Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.049186

ARTICLE

Posture Detection of Heart Disease Using Multi-Head Attention Vision Hybrid (MHAVH) Model

Hina Naz[1], Zuping Zhang[1,*], Mohammed Al-Habib[1], Fuad A. Awwad[2], Emad A. A. Ismail[2] and Zaid Ali Khan[3]

[1]School of Computer Science and Engineering, Central South University, Changsha, 410003, China

[2]Department of Quantitative analysis, College of Business Administration, King Saud University, P.O. Box 71115, Riyadh, 11587, Saudi Arabia

[3]Electrical and Computer Engineering, University of Victoria, Victoria, V9A1B8, Canada

*Corresponding Author: Zuping Zhang. Email: zpzhang@csu.edu.cn

Received: 29 December 2023 Accepted: 27 March 2024 Published: 15 May 2024

ABSTRACT

Cardiovascular disease is the leading cause of death globally. This disease causes loss of heart muscles and is also responsible for the death of heart cells, sometimes damaging their functionality. A person's life may depend on receiving timely assistance as soon as possible. Thus, minimizing the death ratio can be achieved by early detection of heart attack (HA) symptoms. In the United States alone, an estimated 610,000 people die from heart attacks each year, accounting for one in every four fatalities. However, by identifying and reporting heart attack symptoms early on, it is possible to reduce damage and save many lives significantly. Our objective is to devise an algorithm aimed at helping individuals, particularly elderly individuals living independently, to safeguard their lives. To address these challenges, we employ deep learning techniques. We have utilized a vision transformer (ViT) to address this problem. However, it has a significant overhead cost due to its memory consumption and computational complexity because of scaling dot-product attention. Also, since transformer performance typically relies on large-scale or adequate data, adapting ViT for smaller datasets is more challenging. In response, we propose a three-in-one steam model, the Multi-Head Attention Vision Hybrid (MHAVH). This model integrates a real-time posture recognition framework to identify chest pain postures indicative of heart attacks using transfer learning techniques, such as ResNet-50 and VGG-16, renowned for their robust feature extraction capabilities. By incorporating multiple heads into the vision transformer to generate additional metrics and enhance heart-detection capabilities, we leverage a 2019 posture-based dataset comprising RGB images, a novel creation by the author that marks the first dataset tailored for posture-based heart attack detection. Given the limited online data availability, we segmented this dataset into gender categories (male and female) and conducted testing on both segmented and original datasets. The training accuracy of our model reached an impressive 99.77%. Upon testing, the accuracy for male and female datasets was recorded at 92.87% and 75.47%, respectively. The combined dataset accuracy is 93.96%, showcasing a commendable performance overall. Our proposed approach demonstrates versatility in accommodating small and large datasets, offering promising prospects for real-world applications.

KEYWORDS

Image analysis; posture of heart attack (PHA) detection; hybrid features; VGG-16; ResNet-50; vision transformer; advance multi-head attention layer

1 Introduction

Each year, millions of individuals pass away from heart problems, with factors such as poor nutrition, unhealthy lifestyle, and pollution contributing considerably to its prevalence [1]. In multiple countries, like Italy and Japan, more than 20% of the population is over the age of 65, and the proportion (%) of heart attacks (HAs) is rising daily. It is one of the deadliest disease, accounting for 16% of global fatalities. In 2000, the death rate rose quickly, and by 2019, the World Health Organization (WHO) indicated that the death toll from cardiovascular disease was between 2 to 8.9 million. People who live alone find no one nearby to help them in an emergency in case of experiencing heart attack symptoms [2,3]. This vulnerability exposes them to the risk of making an emergency call for assistance. Many are unaware of the signs and symptoms of a myocardial infarction. Lack of awareness regarding these warning signs and symptoms could elevate the risk of death. Patient require quality care when experiencing such symptoms because they may not be able to detect them early. However, not receiving medical assistance care in hospitals could jeopardize a patient's life. The primary symptoms of congestive heart failure include excruciating chest pain, upper body discomfort, exhaustion, and shortness of breath [4] as shown in Fig. 1. However, for the identification of cardiovascular diseases, [5] a person does not need multiple or complex devices to detect congestive heart failure. According to a recent study, a set of images produced by individuals suffering from heart problems was utilized. A thorough review of deep learning (DL) applications in a variety of disciplines has been provided by recent research. DL was employed in the previous studies to train the algorithm, enabling it to examine the dataset and determine whether or not the person was suffering from heart muscle injury [6,7]. Image segmentation is crucial across domains such as medical image processing, object recognition techniques, clustering, and surveillance [8]. Advanced techniques, including stochastic approaches and DL integration, improve segmentation performance. However, these models might perform well for certain image classes but encounter challenges with datasets due to image diversity. Critical factors affecting segmentation performance include noise and intensity in-homogeneity within the image [9].

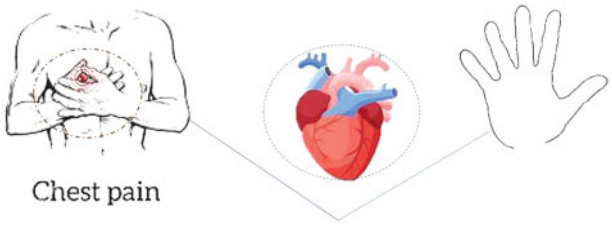

Figure 1: This image shows the relation between heart attack and chest pressure

Our research, we propose a method that utilizes deep learning (DL) techniques to detect ongoing heart attacks. The dataset employed in this technique consists of individuals exhibiting postures associated with heart attacks, which is then divided into two categories. This dataset is used for training with DL to determine whether a person in any posture is experiencing cardiovascular disease. Our study suggests that employing a hybrid model to predict cardiac delays results in a powerful classification ability.

Our contributions are as follows: Dataset Splitting: Initially, we divided the dataset into two subsets based on gender, male and female derived from the original combined dataset. Introduction of MHAVH Model: We introduced a novel Multi-Head Attention Vision Transformer Hybrid (MHAVH) model, designed to enhance classification performance. This model is trained on all three

datasets male, female, and combined to capitalize on the diverse data sources available. The drawbacks of the previous study are outlined below:

a) The initial exploration into posture-based heart attack detection was documented in 2019 [3]. The subsequent 2021 research delved further into posture detection [10]. However, it is noteworthy that the majority of prior studies did not focus on posture-based cardiac arrest detection until this recent technological advancement, as evidenced by the latest paper published in 2022 [6].
b) Researchers aim to design specialized hardware for the challenging task of posture-free diagnosis of coronary infarction. This hardware should seamlessly integrate with existing models, as there has been limited exploration in posture-based heart disease diagnosis. Previous studies have primarily focused on conventional diagnostic tools such as Electrocardiograph (ECG) machines and magnetic resonance imaging (MRI) [11,12]. The envisioned hardware would cater to the unique demands of posture-free diagnosis, enhancing the accuracy and efficiency of detecting coronary infarction.
c) However, there is not too much HA data with posture available on the internet, and ViT-based techniques might not be able to extract features as well. Furthermore, redundant associations between many local patches may be produced by ViT's self-attention mechanism, which would negatively impact the performance [13].
d) The presence of imbalanced data between two distinct classes poses challenges during neural network training, resulting in inaccurate detection and potentially misleading analysis. It is essential to mitigate this imbalance to ensure a more reliable and unbiased model learning process [14].

This paper focuses on enhancing the MHAVH deep learning algorithm for vital sign detection associated with heart attacks, emphasizing object detection's role in assessing chest pain severity. The modified MHAVH technique tailored for posture-based chest pain detection shows promising results, particularly in hospital settings, aided by data augmentation techniques. Leveraging Residual Neural Network (ResNet-50), Visual Geometry Group (VGG-16), and vision transformer (ViT), this hybrid model combines multiple approaches to improve overall performance, with data augmentation playing a crucial role in enhancing performance on image datasets.

Initially, we implemented these models individually, starting with VGG-16, then ResNet-50, and the ViT. However, we encountered overfitting issues. To address this issue, we experimented using multiple combinations (hybrid) of different models before ultimately implementing the MHAVH model. One of the prominent hybrid approaches we used was based on a combination of ViT, MobileNetV2, and VGG-16. Although it resolved overfitting issues. However, we did not get satisfactory results in accuracy and performance. In further experiments, we adjusted the parameters of each model, such as the number of heads and selected mid-head attention. We also determined that changing the number of heads directly impacted the image patch size in the ViT model. Additionally, we optimized parameters like dropout and replaced the Gaussian error linear unit (GeLU) with rectified linear unit (ReLU) activation functions, results into improved performance. Despite these adjustments, we found the combination (hybrid) model's results were insufficient. Then, we further explored data augmentation techniques to improve performance, which significantly impacted our experiments while comparing results with each approach. That is why we proposed and used the MHAVH model in our experiments, as it showed satisfactory results, thereby concluding our research efforts. Our main goals related to the proposed study are as follows:

a) Using machine learning (ML), we proposed a method for detecting congenital heart disease in red, green, blue (RGB) color images.
b) We designed a novel three-in-one stream fuse model called MHAVH by collecting posture and feature extraction data for every class of images, followed by data augmentation techniques at the pre-processing stage.
c) Two pre-trained transfer learning models (VGG-16 and ResNet-50) combined with ViT are employed to enhance the model's performance and resolve unbalanced data issues.
d) We performed Deep Neural Network (DNN) classification and conducted many experiments to improve model performance metrics using ablation studies, quantitative analysis, and qualitative methods to represent the proposed model's effectiveness.
e) To validate our model in the detection of HA.

Hence, to achieve the primary goal with novelty, we made some developments in this study, the main contributions are as follows:

a) Contribution 1: We proposed a model named MHAVH, this is a Three-in-one stream model. In this proposed model, we introduced a fusion method that combine three different models simultaneously: Vision transformer, VGG-16, and ResNet-50. The novelty of this research is combining three separate models. The milestone study, known as ViT, the vision transformer (ViT), represents a significant advancement in the image classification, leveraging its self-attention mechanisms [4]. VGG-16 is an open-source framework mostly employed in a range of multiple research scenarios [15]. It is based on a sequence of six steps. There are two convolutional layers (CLs) in the first two stages. The first is max-pooling (MP), which has a stride of two (2). The other is three (3) CLs and one (1) MP layer of stride two (2) are present for the next three steps. ResNet-50 needs two stages to connect to the feature extraction network [16]. Average pooling should come first; the fully connected layer should come after average pooling.
b) Contribution 2: We designed advanced multi-head attention, which leads to a novel and efficient end-to-end deep method. We increased the number of heads in advance of multi-head attention. We found out that the intermediate size number of heads is neither too small nor too big for better results.
c) Contribution 3: We split datasets into two categories: Male and female. We utilized both categorical datasets along with a combined (original) dataset.
d) Contribution 4: We employed pre-trained convolutional neural network (CNN) models such as VGG-16, ResNet-50, and ViT in combination with deep neural networks (DNN). This hybrid model approach was utilized to enhance detection capability.

The paper is organized into several sections as follows: Section 2 provides a detailed summary of prior research conducted in this field; Section 3 presents a comprehensive methodology for the proposed framework; Section 4 discusses the results achieved; and Section 5 concludes the paper and outlines avenues for future research.

2 Background and Related Work

Image processing is a vital technique, which is being utilized and implemented across various domains of biology. It helps to detect and analyze problems related to object detection and multiple disease detection, such as heart attacks. This paper considers RGB images of posture-based HA detection, as many researchers have worked on heart disease detection using machine and deep

learning techniques. In this section, we categorized our discussion into four different sections: (1) Machine Learning; (2) Deep Learning and Heart Attack Survey; (3) Color Images; (4) Transformers, and Multi-Head Attention.

2.1 Machine Learning

A noninvasive-based technique is an approach in an ML-based system for predicting HA disease by utilizing a dataset of heart posture. In this method, they used seven different ML techniques: A process of cross-validation, algorithms of feature selection, and classifiers that perform well on evaluation criteria such as accuracy, sensitivity, precision, Matthews' correlation ratio and calculated the time of execution [10]. Li et al. employed logistic regression, support vector machine, and extreme gradient boosting to predict heart-related conditions in patients. Their study, conducted between July 2016 and December 2019 on 3759 expectant patients, yielded a prediction model with an accuracy of 0.920, an F1-score of 0.571, and a recall of 0.789 [17]. Another approach is based on statistical classification using ML methods after the handmade segmentation. However, this approach is labor-intensive and computationally expensive, this approach requires a large number of experienced workers, and is error-prone [18].

2.2 Deep Learning and Heart Attack Survey

The leading contributors to cardiovascular disease and the leading cause of high death rates globally are valvular heart disorders (VHDs) [19]. Jamil et al. introduced three distinct frameworks leveraging deep learning to accurately detect valvular heart diseases (VHD) using both One-Dimension (1D) and Two-Dimension (2D) phonocardiogram (PCG) signals. Their proposed method involves an efficient multi-scale convolutional neural network (CNN) for precise VHD identification, aiming to improve early diagnosis and treatment outcomes [19]. This technique has been used to extract the distinguishable features of multiple non-overlapping frequency bands of electroencephalogram signals from several scales for the motor imagery brain-computer Interface (MI-BCI) classifier [20]. In another research, a CNNs and support vector machine (SVM) based hybrid model was used in which CNN acts as a feature extractor and SVM as a binary Identifier. The technique for this method, which is based on handwritten digits, was trained and tested using the Modified National Institute of Standards and Technology (MNIST) dataset [21]. The author proposed a model based on deep convolutional neural network (DCNN) architecture codenamed inception that attains the new state of the art for the classification of images using the ImageNet dataset [22].

2.3 Transformer and Multi-Head Attention

A fusion method [23] proposed an architecture of transformers to combine multi-modal features effectively. It developed a fused multi-modal presentation for multi-label video emotion detection from the video dataset. Wang et al. developed a network based on the transformer's encoder-decoder structure. The 3D CNN [24,25] is employed to extract or retrieve the spatial feature map from the global feature modeling carefully transformed by the input transformer. Simultaneously, the decoder predicts the precise segmentation map by performing progressive up-sampling and utilizing the features or attributes embedded in the transformer. The popularity of transformers [26] is attributed to their self-attention mechanism, which calculates a weighted average of features from other tokens based on similarity scores between pairs of tokens. This allows each token to obtain a contextual representation by attending to others in the sequence [27]. Multi-head attention [28] displays a specific orientation of attention heads and proposes multiple phenomena based on subspace and attended positions. Each attention head displays characteristics that are unique from others. Multiple heads [29]

are concatenated to compute the final output, and multi-head expands single-head attention to capture various attention patterns and increase transformer performance. The study is organized into different phases: Split data, data-augmentation phase, model preparation process, testing, and results.

3 Methodology

Here is a detailed review of the presented MHAVH model working principle for detecting HA detection into two classes from a posture-based database. The method begins by downloading an online dataset [3], which is then divided into two classes for binary classification. To ensure optimal results, various pre-processing techniques are applied to the image dataset. Moreover, a three-in-one stream framework is developed, incorporating ResNet-50, VGG-16, and ViT architecture. Notably, enhancements and improvement are made to the Multi-Head Self-Attention layer. Fig. 2 displays the step-by-step process of the developed model.

Figure 2: The proposed system describes the processes of HA classification

3.1 Multi-Head Attention Hybrid (MHAVH) Model

In this section, we demonstrated the MHAVH proposed framework responsible for binary classification for HA detection in our research. We further discussed VGG-16, ResNet-50, and vision transformer in detail and how we associated deep transfer learning models and the vision transformer model. Furthermore, a deep neural network (DNN) were utilized to perform classification.

3.1.1 Framework

This architecture of MHAVH is a combination of three-in-one feature extraction models. As shown in Fig. 3, the two pre-trained transfer learning models, VGG-16 and ResNet-50, work alongside a self-attention-based model (ViT). This three-in-one-stream network model configura-tion provides robustness for the extraction of hybrid features. In other sub-sections, we discussed each approach used in the MHAVH model in detail.

3.1.2 ResNet-50

It is comparable to VGG-16, except that ResNet-50 has additional identity mapping capabilities. ResNet-50 predicts the delta that is required to reach the final prediction from one layer to the next [30]. Our model used ResNet-50 because of its specific capabilities to deal with the vanishing gradient problem. Fig. 4 shows the structure of RestNet-50, which is based on its 50 layers within five blocks (Conv1, Conv2, . . . , Conv5). For each of these blocks, the residual function F has three convolution layers with the following dimensions: (1 × 1), (3 × 3), and (1 × 1). Eqs. (1)–(3) represent how to calculate the output Z and add the input X of this block with the residual function (F), where the weight matrix (Wi) of three successive convolution layers is updated with X by F. During the feature extraction, the image (I), with the image shape of (256 × 256 × 3), is fed into the ResNet-50 model, and the output $\text{ResNet}_{\text{conv5\_block3\_out\_layer}}$ is utilized for the final output. The extracted feature,

ResNet $_{extracted\_feature}$ is converted to produce a 1D vector output of NResNet with the use of flattened layer.

$$Z = (X) + F(X, Wi) \tag{1}$$

$$ResNet_{extracted\_feature} = Resnet_{conv5\_block3\_out\_layer} \tag{2}$$

$$N_{Resnet} = flatten\left(Resnet_{extracted\_feature}\right) \tag{3}$$

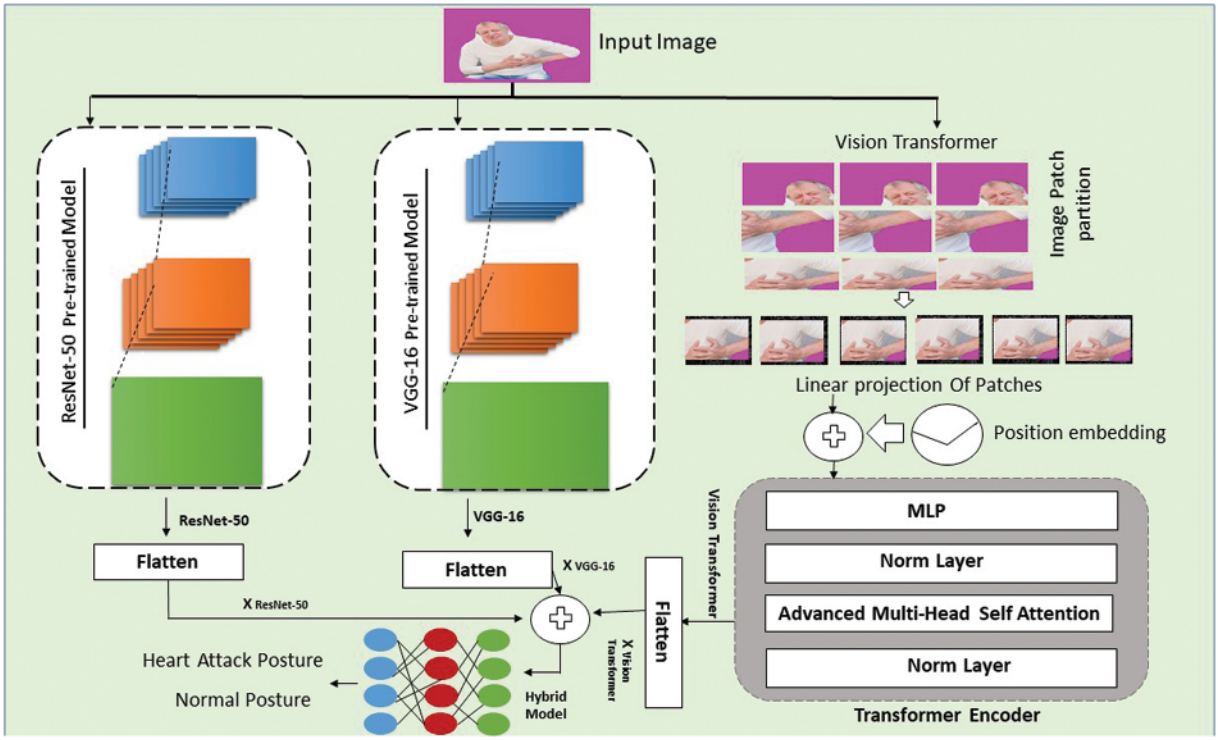

Figure 3: Flow of proposed hybrid MHAVH framework

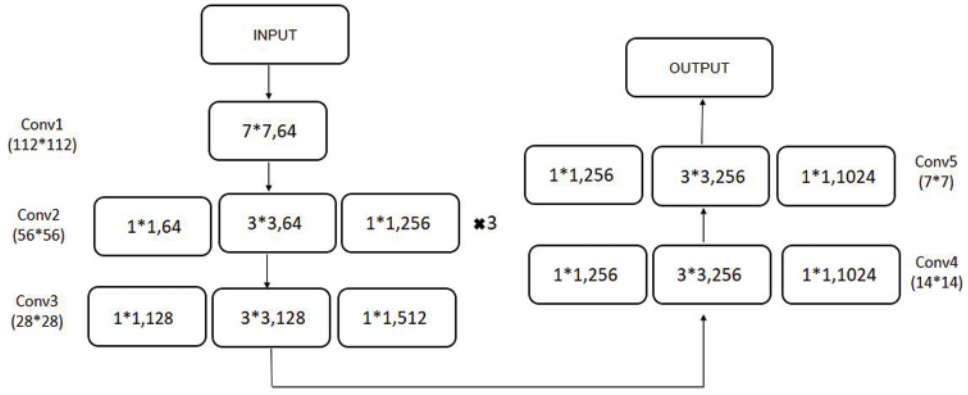

Figure 4: Basic data flow diagram of ResNet-50

3.1.3 VGG-16

The VGG-16 network was trained on the ImageNet database [30]. Its extensive training method provides excellent accuracy even with small image datasets. VGG-16 contains 16 layers, which is responsible for better results compared to the AlexNet architecture, which substitutes the large kernel size with a 3 × 3 filter set. Eq. (4) shows a formulated representation of $VGG_{extracted\_feature}$, which is an outcome of $VGG_{mixed5\_dense\_layer}$. Eq. (5) explains how to obtain N_{VGG} as an output of converting from $VGG_{extracted\_feature}$ after applying flattening technique.

$$VGG_{extracted\_feature} = VGG_{mixed5\_denselayer} \tag{4}$$

$$N_{VGG} = flatten\left(VGG_{extracted\_feature}\right) \tag{5}$$

3.1.4 Vision Transformer

The transformer is one of the popular method of deep learning commonly used to order image patches. Vision transformers (ViT) have shown better performance in the field of image classification, particularly when pre-trained on large dataset and applied to various image reorganization benchmarks such as CIFAR-100, VTA and ImageNet. They require considerably less computational resources during training compared to other methods. ViT is a neural network type with a simple design capable of processing various modalities such as images, text, and speech using processing blocks that are similar. It demonstrates brilliant capabilities when applied to very large capacity networks and extremely large datasets. ResNet can be employed to identify and categorize ViT with a greater level of accuracy. The original transformer model used in natural language processing (NLP), and is served as a basis for ViT. While the input of ViT is a 1D series of tokens. However, pictures are 2D, and ViT models distribute images into little parts of 2D patches and images converted into the patches as tokens, as executed by the first transformer model natural language processing (NLP). The height (H), width (W), and the number of channel (C) is separated into smaller 2D patches to organize the input image similarity to NLP. This creates patches (N = HW/P∗2) with a size of pixel P × P [8]. Before forwarding the image patches to the encoder, flattening, sequence, learning, and patch embedding are performed in the following sequence: Each patch of input data is flattened and then converted into a vector. The embedded input image patches are created by projecting the flattened patches into dimensions (D) using linear projection. A learnable embedding X-class is prefixed to the order of the embedded input patches. The outcome N is consistent with the X-class values. Finally, 1D positional embedding (Epos) are learned during the training process and are inserted into the patch embedding's to prefix positional information to the image. The embedding vectors created as the outcome of these operations are donated by (φo). The transformer encoder, which is a stack of identical layers (L), is fed (xφ) to perform the classification task. The value of the X-class at the Lth layer of the encoder output is then sent to the classification head. Multilayer perceptrons (MLPs) are used to create the classification head during pre-training, and a single linear layer is utilized during fine tuning. As the classification head, the MLP carries out the ReLU non-linearity shows in Eq. (6). Overall, the encoder components of the first NLP transformer architectures were utilized by the vision transformer. A sequence of 32 × 32 embedded input patches, positional data, and a learnable class embedding is suspended in the sequence are fed into the encoder. To smaller size of the patch effect the model, the performance will be improved but the computational cost will be higher. The choice of a 32 × 32 patch size was selected due to its robustness against the loss of performance and computational complexity. A classification head, connected to the encoder's output shows in Eq. (7), collects the learnable class-embedding values and utilizes it to generate a classification output based on its state.

$$Transformer_{extractedfeature} = ReLU(MHAVH) \tag{6}$$

$$N_{Transformer} = flaten\left(Transformer_{extracted\_feature}\right) \tag{7}$$

In Table 1, we have presented the unique parameters used to get desired output from vision transformer. These parameters were pruned based on various experiments with the flow we discussed above in detail.

Table 1: Selected optimum parameters of the vision transformer

Parameter	Value
Image size	256
Patch size	32
Number of heads	16
Number of patches	64
Projection dimension	64
Transformer unit	(128, 64)
Dropout rate	0.1

3.1.5 Advance Multi-Head Self-Attention Layer

To compare with the results reported by previous work [31], we experimented our model and successfully improved its performance. We introduced an advanced multi-head attention layer, which employed a ReLU activation function instead of GeLU, as in the original ViT, to randomly modify the Multi-Branch (MB) architecture sub-paths in DL mode. We increased the number of heads in advanced multi-head attention to address the issue of patch size generation, which significantly impacts the models performance and enhances its capability. We observed a trend against mid-head attention, where the number of heads increased or decreased, it directly impacts the dimension of the key (dk) required to get its value. Given that the dk is dk = dv = dimension of key/number of heads. We increased the number of heads for our model based on its specific requirements. We decided that our number of heads should be intermediate-head sized, neither too small nor too big. Selection of a smaller size is directly impacts dk and the training and testing accuracies are affected. If a larger size is selected, dk can skip out the important information. Consequently, we settled on several heads with a size of 16, which provided the best outcome.

$$A_j = Softmax\left(\frac{Q_i K_t^i}{\sqrt{d_k}}\right) \tag{8}$$

3.1.6 DCNN Classifier

In this stage, enhancing the model's capability and gaining a more intricate understanding becomes possible. The final classifier is a DNN composed of a concatenation of VGG-16, vision transformer, and ResNet-50. The DNN model consists of 5 layers: 4 dense layers with GeLU activations of sizes 128, 64, and 32 units, respectively, along with a batch normalization layer. The last dense layer has 1 unit with a sigmoid activation function for binary classification. Fig. 5 illustrates the operational structure of the deep neural network classifier.

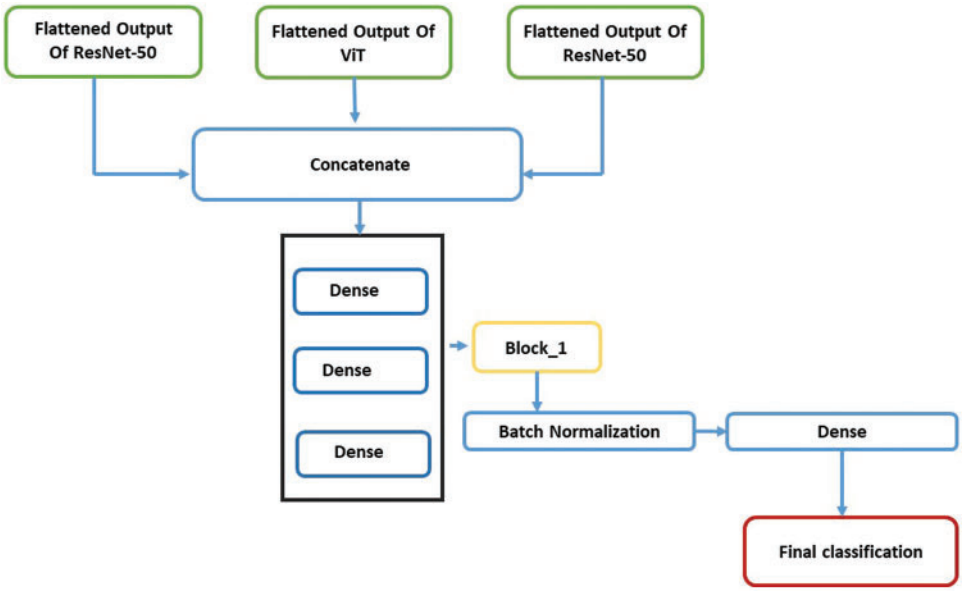

Figure 5: The structural configuration of a deep neural network (DNN) classifier anticipates the class based on the extracted features

In Algorithm 1, the brief explanation of DNN outlines the operational structure of the Multi-Head Attention-Vision Hybrid (MHAVH) model. The first two lines define the input and output of the model, specifying the creation of a function with the parameter 'i'. Subsequently, the algorithm details the augmentation process applied to both the training and testing datasets. We extracted features from each of the given steps 3, 4, and 5 (ResNet-50, VGG-16, and vision transformer). After extracting the features from each model, we used a DNN classifier to classify them against those extracted features. In the initial step of the DNN, concatenation is performed to merge the outputs (extracted features) of the three models in steps 3, 4, and 5 into a single entity named F_{hybrid} in step 6. After this concatenation, we used F_{hybrid} as an input to the dense layers to generate x1, as elucidated in step 7 while classifying with DNN. We used Block1, which contains a combination of a normalization layer and a dense layer, shown in step 8, and applied it to x1, demonstrated in step 9. Finally, the dense layer used softmax to the output x2 for the final classification.

Algorithm 1: Procedure of MHAVH Hybrid Model

Input	Train dataset (d_{test}) Test dataset (d_{train}) image size (i_{size}) epoch (e), class mode (c_{mode}) batch size (b_{size})
Output	classified labeled images (Normal and Heart attack posture) L, Performance metrics M(if ground truth is available)
Step 1.	Function classification(i)
Step 2.	$i \leftarrow img\_datagenerator(t_{dataset}, v_{dataset})$ // rescaling and augmentation
Step 3.	$R_{feature} \leftarrow (i, N_{resnet50})$ // Eq. (3)
Step 4.	$V_{feature} \leftarrow (i, N_{vgg})$ // Eq. (5)
Step 5.	$T_{feature} \leftarrow (i, N_{Transformer})$ // Eq. (7)
Step 6.	$F_{hybrid} \leftarrow concatenate(R_{feature}, V_{feature}, T_{feature})$
Step 7.	$x1 \leftarrow dense(F_{hybrid})$ // DNN operation

(Continued)

Algorithm 1 (continued)

Step 8. Block 1 ← J(batch normalization, dense) // DNN operation
Step 9. x2 ← Block 1(x1) // DNN operation
Step 10. prediction ← softmax(x2)
Step 11. n ← 1
Step 12. **while n ≤ do**
hybrid fit(d_{test}, d_{train}, e, b_{size}, c_{mode}, i_{size}) // Train the MHAVH model
end while
Step 13. accuracy, precision, F1-score // model MHAVH Evaluate matric (d_{test})
Step 14. return accuracy, precision, F1-score

The model's parameters refer to the trainable weights and biases within the layers. In the Dense layers, each unit has weights connected to all units of the previous layer (plus a bias term). The BatchNormalization layer contains parameters for scaling and centering the activations as well as moving averages for these parameters.

3.2 Dataset Description

All the heart attack images used in this study were taken from the Heart-Attack-Detection-In-Images 2019 dataset. We split the complete dataset into three subsets and archived them in three zip files, male and female, and a combination of both. The female dataset is the first dataset with two different classes, the same as the others. This set of images has 10,256 number of images. The male dataset is the second dataset we used. It has 16,983 images and is relevant to males only. The combined dataset is the third dataset used for training and testing, which is available to download from this link: https://github.com/Turing-IA-IHC/Heart-Attack-Detection-In-Images2019. It comprises both males and females and has 31,920 images. All three datasets have images with postures (heart attack) and without postures (without heart attack), 80% of images from both classes in the dataset are used for training and 20% for testing. Our images have a 256 × 256 pixels' resolution, totaling 65,536 RGB images.

Fig. 6 shows the postures in the dataset with examples of both males and females. Some exhibit symptoms, while others do not. We utilized two classes to perform binary classification using our model.

Figure 6: Examples of dataset image samples

3.3 Implementation Detail

As a result, the suggested MHAVH models virtually all have their crucial hyperparameter automatically set utilizing the stochastic gradient descent (SGD) approach. Another architectural hyperparameter is the number of pooling layers, fully linked layers, filters, size of the filters, and activation function, for example, ReLU. In addition, we have performed fine hyperparameter tuning on multiple parameters like momentum, L1 or L2 regularization, batch size, and learning rate. We are employing the binary cross-entropy function as the objective function. This loss function is commonly used for binary classification tasks. It used a sigmoid activation function on the model's output layer. Binary cross-entropy compares the predicted probabilities and the actual binary labels to compute the loss, guiding the model to reduce the error in its predictions during training. We implemented this in our experiments because it works well when the model's output is a probability value within the range [0, 1], which aligns with the binary nature of the classification task. After MHAVH is trained over three epochs due to high graphics processing unit (GPU) constraints, the test set is utilized to estimate model's simulated time. Our model takes about 30 min to train, meaning each iteration takes approximately 10 min or more to complete. For the DNN, we utilized four dense layers with 3660 parameters and one batch normalization layer comprising 120 parameters.

Table 2 outlines the iteration parameters in each experiment conducted on the author's personal computer with Python 3.7. The experiments were trained and tested using Google Collaboratory GPU with TensorFlow. The expanding usage of CNNs in medical image analysis has raised concerns regarding increasing network depth, image quality, and computational costs. Optimizing hyperparameters can mitigate computational expenses and enhance the efficacy of findings.

Table 2: Selected hyper-parameters setting for the proposed MHAVH model

Hyperparameter	Values
Loss-function (L)	Binary cross-entropy
Epoch (e)	3
Utilized-optimizer (O)	SGD
Batch-size (b)	32
Learning-rate (lr)	0.001

3.4 Performance Matrix for Classification

Three CNNs were trained and evaluated based on accuracy, sensitivity, specificity, and F1-score. Results were summarized in Table 3, and a confusion matrix depicted findings from Class A and Class B image detection tests. Eqs. (9)–(11) were utilized to calculate classification accuracy, specificity, and sensitivity using true positive ($T_{positive}$), true negative ($T_{Negative}$), false negative ($F_{Negative}$), and false positive ($F_{positive}$) values.

$$Precision = \frac{T_{positve}}{F_{positive} + T_{positive}} \tag{9}$$

$$Accuracy = \frac{T_{positive} + T_{Negative}}{F_{positive} + T_{positive} + T_{Negaitive} + F_{positive}} \tag{10}$$

$$F1 - Score = \frac{2 * specificity * sensitivity}{specificity + sensitivity} \tag{11}$$

Table 3: Accuracy, precision and F1-score of implemented model

Datasets	Predicted class	Precision	F1-score	Accuracy
Male	PHA and PNH	87.39%	92.72%	92.87%
Female	PHA and PNH	81.62%	75.98%	75.47%
Combine	PHA and PNH	89.62%	94%	93.96%

4 Results and Analysis

The test set is utilized to evaluate the model after being trained for three epochs due to GPU constraints, with each iteration taking approximately 10 min. This implies that the model undergoes roughly half an hour of training. In Fig. 7a, the curve represents the proposed model on the combined dataset, while Fig. 7b displays the curve for the proposed MHAVH framework on the female dataset, and Fig. 7c shows the male dataset. The testing set is employed to evaluate the model's performance. In Fig. 7c, the training accuracy ranges between 89.59% to 99.40%, and the validation accuracy varies between 71.74% to 93.69% for the combined dataset. In Fig. 7b, the training accuracy fluctuates from 80.14% to 98.97%, and the testing accuracy ranges from 45.40% to 75.47%. In Fig. 7c, representing the male dataset, the training accuracy varies from 91.39% to 99.33%, and the testing accuracy for the male dataset lies in the range between 58.09% to 92.88%. However, Fig. 7 indicates that the model achieves a higher level of accuracy when trained on the combined dataset, respectively.

Fig. 8 depicts the training and testing losses of the MHAVH model, demonstrating a decrease in both training and testing losses. In Fig. 8a, the training accuracy quickly reaches 1.0, while the testing accuracy plateaus below that level. This indicates potential overfitting, as the model performs accurately on the training data but not on the testing data, the loss graph supports this observation, showing a very low training loss that continues to decrease, while the testing loss starts to plateau. The training loss decreases from 0.2208 to 0.0186, and the testing loss decreases from 0.8310 to 0.2635. Similarly, in Fig. 8b, the training accuracy again reaches to 1.0, whereas the testing accuracy is significantly lower and seems to plateau. This is a classic sign of overfitting, where the model has learned the training data too well, including its noise and outliers, which does not generalize well to the testing data. The loss for the female dataset shows a training loss nearing zero, while the testing loss decreases at a much slower rate. The training loss varies from 0.3931 to 0.0337, and the testing loss varies from 1.1812 to 0.6161. In Fig. 8c, similar to the other two, the training accuracy reaches to 1.0, with the testing accuracy being lower. The loss graph shown a clear gap between the training and the testing loss, suggesting overfitting, the training loss varies from 0.2016 to 0.0207, and the testing loss reaches from 0.7349 to 0.3087. The primary objective is to minimize both training and testing losses, ensuring the model delivers accurate predictions.

Our evaluation of the model's performance on separate datasets for males, females, and combined groups reveals a tendency for overfitting. As shown in Figs. 7 and 8, the model achieves a training accuracy of 1.0 across all the datasets, indicating a perfect fit of the training data. However, the testing accuracy is notably lower in each case, suggesting that the model does not generalize as effectively to unseen data. The corresponding loss graphs further support this observation, where the training loss approaches zero, but the testing loss plateaus or decreases at a slower rate.

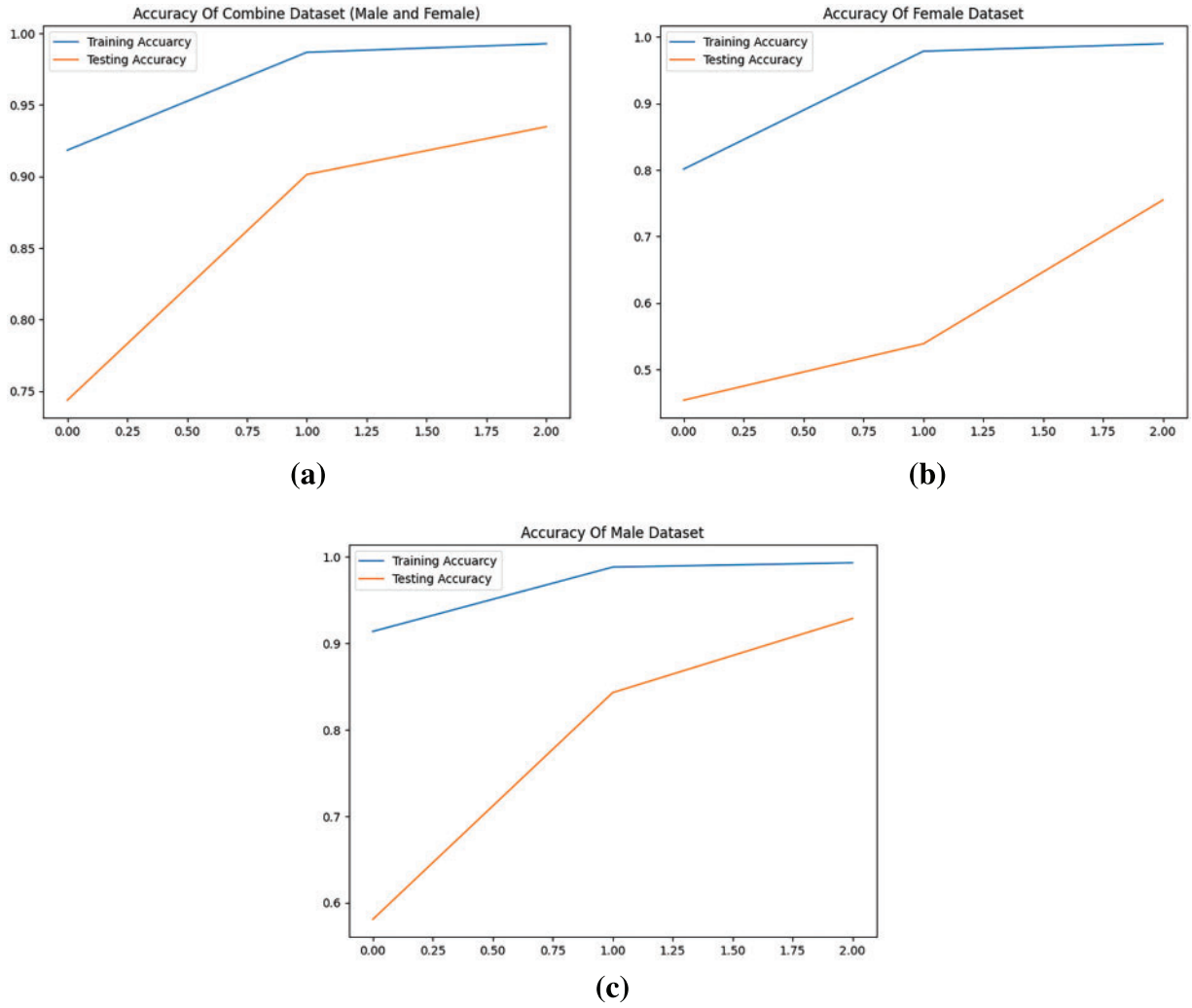

Figure 7: The accuracy curves illustrate the dynamic changes in both training and testing sets with respect to each epoch. The x-axis shows the number of epochs, and the y-axis shows the accuracy of the training and testing of the model. The curves are presented separately for (a) Combined data, (b) Female data, and (c) Male data

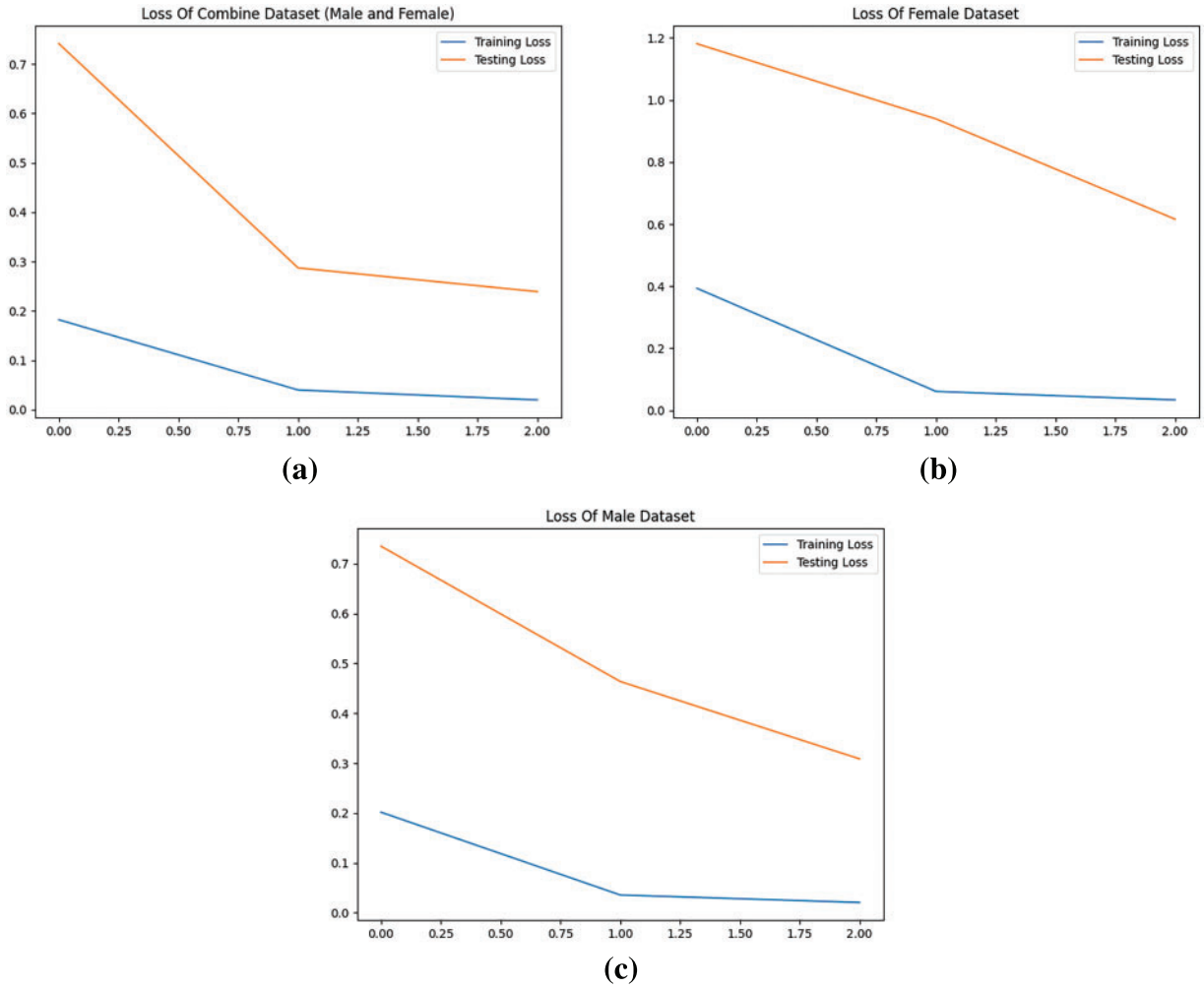

Figure 8: The loss curve follows a descending trajectory

4.1 Quantitative Analysis

The model's performance was evaluated using quantitative metrics such as accuracy, precision, and F1-score on the test set, as detailed in Table 3. For the combined dataset, the model achieved a notable accuracy of 93.69%, with an F1-score of 94% and precision of 89.62%. The male dataset demonstrated an accuracy of 87.39% and an F1-score of 92.72%, while the smaller female dataset exhibited lower performance metrics: Approximately 75.47% accuracy, 75.98% F1-score, and 81.62% precision. Overall, appropriate dataset combination led to higher accuracy, highlighting the effectiveness of the model in classifying heart disease while considering postures of normal human (PNH) and posture of heart attack (PHA).

4.2 Ablation Studies

We proposed the MHAVH framework to justify the effectiveness of the multi-head self-attention component. Multiple experiments are performed on many other networks. We have compared the ResNet-50, VGG-16, ViT, and compact model (MobileNetV2, VGG-16, vision transformer) results

with our implemented model in the experiments, the dataset tested predicted results of each method are shown in Fig. 9. It is evident that the MHAVH method exhibits the highest accuracy when using all three datasets compared to the other models. These results are further detailed in Table 4. To demonstrate the efficiency of our proposed MHAVH model, we compared its performance with and without a vision transformer. Initially, we independently evaluated the accuracy of each transfer learning model (ResNet-50, VGG-16). The lowest testing accuracy that the ResNet-50 model accomplished with the male, female, and combined datasets is (45.22%, 53.84%, and 49.95%). This indicates that VGG-16 outperforms ResNet-50, but we could not obtain high accuracy alone. After that, it was tested utilizing the ViT and the accuracy was found improve over the transfer learning model. The combined male and female datasets all produced higher accuracies than the previous results, with 70.28%, 56.93%, and 73.04%. Then, we fused ViT with MobileNetV2 and VGG-16 and compared the results to the transfer learning model and vision transformer, getting an approximate accuracy of 87.48% with the male dataset, 84.29% with the female dataset, and 90.60% with the combined dataset, but it was still insufficient. Lastly, we conducted tests using our MHAVH model to outperform other models' accuracy and performance.

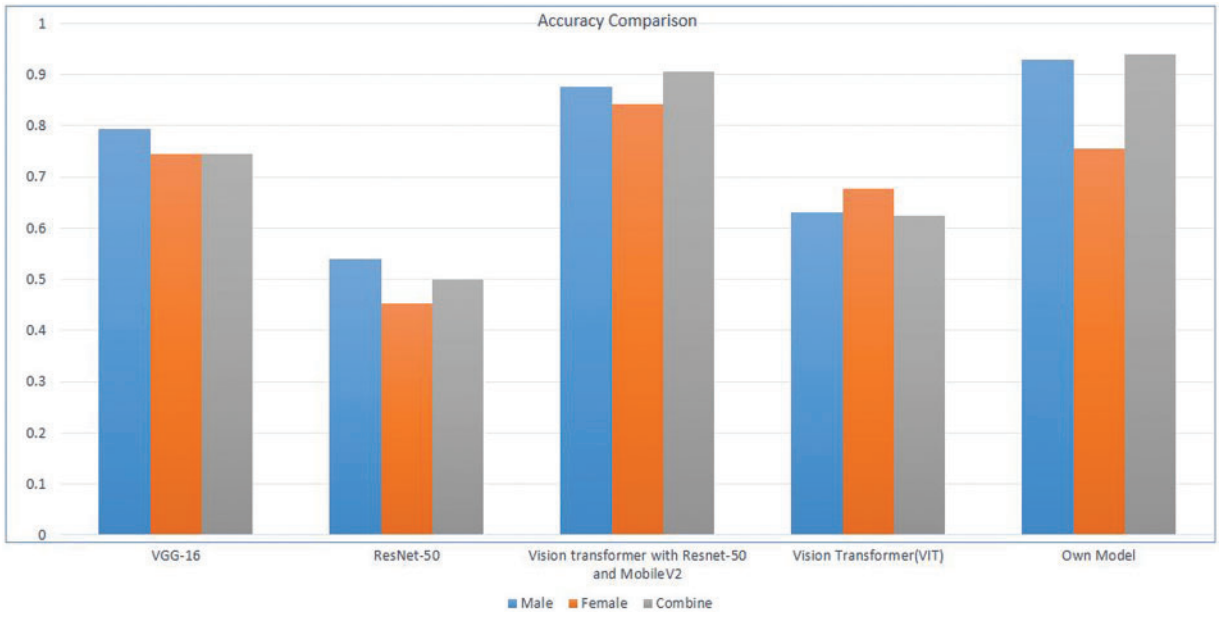

Figure 9: Different model accuracy comparison graph

Table 4: Prediction of each and every class from dataset images on different models

Model	Accuracy	Datasets	Classes	Images
VGG-16 testing	78.44	Female	**PHA**	10256
RestNet-50 testing	45.22		**PNH**	
Vision transformer	56.93			
ViT with MobileNetV2 and VGG-16	84.29			
Our model (MHAVH)	**75.47**			

(Continued)

Table 4 (continued)

Model	Accuracy	Datasets	Classes	Images
VGG-16 testing	72.5	Male	**PHA**	16983
RestNet-50 testing	53.84		**PNH**	
Vision transformer	70.28			
ViT with MobileNetV2 and VGG-16	87.48			
Our model (MHAVH)	**92.87**			
VGG-16 testing	74.52	Combine	**PHA**	16983
RestNet-50 testing	49.95		**PNH**	
Vision transformer	73.04			
ViT with MobileNetV2 and VGG-16	90.6			
Our model (MHAVH)	**93.96**			

4.3 Evaluation Metrics with and without Augmentation

During the augmentation process, several augmentation techniques were applied to the images. The augmented image parameters were set according to the specifications outlined in Table 5, which included a 20% share range and 20% zoom. Additionally, the images were rescaled by a factor of 1/225.0, and the fill mode applied was the nearest.

Table 5: The configuration of parameters for image augmentation

Image augmentation technique	Value
Rescaling	1/225.0
Shear range	0.2
Zoom range	0.2
Fill mode	Nearest

The analysis reveals that employing augmentation improves accuracy compared to scenarios without it. Without augmentation, accuracy is lower for the female dataset (74.01%) compared to augmented results. For male and combined datasets, accuracy without augmentation is 88.14% and 90.72%, respectively, and is found from 92.87% and 93.96% with augmentation. This consistent trend highlights the potential for superior image recognition quality through augmented data, as demonstrated in Table 6.

Table 6: Comparative analysis of MHAVH model performance with and without augmentation

Analysis	PREC	F1	ACCU	Train images	Test images	Datasets
Non-augmentation	99.28	85.27	88.14			
Augmentation	87.39	92.72	92.87	14554	2429	Male

(Continued)

Table 6 (continued)

Analysis	PREC	F1	ACCU	Train images	Test images	Datasets
Non-augmentation	87.06	72.11	74.01			
Augmentation	81.62	75.98	75.47	7932	2324	Female
Non-augmentation	91.13	90.68	90.72			
Augmentation	89.62	94	93.96	22344	4788	Combine

5 Model Evaluation and Discussion

In the trial of observing the essential capability of the proposed framework, as mentioned in Table 7, the MHAVH model is tested with a posture-based dataset proposed by Rojas-Albarracin et al. [3], which contains two classes with heart-attack and normal posture-based images. The overall higher accuracy and F1-scores are 93.96% and 94.0%, respectively. Therefore, this indicates that our proposed model performed well on heart disease classification.

Table 7: A performance of MHAVH model on male, female and combine datasets

Datasets	Predicted class	F1-score	Accuracy
Male	PHA and PNH	92.72%	92.87%
Female	PHA and PNH	75.98%	75.47%
Combine [3]	PHA and PNH	94%	93.96%

5.1 Comparison of Posture-Based Monitoring Systems

A comparative analysis of the proposed MHAVH framework performance with the previous related work on posture-based datasets [3] is performed as represented in Table 8. The results are generated on this dataset for all the pre-existing models to compare performance excellently with the proposed framework. The proposed MHAVH framework achieved better results with an accuracy of 93.96%. This demonstrates that the study of the model differentiates from other state-of-the-art methods. Our model gets better results, including CNN [32], ResNet-50 [6], You Only Look Once, Version 3 (YOLOv3) [33] Multi-Fusion method [34], and Vision Transformer [35]. At the same time, the performance of the other models was substantially good, with more than 80% accuracy, respectively. The ViT model performed ineffectively, with the lowest accuracy of 80%. The test results gave insights into the limited accuracy of a posture-based approach, such as the ViT, for the Blanket-Penetrating Sleep Posture. Table 8 shows that our recommended MHAVH framework performed well on heart attack posture-based detection with an accuracy (ACCU) of 93.96%.

Table 8: Comparison of existing posture based monitoring system

Ref.	Model	ACCU	Background	Dataset
[36]	CNN	**92**	Color and depth images	Own capture (Hand posture and gesture dataset)

(Continued)

Table 8 (continued)

Ref.	Model	ACCU	Background	Dataset
[32]	CNN	**74**	RGB image	Captured with his own camera using Kinect
[35]	ViT	**80**	RGB color images	Blanket-Penetrating sleep posture images using a triple ultra-wideband radar system
[6]	Modified Resnet-50	**92**	RGB-d color images	Heart-attack-detection-in-images
[33]	YOLOv3	**88** **83**	RGB input	Wave Punch (Posture-based images)
[34]	Multi-fusion method	**85.7**	–	Capture with smartphone and RFID device (Posture-based images)
[37]	CNN	**65**	RGB input	YouTube data fall images (human posture images)
[3]	CNN	**91.7**	RGB	Own internet-images (Heart-attack-detection-in-images
OWN MODEL	MHAVH	**93.96** **92.87** **75.47**	RGB input	Combine dataset Male dataset Female dataset

In Fig. 10, we demonstrated the visual comparison of multiple models *vs*. experiments with their results on posture-based dataset (not limited to heart attack posture based), comparing each with our proposed MHAVH model.

In Table 9, statistical data in terms of accuracy, precision, and F1-score based on the heart attack posture based datasets. On comparison, we found that our proposed model performed well compared to previously reported models.

Table 9: Comparison of the proposed MHAVH framework performance with existing state-of-the-art methods on the heart attack posture datasets

Ref.	Model	Accuracy	Precision	F1-score
[3]	CNN	91.75	90.84	45.91
[6]	Modified ResNet-50 network	92	90	92
[38]	Faster R-CNN	–	78.5	80.06
Our	MHAVH	**93.96**	89.62	**94**

5.2 Comparative Analysis with Other Hybrid Models

Table 10 presents a diverse array of deep-learning hybrid models. Notable combinations include Inception-V3, ResNet-50, and Vision Transformer for retinal disease detection, achieving an accuracy of 92.37% in 2023 [18]. Another hybrid model, combining a Convolutional Neural Network (CNN) with a vision transformer-based framework, achieves an 89.6% accuracy for surveillance anomaly detection in the same year [39]. Inception combined with convolutional vision transformers achieves 77.54% accuracy for plant disease identification [40]. Vision transformer paired with a Deep Convolutional Neural Network-based Model attains a 70.24% accuracy for classifying X-ray images [41]. In 2021, InceptionV3 combined with SVM achieved 94.8% accuracy for malaria cell-image classification [42]. ResNet, InceptionV3, and CNN combined reach an impressive 99.2% accuracy for diabetic eye disease detection in 2023 [43]. VGG, DenseNet, and logistic regression achieved an 87% accuracy for breast cancer detection in 2023 [44]. A convolutional neural network combined with long short-term memory (LSTM) for emotion recognition attained a 92.38% accuracy in 2022 [45]. Lastly, a hybrid model combining SVM, logistic regression, decision trees, Naive Bayes, random forest, and deep learning classifiers achieved a remarkable 98.14% accuracy 2021 for posture detection [46]. The chart shows how adaptable and successful hybrid deep learning models are in various computer vision and healthcare applications.

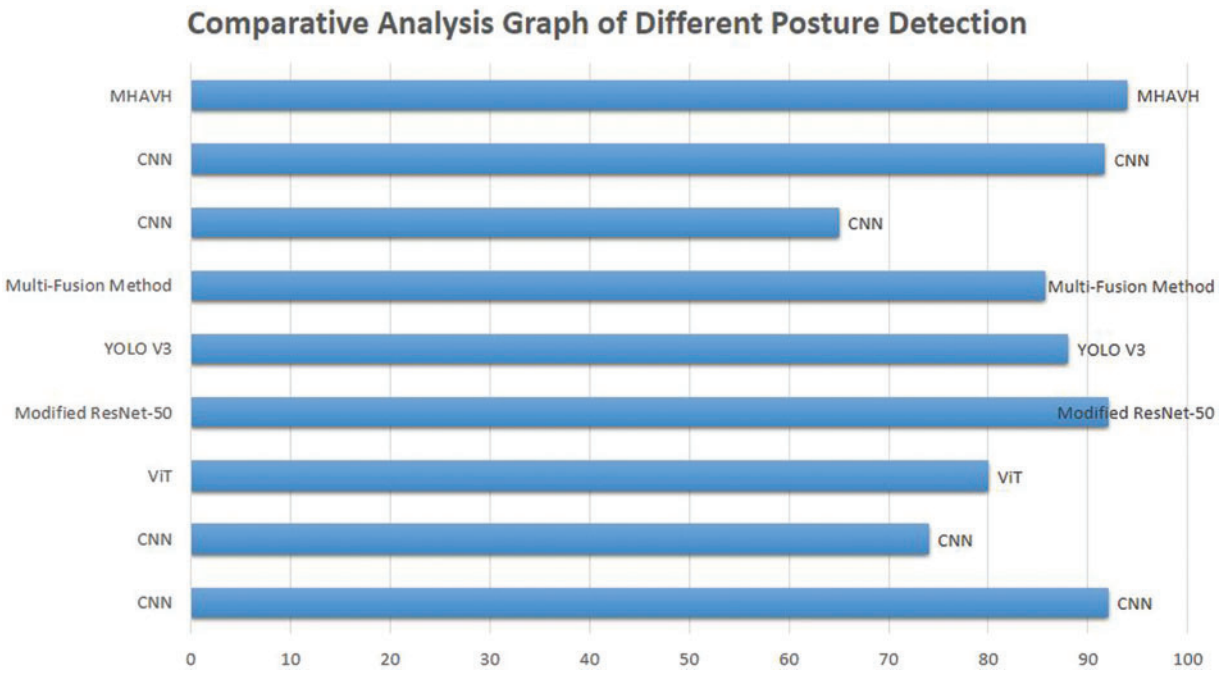

Figure 10: The graph shows the best accuracy achieved by our model as compared to other model

Table 10: Comparison of the proposed framework performance with existing other hybrid model state-of-the-art methods

Ref.	Hybrid models	ACCU	Year	Technique	Purpose of research
[17]	Inception-V3 + ResNet-50 + Vision Transformer	92.37	2023	Conv-Vit	Retinal disease detection
[38]	Convolution neural network (CNN) + vision transformer-based framework	89.6	2023	TransCNN	Surveillance anomaly detection
[39]	Inception + convolutional vision transformers	77.54	2023	–	Plant disease identification
[40]	Vision transformer + Deep convolutional neural network-based model	70.24	2023	–	Classifying X-ray Images
[41]	InceptionV3 + SVM	94.8	2021	–	Malaria cell-image classification
[42]	ResNet + InceptionV3 + CNN	99.2	2023	–	Diabetic eye disease
[43]	VGG + DenseNet + Logistic regression	87	2023	–	Breast cancer detection
[44]	Convolutional neural network + LSTM	92.38	2022	CNN-LSTM-attention (CLA)	Emotion recognition
[45]	Support vector machine (SVM) + logistic regression + k-nearest neighbors algorithm (KNN), decision + tree, Naive Bayes, random forest + deep learning classifiers	98.14	2021	Hybrid (CNN)	Posture detection framework
Own model	VGG-16 + ResNet-50 + Vision transformer	93.96		MHAVH	Heart attack posture

6 Conclusion and Future Work

We proposed a Multi-Head Attention Vision Hybrid (MHAVH) model for an efficient posture-based heart attack detection technique. VGG-16 achieved only 74.61% accuracy with the female dataset and 75.34% with the male dataset. Subsequently, we applied RestNet-50, which provided

an accuracy of 55.25% with the male dataset and 43.80% with the female dataset. In comparison to the vision transformer combined with MobileNetV2 and VGG-16 on the female dataset, the accuracy was 77.62%; on the male dataset, it was 82.17%, and with the combined dataset, the given accuracy was 88.74%. We used the MHAVH model, which obtained 89.21% accuracy with the male dataset. However, with female and combined datasets, we got an accuracy of 74.04% and 93.96%, respectively. The primary purpose of presenting this method is to control the death ratio and save human lives at early stages. This technique uses an RGB image with three datasets based on male and female, and it combines datasets with a data augmentation technique on each image dataset for the best result and increased training accuracy. This model also provides good accuracies with large and small datasets. Despite these efforts, the model's complexity and the datasets nature have presented significant challenges. To improve the model's generalization capability, further research is required to explore additional strategies, such as gathering more diverse data and simplifying the model architecture.

Acknowledgement: Researchers Supporting Project Number (RSPD2024R576), King Saud University, Riyadh, Saudi Arabia.

Funding Statement: This project is funded by King Saud University, Riyadh, Saudi Arabia.

Author Contributions: Study conception and design: Hina Naz; data collection: Hina Naz, Zaid Ali Khan Fuad A. Awwad; analysis and interpretation of results: Zaid Ali Khan, Mohammed Al-Habib, A. A. Ismail; draft manuscript preparation: Ismail, Zaid Ali Khan, Hina Naz, Mohammed Al-Habib; supervision, review, and project administration: Zuping Zhang. All authors reviewed the results and approved the final.

Availability of Data and Materials: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Gurjar and N. A. Sarnaik, "Heart attack detection by heartbeat sensing using Internet of Things: IoT," *Heart*, vol. 5, no. 3, 2018.

[2] Word Health Organization, "The top 10 causes of death," WHO. Accessed: Mar. 18, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[3] G. Rojas-Albarracin, M.Á. Chaves, A. Fernandez-Caballero, and M. T. Lopez, "Heart attack detection in colour images using convolutional neural networks," *Appl. Sci.*, vol. 9, no. 23, pp. 5065, 2019. doi: 10.3390/app9235065.

[4] H. Zhou, S. Huang, and Y. Xu, "Inceptr: Micro-expression recognition integrating inception-CBAM and vision transformer," *Multimed. Syst.*, vol. 29, no. 6, pp. 3863–3876, 2023. doi: 10.1007/s00530-023-01164-0.

[5] S. Aminizadeh *et al.*, "Opportunities and challenges of artificial intelligence and distributed systems to improve the quality of healthcare service," *Artif. Intell. Med.*, vol. 149, pp. 102779, 2024.

[6] Z. Ahmed, A. Irtaza, A. Mehmood, and M. F. Saleem, "An improved deep learning approach for heart attack detection from digital images," in *2022 Int. Conf. Front. Inf. Technol. (FIT)*, IEEE, 2022, pp. 261–266.

[7] Z. Amiri, A. Heidari, N. J. Navimipour, M. Esmaeilpour, and Y. Yazdani, "The deep learning applications in IoT-based bio-and medical informatics: A systematic literature review," *Neural Comput. Appl.*, vol. 36, pp. 1–41, 2024. doi: 10.1007/s00521-023-09366-3.

[8] M. Al-Habib, D. Huang, M. Al-Qatf, and K. Al-Sabahi, "Cooperative hierarchical framework for group activity recognition: From group detection to multi-activity recognition," in *Proc. 2019 8th Int. Conf. Sof. Comput. Appl.*, 2019, pp. 291–298.
[9] A. Rahman *et al.*, "Power mean based image segmentation in the presence of noise," *Sci. Rep.*, vol. 12, no. 1, pp. 21177, 2022. doi: 10.1038/s41598-022-25250-x.
[10] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mob. Inf. Syst.*, vol. 2018, pp. 1–21, 2018. doi: 10.1155/2018/3860146.
[11] S. Aziz, S. Ahmed, and M. S. Alouini, "ECG-based machine-learning algorithms for heartbeat classification," *Sci. Rep.*, vol. 11, no. 1, pp. 18738, 2021. doi: 10.1038/s41598-021-97118-5.
[12] H. Kabasawa, "MR imaging in the 21st century: Technical innovation over the first two decades," *Magn. Reson. Med. Sci.*, vol. 21, no. 1, pp. 71–82, 2022. doi: 10.2463/mrms.rev.2021-0011.
[13] A. Trockman and J. Z. Kolter, "Patches are all you need?," arXiv preprint arXiv:2201.09792, 2022.
[14] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, 2019. doi: 10.1186/s40537-019-0192-5.
[15] H. Malik and T. Anees, "BDCNet: Multi-classification convolutional neural network model for classification of COVID-19, pneumonia, and lung cancer from chest radiographs," *Multimed. Syst.*, vol. 28, no. 3, pp. 815–829, 2022. doi: 10.1007/s00530-021-00878-3.
[16] D. He and C. Xie, "Semantic image segmentation algorithm in a deep learning computer network," *Multimed. Syst.*, vol. 28, no. 6, pp. 2065–2077, 2022. doi: 10.1007/s00530-020-00678-1.
[17] Y. X. Li *et al.*, "Novel electronic health records applied for prediction of pre-eclampsia: Machine-learning algorithms," *Pregnancy Hypertens.*, vol. 26, no. 6, pp. 102–109, 2021. doi: 10.1016/j.preghy.2021.10.006.
[18] P. Dutta, K. A. Sathi, M. A. Hossain, and M. A. A. Dewan, "Conv-ViT: A convolution and vision transformer-based hybrid feature extraction method for retinal disease detection," *J. Imaging*, vol. 9, no. 7, pp. 140, 2023. doi: 10.3390/jimaging9070140.
[19] S. Jamil and A. M. Roy, "An efficient and robust phonocardiography (PCG)-based valvular heart diseases (VHD) detection framework using vision transformer (VIT)," *Comput. Biol. Med.*, vol. 158, pp. 106734, 2023. doi: 10.1016/j.compbiomed.2023.106734.
[20] A. M. Roy, "An efficient multi-scale CNN model with intrinsic feature integration for motor imagery EEG subject classification in brain-machine interfaces," *Biomed. Signal Process. Control*, vol. 74, pp. 103496, 2022. doi: 10.1016/j.bspc.2022.103496.
[21] S. Ahlawat and A. Choudhary, "Hybrid CNN-SVM classifier for handwritten digit recognition," *Procedia Comput. Sci.*, vol. 167, pp. 2554–2560, 2020. doi: 10.1016/j.procs.2020.03.309.
[22] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[23] H. D. Le, G. S. Lee, S. H. Kim, S. Kim, and H. J. Yang, "Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning," *IEEE Access*, vol. 11, pp. 14742–14751, 2023. doi: 10.1109/ACCESS.2023.3244390.
[24] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assis. Intervent. (MICCAI)*, 2021, pp. 109–119.
[25] H. Fan *et al.*, "Multiscale vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6824–6835.
[26] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," arXiv preprint arXiv:1606.01933, 2016.
[27] Z. Lin *et al.*, "A structured self-attentive sentence embedding," arXiv preprint arXiv:1703.03130, 2017.
[28] J. Li, Z. Tu, B. Yang, M. R. Lyu, and T. Zhang, "Multi-head attention with disagreement regularization," arXiv preprint arXiv:1810.10183, 2018.
[29] T. M. Nguyen *et al.*, "Improving transformers with probabilistic attention keys," in *Int. Conf. Mach. Learn.*, PMLR, 2022, pp. 16595–16621.

[30] D. Theckedath and R. Sedamkar, "Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks," *SN Comput. Sci.*, vol. 1, no. 2, pp. 79, 2020. doi: 10.1007/s42979-020-0114-9.
[31] Y. Huang, "ViT-r50 GAN: Vision transformers hybrid model based generative adversarial networks for image generation," in *2023 3rd Int. Conf. Consum. Electron. Comput. Eng. (ICCECE)*, IEEE, 2023, pp. 590–593.
[32] K. Adhikari, H. Bouchachia, and H. Nait-Charif, "Activity recognition for indoor fall detection using convolutional neural network," in *2017 Fifteenth IAPR Int. Conf. Mach. Vis. Appl. (MVA)*, IEEE, 2017, pp. 81–84.
[33] M. F. R. Lee, Y. C. Chen, and C. Y. Tsai, "Deep learning-based human body posture recognition and tracking for unmanned aerial vehicles," *Process*, vol. 10, no. 11, pp. 2295, 2022. doi: 10.3390/pr10112295.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[35] D. K. H. Lai *et al.*, "Vision Transformers (ViT) for blanket-penetrating sleep posture recognition using a triple ultra-wideband (UWB) radar system," *Sens.*, vol. 23, no. 5, pp. 2475, 2023. doi: 10.3390/s23052475.
[36] J. Pyo, S. Ji, S. You, and T. Kuc, "Depth-based hand gesture recognition using convolutional neural networks," in *2016 13th Int. Conf. Ubiquitous Robots Ambient Intell. (URAI)*, IEEE, 2016, pp. 225–227.
[37] Y. Fan, M. D. Levine, G. Wen, and S. Qiu, "A deep neural network for real-time detection of falling humans in naturally occurring scenes," *Neurocomput.*, vol. 260, pp. 43–58, 2017. doi: 10.1016/j.neucom.2017.02.082.
[38] H. Mohan, P. Rao, H. S. Kumara, and S. Manasa, "Non-invasive technique for real-time myocardial infarction detection using faster R-CNN," *Multimed. Tools Appl.*, vol. 80, no. 17, pp. 26939–26967, 2021. doi: 10.1007/s11042-021-10957-2.
[39] W. Ullah, T. Hussain, F. U. M. Ullah, M. Y. Lee, and S. W. Baik, "TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection," *Eng. Appl. Artif. Intell.*, vol. 123, pp. 106173, 2023. doi: 10.1016/j.engappai.2023.106173.
[40] S. Yu, L. Xie, and Q. Huang, "Inception convolutional vision transformers for plant disease identification," *Internet Things*, vol. 21, pp. 100650, 2023. doi: 10.1016/j.iot.2022.100650.
[41] O. Uparkar, J. Bharti, R. Pateriya, R. K. Gupta, and A. Sharma, "Vision transformer outperforms deep convolutional neural network-based model in classifying X-ray images," *Procedia Comput. Sci.*, vol. 218, pp. 2338–2349, 2023. doi: 10.1016/j.procs.2023.01.209.
[42] M. A. Reddy, G. S. S. R. Krishna, and T. T. Kumar, "Malaria cell-image classification using inceptionV3 and SVM," *Int J Eng Res Technol. (IJERT)*, vol. 10, no. 8, pp. 2021, 2021.
[43] M. M. Vespa and C. A. Kumar, "Diabetic eye disease in computerised tomography of feature extraction and classification in hybrid neural network," *Comput. Methods Biomech. Biomed. Eng.: Imag. Vis.*, pp. 1–13, 2024.
[44] S. Singh *et al.*, "Hybrid models for breast cancer detection via transfer learning technique," *Comput. Mater. Contin.*, vol. 74, pp. 3063–3083, 2022.
[45] H. Jiang, D. Wu, X. Tang, Z. Li, and W. Wu, "EEG Emotion recognition using an attention mechanism based on an optimized hybrid model," *Comput. Mater. Contin.*, vol. 73, no. 2, pp. 2697–2712, 2022. doi: 10.32604/cmc.2022.027856.
[46] S. Liaqat, K. Dashtipour, K. Arshad, K. Assaleh, and N. Ramzan, "A hybrid posture detection framework: Integrating machine learning and deep neural networks," *IEEE Sens. J.*, vol. 21, no. 7, pp. 9515–9522, 2021. doi: 10.1109/JSEN.2021.3055898.