IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 April 2024, accepted 25 June 2024, date of publication 28 June 2024, date of current version 8 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3420444

APPLIED RESEARCH

VAE-Driven Multimodal Fusion for Early Cardiac Disease Detection

JUNXIN WANG, (Member, IEEE), JUANEN LI, RUI WANG, AND XINQI ZHOU
Faculty of Arts and Sciences, Beijing Normal University, Beijing 100875, China
Corresponding author: Junxin Wang (202111079160@mail.bnu.edu.cn)

ABSTRACT This study presents a novel multimodal deep learning model designed to improve early detection and diagnosis of chronic cardiac conditions such as Severe Left Ventricular Hypertrophy (SLVH) and Dilated Left Ventricle (DLV). Leveraging nearly 70,000 medical records from Columbia University Irving Medical Center, the model integrates early-stage CXR structured data and chest X-ray imagery, employing SMOTE to correct data imbalances. The model utilizes the pre-trained EfficientNetB3 for image feature extraction, enhanced with SE-Block and CBAM attention mechanisms, while Transformer Encoder layers enrich the structured data representation. Notably, it incorporates Variational Autoencoders (VAEs) to encode both types of data into a cohesive low-dimensional latent space, facilitating an innovative multimodal fusion for cardiac disease risk classification. Ablation studies validate the essential role of each component, with VAE-driven feature fusion significantly boosting accuracy and stability (increasing by 5.43% for SLVH and 14.13% for DLV datasets). The model outperforms existing advanced multimodal frameworks, showing a marked improvement in accuracy, recall, precision, and F1 scores. Specifically, it surpasses the leading CLIP model by 1.56% and 0.68% in accuracy for 90—270 day SLVH and DLV datasets, respectively. High AUC values across various disease stages highlight the model's robustness, demonstrating consistently superior performance in disease progression prediction. These results underscore the potential of integrating multimodal data with advanced deep learning techniques to significantly enhance the diagnostic capabilities of medical tools, paving the way for better early cardiac disease interventions and patient outcomes.

INDEX TERMS Multimodal deep learning, early cardiac disease detection, SMOTE, EfficientNetB3, Transformer encoder, variational autoencoders embedding, medical diagnostics.

I. INTRODUCTION

The critical importance of early detection in cardiac care cannot be overstated. Timely diagnosis of cardiac conditions is pivotal for effective treatment planning, significantly reducing mortality rates and improving patient outcomes. In recent years, the advent of deep learning has poised to revolutionize the field of cardiac care, offering novel methodologies for accurate and early diagnosis. Deep learning's ability to analyze complex medical data sets, including imaging and electronic health records, has shown promising results in identifying subtle patterns that precede overt cardiac diseases [1], [2], [3].For instance, EfficientNetB3, a deep learning model known for its efficiency and accuracy in image processing, has been successfully applied in extracting features from chest X-ray images for cardiac disease prediction [4].

Despite the promising advancements brought about by deep learning in cardiac care, the field faces significant challenges, particularly concerning early disease prediction. One of the primary obstacles is the reliance on single-modality data analysis. Traditional diagnostic methods typically focus on a singular type of data, such as imaging or clinical measurements, which may not capture the multifaceted nature of cardiac diseases. This limitation becomes apparent in the context of diseases like severe left ventricular hypertrophy (SLVH) and dilated left ventricle (DLV), where the integration of diverse data types could significantly enhance diagnostic accuracy [5]. Furthermore, the current state-of-the-art models, including VisualBert and CLIP, although groundbreaking, fall short in seamlessly integrating multimodal data for cardiac disease prediction,

The associate editor coordinating the review of this manuscript and approving it for publication was Yudong Zhang.

highlighting a critical research gap that needs to be addressed [6], [7].

The paramount objective of this study is to address the prevailing challenges in early cardiac disease prediction by harnessing the power of multimodal deep learning. Specifically, the research aims to integrate chest X-rays and echocardiographic data within a unified framework to enhance the prediction accuracy of severe left ventricular hypertrophy (SLVH) and dilated left ventricle (DLV). This approach seeks to overcome the limitations of single-modality data analysis by capitalizing on the complementary information provided by these two diagnostic modalities. Through the development and implementation of an advanced deep learning model, this study endeavors to set a new benchmark in the precision of cardiac disease prediction, paving the way for timely and more effective clinical interventions.

Overall, we summarize our contributions as follows:

This study presents a novel multimodal deep learning framework tailored for the early detection and diagnosis of complex cardiac conditions, utilizing a unique integration of advanced computational techniques. Central to our innovation is the employment of Variational Autoencoders (VAEs) for the fusion of multimodal data [10], a method that stands out for its ability to encode heterogeneous data types—including chest X-ray imagery and structured echocardiographic measurements—into a coherent, low-dimensional latent space. This approach not only preserves essential diagnostic details but also enhances the model's capacity for interpreting and synthesizing diverse data streams.

To support the VAE's functionality, EfficientNetB3 is integrated to extract high-level features from chest X-ray images [4]. Its state-of-the-art efficiency in processing medical images ensures that intricate patterns indicative of cardiac abnormalities are captured with high fidelity. This detailed image analysis is complemented by sophisticated attention mechanisms, specifically SE-Block and CBAM [8], [9]. These mechanisms refine the feature maps generated by EfficientNetB3, focusing the model's capacity on the most clinically relevant features, thereby improving the predictive accuracy of subsequent analyses.

The strategic use of Transformer Encoder layers further enriches the representation of structured CXR data, enabling a more detailed and effective integration within the VAE framework. The culmination of these technologies in a unified model facilitates a groundbreaking approach to data fusion, leveraging the complementary strengths of each modality to offer a comprehensive, nuanced understanding of cardiac health.

II. LITERATURE

A. DEEP LEARNING IN CARDIOLOGY

The integration of deep learning (DL) into cardiology, particularly in diagnostic imaging, has revolutionized the way cardiovascular diseases are detected and diagnosed. Over the past three years, numerous studies have underscored the potential of DL to enhance the accuracy and efficiency of diagnostic processes. For example, Alshammari et al. demonstrated the effectiveness of machine learning algorithms in identifying cardiovascular diseases through echocardiography, suggesting a significant shift towards more automated and precise diagnostic methods [11].

Another notable advancement was made by Zhang et al. (2022), who developed a deep learning model capable of detecting myocardial infarction with higher accuracy than traditional methods. This study highlights the importance of DL in analyzing complex cardiac imaging data to provide early and accurate diagnoses [12].

Furthermore, the role of DL in enhancing cardiac MRI analysis was explored by Liu et al. (2021), who implemented a convolutional neural network (CNN) model to improve the detection of cardiac structural abnormalities. Their work emphasizes the growing reliance on DL technologies to interpret cardiac MRI data, offering new avenues for diagnosing heart diseases [13].

B. MULTIMODAL LEARNING

Multimodal learning, which involves integrating data from multiple sources or modalities, has emerged as a crucial approach within medical imaging, including cardiology. The synergy between different types of medical imaging data, such as echocardiography, MRI, and CT scans, provides a more comprehensive view of the heart's structure and function.

Ghosh and Jayanthi (2021) presented a novel approach to multimodal image fusion, combining images from different modalities to enhance the diagnostic process. Their work underscores the potential of multimodal learning to offer more detailed and accurate insights into cardiac health [14].

Additionally, the application of multimodal learning extends beyond image fusion. Hashmi et al. explored the integration of text and image data in diagnosing cardiac conditions, demonstrating the benefits of combining visual information with clinical notes to improve diagnostic accuracy [15].

Multimodal learning's significance is further highlighted in the context of interventional cardiology. The fusion of real-time imaging data with preoperative scans offers unparalleled guidance during procedures, enhancing surgical outcomes and patient safety [16], [17], [18].

C. STATE-OF-THE-ART MULTIMODAL MODELS

VisualBert has made significant strides in vision-language tasks, enabling more nuanced interactions between visual data and natural language. Oza and Kambli work, ''Pixels to Phrases: Evolution of Vision Language Models,'' highlights its utility in understanding complex visual scenes through natural language descriptions [19].Despite its advancements, VisualBert's limitations lie in its heavy reliance on large

annotated datasets and potential biases inherited from pre-training data [20], [21].

The combination of Residual Networks (ResNet) and Multi-Layer Perceptrons (MLP) has been applied across various domains, including medical imaging for disease detection [22], [23]. Baybars et al. demonstrate the effectiveness of ResNet combined with MLP in the detection of tongue anomalies, showca [24]. However, this approach can be computationally intensive and may require fine-tuning for specific application

CLIP (Contrastive Language–Image Pre-training)offers a versatile framework for understanding and generating natural language descriptions of images, bridging the gap between visual and textual data [25], [26]. It has shown promise in diverse applications, from enhancing search capabilities to aiding creative processes [27]. However, CLIP's performance can vary significantly across different datasets and tasks, indicating a need for adaptive approaches and broader training data [28].

MURAL (MUltimodal, MUltitask Representations Across Languages) significantly advances multimodal understanding and multitask learning, particularly in cross-linguistic contexts. By leveraging both image-caption pairs and billions of translation pairs, MURAL extends the capabilities of ALIGN, a prior state-of-the-art dual encoder learned from 1.8 billion noisy image-text pairs, to not only match or exceed ALIGN's performance on well-resourced languages but also significantly improve performance on under-resourced languages [29].Although MURAL advances multimodal understanding across languages, its application to medical imaging is limited. Specifically, MURAL's generalist approach might not capture the nuanced details critical in medical diagnostics. The model's reliance on extensive, diverse datasets poses challenges in the medical field, where data privacy and the specificity of medical terminologies are paramount. Adapting MURAL for medical use requires significant customization, including integrating domain-specific knowledge and ensuring high precision and sensitivity crucial for medical diagnostics.

D. RESEARCH GAP

Despite these advancements, a significant gap remains in the effective combination of radiographic and echocardiographic data for cardiac disease prediction. Current models tend to excel in either image analysis or textual data interpretation but often fall short in seamlessly integrating these two modalities [30]. The challenge lies in the development of models that can not only analyze multimodal data concurrently but also understand the intricate relationships between these data types in the context of cardiac health [31].

A more effective method is needed to leverage the complementary strengths of radiographic imaging, which provides detailed anatomical information, and echocardiography, which offers dynamic insights into cardiac function. Such an approach would enable a more holistic and accurate prediction of cardiac diseases, moving beyond the limitations of single-modality analysis [32].

III. METHODOLOGY

A. DATA COLLECTION AND PREPARATION

1) DATA COLLECTION

The initial dataset comprised over 70,000 medical records from Columbia University Irving Medical Center, spanning from January 2013 to August 2018 [33]. This vast collection encapsulated a diverse array of patient encounters, inclusive of varying degrees of cardiac health and disease progression. Our selection criterion aimed at isolating instances where patients had undergone both a chest X-ray and an echocardiogram within a 12-month period, ensuring a robust linkage between radiographic imagery and echocardiographic measurements. This filtration yielded 71,589 unique chest X-rays across 24,689 patients, each annotated with echocardiographic insights into left ventricular hypertrophy and dilated left ventricle conditions, among other pathologies.

2) DATA PREPARATION AND PREPROCESSING

The preparatory phase involved several critical steps to render the data amenable to high-throughput deep learning analysis, as seen in Figure 1:

①**Exclusion of Single Visits:** Records pertaining to patients with only a single hospital visit were omitted to focus on longitudinal health trajectories.

②**Labeling for Disease Progression:** The dataset was stratified based on the patient's transition across health states—never sick to sick (labeled as Yes) and never sick to not sick (labeled as No), providing a binary classification framework essential for training predictive models.

③**Temporal Grouping:** To capture the subtle nuances in the progression of diseases, we have stratified our data into six subsets based on the time intervals calculated from the initial detection of the disease to subsequent screenings. These intervals are segmented into 0 to 90 days, 90 to 270 days, 270 to 540 days, 540 to 900 days, 900 to 1440 days, and beyond 1440 days, with the division primarily influenced by the right-skewed distribution of time intervals. Given that our prognosis encompasses two distinct conditions, namely Severe Left Ventricular Hypertrophy (SLVH) and Left Ventricular Dilation (DLV), the dataset has been further delineated into twelve unique subsets, each representing different diseases across varying timeframes.

④**Addressing Data Imbalance:** Given the preponderance of non-disease instances (label No), we employed Synthetic Minority Over-sampling Technique (SMOTE) to enrich our dataset with 8,000 synthesized records mirroring the characteristics of the disease-present instances (label Yes), thereby rectifying the imbalance and enhancing the model's learning capacity.The following is the detailed usage process of the SMOTE algorithm [34], [35].

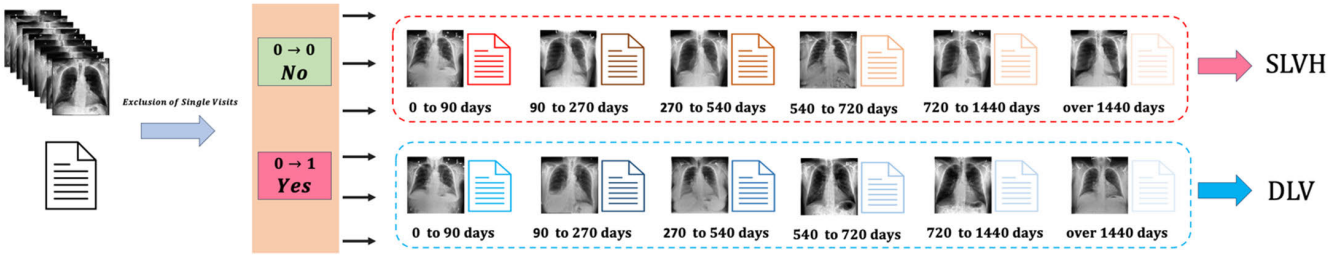

FIGURE 1. Image of the early-stage CXR structured data and chest X-ray image data preparation and preprocessing process.

For our dataset D consisting of N_{min} minority class samples, the SMOTE algorithm generates S synthetic samples as follows:

a) **For each sample x_i in the minority class**, find its k nearest neighbors in the minority class, forming a set $N_k(x_i)$.

b) **Synthetic Sample Generation:** For each x_i generate s_i.synthetic samples using the formula:

$$x_{new} = x_i + \lambda \cdot (x_z - x_i) \tag{1}$$

where x_z is a randomly selected neighbor from $N_k(x_i)$ and λ is a random number between 0 and 1.

c) **Repetition:**Repeat the process until S synthetic samples are generated.

The choice of k (the number of nearest neighbors) and S (the number of synthetic samples to be generated) are crucial hyperparameters that influence the effectiveness of SMOTE in addressing class imbalance [36].

For the subset of the data for each disease for each time period, our objective was to balance the distribution between positive (disease-present) and negative (disease-absent) instances. Given the significant imbalance, with positive instances being the minority, we applied SMOTE as follows:

- **Identification of Positive Instances:** We identified all instances labeled as (indicating the presence of either SLVH or DLV) in the dataset.
- **Nearest Neighbor Calculation:** For each positive instance, we calculated its nearest neighbors within the positive class, utilizing the Euclidean distance in the feature space composed of age, sex, and echocardiographic measurements (IVSd, LVIDd, LVPWd).
- **Synthetic Data Generation:** Following the SMOTE formula, we generated 8,000 synthetic positive instances to augment the original dataset, ensuring a balanced representation of both classes. This was achieved by interpolating between each positive instance and its nearest neighbors, introducing nuanced variations within the positive class.

⑤**Integration of Multimodal Data:**The training set was enriched with both image data (chest X-rays) and structured data (patient age, sex, and echocardiographic measurements—IVSd, LVIDd, LVPWd), laying the groundwork for a comprehensive multimodal learning approach.

B. EFFICIENTNETB3 FOR IMAGE PROCESSING

EfficientNet, introduced by Mingxing Tan and Quoc V. Le in their landmark paper, represents a paradigm shift in the design of convolutional neural networks (CNNs) through systematic scaling of network dimensions. The core principle of EfficientNet is to balance network depth, width, and resolution, which are crucial factors affecting the model's performance and efficiency. This balance is achieved by compound scaling, which uniformly scales these dimensions with a set of fixed coefficients, derived from a principled search using a simple yet effective compound coefficient.Meanwhile,Chest X-rays are rich in detail, demanding a model capable of discerning subtle patterns indicative of cardiac conditions. EfficientNetB3's depth and convolutional operations are adept at capturing these intricacies, translating into more accurate disease identification. Its architecture facilitates a comprehensive feature extraction process, crucial for the detection of conditions such as severe left ventricular hypertrophy (SLVH) and dilated left ventricle (DLV).

The compound scaling method is encapsulated by the formula:

$$\text{depth}: d = \alpha^{\phi} \tag{2}$$

$$\text{width}: w = \beta^{\phi} \tag{3}$$

$$\text{resolution}: r = \gamma^{\phi} \tag{4}$$

where:

- d, w, and r are factors to scale the network's depth, width, and resolution, respectively.
- ϕ is a user-specified coefficient that controls how much the network's resources are increased.
- α, β, and γ are constants that determine how to allocate resources efficiently to each of the dimensions, under the constraint that $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ and $\alpha \geq 1$, $\beta \geq 1$, $\gamma \geq 1$.

Applying EfficientNetB3 in our study, to the specific project involves several key steps(as seen in Figure 2), tailored to exploit the model's efficiency and accuracy in analyzing medical images:

①**Preprocessing and Input Configuration**

EfficientNet models, including B3, are pre-trained on ImageNet, requiring input images of a specific resolution. For EfficientNetB3, the input resolution is 300×300 pixels. In our project, images are resized to 224×224 for

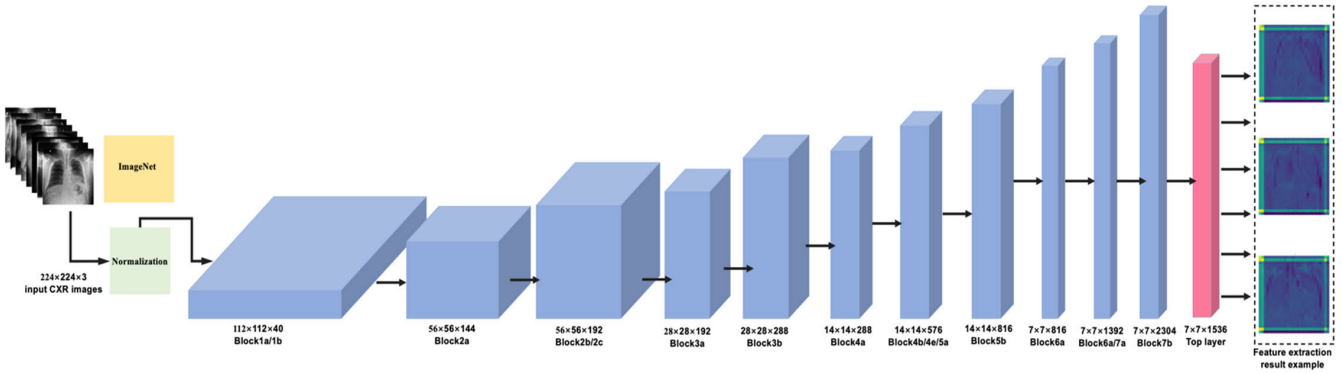

FIGURE 2. The application process of EfficientNetB3 pre-training model in our model framework.

compatibility:

$$\text{Image Preprocessing} : \text{Resizing to } 224 \times 224 \tag{5}$$

This step ensures that the input images match the network's expected input dimensions, allowing for optimal feature extraction without unnecessary distortion or information loss.

②Feature Extraction Layer

The EfficientNetB3 model is utilized as a feature extractor where the convolutional base is followed by custom top layers designed for the specific task. The model's output serves as an enhanced feature representation of the input images, capturing both high-level and fine-grained details relevant to identifying cardiac conditions.

③Integration with Attention Mechanisms

Post-EfficientNetB3 feature extraction, attention mechanisms such as SE-Block and CBAM are applied. These mechanisms refine the feature maps by emphasizing important features and suppressing irrelevant ones, enhancing the model's focus on critical image areas indicative of disease.

④Fusion with Structured Data

The extracted and attention-refined image features are then concatenated with processed structured data (e.g., clinical parameters like IVSd, LVPWd, LVIDd). This multimodal fusion leverages both the spatial characteristics from the X-rays and the clinical insights from structured data, providing a comprehensive feature set for classification.

⑤Classification and Model Training

A classification head, consisting of fully connected layers and activation functions, is appended to process the combined features. The model is trained using a binary cross-entropy loss function, with additional considerations for class imbalance addressed through techniques like SMOTE or class weighting.

$$\text{Loss Function} : \text{Binary Cross} - \text{Entropy}$$

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{y}_i\right) + (1 - y_i)\log\left(1 - \hat{y}_i\right)\right] \tag{6}$$

where:

- L is the loss function.
- N is the number of samples.
- y_i is the true label.
- $\hat{y}_i$ is the predicted label.

C. ATTENTION MECHANISM IMPLEMENTATION

The implementation of attention mechanisms such as Squeeze-and-Excitation (SE) Blocks and Convolutional Block Attention Module (CBAM) represents a sophisticated approach to enhance the representational power of convolutional neural networks (CNNs). These mechanisms focus the model's attention on relevant features within an image, significantly improving performance for complex tasks.

1) SQUEEZE-AND-EXCITATION (SE) BLOCK

The SE block re-calibrates channel-wise feature responses by explicitly modeling interdependencies between channels. The process can be distilled into two key operations: squeeze and excitation.

a: SQUEEZE OPERATION

The squeeze operation aggregates the spatial information of each channel into a single descriptor by employing global average pooling, reducing the feature map F∈R^(H×W×C) to a vector z∈R^C with its c-th element computed as:

$$z_c = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{ijc} \tag{7}$$

b: EXCITATION OPERATION

The excitation operation captures channel-wise dependencies through a self-gating mechanism, consisting of two fully connected (FC) layers and a sigmoid activation:

$$s = \sigma\left(g\left(z, W\right)\right) = \sigma\left(W_2\delta\left(W_1 z\right)\right) \tag{8}$$

Here, σ denotes the sigmoid activation function, δ denotes the ReLU activation, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ represent the weights of the FC layers, and r is the reduction ratio, controlling the capacity and computational cost.

The final output of the SE block, $\tilde{X}$, is obtained by rescaling the original feature map F with the activations s:

$$\tilde{X}_c = F_c \cdot s_c \tag{9}$$

2) CONVOLUTIONAL BLOCK ATTENTION MODULE (CBAM)

CBAM sequentially applies channel and spatial attention mechanisms to refine the feature map based on inter-channel and spatial relationships, enhancing the model's focus on informative features.

a: CHANNEL ATTENTION

Channel attention focuses on meaningful channels by exploiting the global spatial information of feature maps. It is computed as the sum of max-pooling and average-pooling operations followed by a shared MLP:

$$M_c(F) = \sigma\left(MLP\left(AvgPool\left(F\right)\right) + MLP\left(MaxPool\left(F\right)\right)\right) \tag{10}$$

where *AvgPool* and *MaxPool* are global average and max pooling operations across spatial dimensions, respectively, and σ is the sigmoid function.

b: SPATIAL ATTENTION

Spatial attention highlights informative spatial locations by using the channel-wise aggregated information:

$$M_s(F) = \sigma\left(f^{7\times7}\left(AvgPool_c\left(F\right) \oplus MaxPool_c\left(F\right)\right)\right) \tag{11}$$

where $f^{7\times7}$ represents a convolution operation with a filter size of 7×7, $AvgPool_c$ and $MaxPool_c$ are average and max pooling operations along the channel axis, and $\oplus$ denotes concatenation

In our study, SE-Block and CBAM are integrated right after the feature extraction layer of EfficientNetB3. This strategic placement ensures that the refined feature maps, emphasizing critical regions and channels relevant to identifying cardiac conditions, are utilized for subsequent analysis and classification. The combination of these attention mechanisms not only boosts the model's interpretative ability but also aligns with the intricate requirements of medical imaging tasks, where discerning subtle features can be crucial for accurate diagnosis.

D. TRANSFORMER ENCODER FOR STRUCTURED DATA

The adaptation of Transformer encoders for processing structured echocardiographic data in our projects described encapsulates a pivotal advancement in leveraging deep learning to capture complex patterns within multimodal datasets. The Transformer encoder, initially conceived for natural language processing tasks, has been reimagined for its utility in analyzing structured medical data, offering a nuanced approach to understanding the intricate relationships between different clinical measurements.

The Transformer encoder is built upon the principle of self-attention, enabling the model to weigh the importance

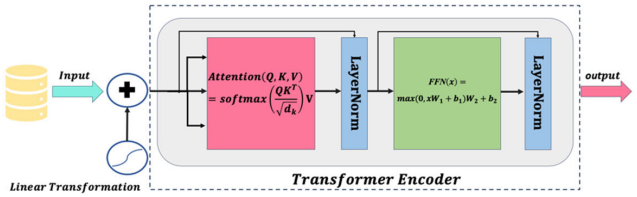

FIGURE 3. The schematic diagram of transformer encoder processing structured data.

of different input features dynamically. This mechanism is particularly adept at handling sequential or structured data, where the inter-feature relationships contribute significantly to the predictive outcome.

Formulation of Transformer Encoder for Structured Data: Given a set of structured input features $X \in \mathbb{R}^{N\times F}$, where N is the number of samples and F is the number of features (e.g., echocardiographic measurements such as IVSd, LVPWd, LVIDd), the Transformer encoder processes this data through the following stages (as seen in Figure 3):

①**Input Linear Transformation:**The input features are first linearly transformed to higher-dimensional space to facilitate more complex interactions:

$$X' = XW^e + b^e \tag{12}$$

where $W^e \in \mathbb{R}^{F\times D}$ and $b^e \in \mathbb{R}^D$ are the weights and bias of the linear transformation, respectively, and D is the dimensionality of the transformed space.

②**Self-Attention Mechanism:**The core of the Transformer encoder is the self-attention mechanism, which allows each feature to interact with every other feature, weighted by their calculated significance:

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{13}$$

Here, $Q = X'W^Q$, $K = X'W^K$, and $V = X'W^V$ are the query, key, and value matrices obtained by projecting X' onto different spaces, and d_k is the dimensionality of the key vectors, used for scaling.

③**Position-wise Feed-forward Networks:**The output from the self-attention mechanism is then passed through a position-wise feed-forward network (FFN) for each position separately and identically:

$$\text{FFN}\left(x\right) = \max\left(0, xW_1 + b_1\right)W_2 + b_2 \tag{14}$$

This non-linear transformation further enhances the model's ability to learn complex patterns.

④**Layer Normalization and Residual Connections:**Both the attention outputs and the FFN outputs are supplemented with residual connections followed by layer normalization:

$$\text{LayerNorm}\left(x + \text{Sublayer}\left(x\right)\right) \tag{15}$$

These components help in stabilizing the learning process and improving the model's convergence.

In the context of analyzing echocardiographic measurements, the Transformer encoder's capacity to discern intricate patterns and relationships among the structured data becomes invaluable. By treating each echocardiographic measurement as a distinct feature within the input space, the Transformer encoder can effectively identify and amplify the salient signals predictive of cardiac conditions.The practical implementation involves feeding the preprocessed and standardized echocardiographic measurements as input to the Transformer encoder, configured with an appropriate number of heads and dimensionality to match the complexity of the data. The encoder's output serves as a rich, contextualized representation of the structured data, which is then concatenated with features extracted from corresponding chest X-ray images via models like EfficientNetB3, further augmented with attention mechanisms such as SE-Blocks and CBAM.

E. VARIATIONAL AUTOENCODER EMBEDDING

The integration of Variational Autoencoders (VAEs) into our projects represents a significant leap in embedding high-dimensional data into a more tractable, low-dimensional latent space. This process is crucial for the nuanced fusion of multimodal data, enabling a more effective amalgamation of features from distinct data sources. VAEs, by design, offer a probabilistic manner for describing an observation in latent space, thereby facilitating this dimensional reduction with an emphasis on generating new data points with similar characteristics

VAEs consist of two main components: an encoder and a decoder. The encoder maps inputs to a latent distribution parameterized by mean (μ) and variance (σ^2), while the decoder reconstructs the input data from this latent representation

1) ENCODER

Given an input x, the encoder produces two parameters in a latent space, z, which are μ and σ^2, representing the mean and variance, respectively:

$$\mu, \sigma = f_{encoder}(x) \tag{16}$$

where $f_{encoder}$ is a neural network.

2) REPARAMETERIZATION TRICK

To enable backpropagation through random nodes, VAEs employ the reparameterization trick, where a sample z from the latent space is expressed as:

$$z = \mu + \sigma \odot \epsilon \tag{17}$$

Here, ϵ is an element-wise product with a random noise sampled from a standard normal distribution, $\mathcal{N}(0, I)$.

3) DECODER

The decoder part of the VAE takes the latent representation z and reconstructs the input x':

$$x' = f_{decoder}(z) \tag{18}$$

where $f_{decoder}$ is another neural network.

TABLE 1. Specific parameters of the VAE module.

Module	Layer	Activation function	Number of neurons
Encoder first layer	Dense 1	Relu	128
Encoder second layer	Dense 2	\	128(latent_dim*2)
Decoder first layer	Dense 1	Relu	128
Decoder second layer	Dense 2	\	Input dimension

4) LOSS FUNCTION

The VAE is trained to minimize the reconstruction loss between the input and output and a regularization term given by the Kullback-Leibler (KL) divergence, which enforces the latent space to approximate a standard normal distribution:

$$\mathcal{L}(x, x') = -\mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] + KL\left(q(z|x) \parallel p(z)\right) \tag{19}$$

where:

1) The first term is the reconstruction loss (e.g., binary cross-entropy or mean squared error).
2) The second term, $KL(q(z|x) \parallel p(z))$, is the KL divergence between the learned latent distribution $q(z|x)$ and the prior distribution $p(z)$, typically assumed to be a standard normal distribution $\mathcal{N}(0, I)$.

Design Method of the VAE Model (Module parameters can be seen in Table 1):

- Latent Space Dimension (latent_dim):64 is chosen as the dimension of the latent space to ensure that the latent variables can effectively represent the input data while controlling the complexity of the model.
- Encoder Design:The encoder is designed with two fully connected layers (Dense Layers). The first layer is used to extract high-dimensional features, and the second layer generates the mean and logarithm of the variance for the latent space.
- Reparameterization Trick:Sampling is performed based on the mean and logarithm of the variance output by the second layer, which is essential for implementing the core functionality of the variational autoencoder.
- Decoder Design:The decoder uses fully connected layers symmetric to the encoder. It restores the variables in the latent space to the same dimensions as the input data, ensuring that the input data can be reconstructed.
- KL Divergence Loss Calculation:KL divergence loss is added to the total loss of the model to ensure that the distribution of the latent space approaches a standard normal distribution.

In the context of our study, VAEs are employed to embed both the high-dimensional features extracted from chest X-ray images and structured echocardiographic measurements into a cohesive, low-dimensional latent space. This embedding process allows for the efficient integration of heterogeneous data modalities, enhancing the model's

capacity to capture and leverage complex, multimodal patterns indicative of cardiac conditions (as seen in Figure 4).

F. FORMULATION OF THE MULTIMODAL FUSION STRATEGY

The Innovative Multimodal Fusion Strategy in the context of deep learning leverages the strengths of Variational Autoencoders (VAEs), Transformer Encoders, and attention mechanisms (such as SE-Blocks and CBAM) to create a sophisticated method for combining and processing multimodal data. This strategy is pivotal for tasks like disease risk prediction, where integrating high-dimensional image data with structured clinical measurements is crucial for accurate medical diagnosis.

The multimodal fusion strategy involves several key components (as seen in Figure 5), each contributing uniquely to the model's ability to understand and synthesize information from diverse data sources effectively. The data structure and dimension of each neural network module can be seen in Table 2.

①Variational Autoencoder (VAE) Embedding

VAEs are employed to map both image features and structured data into a latent space. The encoder part of the VAE for each modality is defined as follows:

For image data (x_{img}) and structured data (x_{feat}):

1) ENCODER

$$q_\phi(z \mid x) = \mathcal{N}\left(z; \mu(x), \sigma^2(x)\right) \tag{20}$$

where $\mu(x)$ and $\log\sigma^2(x)$ are outputs of dense layers applied to the input data x, representing the mean and log variance of the latent distribution.

2) REPARAMETERIZATION

$$z = \mu + \sigma \odot \epsilon, \epsilon \sim \mathcal{N}(0, I) \tag{21}$$

3) KL DIVERGENCE LOSS

$$\mathcal{L}_{KL} = -\frac{1}{2}\sum_{k=1}^{K}\left(1 + \log\left(\sigma_k^2\right) - \mu_k^2 - \sigma_k^2\right) \tag{22}$$

②Attention Mechanisms

SE-Block recalibrates channel-wise features by applying a squeeze and excitation operation, effectively allowing the network to perform dynamic channel-wise feature recalibration.

CBAM sequentially applies channel and spatial attention mechanisms, enhancing the representation of important features while suppressing less useful ones.

③Transformer Encoder for Structured Data

The Transformer encoder captures complex patterns and relationships within the structured data through self-attention mechanisms:

Self-Attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{23}$$

where Q, K, and V are the query, key, and value matrices derived from the input data, and d_k is the scaling factor.

Position-wise Feed-forward Networks: Enhance the representation capability for each position in the sequence independently.

④Fusion of Embedded Features

The latent representations from the VAEs (z_{img} and z_{feat}) are concatenated to form a unified feature vector:

Combined Embedding:

$$z_{combined} = [z_{img}; z_{feat}] \tag{24}$$

This combined embedding is then passed through dense layers for classification.

In our study, this fusion strategy is meticulously applied as follows:

- **Image Features**: Extracted via EfficientNetB3, enhanced by SE-Block and CBAM for attention-focused feature refinement.
- **Structured Data**: Processed through a Transformer Encoder to capture complex patterns in clinical measurements.
- **VAE Embeddings**: Both sets of features are embedded into a latent space using VAEs, ensuring that the multimodal data is represented in a form that facilitates effective fusion.
- **Fusion and Classification**: The concatenated embeddings form a comprehensive feature set, which is then utilized for the final classification task, predicting the presence of cardiac conditions.

This innovative fusion strategy not only capitalizes on the unique strengths of each component—VAE embeddings for dimensionality reduction and generative representation, attention mechanisms for feature refinement, and Transformer encoders for capturing sequential relationships—but also harmonizes these elements to maximize the predictive performance of the our model.

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENT 1: SINGLE DATA SOURCE MODELS

The aim of Experiment 1 was to methodically assess the predictive performance of models using singular data sources—namely structured clinical data and chest X-ray (CXR) images—for the classification of markers indicative of early-stage cardiac diseases, specifically Severe Left Ventricular Hypertrophy (SLVH) and Dilated Left Ventricle (DLV), within a delineated timeframe of 90 to 270 days.

A bifurcated analytical approach was adopted:

- **Structured Data Model:**This model was architected employing a neural network framework [37] tailored to process structured clinical variables (age, sex, interventricular septal thickness at end-diastole [IVSD], left ventricular posterior wall diameter [LVPWD], and left ventricular internal diameter at end-diastole [LVIDD]). The model's design was oriented towards extracting predictive insights from clinical parameters critical in the early diagnosis of cardiac pathologies.

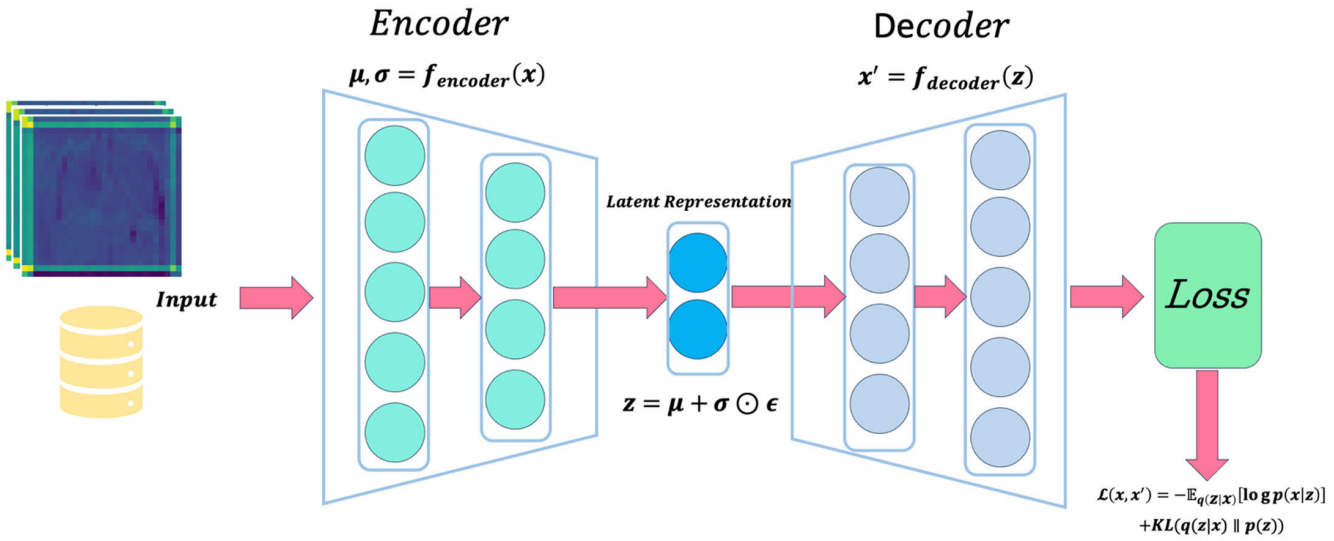

FIGURE 4. Schematic diagram of variational autoencoder (VAE) embeddings encoding image features and reconstructed data features.

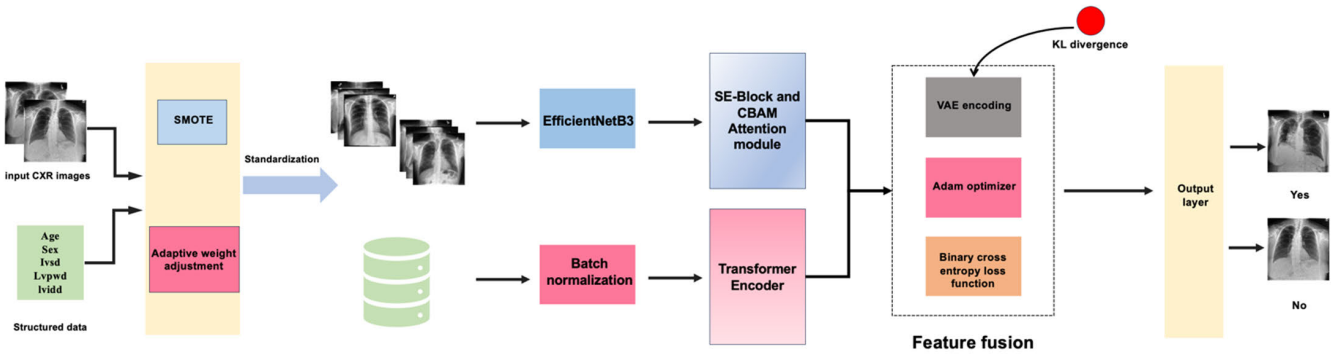

FIGURE 5. Framework diagram of multi-modal deep learning fusion strategy.

- **Image Data Model:**For the analysis of CXR images, the EfficientNet architecture was employed, chosen for its demonstrated efficacy in balancing model depth with computational efficiency [38], [39]. This model was expected to leverage the nuanced visual indicators present in CXR images that signify early cardiac abnormalities.

Both models were subjected to a comprehensive evaluation regimen encompassing training, validation, and testing to ensure statistical rigor and generalizability of the findings.

1) RESULTS

Analysis: The structured data model showcased robust predictive capabilities across both disease conditions, underpinned by high accuracy values, indicative of the model's discriminative power between disease and no-disease states. Conversely, while the image data model exhibited superior precision, particularly in the context of SLVH disease, it manifested limitations in recall. This discrepancy underscores a critical shortfall in using image data in isolation: a propensity to overlook positive cases, thereby necessitating the exploration of integrative approaches that amalgamate the strengths of both data modalities (as seen in Table 3, Figure 6, 7).

The differential performance of the models, especially in recall metrics, accentuates the inherent challenge in singular data source utilization for early disease detection. Structured data, while offering a broad spectrum of clinical insights, may lack the specificity afforded by the granular visual details in CXR images. However, the image data model's lower recall rate signals a crucial need for augmenting specificity without compromising sensitivity.

2) CONCLUSION

Experiment 1 revealed the comparative limitations of using single-source data models, underscoring the complexity of cardiac disease markers which cannot be fully captured through either structured data or image data alone. While structured data models exhibited high accuracy, they lacked the nuanced visual analysis capability that image data models

TABLE 2. Structure, dimensions, and meanings of input and output signals for each neural network module.

Module Name	Input Signal Structure	Input Signal Dimension	Output Dimension	Meaning
EfficientNetB3	Image Input	(224, 224, 3)	(7, 7, 1536)	7x7 feature map, 1536 channels
SE-Block	EfficientNetB3 Output	(7, 7, 1536)	(7, 7, 1536)	Same as above
CBAM-Block	SE-Block Output	(7, 7, 1536)	(7, 7, 1536)	Same as above
GlobalAveragePooling2D	CBAM-Block Output	(7, 7, 1536)	(1536,)	Vector, 1536 elements
Dense (x_feat 1)	Feature Input	(3,)	(64,)	Vector, 64 elements
BatchNormalization	Dense (x_feat 1) Output	(64,)	(64,)	Vector, 64 elements
Dropout	BatchNormalization Output	(64,)	(64,)	Vector, 64 elements
Reshape	Dropout Output	(64,)	(1, 64)	Matrix, 1x64
Transformer Encoder	Reshape Output	(1, 64)	(1, 64)	Same as above
Flatten	Transformer Encoder Output	(1, 64)	(64,)	Vector, 64 elements
VAE Encoder (Image)	GlobalAveragePooling2D Output	(1536,)	(64,)	Vector, 64 elements
VAE Encoder (Feature)	Flatten Output	(64,)	(64,)	Vector, 64 elements
Concatenate	VAE Encoder (Image) and (Feature) Output	[(64,), (64,)]	(128,)	Vector, 128 elements
Dense (combined_embedded)	Concatenate Output	(128,)	(256,)	Vector, 256 elements
BatchNormalization	Dense (combined_embedded) Output	(256,)	(256,)	Vector, 256 elements
Dropout	BatchNormalization Output	(256,)	(256,)	Vector, 256 elements
Dense (predictions)	Dropout Output	(256,)	(1,)	Scalar, 1 element (classification probability)

TABLE 3. Predicting 90 to 270-day SLVH and DLV disease model outcomes using a single data source.

Data	Accuracy	Recall	Precision	F1
SLVH Structured data	0.9467	0.9424	0.9034	0.9225
SLVH CXR images	0.9680	0.9155	0.9886	0.9507
DLV Structured data	0.9467	0.9424	0.9034	0.9225
DLV CXR images	0.9680	0.9155	09886	0.9507

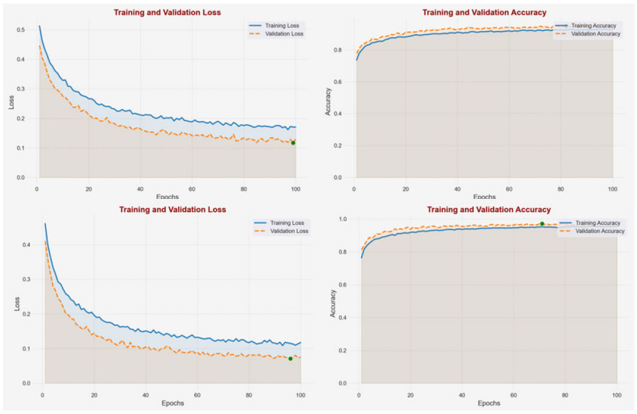

FIGURE 6. Loss and accuracy trends during model training using structured data to predict the disease risk of SLVH disease (top image) and DLV disease (bottom image) from 90 to 270 days.

provided, albeit with lower recall rates. This discrepancy highlights the complementary nature of the two data types and the necessity for their integration to achieve comprehensive diagnostic performance.

B. EXPERIMENT 2: ABLATION STUDY

The crux of our ablation study lies in deconstructing our composite model to elucidate the individual contributions of its core components towards the overall predictive performance, particularly in the nuanced realm of cardiac disease detection within the specified temporal window of 90 to 270 days for SLVH and DLV diseases.

1) OBJECTIVE

This study meticulously dissects our model to validate the indispensability and efficacy of its architectural innovations, namely the SE-block and CBAM attention mechanisms, the

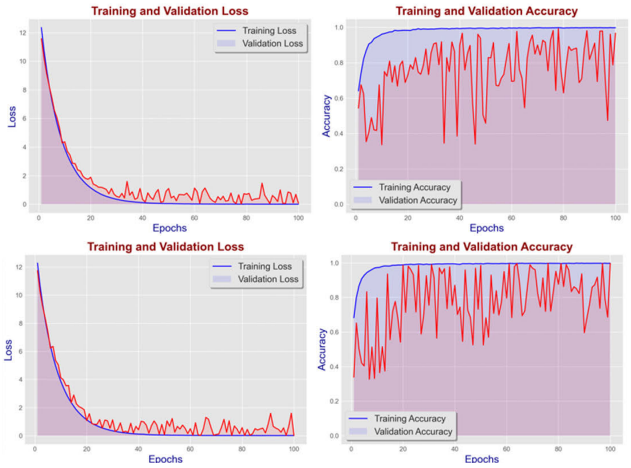

FIGURE 7. Loss and accuracy trends during model training using CXR image data to predict the disease risk of SLVH disease (top image) and DLV disease (bottom image) from 90 to 270 days.

Transformer Encoder layer, and the Variational Autoencoder (VAE)-based feature embedding.

2) METHODOLOGY

Adopting a methodical approach, we sequentially dismantled each aforementioned component from our full-fledged model to discern their singular effects on the model's precision, recall, and F1 score. This systematic removal not only sheds light on the functionality and performance enhancement attributed to each component but also underscores their collective synergy in achieving state-of-the-art results.

3) RESULTS AND DISCUSSION

Conclusion: **Experiment 2**'s ablation study provided concrete evidence of the integral roles played by each component of our model's architecture. The omission of SE-Block and CBAM significantly impacted the model's ability to focus on and refine critical features within X-ray images, resulting in a notable decrease in recall and precision. Similarly, the exclusion of the Transformer Encoder and VAE embeddings compromised the model's ability to synthesize contextual information from structured data and efficiently encode multimodal data into a cohesive feature space, respectively(as seen in Table 4, 5). These findings affirm the architectural choices made in developing our model, each contributing uniquely to its overall efficacy.

C. EXPERIMENT 3: COMPARISON WITH POPULAR MULTI-MODAL MODELS

In the context of leveraging multimodal data for early cardiac disease detection, this study introduces a novel deep learning architecture that significantly advances the state of the art. Our model is meticulously evaluated against four established multimodal models: VisualBert, Resnet+MLP and CLIP model. This comparative analysis spans two distinct datasets, corresponding to Severe Left Ventricular Hypertrophy (SLVH) and Dilated Left Ventricle (DLV) within a critical timeframe of 90 to 270 days.

TABLE 4. Model results from an ablation experiment using multi-source data to predict 90- to 270-day SLVH disease risk.

Ablation component	Accuracy	Recall	Precision	F1
SE-Block & CBAM	0.9641	0.8992	0.9936	0.9441
Transformer Encoder	0.9353	0.8534	0.9834	0.9245
VAE feature embedding	0.9409	0.8647	0.9555	0.9078
Our complete model	0.9952	0.9923	0.9933	0.9928

TABLE 5. Model results from an ablation experiment using multi-source data to predict 90- to 270-day DLV disease risk.

Ablation component	Accuracy	Recall	Precision	F1
SE-Block & CBAM	0.9667	0.8958	0.9945	0.9448
Transformer Encoder	0.9347	0.8633	0.9743	0.9487
VAE feature embedding	0.8551	0.9647	0.6985	0.8103
Our complete model	0.9964	0.9899	0.998	0.9944

Methodological Overview: **VisualBert** integrates visual and textual cues using a transformer-based approach, leveraging the inherent strengths of transformers in handling sequential data. Despite its proficiency in extracting complex inter-modal relationships, its application to clinical imagery and structured data presents challenges in terms of computational intensity and adaptability to specific medical contexts.

Resnet+MLP combines the deep feature extraction capabilities of Resnet with the versatility of MLPs to process structured data. While effective, this approach often falls short in fully capturing the nuanced correlations between clinical parameters and visual markers inherent in medical images.

CLIP employs contrastive learning to align image and text representations in a shared embedding space. Although CLIP demonstrates remarkable generalization across diverse visual tasks, its performance in medical applications is constrained by the specificity and complexity of clinical images and annotations.

Our proposed model outstrips the aforementioned frameworks across key performance metrics—accuracy, precision, recall, and F1 score—while maintaining exemplary computational efficiency(as seen in Table 6, 7). Distinct advantages of our model include:

TABLE 6. Comparative results with other models for predicting 90 to 270-day SLVH disease risk using multi-source data.

Model	Accuracy	Recall	Precision	F1
Visual Bert	0.9512	0.9368	0.9535	0.9444
Resnet+MLP	0.9221	0.9041	0.8599	0.8815
CLIP	0.9802	0.9827	0.9597	0.9710
Our model	0.9958	0.9933	0.9942	0.9938

TABLE 7. Comparative results with other models for predicting 90 to 270-day DLV disease risk using multi-source data.

Model	Accuracy	Recall	Precision	F1
Visual Bert	0.9757	0.9634	0.9743	0.9688
Resnet+MLP	0.9134	0.9042	0.8601	0.8911
CLIP	0.9896	0.9879	0.9800	0.9839
Our model	0.9964	0.9899	0.9980	0.9942

- **Elevated Performance Metrics:**Across both SLVH and DLV conditions, our model consistently achieves superior performance. This is attributable to its innovative data fusion strategy, which effectively integrates disparate data modalities, enhancing the model's predictive accuracy and reliability(as seen in Figure 8, 9, 10)
- **Computational Efficiency:** Unlike the Multimodal MURAL Model, our architecture is optimized for reduced computational load without sacrificing performance. This efficiency enables its application in real-time clinical diagnostics, a crucial factor for early disease detection and intervention.
- Enhanced Data Integration: By synthesizing insights from both structured clinical data and medical imagery, our model captures a comprehensive view of disease markers. This holistic approach ensures a nuanced analysis, pivotal for accurate disease characterization.
- **Generalizability and Scalability:**The utilization of pre-trained networks, tailored through fine-tuning to specific medical datasets, endows our model with robust generalizability. This facilitates its application across various cardiac conditions, underscoring its potential for broader clinical adoption.

Experiment 3 set our model in comparison with existing multi-modal models, including VisualBert, Resnet+MLP, and the CLIP model. Our model demonstrated superior performance across all metrics, with marked improvements in precision, recall, and F1 scores. Notably, our model achieved these results with significantly greater computational efficiency than the CLIP model, highlighting its suitability for real-world clinical applications where both accuracy and processing time are critical considerations.

D. EXPERIMENT 4: MULTIMODAL MODEL PERFORMANCE ACROSS VARIOUS DISEASE STAGE

Objective: This experiment aims to evaluate the robustness and adaptability of our proposed multimodal deep learning model, across different stages of two specific cardiac conditions: Severe Left Ventricular Hypertrophy (SLVH) and Dilated Left Ventricle (DLV). By assessing model performance over varied disease timelines, we seek to establish the model's efficacy in providing consistent and reliable risk predictions across progressive stages of cardiac diseases.

Data for both SLVH and DLV diseases were categorized into six distinct time intervals ranging from 0 to over 1440 days. The model was trained separately on each subgroup to predict the risk of disease progression. The model integrated structured clinical data with chest X-ray (CXR) imaging data, utilizing our established framework of SE-blocks, CBAM attention mechanisms, Transformer Encoder layers, and VAE-based feature embeddings. Model performance was quantitatively evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC), a robust indicator of diagnostic accuracy.

Results: The AUC values obtained from our model across different time intervals are visualized in Figure 3, providing a comparative insight into the model's performance across early to late stages of disease progression. The visualization highlights consistent high AUC scores(as seen in Figure 11), underscoring the model's capability to maintain high diagnostic precision irrespective of disease stage.

The analysis of the AUC trends revealed several key insights:

- Consistency Across Stages: Our model demonstrated robust performance with AUC values consistently above 0.98 across all time intervals for both diseases. This consistency is indicative of the model's strong generalizability and its effectiveness in capturing relevant disease markers at various stages of progression.
- Early Detection Capabilities: Notably, the model exhibited particularly high accuracy(as seen in Table 8) in the early stages (0-270 days), crucial for timely intervention and management of cardiac conditions. This suggests that the integrative data approach effectively captures early subtle changes in clinical and imaging markers.
- Performance in Chronic Stages: In the later stages (over 900 days), where clinical manifestations might be more pronounced, the model similarly maintained high accuracy, demonstrating its utility in ongoing disease monitoring and management.

Experiment 4 demonstrated the robustness of our model, across various stages of cardiac diseases, from acute to chronic phases. The model consistently achieved high AUC values, indicating its efficacy in detecting and monitoring Severe Left Ventricular Hypertrophy (SLVH) and Dilated Left Ventricle (DLV) over extended periods. This performance suggests our model's capability to adapt to evolving disease markers.

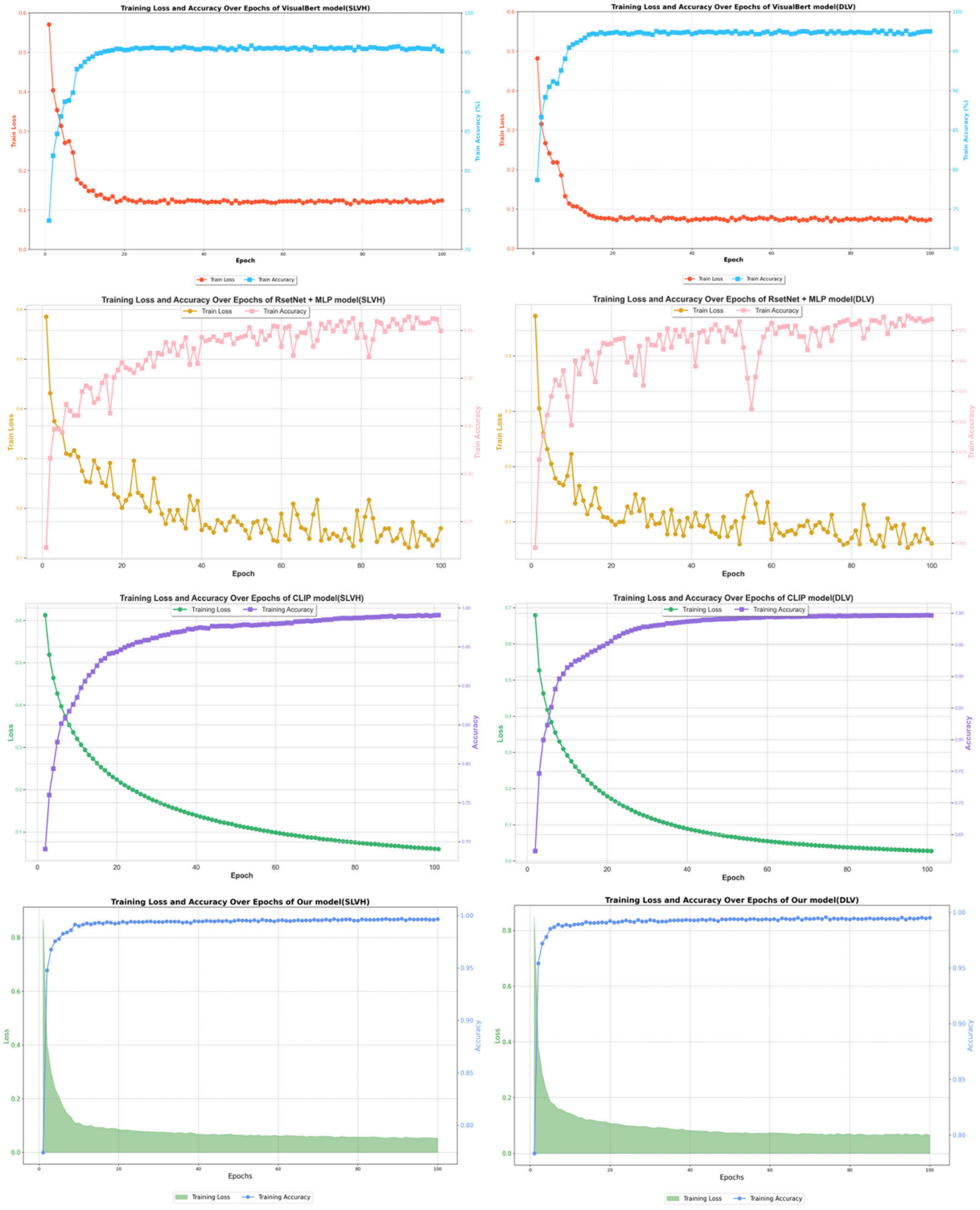

FIGURE 8. Each model (the first row is the VisualBert model, the second row is the Resnet+MLP model, the third row is the CLIP model, and the fourth row is our model) uses multimodal data to predict SLVH disease (left) and DLV disease (left) within 90 to 270 days (Right) Loss and accuracy trends during model training for disease risk.

V. DISCUSSION

The comprehensive suite of experiments conducted in this study delineates the efficacy of a novel multimodal deep learning model designed for early detection of cardiac diseases such as Severe Left Ventricular Hypertrophy (SLVH) and Dilated Left Ventricle (DLV). The integrated approach

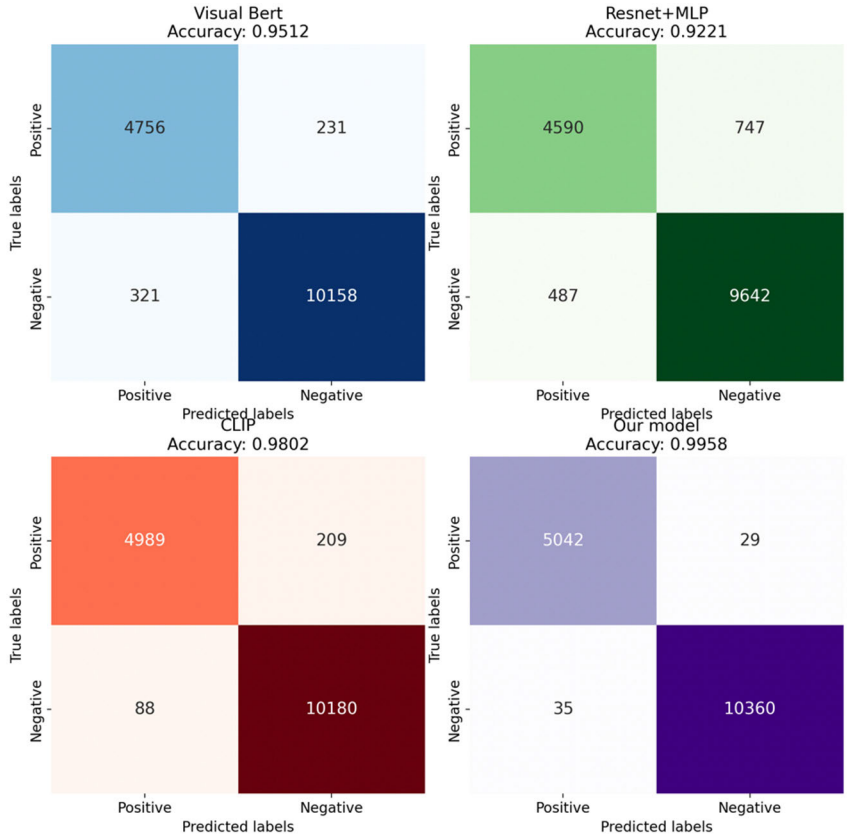

FIGURE 9. Mixture matrix of individual models predicting SLVH disease risk over 90 to 270 days using multimodal data.

TABLE 8. Model performance of our model at different disease stages in SLVH and DLV.

Segmented interval	Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1
	SLVH				DLV			
0-90 days	0.9987	0.9976	0.9976	0.9976	0.9971	0.9943	0.9929	0.9936
90-270 days	0.9952	0.9923	0.9933	0.9928	0.9964	0.9899	0.998	0.9944
270-540 days	0.9901	0.9780	0.9935	0.9935	0.9949	0.9878	0.9944	0.9911
540-900 days	0.9928	0.9866	0.9958	0.9912	0.9934	0.987	0.9953	0.9911
900-1440 days	0.9927	0.9862	0.9967	0.9914	0.9961	0.9907	0.9991	0.9949
Over 1440 days	0.9958	0.9953	0.9969	0.9961	0.9954	0.9916	0.9968	0.9942

employing Variational Autoencoders (VAEs) for data fusion underscores a significant advance in the field, enabling the nuanced integration of heterogeneous data types—namely structured clinical data and chest X-ray imagery.

A. DISCUSSION OF RESULTS

Our findings highlight the model's superior diagnostic accuracy compared to traditional single-source and other advanced multimodal models. This is evidenced by its consistently higher performance metrics across various datasets and disease stages. The integration of SE-Block and CBAM attention mechanisms, along with Transformer Encoders, effectively enhances the model's capability to discern and synthesize critical features from both structured and image data. The VAEs play a crucial role in encoding this information into a low-dimensional latent space, thereby preserving essential diagnostic details while facilitating an efficient computational process.

B. ADVANTAGES AND INNOVATIONS

The model's architecture leverages the strengths of each component to address previous limitations seen in cardiac

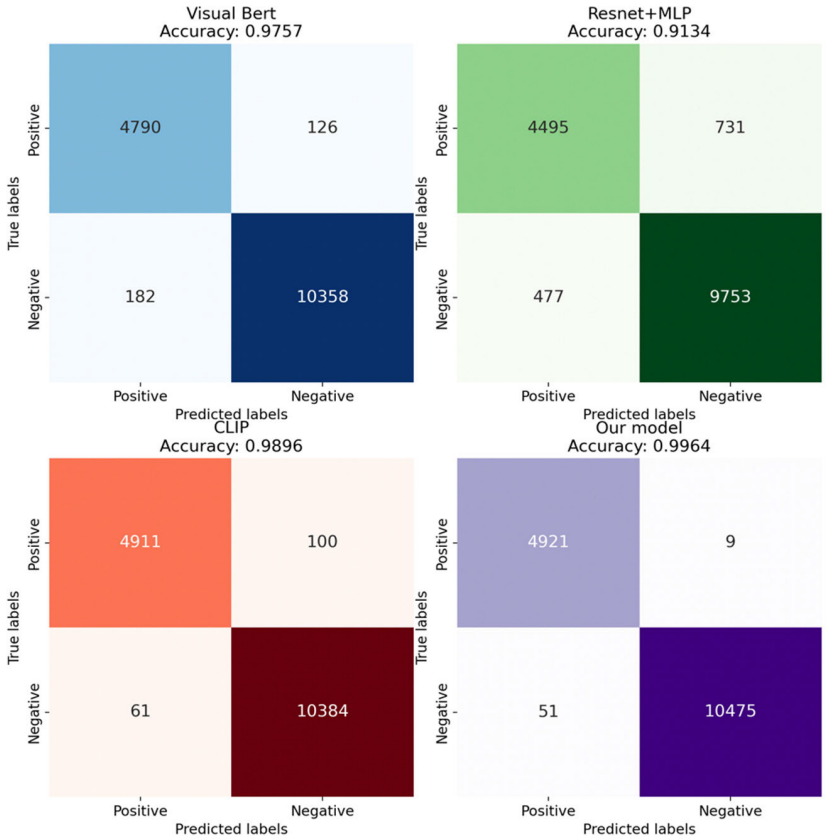

FIGURE 10. Mixture matrix of individual models predicting DLV disease risk over 90 to 270 days using multimodal data.

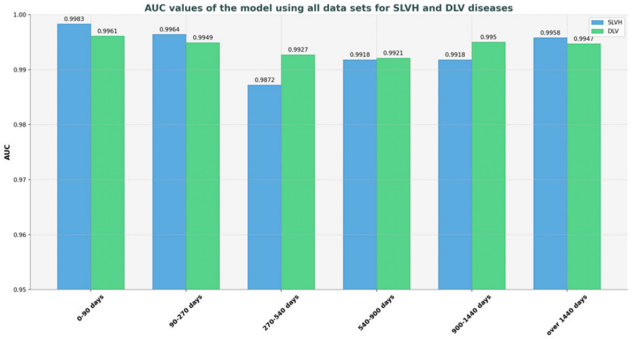

FIGURE 11. Visualization of AUC values for our model across different disease stages for SLVH and DLV.

diagnostics. By combining attention-driven feature refinement and sophisticated data embedding strategies, it achieves a significant improvement in accuracy and precision. On the other hand, the model demonstrates robust adaptability and reliability in predicting disease progression across different stages, which is vital for timely clinical intervention and ongoing patient management.

C. LIMITATIONS AND FUTURE WORK

1) LIMITATIONS

Despite achieving significant advancements in multimodal deep learning for medical diagnostics, our study encounters several limitations that warrant discussion. Firstly, the dataset utilized, while comprehensive, represents a confined scope, predominantly focusing on cardiac diseases within specific demographic and clinical settings [40],. The potential for model generalization across diverse populations and varied clinical conditions remains to be thoroughly evaluated through external validation [41].

Moreover, the study acknowledges the inherent challenge of potential biases within the dataset, which may arise from skewed distributions of disease prevalence, imaging techniques, or demographic characteristics. Such biases could inadvertently influence model training, leading to disparities in diagnostic performance across different patient groups [42].

2) FUTURE DIRECTIONS

Future research endeavors should aim to address these limitations while pushing the boundaries of multimodal recognition in medicine. Expanding the dataset to encompass a broader spectrum of diseases, patient demographics, and clinical scenarios would not only enhance the model's generalizability but also facilitate its application in diverse medical contexts.

Further architectural refinements are anticipated, with an emphasis on exploring novel AI methodologies such as graph neural networks (GNNs) for more effective data integration and Generative Adversarial Networks (GANs) for enriched

data augmentation [43]. The exploration of additional data modalities, including genomic, proteomic, and electronic health record (EHR) data, could unveil deeper insights into disease mechanisms and patient health, enriching the multimodal analysis framework.

Moreover, the translational potential of the proposed model extends beyond cardiac diseases to encompass a wide array of clinical conditions, from oncology to neurodegenerative diseases. Leveraging emerging AI techniques such as federated learning could further refine the model's diagnostic accuracy while ensuring data privacy and security in multi-institutional collaborations [44].

VI. CONCLUSION

In concluding our exploration into multimodal deep learning for early cardiac disease detection, we presented a model that integrates structured and image data through advanced deep learning techniques. This study demonstrated the potential of combining SE-Block and CBAM for attention-driven feature enhancement, Transformer Encoders for integrating structured data, and VAE embeddings for efficient feature fusion, resulting in a model that exhibits significant improvements in accuracy, precision, recall, and F1 scores compared to existing benchmarks.

Our findings indicate that attention mechanisms are crucial for emphasizing important features within images, Transformer Encoders effectively synthesize structured data for a comprehensive analysis, and VAE embeddings provide a robust method for data representation. However, the study acknowledges limitations, including the dataset's specificity and the need for broader validation, which points to future research directions.

Future work will aim to refine the model architecture, explore additional data modalities, and extend the applicability of the model to other clinical conditions. The potential of integrating emerging AI techniques to enhance the model's diagnostic accuracy remains an exciting avenue for research.

This study contributes to the field of medical diagnostics by providing a more accurate and efficient tool for early disease detection, which could significantly impact patient care. The advancement of multimodal deep learning models offers promising prospects for improving diagnostic processes and patient outcomes in healthcare.

REFERENCES

[1] W. Alshammari and F. Saleem. (2024). *A Machine Learning Framework for Early Detecting the Likelihood of Cardiovascular Disease in a Patient Using Multi-Attributes*. Platform Almanhal. [Online]. Available: https://platform.almanhal.com/Files/4/251461

[2] R. G. Schwartz, J.-P. Iskandar, and P. Soman, "Advances in clinical care with contemporary cardiac SPECT," *J. Med. Imag. Radiat. Sci.*, vol. 55, no. 2, pp. S64–S80, Jun. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1939865424000547

[3] H. M. Shepherd, J. Heaton, T. Marghitu, and L. Bai, "Applying deep learning in heart failure: Hospital readmission is not like other health quality metrics," *MedRxiv*, Mar. 2024, doi: 10.1101/2024.03.27.24304999.

[4] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114. [Online]. Available: http://proceedings.mlr.press/v97/tan19a.html

[5] M. McGilvray, J. Heaton, A. Guo, M. F. Masood, B. P. Cupps, M. Damiano, M. K. Pasque, and R. Foraker, "Electronic health record-based deep learning prediction of death or severe decompensation in heart failure patients," *JACC, Heart Failure*, vol. 10, no. 9, pp. 637–647, 2022.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, in Proceedings of Machine Learning Research, vol. 139, 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[7] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[9] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[10] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, 2014.

[11] L. Kumar, C. Anitha, V. N. Ghodke, N. Nithya, V. A. Drave, and F. Azmath, "Deep learning based healthcare method for effective heart disease prediction," *EAI Endorsed Trans. Pervasive Health Technol.*, vol. 9, pp. 1–6, Oct. 2023, doi: 10.4108/eetpht.9.4283.

[12] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. L. Martel, "Loss Odyssey in medical image segmentation," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102035, doi: 10.1016/j.media.2021.102035.

[13] N. Zhang, G. Yang, Z. Gao, C. Xu, Y. Zhang, R. Shi, J. Keegan, L. Xu, H. Zhang, Z. Fan, and D. Firmin, "Deep learning for diagnosis of chronic myocardial infarction on nonenhanced cardiac cine MRI," *Radiology*, Apr. 2019, doi: 10.1148/radiol.2019182304.

[14] T. Ghosh and N. Jayanthi, "An efficient dense-Resnet for multimodal image fusion using medical image," *Multimedia Tools Appl.*, vol. 80, no. 1, pp. 1357–137, 2021, doi: 10.1007/s11042-024-18974-7.

[15] A. U. R. Hashmi, D. Mahapatra, and M. Yaqub, "Envisioning MedCLIP: A deep dive into explainability for medical vision-language models," 2022, *arXiv:2203.18996*.

[16] S. Celi, N. Martini, L. E. Pastormerlo, V. Positano, and S. Berti, "Multimodality imaging for interventional cardiology," *Current Pharmaceutical Design*, vol. 23, no. 22, pp. 3285–3300, Sep. 2017, doi: 10.2174/1381612823666170704171702.

[17] T. Wolf, "Artificial intelligence in interventional cardiology," *EMJ Int. Cardiol.*, Jun. 2022, doi: 10.33590/emjintcardiol/22F0628.

[18] B. L. van der Hoeven, M. J. Schalij, and V. Delgado, "Multimodality imaging in interventional cardiology," *Nature Rev. Cardiology*, vol. 9, no. 6, pp. 333–346, Jun. 2012. [Online]. Available: https://www.nature.com/articles/nrcardio.2012.14

[19] J. Oza and G. Kambli, "Pixels to phrases: Evolution of vision language models," *AuthoreaPreprints*, Mar. 2024, doi: 10.36227/techrxiv.171078045.57266373/v2.

[20] L. Harold Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," 2019, *arXiv:1908.03557*.

[21] J. Cho, J. Lei, H. Tan, and M. Bansal, "Unifying vision-and-language tasks via text generation," 2021, *arXiv:2102.02779*.

[22] S. Lu, S. Wang, and Y. Zhang, "Detecting pathological brain via ResNet and randomized neural networks," *Heliyon*, vol. 6, no. 12, 2020, Art. no. e05625, doi: 10.1016/j.heliyon.2020.e05625.

[23] Z. Zheng, H. Zhang, X. Li, S. Liu, and Y. Teng, "ResNet-based model for cancer detection," in *Proc. IEEE Int. Conf. Consum. Electron. Comput. Eng. (ICCECE)*, Jan. 2021, pp. 325–328, doi: 10.1109/ICCECE51280.2021.9342346.

[24] S. C. Baybars, M. H. Duran, and S. A. Tuncer, "Detection of tongue anomalies using convolutional neural networks," *SSRN J.*, Mar. 2024. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4772603

[25] C.-W. Xie, S. Sun, X. Xiong, Y. Zheng, D. Zhao, and J. Zhou, "RA-CLIP: Retrieval augmented contrastive language-image pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19265–19274. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Xie_RA-CLIP_Retrieval_Augmented_Contrastive_Language-Image_Pre-Training_CVPR_2023_paper.html

[26] J. Zhou, L. Dong, Z. Gan, L. Wang, and F. Wei, "Non-contrastive learning meets language-image pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11028–11038. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Zhou_Non-Contrastive_Learning_Meets_Language-Image_Pre-Training_CVPR_2023_paper.html

[27] Y. Li, H. Wang, Y. Duan, H. Xu, and X. Li, "Exploring visual interpretability for contrastive language-image pre-training," 2022, *arXiv:2209.07046*.

[28] A. Radford. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. OpenAI. [Online]. Available: https://proceedings.mlr.press/v139/radford21a

[29] A. Jain, M. Guo, K. Srinivasan, T. Chen, S. Kudugunta, C. Jia, Y. Yang, and J. Baldridge, "MURAL: Multimodal, multitask retrieval across languages," 2021, *arXiv:2109.05125*.

[30] B. A. Abdelghani, S. Fadal, S. Bedoor, and S. Banitaan, "Prediction of heart attacks using data mining techniques," in *Proc. 21st IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2022, pp. 951–956, doi: 10.1109/ICMLA55696.2022.00159.

[31] A. Elhendy, D. W. Mahoney, R. B. McCully, J. B. Seward, K. N. Burger, and P. A. Pellikka, "Use of a scoring model combining clinical, exercise test, and echocardiographic data to predict mortality in patients with known or suspected coronary artery disease," *Amer. J. Cardiol.*, vol. 93, no. 10, pp. 1223–1228, May 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000291490400219X

[32] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8740989

[33] S. Bhave et al., "Deep learning to detect left ventricular structural abnormalities in chest X-rays," *Eur. Heart J.*, vol. 45, no. 22, pp. 2002–2012, Jun. 2024, doi: 10.1093/eurheartj/ehad782.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[35] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing*. Berlin, Germany: Springer, 2005, pp. 878–887. [Online]. Available: https://link.springer.com/chapter/10.1007/11538059_91

[36] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[37] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, Jan. 2019, doi: 10.1038/s41591-018-0316-z.

[38] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*.

[39] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, p. 2402, Dec. 2016, doi: 10.1001/jama.2016.17216.

[40] Y. Zhang, H. Wu, H. Liu, L. Tong, and M. D. Wang, "Mitigating the effect of dataset bias on training deep models for chest X-rays," in *Proc. Image Video Process.*, 2019. [Online]. Available: https://www.researchgate.net/publication/336577578_Mitigating_the_Effect_of_Dataset_Bias_on_Training_Deep_Models_for_Chest_X-rays

[41] K.-H. Thung, P. Yap, and D. Shen, "Multi-stage diagnosis of Alzheimer's disease with incomplete multimodal data via multi-task deep learning," in *Proc. 3rd Int. Workshop Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, vol. 10553, Quebec City, QC, Canada. Cham, Switzerland: Springer, 2017, pp. 160–168, doi: 10.1007/978-3-319-67558-9_19.

[42] I. Galić and M. Habijan, "Deep learning in medical image analysis for personalized medicine," in *Proc. Int. Symp. ELMAR*, Sep. 2023, pp. 207–212, doi: 10.1109/elmar59410.2023.10253934.

[43] L. Peng, N. Wang, N. Dvornek, X. Zhu, and X. Li, "FedNI: Federated graph learning with network inpainting for population-based disease prediction," *IEEE Trans. Med. Imag.*, vol. 42, no. 7, pp. 2032–2043, Jul. 2022, doi: 10.1109/TMI.2022.3188728.

[44] I. Pejic, R. Wang, and K. Liang, "Effect of homomorphic encryption on the performance of training federated learning generative adversarial networks," 2022, *arXiv:2207.00263*.

JUNXIN WANG (Member, IEEE) is currently pursuing the bachelor's degree with Beijing Normal University. He is a Scientific Research Assistant with the Statistics and Data Science Research Center, Institute for Advanced Study of Natural Sciences, Beijing Normal University. He is assisting the Educational Science and Technology Center of the Institute for Advanced Study of Humanities and Social Sciences in completing the National Social Science Fund's "13th Five-Year Plan" Education Youth Project. His research interests include image recognition, image processing, and remote sensing image analysis

JUANEN LI is currently pursuing the bachelor's degree with Beijing Normal University. He is preparing for direct doctoral admission through a recommendation program with Tsinghua University. He is a Ph.D. Student with Zhejiang University on projects involving automated vulnerability detection in smart contracts through fuzzing. Additionally, he is collaborating with a Ph.D. Student with Oxford University on research related to AI security. His research interests include fuzzing for smart contracts and adaptive upstream adjustments to pre-trained models.

RUI WANG is currently pursuing the bachelor's degree with Beijing Normal University, mainly studying data science and big data technology. He is the Scientific Research Assistant Management of the Graduate Center of Beijing Normal University to assist in the completion of the Natural State Fund Project. He is currently applying for the project of Harbin University of Technology about cloud computing. His research involves a distributed system architecture and edge/fog calculation.

XINQI ZHOU is currently pursuing the bachelor's degree with Beijing Normal University. He is a Scientific Research Assistant with Huitong College, Beijing Normal University, assisting his Tutor in completing the National Natural Science Foundation of China Youth Project. The author is preparing to apply for the data science program with Carnegie Mellon University. His main research interests include data processing, data analysis, and statistics.