

Neural Network-Based Multimodal Fusion Model for Oncology Outcome Prediction: An Attention-Weighted Encoder-Fusion-Predictor Framework

Redeemer Pyagbara Dobe, Chima Godknows Igiri, Onate Egerton Taylor

Department of Computer Science, Faculty of Sciences, Rivers State University, Nkpulu-Oroworukwo, Port Harcourt, Nigeria
dobe.redeemer@rsu.edu.ng, igiri.chima@ust.edu.ng, taylor.onate@ust.edu.ng

Abstract:

This study developed a neural network-based multimodal fusion system for survival-risk prediction and auxiliary event-free classification. The system integrates histopathology feature vectors, RNA sequencing gene-expression data, and structured clinical variables using modality-specific encoders, attention-weighted fusion, a Cox survival-risk head, and an event-free classification head. The model was implemented in Python using PyTorch and evaluated using TCGA-BRCA, TCGA-LUAD, and a simulated local cohort. In TCGA-BRCA 5-fold cross-validation, the system achieved a mean concordance index (C-index) of 0.6573, classification accuracy of 0.8683, calibration error of 0.0271, and a worst single-missing-modality C-index drop of 0.0427. Genomic and histopathology modalities received the highest attention weights. However, transfer performance was weak on TCGA-LUAD and the simulated local cohort, with C-indices of 0.4682 and 0.4877, respectively. The findings support the model for breast cancer survival-risk prediction within TCGA-BRCA, while broader application requires cancer-specific retraining and validation using real Nigerian oncology data.

Keywords: multimodal data fusion; oncology prediction; attention mechanism; survival analysis; neural networks; TCGA-BRCA

1. INTRODUCTION

The care of persons diagnosed with malignant tumours requires precision in diagnosis, treatment planning, and prognostic assessment. In oncology settings, the availability of multiple kinds of patient data — radiological scans, histopathological slides, molecular assays, and electronic health records — has increased in recent years. This multiplicity of data sources creates both an opportunity to draw on diverse evidence and a challenge of combining heterogeneous information into a single, clinically meaningful output. Current analytic practice frequently treats each modality separately, limiting the value that could arise from integration. Multimodal data fusion responds to this limitation by constructing computational models that process more than one kind of data simultaneously, an approach associated in oncology with improved outcome prediction and treatment-response modelling (Wang et al., 2021).

Deep neural networks have become widely used for this purpose because they can extract latent features from high-dimensional data, including images, genomic sequences, and clinical records, and can learn patterns that are difficult to capture with traditional statistical methods. Several studies report that multimodal fusion models can outperform unimodal counterparts under specific evaluation conditions (Nakach et al., 2024); pan-cancer analyses integrating histology and genomic features suggest that combining tissue morphology with molecular data improves prognostic prediction relative to a single data source (Chen et al., 2022). Recent oncology-specific studies reinforce this pattern: a multicentre hepatocellular-

carcinoma study combining CT-derived deep-learning features with clinical data achieved concordance indices of 0.74 and 0.69 for overall and progression-free survival respectively (Xia et al., 2024), and a transformer-based multimodal architecture outperformed unimodal approaches for glioblastoma survival prediction (Gomaa et al., 2024).

Fusion strategies are commonly grouped into early, intermediate, and late fusion. Early fusion concatenates features from all modalities before prediction; late fusion combines independent per-modality outputs only at the decision stage; intermediate fusion processes each modality through a dedicated encoder before merging representations at a middle stage, offering a compromise between the two extremes (Baltrušaitis et al., 2019). The choice of strategy materially affects performance (Nakach et al., 2024), and each carries a distinct set of trade-offs: early fusion risks noise amplification in high-dimensional concatenated vectors, whereas late fusion may fail to capture deep cross-modal interactions.

Despite this promise, several operational barriers persist. Imaging, clinical, and molecular data follow different distributions, formats, and feature spaces, so effective fusion requires alignment, normalisation, and representation in a shared analytic space (Huang et al., 2020). Labelled oncology datasets are frequently small relative to model complexity, raising concerns about overfitting and generalisability (Kelly et al., 2019), and interpretability of fused models remains a barrier to clinical adoption because clinicians require transparency about how inputs drive predictions (Holzinger et al., 2019). In the

Nigerian healthcare environment specifically, resource constraints and data fragmentation often restrict the use of advanced analytic methods, motivating a design that can be evaluated explicitly under complete- and missing-modality conditions.

Existing multimodal oncology studies provide useful fusion models, but there remains limited documentation of a reproducible, modular reference architecture that integrates histopathology feature representations, molecular profiles, and structured clinical records for survival-risk and event-free outcome prediction under resource-constrained conditions. System-level evaluation methodologies that explicitly account for missing modalities, variable data quality, and resource limits remain incomplete, and the implementation requirements and technical barriers to deploying such systems in resource-constrained settings are not well characterised. This study addresses these gaps.

The aim of this study was to design and evaluate a neural network-based multimodal fusion system for oncology outcome prediction. The specific objectives were to: (i) design a neural network capable of integrating histopathology feature representations, molecular profile encodings, and clinical-record sequences through an attention-weighted fusion mechanism; (ii) develop a histopathology feature-encoding stream for fusion with the other modalities; (iii) implement the multimodal fusion model in Python with modality-specific encoders, attention-weighted fusion, and dual survival/classification prediction heads; and (iv) evaluate system performance using the concordance index, event-free classification accuracy, calibration, and missing-modality robustness metrics on publicly available oncology datasets.

The contribution of the study is twofold: it documents a modular encoder-fusion-predictor reference framework for integrating histopathology, genomic, and clinical data within a single neural network, and it provides a system-level evaluation workflow spanning cross-validation, cross-dataset transfer, temporal validation, modality ablation, and calibration analysis for testing such systems under complete-data and missing-modality conditions.

Multimodal representation theory holds that different data modalities — imaging, genomics, clinical records, and histology reflect related facets of the same underlying biological reality, and that a shared latent representation can unify these facets to improve predictive performance (Sui et al., 2012; Baltrušaitis et al., 2019). Because a tumour's manifestation spans macro-scale imaging, micro-scale histological architecture, molecular aberrations, and patient-level variables, no single modality is expected to capture disease heterogeneity fully; only joint representations that exploit cross-modal synergy are expected to yield stable predictions of prognosis or treatment response (Waqas et al., 2024). Early fusion concatenates modality embeddings prior to prediction (Cheerla & Gevaert, 2019; Nakach et al., 2024); intermediate fusion builds modality-specific encoders

before merging latent representations, balancing modality-specific learning with integration (Waqas et al., 2024); and late fusion combines independent per-modality predictions, which is useful when modalities differ greatly in scale or availability (Baltrušaitis et al., 2019). Empirical work shows that a poorly informative modality can degrade performance when fused indiscriminately with stronger modalities, so fusion should be selective rather than exhaustive (Li et al., 2022).

Deep learning theory, grounded in the universal approximation capability of sufficiently large neural networks (Goodfellow et al., 2016), motivates the use of hierarchical feature extraction for oncology data whose relationships among imaging, molecular, and clinical variables resist simple parametric description. Reviews of medical deep learning report that deep architectures outperform traditional pipelines across imaging modalities (Sistaninejhad et al., 2023), but also identify recurring weaknesses: inadequate annotated data, dataset bias, limited external validation, and poor interpretability (Egger et al., 2022; Holzinger et al., 2019). These weaknesses are amplified in low-resource clinical settings, where data volume, quality, and standardisation are often limited (Kelly et al., 2019), and where cross-population external validation is rarely reported (Liu et al., 2019; Nagendran et al., 2020).

Empirically, multimodal fusion has produced measurable gains in oncology survival prediction. A pan-cancer model integrating clinical, genomic, transcriptomic, and whole-slide histology data through unsupervised encoders achieved a concordance index of 0.78 (Chaudhary et al., 2018); a co-attention histology-genomic fusion model (PORPOISE) reported a concordance index of approximately 0.645 for breast cancer (Chen et al., 2022); and a hierarchical bilinear-fusion model (HFBSurv) reported approximately 0.766 on comparable histology-genomic tasks (Li et al., 2022). Reviews of multimodal machine learning in health report accuracy gains of 6–33% for fused models relative to unimodal baselines, depending on task and dataset (Soenksen et al., 2022; Waqas et al., 2024), supporting the general hypothesis that integrating complementary data channels reduces uncertainty relative to any single source (Zitnik et al., 2019). At the same time, reviews caution that most fusion studies use curated, high-income-country datasets with complete modalities, so real-world missingness, heterogeneous data quality, and limited external validation can erode the theoretical information gain of fusion (Waqas et al., 2024; Huang et al., 2020; Gimeno et al., 2023).

In sub-Saharan African contexts, structural barriers including delayed diagnosis, workforce shortages, and fragmented data infrastructure constrain both cancer care and the adoption of AI-based decision support (Lombe et al., 2023; Nnaji, 2024). Survey evidence from Nigerian oncology practice indicates cautious but growing interest in AI-assisted tools among practitioners, alongside concerns about data availability, validation, and trust (Olawade et al.,

2025). Taken together, the literature indicates that attention-weighted, modality-specific fusion architectures are theoretically motivated and empirically competitive with alternative fusion strategies, but that documented evaluation of such systems under missing-modality and resource-constrained conditions — particularly outside high-income-country data environments — remains limited. This gap motivates the modular, missing-modality-aware architecture developed in the present study.

2.0 Materials and Methods

The materials used and procedures followed in this study are described below

2.1 Research and Design Approach

The study adopted a design science research methodology, appropriate where the researcher identifies a practical problem, defines system objectives, designs and implements an artefact, evaluates its performance, and communicates the outcome (Hevner et al., 2004; Peffers et al., 2007; Gregor & Hevner, 2013). The artefact in this study is a modular multimodal neural network that integrates histopathology feature vectors, gene-expression features, and structured clinical variables for survival-risk prediction and auxiliary event-free classification. The design followed an encoder-fusion-predictor pattern: each modality is processed by a dedicated encoder before an attention-fusion module combines the resulting embeddings into a single patient-level representation, which is then passed to two prediction heads. This pattern was chosen because the three modalities differ structurally — histopathology and genomic inputs are fixed-length numerical vectors, whereas clinical data form a short ordered sequence — so a single, uniform encoding scheme would not represent all three adequately. Missing-modality robustness was treated as a first-class design requirement: modality masks were supported throughout, modality dropout was applied during training, and all-modality, unimodal, and pairwise-modality scenarios were evaluated after training (Section 3.4). The study is experimental and retrospective; no real-time clinical deployment was conducted. All evaluation was performed on publicly available TCGA-derived data and a simulated local cohort, so the method tests whether the proposed architecture can be implemented and evaluated reproducibly, rather than whether it is ready for clinical use.

2.2 Datasets

Three cohorts were used (Table 1). TCGA-BRCA served as the primary dataset for fitting preprocessing parameters, selecting genomic features, training the model, and evaluating the study targets. TCGA-LUAD served as a cross-dataset transfer-test cohort. A Local Cohort, generated from TCGA-BRCA distributional templates, was used for stress-testing under a simulated resource-constrained scenario; it is not a substitute for real local clinical data and its results are interpreted accordingly.

Table 1. Datasets used for evaluation.

Dataset	Patients	Event rate (%)	Median survival (months)	Event-free rate (%)	Role in study
TCGA-BRCA	759	12.25	19.42	87.75	Primary validation dataset
TCGA-LUAD	503	36.18	21.58	63.82	Cross-dataset transfer test
Local Cohort	150	41.33	24.73	58.67	Resource-constrained stress test

Each patient record comprised three modality groups plus outcome variables (Table 2). Histopathology was represented by stored 2,048-dimensional feature vectors, standardised with scaling parameters fitted on TCGA-BRCA. Genomic data comprised the top 1,000 variance-ranked RNA-seq genes selected from TCGA-BRCA, with missing-value handling and z-score standardisation. Clinical data were encoded as a 5 x 3 sequence: demographic variables (age, gender); pathological/T/N stage; ER, PR, and HER2 biomarker status; treatment- and metastasis-related indicators; and interaction terms (e.g., age-stage, ER-HER2, radiation-chemotherapy).

Table 2. Feature representation by modality.

Modality	Representation	Dimensions	Preprocessing
Histopathology	Stored feature vectors	2,048	Standardised using TCGA-BRCA-fitted scaling
Genomic	Top variance-ranked RNA-seq (HiSeqV2) genes	1,000	Variance ranking, missing-value handling, z-score standardisation
Clinical	Structured clinical sequence	5 x 3	Encoded clinical variables and interaction features

Standardisation of the genomic and histopathology feature matrices used z-score scaling, $x' = (x - \mu) / \sigma$, where μ and σ are the feature mean and standard deviation fitted on TCGA-BRCA and then applied unchanged to TCGA-LUAD and the Local Cohort, preserving separation between model fitting and transfer evaluation. Patient records were matched across modalities using the first twelve characters of the TCGA barcode, and records without usable survival outcomes were removed before modelling. The survival task used continuous survival time with a binary event indicator (0 = censored, 1 = observed death); the auxiliary classification task used a derived binary event-free label (1 = event-free at last follow-up). Direct treatment-response labels were not consistently available in the evaluated datasets and were therefore not used as a classification endpoint.

2.3 Model Architecture

The proposed model receives three input modalities and processes each with a dedicated encoder before attention-weighted fusion and dual-head prediction (Figure 1). Histopathology and genomic inputs, represented as fixed-length numerical vectors, are each encoded by a multilayer perceptron (2,048 $\rightarrow$ 512 $\rightarrow$ 256 and 1,000 $\rightarrow$ 512 $\rightarrow$ 256 respectively, each with batch normalisation, ReLU activation, and dropout). Clinical input, an ordered 5 x 3 sequence, is encoded by a two-layer bidirectional LSTM projected to a 256-dimensional embedding. The three 256-dimensional embeddings are stacked and passed through an attention scorer (256 $\rightarrow$ 64 $\rightarrow$ 1) followed by a softmax, yielding per-modality attention weights that are used to compute a weighted sum, i.e., the fused patient-level representation. This fused representation feeds two heads: a Cox proportional-hazards risk head producing a continuous survival-risk score, and a sigmoid event-free classification head producing an event-free probability. Table 3 summarises the architecture; layer-by-layer tensor dimensions are given in Figure 1.

Table 3. Summary of the multimodal model architecture.

Component	Architecture
Histopathology encoder	MLP 2048 $\rightarrow$ 512 $\rightarrow$ 256 with batch normalisation and dropout
Genomic encoder	MLP 1000 $\rightarrow$ 512 $\rightarrow$ 256 with batch normalisation and dropout
Clinical encoder	2-layer bidirectional LSTM over a 5 x 3 clinical sequence, projected to 256
Fusion module	Attention-weighted fusion over three 256-dimensional modality embeddings
Prediction heads	Cox proportional-hazards risk head; sigmoid event-free classification head

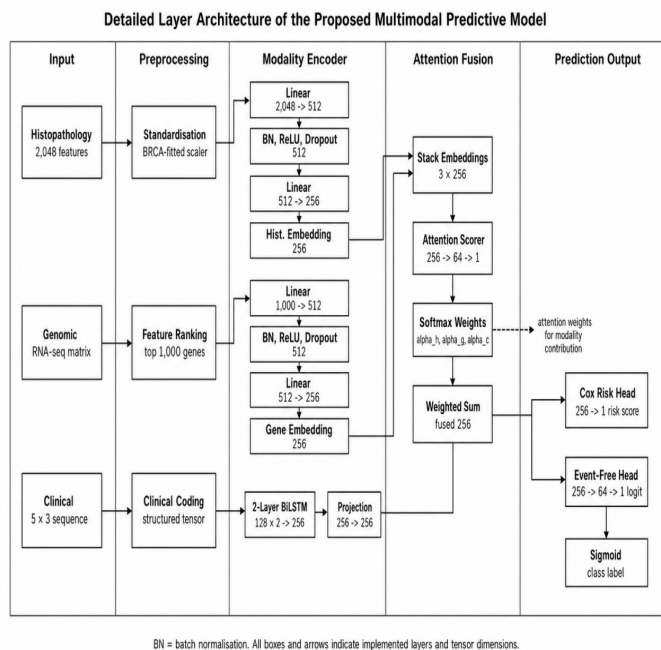


Figure 1. Detailed layer architecture of the proposed attention-weighted encoder-fusion-predictor model. BN = batch normalisation.

The attention-weighted fusion mechanism was chosen in preference to simple feature concatenation because concatenation treats all modalities as equally important, provides no explicit modality weighting for interpretability, assumes complete data availability across modalities, and can increase overfitting risk by joining all features into one high-dimensional vector when the number of aligned multimodal records is limited. Attention-weighted fusion instead learns a patient-specific contribution for each modality and remains defined when a modality is masked as unavailable, which supports both interpretability and missing-modality robustness.

2.4 Training and Missing-Modality Simulation

The model was trained with the Adam optimiser using a joint objective combining a Cox partial-likelihood loss for the survival-risk head and a weighted binary cross-entropy loss for the event-free classification head. Class imbalance in the event-free label was addressed through positive-class weighting derived from the training-fold event distribution. During training, modality dropout was applied stochastically: for a subset of training batches, one of the three modality masks was set to zero with fixed probability, exposing the model to single-missing-modality conditions during optimisation and reflecting incomplete data availability that can occur in practice.

Post-hoc probability calibration was fitted on validation-fold logits using logistic (Platt) scaling (Platt, 1999) and applied to test-fold predictions before computing classification and calibration metrics.

2.5 Experimental Protocol and Evaluation Metrics

Performance was assessed through five complementary procedures. (1) Five-fold stratified cross-validation on TCGA-BRCA, with an internal 80/20 train/validation split within each training fold used for model selection and probability calibration, and out-of-fold predictions used for pooled evaluation. (2) Cross-dataset transfer testing, in which the BRCA-trained model (with TCGA-BRCA-fitted preprocessing parameters) was applied unchanged to TCGA-LUAD and to the simulated Local Cohort. (3) Temporal validation, training on TCGA-BRCA patients diagnosed before 2010 and testing on patients diagnosed from 2010 onward. (4) Modality ablation, evaluating the all-modality configuration against three single-modality and three pairwise-modality configurations by masking the corresponding encoder inputs at inference. (5) Attention-weight analysis, summarising the softmax attention weights assigned to each modality across the evaluation set.

Survival discrimination was evaluated using Harrell's concordance index (C-index) (Harrell et al., 1982), computed on the Cox risk score against observed survival time and event status. Event-free classification was evaluated using accuracy, precision, recall, F1-score, and area under the receiver-operating-characteristic curve (AUC-ROC), with the decision threshold selected to optimise the F1-score on validation-fold predictions. Calibration was evaluated as the mean absolute difference

between predicted and observed event-free probability across probability bins. Missing-modality robustness was evaluated as the drop in C-index for each single-missing-modality configuration relative to the all-modality configuration. Pre-specified performance targets, set prior to evaluation, were: C-index ≥ 0.65 ; classification accuracy ≥ 0.80 ; calibration error ≤ 0.15 ; and worst single-missing-modality C-index drop ≤ 0.05 .

The implementation used Python with PyTorch for model definition and training, scikit-learn for auxiliary metrics and calibration, and pandas/NumPy for data handling; the codebase was organised into separate modules for data loading, preprocessing, model definition, training, evaluation, and visualisation to support reproducibility. The system was designed to run on standard computing hardware without a dedicated GPU, consistent with the study's interest in feasibility under resource-constrained conditions.

2.6 Ethical Considerations

The study used de-identified, publicly available TCGA data and a synthetic simulated cohort; no identifiable patient information was collected or processed, and no primary human-subjects data collection was undertaken. Because the system was evaluated retrospectively and was not deployed clinically, findings are reported strictly as evidence of technical feasibility and should not be interpreted as evidence of clinical validity.

3.0 Results

The aligned TCGA-BRCA cohort contained 759 patients with an observed death-event rate of 12.25% and a median survival of 19.42 months. TCGA-LUAD contained 503 patients with a higher event rate of 36.18%, and the simulated Local Cohort contained 150 patients with an event rate of 41.33%. These differences in event rate and survival distribution are relevant to interpreting subsequent cross-dataset comparisons.

3.1 Cross-Validation Performance on TCGA-BRCA

Table 4 presents five-fold cross-validation results for the primary TCGA-BRCA evaluation. The model achieved a mean C-index of 0.6573, mean classification accuracy of 0.8683, mean precision of 0.8773, mean recall of 0.9880, mean F1-score of 0.9293, mean AUC-ROC of 0.6473, and mean calibration error of 0.0271. The high recall indicates that the classifier identified most event-free cases; the pooled confusion matrix (658 true positives, 92 false positives, 8 false negatives, 1 true negative) shows that observed death-event cases were harder to separate, consistent with the low TCGA-BRCA event rate.

Table 4. TCGA-BRCA five-fold cross-validation results.

Fold	C-index	Accuracy	Precision	Recall	F1-score	AUC-ROC	Calib. error
1	0.6328	0.8487	0.8776	0.9627	0.9181	0.5916	0.0441
2	0.6253	0.8618	0.8784	0.9774	0.9253	0.6858	0.0277
3	0.6224	0.8750	0.8750	1.0000	0.9333	0.6656	0.0023
4	0.6993	0.8750	0.8750	1.0000	0.9333	0.6601	0.0242
5	0.7067	0.8808	0.8808	1.0000	0.9366	0.6333	0.0372
Mean	0.6573	0.8683	0.8773	0.9880	0.9293	0.6473	0.0271

3.2 Cross-Dataset Transfer and Resource-Constrained Stress Testing

Table 5 and Figure 2 compare performance across the three evaluation datasets using the BRCA-trained model and BRCA-fitted preprocessing parameters. Performance declined substantially outside TCGA-BRCA: C-index fell to 0.4682 on TCGA-LUAD and 0.4877 on the Local Cohort, both close to chance-level discrimination, with correspondingly higher calibration error (0.2915 and 0.3160 respectively).

Table 5. Performance across evaluation datasets.

Dataset	C-index	Accuracy	Precision	Recall	F1-score	Calib. error
TCGA-BRCA (5-fold CV mean)	0.6573	0.8683	0.8773	0.9880	0.9293	0.0271
TCGA-LUAD (transfer)	0.4682	0.6243	0.6352	0.9657	0.7664	0.2915
Local Cohort (transfer)	0.4877	0.5867	0.5867	1.0000	0.7395	0.3160

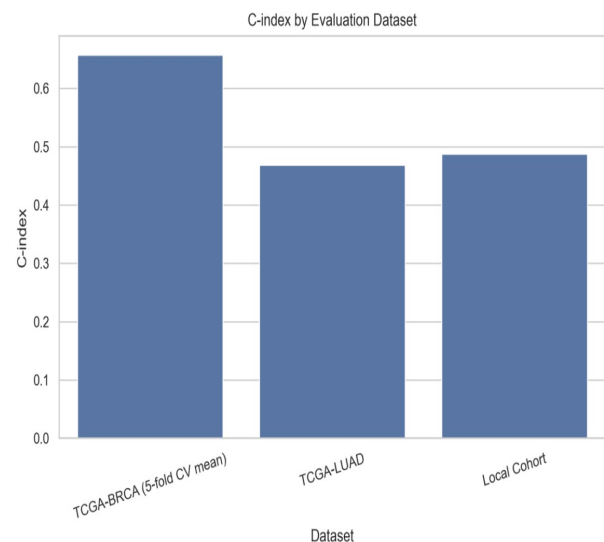


Figure 2. Concordance index (C-index) by evaluation dataset.

3.3 Modality Ablation

Table 6 reports performance under different modality-availability settings, obtained by masking encoder inputs at inference. The worst single-missing-modality C-index drop was 0.0427 (histopathology + clinical, i.e., genomic

masked), below the pre-specified 0.05 threshold, indicating that the attention-fusion mechanism together with modality-dropout training preserved survival-ranking performance when one modality was unavailable. The clinical-only configuration produced a markedly larger C-index drop of 0.2269, indicating that the simplified five-step clinical sequence was not, by itself, a sufficient survival predictor in this implementation, even though it achieved the highest standalone classification accuracy (0.8775) among single-modality settings.

Table 6. Modality ablation results.

Modality setting	Available modalities	C-index	Accuracy	C-index drop
All modalities	Histopathology + genomic + clinical	0.9052	0.7839	0.0000
Histopathology only	Histopathology	0.8612	0.6759	0.0440
Genomic only	Genomic	0.8758	0.7246	0.0294
Clinical only	Clinical	0.6783	0.8775	0.2269
Histopathology + genomic	Histopathology + genomic	0.9042	0.7536	0.0010
Histopathology + clinical	Histopathology + clinical	0.8625	0.7813	0.0427
Genomic + clinical	Genomic + clinical	0.8748	0.7602	0.0304

3.4 Attention Weight Analysis

Table 7 and Figure 3 summarise the learned attention weights. The genomic modality received the highest mean weight (0.4499), followed by histopathology (0.4010), with clinical data receiving the lowest mean weight (0.1491). This pattern indicates that, in the fitted TCGA-BRCA model, molecular and tissue-derived representations carried more predictive signal than the simplified clinical sequence, while all three modalities contributed to the fused representation.

Table 7. Attention weight distribution across modalities.

Modality	Mean attention weight	Standard deviation
Histopathology	0.4010	0.1030
Genomic	0.4499	0.1004
Clinical	0.1491	0.0335

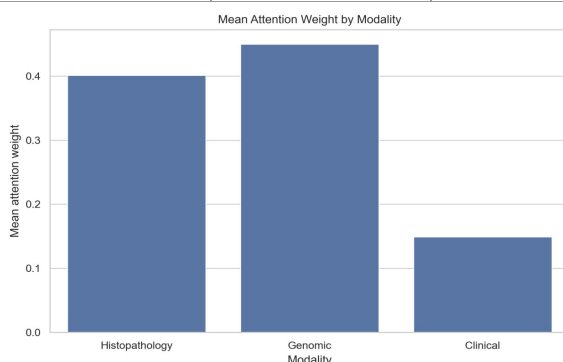


Figure 3. Mean learned attention weight by modality.

3.5 Temporal Validation

Training on 538 patients diagnosed before 2010 and testing on 221 patients diagnosed from 2010 onward produced a test C-index of 0.8198 (training C-index 0.9138), indicating that survival-risk ranking remained strong across this temporal split within TCGA-BRCA. This finding supports the internal stability of the ranking component but, being retrospective and dataset-internal, does not substitute for prospective external validation.

3.6 Performance Against Pre-Specified Targets

Table 8 compares the primary TCGA-BRCA cross-validation results with the performance targets defined in Section 3.5. All four targets were met.

Table 8. Observed performance against pre-specified targets.

Target	Threshold	Observed value	Outcome
C-index	0.65	0.6573	Met
Classification accuracy	0.80	0.8683	Met
Calibration error	0.15	0.0271	Met
Single-modality C-index drop	0.05	0.0427	Met

4.0 Discussion

The results obtained in this study are discussed below.

4.1 Interpretation of Findings

The primary result of the study was the TCGA-BRCA cross-validated C-index of 0.6573, exceeding the pre-specified target of 0.65 and indicating that the fused model ranked patients by survival risk better than chance, within a range consistent with related single-cancer multimodal survival prediction work. The event-free classification results — accuracy 0.8683, precision 0.8773, recall 0.9880, F1-score 0.9293 — should be interpreted alongside the pooled confusion matrix rather than as a standalone measure, because the low TCGA-BRCA event rate (12.25%) produces an imbalanced classification problem in which event-free cases dominate. Post-hoc calibration reduced calibration error to 0.0271, well below the 0.15 target, indicating that predicted event-free probabilities were reasonably aligned with observed frequencies within the cross-validation setting.

The attention-weight analysis indicates that the model did not treat the three modalities equally: genomic and histopathology representations carried more predictive signal than the simplified clinical sequence. This is consistent with the modality-ablation results, in which the clinical-only configuration produced the largest C-index drop (0.2269), suggesting that the current five-step clinical encoding — while sufficient as a contributing signal within the fused model — does not by itself capture the complexity of longitudinal clinical information such as treatment timelines, comorbidities, and laboratory results. The finding does not imply that clinical data are unimportant; rather, it identifies a specific representational limitation to be addressed in future work.

The missing-modality design requirement was supported by the ablation results: the worst single-missing-modality C-index drop (0.0427) remained below the 0.05 threshold, showing that attention-weighted fusion combined with modality-dropout training allowed the model to tolerate the loss of any one modality. This property is directly relevant to resource-constrained oncology settings, including many Nigerian clinical environments, where complete histopathology, molecular, and clinical data are not always available for every patient.

Cross-dataset transfer was limited: applying the BRCA-trained model to TCGA-LUAD and to the simulated Local Cohort produced C-index values of 0.4682 and 0.4877 respectively, close to chance-level discrimination. This is an expected consequence of the differences in tumour biology, molecular drivers, treatment patterns, and event rates between breast and lung adenocarcinoma, and indicates that the learned representations were cancer-type specific rather than generally transferable. Temporal validation, by contrast, produced a substantially higher C-index (0.8198) on the 2010-onward TCGA-BRCA subset, supporting stability within the training cancer type even though this remains a retrospective, dataset-internal test. The simulated Local Cohort result should be read as evidence about the evaluation pipeline's behaviour under a stress-test scenario rather than as evidence about performance on real Nigerian oncology data, which the synthetic cohort cannot fully represent.

4.2 Comparison with Existing Systems

Table 9 situates the best comparable result from this study — the TCGA-BRCA C-index of 0.6573 — against selected multimodal oncology survival-prediction systems reported in the literature. The proposed system performed within the lower-to-middle range of these comparators: its result was close to the BRCA-specific figure reported for the co-attention PORPOISE model (Chen et al., 2022) but lower than pan-cancer and more complex fusion approaches (Cheerla & Gevaert, 2019; Vale-Silva & Rohr, 2021; Li et al., 2022). These comparisons should be treated cautiously given differences in datasets, feature-extraction pipelines, validation protocols, and cancer-type scope across studies.

Table 9. Comparison of the proposed system with existing multimodal oncology survival-prediction systems.

Study/system	Main approach	Data modality	Reported C-index	Comparison
Cheerla & Gevaert (2019)	Multimodal deep learning, pan-cancer survival	Histology + genomic + clinical	~0.78	Higher, broader pan-cancer protocol
Vale-Silva & Rohr (2021)	Multimodal pan-cancer survival modelling	Histology + genomic + clinical	~0.82	Higher, different data/protocol
Chen et al. (2022), PORPOISE	Co-attention multimodal fusion	Whole-slide image + genomic	~0.645	Comparable to proposed system

Study/system	Main approach	Data modality	Reported C-index	Comparison
Li et al. (2022), HFBSurv	Hierarchical bilinear fusion	Histology + genomic	~0.766	Higher, more complex fusion
Proposed system	Attention-weighted encoder-fusion-predictor	Histopathology + genomic + clinical	0.6573	Meets target; interpretable weights

The principal contribution of the proposed system is not that it outperforms these comparators, but that it provides a modular, reproducible encoder-fusion-predictor implementation that integrates three data modalities, reports interpretable attention weights, and explicitly evaluates missing-modality behaviour — characteristics that are directly relevant where data completeness varies across patients and institutions. The comparison also points to specific avenues for improvement: higher-performing systems in the literature typically use pathology-specific feature extractors, more expressive cross-modal interaction mechanisms (e.g., bilinear or cross-attention fusion), or biologically informed genomic representations, any of which could plausibly improve on the generic feature representations used here.

4.3 Limitations

Several limitations qualify these findings. Histopathology was represented using stored, pre-extracted 2,048-dimensional feature vectors rather than raw whole-slide images processed through a pathology-specific encoder, which may understate the achievable contribution of the histopathology modality. Genomic features were selected by variance ranking rather than by biologically curated gene panels or pathway scores. The clinical representation was intentionally simplified to five structured steps and does not include treatment timelines, comorbidities, or longitudinal follow-up. The Local Cohort was synthetic and cannot substitute for real local clinical data. Finally, all evaluation was retrospective and dataset-internal; no prospective or externally sourced clinical validation was performed, and the system was not evaluated as part of a clinical workflow.

5.0 Conclusion and Future Work

This study designed, implemented, and evaluated a modular, attention-weighted, multimodal neural network for oncology outcome prediction, integrating histopathology feature representations, genomic expression data, and structured clinical records through modality-specific encoders and an attention-fusion mechanism with dual Cox survival-risk and event-free classification heads. On the primary TCGA-BRCA validation cohort, the system met all four pre-specified performance targets: C-index (0.6573 vs. target 0.65), classification accuracy (0.8683 vs. target 0.80), calibration error (0.0271 vs. target 0.15), and

worst single-missing-modality C-index drop (0.0427 vs. target 0.05). Attention analysis showed that genomic and histopathology representations contributed more strongly than the simplified clinical sequence, and modality-ablation analysis showed that the system tolerated the loss of any single modality within the pre-specified threshold, a property relevant to resource-constrained oncology settings.

At the same time, cross-dataset transfer to TCGA-LUAD (C-index 0.4682) and to a simulated Local Cohort (C-index 0.4877) was weak, indicating that the learned representations were specific to the training cancer type and setting. The system should therefore be regarded as a defensible, target-meeting prototype for breast-cancer outcome prediction on public multimodal data, and not as a validated, general-purpose, or pan-cancer clinical tool.

Future work should prioritise five directions: (i) validating the system on real Nigerian and other resource-constrained oncology cohorts, since the simulated Local Cohort cannot represent the full biological, clinical, infrastructural, and follow-up characteristics of any real population; (ii) replacing stored generic histopathology feature vectors with pathology-specific foundation-model or CNN-derived embeddings; (iii) refining genomic feature selection using biologically informed gene signatures, molecular subtype markers, or curated cancer gene panels; (iv) expanding the clinical representation to include treatment timelines, comorbidities, laboratory results, performance status, and follow-up history; and (v) evaluating cancer-type-specific retraining or domain adaptation, together with alternative fusion mechanisms such as cross-attention or bilinear fusion, before the system is applied outside TCGA-BRCA. These contributions are intended to support the broader goal of data-driven, equitable oncology decision support consistent with Sustainable Development Goal 3 (health and well-being), while underscoring that cancer-type-specific and setting-specific validation remains a prerequisite for clinical use.

Declarations

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of interest: The authors declare no conflict of interest.

Data availability: The TCGA-BRCA and TCGA-LUAD datasets analysed in this study are publicly available via the UCSC Xena platform (Goldman et al., 2020). The simulated Local Cohort was generated for this study from TCGA-BRCA distributional templates and is available from the corresponding author on reasonable request.

Ethical approval: This study used de-identified, publicly available data and a synthetic dataset; it did not involve new data collection from human participants.

References

- Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423-443. <https://doi.org/10.1109/TPAMI.2018.2798607>
- Chaudhary, K., Poirion, O. B., Lu, L., & Garmire, L. X. (2018). Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, 24(6), 1248-1259.
- Cheerla, A., & Gevaert, O. (2019). Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*, 35(14), i446-i454.
- Chen, R. J., Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Noor, Z., & Mahmood, F. (2022). Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8), 876-885.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187-202.
- Egger, J., Gsaxner, C., Pepe, A., Pomykala, K. L., Jonske, F., Kurber, M., & Chen, X. (2022). Medical deep learning: A systematic meta-review. *Computer Methods and Programs in Biomedicine*, 221, 106874.
- Gimeno, M., Sada del Real, K., & Rubio, A. (2023). Precision oncology: A review to assess interpretability in several explainable methods. *Briefings in Bioinformatics*, 24(4), bbad200. <https://doi.org/10.1093/bib/bbad200>
- Goldman, M. J., Craft, B., Hastie, M., Repecka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., Zhu, J., & Haussler, D. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology*, 38, 675-678. <https://doi.org/10.1038/s41587-020-0546-8>
- Gomaa, M., Ba-Masq, A., Al-Zahrani, M., Al-Turaiki, I., & Al-Razgan, M. (2024). Comprehensive multimodal deep learning survival prediction enabled by a transformer architecture: A multicenter study in glioblastoma. *Cancers*, 16(5), 948.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2), 337-355.
- Harrell, F. E., Jr., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, 247(18), 2543-2546.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75-105.
- Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Muller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312.
- Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2020). Fusion of medical imaging and

electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digital Medicine*, 3(1), 136. <https://doi.org/10.1038/s41746-020-00341-z>

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1), 1-9.

Li, Z., Wu, F., Li, S., & Wang, J. (2022). HFBSurv: Hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. *Bioinformatics*, 38(9), 2587-2594.

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., & Keane, P. A. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet Digital Health*, 1(6), e271-e297.

Lombe, D. C., Mwamba, M., Msadabwe, S., Bond, V., Simwinga, M., Ssemata, A. S., Muhumuza, R., Seeley, J., Mwaka, A. D., & Aggarwal, A. (2023). Delays in seeking, reaching and access to quality cancer care in sub-Saharan Africa: A systematic review. *BMJ Open*, 13(4), e067715. <https://doi.org/10.1136/bmjopen-2022-067715>

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., & Maruthappu, M. (2020).

Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*, 368, m689.

Nakach, F.-Z., Idri, A., & Goceri, E. (2024). A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification. *Artificial Intelligence Review*, 57, Article 327. <https://doi.org/10.1007/s10462-024-10984-z>

Nnaji, N. D. (2024). Impact and challenges of artificial intelligence integration in the African health sector: A review. *Trends in Medical Research*, 19(1), 220-235.

Olawade, D. B., Ojo, I. O., Oisakede, E. O., Joel-Medewase, V. I., & Wada, O. Z. (2025). Artificial intelligence in Nigerian oncology practice: A qualitative exploration of oncologists' perspectives. *Journal of Cancer Policy*, Article 100626. <https://doi.org/10.1016/j.jcpo.2025.100626>

Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45-77.