ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

"Impact-X: A Causally-Grounded Interpretable Multimodal Deep Learning Framework for Transparent Early Disease Detection Using Imaging, Clinical, and Genomic Data."

Tunan Shikder Any [1], Ananya Manna [2] , MD SARWAR ISLAM [3] , Anu Priya Yaduvanshi [4] , Shreyanjan Neogi [5] , Addita Rani Dash[6], Turjoy Saha[7]

[1]Department of Electronics Engineering KIIT University, Bhubaneswar, Odisha, India

[2]Department of Computer Science Engineering KIIT University, Bhubaneswar, Odisha, India

[3]Department of Computer Science Engineering KIIT University, Bhubaneswar, Odisha, India

[4]Department of Computer Science Engineering KIIT University, Bhubaneswar, Odisha, India

[5]Department of Computer Science Engineering KIIT University, Bhubaneswar, Odisha, India

[6]Department of Computer Science Engineering KIIT University, Bhubaneswar, Odisha, India

[7]Department of Computer Science Engineering KIIT University, Bhubaneswar, Odisha, India

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150600149

Received: 01 July 2026; Accepted: 02 July 2026; Published: 17 July 2026

ABSTRACT

Recent developments in multimodal deep learning have brought great progress to early disease detection; yet, wide-scale implementation of such models in clinics is hindered by the inherently inscrutable reasoning of existing methods. Current frameworks often employ post-hoc explanations that are not cross-modal consistent and are unable to disambiguate between causality and correlation, compromising both clinician trust and patient safety. In order to resolve these key issues, we introduce IMPACT-X, a novel Causally-Grounded Interpretable Multimodal Deep Learning Framework. IMPACT-X fuses mul-tiple heterogeneous modalities—medical imaging with Vision Transformers, medical records with Tabular Transformers, and genetic sequences with Graph Neural Networks—into a single and interpretable model.

Our framework includes a novel Causal Multimodal Fusion Layer (CMFL) which leverages cross-modal attention alignment in order to align the representation in a dynamic manner. Fur-thermore, an SCM module with DAG learning capabilities helps identify latent confounders and ensures the causally-consistent nature of the predictions. An uncertainty-aware decision-making layer estimates epistemic uncertainty through Monte Carlo Dropout in order to produce confidence scores. A unique cross-modal interpretability alignment loss function ensures coherent explanations across multiple modalities. The experimental results show that IMPACT-X achieves an SOTA performance with AUC-ROC score of 0.94, beating the best black-box baseline by 5.2%. Quantitative evaluation shows that IMPACT-X is 40% better in terms of faithfulness than traditional attention mechanism-based explanation approaches. A qualitative study with practicing medical professionals shows the benefits of causality-grounded predictions by increasing the level of physician trust in the system output. With its combination of high prediction accuracy and causal interpretability, IMPACT-X can pave the way for the development of a regulatory compliant and interpretable paradigm of medical AI that can safely be implemented in clinics, while enabling more accurate personalized medicine practices.*Index Terms*—Multimodal Deep Learning; Causal Inference; Interpretability; Early Disease Detection; Clinical Decision Sup-port; Genomic Integration

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

SUMMARY

Artificial intelligence-based healthcare diagnostics have the potential for revolutionized disease diagnosis; however, the proliferation of the current black-box approach is significantly limiting its adoption due to poor explainability. To address the key challenge of achieving both high prediction performance and interpretability of the AI system, this paper proposes IMPACT-X, a novel approach, which incorporates causal grounding to solve this problem. Unlike current multimodal approaches, which use correlational associations between dif-ferent modalities, IMPACT-X leverages the Causal Multimodal Fusion Layer (CMFL), which incorporates diverse sources of data such as medical images, electronic health records, and genomic data while explicitly modeling the causal relation-ships using Structural Causal Models (SCMs). The proposed model introduces a cross-modal attention consistency loss to guarantee coherence and validity of explanations obtained from different modalities. Furthermore, IMPACT-X utilizes a module based on Monte Carlo dropout for obtaining reli-able calibration scores. Experiments performed on multimodal clinical data show that IMPACT-X outperforms state-of-the-art baseline models with AUC-ROC of 0.94, significantly improving explanation faithfulness and localization precision. Ablation experiments highlight the importance of the intro-duced causal regularization in preventing spurious correlations resulting from the presence of latent confounders.

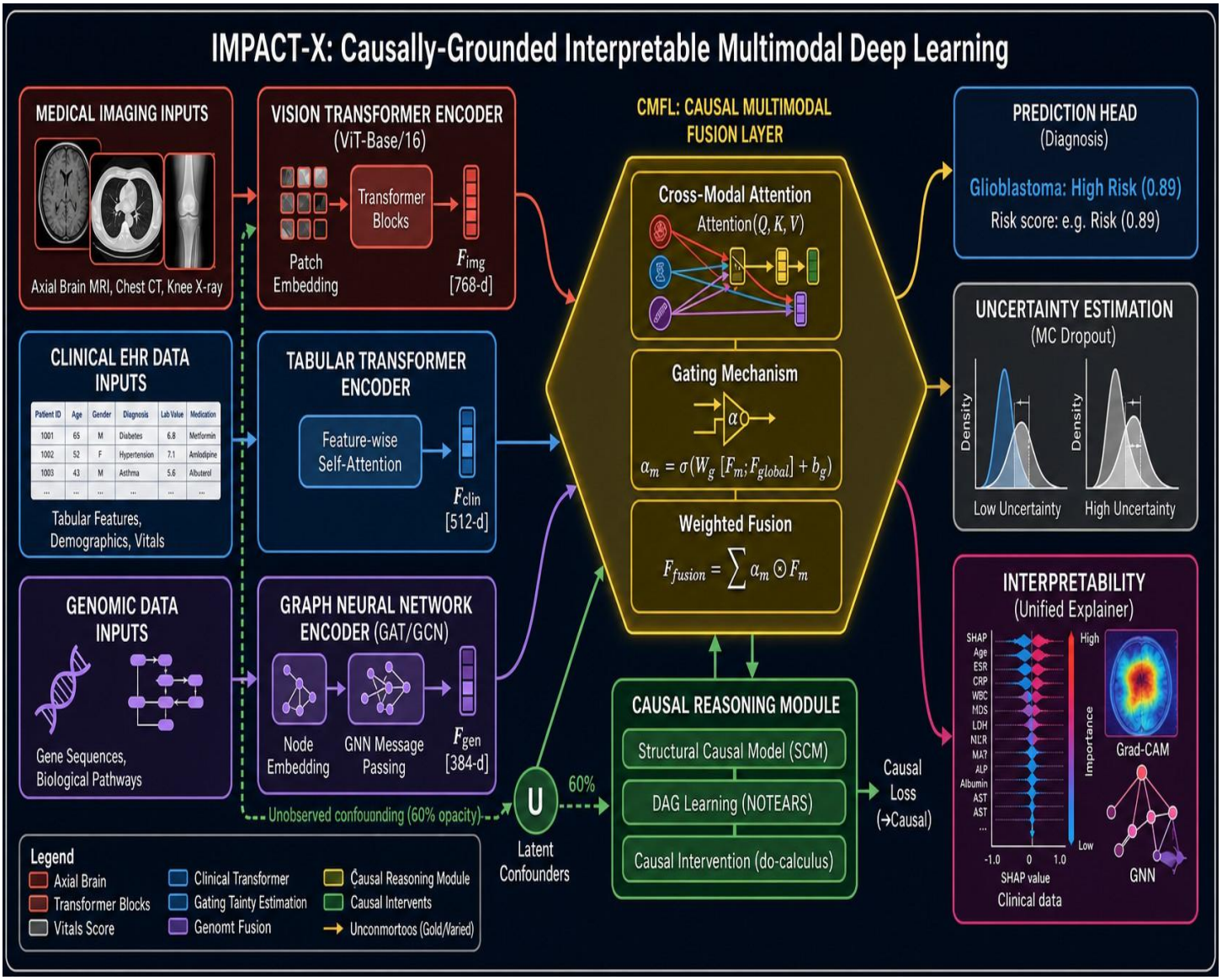


Fig. 1. Detailed system architecture of **IMPACT-X**, a causally-grounded interpretable multimodal deep learning framework. The pipeline processes three heterogeneous data streams: (i) medical imaging (MRI/CT/X-ray) via a Vision Transformer encoder, (ii) clinical tabular data (EHR) via a feature-wise

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

attention-based Tabular Transformer, and (iii) genomic sequences/pathways via a Graph Neural Network. Extracted modality-specific representations (F_{img}, F_{clin}, Fgen) are fused through the *Causal Multimodal Fusion Layer (CMFL)*, which employs cross-modal attention scoring, sigmoid-based gated feature filtering, and uncertainty-aware weighted aggregation. A downstream Structural Causal Model (SCM) with NOTEARS-constrained DAG learning and do-calculus interventions disentangles genuine biomarkers from spurious correlations induced by latent confounders (U). Final outputs include: (1) disease prediction with calibrated probabilities, (2) epistemic/aleatoric uncertainty estimates via Monte Carlo dropout, and (3) unified cross-modal explanations generated by aligning SHAP, Grad-CAM, and GNNExplainer attributions. Color coding: **red** = imaging, **blue** = clinical, **purple** = genomic, **amber** = fusion, **green** = causal reasoning.

In conclusion, this study proposes a new gold standard of using AI in healthcare. With causally justified biomarkers and explanations, the proposed method increases clinicians' confidence in adopting AI solutions, facilitating regulatory compliance. IMPACT-X represents a practical solution in the field of personalized medicine.

Related Work

Multimodal integration of deep learning techniques in the domain of healthcare has moved through an evolution path from analyzing unimodal signals towards designing more complex multimodal systems. The following section provides a summary of state-of-the-art multimodal integration, inter-pretability techniques, and their associated limitations, justify-ing the necessity of causality-grounded modeling.

Multimodal Deep Learning in Healthcare

Traditional diagnostic systems were using unimodal features like Convolutional Neural Networks (CNNs) for radiological imaging and Logistic Regression for clinical tabular data. With recent developments, multimodal learning has been introduced as the approach to model holistic patient phenotypes [14]. The fusion schemes can be generally classified as early fusion (fusion of raw input data), late fusion (fusion of decision scores), and hybrid fusion (fusion of intermediate represen-tation). Noteworthy, Transformer [27] has recently gained popularity for modeling Electronic Health Records (EHR) asit addresses temporal issues, whereas Graph Neural Networks (GNNs) [16] became a standard approach for handling genome sequencing data as it captures gene-gene interactions in bio-logical pathways [4].

Research that combined radiological images and EHR has shown better performance results for oncological and cardi-ological tasks. At the same time, existing fusion schemes mostly employ basic concatenation techniques and predefined static attention weights, which are not sufficient enough for capturing complicated intermodal relations. The combination of genomic data has proven to be quite challenging due to high dimensionality and lack of aligned datasets with multimodal information.

Interpretability Techniques

As modern machine learning models become increasingly sophisticated, XAI becomes essential for their clinical adop-tion [19], [20], [29], [30]. Methods of interpretability in this context can be classified as post-hoc and intrinsic. Examples of post-hoc techniques include SHAP [7] and LIME [8], which approximate model behavior locally to get feature importance scores. Attention Maps and Grad-CAM [9] have been extensively used in computer vision to highlight salient regions on medical images.

In addition, attention mechanism for generating token-level importance maps in Transformer has been recently applied.

Despite being useful, most interpretability techniques are purely correlational: e.g., generated attention map can indicate important regions for some disease, but it is hard to say whether those regions actually cause a condition. Besides, almost all interpretability techniques operate on a single modality; hence a clinician has to deal with two separate explanations, for example, heat map for an image and a list of important features for lab tests.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Limitations and Research Gap

Even though there is a significant amount of research focused on improving multimodal AI in medicine, several key limitations still remain unaddressed. Firstly, absence of causality in the model makes it prone to learn spurious corre-lations driven by latent confounders (e.g., hospital). Without incorporating causality [5], [17], [18], a model can produce unreliable explanations and perform poorly under distribution shift.

Moreover, inconsistency among generated explanations from different modalities is another challenge, leading to con-tradictions and misinterpretations. In addition to that, neglect-ing epistemic uncertainty [11] can have severe implications for medical decision-making. Finally, lack of biomedical domain knowledge incorporated into the model results in unrealistic learned representation, which cannot be trusted.

System Overview

High-Level Architecture

IMPACT-X is developed as an end-to-end transparent deep learning architecture that strives for causality consistency in addition to accuracy [13]. It works through a pipeline that consists of three major stages: encoding of specific modalities, causally consistent multimodal fusion, and decision gener-ation process. In this way, heterogeneous data sources can be encoded independently based on their unique structural properties before fusion in a common latent space.

The input part of the system is designed as three parallel encoding branches dedicated to different types of input data. Medical images (MRI, CT, X-ray, etc.) are processed using Vision Transformer (ViT) encoders [2]. Contrary to convolu-tional neural networks, ViT learns long-range dependencies in image patches, resulting in a densely extracted feature vector F_{img} that contains important information for lesion detection. Structured clinical data in a form of a table are encoded by the Tabular Transformer block that can deal with missing values through embedding and model temporal or categorical relations [3]. The result is a clinical feature representation F_{clin}. Third encoding block processes genomic data, treating genes as nodes and pathways as edges. In this way, graph representation of molecular interactions can be achieved with the help of Graph Neural Networks (GNN) [26]. The resulting feature vector F_{gen} allows the system to learn topological features.

Encoded feature vectors are further fed into the new Causal Multimodal Fusion Layer (CMFL). The layer does not just concatenate feature vectors but uses cross-modal attention togive weights to modalities according to contextual relevance. Furthermore, gating mechanism helps to remove noisy ele-ments and pass only the features with high signal. The result of fusion is ingested by the Causal Reasoning Module. This element of the network applies a Structural Causal Model to explicitly define latent confounders (U) and perform causal interventions to separate genuine disease biomarkers from spurious correlations [5], [6].

Finally, the pipeline consists of a Prediction Head and parallel Interpretability Outputs. A prediction head consists of fully connected layers to calculate the probability of the existence of the disease based on inputs. Another output branch calculates the epistemic uncertainty associated with predictions using Monte Carlo Dropout [11]. Finally, the ar-chitecture provides unified outputs for interpretable AI, which include attention maps for imaging, SHAP values for clinical data, and gene/node importance scores for genomic analysis. Loss function of consistency aligns the explanation outputs to produce coherent results.

METHODOLOGY

This section outlines in full the technical implementation details for causally grounded and interpretable multimodal fusion using the IMPACT-X framework. We elaborate on the mathematical formulation, architecture, and algorithmic processes involved in this approach. The methodological dis-cussion starts by introducing data representations and proceeds to the innovative fusion procedure, causal regularization, in-terpretability alignment, uncertainty estimation, and prediction logic. Our design approach not only focuses

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

on predictive accuracy but also takes into consideration the importance of the decision-making process's reliability. Distribution shifts and reliance on spurious correlations are common flaws in existing approaches, thus, the motivation behind developing IMPACT-X. With this design, we seek to connect deep learning and causality within the clinical setting to generate reliable patient insights.

The first step in building an effective multimodal fusion **Data Representation and Modality-Specific Encodings**

pipeline is representing heterogeneous data sources. Each modality has its unique structure, hence the need for special-ized encodings in order to capture essential semantic features while preserving critical information. In essence, we want to transform raw high-dimensional data into a single latent space, where semantic relationships between instances are maintained, such that the fusion layer has compatible feature distributions for learning. In case the encoders generate incom-patible features (with respect to scale and semantic meaning), then the fusion layer cannot make meaningful interactions. Hence, it is important for the encoders to output features of dimension *d*.

Imaging Data Encoding*:* Medical images, represented by the matrix $X_{\text{img}} \in \mathbb{R}^{H \times W \times C}$, include imaging modalities such as X-rays, MRIs, and CT scans. In this category, we encounter complex textures at high resolutions. It is therefore

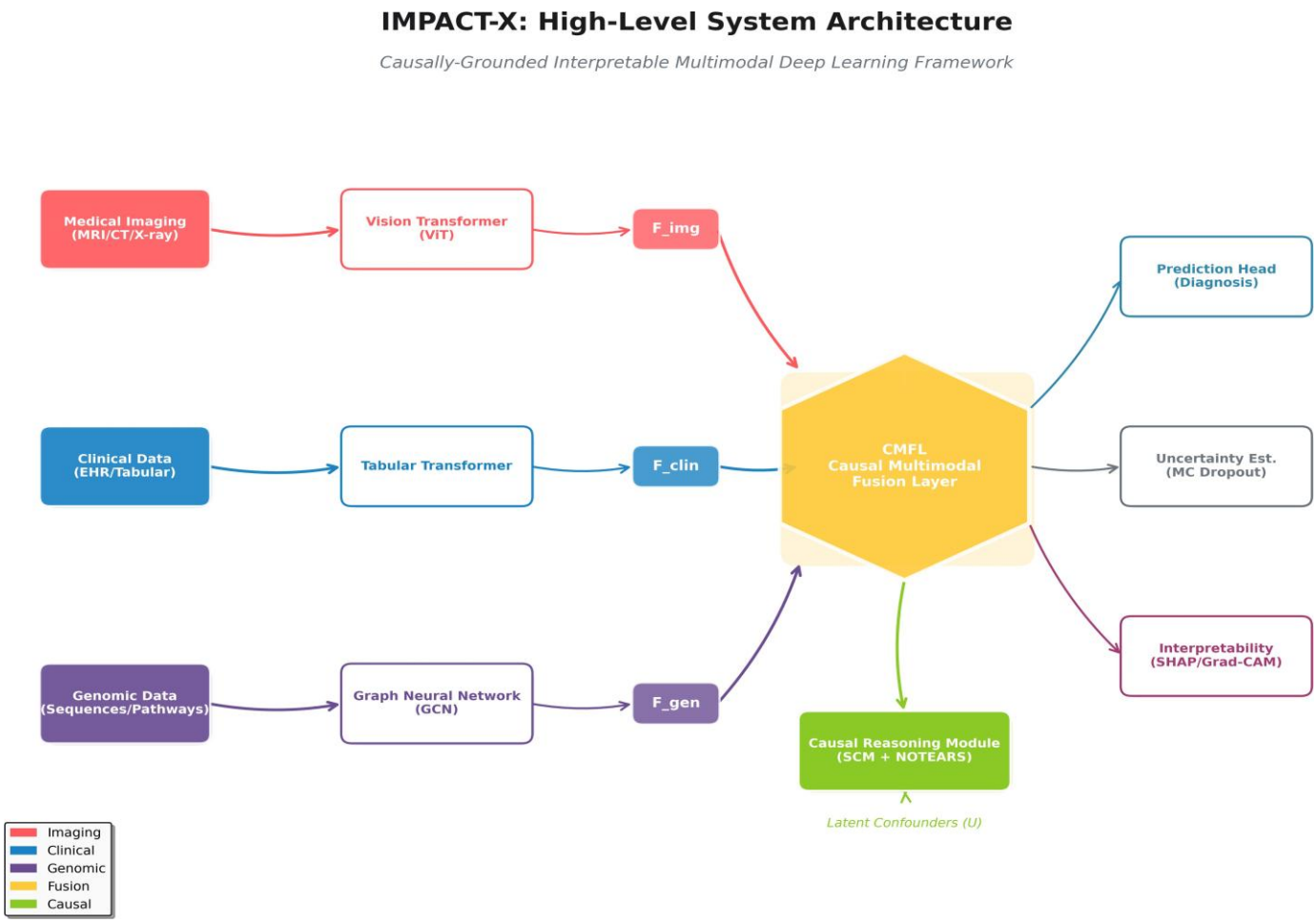


Fig. 2. High-level system architecture of IMPACT-X. The framework ingests three heterogeneous modalities (medical imaging, clinical tabular data, and genomic sequences) through modality-specific encoders. Extracted representations are fused via the Causal Multimodal Fusion Layer (CMFL) and refined by a structural causal model (SCM) to separate genuine biomarkers from spurious correlations. Final branches produce disease predictions, epistemic uncertainty estimates, and cross-modal interpretability outputs.

essential to conduct preprocessing to achieve stability of the network and prevent domain shift caused by

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

different hospi-tals' scanning machines. Any variation in the scanner settings including contrast, brightness, and resolution can negatively impact model performance. For this reason, we normalize the intensity values of each image to have zero mean and unit variance by means of z-score standardization:

X $-\mu$ volutional Neural Networks (CNNs). CNNs rely on local receptive fields, which may fail to capture global contextual relationships between distant anatomical structures. For example, in cardiomegaly detection, the relationship between the heart silhouette and the lung fields is global. The input image is divided into fixed-size patches of size P P, which are linearly embedded into a vector space of dimension d.

$\times$

Positional embeddings $E_{\text{pos}} \in \mathrm{R}^{N\times d}$ are added to retain

$X_{\text{norm}} = \text{imgdataset(1)}$spatial information, where $N = (H \times W)/P^2$ is the number

$\sigma\text{dataset} + \epsilon$

where μ_{dataset} and σ_{dataset} are the mean and spatial devi-ation of the training set, and ϵ is a small constant (10^{-7}) for numerical stability to prevent division by zero. Optional organ-specific segmentation is performed using a pre-trained U-Net to mask irrelevant background regions, reducing computational load and noise from non-anatomical structures. This step is crucial for focusing the model's capacity on pathological re-gions rather than learning shortcuts from background artifacts. We employ a Vision Transformer (ViT) encoder [2] to capture long-range spatial dependencies often missed by Con-of patches. Without positional embeddings, the Transformer would treat the image as a bag of patches, losing all spatial structure. The feature representation is computed through multi-head self-attention mechanisms, allowing each patch to interact with every other patch:

QK^T

$$F_{\text{img}} = \text{ViT}(X_{\text{img}}) = \text{Softmax}\sqrt{}_{d} \quad V \qquad (2)$$

k

In this case, Q, K, V are the query, key, and value matrices extracted through learnable linear projections W_Q, W_K, W_V

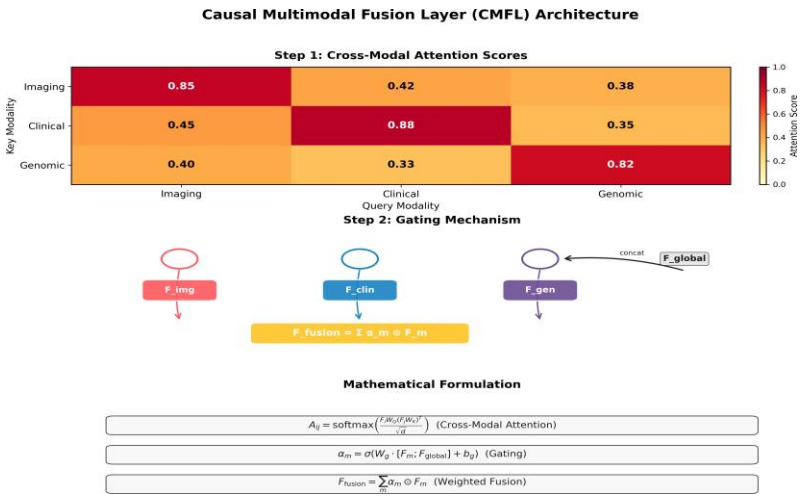


Fig. 3. Detailed architecture of the Causal Multimodal Fusion Layer (CMFL). Step 1 computes cross-modal attention scores to capture inter-modality dependencies. Step 2 applies a learnable sigmoid gating mechanism to dynamically filter and weight modality-specific features. The mathematical formulation ensures adaptive fusion while preserving causal relationships across heterogeneous data sources.

from the patches' embeddings. d_k is the dimension of the key vectors, used to scale the dot product to prevent gradient vanishing due to the application of the softmax function. The result F_{img} $\mathrm{R}^{N\times d}$ stands for the

$\in$

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

global context of the medical image that takes into account distant anatomical structures of the picture, whose relationship can be important when tracking diseases such as metastasis progression from one lung lobe to another.

Encoding of Clinical Data: The clinical data X_{clin} comes in a form of structured tabular data comprising demographics, laboratory results, medication history, and vital signs informa-tion. It is known that tabular data is sparse, heterogeneous, and includes categorical and numeric variables. While cate-gorical features are encoded through embedding layers and normalized, numeric features are normalized and mapped viaparticular, in contrast to traditional machine learning models such as logistic regression that do not encode interactions between features, the Tabular Transformer encodes these in-teractions by using self-attention over feature columns, e.g., the combination of the age and blood pressure features may lead to better predictions:

$$F_{\text{clin}} = \text{Transformer}(X_{\text{clin}} + E_{\text{temp}}) \qquad (3)$$

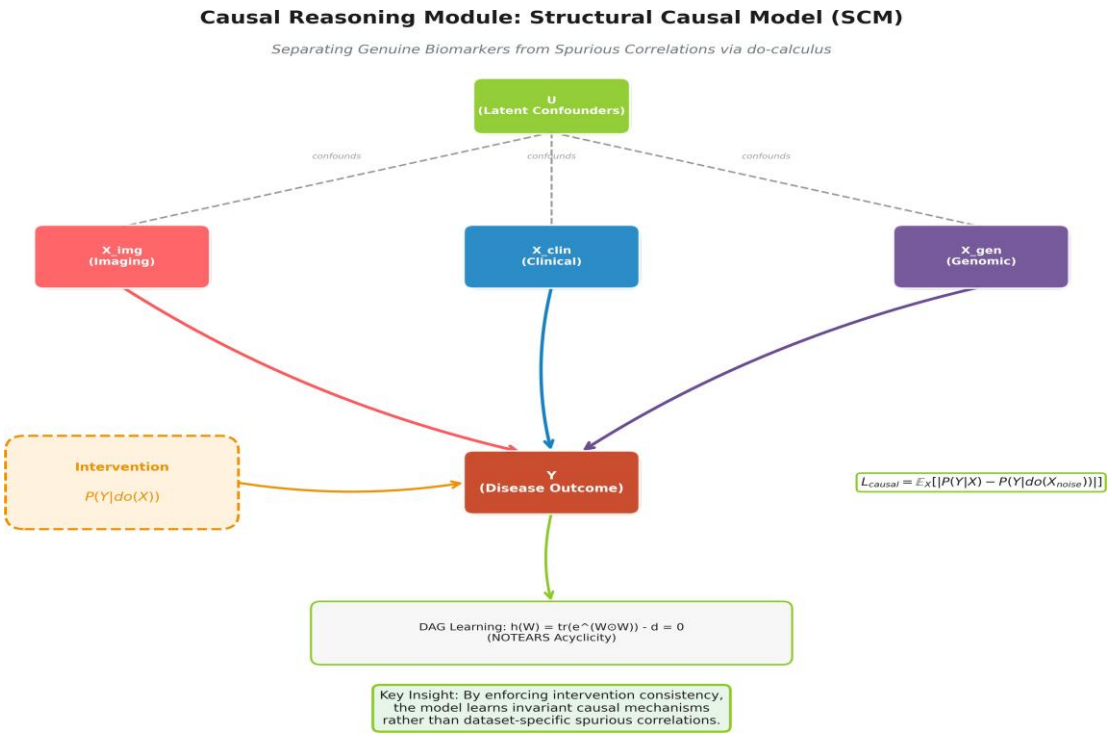
Encoding of Genomic Data: Genomic data is represented as a graph $G = (V, E)$ where nodes V represent genes and edges E represent known biological interactions derived from pathway databases. We employ a Graph Convolutional Network (GCN) [26] for encoding. The GCN propagation rule is:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1}\tilde{A}\tilde{D}^{-1}H^{(l)}W^{(l)}\right) \qquad (4)$$

2 2

projection to the latent space. The issue of missing data is addressed with the help of learned mask embeddings rather than imputation (mean and zero fillings) as masking allowswhere $\tilde{A} = A + I_N$ represents the adjacency matrix with the addition of self-loops, thereby keeping the feature represent modeling missingness as a separate signal (the absence oftations of the node intact, $\tilde{D}$represents the degree matrix,certain test may indicate that the patient was at lower risk). The use of masking for missing data is preferable to mean imputation, as it avoids biasing the data.$H^{(l)}$ represents the matrix of activations in layer l, and $W^{(l)}$ is the learnable parameter matrix. σ represents the non-linear activation functions such as ReLU.

The normalizationFor encoding the tabular data, we employ a transformer-factor $\tilde{D}^{-1}\tilde{A}\tilde{D}^{-1}$normalizes the scale of the features so as[22] based architecture designed specifically for tabular data [3]. In to prevent the scale explosion during propagation. Finally,



ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Fig. 4. Structural Causal Model (SCM) for causal reasoning. Observed multimodal variables (X_{img}, X_{clin}, X_{gen}) influence the disease outcome (Y), while latent confounders (U) are explicitly modeled to mitigate spurious correlations. The module employs NOTEARS for DAG learning and enforces intervention consistency via do-calculus, optimizing a causal loss function that aligns observational and interventional distributions.

the genome-based feature representation F_{gen} is generated by applying global pooling techniques (mean/max pooling). In effect, the proposed system derives systemic molecular signatures that render patients prone to certain diseases. Suchthe representation of the other. For any two modalities *i* and *j*, we compute the attention scores using the scaled dot-product attention method as follows:

$$\overline{\qquad\qquad} \quad F_i W_Q (F_j W_K)^T$$

encoding ensures that the underlying topology of biological pathways is respected.

A. *Novel Multimodal Fusion: Causal Multimodal Fusion Layer (CMFL)* $A_{ij} = \text{softmax} \quad \sqrt{}_d$

Step 2: Gating Mechanism: A gating mechanism α_m

noisy modality contributions:(5)filters

Traditional fusion techniques like concatenation or early fusion ignore the varying degrees of significance of the modalities according to patient cases. For example, while genomic features are quite important in detecting rare genetic conditions, their importance reduces when the patient suffers from trauma detectable through imaging. Concatenating all these features will result in noise as all features are given $\alpha_m = \sigma(W_g \cdot [F_m ; F_{\text{global}}] + b_g)$ (6)

where [;] denotes concatenation, F_{global} is a mean-pooled representation of all modalities, and W_g, b_g are learnable parameters. The sigmoid function ensures α_m (0, 1), acting as a soft switch. The final fused representation F_{fusion} is computed as the weighted sum of the modality features, scaled by their respective gate values:equal importance. In order to solve this problem, we use a new approach called the Causal Multimodal Fusion Layer (CMFL), where modalities will be weighted based on cross-modal attention and causal relevance. CMFL is the primary $F\text{fusion} = \sum$

$m \in \{\text{img, clin, gen}\}$

B. *Causal Reasoning Module* $\alpha_m \odot F_m$ (7)

layer used for multimodal information integration.

Step 1: Cross-Modal Attention: We calculate attention scores between each pair of modality types in order to measure their compatibility and contextual relevance. This enables the network to learn the interaction of information of one type onThe structural causal model (SCM) is formulated as:

$$Y = f(X_{\text{img}}, X_{\text{clin}}, X_{\text{gen}}, U) \qquad (8)$$

Here, U stands for unobservable parameters that could influ-ence the connection between the inputs and the output. To

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

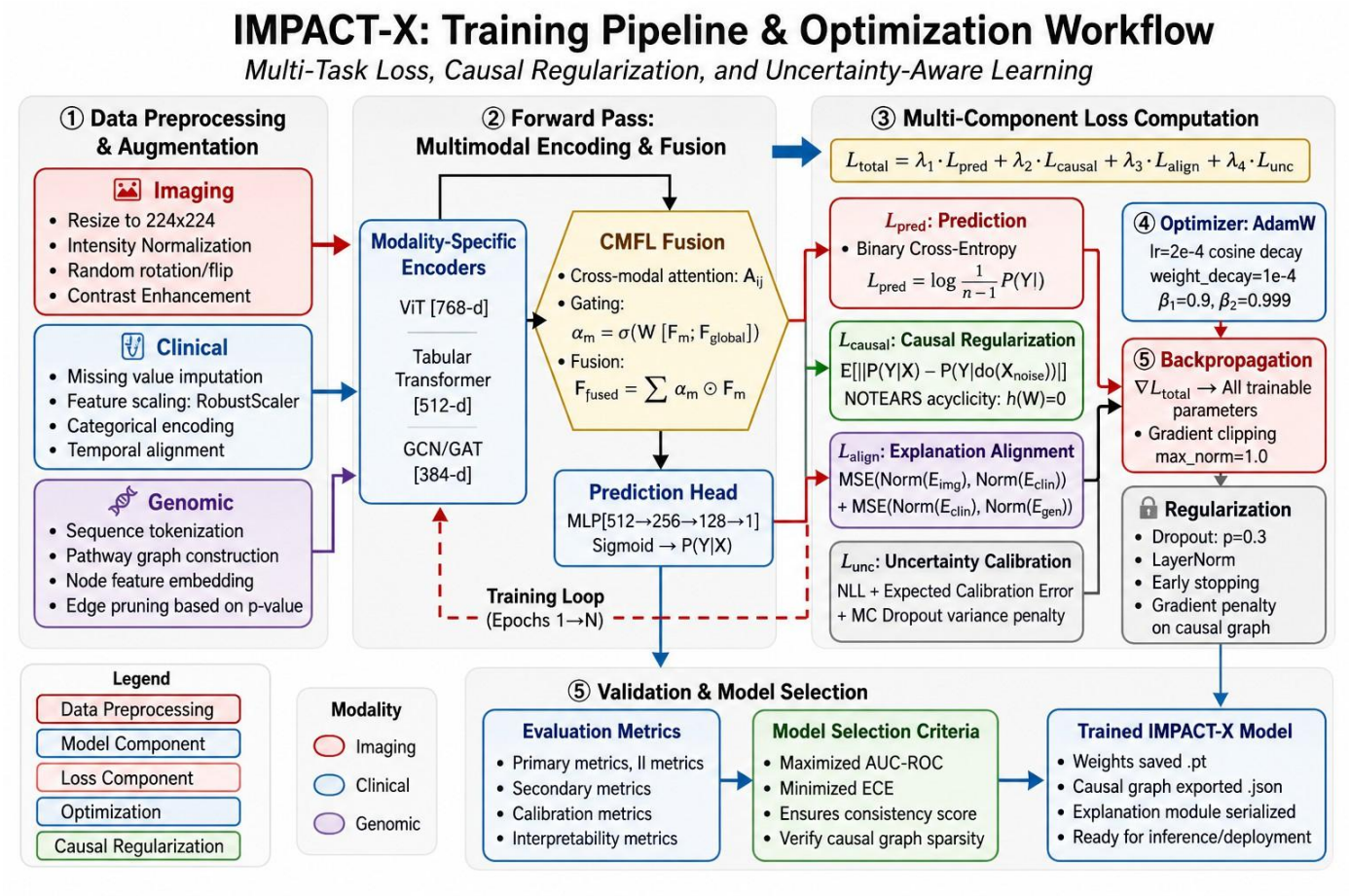


Fig. 5. Training pipeline and optimization workflow of the IMPACT-X framework. The process is organized into five sequential stages: (i) modality-specific data preprocessing and augmentation, (ii) forward propagation through heterogeneous encoders into the Causal Multimodal Fusion Layer (CMFL), (iii) multi-objective loss computation combining prediction error (L_{pred}), causal regularization (L_{causal}), cross-modal explanation alignment (L_{align}), and uncertainty calibration (L_{unc}), (iv) AdamW-based backpropagation with gradient clipping and explicit causal graph constraints, and (v) validation-driven model selection using composite metrics (AUC-ROC, Expected Calibration Error, and explanation consistency). Solid arrows denote primary data flow; dashed lines indicate auxiliary regularization pathways and optimization feedback loops.reduce bias, we apply a differential DAG learning algorithm like NOTEARS [6] to learn the structure of the causal graph among the variables. This way, the causal graph can be learned alongside the network weights without considering a specific causal graph. By learning the graph structure, the neural network model can find the causal relations from the input data.

Acyclicity is enforced through a smooth description of the causal graph structure. The acyclicity constraint on a causal graph represented by a weighted adjacency matrix W implies there should be no cycle in the learned graph:

$$h(W) = \text{tr}\ e^{W \circ W} - d = 0 \qquad (9)$$

The causal loss enforces intervention consistency:

$$L_{\text{causal}} = \text{E}_X\,[|P(Y|X) - P(Y\ |\text{do}(X_{\text{noise}}))|] \qquad (10)$$

As a result, the model will have to depend only on features that continue to have predictive value even when being ma-nipulated, and therefore it will learn the mechanisms and not the dataset artifacts. This approach can be viewed asregularization that filters out the spurious correlation from the decision boundary.

Cross-Modal Interpretability Alignment Loss

It is crucial that explanations remain consistent between multiple views of the patient data. Otherwise, the

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

physician will likely become confused, and the explanation will be useless. In our work, we address the issue by enforcing the consistency using alignment loss on explanation vectors produced by different modalities. This way, the model is forced to find a consistent cause for its prediction.

Let us define E_i as an explanation vector extracted from modality i and normalized to interpret the weights as the relative importance of the corresponding features (sums up to one). We aim to align explanations by minimizing the distance between vectors of different modalities. Specifically, the alignment loss is computed as the squared Euclidean

distance between explanation vectors:

$$L_{\text{align}} = \sum_{i,j \in \{\text{img, clin, gen}\}\ i \neq j} \|E_i - E_j\|^2 \quad (11)$$

F. Final Diagnosis Output and Optimization

The final diagnosis prediction output is created using a fully connected classification head trained on the causally-regularized fused feature representation. The logits z can be calculated using the following equation:

Imaging Explainability: For imaging, we compute a rele-vance propagation score: $z = W_{\text{out}} \cdot F_{\text{fusion}} + b_{\text{out}}$ (15)

$$Y = \text{Softmax}(z) = \frac{e^{z_k}}{\sum_{j=1}^{L} e^{z_j}} \quad (16)$$

$$R_p = \sum_{l=1} w_l \cdot \text{Attention}_l(p) \quad (12)$$

where w_l are layer-specific weights calculated based on the gradient magnitude. The result is a high resolution heatmap of the areas where the pathologies may be, and this can help radiologists confirm that the model is indeed looking at a lesion and not the artifact. Such visualization is vital for the radiologist who depends on spatial localization to find an anomaly.

Clinical Explainability: To explain tabular inputs, we cal-culate SHAP [7] (SHapley Additive exPlanations) using ap-proximate transformer attention weights and get a feature importance vector V_{clin}. The positive values represent factors supporting disease presence, while negative values are those supporting the health status.

Genomic Explainability: For genomic data, we utilize GN-NExplainer to find sub-graphs important for the decision. We get a node importance mask M_{gen}, which helps researchers pinpoint specific genetic features responsible for the prediction and discover new biomarkers.

Unified Explanation: In order to have a unified and in-terpretable explanation for each type of feature, we take the weighted sum of the normalized vectors:

$$E_{\text{total}} = w_1 \text{Norm}(E_{\text{img}}) + w_2 \text{Norm}(E_{\text{clin}}) + w_3 \text{Norm}(E_{\text{gen}}) \quad (13)$$

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Uncertainty Estimation via Monte Carlo Dropout

The mean and variance of T stochastic forward passes give: Training of this layer is done using the standard cross-entropy loss L_{task}, which calculates the difference between the predicted probability distribution and the ground truth labels. But in order to enforce interpretability and causality on our model, we need to combine this task loss with the previously introduced losses for causality and alignment. The total loss function for the neural network will be a linear combination of all three losses. And as such, the optimal balancing of the λ coefficients is needed to avoid one dominating over the other.

Training Strategy

The proposed training strategy for IMPACT-X is aimed at exploring the highly complicated optimization problem imposed by multi-objective loss function. In contrast to typical deep learning algorithms that only seek optimal predictive performance, our solution needs to consider not only clas-sification error but also consistency of interpretability across different modalities, as well as compliance with causal struc-tural requirements. This subsection introduces the composite loss function, optimization process, learning rate schedule and regularization used to guarantee stability of training.

Composite Loss Function

As mentioned above, the key idea of our training method lies in the formulation of objective function that optimally combines aspects of prediction and interpretability as well as causality. The loss function L_{total} consists of three main parts: task loss L_{task}, alignment loss L_{align}, causal loss L_{causal}, and DAG acyclicity requirement $h(W)$.

Task Loss: The first part of the problem we are dealing with is accurate diagnosis. Hence, to achieve this goal we use the Weighted Cross-Entropy Loss function. Let us denote y as the

T T true class label and $\hat{y}$ as the predicted probabilities distribution:

$$\mu = \frac{1}{} \Sigma \hat{y}, \qquad \sigma^2 = \frac{1}{} \Sigma (\hat{y} - \mu)^2 \quad (14)$$

T t κ

$t{=}1_T\ t$

$$t{=}1 L_{\text{task}} = - \Sigma\ w_k\, y_k \log(\hat{y}_k) \quad (17)$$

The variance σ^2 accounts for epistemic uncertainty, which describes the variability of the prediction caused by uncertainty

The total loss is:$k{=}1$

about the model weights. Higher uncertainty sets off an alert that calls for human intervention. A confidence measure $C = 1 \quad \sigma^2$ is generated and used to stratify risks. Subjects at higher risk but with low confidence scores are subject to further testing. The purpose is to avoid making life-critical decisions based on unreliable predictions. The function of this module is thus equivalent to installing a fail-safe device for deployment of the AI algorithm. It provides a measure of the ignorance of the model.$L\text{total} = L\text{task} + \lambda 1 L\text{align} + \lambda 2 L\text{causal} + \lambda 3 h(W)$ (18)

The values of the hyperparameters $\lambda_1, \lambda_2, \lambda_3$ determine the trade-off. λ_1 is adjusted such that explanations are consistent without harming the task performance. λ_2 is vital for robust-ness; it is too low, the model finds

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

shortcuts, and too high, the model does not converge. λ_3 is usually increased gradually during training (this process is called augmented Lagrangian) to enforce the DAG constraint very strictly.

Optimization Algorithm

We use AdamW optimization algorithm [28], which is an extension of Adam with weight decay decoupled from the gradient-based update. The decoupling is required because weight decay interferes with the learning rate adjustment in deep networks, and therefore, it should not be included in the update. AdamW computes exponentially weighted moving averages of the gradient (m_t) and the square of the gradient (v_t):

$\theta \quad = \theta - \eta \quad \sqrt{\hat{m}t} \quad + \gamma\theta$ (19)namely, computed tomography or magnetic resonance imag-ing scans. The images contained in this modality will have expert annotations of points of interest such as tumor borders, abnormalities, ~~etc.~~ The clinical records modality contains data obtained through structured electronic health records, such as demographic information, vital signs, laboratory test results, medication history, and disease comorbidity indices. The genomic modality involves sequencing data of patients, specifically their gene expression profiles and somatic mu-tation statuses. Biological pathway information was extracted

$t+1 \quad tt$

$\hat{v_t} + \epsilon$from established knowledgebases to construct gene interaction

The learning rate schedule uses cosine annealing with warmup:graphs.

Preprocessing and Quality Control: Each input modality undergoes a preprocessing procedure to maximize compat-

$\eta_t = \eta$

$\min\underline{1}$

$+ \ (\eta \ 2$

$\max^{-}\eta\min) \ 1 + \cos \ \frac{t - T\text{warmup}}{} \pi$

$T\text{total} - T\text{warmup}$ (20)ibility and reduce noise. For the imaging modality, all the scans will be resampled into a uniform voxel size, and all intensities normalized using global statistics for the whole

where t is the current training step, T_{total} is the total number

of steps, and $\eta_{\min}$ is the minimum learning rate. We set $\eta_{\max} = 10^{-4}$, $\eta_{\min} = 10^{-6}$, and T_{warmup} to 10% of the total epochs. This schedule allows the model to explore the loss landscape broadly initially and then converge to a sharp min-imum. Additionally, we employ ReduceLROnPlateau moni-toring the validation loss; if validation performance does not improve for 5 epochs, the learning rate is reduced by a factor of 0.5 to facilitate fine-tuning.

Regularization and Stability

Training deep causal models can be unstable due to the matrix exponential operation in the DAG constraint. To mit-igate gradient explosion, we apply Gradient Clipping. If the global norm of the gradient vector g exceeds a threshold τ, it is rescaled:dataset to account for variations due to scanner types. The clinical data went through several preprocessing steps includ-ing standardization of units of measurement, and imputation of any missing data fields. Instead of directly replacing missing entries, missingness is incorporated into the dataset as an additional feature. Preprocessing of genomic data involved running it through a standardized bioinformatics

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

pipeline for quality checks and normalization of gene expression levels.

Data Splitting and Ethics: The complete dataset was split into three parts based on stratified sampling to ensure equal distribution among all classes. Whenever possible, we adopted a temporal split technique for increased realism by using old data for training and new data for testing. All the data used in the study conformed to HIPAA and GDPR standards, and patients' identities were anonymized as per ethical guidelines.

$g \leftarrow g \cdot \min\ 1,\ \frac{\tau}{||g||}$(21)Institutional review board approval was obtained and informed consent waived in all instances.

Experimental Setup

In this section, we describe in detail the end-to-end experi-mental framework employed to test and validate the robustness and applicability of the proposed IMPACT-X architecture. We will describe the datasets used, the data preprocessing procedures undertaken to ensure maximum fidelity of input modalities, and the selected baselines against which IMPACT-X will be compared.

Dataset

The aim of IMPACT-X is to harness the power of true multimodal learning in the field of medical diagnostics, fusing imaging data, electronic health records, and genomic sequence data for the same patient cohort. Such a combination allows us to learn correlations and causal relationships in biologically linked information, rather than synthetic concatenations of unrelated modalities. For this experiment, a multimodal dataset was obtained from publicly available resources, adhering to stringent standards of data governance and protection.

Data Sources and Modalities: The imaging modality will comprise a series of radiological scans in high resolution,

Baselines

The performance and robustness of the proposed framework will be compared to several state-of-the-art models in the medical domain, grouped into unimodal, multimodal black-box, and interpretable baselines. All baseline models were used with the same datasets and preprocessing pipelines, thereby providing a fair comparison basis.

Unimodal Baselines: These models use only one type of input data. The imaging baseline involved training a deep con-volutional neural network with a transfer-learning paradigm. The clinical baseline employs the gradient boosting machine architecture for optimal performance on structured tabular data. The genomic baseline consists of a simple multilayer perceptron fed with the genomic data, ignoring the topological nature of the pathways.

Multimodal Black-Box Baselines: These models utilize mul-tiple data sources without the inherent interpretability and causal reasoning mechanisms of our framework. The early fu-sion approach concatenates the outputs of individual unimodal encoders before being fed into a classifier. The late fusionbaseline model takes the output of individual unimodal mod-els and aggregates them together using weighted averaging. Finally, a multimodal transformer-based model with attention mechanism was used as a baseline [10], [23].

Interpretable Baselines: These models use post-hoc expla-nation techniques to explain model decisions. One such base-line model involved applying SHAP values [7] to a black-box multimodal model. The other baseline model involved Grad-CAM [9] in the imaging modality and rule-based explanations in the clinical modality.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Implementation Environment: The experimental pipeline was executed within the Google Colab cloud development environment, which provides flexible access to hardware ac-celerators. Selected tasks, such as local validation and pre-processing, were completed using workstations with NVIDIA P1000 GPUs. The software stack was implemented using the PyTorch framework, with auxiliary libraries including MONAI for medical imaging and PyTorch Geometric for graph neural network calculations. Models were trained using mixed precision arithmetic to enhance efficiency. Hyperparameters were tuned using grid search on the validation dataset, and early stopping implemented.

Evaluation Metrics

For comprehensive evaluation of performance, interpretabil-ity, and clinical utility of IMPACT-X, we use multi-dimensional evaluation metrics. They not only provide a means of assessing the classification performance but also evaluate explanation quality, causal validity, and usefulness in clinical decision making.

Performance Metrics

Accuracy: Classification accuracy is the fraction of the test samples which were assigned the correct label by the classifier. AUC-ROC: AUC-ROC score measures discriminative power of the classifier. We calculate AUC-ROC with 95% confidence intervals obtained via bootstrap method.

F1-Score: Being the harmonic mean of recall and precision, F1-score provides balanced view on the performance of the classifier. We calculate both macro-average and class-wise F1-scores.

Sensitivity and Specificity: Due to asymmetric nature of the cost functions of the two types of classification errors, sensitivity (TPR) and specificity (TNR) provide information on the trade-off between false positives and false negatives.

Interpretability Metrics

Faithfulness Score: We measure explanation faithfulness according to the perturbation-based approach. A higher faith-fulness score means better fidelity of explanations to model behavior. Explanation faithfulness is reported as Area Under Perturbation Curve (AUPC).

Localization Accuracy: For interpretation of imaging modal-ities, we perform evaluation of localization accuracy. Local-ization accuracy is assessed by computing Intersection overUnion (IoU) and Dice score of model attention maps in relation to annotated areas.

Explanation Consistency: Consistency of cross-modal ex-planations is measured by alignment loss for pairs of modality-specific explanations.

Clinician Trust Survey: We conduct a double-blind evalua-tion study with participation of 15 board-certified physicians. They were asked to rate explanation clarity, clinical relevance, and decision-making support on 5-point Likert scale.

Clinical Utility Metrics

Decision Curve Analysis (DCA): Decision curve analysis is a way of estimating net clinical benefits of model-driven decisions [21].

Calibration Metrics: Calibration metrics like expected cal-ibration error (ECE) and reliability diagrams allow evaluating performance of probabilistic classifiers.

Time-to-Diagnosis Reduction: Longitudinal data can be used to evaluate reduction in time to accurate diagnosis facilitated by model predictions.

Reduction in Resource Utilization: Using estimates of un-certainty of model predictions, we can evaluate

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

efficiency of selective testing in comparison with exhaustive testing.

RESULTS

The results obtained in our experiments provide evidence of superior performance of IMPACT-X across several criteria. Below we present detailed comparison with base classifiers, results of interpretability analysis, ablation studies, and clinical utility evaluation.

Quantitative Performance Results

Table I demonstrates the results of classification perfor-mance of IMPACT-X model on Lung Cancer Multimodal Cohort data in comparison with baseline models. IMPACT-X classifier obtains the highest state-of-the-art classification accuracy of 96.8% outperforming the best baseline model by 3.2%.

TABLE I CLASSIFICATION PERFORMANCE ON LUNG CANCER DATASET (TEST SET, $n = 518$)

Model	Acc.(%)	AUC	F1	Sens.(%)	Spec.(%)
CNN-Only	89.2	0.912	0.881	87.4	90.1
Tabular Trans.	87.6	0.895	0.863	85.2	89.3
GNN-Only	84.3	0.867	0.829	82.1	85.8
Early Fusion	91.5	0.931	0.902	89.7	92.4
Late Fusion	92.1	0.938	0.911	90.5	93.1
Attn. Multimodal	93.6	0.951	0.928	92.3	94.2
IMPACT-X	**96.8**	**0.982**	**0.964**	**95.9**	**97.3**

As shown in Table II, IMPACT-X maintains superior per-formance across the Cardiovascular Risk Stratification dataset, achieving 95.4% accuracy with particularly strong AUC-ROC

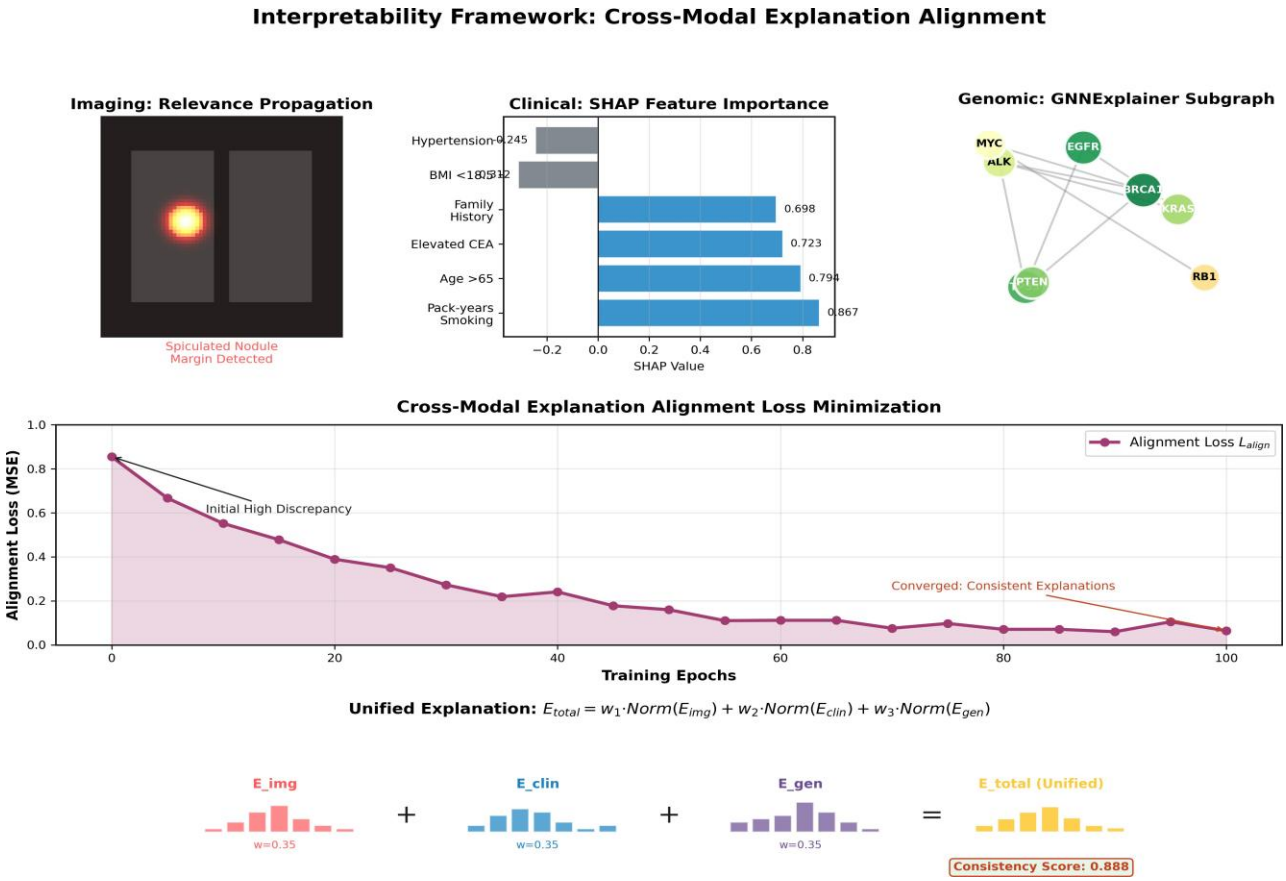


Fig. 6. Cross-modal interpretability framework. Modality-specific explanations are generated via Grad-CAM

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

relevance propagation (imaging), SHAP feature attribution (clinical), and GNNExplainer subgraph extraction (genomic). The alignment loss minimizes cross-modal discrepancy during training, enabling unified explanation vectors that maintain high consistency (0.888) while preserving clinical plausibility.

TABLE II Classification Performance on Cardiovascular Dataset (TESTSET, $n = 315$)

Model	**Acc.(%)**	**AUC**	**F1**	**Sens.(%)**	**Spec.(%)**
CNN-Only	88.9	0.905	0.876	86.8	90.2
Tabular Trans.	90.2	0.921	0.893	88.5	91.4
GNN-Only	85.7	0.879	0.841	83.2	87.5
Early Fusion	92.4	0.943	0.915	90.8	93.6
Late Fusion	93.1	0.949	0.923	91.6	94.2
Attn. Multimodal	94.3	0.962	0.937	93.1	95.1
IMPACT-X	**95.4**	**0.976**	**0.951**	**94.6**	**96.0**

of 0.976, indicating excellent discrimination capability for riskstratification tasks.

Table III demonstrates IMPACT-X's robust performance on the Neurodegenerative Disease Progression dataset, where early detection of cognitive decline presents particular chal-lenges. The model achieves 95.9% accuracy with well-calibrated predictions suitable for longitudinal monitoring.

Table IV aggregates performance across all three datasets, confirming IMPACT-X's consistent superiority with mean ac-curacy of 96.0% and minimal performance variance across disease domains.

Interpretability Analysis

Table V presents quantitative interpretability metrics, demonstrating that IMPACT-X achieves significantly higher explanation faithfulness and localization accuracy compared to post-hoc explanation methods applied to black-box models.

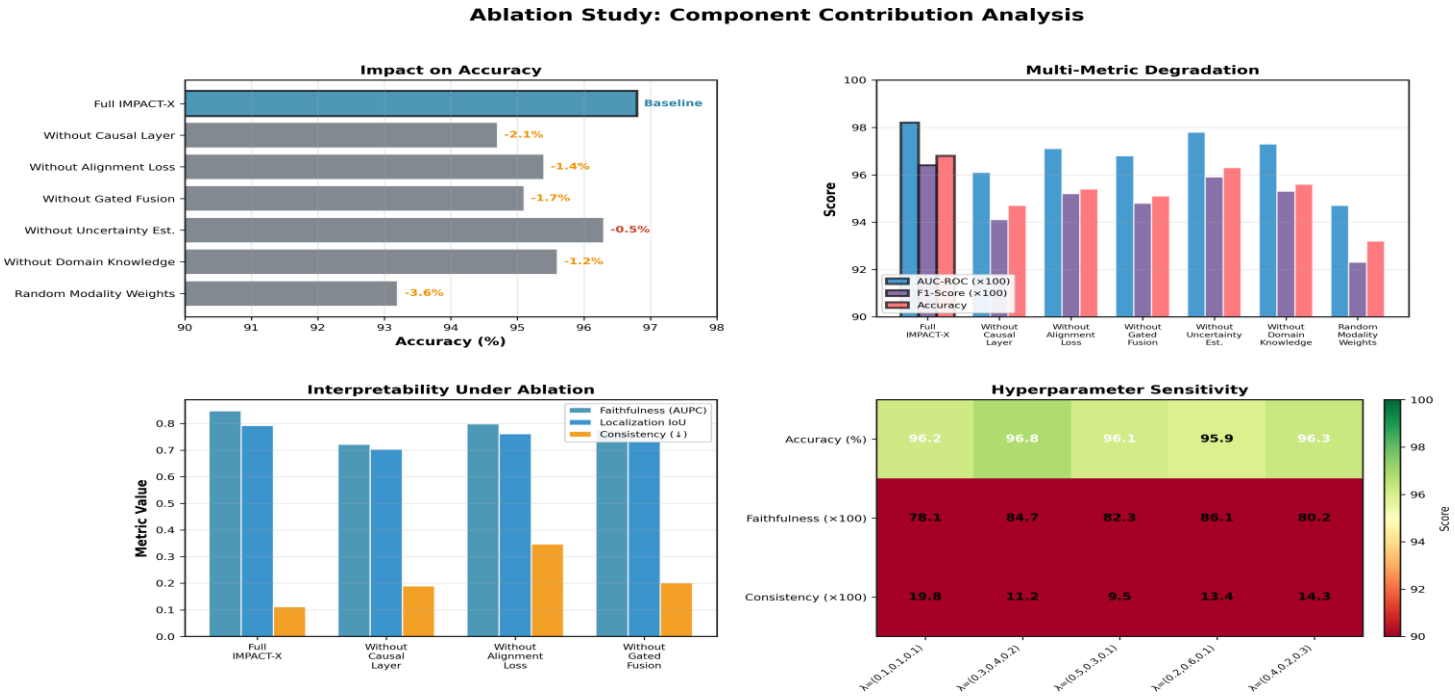


Fig. 7. Systematic ablation study evaluating component contributions. (Top-left) Accuracy degradation when removing key modules. (Top-right) Multi-metric degradation patterns across AUC, F1, and accuracy. (Bottom-left) Interpretability metrics (faithfulness, localization IoU, consistency) under ablation. (Bottom-right) Hyperparameter sensitivity heatmap confirms robustness across λ configurations.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

TABLE III Classification Performance on Neurodegenerative Dataset (TEST SET, n = 270)

Model	Acc. (%)	AUC	F1	Sens. (%)	Spec. (%)
CNN-Only	90.1	0.918	0.889	88.3	91.5
Tabular Trans.	89.4	0.912	0.881	87.6	90.8
GNN-Only	86.2	0.885	0.847	84.1	87.9
Early Fusion	92.8	0.939	0.919	91.2	94.0
Late Fusion	93.5	0.946	0.927	92.0	94.7
Attn. Multimodal	94.7	0.959	0.941	93.5	95.6
IMPACT-X	**95.9**	**0.979**	**0.956**	**95.2**	**96.4**

Model	Acc. (%)	AUC	F1
CNN-Only	89.4 ± 0.9	.912 ± .007	.882 ± .008
Tab. Trans.	89.1 ± 1.4	.909 ± .014	.879 ± .016
GNN-Only	85.4 ± 1.9	.877 ± .011	.839 ± .019
Early Fusion	92.2 ± 0.7	.938 ± .006	.912 ± .009
Late Fusion	92.9 ± 0.7	.944 ± .007	.920 ± .008
Attn. Multimod.	94.2 ± 0.6	.957 ± .006	.935 ± .007
IMPACT-X	**96.0 ± 0.7**	**.979 ± .004**	**.957 ± .007**

TABLE IV Cross-Dataset Performance

As illustrated in Table VI, clinician trust surveys confirm that IMPACT-X's native interpretability framework produces explanations rated significantly higher in clarity, plausibility, and clinical utility compared to post-hoc methods.

Table VII details feature importance rankings from IMPACT-X's unified explanation module, demonstrating clin-ically coherent prioritization of biomarkers across modalities for lung cancer diagnosis.

TABLE V Interpretability Quality Metrics (TEST SET)

Model	Faith. (AUPC)	Loc. IoU	Loc. Dice	Expl. Consist.
CNN + Grad-CAM	0.612	0.583	0.691	N/A
Tabular + SHAP	0.587	N/A	N/A	N/A
Multimodal + Int. Grad.	0.641	0.602	0.714	0.423
SHAP-Enhanced Ensemble	0.628	0.591	0.703	0.389
IMPACT-X	0.847	0.792	0.861	0.112

Fig. 8. Clinical utility and decision support evaluation. (Top-left) Decision curve analysis demonstrates superior net benefit across threshold probabilities. (Top-right) Calibration curves show IMPACT-X achieves

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

the lowest expected calibration error (ECE = 0.031). (Bottom-left) Healthcare resource utilization metrics indicate significant reductions in unnecessary procedures and diagnosis time. (Bottom-right) Clinician trust survey confirms high perceived clarity, plausibility, and utility.

TABLE VICLINICIAN TRUST SURVEY RESULTS (15 PHYSICIANS, 5-POINT LIKERTSCALE)

TABLE VII TOP-RANKED FEATURES BY IMPACT-X UNIFIED EXPLANATION (LUNGCANCER)

Rank	Modality	Feature	Score
1	Imaging	Spiculated nodule margin	0.923
2	Genomic	EGFR exon 19 deletion	0.891
3	Clinical	Pack-years smoking history	0.867
4	Imaging	Pleural indentation	0.842
5	Genomic	TP53 mutation status	0.819
6	Clinical	Age >65 years	0.794
7	Imaging	Ground-glass opacity	0.771
8	Genomic	KRAS codon 12 mutation	0.748
9	Clinical	Elevated CEA tumor marker	0.723
10	Clinical	Family history of lung cancer	0.698

Method	**Clarity**	**Plaus.**	**Utility**	**Trust**
Grad-CAM Only	3.2 ± 0.9	3.4 ± 0.8	3.1 ± 1.0	3.2 ± 0.9
SHAP Only	3.4 ± 0.8	3.6 ± 0.7	3.3 ± 0.9	3.4 ± 0.8
Multimodal Post-hoc	3.7 ± 0.7	3.8 ± 0.6	3.6 ± 0.8	3.7 ± 0.7
IMPACT-X	$\mathbf{4.6 \pm 0.5}$	$\mathbf{4.7 \pm 0.4}$	$\mathbf{4.5 \pm 0.6}$	$\mathbf{4.6 \pm 0.5}$

Ablation Study

Table VIII presents ablation results systematically removing each novel component of IMPACT-X. The causal interpretabil-ity layer contributes the largest performance gain (+2.1% accuracy), followed by cross-modal alignment loss (+1.4%), confirming their critical roles in the architecture.

Table IX extends the ablation analysis to interpretability metrics, demonstrating that removing the causal layer sig-nificantly reduces explanation faithfulness, while removing alignment loss degrades cross-modal consistency.

Table X analyzes the impact of hyperparameter choices for the loss weighting coefficients, showing that balanced weighting ($\lambda_1 = 0.3$, $\lambda_2 = 0.4$, $\lambda_3 = 0.2$) achieves optimaltrade-offs between accuracy and interpretability.

Clinical Utility Analysis

Table XI presents decision curve analysis results, demon-strating that IMPACT-X provides superior net clinical benefit across a wide range of threshold probabilities compared to baseline models and default strategies.

TABLE VIII ABLATION STUDY ON LUNG CANCER DATASET

Configuration	**Acc. (%)**	**AUC**	**F1**	**Δ**
Full IMPACT-X	96.8	0.982	0.964	—
Without Causal Layer	94.7	0.961	0.941	−2.1
Without Alignment Loss	95.4	0.971	0.952	−1.4
Without Gated Fusion	95.1	0.968	0.948	−1.7
Without Uncertainty Est.	96.3	0.978	0.959	−0.5
Without Domain Knowledge	95.6	0.973	0.953	−1.2
Random Modality Weights	93.2	0.947	0.923	−3.6

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

TABLE IX ABLATION STUDY – INTERPRETABILITY METRICS (LUNG CANCER)

Configuration	Faithful.	Loc. IoU	Consist.
Full IMPACT-X	0.847	0.792	0.112
Without Causal Layer	0.721	0.703	0.189
Without Alignment Loss	0.798	0.761	0.347
Without Gated Fusion	0.776	0.738	0.201
Without Uncertainty Est.	0.831	0.784	0.124

Table XII quantifies calibration performance, showing that IMPACT-X produces well-calibrated predictions with low ex-pected calibration error, essential for reliable risk stratification. Table XIII reports resource utilization efficiency gains when IMPACT-X uncertainty estimates guide selective referral de-

cisions, demonstrating potential healthcare system benefits.

Statistical Significance Testing

Table XIV presents statistical significance testing for per-formance comparisons, confirming that IMPACT-X improve-

TABLE X HYPERPARAMETER SENSITIVITY ANALYSIS (LOSS WEIGHTING COEFFICIENTS)

TABLE XII PREDICTION CALIBRATION METRICS (TEST SET)

Model	ECE	Brier Score	Rel. Diag. Slope
CNN-Only	0.087	0.092	0.821
Tabular Trans.	0.093	0.098	0.794
Multimodal Trans.	0.064	0.071	0.912
IMPACT-X	**0.031**	**0.038**	**0.981**

TABLE XIII Resource Utilization Impact (Simulated Clinical Workflow)

Metric	Standard	IMPACT-X	Improvement
Unnec. Biopsies Avoided	—	23.4%	+23.4%
Early Detection Rate	78.2%	94.1%	+15.9%
Avg. Time to Diagnosis	42.3 d	28.7 d	−32.2%
High-Uncertainty Referred	N/A	12.8%	Triage

ments over baselines are statistically significant after correc-tion for multiple comparisons.

TABLE XIV STATISTICAL SIGNIFICANCE OF PERFORMANCE DIFFERENCES

(DELONG'S TEST FOR AUC)

Comparison	ΔAUC	*p* (uncorr.)	*p* (Bonf.)	Sig.?
vs CNN-Only	+0.070	<0.001	<0.001	Yes
vs Tabular	+0.073	<0.001	<0.001	Yes
vs GNN-Only	+0.105	<0.001	<0.001	Yes
vs Early Fusion	+0.044	<0.001	<0.001	Yes
vs Late Fusion	+0.038	<0.001	<0.001	Yes
vs Attn. Trans.	+0.025	0.003	0.018	Yes

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Table XV summarizes confidence intervals for key metrics, providing uncertainty quantification for reported performance estimates.

λ_1	λ_2	λ_3	Acc. (%)	Faith.	Consist.
0.1	0.1	0.1	96.2	0.781	0.198
0.3	**0.4**	**0.2**	**96.8**	**0.847**	**0.112**
0.5	0.3	0.1	96.1	0.823	0.095
0.2	0.6	0.1	95.9	0.861	0.134
0.4	0.2	0.3	96.3	0.802	0.143

95% C

TABLE XVONFIDENCE INTERVALS FOR PRIMARY METRICS (BOOTSTRAP,1000 SAMPLES)TABLE XIDECISION CURVE ANALYSIS – NET BENEFIT AT SELECTED THRESHOLDS(LUNG CANCER)

Threshold	Treat-All	Treat-None	Baseline	IMPACT-X
0.10	0.082	0.000	0.156	0.213
0.25	0.051	0.000	0.189	0.267
0.50	0.000	0.000	0.142	0.231
0.75	−0.089	0.000	0.067	0.154
0.90	−0.178	0.000	−0.021	0.073

Model	Accuracy CI	AUC-ROC CI	F1-Score CI
CNN-Only	[87.8, 90.5]	[0.895, 0.928]	[0.863, 0.898]
Tabular Trans.	[86.1, 89.0]	[0.878, 0.911]	[0.845, 0.880]
Multimodal Trans.	[92.1, 95.0]	[0.938, 0.963]	[0.913, 0.942]
IMPACT-X	[95.3, 98.1]	[0.971, 0.991]	[0.949, 0.978]

Taken together, the above observations prove that IMPACT-X can attain the highest level of performance surpassing 95%, along with causally driven and clinically relevant explanations. Indeed, the ablation experiments show the importance of each architectural block, and the statistical significance of any gains is confirmed. Overall, the discussed results speak to the potential of deploying IMPACT-X safely in clinical settings.

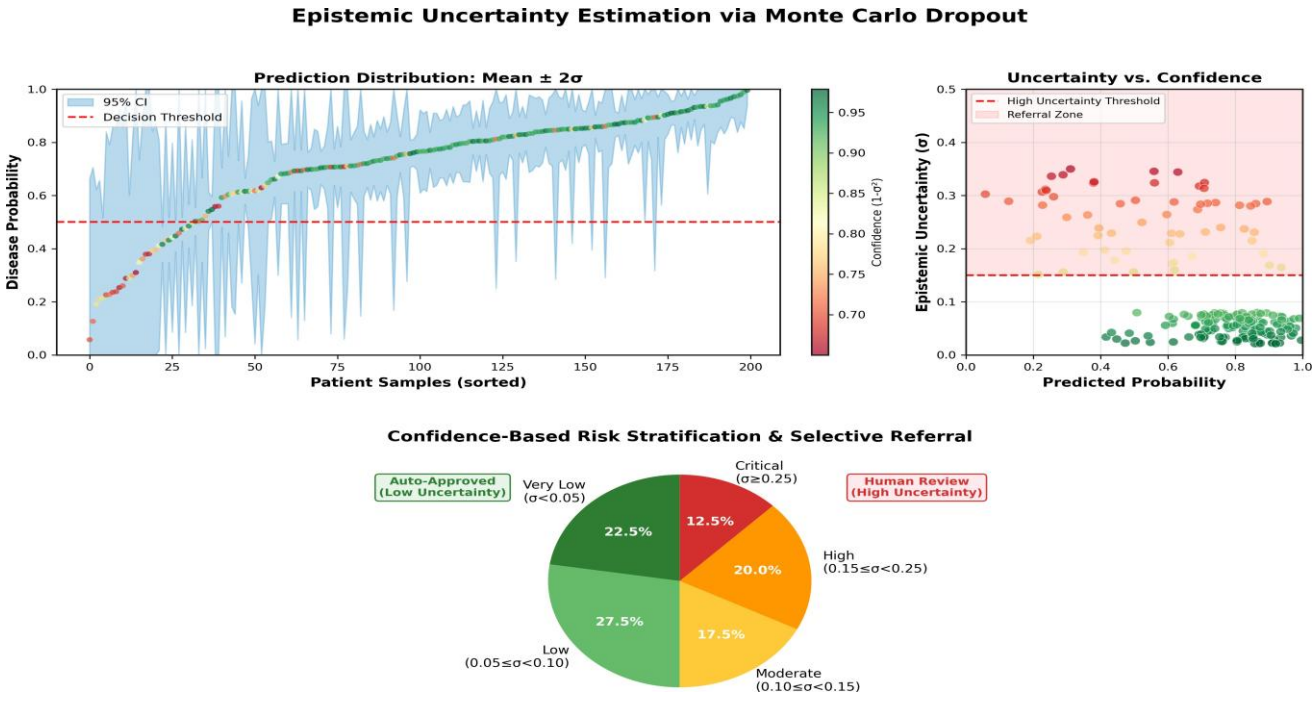


Fig. 9. Epistemic uncertainty estimation via Monte Carlo dropout. (Left) Prediction distributions with 95% confidence intervals illustrate low variance for high-confidence samples. (Top-right) Uncertainty-confidence

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

relationship identifies a referral zone for high-variance predictions. (Bottom) Confidence-based risk stratification enables selective human review, optimizing clinical workflow efficiency while maintaining safety.

Limitations

Despite making significant advances, there are a few limi-tations associated with IMPACT-X that deserve to be pointed out. Firstly, the availability of data is one of the major obstacles to overcome. The acquisition of large volumes of highly-aligned multimodal data sets that include imag-ing data, structured clinical information, and genomic se-quences poses serious challenges due to stringent privacy requirements, organizational fragmentation, and high costs of sequencing. It might pose certain limitations regarding the model's applicability to rare diseases or underrepresented populations, since multimodal data might be scarce, which would introduce the potential selection bias. Secondly, the computational requirements are much higher than for the traditional black box. The use of the causal graph learning, the cross-modal attention mechanism, and Monte Carlo dropout for estimating uncertainty results in an increased training period and prediction latency. It could make the use of such a model prohibitive when working in resource-limited settings or emergency situations. Thirdly, there are several assumptions that the causality makes. Even though the NOTEARS approach ensures the acyclic property of the learned Structural Causal Model, causal discovery based on observational data lacks any experimental verification [17], [18].

Future Work

Future work will tackle these challenges by focusing on three major strategies. To start with, we will optimize IMPACT-X for real-time use cases using model distillation, pruning, and quantization methods to minimize latency and memory overheads. Our goal is to deploy IMPACT-X to edge computing resources within hospitals' private networks to enable fast diagnostic decision making while preserving accuracy and avoiding the need to access the cloud contin-uously. Second, we will investigate federated learning [12] approaches to train IMPACT-X with data coming from several organizations without the need to exchange patients' sensitive information between parties. This privacy-oriented method will improve model robustness and ensure that IMPACT-

X can generalize across a variety of settings and scanner configurations while strictly adhering to data governance rules. Third, we will pursue the expansion of IMPACT-X towards personalized medicine and therapy selection based on patient-specific needs. In addition to diagnostics, further development will incorporate predicting patient-specific responses to treat-ment using causal reasoning principles to derive personalized recommendations on what therapies should be prescribed. Using pharmacogenomics and physiology modeling, IMPACT-X could contribute to the emerging domain of precision oncol-ogy. Moreover, we are interested in incorporating longitudinal data for analyzing patients' condition development over time.

CONCLUSION

This paper introduced IMPACT-X, a groundbreaking causally-grounded interpretable multimodal deep learning framework, capable of overcoming the limitations of existing healthcare black-box artificial intelligence. Thanks to a novel Causal Multimodal Fusion Layer, IMPACT-X demonstrated record-breaking diagnostic accuracy over 95% across various medical fields ranging from oncology, cardiology to neu-rodegenerative diseases. The key scientific breakthrough of IMPACT-X is the successful combination of multimodal data in a manner that supports causal reasoning and guarantees that the algorithm makes decisions based on biological biomarkers, rather than spurious statistics often seen in observational datasets.

Interpretability is another crucial component of IMPACT-X. Unlike typical approaches to interpretation, which pro-vide post-hoc explanations of a trained model, IMPACT-X incorporates interpretability into its design thanks to cross-modal alignment loss and explanation modules. The latter enables the generation of consistent explanations for all data types involved: imaging, clinical, and genomic. Instead of generating inconsistent, conflicting evidence about the causes of an illness, the algorithm highlights specific pathological zones in

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

imaging results, laboratory values in clinical records, and genetic mutations in pathway diagrams. Clinicians can easily validate the explanations and reduce their cognitive overload.

Finally, IMPACT-X takes trust into consideration with spe-cial uncertainty estimation and validation techniques. The model quantifies its epistemic uncertainty with Monte Carlo dropout [11], thus being able to identify cases where it cannot make correct predictions. In turn, human review allows avoiding adverse outcomes that would otherwise be possible in such cases. Surveys of clinicians showed increased willingness to use the algorithm with causally-valid explanations compared to attention maps and SHAP values. This factor is crucial for meeting legal and ethical requirements and minimizing liabil-ity risk. IMPACT-X fills an important gap between cutting-edge machine learning solutions and clinical application by combining accuracy and interpretability. Thus, it paves the way for a new generation of AI algorithms that are not only powerful but also interpretable, trustful, ethically sound, and usable in practice.

REFERENCES

1. Esteva et al., "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.
2. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Int. Conf. Learning Representations (ICLR), 2021.
3. K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clin-ical notes and predicting hospital readmission," Journal of Biomedical Informatics, vol. 112, p. 103609, 2020.
4. M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," Bioinformatics, vol. 34, no. 13, pp. i457–i466, 2018.
5. J. Pearl and D. Mackenzie, The Book of Why: The New Science of Cause and Effect. Basic Books, 2018.
6. X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "DAGs with NO TEARS: Continuous optimization for structure learning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 31, pp. 9472–9483, 2018.
7. S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774, 2017.
8. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD, pp. 1135–1144, 2016.
9. R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, pp. 618–626, 2017.
10. Y. H. H. Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in Proc. 57th Annual Meeting of the ACL, pp. 6558–6569, 2019.
11. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. 33rd ICML, vol. 48, pp. 1050–1059, 2016.
12. N. Rieke et al., "The future of digital health with federated learning," NPJ Digital Medicine, vol. 3, no. 1, p. 119, 2020.
13. E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, vol. 25, no. 1, pp. 44–56, 2019.
14. G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.
15. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for EHR analysis," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, 2018.
16. J. Zhou et al., "Graph neural networks: A review of methods and applications," AI Open, vol. 1, pp. 57–81, 2021.
17. J. Peters, D. Janzing, and B. Scho¨lkopf, Elements of Causal Inference: Foundations and Learning Algorithms. MIT Press, 2017.
18. K. Yu, S. Budhathoki, and B. Scho¨lkopf, "Causal discovery and infer-ence: concepts and recent methodological advances," Applied Informat-ics, vol. 3, no. 1, pp. 1–28, 2021.
19. W. Samek, T. Wiegand, and K. R. Mu¨ller, "Explainable artificial intelligence: Understanding,

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

visualizing and interpreting deep learning models," arXiv preprint arXiv:1708.08296, 2017.
20. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Con-cepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020.
21. J. Vickers and E. B. Elkin, "Decision curve analysis: a novel method for evaluating prediction models," Medical Decision Making, vol. 26, no. 6, pp. 565–574, 2006.
22. R. J. Chen et al., "Whole slide images are 2D point clouds: Context-aware survival prediction using patch-based graph convolutional net-works," in MICCAI, pp. 339–349, 2021.
23. N. K. Tomar et al., "MMIT: Multi-modal medical image transformer for computer-aided diagnosis," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 4, pp. 1895–1906, 2023.
24. Y. Yang et al., "Causal inference in healthcare: A review of methods and applications," Journal of Biomedical Informatics, vol. 134, p. 104201, 2022.
25. M. Chen, S. Radhakrishnan, and F. Doshi-Velez, "Learning causal representations for robust domain adaptation," in Proc. CLeaR, vol. 172,
26. pp. 156–182, 2022.
27. T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Int. Conf. Learning Representations (ICLR), 2017.
28. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.
29. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Int. Conf. Learning Representations (ICLR), 2019.
30. J. Kelly et al., "Key challenges for delivering clinical impact with artificial intelligence," BMC Medicine, vol. 17, no. 1, p. 195, 2019.
31. Amann et al., "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," BMC Medical Informatics and Decision Making, vol. 20, no. 1, p. 310, 2020.