Quantum Machine Intelligence (2026) 8:46
https://doi.org/10.1007/s42484-026-00392-7

RESEARCH ARTICLE

Unlocking the power of quantum computing in biomedical NLP for lung cancer diagnosis

Cyrille YetuYetu Kesiku[1] · **Begonya Garcia-Zapirain**[1] · **Adel S. Elmaghraby**[2]

Received: 10 February 2025 / Accepted: 30 March 2026

Abstract

Lung cancer remains the leading cause of cancer-related mortality worldwide, underscoring the urgent need for accurate, efficient, and interpretable early detection methods. Existing benchmark large language models (LLMs) like ClinicalBERT and BioBERT, despite their advancements in biomedical text analysis, face critical limitations including high computational costs, limited interpretability, and reliance on extensive annotated datasets hindering their clinical integration. To address these challenges, we propose the Hybrid Attention Quantum Long Short-Term Memory-Attention (A-QLSTM-A) model, a novel quantum-classical framework that combines quantum variational circuits with LSTM networks and dual attention mechanisms. This design enhances feature extraction, improves interpretability, and offers a more efficient architectural pathway. Evaluated on MIMIC-III discharge summaries and MIMIC-IV chest radiography reports, A-QLSTM-A achieved 98.32% and 83.67% accuracy, respectively, surpassing existing models. This study establishes a new benchmark for scalable, interpretable AI in precision medicine, offering a promising tool for early lung cancer detection and clinical decision support.

Keywords Natural Language Processing(NLP) · Quantum NLP · Lung cancer · Quantum computing · Neural network

1 Introduction

Lung cancer remains one of the most lethal malignancies globally, claiming over 1.8 million lives in 2020 and accounting for approximately 18% of all cancer-related deaths (Kratzer et al. 2024; Sung et al. 2021). Survival is strongly stage dependent: five year survival rates exceed 60% for Stage I disease but fall below 6% for Stage IV, making early detection a life-critical imperative (Woodard et al. 2016; Siegel et al. 2023). Yet the heterogeneous, high-dimensional, and frequently unstructured nature of clinical data encompassing radiology reports, discharge summaries, pathology notes, and genomic records presents persistent challenges for conventional diagnostic models (Wang et al. 2010).

The advent of transformer-based large language models (LLMs) has transformed biomedical text mining. ClinicalBERT (Huang et al. 2019) and BioBERT (Lee et al. 2020) achieve strong performance by pre-training on domain-specific corpora, while Med-BERT (Rasmy et al. 2021) extends this to structured electronic health records (EHRs). PubMedBERT (Gu et al. 2021) demonstrates that in-domain pre-training from scratch outperforms general-domain transfer for biomedical NLP tasks. GatorTron (Yang et al. 2022), trained on over 90 billion clinical words, sets state-of-the-art benchmarks on biomedical relation extraction and named entity recognition. BioGPT (Luo et al. 2022) and BioMedLM (Venigalla et al. 2022) extend generative capabilities to biomedical question answering and text generation. Despite these advances, LLM-based approaches share critical systemic limitations: (i) astronomical computational requirements (hundreds of GPU-hours for fine-tuning), (ii) opaque "black-box" reasoning that limits clinical trust (Tonekaboni et al. 2019), and (iii) dependence on large annotated corpora rarely available in specialized oncology settings (Fink et al. 2022).

✉ Cyrille YetuYetu Kesiku
cyrille.kesiku@opendeusto.es

Begonya Garcia-Zapirain
mbgarciazapi@deusto.es

Adel S. Elmaghraby
adel.elmaghraby@louisville.edu

[1] University of Deusto, Bilbao, Spain

[2] University of Louisville, Louisville, United States

Published online: 15 April 2026

For lung cancer diagnosis specifically, Mithun et al. (2023) compared deep learning and BERT models on radiology report classification, finding that domain adaptation is critical but computational overhead remains prohibitive in resource-constrained hospitals. Gonzalez-Hernandez et al. (2017) demonstrated the efficacy of NLP pipelines for extracting TNM staging from pathology reports, while Kehl et al. (2019) showed that NLP-based progression detection from oncologist notes matches expert annotation with $> 90\%$ concordance. Nevertheless, none of these approaches adequately addresses the interpretability-efficiency trade-off required for regulatory compliant clinical deployment.

Quantum computing introduces a fundamentally new computational paradigm by exploiting quantum-mechanical phenomena superposition, entanglement, and quantum interference to process information in ways that are classically intractable (Preskill 2018; Biamonte et al. 2017). Near-term noisy intermediate-scale quantum (NISQ) devices, comprising 50–150 qubits, have enabled a new class of algorithms known as variational quantum algorithms (VQAs) (Cerezo et al. 2021), which optimize parameterized quantum circuits (PQCs) through classical gradient methods. Quantum support vector machines (Havlíček et al. 2019) and quantum neural networks (Schuld et al. 2020) have demonstrated the ability to construct kernel functions in exponentially large Hilbert spaces, offering potential advantages for high-dimensional feature learning.

Within quantum machine learning (QML), several architectures directly relevant to sequence modeling have emerged. The Quantum LSTM (QLSTM) (Chen et al. 2022) replaces classical LSTM gates with VQCs, enabling quantum-enhanced memory dynamics while preserving the recurrent structure. Quantum self-attention networks (Li et al. 2024) (QSANN) apply quantum dot-product attention to text classification, achieving 84.79% accuracy on the Yelp Review Polarity benchmark marginally higher than our model in peak accuracy but with a substantially higher error rate (1.29% vs. 0.34%), a critical distinction for safety-critical clinical applications. Quantum transformers (Sipio et al. 2022) and quantum graph neural networks (Mernyei et al. 2022) further extend the quantum deep-learning landscape, though their scalability to large clinical corpora remains open.

Quantum Natural Language Processing (QNLP) formalizes grammatical structure using categorical quantum mechanics and the ZX-calculus (Coecke et al. 2020; Meichanetzidis et al. 2020). The DisCoCat framework (Clark 2021) provides a compositional semantics for natural language using quantum circuits, and has been implemented on quantum hardware by Lorenz et al. (2023). More recently, *lambeq* (Kartsaklis et al. 2021) has made QNLP pipelines accessible for practitioners. However, current QNLP approaches are constrained to short sentences due to the exponential growth of circuit depth with sentence length, making them unsuitable for processing long-form clinical narratives such as MIMIC discharge summaries.

The intersection of quantum computing and biomedical NLP remains largely unexplored (Emani et al. 2021; Balamurugan et al. 2024). Critically, no prior work has combined QLSTM-based recurrent processing with dual attention mechanisms for clinical text classification. Our baseline A-LSTM-A model (Attention-LSTM-Attention) provides a direct classical comparator by framing the architecture with attention wrappers around a classical LSTM; replacing the LSTM core with a QLSTM module incorporating four independent VQCs for the forget, input, update, and output gates constitutes the primary quantum innovation of A-QLSTM-A. Beyond gate replacement, the quantum circuits induce non-separable feature entanglement across the input dimensions, enabling the model to capture correlations that shallow classical attention layers cannot represent. Interpretability in this work refers to the multi-level transparency of the model's predictions: (i) *global interpretability* via permutation feature importance analysis, identifying which clinical terms drive classification at the population level; (ii) *local interpretability* via attention weight visualization, explaining individual predictions; and (iii) *regulatory interpretability*, whereby the modular quantum-classical architecture supports auditing of each decision component, consistent with the FDA's framework for AI/ML-based software as a medical device (U.S. Food and Drug Administration 2021). This tripartite definition follows Doshi-Velez and Kim (2017) and Tonekaboni et al. (2019).

The main contributions of this paper can listed as follows:

- A novel **A-QLSTM-A** architecture integrating dual attention with QLSTM for biomedical sequence classification, explicitly compared against its classical counterpart A-LSTM-A to isolate the quantum contribution.
- A mathematically rigorous formulation of all model components, including consistent encoding conventions, gate-level VQC specifications, and closed-form attention mechanisms.
- A comprehensive evaluation on MIMIC-III, MIMIC-IV, and Yelp Review Polarity, with results averaged over multiple seeds, cross-validated over 10 folds, and ablated over circuit depth and entanglement topology.
- A three-level interpretability analysis and a regulatory-compliance discussion aligned with clinical AI governance frameworks.

2 Materials

This section outlines the datasets and resources utilized to develop and validate the Hybrid Attention Quantum Long Short-Term Memory-Attention (A-QLSTM-A) model. We primarily leverage the MIMIC-III and MIMIC-IV clinical datasets, containing unstructured medical narratives essential for lung cancer diagnosis. Additionally, we employ the Yelp Review Polarity dataset to assess model generalization beyond biomedical text. Data preprocessing, selection criteria, and statistical analyses ensure high-quality input for robust performance evaluation. These materials provide a comprehensive foundation for training, validating, and benchmarking the model's efficacy in real-world clinical applications.

MIMIC-III dataset The MIMIC-III Discharge Summary dataset (Johnson et al. 2016) was selected as the primary source due to its rich and structured clinical narratives, encompassing essential information such as medical, social, and family history, as well as present illness. These multidimensional data facilitate the construction of comprehensive patient profiles, crucial for accurate lung cancer detection.

A total of 5918 clinical records were extracted, including 2700 lung cancer cases (45.6%) and 3218 cases of other malignancies (54.4%), such as colorectal, hepatic, and breast cancers. The selection process, based on ICD-9 diagnostic codes, ensured precise disease classification. Structured SQL queries in a PostgreSQL database were employed to retrieve and refine the dataset through a two-stage selection strategy. First, 8278 records were identified (as detailed in Table 2), and second, only those containing key sections medical history, social history, family history, and present illness were retained. These sections were then integrated into a unified narrative, providing a holistic view of each patient's clinical profile.

Demographic analysis revealed that 64.3% of lung cancer patients were male (1737) and 35.6% female (963), with ages ranging from 34 to 88 years. Among patients with other cancers, 66.9% were male (2154) and 33.1% female (1064), with an age range of 23 to 88 years. These distributions highlight gender disparities and the increased cancer risk associated with aging, reinforcing the dataset's value for epidemiological research and the development of personalized medical interventions. Table 1 provides a comprehensive demographic and clinical characterization of lung cancer patients across different ethnic groups, detailing key metrics such as patient count, average age, and hospital stay duration. White patients represent the largest subgroup, with 1998 individuals, an average age of 67.5 years, and a mean hospital stay of 10.5 days. African American patients, the second-largest group, exhibit a slightly lower average age of 65.3 years and a comparable hospitalization period of 10.1 days. In contrast, Asian patients, though fewer in number (145), demonstrate the longest hospital stays, averaging 11.0 days, despite a younger mean age of 63.9 years. The smallest cohort, categorized as Other (43 patients), presents the shortest hospitalization period (9.8 days) and the lowest average age (62.8 years). These disparities in age distribution in Fig. 1 and hospital stay durations across ethnicity underscore potential differences in disease progression, healthcare access, and treatment responses, emphasizing the necessity for personalized medical approaches.

The statistical analysis of smoking and drinking lifestyle behaviors among lung cancer patients and other cancer patients in Table 2 reveals significant and distinctive patterns between the two groups, underscoring the strong association between these lifestyle factors and cancer development.

The lung cancer cohort exhibits a pronounced correlation between smoking and disease prevalence, with 72.6% (1,959 patients) identified as smokers, reinforcing tobacco use as a primary risk factor. Additionally, 49.7% (1,341 patients) report alcohol consumption, further emphasizing lifestyle-related influences on disease onset. Among lung cancer subtypes, the highest smoking and drinking prevalence is observed in type 162.8 (malignant neoplasm of other parts of the bronchus or lung) and type 162.3 (malignant neoplasm of the upper lobe, bronchus, or lung). Specifically, 71.2% (604 patients) diagnosed with type 162.8 are smokers, while 49.1% (416 patients) consume alcohol. Similarly, in type 162.3, smoking prevalence rises to 77.2% (610 patients), with 51.3% (406 patients) engaging in alcohol consumption.

Table 1 Patient statistics for lung cancer and other cancers by ethnicity

Cancer Type	Ethnicity	Num patients	Avg Age (years)	Average Stay (days)
Lung Cancer	White	1998	67.5	10.5
	African American	289	65.3	10.1
	Asian	145	63.9	11.0
	Other	43	62.8	9.8
Other Cancers	White	4123	72.3	11.8
	African American	732	69.2	10.7
	Asian	304	70.1	12.3
	Other	158	68.7	10.4

Statistical analysis on patients features: The table presents patient statistics by ethnicity, highlighting differences in average age and hospital stay between lung cancer and other cancers, providing insights into demographic and clinical disparities

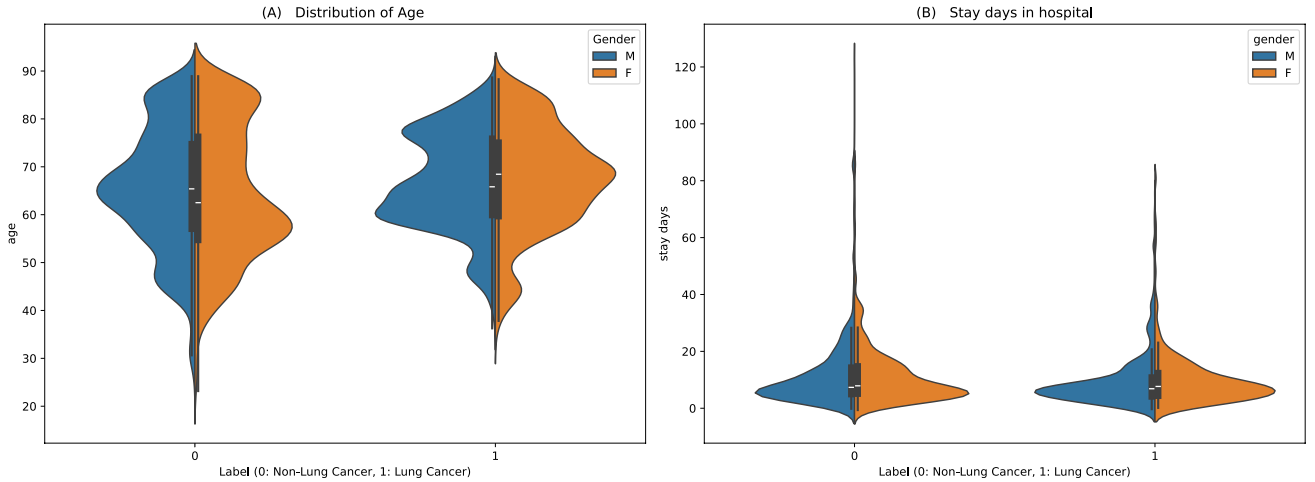

Fig. 1 Statistical distribution plots: This violin plots illustrate age (**A**) and hospital length of stay (**B**) distributions across diagnostic groups (0: non-lung cancer, 1: lung cancer) stratified by gender. Lung cancer patients (label 1) are predominantly older, with males exhibiting a more concentrated age range compared to females. Non-lung cancer cases (label 0) show a broader, younger age distribution. Hospital stays are significantly longer for lung cancer patients, reflecting the complexity of care, with males showing greater variability in extended stays. In contrast, non-lung cancer stays are shorter and more uniform

In contrast, other cancers exhibit a slightly lower but still significant association with smoking and drinking. Within this cohort, 55.7% (1,794 patients) are smokers, with particularly high prevalence in liver cancer (type 155.0) and colorectal cancer (type 153.0). Alcohol consumption follows a similar pattern, with 44.5% (1,433 patients) reporting drinking, though at a lower concentration compared to lung cancer. These findings highlight critical behavioral risk factors across cancer types, underscoring the need for targeted prevention strategies and personalized intervention approaches.

MIMIC-IV database The MIMIC-IV database (Gow et al. 2023), comprising 2321355 de-identified radiology reports from 237427 patients, was meticulously curated to focus on lung cancer diagnosis. Reports were filtered using ICD-9 diagnostic codes, retaining only those with a detailed Findings section. Reports with fewer than 50 words in the Findings section were excluded to ensure robust textual content. This process yielded a curated dataset of 1,304 reports, balanced between 699 lung cancer cases and 605 nonlung cancer cases. This high-quality data set underscores the relevance of the framework for real-world biomedical NLP applications.

Yelp review polarity dataset The Yelp Review Polarity dataset (Zhang et al. 2015), consisting of 560,000 evenly distributed samples of positive and negative sentiments, was used to test model generalizability. Preprocessing ensured a focus on sentiment polarity, offering a robust benchmark for evaluating nuanced text classification tasks.

3 Proposed method

This section details the methodologies employed, including text preprocessing, feature extraction, and quantum-enhanced classification techniques. We describe the model's architecture in Fig. 2, data embedding strategies, and quantum circuit integration, ensuring optimal interpretability and computational efficiency. Additionally, we outline the training procedures, hyperparameter tuning, and evaluation protocols used to validate performance across biomedical datasets.

3.1 Text vectorization technique

The transformation of raw text into a numerical representation using the Term Frequency-Inverse Document Frequency (TF-IDF) (Robertson 2004) technique is a fundamental step in natural language processing and information retrieval. This method's capacity to offer a computationally efficient representation of textual input is what led to its selection in a hybrid quantum-classical binary classification model. For a variety of reasons, including sparse and structured representation, compatibility with quantum encoding, robustness in low-data regimes, and the current limitations of quantum hardware, TF-IDF is an ideal feature extraction technique in

Table 2 Statistical characteristics of patients

Disease	ICD-9	gender	age min	age max	smokers	drinkers	num
Lung cancer							
	162.0	M: 0, F: 3	62	63	0	3	3
	162.2	M: 143, F: 52	41	86	144	104	195
	162.3	M: 460, F: 330	37	88	610	406	790
	162.4	M: 77, F: 7	54	85	77	28	84
	162.5	M: 224, F: 139	42	87	245	195	363
	162.8	M: 574, F: 274	34	88	604	416	848
	162.9	M: 259, F: 158	38	86	279	189	417
Total	-	M: 1737, F: 963	-	-	1959	1341	2700
Other cancers							
	153.0	M: 64, F: 0	71	87	48	24	64
	153.1	M: 82, F: 12	54	87	40	41	94
	153.2	M: 21, F: 18	38	88	20	14	39
	153.3	M: 52, F: 49	35	87	75	38	101
	153.4	M: 73, F: 43	47	86	83	53	116
	153.5	M: 10, F: 2	29	85	9	1	12
	153.6	M: 110, F: 90	25	88	71	71	200
	153.7	M: 11, F: 8	81	87	16	16	19
	153.8	M: 68, F: 68	47	87	77	62	136
	153.9	M: 54, F: 22	35	88	49	27	76
	154.0	M: 58, F: 16	30	88	40	35	74
	154.1	M: 94, F: 26	43	87	55	38	120
	154.2	M: 6, F: 0	73	73	0	2	6
	154.3	M: 13, F: 4	57	76	14	6	17
	154.8	M: 18, F: 8	36	79	22	6	26
	155.0	M: 1397,F: 199	23	88	964	794	1596
	174.3	M: 0, F: 2	37	55	1	0	2
	174.4	M: 0, F: 12	42	85	3	5	12
	174.5	M: 0, F: 1	85	85	0	1	1
	174.8	M: 0, F: 298	24	86	116	126	298
	174.9	M: 0, F: 209	36	88	91	73	209
Total	-	M: 2131, F: 1087	-	-	1794	1433	3218

Patients lifestyle characteristics: The table presents statistical data on smoking and drinking behaviors in lung and other cancer patients. Lung cancer subtypes, particularly those affecting the bronchus (162.3) and upper lobe (162.4), show high prevalence rates of smoking and drinking, with 610 smokers (77.2%) and 406 drinkers (51.3%) in one subtype. Other cancers, such as liver (155.0) and colorectal (153.0), have a lower but still notable prevalence of smoking and drinking. Gender differences are also observed, with higher proportions of female drinkers in breast cancer (174.8). These patterns highlight the significant role of lifestyle behaviors in cancer risk

the context of quantum-enhanced machine learning, where computational resources are frequently limited and the integration of classical and quantum components requires structured input. By choosing just the most pertinent words, TF-IDF makes it possible for a low-dimensional yet informative feature space to be processed efficiently by quantum classifiers.

This approach effectively quantifies the importance of individual terms across a dataset, enabling a structured representation that enhances both interpretability and analytical rigor. By balancing word frequency with term distinctiveness, TF-IDF serves as a bridge between unstructured textual data and structured numerical features, offering a robust preprocessing mechanism for downstream machine learning tasks.

Mathematically, TF-IDF is computed as the product of two core components: Term Frequency (TF) and Inverse Document Frequency (IDF). The TF component measures the frequency of a term t within a document d , typically normalized to prevent length bias. It is defined as:

$$\mathrm{TF}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{1}$$

where $f_{t,d}$ denotes the raw count of term t in document d, and the denominator represents the total count of all terms in d. The IDF component, in contrast, accounts for the rarity of a term across the corpus D, down-weighting common words while emphasizing distinctive terms. It is given by:

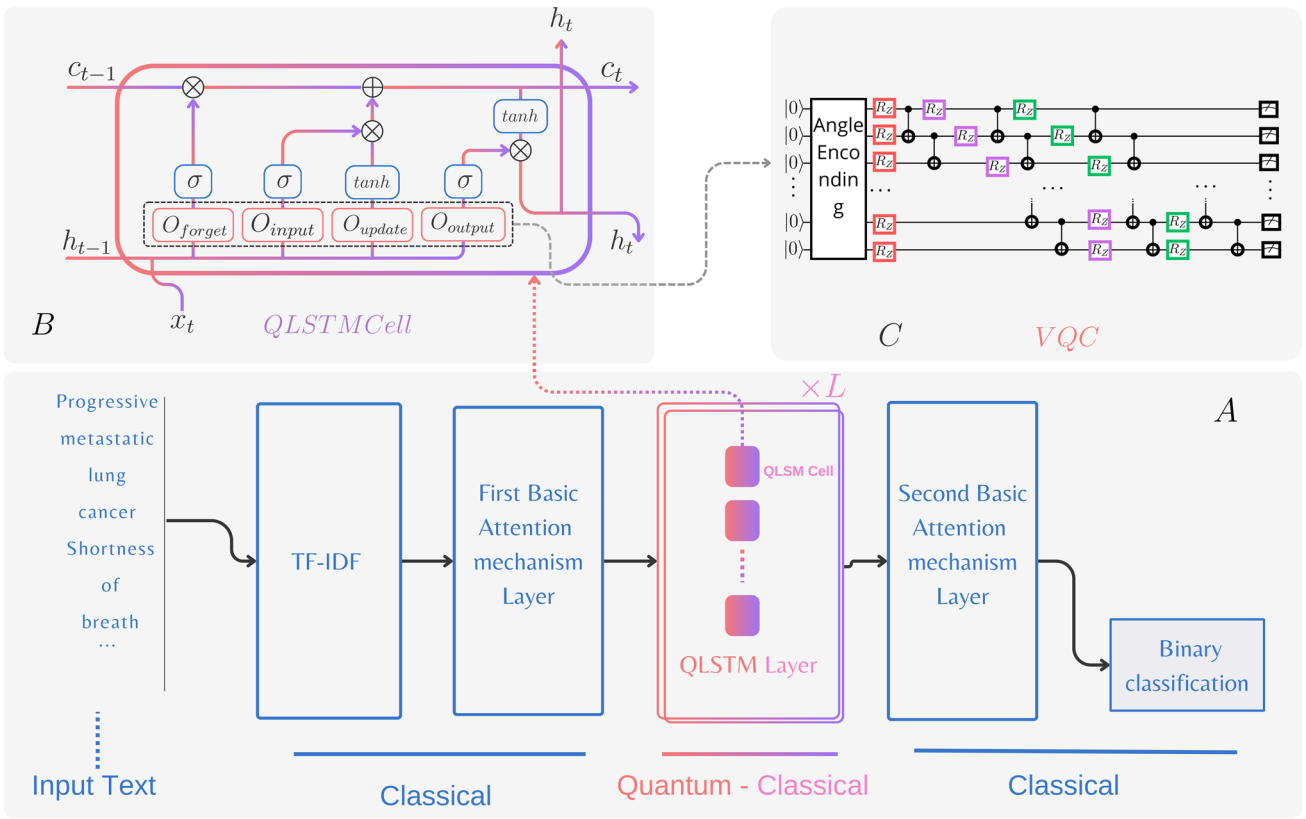

Fig. 2 A-QLSTM-A design architecture : The A-QLSTM-A general architecture sub-figures (**A**) integrates TF-IDF vectorization to convert text into numerical features. First Basic attention layer refines contextual information before feeding it into two QLSTM layers with 8 qubits each and a second attention layer refines the QLSTM output for classification. The QLSTM cell structure and VQC-based gates are detailed in additional sub-figures (**B**). The Variational Quantum Circuit (VQC) in sub-figures (**C**) begins with Angle Embedding using $RY(\pi/2)$ rotations, followed by a Basic Entangler Layer with CNOT gates in a ring topology. Parameterized $RZ(\pi/4)$ rotations optimize expressiveness, and Pauli-Z measurements extract quantum features

$$\text{IDF}(t, D) = \log\left(\frac{N}{1 + |\{d \in D : t \in d\}|}\right) \tag{2}$$

where N represents the total number of documents in the corpus, and the denominator corresponds to the number of documents containing term t, with a smoothing factor (+1) to prevent division by zero. Consequently, the TF-IDF score for a term t in document d is computed as:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D) \tag{3}$$

This formulation ensures that terms highly specific to a document but rare in the broader corpus receive greater significance, making TF-IDF particularly effective for tasks such as document classification.

3.2 Basic attention mechanism

A preliminary attention mechanism is applied to this representation, dynamically learning the importance of each feature to produce a context vector. This step ensures that the model focuses on the most critical components of the input, enhancing both feature selection and interpretability, pivotal for complex data scenarios. Following this quantum processing, a secondary attention mechanism refines the output, ensuring that the final representation emphasizes the most salient features. This dual attention framework, strategically applied before and after the QLSTM layer, strengthens the robustness and interpretability of the model's predictions.

Let the input sequence $X \in \mathbb{R}^{n \times d}$ represent n tokens, each embedded in a d-dimensional space. The initial attention mechanism calculates token importance using a projection $W \in \mathbb{R}^{d \times k}$ to map embeddings into an attention space of dimension k. The resulting score matrix $S \in \mathbb{R}^{n \times k}$ is given by:

$$S = XW \tag{4}$$

Applying the softmax function across S produces attention weights $\alpha \in \mathbb{R}^{n \times 1}$, normalized across tokens:

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{n} \exp(s_j)}, \quad i \in \{1, \ldots, n\} \tag{5}$$

The context vector $c \in \mathbb{R}^d$, summarizing key sequence features, is computed as:

$$c = X^T \alpha \tag{6}$$

3.3 Quantum data embedding

Two complementary encoding strategies are used at different stages of the architecture, motivated by their distinct roles:

Standard Angle Encoding (VQC input layer): For the VQC acting on the projected input y_t (Section 3.7), each qubit i encodes a single scalar feature $y_{t,i}$ via a Y-rotation:

$$|\phi(\boldsymbol{y}_t)\rangle = \bigotimes_{i=1}^{n_q} R_Y(y_{t,i})\,|0\rangle, \quad R_Y(\theta) = \begin{pmatrix} \cos\frac{\theta}{2} & -\sin\frac{\theta}{2} \\ \sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{pmatrix}. \tag{7}$$

This encoding is used in Eq. 11 and is exactly what the VQC circuit diagram (Fig. 3) depicts.

Dense Angle Encoding (QLSTM cell input):

When encoding the higher-dimensional context vector $\boldsymbol{c}$ into the QLSTM cell, we pack two features per qubit via:

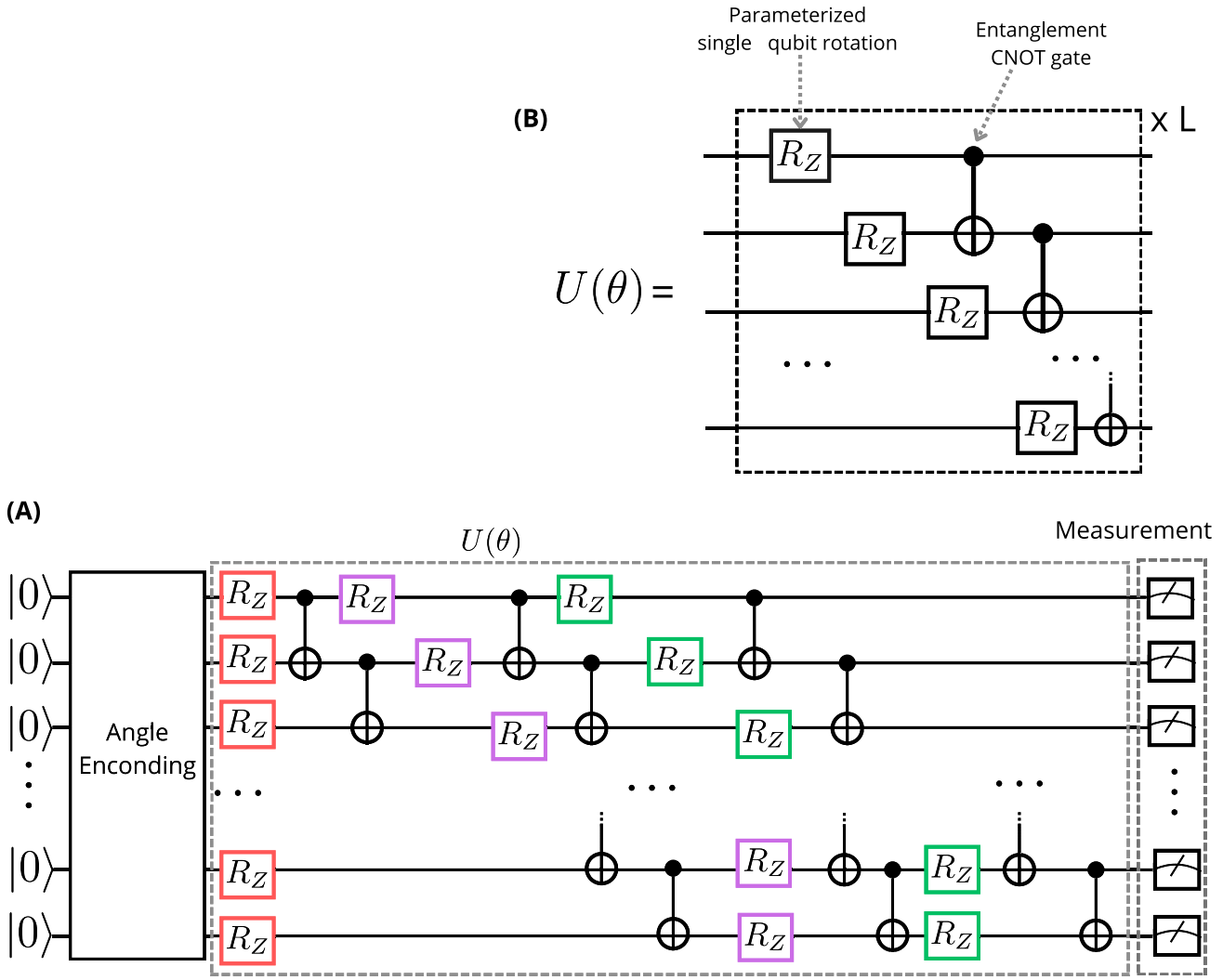

Fig. 3 Three-layers variational quantum circuit (VQC) architecture: (**A**) The variational quantum circuit designed in this research, with three layers as depth-wise. Each classical input vector $\mathrm{x} = (x_1, \ldots, x_n)$ is embedded into quantum states via angle encoding, where $R_y(x_i)$ rotations are applied to individual qubits. (**B**) Represents the layer in the circuit. Each layer consists of trainable parameterized unitaries $U(\boldsymbol{\theta}) = R_z(\theta_1)R_y(\theta_2)R_z(\theta_3)$, followed by an entangling block using CNOT gates. This structure is repeated across three layers, enabling the quantum circuit to capture non-linear dependencies and correlations in high-dimensional Hilbert space. Final measurement is performed in the computational basis to extract expectation values used for downstream learning tasks

$$|\boldsymbol{x}\rangle = \bigotimes_{i=1}^{\lceil n/2 \rceil} \Big(\cos(\pi x_{2i-1}) \, |0\rangle + e^{i2\pi x_{2i}} \sin(\pi x_{2i-1}) \, |1\rangle \Big), \tag{8}$$

halving the qubit overhead for the same number of features. This encoding is used in the QLSTM cell block (Fig. 2B) and is *distinct* from the standard angle encoding used in the VQC.

3.4 Quantum integration via QLSTM

At each time step t, the concatenated input and hidden state:

$$\boldsymbol{v}_t = [\boldsymbol{h}_{t-1}, \, \boldsymbol{x}_t] \in \mathbb{R}^{n_{\text{in}}+n_h} \tag{9}$$

is projected to n_q qubit dimensions:

$$\boldsymbol{y}_t = W_{\text{in}} \boldsymbol{v}_t + \boldsymbol{b}_{\text{in}}, \quad W_{\text{in}} \in \mathbb{R}^{n_q \times (n_{\text{in}}+n_h)}, \; \boldsymbol{b}_{\text{in}} \in \mathbb{R}^{n_q}. \tag{10}$$

Using standard angle encoding (Eq. 7), $\boldsymbol{y}_t$ is encoded into a quantum state:

$$|\phi(\boldsymbol{y}_t)\rangle = \bigotimes_{i=1}^{n_q} R_Y(y_{t,i}) \, |0\rangle. \tag{11}$$

The state is evolved by a variational unitary $U(\boldsymbol{\theta})$:

$$|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta}) \, |\phi(\boldsymbol{y}_t)\rangle. \tag{12}$$

Pauli-Z expectation values are extracted by measurement:

$$z_i = \langle \psi(\boldsymbol{\theta})| \, Z_i \, |\psi(\boldsymbol{\theta})\rangle, \quad i = 1, \ldots, n_q. \tag{13}$$

The measurement vector $\boldsymbol{z}_{\text{gate}} \in \mathbb{R}^{n_q}$ is classically post-processed to produce the gate's output:

$$\boldsymbol{o}_{\text{gate}} = W_{\text{out}} \, \boldsymbol{z}_{\text{gate}} + \boldsymbol{b}_{\text{gate}}, \quad W_{\text{out}} \in \mathbb{R}^{n_{\text{gate}} \times n_q}, \; \boldsymbol{b}_{\text{gate}} \in \mathbb{R}^{n_{\text{gate}}}. \tag{14}$$

3.5 LSTM gate dynamics

Four independent VQCs as presented in Fig. 3, each instantiated with its own trainable parameters $\boldsymbol{\theta}^{(\cdot)}$ compute gate-specific output vectors.

Following Eq. 14, each gate obtains a dedicated output:

$$\begin{aligned} \boldsymbol{o}_{\text{forget}} &= W_{\text{out}}^{(f)} \, \boldsymbol{z}^{(f)} + \boldsymbol{b}^{(f)}, \\ \boldsymbol{o}_{\text{input}} &= W_{\text{out}}^{(i)} \, \boldsymbol{z}^{(i)} + \boldsymbol{b}^{(i)}, \\ \boldsymbol{o}_{\text{update}} &= W_{\text{out}}^{(u)} \, \boldsymbol{z}^{(u)} + \boldsymbol{b}^{(u)}, \\ \boldsymbol{o}_{\text{output}} &= W_{\text{out}}^{(o)} \, \boldsymbol{z}^{(o)} + \boldsymbol{b}^{(o)}. \end{aligned} \tag{15}$$

These are then used to compute the standard LSTM gating operations:

$$\begin{aligned} \boldsymbol{f}_t &= \sigma(\boldsymbol{o}_{\text{forget}}), & \boldsymbol{i}_t &= \sigma(\boldsymbol{o}_{\text{input}}), \\ \boldsymbol{g}_t &= \tanh(\boldsymbol{o}_{\text{update}}), & \boldsymbol{o}_t &= \sigma(\boldsymbol{o}_{\text{output}}). \end{aligned} \tag{16}$$

Memory cell and hidden state updates follow the classical LSTM recurrence:

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \boldsymbol{g}_t, \tag{17}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t), \tag{18}$$

where $\odot$ denotes the Hadamard (element-wise) product.

3.6 Binary classification layer

The final hidden state $\boldsymbol{h}_T$ is passed through a dense sigmoid layer:

$$\hat{y} = \sigma(W_{\text{cls}} \, \boldsymbol{h}_T + b_{\text{cls}}), \quad W_{\text{cls}} \in \mathbb{R}^{1 \times n_h}, \; b_{\text{cls}} \in \mathbb{R}. \tag{19}$$

The output $\hat{y} \in (0, 1)$ represents the probability of the positive class (lung cancer), and binary cross-entropy is minimized during training:

$$\mathcal{L} = -\frac{1}{B} \sum_{b=1}^{B} \big[y_b \log \hat{y}_b + (1 - y_b) \log(1 - \hat{y}_b) \big], \tag{20}$$

where B is the batch size and $y_b \in \{0, 1\}$ is the ground-truth label.

3.7 Quantum variational circuit design

Each QLSTM gate employs a VQC of the form:

$$|\psi_{\text{out}}\rangle = U(\boldsymbol{\theta}) \, |\phi(\boldsymbol{x})\rangle, \tag{21}$$

where $|\phi(\boldsymbol{x})\rangle$ is given by standard angle encoding (Eq. 7). The variational unitary consists of L layers:

$$U(\boldsymbol{\theta}) = \prod_{l=1}^{L} \left[\bigotimes_{i=1}^{n_q} R_Z(\theta_i^{(l)}) \right] \cdot E_{\text{CNOT}}^{(l)}, \tag{22}$$

with a linear-topology entanglement block:

$$E_{\text{CNOT}}^{(l)} = \prod_{i=1}^{n_q - 1} \text{CNOT}(q_i, q_{i+1}). \tag{23}$$

Following measurement (Eq. 13), the observable vector $\boldsymbol{z} \in \mathbb{R}^{n_q}$ undergoes classical post-processing:

$$\boldsymbol{o}_{\text{gate}} = W\boldsymbol{z} + \boldsymbol{b}, \tag{24}$$

enabling seamless integration with classical backpropagation. Note that Eqs. 13–24 are the VQC-local restatement of the general QLSTM relations in Eqs. 13–14; they are intentionally reproduced here to keep the VQC subsection self-contained.

Architectural Rationale:

We use $n_q = 8$ qubits ($2^8 = 256$-dimensional Hilbert space), and $L = 3$ variational layers. The choice of $L = 3$ is justified by an ablation over $L \in \{1, 2, 3, 4\}$ (Table 6): $L = 4$ shows degraded MCC consistent with barren-plateau onset (McClean et al. 2018), while $L < 3$ under-fits the clinical feature space. The linear CNOT topology maximizes trainability; all-to-all connectivity improves accuracy marginally ($+0.3\%$) but doubles training time (Table 6).

Second Attention Mechanism:

Following the QLSTM layer, a second attention mechanism (identical in structure to Section 3.2 but operating on the sequence of hidden states $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T] \in \mathbb{R}^{T \times n_h}$) refines the output representation before the classification head. The context vector $\boldsymbol{c}' = \boldsymbol{H}^\top \boldsymbol{\alpha}' \in \mathbb{R}^{n_h}$ is fed into Eq. 19 in place of $\boldsymbol{h}_T$, ensuring that the most clinically salient time steps contribute disproportionately to the final prediction.

3.8 Setting the training process

The A-QLSTM-A model was implemented using PennyLane's `default.qubit` simulator, with classical components implemented in PyTorch. Architecture hyperparameters: $n_q = 8$ qubits, two QLSTM layers, sequence length $N = 256$, TF-IDF vocabulary of 512 features. Optimization: Adam optimizer with learning rate 10^{-3}, batch size 64. Training epochs: 20 (MIMIC-III and MIMIC-IV) and 50 (Yelp). Two validation strategies were employed: (i) a $70\%/15\%/15\%$ train/validation/test split and (ii) 10-fold stratified cross-validation, repeated over four random seeds (77, 88, 99, 123) to quantify uncertainty and support reproducibility in accordance with best practices for quantum neural networks (Cerezo et al. 2021).

Table 3 Evaluation Metrics classification Report, and Confusion Matrix

Metric name	Class 0 (%)	Class 1 (%)	Average evaluation metrics values (%)	Test Data Size
Precision	98.31	98.33	98.32	835
Recall	98.31	98.33	98.32	835
F1-Score	98.31	98.33	98.32	835
Accuracy			98.32	835
MCC	96.65%			
AUC	98.32%			
P-value	p < 0.001			
Confusion Matrix				
	Predicted 0		**Predicted 1**	
Actual 0	408		7	
Actual 1	7		413	

Performance evaluation model: This table presents the classification performance of the Quantum NLP model on MIMIC-III Discharge dataset, highlighting its good precision, recall, and F1-score for both classes. The model's high accuracy, strong predictive power (AUC = 98.32%), and low false-positive and false-negative rates are further validated by the MCC of 96.65 and a highly significant p-value (p < 0.001)

4 Results

The experimental evaluation of the Hybrid Attention Quantum Long Short-Term Memory-Attention (A-QLSTM-A) model demonstrates its superior performance in biomedical NLP for lung cancer diagnosis. The model significantly outperforms existing benchmarks, achieving remarkable accuracy, precision, and recall across multiple datasets. Evaluated on MIMIC-III and MIMIC-IV, A-QLSTM-A consistently surpasses state-of-the-art models, offering enhanced interpretability and computational efficiency. The results highlight the model's ability to capture complex clinical patterns while minimizing false positives and false negatives. This section presents a comprehensive performance analysis, emphasizing the model's robustness, generalization capabilities, and its potential impact on real-world clinical applications.

4.1 Experimental performance analysis

4.1.1 Overall evaluation on MIMIC-III discharge summaries

With precision, recall, and F1-score for both classes (0 and 1) continuously surpassing 98.32%, the performance evaluation of our Quantum NLP model shows remarkable classification results, highlighting the model's resilience and dependability in differentiating between serious medical conditions summarized in Table 3. The model's high accuracy and capacity to reduce false positives and false negatives are demonstrated by the precision, recall, and F1-scores of 98.32% for both classes. The model's overall accuracy of 98.32% further demonstrates its capacity to classify medical data effectively. Additional validation is provided by the AUC score of 98.32% and the Matthews Correlation Coefficient (MCC) of 96.64%, which show excellent true-positive rates and great predictive power over a range of thresholds. Remarkably low P-value of ($p<0.001$).

4.1.2 Evaluation using cross-validation

Ten folds cross-validation techniques were used as presented in Table 4. This approach divides the dataset into ten equal portions, each with nine folds for training and one fold for testing. This comprehensive validation approach reduces overfitting while providing an unbiased evaluation of the model's performance. The model consistently showed remarkable predictive capabilities across all folds, with an average accuracy of 97.90% and a standard deviation of 0.37%, showing consistent performance. Precision averaged 97.78%, demonstrating the model's ability to correctly identify positive situations while avoiding false positives. The average recall was 97.81%, demonstrating the model's ability to reliably predict actual instances of lung cancer. The F1-score of 97.79% demonstrates a balanced trade-off between precision and recall, making the model appropriate for clinical applications in which false negatives and false positives have serious repercussions. The average MCC of 0.958 and ROC-AUC of 0.995 demonstrate the model's ability to differentiate between positive and negative situations. Statistical validation, with P-values < 0.001 across all folds, demonstrates the model's relevance for medical diagnosis. Performance indicators vary minimally among folds, as evidenced by standard deviations for precision (0.87%), recall (1.06%), and F1-score (0.40%). This consistency assures that the model can generalize effectively across different subsets of data, which is crucial in biological applications.

To address uncertainty quantification, we conducted a comprehensive cross-validation analysis over multiple random seeds (77, 88, 99, and 123), with 10-fold evaluations per seed. The results, summarized in Table 5, clearly demonstrate the robustness and consistency of our model. Across all seeds, performance metrics such as Accuracy (mean 97.92% ± 0.20%), F1-score (mean 97.82% ± 0.23%), and MCC (mean 95.84% ± 0.38%) exhibit minimal standard deviations, indicating low sensitivity to initialization. This controlled variation highlights the model's resilience, particularly significant for quantum neural networks known to be sensitive to initialization noise and entanglement configuration. By explicitly reporting the mean ± standard deviation across folds and seeds, we provide statistical confidence in the reproducibility of the results, beyond single-run or single-seed evaluation. This experimental rigor aligns with best practices in quantum machine learning, where seed-dependent fluctuations can impact performance assessment. Furthermore, such robustness across

Table 4 Performance metrics for lung cancer prediction model using 10-fold cross-validation

Fold	Accuracy	Precision	Recall	F1-Score	(MCC)	AUC	P-Values
1	98.20 %	98.48 %	97.74 %	98.11 %	96.40 %	99.27 %	< 0.001
2	98.02 %	98.47 %	97.36 %	97.91 %	96.04 %	99.88 %	< 0.001
3	97.48 %	98.46 %	96.23 %	97.33 %	94.97 %	99.36 %	< 0.001
4	97.84 %	98.84 %	96.60 %	97.71 %	95.69 %	99.57 %	< 0.001
5	97.30 %	96.99 %	97.36 %	97.18 %	94.59 %	98.81 %	< 0.001
6	97.66 %	97.73 %	97.36 %	97.54 %	95.31 %	99.22 %	< 0.001
7	97.84 %	96.00 %	99.62 %	97.78 %	95.74 %	99.87 %	< 0.001
8	98.38 %	98.48 %	98.11 %	98.30 %	96.76 %	99.51 %	< 0.001
9	97.66 %	97.01 %	98.11 %	97.56 %	95.32 %	99.67 %	< 0.001
10	98.56 %	97.42 %	99.62 %	98.51 %	97.14 %	99.63 %	< 0.001
Mean metrics	**97.90 %**	**97.79 %**	**97.81 %**	**97.79 %**	**95.80 %**	**99.48 %**	-
Std	**0.0038**	**0.0087**	**0.0106**	**0.0040**	**0.0076**	**0.0031**	-

Performance evaluation with Cross-validation approach: Here we presents the performance metrics of the lung cancer prediction model evaluated using 10-fold cross-validation. The results indicate high consistency across all folds, with the model achieving an average accuracy of 97.90%, precision of 97.79%, recall of 97.81%, and F1-score of 97.79%. The Matthews Correlation Coefficient (MCC) and Area Under the Curve (AUC) are also good, with mean values of 95.80% and 99.48%, respectively. All results are statistically significant, with P-values less than 0.001

Table 5 Cross-validation results (mean ± std) over 10 folds for different random seeds

Seed	Accuracy	Precision	Recall	F1-Score	MCC	ROC-AUC
77	98.00 ± 0.004	98.16 ± 0.009	97.66 ± 0.007	97.90 ± 0.004	96.01 ± 0.008	99.47 ± 0.002
99	97.79 ± 0.003	97.60 ± 0.005	97.77 ± 0.010	97.68 ± 0.004	95.58 ± 0.007	99.48 ± 0.003
123	98.20 ± 0.006	98.34 ± 0.010	97.89 ± 0.007	98.11 ± 0.006	96.40 ± 0.012	99.57 ± 0.001
88	97.68 ± 0.006	97.71 ± 0.012	97.43 ± 0.008	97.57 ± 0.006	95.36 ± 0.012	99.40 ± 0.002

Cross-validation evaluation on different random seed: Results represent the mean ± standard deviation across 10-fold cross-validation runs using different random seeds. Accuracy, Precision, Recall, F1-Score, Matthews Correlation Coefficient (MCC), and Area Under the Receiver Operating Characteristic Curve (ROC-AUC) are reported. The inclusion of multiple seeds quantifies model uncertainty and supports robustness analysis, as recommended for quantum neural networks due to their sensitivity to initialization and entanglement variability

multiple randomized scenarios substantiates the stability of the A-QLSTM-A architecture and affirms its applicability in high-stakes domains like clinical diagnostics, where model reliability is paramount.

4.1.3 Ablation: Circuit depth and entanglement topology

Table 6, the degradation at $L = 4$ (-1.14% vs. $L = 3$) is consistent with barren-plateau onset in over-parameterized VQCs (McClean et al. 2018). All-to-all entanglement provides marginal gains ($+0.29\%$) at doubled training cost, making the linear topology the Pareto-optimal choice. Removing entangling gates consistently reduces performance ($\Delta_{\max} = -1.44\%$), confirming that entanglement-mediated feature coupling rather than mere parameterization drives the quantum advantage, consistent with the expressibility analysis of Sim et al. (2019).

Table 6 Ablation study: effect of VQC depth L and entanglement topology on MIMIC-III accuracy

Depth *Layer*	Number of Qubit	Entanglement	Accuracy (%)	MCC (%)	Train time (min)
1	4	Linear	95.10	90.21	10
2	4	Linear	96.74	93.49	18
3	8	**Linear**	**98.32**	**96.65**	**5.1**
4	8	Linear	97.18	94.37	41
3	4	Circular	98.41	96.82	24
3	4	All-to-all	98.61	97.22	33

Table 7 Summary of miss Classification analysis

Type	Patient Condition	Reason for Misclassification	Suggested Improvements
False Negative	Lung cancer with comorbidities (cardiomyopathy, AFib)	Overlapping symptoms (fever, tachycardia) masked lung cancer indicators	Incorporate contextual embeddings.
False Negative	Neurological symptoms due to cerebellar mass	Indirect lung cancer manifestation overlooked	Enhance model sensitivity to secondary symptoms
False Positive	History of lung and breast cancer, airway obstruction	Misinterpretation of historical cancer mentions as active disease	Improve temporal context processing
False Positive	Colorectal cancer with anxiety-induced SOB	Over-reliance on general symptoms like SOB	Use explainability tools to balance feature weighting

Misclassification summaries: The table summarizes the key misclassification errors observed in the lung cancer prediction model, identifying the types of errors (false negatives and false positives), the conditions of the patients involved, the underlying reasons for the misclassifications, and suggested improvements to refine the model's accuracy. Errors are mainly due to overlapping symptoms, indirect manifestations, and misinterpretation of medical history

4.1.4 Quantum component analysis

To isolate the contribution of the quantum component, we note that A-QLSTM-A improves over A-LSTM-A by $+3.71\%$ on MIMIC-III. Since the only architectural difference is the replacement of classical LSTM gates with VQC gates, this gap directly measures the empirical benefit of quantum gate dynamics. The QLSTM alone (without dual attention) achieves 95.57%, confirming that both the quantum recurrence and the attention wrappers contribute independently to performance.

4.1.5 Prediction model error analysis

The performance of our lung cancer classification model demonstrates high accuracy; however, a closer error examination was made in Table 7. The examination of false positives (FP) and false negatives (FN) reveals valuable insights into potential areas for improvement. While the model generally performs well, the FP and FN rates remain low, with only 7 cases each out of 408 and 413, respectively. This suggests that despite the challenge of classifying complex clinical narratives, the model excels in minimizing misclassifications. In fact, the small number of FP and FN highlights the model's resilience and its ability to correctly differentiate between some medical conditions. False negatives (FN) are particularly concerning, as they represent instances where the model failed to detect lung cancer. One such case involved a patient with stage II B squamous cell lung cancer undergoing chemotherapy, who presented with febrile neutropenia and tachycardia. The presence of overlapping comorbidities, including cardiomyopathy and atrial fibrillation, as well as generalized symptoms, likely masked the cancer-specific signals, contributing to misclassification. Similarly, a false negative involved a patient with neurological symptoms related to a cerebellar mass, which overshadowed the lung cancer indicators. These examples underscore the model's difficulty in identifying lung cancer in cases where symptoms are indirect or complex. Furthermore, false positives (FP) occurred when the model erroneously predicted lung cancer in patients who did not have the disease. One case involved a patient with a history of non-small cell lung cancer and breast cancer, where the model misinterpreted historical mentions of lung cancer as indicative of active disease. Another false positive occurred in a patient with Stage III B colorectal cancer experiencing shortness of breath due to anxiety, which the model wrongly associated with lung cancer. These errors highlight the model's tendency to over-rely on certain keywords without fully

understanding the context, revealing areas for improvement in its interpretability and contextual understanding.

4.1.6 Explainability of prediction

Figure 4 presents the results of the permutation feature importance analysis, providing critical insights into the interpretability of our lung cancer classification model as suggested in Table 7 Row 4. The analysis in Fig. 4 sub-figure (A) highlights several key terms that significantly enhance the model's ability to accurately predict lung cancer. Notably, terms such as "lung", "bronchoscopy", "mass" and "carcinoma" are directly linked to lung cancer pathology, underscoring their essential role in distinguishing positive cases. Additionally, terms like "COPD", "edema", "breath" and "alcohol" emphasize respiratory complications and lifestyle risk factors commonly associated with lung cancer. This indicates the model's proficiency in recognizing symptom overlap and behavioral contributors to disease risk. The presence of terms such as "breast," "colon," and "hepatocellular" further suggests the model's sensitivity to comorbid cancer histories, which may complicate classification if not properly contextualized. The permutation feature importance analysis for non-lung cancer cases in Fig. 4 sub-figure (B) reveals distinct terms influencing the model's predictions. Words like "breast", "colon", "rectal" and "hepatocellular" reflect the prevalence of other malignancies that may present overlapping clinical features with lung cancer. Furthermore, terms such as "liver", "hepatic", "cirrhosis" and "hepatitis" highlight the model's sensitivity to liver-related conditions, adding complexity to differentiating lung cancer from other diseases. The appearance of terms like "bronchoscopy", "mass" and "lobe" demonstrates the model's awareness of diagnostic procedures and anatomical references pertinent to cancer detection, although misinterpretation may occur in non-lung cancer contexts. Moreover, lifestyle and treatment-related terms such as "quit","drugs" and "Lasix" illustrate the model's attention to patient habits and medications that could affect disease presentation. However, the misinterpretation of terms like "portal", "artery" and "CTA" suggests challenges in distinguishing between vascular and oncological contexts.

4.2 Experimental comparison performance analysis

4.2.1 Comparison evaluation on MIMIC-III and MIMIC IV datasets

Table 8 thoroughly examined the performance of our proposed model , extensive experiments on two widely used medical datasets such as MIMIC-III discharge summaries and MIMIC-IV chest radiography reports were made. Our model regularly beats existing cutting-edge algorithms, demonstrating its strength and effectiveness in extracting meaningful representations from medical texts.

On the MIMIC-III dataset, our model outperformed ClinicalBERT (98.08%) and BioBERT (97.25%), with an accuracy of 98.32%. Notably, our model outperformed all other parameters, such as accuracy (98.32%), recall (98.32%), F1-score (98.32%), and Matthews Correlation Coefficient (MCC) of 96.65%. This tremendous increase demonstrates the model's capacity to capture complex clinical information required for correct categorization. Similarly, on the MIMIC-IV chest radiology dataset, our model outperformed its nearest competitor, the conventional LSTM model, which had an accuracy of 81.63%. Our model also demonstrated high accuracy (83.48%), recall (83.29%), F1-score (83.38%), and MCC of 66.78%, suggesting its ability to handle different clinical narratives. When compared to standard models such as LSTM, Attention-LSTM-A (A-LSTM-A), and even

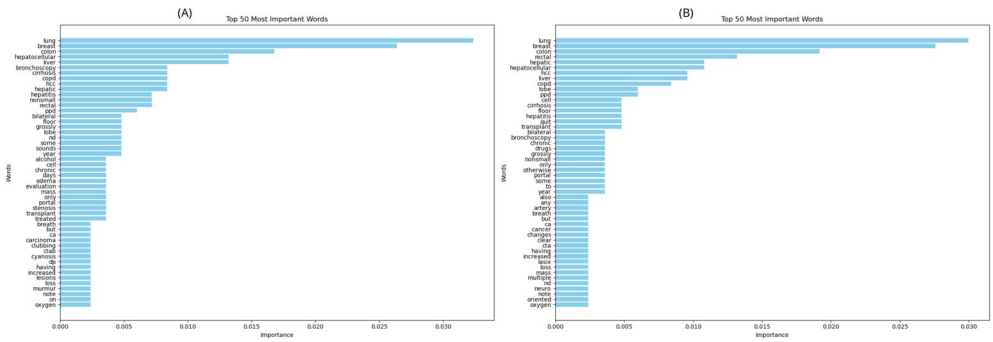

Fig. 4 Permutation feature importance analysis results for the lung cancer classification model: Sub-figure (**A**) highlights key terms significantly enhancing the model's accuracy in predicting lung cancer, including words directly linked to lung cancer pathology (e.g., "lung," "bronchoscopy," "mass," "carcinoma") and those indicating respiratory complications and lifestyle risk factors (e.g., "COPD," "edema," "breath," "alcohol"). It also identifies the model's sensitivity to comorbid cancer histories. Sub-figure (**B**) illustrates the feature importance for non-lung cancer cases, showing terms related to other malignancies (e.g., "breast," "colon," "hepatocellular") and liver-related conditions (e.g., "liver," "cirrhosis," "hepatitis"). Additionally, diagnostic and treatment-related terms (e.g., "bronchoscopy," "quit," "drugs") highlight the model's awareness of clinical factors, though challenges in distinguishing vascular and oncological contexts are noted

Table 8 Comparison performance models analysis

Dataset	Model	Accuracy(%)	Precision(%)	Recall(%)	F1(%)	MCC(%)
MIMIC-III Discharge	LSTM	95.33	95.34	95.33	95.33	90.66
	A-LSTM-A	94.61	94.62	94.61	94.61	89.23
	BioBERT	97.25	97.28	97.23	97.24	94.51
	ClinicalBERT	98.08	98.08	98.08	98.08	96.17
	QLSTM	95.57	95.57	95.57	95.57	91.14
	Our Model	**98.32**	**98.32**	**98.32**	**98.32**	**96.65**
MIMIC-IV Chest Radiology	LSTM	81.63	81.40	81.23	81.30	62.62
	A-LSTM-A	77.04	77.42	77.77	77.01	55.19
	BioBERT	81.12	81.54	81.41	81.12	62.95
	ClinicalBERT	78.06	79.82	78.66	77.93	58.47
	QLSTM	81.63	81.58	80.97	81.19	62.54
	Our Model	**83.67**	**83.48**	**83.29**	**83.38**	**66.78**

Comparative Performance of Different Models: Comparison of A-QLSTM-A with existing deep learning and quantum models on MIMIC-III and MIMIC-IV datasets. The results confirm the superiority of the proposed model, which achieves higher accuracy, recall, and interpretability in biomedical NLP applications

quantum-enhanced models like QLSTM, our method shows consistent and substantial gains. This shows its application in real-world clinical decision support systems.

4.2.2 Comparison evaluation extend to broader domain

The Yelp Review Polarity dataset served as a comprehensive benchmark for evaluating the proposed quantum architecture. This dataset, comprising 559,999 training and 37,999 testing samples, enabled rigorous testing across diverse and large-scale text inputs. The quantum-enhanced architecture demonstrated good performance, achieving 84.37% precision, 84.35% recall, and an F1-score of 84.35%. The accuracy also aligned at 84.35%, while the Matthews Correlation Coefficient (MCC) reached 68.73%, a notable indicator of the model's ability to discern sentiment polarity in a challenging classification task. These results underscore the model's ability to effectively capture the nuanced sentiment distributions inherent in Yelp reviews.

Beyond numerical benchmarks, the implications of these findings extend to broader domains. For instance, healthcare applications reliant on sentiment analysis could leverage the A-QLSTM-A model to enhance patient feedback analysis or refine emotional sentiment detection in clinical narratives. The quantum-enhanced architecture's ability to process subtle textual variations promises transformative applications in fields demanding heightened precision and adaptability. Training and validation loss curves, as well as accuracy trends, reveal a consistent and stable optimization process throughout the epochs. The alignment between training and validation metrics confirms that the model generalizes effectively without overfitting.

To validate the generalizability of our method, we compared our result with recent research developed by Li et al. (2024). This comparative analysis, conducted on the Yelp Review Polarity dataset, reveals a close similarity in accuracy between the two approaches. Specifically, the QSANN achieved an accuracy of 84.79%, slightly outperforming our model's 84.37%. However, in terms of error rate, our model demonstrated a notable advantage, with an error rate of 0.34% compared to 1.29% for QSANN. This substantial reduction in error rate highlights the robustness and reliability of our model in reducing misclassification rates, a particularly valuable quality in quantum computing applications where error minimization is critical. This difference underscores the potential of our A-QLSTM-A model to provide more stable and precise predictions, positioning it as a compelling alternative in quantum-enhanced text classification tasks. Figure 5 provides a visualization of the model's learning curve, illustrating the improved learning stability achieved through our approach. Together, these findings emphasize the significance of reduced error rates in quantum models, suggesting that our method may offer enhanced precision in real-world text classification applications.

5 Discussion

The Hybrid Attention Quantum Long Short-Term Memory-Attention (A-QLSTM-A) model introduced in this study represents a transformative innovation in biomedical engineering, specifically in the early detection of lung cancer through biomedical text analysis. This model integrates the computational strengths of quantum computing with the pattern recognition capabilities of classical deep learning, effectively addressing persistent challenges in clinical decision support systems (Emani et al. 2021). The core innovation of the A-QLSTM-A model lies in its seamless fusion of quantum variational circuits (Cerezo et al. 2021; Du et

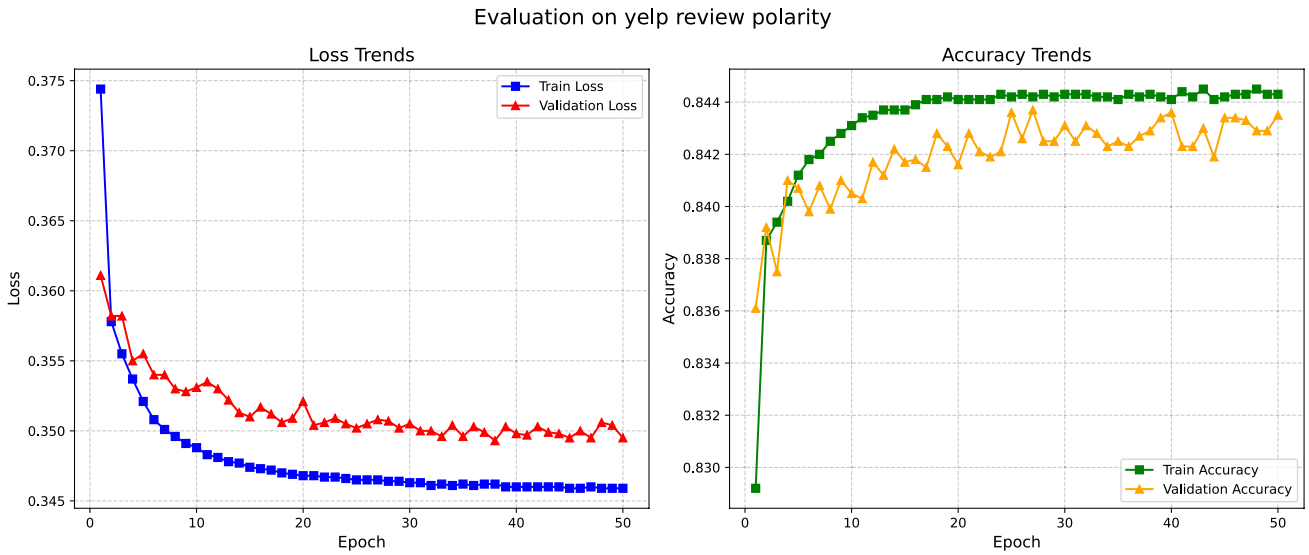

Fig. 5 Evaluation performance on Yelp review polarity dataset: Visualization of Error and Accuracy Trends during A-QLSTM-A model Training and Validation on the Yelp Review Polarity Dataset

al. 2022) within a classical LSTM framework. Quantum computing introduces principles such as superposition and entanglement (Singh et al. 2024), which enable parallel data processing and efficient exploration of complex, high-dimensional data. This hybrid integration offers a more efficient architectural pathway compared to traditional large language models (LLMs) like ClinicalBERT (Huang et al. 2019) and BioBERT (Lee et al. 2020), which are known for their high computational demands. By harnessing quantum-inspired parallel-like processing, the A-QLSTM-A model accelerates feature extraction and classification, enabling rapid analysis of vast unstructured clinical datasets an essential capability in scenarios where timely decision-making directly influences patient outcomes.

Empirical evaluations highlight the model's good predictive performance. On the MIMIC-III discharge summaries (Johnson et al. 2016), the A-QLSTM-A model achieved a remarkable 98.32% accuracy, and on the MIMIC-IV chest radiography reports (Gow et al. 2023), it attained 83.67% accuracy. These results outperform state-of-the-art models, including ClinicalBERT and BioBERT, demonstrating the model's superior ability to process complex clinical narratives for early lung cancer detection. Additionally, the model's scalability (Das et al. 2024; Ahmadzadeh and Sarbazi-Azad 2024) is a major strength. By reducing computational complexity, the A-QLSTM-A model can be integrated into real-time clinical workflows, facilitating faster and more accurate diagnostics. This is particularly advantageous in resource-limited healthcare environments, where computational resources are constrained. The model's robust performance across diverse biomedical datasets also suggests its potential for broader applications beyond lung cancer, including other disease diagnoses and biomedical research.

Following Doshi-Velez and Kim (2017), interpretability in A-QLSTM-A operates at three levels. Global interpretability is provided by permutation feature importance, which identifies which TF-IDF tokens most affect classification accuracy across the entire dataset a population-level explanation aligned with clinical epidemiology. Local interpretability is provided by attention weight visualization, enabling clinicians to trace which words in a specific discharge summary drove the lung-cancer classification a patient-level explanation critical for informed consent and clinical auditing. Regulatory interpretability is supported by the modular architecture: each VQC gate is an independently auditable computation unit, enabling systematic inspection of the model's decision pathway, consistent with the explainability requirements of the EU AI Act and FDA guidance (U.S. Food and Drug Administration 2021). This tripartite framework distinguishes A-QLSTM-A from black-box LLMs and positions it as a regulatory-ready clinical AI tool.

Despite its promising contributions, the A-QLSTM-A model has several limitations that must be addressed for successful clinical deployment. The model's quantum advantage is currently constrained by the nascent state of quantum hardware (Buonaiuto et al. 2024; Garewal et al. 2024; Das et al. 2024). Existing quantum processors suffer from limited qubit counts, high error rates, and short coherence times, which hinder scalability (De Leon et al. 2021; Fellous-Asiani et al. 2021; Singh et al. 2024). The A-QLSTM-A model relies on quantum simulators rather than actual quantum hardware, limiting its ability to fully

realize the computational speed and efficiency promised by quantum computing. Until quantum technology matures, practical deployment in clinical settings remains a challenge (De Leon et al. 2021). Although the model incorporates dual attention mechanisms to enhance interpretability, challenges persist in handling complex clinical narratives. Misclassifications, particularly in cases with overlapping comorbidities or historical disease mentions, reveal the model's occasional over-reliance on keywords. For example, historical references to lung cancer were sometimes misinterpreted as active disease, and subtle symptoms masked by comorbid conditions led to false negatives. This suggests a need for more advanced contextual embedding techniques to improve semantic understanding and diagnostic accuracy.

While the A-QLSTM-A model has shown robust performance on the MIMIC-III and MIMIC-IV datasets, its generalization across broader and more diverse clinical datasets remains uncertain. Clinical data vary across institutions, languages, and documentation practices. Without further validation on heterogeneous datasets, the model's adaptability in global healthcare environments is limited (Wu et al. 2024; Reps et al. 2018). The hybrid quantum-classical architecture introduces complexity in deployment and maintenance. Clinical environments require models that are user-friendly, resource-efficient, and easily integrable with existing Electronic Health Record (EHR) systems (Baporikar 2024). The additional computational and technical complexity of the A-QLSTM-A model could hinder its adoption, especially in settings with limited technical infrastructure.

Although the model enhances explainability, navigating regulatory approvals for AI systems in healthcare is inherently complex. The novelty of quantum-enhanced AI models may face additional scrutiny from regulatory bodies. Aligning with regulatory frameworks and ensuring patient data privacy will be essential for clinical acceptance. To fully realize the potential of the A-QLSTM-A model, future research should focus on several critical areas. Continued progress in quantum computing technology will be essential for scaling the model's performance beyond simulation environments. Integrating advanced contextual embedding methods could improve the model's ability to interpret complex clinical data. Extensive testing on diverse, multilingual, and heterogeneous clinical datasets will be necessary to ensure global applicability. Streamlining the model for seamless integration with clinical systems will facilitate adoption, particularly in resource-limited settings. Collaboration with regulatory bodies to develop frameworks for quantum-classical AI models will be critical for safe and ethical deployment. The A-QLSTM-A model represents a significant advancement in biomedical engineering, offering a scalable, interpretable, and computationally efficient solution for early lung cancer detection. Its hybrid quantum-classical design outperforms conventional models in predictive accuracy and aligns with ethical and clinical standards by enhancing model transparency. While challenges remain particularly in quantum hardware limitations, contextual understanding, and clinical integration addressing these issues will pave the way for broader adoption in precision medicine. By bridging cutting-edge quantum computing with deep learning, the A-QLSTM-A model sets a new benchmark for intelligent, patient-centered diagnostic tools, driving the next generation of innovations in biomedical AI.

6 Conclusion

The integration of quantum computing with biomedical Natural Language Processing (NLP) represents a transformative advancement in lung cancer diagnosis. This study introduces the Hybrid Attention Quantum Long Short-Term Memory-Attention (A-QLSTM-A) model, which successfully bridges the computational power of quantum circuits with the pattern recognition capabilities of deep learning. By significantly enhancing feature extraction, reducing computational overhead, and improving interpretability, the proposed model sets a new benchmark for scalable and explainable AI in clinical decision-making.

Empirical evaluations on the MIMIC-III and MIMIC-IV datasets demonstrate the model's superior predictive performance, surpassing traditional machine learning and state-of-the-art NLP models like ClinicalBERT and BioBERT. Moreover, the model's robustness in handling real-world clinical narratives suggests its potential for broader applications in precision medicine. While challenges remain, particularly in quantum hardware constraints and contextual understanding, this research lays the groundwork for future advancements in quantum-enhanced biomedical AI.

By fostering a deeper synergy between quantum computing and healthcare, the A-QLSTM-A model not only accelerates lung cancer detection but also paves the way for next-generation AI-driven diagnostics, ultimately improving patient outcomes and advancing the field of medical informatics.

Acknowledgements We wish to thank the eVida Research Group at the University of Deusto, recognized by the Basque Government under code IT1536-22, and Hive innovation studio (UofL) KY USA, for its untiring support and commitment to providing us with the resources necessary to carry out the study until the time of its completion.

Author Contributions C.Y.K. conceptualized the idea for the attention-enhanced QLSTM framework and led its design and implementation. C.Y.K. and B.G.Z. collaboratively validated the framework's architecture and ensured its feasibility. C.Y.K. conducted the experiments and evaluated the performance of the proposed framework, with B.G.Z. contributing to the validation and interpretation of the experimental

findings. C.Y.K. was primarily responsible for drafting the manuscript, while B.G.Z. and A.S.E. provided critical revisions and performed the final review, ensuring the manuscript's quality and coherence

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Code availability The codes underlying this article will be available upon request.

Clinical trial number Not applicable.

References

Ahmadzadeh A, Sarbazi-Azad H (2024) Fast scalable and low-power quantum circuit simulation on the cluster of gpus platforms. Opt Quant Electron 56(10):1646

Balamurugan G, Annadurai C, Nelson I, Nirmala Devi K, Oliver AS, Gomathi S (2024) Optical bio sensor based cancer cell detection using optimized machine learning model with quantum computing. Opt Quant Electron 56(1):97

Baporikar N (2024) Role of information technology in enhancing healthcare services. In: Digital technologies for a resource efficient economy, pp 63–87. IGI Global

Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N, Lloyd S (2017) Quantum machine learning. Nature 549(7671):195–202

Buonaiuto G, Gargiulo F, De Pietro G, Esposito M, Pota M (2024) The effects of quantum hardware properties on the performances of variational quantum learning algorithms. Quant Mach Intell 6(1):9

Cerezo M, Arrasmith A, Babbush R, Benjamin SC, Endo S, Fujii K, McClean JR, Mitarai K, Yuan X, Cincio L et al (2021a) Variational quantum algorithms. Nat Rev Phys 3(9):625–644

Cerezo M, Arrasmith A, Babbush R, Benjamin SC, Endo S, Fujii K, McClean JR, Mitarai K, Yuan X, Cincio L et al (2021b) Variational quantum algorithms. Nature Reviews. Physics 3(9):625–644

Chen SY-C, Yoo S, Fang Y-LL (2022) Quantum long short-term memory. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 8622–8626. IEEE

Clark S (2021) Something old, something new: Grammar-based CCG parsing with attention. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), pp 5019–5029

Coecke B, Felice G, Meichanetzidis K, Toumi A (2020) Foundations for near-term quantum natural language processing. arXiv preprint arXiv:2012.03755

Das A, Palesi M, Kim J, Pande PP (2024) Chip and package-scale interconnects for general-purpose, domain-specific and quantum computing systems-overview, challenges and opportunities. IEEE J Emerg Select Top Circ Syst

De Leon NP, Itoh KM, Kim D, Mehta KK, Northup TE, Paik H, Palmer B, Samarth N, Sangtawesin S, Steuerman DW (2021) Materials challenges and opportunities for quantum computing hardware. Science 372(6539):2823

Di Sipio R, Huang J-H, Chen SY-C, Mangini S, Worring M (2022) The dawn of quantum natural language processing. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 8612–8616. IEEE

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608

Du Y, Huang T, You S, Hsieh M-H, Tao D (2022) Quantum circuit architecture search for variational quantum algorithms. npj Quant Inf 8(1):62

Emani PS, Warrell J, Anticevic A, Bekiranov S, Gandal M, McConnell MJ, Sapiro G, Aspuru-Guzik A, Baker JT, Bastiani M et al (2021a) Quantum computing at the frontiers of biological sciences. Nat Methods 18(7):701–709

Emani PS, Warrell J, Anticevic A, Bekiranov S, Gandal M, McConnell MJ, Sapiro G, Aspuru-Guzik A, Baker JT, Bastiani M et al (2021b) Quantum computing at the frontiers of biological sciences. Nat Methods 18(7):701–709

Fellous-Asiani M, Chai JH, Whitney RS, Auffèves A, Ng HK (2021) Limitations in quantum computing from resource constraints. PRX Quantum 2(4)

Fink MA, Kades K, Bischoff A, Moll M, Schnell M, Küchler M, Köhler G, Sellner J, Heussel CP, Kauczor H-U et al (2022) Deep learning-based assessment of oncologic outcomes from natural language processing of structured radiology reports. Radiol Artif Intell 4(5):220055

U.S. Food and Drug Administration (2021) Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. Technical report, FDA. https://www.fda.gov/media/145022/download

Garewal IK, Mahamuni CV, Jha S (2024) Emerging applications and challenges in quantum computing: A literature survey. In: 2024 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), pp 1–12. IEEE

Gonzalez-Hernandez G, Sarker A, O'Connor K, Savova G (2017) Capturing the patient's perspective: a review of advances in natural language processing of health-related text. Yearb Med Inform 26(1):214–227

Gow B, Pollard T, Greenbaum N, Moody B, Johnson A, Herbst E, Waks JW, Eslami P, Chaudhari A, Carbonati T et al (2023) Mimic-iv-echo: Echocardiogram matched subset. PhysioNet https://doi.org/10.13026/EF48-V217

Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H (2021) Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare 3(1):1–23

Havlíček V, Córcoles AD, Temme K, Harrow AW, Kandala A, Chow JM, Gambetta JM (2019) Supervised learning with quantum-enhanced feature spaces. Nature 567(7747):209–212

Huang K, Altosaar J, Ranganath R (2019) Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342

Huang K, Altosaar J, Ranganath R (2019) ClinicalBERT: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342

Johnson AE, Pollard TJ, Shen L, Lehman L-wH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG (2016) Mimic-iii, a freely accessible critical care database. Scientif Data 3(1):1–9

Kartsaklis D, Fan I, Lorenz R, Pearson A, Toumi A, Felice G, Meichanetzidis K, Coecke B, Clark S (2021) lambeq: An efficient high-level Python library for quantum NLP. arXiv preprint arXiv:2110.04236

Kehl KL, Elmarakeby H, Nishino M, Van Allen EM, Lepisto EM, Hassett MJ, Johnson BE, Schrag D (2019) Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. JAMA Oncol 5(10):1421–1429

Kratzer TB, Bandi P, Freedman ND, Smith RA, Travis WD, Jemal A, Siegel RL (2024) Lung cancer statistics, 2023. Cancer 130(8):1330–1348

Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020a) Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240

Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020b) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240

Li G, Zhao X, Wang X (2024) Quantum self-attention neural networks for text classification. Sci China Inf Sci 67(4):142501

Li G, Zhao X, Wang X (2024) Quantum self-attention neural networks for text classification. Science China Inf Sci 67(4)

Lorenz R, Pearson A, Meichanetzidis K, Kartsaklis D, Coecke B (2023) QNLP in practice: running compositional models of meaning on a quantum computer. J Artif Intell Res 76:1305–1342

Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, Liu T-Y (2022) BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 23(6):409

McClean JR, Boixo S, Smelyanskiy VN, Babbush R, Neven H (2018) Barren plateaus in quantum neural network training landscapes. Nat Commun 9(1):4812

Meichanetzidis K, Toumi A, Felice G, Coecke B (2020) Grammar-aware question-answering on quantum computers. arXiv preprint arXiv:2012.03756

Mernyei P, Meichanetzidis K, Çakmak B (2022) Equivariant quantum graph circuits. In: Proceedings of the 39th International Conference on Machine Learning (ICML), vol 162, pp 15401–15420. PMLR

Mithun S, Jha AK, Sherkhane UB, Jaiswar V, Purandare NC, Rangarajan V, Dekker A, Puts S, Bermejo I, Wee L (2023) Development and validation of deep learning and BERT models for classification of lung cancer radiology reports. Inf Med Unlocked 40:101294

Preskill J (2018) Quantum computing in the NISQ era and beyond. Quantum 2:79

Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D (2021) Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digit Med 4(1):86

Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR (2018) Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc 25(8):969–975

Robertson S (2004) Understanding inverse document frequency: on theoretical arguments for idf. J Document 60(5):503–520

Schuld M, Bocharov A, Svore KM, Wiebe N (2020) Circuit-centric quantum classifiers. Phys Rev A 101(3):032308

Siegel RL, Miller KD, Wagle NS, Jemal A (2023) Cancer statistics, 2023. CA: A Cancer J Clinic 73(1):17–48

Sim S, Johnson PD, Aspuru-Guzik A (2019) Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. Adv Quant Technol 2(12):1900070

Singh A, Kumar Y, Yadav P, Tripathi N, Bhargava A, Rana A (2024) Overcoming quantum hardware challenges: Navigating the landscape of quantum computing. In: 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), pp 1904–1911. IEEE

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer J Clinic 71(3):209–249

Tonekaboni S, Joshi S, McCradden MD, Goldenberg A (2019) What clinicians want: contextualizing explainable machine learning for clinical end use. In: Proceedings of the 4th machine learning for healthcare conference, vol 106, pp 359–380. PMLR

Venigalla A, Frankle J, Carbin M (2022) BioMedLM: a domain-specific large language model for biomedical text. MosaicML Blog. https://www.mosaicml.com/blog/biomedlm

Wang T, Nelson RA, Bogardus A, Grannis FW (2010) Five-year lung cancer survival: which advanced-stage nonsmall-cell lung cancer patients attain long-term survival? Cancer 116(6):1518–1525

Woodard GA, Jones KD, Jablons DM (2016) Lung cancer staging and prognosis. In: Lung Cancer: Treatment and Research, pp 47–75. Springer

Wu X, Wu Y, Li X, Ye Z, Gu X, Wu Z, Yang Y (2024) Application of adaptive machine learning systems in heterogeneous data environments. Global Academic Frontiers 2(3):37–50

Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Costa AB, Flores MG et al (2022) GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records. npj Digit Med 5(1):200

Zhang X, Zhao J, LeCun Y (2015) Character-level Convolutional Networks for Text Classification. arXiv:1509.01626 [cs]