography with structured variables discriminates better than either modality alone (Huang et al., 2020), and attention lets a model weight streams rather than concatenate them (Vaswani et al., 2017). However, fusion typically occurs on structured fields extracted from free text, eliding document digitization, and most works do not comment on performance when a modality is absent—the typical operating condition in low-resource settings.

Quantum machine learning. Quantum methods encode information in a Hilbert space whose dimension is exponential in the number of qubits (Biamonte et al., 2017). Noisy intermediate-scale quantum (NISQ) devices are constrained to shallow circuits, so proposals typically use variational circuits trained by classical optimization (Cerezo et al., 2021; Benedetti et al., 2019), differentiable by parameter-shift rules (Mitarai et al., 2018) and embeddable in classical automatic differentiation frameworks (Bergholm et al., 2018). Two results limit immediate practical impact: gradients of randomly initialized circuits vanish exponentially with depth, constraining depth (McClean et al., 2018), and the qubit encoding determines which functions can be represented (Schuld et al., 2021). Thus, realistic algorithms target vision tasks with simplified circuits (Havlíček et al., 2019).

Research gap. Table 1 cross-references prior work against five capabilities necessary for programmable TB pipelines; no single work covers all five. The goal is not to incrementally improve the state of the art but to define a system with components chosen for low-resource deployment.

3 Proposed HQAI-TB Framework

3.1 Overall Architecture

HQAI-TB processes two evidence streams through separate encoders, pools their representations, and classifies TB with a hybrid quantum-classical head (Fig. 1). The imaging stream is conditioned on the estimated quality and encoded with a residual network; the documentary stream extracts text

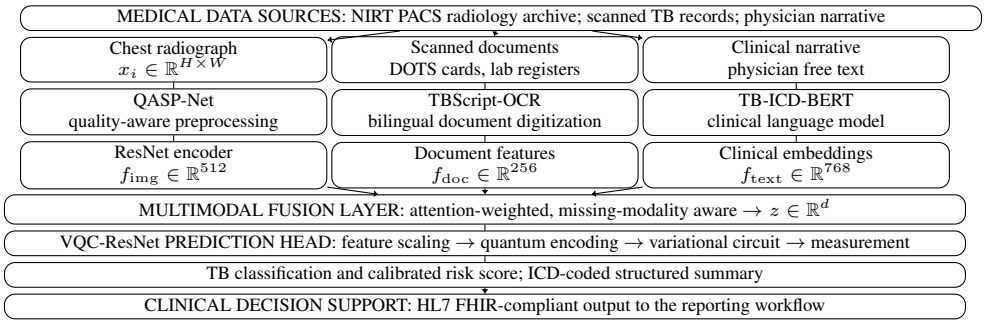

Figure 1: Overall architecture of the HQAI-TB multimodal hybrid quantum–classical AI framework.

from the scanned forms and encodes it with a clinical language model. Fusion operates on the representations rather than the raw evidence, which enables independent training of each encoder and allows dropping a stream at inference.

Formally, let x_i denote a chest radiograph, let $D = \{d_1, \ldots, d_m\}$ denote the scanned clinical documents for the same patient episode, and let $y \in \{0, 1\}$ be the TB label given by the microbiological reference standard. The framework estimates $\hat{y} = P(y = 1 \mid x_i, D)$ as the composition of four learned maps and one fusion operator. Three design notes are in order: preprocessing is a function of the image and its estimated quality, not the image alone; the quantum component only enters the prediction head, making the framework quantum-backend agnostic; and fusion weights are conditioned on each modality's presence, so an absent stream receives zero attention rather than a fixed vector.

3.2 QASP-Net: Quality-Aware Adaptive Image Preprocessing

Radiographs from mixed acquisition devices vary widely in noise, exposure, contrast, and motion artifacts, and no single enhancement chain is optimal across all. Contrast-limited adaptive histogram equalization (Zuiderveld, 1994) with a clip level optimized for a properly exposed film will amplify quantum noise in an underexposed one, while aggressive denoising will destroy the micronodular texture typical of early TB. QASP-Net thus conditions enhancement on estimated image quality.

A no-reference quality estimator $Q(\cdot)$, built on natural-scene-statistic descriptors from the BRISQUE (Mittal et al., 2012) and NIQE (Mittal et al., 2013) families plus a variance-of-Laplacian sharpness term, produces a compact quality vector

$$q = Q(x_i) = [q_n, q_c, q_b, q_l]^T \in \mathbb{R}^4, \tag{1}$$

whose components summarize estimated noise level, contrast, blur and illumination uniformity. Instead of a discrete decision on q, QASP-Net uses it to weight K differentiable enhancement operators $E_1, \ldots, E_K$: contrast-limited adaptive histogram equalization, bilateral filtering, non-local means denoising, unsharp masking and the identity. The enhanced image is the convex combination

$$x' = f(x_i, q) = \sum_k w_k(q)\, E_k(x_i), \tag{2}$$

with gating weights produced by a small fully connected network and constrained to the simplex,

$$w(q) = \mathrm{softmax}(W_g q + b_g), \qquad \sum_k w_k(q) = 1, \tag{3}$$

so the operator mixture varies continuously with quality, and the identity branch lets a high-quality image pass essentially untouched. Intensity is then standardized to remove inter-scanner offset,

$$x'' = \frac{x' - \mu(x')}{\sigma(x')}. \tag{4}$$

A purely task-driven objective would let the gate sacrifice diagnostically relevant structure for classification convenience, so training adds an explicit structure-preservation penalty on the gradient field,

$$\mathcal{L}_{\text{QASP}} = \mathcal{L}_{\text{task}} + \lambda \|\nabla x'' - \nabla x_i\|_1, \tag{5}$$

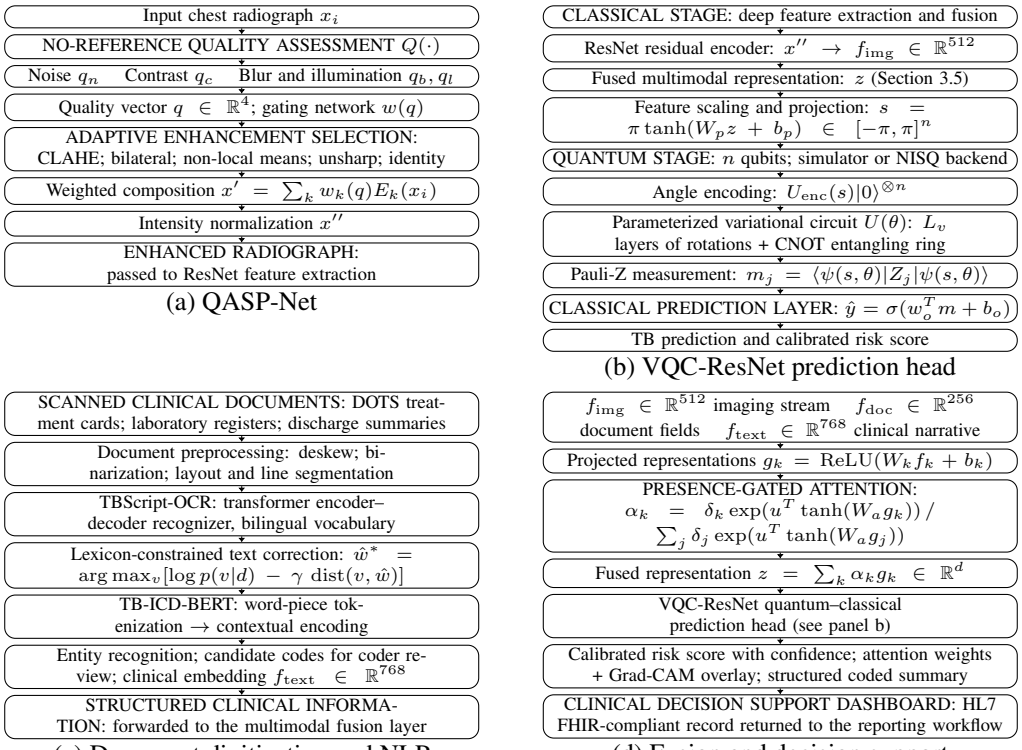

Figure 2: (a) Quality-aware adaptive preprocessing in QASP-Net: the gating network picks a continuous mixture of enhancement operators, conditioned on the estimated quality vector. (b) The hybrid quantum-classical VQC-ResNet prediction head: gradients pass through the quantum stage by the parameter-shift rule, so the classical and quantum parameters are optimized together. (c) Clinical document digitization and NLP pipeline; coding outputs are review candidates, not final assignments. (d) Multimodal feature fusion and decision-support pipeline: the presence indicator δ_k drops an absent modality from the attention normalization instead of substituting a zero vector.

where λ balances enhancement freedom against fidelity to the original anatomy. The gate parameters W_g, b_g and λ are learned jointly with the encoder, and no manual quality annotation is needed, which matters because public TB radiograph collections carry none. Fig. 2(a) shows the pipeline.

3.3 VQC-ResNet: Hybrid Quantum-Classical Prediction

The prediction head pairs a classical encoder with a quantum classifier. The encoder is a residual convolutional network—identity shortcuts are what make deep encoders trainable without degradation (He et al., 2016)—and each block computes

$$h_{l+1} = \text{ReLU}(h_l + F(h_l; W_l)), \tag{6}$$

where F is the residual branch and W_l its parameters. Global average pooling over the final block gives the image representation

$$f_{\text{img}} = \text{GAP}(h_L) \in \mathbb{R}^{512}. \tag{7}$$

The quantum classifier acts not on f_{img} directly but on the fused representation z of Section 3.5, by which point imaging and clinical evidence have already been combined. A circuit of n qubits admits only n rotation angles per encoding layer, so z is first reduced and bounded by a learned linear projection followed by a squashing nonlinearity,

$$s = \pi\ \tanh(W_p z + b_p) \in [-\pi, \pi]^n. \tag{8}$$

The scaling is a modelling decision, not a formality: how classical data enters a variational circuit determines what functions it can express, and unbounded angles alias onto the same rotations

(Schuld et al., 2021). Each component s_j is then encoded by angle embedding on qubit j, optionally repeated over L_e layers to raise the degree of the induced Fourier representation,

$$|\psi_x\rangle = U_{\text{enc}}(s)|0\rangle^{\otimes n} = \prod_j R_y(s_j)|0\rangle^{\otimes n}. \tag{9}$$

A hardware-efficient variational ansatz $U(\theta)$ of depth L_v follows. Each layer combines single-qubit rotations with a ring of CNOT entanglers, giving the parameterized state

$$|\psi(s,\theta)\rangle = U(\theta)U_{\text{enc}}(s)|0\rangle^{\otimes n}. \tag{10}$$

Depth is kept shallow on purpose. Randomly initialized deep circuits suffer exponentially vanishing gradients, so an ansatz cannot be made arbitrarily expressive without becoming untrainable (McClean et al., 2018). Readout takes the expectation of the Pauli-Z observable on each qubit,

$$m_j = \langle\psi(s,\theta)|Z_j|\psi(s,\theta)\rangle \in [-1,1], \tag{11}$$

and the measurement vector $m = [m_1,\dots,m_n]^T$ feeds a classical output layer, giving the prediction

$$\hat{y} = \sigma(w_o^T m + b_o), \tag{12}$$

with σ the logistic function. Training minimizes binary cross-entropy over the joint parameter set $\{W_l, W_p, \theta, w_o\}$. Circuit-parameter gradients come analytically from the parameter-shift rule,

$$\frac{\partial\langle Z_j\rangle}{\partial\theta_k} = \frac{1}{2}\left[\langle Z_j\rangle_{\theta_k+\pi/2} - \langle Z_j\rangle_{\theta_k-\pi/2}\right], \tag{13}$$

which costs two extra circuit evaluations per parameter and composes with classical backpropagation, so the hybrid model trains end to end inside a single automatic-differentiation graph (Mitarai et al., 2018; Bergholm et al., 2018). On noisy hardware, the exact rule would give way to simultaneous perturbation stochastic approximation (Spall, 1992), which estimates the full gradient from two evaluations however many parameters there are. Fig. 2(b) shows the head. We claim no quantum advantage: whether the circuit beats a classical head of matched parameter count is a question the protocol of Section 4 is built to answer, and the literature urges caution (Huang et al., 2021).

3.4 Clinical Document Intelligence: TBScript-OCR and TB-ICD-BERT

TBScript-OCR is an encoder-decoder recognizer in the TrOCR family (Li et al., 2023), adapted along three axes. Geometric normalization corrects skew and segments lines on scanned forms whose layout is tabular rather than free-flowing. The tokenizer and decoder vocabulary are extended to cover code-switched Tamil-English text, which engines that segment on Latin script alone handle poorly (Smith, 2007). And a domain lexicon of TB drug names, regimen abbreviations, laboratory assay identifiers and NTEP form field labels drives a constrained post-correction step: for a decoded hypothesis $\hat{w}$ the emitted string is

$$\hat{w}^* = \arg\max_{v\in V}\left[\log p(v \mid d) - \gamma\ \text{dist}(v,\hat{w})\right], \tag{14}$$

where V is the union of the general vocabulary and the domain lexicon, distance is normalized edit distance, and γ controls how far the correction may drift from the visual evidence. The constraint matters clinically: a one-character error in a drug name or a sputum grade changes the meaning outright, whereas the same error in free narrative usually does not.

TB-ICD-BERT consumes the recovered text. It is a transformer encoder (Vaswani et al., 2017), initialized from biomedical and clinical pre-training (Lee et al., 2020; Alsentzer et al., 2019) and then adapted on TB-specific institutional text. Word-piece tokenization is followed by contextual encoding, and the pooled representation of the classification token gives the clinical embedding

$$f_{\text{text}} = \text{BERT}_{[\text{CLS}]}(\hat{w}_1,\dots,\hat{w}_T) \in \mathbb{R}^{768}. \tag{15}$$

Two task heads sit on top. A token-level head recognizes clinical entities—symptoms, anatomical sites, drug regimens, comorbidities, laboratory findings. A multi-label head over the tuberculosis chapter of the diagnostic classification produces candidate codes with per-label attention (Mullenbach et al., 2018), so every proposed code carries the text span that supports it. The system produces suggestions for a certified coder to review; it is not a validated coding instrument, and no equivalence to human coding is claimed. While the design targets eventual migration to ICD-11, the prototype runs against the ICD-10 tuberculosis chapter, because that is the classification for which reference-coded institutional data exist. Fig. 2(c) shows the stream.

3.5 Multimodal Feature Fusion

Fusion has to combine three representations that differ in dimension, provenance and reliability: f_{img} from the imaging stream, f_{doc} summarizing document-derived structured fields, and f_{text} from the clinical language model. Each is first projected to a common space of dimension d,

$$g_k = \text{ReLU}(W_k f_k + b_k) \in \mathbb{R}^d, \quad k \in \{\text{img}, \text{doc}, \text{text}\}. \tag{16}$$

A fixed linear combination $z = \alpha g_{\text{img}} + \beta g_{\text{text}} + \gamma g_{\text{doc}}$ with $\alpha+\beta+\gamma = 1$ is kept as a baseline, but it weights every patient the same way. How informative each stream is varies case by case—a radiograph may be unreadable, or a treatment card may be the only surviving record of prior therapy—so the weights are instead produced per sample by an attention mechanism gated on modality presence,

$$\alpha_k = \frac{\delta_k \exp\big(u^T \tanh(W_a g_k)\big)}{\sum_j \delta_j \exp(u^T \tanh(W_a g_j))}, \tag{17}$$

where $\delta_k \in \{0, 1\}$ indicates whether modality k is present for this episode, and u, W_a are learned. The fused representation is

$$z = \sum_k \alpha_k g_k \in \mathbb{R}^d. \tag{18}$$

The presence gate is what makes the missing-modality case well defined. Setting $\delta_k = 0$ removes modality k from the normalization rather than contributing a zero vector, so attention mass redistributes over the modalities that are present and the fused vector keeps its scale; zero-filling an absent stream would instead hand the classifier a systematically shrunken input. Modality dropout during training stops the encoders from co-adapting to always having all three. The fused vector z feeds (8) and hence the quantum head, and the weights α_k are kept as an interpretability signal showing which evidence dominated a decision—complemented, for imaging, by gradient-weighted class activation maps (Selvaraju et al., 2017). Fig. 2(d) shows the fusion and decision-support path; Algorithm 1 (Appendix A) summarizes inference, and Appendix B analyzes computational cost.

4 Evaluation Design and Preliminary Measurements

HQAI-TB as specified in Section 3 has not been trained end to end, evaluated on patient data, or clinically validated. What exists is a set of component prototypes and a pre-specified protocol.

4.1 Datasets and Reference Standards

We use only public collections with documented provenance, plus one institutional corpus (Table 3, Appendix C). The Montgomery County and Shenzhen sets (Jaeger et al., 2014) are the standard public TB benchmarks (138 and 662 frontal radiographs with radiologist labels); being small, they serve as evaluation rather than pre-training data. ChestX-ray14 (Wang et al., 2017) (112,120 images from 30,805 patients) supplies thoracic pre-training weights, though its labels are report-derived. Rahman et al. (2020) adds a further public TB cohort, public de-identified critical-care notes (Johnson et al., 2016) support language-model adaptation, and the tuberculosis chapter of the classification supplies the code label space. The public imaging sets and the institutional form corpus serve separate components and are not paired at the patient level; end-to-end multimodal validation requires paired radiograph-document data and remains future work.

Institutional data come in two kinds that must not be conflated: a collected pilot corpus of 500 printed and handwritten TB forms, which underlies the one measurement in Section 4.4, and a specified but uncollected larger corpus (5,000 annotated radiographs and 10,000 clinical documents, followed by a pilot at two programme district centers), every quantity of which is a target, not an observation. For imaging and fusion endpoints the reference standard is microbiological (smear microscopy, cartridge-based nucleic acid amplification and line-probe assay, per national programme protocol); a consensus radiologist read is recorded separately as a comparator rather than ground truth. The coding endpoint is referenced to a certified coder.

4.2 Implementation and Simulation Environment

Prototypes are written in Python, with PyTorch (Paszke et al., 2019) for the classical encoders and PennyLane (Bergholm et al., 2018) for the variational circuit; the quantum layer is a differentiable

module, so both parameter sets update in one graph. Classical training runs on the institutional GPU cluster (NVIDIA V100 and A100). Every quantum result here comes from noiseless state-vector simulation; no physical quantum hardware has been used. The prototype uses $n = 8$ qubits and ansatz depth $L_v = 4$; the encoding-layer setting L_e is not reported here. A superconducting backend is planned for validation, where the same circuit will be re-evaluated under device noise. No timing or throughput figures are given, as none have been recorded under controlled conditions.

4.3 Metrics and Statistical Procedure

Metrics are fixed in advance (Table 4, Appendix C). Imaging and fusion endpoints use AUC-ROC, sensitivity, specificity, balanced accuracy and F1, with sensitivity dominating operationally, since a missed case carries transmission consequences that a false alarm does not; calibration is assessed by Brier score and reliability curves, since a well-ranked but poorly calibrated score is unsafe to put in front of a clinician. Digitization uses character and word error rate stratified by script type and degradation; coding uses micro- and macro-averaged precision, recall, F1 and coder agreement; the computational characterization reports parameters, qubits, depth and measured latency. The analysis plan specifies five-fold cross-validation with patient-level splits, bootstrap confidence intervals, and McNemar's test for paired comparisons against each ablated variant; no formal sample-size calculation is reported in this prototype-stage manuscript. Subgroup performance is audited by sex, age band and severity, and $\kappa \geq 0.80$ against the reference coder is the coding acceptance criterion.

4.4 Preliminary Prototype Measurements

Four component prototypes exist, at varying maturity. QASP-Net's quality-assessment front end was evaluated on the ChestX-ray14 and Montgomery sets to establish that the descriptors of (1) compute stably across acquisition conditions; no diagnostic endpoint was measured. The variational circuit was trained in simulation on standard benchmarks and the residual backbone evaluated on public TB radiographs, never both under the protocol of Table 4, so no comparative figure is reported.

One quantitative measure is reported: the digitization module, initialized from a pre-trained transformer recognizer (Li et al., 2023) and fine-tuned on the 500-form institutional pilot corpus, achieves approximately 82% character-level accuracy (roughly 18% character error rate) on printed bilingual forms. Four qualifications apply: it is reported on a single institution's single form family; it applies only to printed text, where error rates should be lower than across handwriting; the transcription reference was generated by internal annotators without adjudicated double entry; and it is not stratified by degradation type as required by Table 4. It is reported as the framework's only measured quantity, not as a measure of clinical utility: an 18% character error rate sits well above any downstream coding system's tolerance for errors not corrected by the ontology-constrained mechanism of Section 3.4, whose effect on the error rate has not yet been measured; until it is, confidence filtering and human review would be required before downstream use.

The coding module was initialized from a biomedical language model (Lee et al., 2020) and fine-tuned on the tuberculosis chapter of the classification, with a multi-label prototype evaluated on synthetic clinical notes; since those notes were synthetic, the quantities described their generator, not clinical text, and are not reported. No end-to-end quantitative result exists; this paper therefore has no performance table, ROC curve, confusion matrix, or significance test for the framework.

4.5 Ablation Protocol

The ablation (Table 2) is designed to make the framework's claims individually falsifiable, not to prove staged improvements upon an obviously wrong baseline. Configuration A is the classical unimodal baseline; B isolates quality-conditional preprocessing; C replaces A's linear head with the variational quantum head, isolating the quantum component at fixed features, with a classical head of matched parameter count as an additional comparator; D tests that the contributions are additive and not redundant; and E, the full HQAI-TB, isolates the documentary evidence as the flagship contribution. A component is retained if the AUC improvement has a bootstrap confidence interval excluding zero on the held-out institutional cohort and survives the paired McNemar comparison against the configuration below it in Table 2. If the quantum head proves ineffective—as the barren plateau and encoding dependence of McClean et al. (2018) and Schuld et al. (2021) suggest it may

Table 2: Ablation protocol and retention criteria.

Config.	Components	Hypothesis isolated	Comparator	Retention criterion
A	ResNet-50, unenhanced radiograph	Classical unimodal baseline	–	Reference configuration
B	A + QASP-Net	Quality-conditioned preprocessing helps	vs. A	ΔAUC > 0, bootstrap CI excludes 0
C	A + variational quantum head	Quantum head helps at fixed features	vs. A and a parameter-matched classical head	ΔAUC > 0, CI excludes 0; McNemar $p < 0.05$
D	A + QASP-Net + quantum head	Contributions are additive, not redundant	vs. B and C	Exceeds both single-component variants
E	Full HQAI-TB (D + document stream + fusion)	Documentary evidence adds signal beyond imaging	vs. D	ΔAUC > 0, CI excludes 0; calibration not degraded

well be—the responsibility is on the author to report it as such rather than equivocating around the clinical utility of the retained components.

4.6 Discussion

Underneath the framework is an empirical claim about where the diagnostic information is in the first place. In a programmatic setting the radiograph is often the only evidence in machine-accessible digital form, while the laboratory result and treatment history remain on paper; if this is correct, the binding constraint is access to modalities, not classifier capacity, making digitization-to-decision, rather than a deeper imaging model, the appropriate target for improvement. Table 2 is the primary battleground for this claim, as configuration E differs from D only by the documentary stream. The variational head is a parameterization—a bounded-depth circuit on a small register realizes a function class that a linear classifier of comparable parameter count does not straightforwardly express (Schuld et al., 2021). Gradient variance decreases exponentially with depth under generic initialization (McClean et al., 2018), and a classical variant of the classifier given the same data has already found only weak separations that vanish under distribution shifts (Huang et al., 2021); the head is a hypothesis, tested at fixed features with an explicit retention criterion. Explainability and deployment are prerequisites, not optional add-ons: gradient-weighted localization (Selvaraju et al., 2017) provides spatial attribution, the fusion weights of (17) indicate which stream(s) caused a positive prediction, and the coding head can report the span of text supporting each code. None of these individually constitutes an explanation; deployment additionally requires ethics approval, a data-sharing agreement, identifier redaction, and fairness auditing.

5 Conclusion and Future Directions

The paper presents HQAI-TB, a hybrid quantum-classical framework processing radiographic and documentary evidence through separate channels, fused in a missing-modality-aware architecture and analyzed with a variational quantum head; the model's main cost is in the classical language component. We report a preliminary OCR character-level accuracy of approximately 82% on a 500-form pilot corpus but make no clinical or quantum advantage claims. Short-term priorities are assembling and annotating institutional cohorts, end-to-end training, and the five-configuration ablation; future work includes evaluating the variational head on hardware, federated training across sites, and improving handwriting recognition, given the current 18% character error rate on printed forms.

AI Use Statement

In this work, Large language models were used to assist with literature/reference retrieval, writing refinement and language polishing, and organization and formatting of portions of the manuscript. The authors reviewed and verified the resulting content and are fully responsible for the accuracy, originality, and final version of the manuscript. AI tools were not used to conduct experiments, generate experimental results, perform mathematical derivations, develop the methodology, or analyze experimental data.

REFERENCES

J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol. Multimodal biomedical ai. *Nature Medicine*, 28:1773–1784, 2022.

Emily Alsentzer et al. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop (NAACL)*, pp. 72–78, 2019.

Marcello Benedetti, Erika Lloyd, Stefan Sack, and Matthias Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, 2019.

Ville Bergholm et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*, 2018.

Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.

C. C. Boehme, P. Nabeta, D. Hillemann, M. P. Nicol, S. Shenai, F. Krapp, J. Allen, R. Tahirli, R. Blakemore, R. Rustomjee, A. Milovic, M. Jones, S. M. O'Brien, D. H. Persing, S. Ruesch-Gerdes, E. Gotuzzo, C. Rodrigues, D. Alland, and M. D. Perkins. Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the xpert mtb/rif test for diagnosis of tuberculosis and multidrug resistance: A multicentre implementation study. *Lancet*, 377(9776):1495–1505, 2011.

M. Cerezo et al. Variational quantum algorithms. *Nature Reviews Physics*, 3:625–644, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

V. Havlíček et al. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747): 209–212, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017.

Hsin-Yuan Huang, Michael Broughton, M. Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R. McClean. Power of data in quantum machine learning. *Nature Communications*, 12:2631, 2021.

S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *npj Digital Medicine*, 3:136, 2020.

Stefan Jaeger, Sema Candemir, Sameer Antani, Y.-X. J. Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 4(6):475–477, 2014.

Alistair E. W. Johnson et al. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3: 160035, 2016.

A. Kline et al. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5:171, 2022.

Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2): 574–582, 2017.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

J. Lee et al. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

M. Li et al. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13094–13102, 2023.

Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9:4812, 2018.

Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, 2018.

Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

Anish Mittal, R. Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of NAACL-HLT*, pp. 1101–1111, 2018.

F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, and D. Pfeiffer. Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization. *Scientific Reports*, 9:6268, 2019.

Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 8024–8035, 2019.

Z. Z. Qin et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Scientific Reports*, 9:15000, 2019.

T. Rahman et al. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601, 2020.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, 2021.

Ramprasaath R. Selvaraju, Abhishek Cogswell, Abhijit Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Ray Smith. An overview of the tesseract ocr engine. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 629–633, 2007.

James C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

Ashish Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5998–6008, 2017.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2097–2106, 2017.

World Health Organization. *Global Tuberculosis Report 2024*. WHO, Geneva, Switzerland, 2024.

Karel Zuiderveld. Contrast limited adaptive histogram equalization. In Paul S. Heckbert (ed.), *Graphics Gems IV*, pp. 474–485. Academic Press, San Diego, CA, USA, 1994.

A HQAI-TB INFERENCE ALGORITHM

Algorithm 1 HQAI-TB diagnostic pipeline (inference).
Input: radiograph x_i; document set D; trained parameters $\{W_l, W_g, W_p, \theta, W_k, u, W_a, w_o\}$.
Output: TB risk $\hat{y}$; structured coded summary S; attention weights α.

$q \leftarrow Q(x_i)$ // no-reference quality vector, (1)
$w \leftarrow \mathrm{softmax}(W_g q + b_g)$ // enhancement gating, (3)
$x' \leftarrow \sum_k w_k E_k(x_i)$; $x'' \leftarrow \mathrm{normalize}(x')$ // (2), (4)
$f_{\mathrm{img}} \leftarrow \mathrm{GAP}(\mathrm{ResNet}(x''))$ // (6), (7)
for each $d \in D$ **do**
  $\hat{w} \leftarrow \text{TBScript-OCR}(\mathrm{deskew}(d))$ // document digitization
  $\hat{w}^* \leftarrow$ lexicon-constrained correction$(\hat{w})$ // (14)
end for
$(E, C, f_{\mathrm{text}}) \leftarrow \text{TB-ICD-BERT}(\hat{w}^*)$ // entities, candidate codes, embedding, (15)
$f_{\mathrm{doc}} \leftarrow \mathrm{structure}(E)$ // document-derived fields
$\delta \leftarrow$ modality presence indicators
$g_k \leftarrow \mathrm{ReLU}(W_k f_k + b_k)$ for $k \in \{\mathrm{img}, \mathrm{doc}, \mathrm{text}\}$ // (16)
$\alpha \leftarrow$ presence-gated attention(g, δ); $z \leftarrow \sum_k \alpha_k g_k$ // (17), (18)
$s \leftarrow \pi \tanh(W_p z + b_p)$ // bounded scaling, (8)
$|\psi\rangle \leftarrow U(\theta) U_{\mathrm{enc}}(s)|0\rangle^{\otimes n}$ // encode and execute VQC, (9), (10)
$m_j \leftarrow \langle\psi| Z_j |\psi\rangle$ for $j = 1, \ldots, n$ // measurement, (11)
$\hat{y} \leftarrow \sigma(w_o^T m + b_o)$ // TB prediction, (12)
$S \leftarrow \mathrm{render}(E, C, \hat{y}, \alpha)$ // decision-support summary
return $\hat{y}, S, \alpha$

B COMPUTATIONAL COMPLEXITY

Cost is dominated by the classical components. Quality estimation and each enhancement operator are linear in pixel count, so QASP-Net contributes $O(K \cdot HW)$ for $K = 5$ operators—negligible next to the encoder. A ResNet-50 backbone carries about 25.6 million parameters and on the order of 4 GFLOPs per 224×224 forward pass (He et al., 2016), and a base-sized transformer encoder for clinical text adds roughly 110 million parameters, with attention costing $O(T^2 d)$ in sequence length T (Devlin et al., 2019; Vaswani et al., 2017). Fusion is a handful of dense projections at $O(\sum_k d_k d)$.

The quantum head, by contrast, is tiny. With n qubits and L_v variational layers at two rotations per qubit per layer, $|\theta| = 2nL_v$, the two-qubit gate count is $O(nL_v)$, and a ring entangler keeps parallelized depth at $O(L_v)$. At $n = 8$ and $L_v = 4$ that is 64 quantum parameters against 25.6 million in the image encoder—a rounding error on model size. The cost is in evaluations. State-vector simulation is $O(2^n)$ per shot, cheap at $n = 8$ but closing off naive scaling, and parameter-shift training needs $2|\theta| = 128$ circuit executions per sample per optimizer step, which is why simultaneous perturbation would replace it on physical hardware (Spall, 1992).

C DATASET AND ENDPOINT TABLES

Table 3: Datasets and planned evaluation configuration.

Resource	Content	Role in this work	Status
Montgomery County CXR (Jaeger et al., 2014)	138 frontal radiographs, radiologist TB labels	Imaging evaluation set	Public; used in component proof-of-concept
Shenzhen CXR (Jaeger et al., 2014)	662 frontal radiographs, radiologist TB labels	Imaging evaluation set	Public; selected for framework validation
ChestX-ray14 (Wang et al., 2017)	112,120 frontal images, 30,805 patients	Thoracic pre-training only; labels report-derived	Public; used in component proof-of-concept
TB chest radiography set (Rahman et al., 2020)	Public TB and normal radiographs	Additional imaging cohort	Public; selected for framework validation
De-identified clinical notes (Johnson et al., 2016)	Free-text critical-care documentation	Clinical language-model adaptation	Public; selected for framework validation
Institutional TB form corpus	500 printed and handwritten bilingual forms	OCR fine-tuning and character-accuracy measurement	Collected and used (see Section 4.4)
Institutional radiograph cohort	Target 5,000 annotated radiographs	End-to-end imaging and fusion validation	Not collected; selected for future validation
Institutional document cohort	Target 10,000 clinical documents	OCR, NLP and fusion validation	Not collected; selected for future validation

Table 4: Evaluation endpoints, reference standards and analysis plan.

Module / endpoint	Primary metrics	Reference standard	Analysis
QASP-Net (preprocessing)	BRISQUE, NIQE; downstream AUC gain	No-reference IQA; TB label	Paired comparison against fixed CLAHE pipeline
VQC-ResNet (classification)	AUC-ROC, sensitivity, specificity, balanced accuracy, F1	Microbiological composite	5-fold patient-level CV; bootstrap 95% CI; McNemar
TBScript-OCR (digitization)	Character and word error rate	Manual transcription	Stratified by script type, blur, skew, contrast
TB-ICD-BERT (coding)	Micro / macro precision, recall, F1; agreement	Certified coder assignment	Per-code and aggregate; acceptance at $\kappa \geq 0.80$
Fusion layer	AUC-ROC; Brier score; reliability curve	Microbiological composite	Compared against best single-modality branch
Missing-modality behaviour	AUC under each modality withheld	Microbiological composite	Degradation profile relative to full input
Computational profile	Parameters, qubits, circuit depth, latency	Not applicable	Measured on stated hardware; simulator and device