MMPF: Multi-Modal Perception Framework for Abnormal Medical Condition Detection

Chuyi Zhong[1,2,3], **Dingkang Yang**[1,2,3], **Peng Zhai**[1,2,3], **Lihua Zhang**[1,2,3,4,5*]
[1]Academy for Engineering and Technology, Fudan University
[2]Cognition and Intelligent Technology Laboratory (CIT Lab)
[3]Institute of Metaverse & Intelligent Medicine, Fudan University
[4]Jilin Provincial Key Laboratory of Intelligence Science and Engineering, Changchun, China
[5]Engineering Research Center of AI and Robotics, Ministry of Education, Shanghai, China
{cyzhong20, dkyang20, pzhai, lihuazhang}@fudan.edu.cn

Abstract

As the global population ages and the incidence of chronic diseases increases, the demand for early detection of abnormal medical conditions is increasing. Traditional health monitoring methods often require significant resources and specialized personnel, limiting their widespread use. Leveraging advancements in AI technologies, this study proposes a non-invasive method for detecting abnormal medical conditions from image data. A multimodal perception framework is introduced, integrating features from various modalities, including facial expressions and body postures, to enhance detection accuracy. The framework employs a Cascaded Squeeze-Excitation (CSE) module, consisting of Adaptive and Multi-modal Squeeze-Excitation components, to capture complex feature dependencies and improve cross-modal performance. Extensive experiments demonstrate the effectiveness of this approach, showing improved performance over existing methods. In addition, a new dataset that encompasses a wide range of medical conditions has been released, providing a valuable resource for future research in this domain.

Introduction

With the aging population and the increasing prevalence of chronic diseases, the timely detection and identification of abnormal medical conditions have become crucial. According to the United Nations, by 2050, the global population over 60 is expected to reach 2 billion, with those over 65 reaching 1.5 billion (He et al. 2016b). This rapid societal aging will lead to various challenges, including an increase in the number of elderly people living alone, an uneven distribution of medical resources, and a shortage of nursing staff in care homes (Goodman et al. 2016). For elderly individuals living alone, the risk of death due to undetected sudden illness, thus missing the optimal time for intervention, is more dangerous than the risk posed by loneliness.

Early detection of abnormal medical conditions can identify potential signs of disease immediately, effectively preventing disease progression and potentially saving lives. Traditional health monitoring methods are mainly based on medical equipment and professional healthcare staff, which are time consuming, labor intensive, and difficult to implement widely in areas limited in resources (Baig and Gholamhosseini 2013; Mshali et al. 2018). Recently, with the rapid advances in computer vision and deep learning technologies, the detection of abnormal medical conditions through image data has emerged as a new research focus, surpassing traditional sensor-based methods. Unlike sensors, which can be redundant, noisy, and inconvenient to wear (Lara and Labrador 2012), images offer a resource-efficient data source that has been widely used in behavior recognition and other fields (Beddiar et al. 2020; Demrozi et al. 2020). Among these, abnormal medical condition recognition is an emerging area of interest. Previous studies have provided timely warnings by detecting falls in individuals (Harrou et al. 2017; Alam et al. 2022; Khraief, Benzarti, and Amiri 2020). Other research has extracted features from images to determine potential heart attack conditions (Rojas-Albarracin et al. 2019; Mohan et al. 2021). However, these methods primarily extract broad features from images without capturing the detailed external signs and specific abnormalities of individuals.

Research in this area has highlighted the significant role of facial expressions in conveying emotional states and indicating a person's current health status. Notably, when assessing an individual's health, negative changes in facial expressions often serve as the most intuitive outward manifestations of distress and physical discomfort (Prkachin 2009). Previous studies have frequently overlooked the critical role of facial expressions in detecting abnormal conditions. Additionally, body posture and scene context can provide valuable clues. Currently, no existing methods simultaneously account for the contributions of both facial expressions and body postures in detecting abnormal medical conditions, which may hinder progress and enhancement in this field.

Inspired by these insights, this study proposes a non-invasive method for rapidly detecting potential abnormal medical conditions from images, without the need for specialized equipment or medical staff. The main contributions of this work are as follows:

- A multi-stream framework is proposed, which integrates multi-modal features such as localized facial expressions and body postures to enhance the detection of potential medical conditions.

*Corresponding author

- A Cascade Squeeze-Excitation (CSE) modular structure is introduced, consisting of an Adaptive Squeeze-Excitation (ASE) module and a Multi-modal Squeeze-Excitation (MSE) module. This structure allows the network framework to better capture multi-modal and multi-dimensional features, thereby improving detection performance. Extensive experiments and studies confirm the effectiveness and superiority of the proposed network framework and modular structure.
- A novel dataset has been released, encompassing a wide range of medical conditions, providing a robust database for future medical condition detection tasks.

Related Works

Vision-based Abnormal Behavior / Medical Condition Detection

Vision-based methods for detecting abnormal behavior typically use various types of cameras, including single RGB, infrared, depth, and camera array-based 3D systems (Yadav et al. 2021). These methods capture relevant behavioral features for identification. The Microsoft Kinect series of RGB-D cameras (Zhang 2012), known for their low cost and ease of installation, has been widely adopted for behavioral data acquisition and recording. This adoption has led to numerous methods for behavioral recognition and fall detection. Camera array-based 3D methods are valued for their multi-view and occlusion-free capabilities. Advances in depth cameras have also propelled pose estimation techniques for skeletal keypoints, such as OpenPose (Cao et al. 2017) and AlphaPose (Fang et al. 2017). These developments have facilitated applications in activity recognition (Noori et al. 2019; Ghazal et al. 2019), fall detection (Huang et al. 2018), and gait analysis (D'Antonio et al. 2020).

Image-based methods for detecting abnormal medical conditions have attracted significant research interest, leveraging advances in computer vision and deep learning to analyze visual data for signs of distress or health issues. For example, Rojas et al. (2019) collected a dataset of images of people experiencing heart attacks and normal conditions, using convolutional neural networks (CNNs) to extract features for potential heart attack detection. Mohan et al. (2021) used the Faster R-CNN (Ren et al. 2015) framework, trained on NTU RGB+D (Shahroudy et al. 2016) and custom datasets, to detect chest pain and falls. Kim et al. (2021) applied a pre-trained ResNet (He et al. 2016a) to identify patients performing physical rehabilitation movements. McCay et al. (McCay et al. 2020) conducted early diagnosis of cerebral palsy in infants by detecting agitated movements from video frames.

Despite these advancements, vision-based detection methods face several limitations. Depth cameras and camera array approaches, while offering more dimensional information, come with higher costs and require more spatial accommodation, which limits their generalizability. Additionally, many existing methods are focused on detecting a single specific medical condition and cannot perform a broader range of detection tasks. They often extract broad, global features from images and overlook valuable details, such as facial expressions, which significantly enhance detection tasks in real-world applications. Furthermore, current model structures and feature extraction methods often fail to balance various features, such as postures and expressions, leading to unstable performance under complex conditions. Consequently, these approaches face challenges in task complexity, multi-modal integration, and real-time robustness, indicating a need for further improvement.

Multi-modal Fusion

Multi-modal fusion methods, combining data from images, skeletal data, and sounds, are gaining popularity in emotion perception and behavior recognition to improve detection accuracy and robustness. Early multi-modal fusion techniques often used simple feature concatenation from different modalities. Baltrusaitis et al. (2018) provided a comprehensive survey of multi-modal machine learning, noting that early fusion strategies merged features at the input level. Although this approach is straightforward, it suffers from a lack of interaction between different feature types, resulting in diluted representations and sub-optimal performance. An alternative is late fusion, which combines the outputs of independently trained models for each modality (Gadzicki, Khamsehashari, and Zetzsche 2020). While this method allows for independent optimization of each modality, it may not capture the synergistic relationships between them. Although late fusion mitigates some limitations of early fusion, it still falls short in effectively integrating cross-modal dependencies. To overcome these limitations, attention mechanisms have been introduced to dynamically weigh the importance of each modality, showing improved performance in tasks such as sentiment analysis and emotion recognition.

Building on these foundations, this work proposes the Cascade Squeeze-Excitation (CSE) module, which incorporates both channel-wise and modality-wise features. This module enables the network to adaptively emphasize the most informative features within each modality while integrating them across modalities. By capturing intricate feature inter-dependencies, the CSE module provides more robust feature descriptions. When combined with our multi-stream framework, our approach effectively handles complex multi-modal data, significantly enhancing the detection of abnormal medical conditions.

Methodology

Overview

In this study, a novel approach for detecting abnormal medical conditions through multi-modal data integration is introduced. The method involves extracting faces from images to analyze facial expressions, assigning both emotional state labels and condition labels. 2D poses are estimated and 3D poses are inferred using a lifting network to address potential torso occlusions. A multi-stream framework, combined with a Cascaded Squeeze-Excitation (CSE) module, enhances feature extraction and fusion across various modalities, including whole images, facial data, and posture information. Additionally, a random joint occlusion module

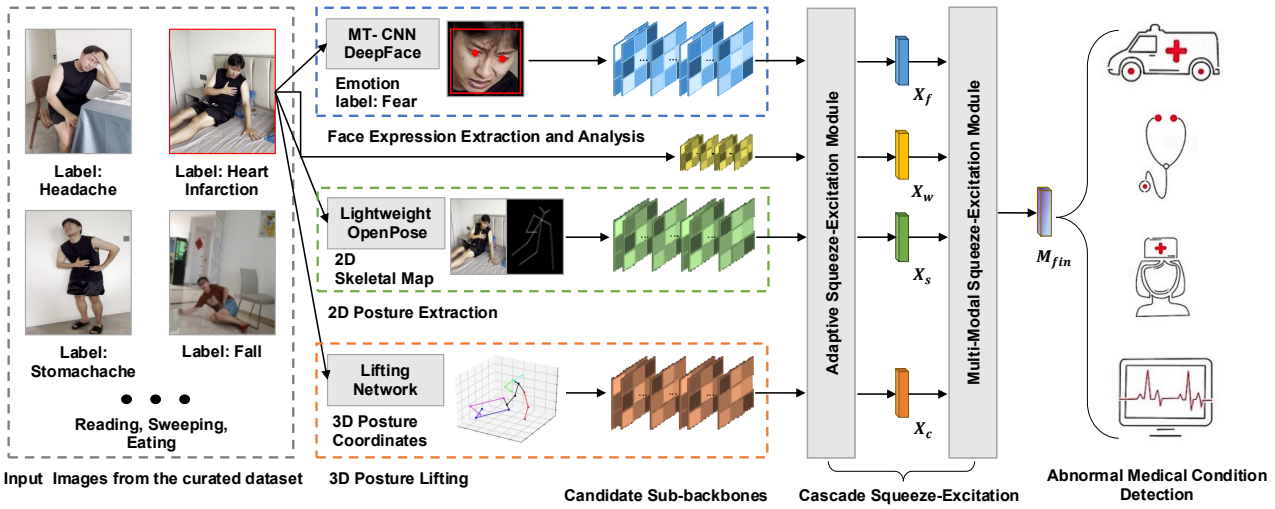

Figure 1: The architecture of the proposed methodology .

is employed for data augmentation during training, simulating occlusion scenarios to improve the model's performance in real-world conditions. The architecture of the proposed methodology is shown in Figure 1.

Facial Expression Extraction and Analysis

Facial expressions are crucial for conveying emotional states and serve as significant indicators of a person's health status (McLennan et al. 2019). Research indicates that negative changes in facial expressions during illness often reflect pain and can be a key indicator of medical conditions (Bargshady et al. 2020; Weitz et al. 2019). While facial expressions have been widely utilized in emotion recognition tasks (Canal et al. 2022), their role in behavioral recognition has also been explored (Islam et al. 2023). This approach incorporates a dedicated facial analysis component within the framework to leverage the information provided by facial expressions for detecting abnormal health conditions.

The facial branch is designed to detect, extract, and align face regions from input images. To achieve this, a deep cascading multitasking framework known as MT-CNN (Zhang et al. 2016) is employed. This framework utilizes a cascading architecture with three levels of deep convolutional networks to predict face and landmark locations from coarse to fine. By leveraging the intrinsic relationship between detection and alignment, MT-CNN enhances overall performance. Its accuracy and speed, along with its robustness to various angles, lighting conditions, and occlusions, make it a preferred choice for face extraction and alignment.

Once facial images are obtained using MT-CNN, facial expressions are analyzed with DeepFace (Taigman et al. 2014), a lightweight facial attribute analysis tool developed by Facebook. DeepFace assigns one of seven discrete expression labels—Anger, Fear, Neutrality, Sadness, Disgust, Happiness, and Surprise—to each facial image. Unlike previous works (Rojas-Albarracin et al. 2019; Mohan et al. 2021), this approach incorporates expression labels in addition to labeling the medical conditions of the images. This creates a multi-label classification problem, where each image is assigned multiple labels reflecting both its emotional state and medical condition.

2D Posture Extraction & 3D Posture Lifting

In addition to facial expressions, body posture provides crucial insights into an individual's current health status. Early signs of certain life-threatening conditions, such as heart disease and cerebral infarction, often manifest through specific postural symptoms. Common symptoms include chest pain and tightness (84%), pain and numbness in the upper extremities (56.1%), shortness of breath (47.9%), headache and dizziness (17.3%), sweating (13.8%), and back pain (5.2%) (Smith et al. 2002; Birnbach, Höpner, and Mikolajczyk 2020). These symptoms often correlate with characteristic postures. For instance, individuals experiencing chest pain may bend their bodies and place their hands on their chests, while those suffering from headaches or dizziness might cover their foreheads with their hands. Similarly, individuals with abdominal pain often place their hands on their abdomen. These specific postures are valuable for assessing health conditions, highlighting the importance of incorporating a posture analysis component into the framework.

2D Pose Estimation Several well-established methods for human pose estimation include OpenPose, AlphaPose, and the Stacked Hourglass (SH) network (Newell, Yang, and Deng 2016). Given the potential application scenarios of the proposed approach on mobile platforms, which are constrained by computational power and storage resources, Lightweight OpenPose (Osokin 2018) is selected for 2D pose estimation. This lightweight model offers comparable accuracy to OpenPose while requiring fewer computational resources, enabling faster real-time processing. For a given image $I \in \mathbb{R}^{H \times W \times C}$, the feature map F can be extracted through a lightweight convolutional neural network (CNN). Then, the heatmaps H_k for keypoint k and Part Affinity

Fields (PAFs) P_l for connection l can be predicted as:

$$H_k = \sigma(W_k * F), P_l = \tau(W_l * F), H_k, P_l \in \mathbb{R}^{H' \times W'}. \quad (1)$$

Here, W_k and W_l represent the convolutional filters, while σ and τ are the activation functions. Non-Maximum Suppression (NMS) is applied on the heatmaps H_k to determine the candidate keypoint positions $\{\hat{x}_k\}$. These keypoints are then associated with the predicted PAFs P_l to construct the skeletal structure. For each pair of keypoints k_1 and k_2, the connection score is computed as:

$$\text{Score}(k_1, k_2) = \int_{\hat{x}_{k_2}}^{\hat{x}_{k_1}} \left(P_l(x) \cdot \frac{\hat{x}_{k_2} - \hat{x}_{k_1}}{\|\hat{x}_{k_2} - \hat{x}_{k_1}\|} \right) dx. \quad (2)$$

This integral evaluates the alignment of the PAF values with the direction vector between keypoints, facilitating a robust connection between them.

3D Pose Estimation The 2D pose provides only planar position information and does not fully capture the three-dimensional structure and depth variation of the human body. To enhance detection accuracy and robustness, a 3D pose estimation is introduced within the pose branch. The 3D pose offers more precise joint positions and angles, improving depth perception and addressing limitations associated with 2D pose estimation, such as handling occlusions, visual changes, and variations in clothing. This addition significantly enhances the robustness of the method in complex scenarios.

This work employs a Lifting Network (Tome, Russell, and Agapito 2017) to perform 3D pose estimation directly from a single image. This method first estimates 2D pose from the RGB image and then uses the lifting network to transform the 2D joint positions, $P_{2D} = \{(x_k, y_k) \mid k \in \{1, 2, \ldots, K\}\}$, to the 3D joint positions in Cartesian space, $P_{3D} = \{(x_k, y_k, z_k) \mid k \in \{1, 2, \ldots, K\}\}$.

The lifting network comprises several hidden layers, where each layer is computed as:

$$h_i = \sigma(W_i h_{i-1} + b_i), \quad i = 1, \ldots, n. \quad (3)$$

Here, σ denotes the nonlinear activation function, and W_i and b_i are the weight and bias parameters for the respective layer. The final output layer maps the output of the hidden layers to 3D joint positions:

$$P_{3D} = W_{\text{out}} h_n + b_{\text{out}}. \quad (4)$$

During training, labeled 2D and 3D joint position data are used to optimize the network parameters by minimizing the Euclidean distance between the predicted and true 3D joint positions. The loss function is defined as follows:

$$L = \sum_{k=1}^{K} \left\| \hat{p}_{3D,k} - p_{3D,k} \right\|^2, \quad (5)$$

where $\hat{p}_{3D,k}$ is the predicted 3D position, $p_{3D,k}$ is the ground truth 3D position of the k-th joint, and K represents the total number of joints. The 2D pose and 3D pose obtained from the posture branch will be used for the anomaly detection in the data format of the skeletal map and 3D coordinates, which are complementary.

Cascade Squeeze-Excitation Module

Inspired by the Squeeze-Excitation Network (SENet) (Hu et al. 2019), a novel modular structure called the Cascade Squeeze-Excitation (CSE) module is proposed. The CSE module comprises two components: the Adaptive Squeeze-Excitation (ASE) module and the Multi-modal Squeeze-Excitation (MSE) module, which operate on each modality input as well as at the final multi-modal fusion stage. The SENet has demonstrated outstanding performance in ImageNet (Deng et al. 2009) databases by explicitly capturing channel dependencies between feature maps, showing that learning channel-level nonlinear attention enhances feature discrimination. Despite its success, SENet has some limitations. Specifically, SENet employs global average pooling to integrate spatial information. While global average pooling provides a comprehensive view of the inputs, it only considers global information and fails to capture inter-channel interactions, making the learned nonlinear attention susceptible to noise. To address these issues, an adaptive mechanism is introduced in the Squeeze-Excitation module to model associations between different channels more effectively, thereby capturing complex feature dependencies and improving the robustness and accuracy of feature representation.

Adaptive Squeeze-Excitation Module For each input modal feature $X_i \in \mathbb{R}^{H \times W \times C}$, the feature vector after global average pooling $z_i \in \mathbb{R}^C$ can be obtained as follows:

$$z_i = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} X_{i,h,w,c}, \quad (6)$$

where H and W are the height and width of the feature map, and C is the number of channels.

In the squeeze phase, an adaptive attention mechanism is introduced that computes the corresponding query, key, and value variables through three fully connected layers: $q_i = \text{ReLU}(W_q z_i + b_q)$, $k_i = \text{ReLU}(W_k z_i + b_k)$, $v_i = \text{Sigmoid}(W_v z_i + b_v)$, where W_q, W_k, W_v, b_q, b_k, and b_v are the weights and biases of the fully connected layers for the query, key, and value variables, respectively. γ is the reduction ratio in the squeeze phase.

Next, the attention weights are computed as:

$$\omega_i = \text{softmax}\left(\frac{q_i \cdot k_i^T}{\sqrt{d_k}} \right), \quad (7)$$

where $d_k = \frac{C}{\gamma}$ is the scaling factor used to stabilize the gradients. The reshaped excitation vector is then obtained as:

$$e_i = \text{reshape}(\omega_i \cdot v_i) \in \mathbb{R}^{1 \times 1 \times C}. \quad (8)$$

The weighted excitation feature matrix is subsequently obtained by multiplying the input feature matrix with the excitation vector: $X_i' = X_i \odot e_i$, where $\odot$ denotes element-wise multiplication. This process ensures that the network adaptively emphasizes the most informative features, enhancing the model's ability to capture complex feature dependencies and improving overall feature discrimination.

Multi-modal Squeeze-Excitation Module The Multi-modal Squeeze-Excitation (MSE) module follows the same principles as the Adaptive Squeeze-Excitation (ASE) module, aiming to fuse features from multiple modalities by concatenating and weighting the global features. This approach enables a comprehensive utilization of multi-modal features. Let M_i denote the input feature of the i-th modality. We concatenate the features from all input modalities as follows:

$$M_{\text{concat}} = \text{Concat}(M_1, \ldots, M_n) \in \mathbb{R}^{H \times W \times (n \cdot C)}. \quad (9)$$

Global average pooling is then performed on the concatenated features to obtain the global feature vector:

$$Z = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} M_{\text{concat},i,j} \in \mathbb{R}^{n \cdot C}. \quad (10)$$

Similarly, the query, key, and value vectors can be obtained as follows: $Q = \text{ReLU}(W_q Z + B_q)$, $K = \text{ReLU}(W_k Z + B_k)$, $V = \text{Sigmoid}(W_v Z + B_v)$, where $W_q, W_k \in \mathbb{R}^{(n \cdot C) \times \frac{n \cdot C}{\gamma}}$ and $W_v \in \mathbb{R}^{(n \cdot C) \times (n \cdot C)}$. The attention weights are then calculated as:

$$A = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{\frac{n \cdot C}{\gamma}}}\right). \quad (11)$$

Subsequently, the excitation vector is derived:

$$E = \text{reshape}(A \cdot V, (1, 1, n \cdot C)). \quad (12)$$

Finally, the feature representation for each modality is obtained as:

$$M_i' = M_i \odot E_i, \quad E_i \in \mathbb{R}^{1 \times 1 \times C}. \quad (13)$$

This process ensures that each modality's features are enhanced and weighted appropriately based on the learned attention, leading to improved multi-modal feature integration.

Random Joint Occlusion Module

To enhance the robustness of our model in handling complex real-world scenarios, this work introduces a random joint occlusion module during the training phase. This module simulates random occlusions due to various factors such as other objects, body parts, or environmental conditions.

First, the 2D pose in the image I can be obtained through the lightweight OpenPose network in the posture branch and get the 2D coordinates of each key point, $P_{\text{2D}} = \{(x_k, y_k) \mid k \in \{1, 2, \ldots, K\}\}$, as well as the visibility, $V(P_{\text{2D}}) \in \{0, 1\}^K$. If the key point is visible and the random probability is less than 0.5 (i.e., 50% probability of occlusion), a random occlusion block will be generated to occlude the corresponding joint point according to the size of the image, $s = \text{int}(\min(H, W) \times r)$, where r is the scale range of the occlusion block relative to the original image size. Then determine the boundaries and generate the occlusion block: $x_1 = \max(0, x - s), y_1 = \max(0, y - s)$, $x_2 = \min(W, x + s), y_2 = \min(H, y + s)$; $I[y_1 : y_2, x_1 : x_2] = 0$.

During the training process, the random joint occlusion module can generate rich, unique occlusion patterns in the training samples, improving the network's adaptability to various occlusion situations and enhancing our model's robustness. Figure 2 shows a schematic of the proposed Random Joint Occlusion module.

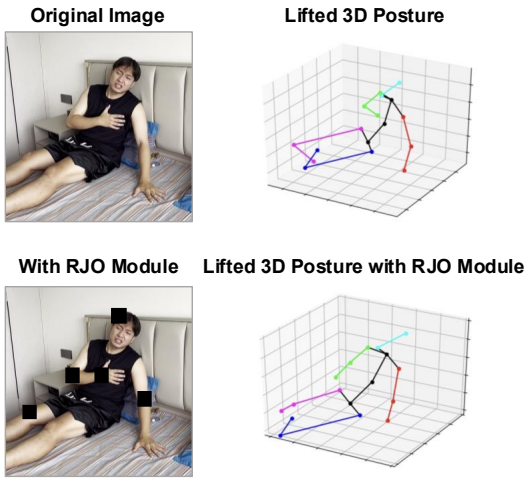

Figure 2: The schematic of the Random Joint Occlusion Module.

Experiments

Datasets

This study develops and releases a novel image dataset to address the lack of benchmark datasets for medical condition detection. The dataset was carefully selected after extensive research and expert consultation and includes symptom presentation categories of common diseases: infarction (cardiovascular disease), headache (migraine), stomach pain (abdominal pain), and falls (stroke and fainting). The dataset presents symptoms in a comprehensive visual form, focusing on facial expressions and body postures to ensure the content is reliable and informative. To enhance generalizability, images of common daily activities considered normal condition categories are also included: sweeping, reading, and eating. This versatile dataset can be used for abnormal and normal condition detection tasks. Each category contains 750 images from authorized web content (randomly split) and 750 reproduced images imitated by experimenters under expert guidance (split by the individual to avoid overlap). The total size of the dataset comes to 10,500 images, with a split ratio of 70%:15%:15%.

Implementation Details

The framework is implemented using TensorFlow (Pang, Nijkamp, and Wu 2020) with Keras (Géron 2022), and training is conducted on two Nvidia Quadro RTX 8000 GPUs. The AdamW optimizer (Loshchilov and Hutter 2017) is employed, with an initial learning rate of $1e^{-3}$ and weight decay of $1e^{-4}$. Models are trained with a batch size of 32 for 50 epochs. The input dimensions are set to $(3, 224, 224)$ for images and $(1, 3, 17)$ for 3D posture coordinates. Hyperpa-

Modality				Acc.	F1	Emo CG-Acc.	Emo CG-F1	Emo&Cond CG-Acc.	Emo&Cond CG-F1	ID
Whole Image	Face Image	2D Skeletal Map	3D Coordinates							
	VGG-19		3DCNN	86.52	85.38	91.12	88.72	92.74	91.53	1
	ResNet512		3DCNN	87.90	86.84	91.75	89.80	92.37	91.85	2
	DenseNet		3DCNN	87.91	<u>86.87</u>	<u>92.15</u>	<u>90.38</u>	<u>93.88</u>	<u>91.91</u>	3
	EfficientNet		3DCNN	<u>88.13</u>	**87.74**	**93.11**	**91.79**	**94.02**	**92.30**	4
	MobileNetV2		3DCNN	**88.71**	86.25	91.80	89.20	92.37	90.57	5
	EfficientNet		VoxelNet	87.63	86.55	91.14	88.96	92.12	90.67	6
	ResNet512		I3DNet	87.97	85.59	92.13	90.27	92.07	91.18	7
	MobileNetV2		I3DNet	87.30	85.96	91.02	89.30	92.17	90.73	8
EfficientNet	MobileNetV2	ResNet512	VoxelNet	87.50	85.47	90.67	88.15	92.24	90.69	9

Table 1: Experimental results of different backbone arrangement. The best results are marked in **bold** and the second-best are marked <u>underlined</u>.

Modality Streams				Acc.	F1
Whole Image	Face Image	2D Skeletal Map	3D Coordinates		
✓				79.83	78.90
	✓			52.81	49.55
		✓		62.50	61.93
			✓	63.15	62.32
✓	✓			83.42	82.88
✓	✓	✓		86.77	86.10
✓	✓	✓	✓	**88.13**	**87.74**

Table 2: Experimental results with different modality streams.

rameters are selected based on performance on the validation set, and the Cascade Squeeze-Excitation module is used as the default experimental setup. Additionally, for real-world application scenarios, emotional coarse-graining (Emo CG: Positive, Neutral, Negative) and conditional coarse-graining (Cond CG: Infarction, Headache, Stomachache, Fall, Normal) are also evaluated. Performance is measured using classification accuracy (Acc.) and the weighted F1 score (F1).

Experimental Results and Analyses

Model Zoo Classical 2D convolutional neural networks like ResNet, VGG (Simonyan and Zisserman 2014), and DenseNet (Huang et al. 2017) have demonstrated significant success in image-based recognition tasks. In this study, these architectures were utilized with minimal modifications. Additionally, lightweight networks designed for mobile platforms, including MobileNet (Sandler et al. 2018) and EfficientNet (Tan and Le 2020), were incorporated due to their resource efficiency, which is particularly advantageous for medical condition detection tasks. For feature extraction from skeletal keypoints, 3D CNN (Tran et al. 2015), I3DNet (Carreira and Zisserman 2017), and VoxelNet (Zhou and Tuzel 2018) were employed. Each model was evaluated under different pattern arrangements, with the Cascade Squeeze-Excitation (CSE) module and the random joint occlusion module as the default configurations.

Several key insights were derived from the experimental results presented in Table 1: (1) Although deeper network architectures generally yield slightly better results, the performance gain is marginal. This finding suggests that the effectiveness of the proposed method is not heavily dependent on the choice of backbone network, allowing for flexibility in model selection without significantly compromising results. (2) The coarse-grained evaluation metrics (CG-Acc./F1) consistently outperform standard accuracy and F1 score metrics. Notably, the improvement in overall detection performance is more pronounced in emotional coarse-graining. This may be due to the dataset primarily consisting of images captured in real-world environments, where fine-grained distinctions between emotions are challenging due to the limited richness and precision of the facial information modality. Despite these challenges, the method demonstrates effectiveness in recognizing coarse-grained emotional states and robustness in less controlled settings. (3) Resource-efficient model combinations achieve competitive or even superior results compared to denser network structures. This finding indicates that performance and real-time processing requirements can be balanced when selecting models for specific application scenarios and platforms.

Impact of Distinct Modalities To evaluate the impact of different input modalities, the highest-performing combination (ID 4) from Table 1 was selected as a baseline, and additional modalities were incrementally incorporated for testing. The results, summarized in Table 2, reveal several key insights: (1) The whole image, when used as an isolated input, captures the overall context and information. In contrast, relying solely on facial expressions, 2D skeletal maps, or 3D posture coordinates results in sub-optimal performance due to the limited and incomplete nature of these individual modalities. Notably, body posture modalities convey more relevant information about medical conditions compared to facial expressions. (2) As additional modalities are progressively integrated, both facial expressions and body postures contribute to a more comprehensive

Baseline	Acc.
B1: MobileNet+3DCNN+Concat.	86.73
B2: MobileNet+3DCNN+SE+Concat.	87.15
B3: MobileNet+3DCNN+MSE	88.12
B4: MobileNet+3DCNN+ASE+Concat.	87.81
B5: MobileNet+3DCNN+ASE+Add.	87.78
B6: MobileNet+3DCNN+ASE+MSE	**88.71**

Table 3: Ablation study for Cascade Squeeze-Excitation Modules. Only Acc. are reported due to similar results to F1.

understanding of the individual's current medical condition. (3) Although 3D posture coordinates are derived from 2D skeletal maps, they provide enriched spatial information that complements the 2D maps, thereby further enhancing the model's performance.

Effectiveness of CSE Modules As previously mentioned, the Adaptive Squeeze-Excitation (ASE) Module and the Multi-modal Squeeze-Excitation (MSE) Module are integral components of the proposed CSE module. To assess their effectiveness, an ablation study was conducted using MobileNet, a lightweight model that balances performance and training efficiency, as the baseline. In this study, **Concat** refers to the concatenation feature fusion operation, **Add** denotes the additive feature fusion operation, and **SE** references the Squeeze-and-Excitation (SENet) mechanism. The results in Table 3 demonstrate the superior performance of the proposed ASE and MSE modules in enhancing the expressive power of multi-modal features: (1) Compared to traditional concatenation and additive fusion strategies, the MSE module shows marked performance improvement (as seen when comparing B4, B5, and B6). (2) The ASE module outperforms SENet, validating the effectiveness and value of the adaptive strategy (as seen when comparing B2 and B4). (3) While the ASE module alone is slightly less effective than the MSE module, likely due to limited variability among channel features within a single modality, their combination synergistically enhances the representation of multi-modal feature fusion.

Comparison of Other Methods Previous related works (Rojas-Albarracin et al. 2019; Mohan et al. 2021; Khraief, Benzarti, and Amiri 2020; Gul et al. 2020) are summarized, noting that differences in classification tasks and datasets used make direct comparisons of results challenging. Methods from these studies were replicated based on the original papers and comparison experiments were conducted on the newly developed dataset and the NTU RGB+D dataset for the medical condition categories (A43: Falling Down, A44: Headache, A45: Chest Pain, and A46: Back Pain). The experimental results listed in Table 4 demonstrate the clear superiority of the proposed method. While the performance on the NTU RGB+D dataset is slightly lower than on our dataset, this may be attributed to the incomplete representation of faces and body postures due to the multiple viewpoints in NTU RGB+D data, and the fact that experiment

Method	Acc.	
	Our Dataset	NTU
(Rojas-Albarracin et al. 2019)	79.92	70.26
(Mohan et al. 2021)	81.37	73.40
(Khraief, Benzarti, and Amiri 2020)	73.20	68.55
(Gul et al. 2020)	82.41	77.18
Ours (ID4)	**88.13**	**82.51**

Table 4: Comparison results with other methods.

Baseline	F1	
	with RJO.	w/o RJO.
MobileNet+3DCNN+CSE	**86.25**	84.73

Table 5: Experimental results with/without Random Joint Occlusion Module during the training process.

participants may not have accurately replicated real facial expressions when simulating medical conditions. Nevertheless, our approach remains highly competitive, proving its effectiveness even in challenging scenarios with incomplete information.

Benefit from Random Joint Occlusion Module The effectiveness of the proposed Random Joint Occlusion Module is evaluated by comparing the performance of the method with and without this module. The results, as presented in Table 5, clearly demonstrate the positive impact of the module on the overall performance of the model. By generating diverse occlusion patterns during training, the module significantly enhances the model's robustness, enabling it to generalize more effectively to real-world scenarios where occlusions are common.

Conclusion and Discussion

This research introduces a pioneering Multi-Modal Perception Framework (MMPF) designed for the detection of abnormal medical conditions from images. The proposed framework integrates various modalities, such as facial expressions and body postures, to improve detection accuracy. By employing the Cascade Squeeze-Excitation (CSE) module, the model effectively captures intricate feature dependencies, thereby improving cross-modal performance. Extensive qualitative and quantitative experiments conducted on both a newly curated dataset and established benchmarks demonstrate the superiority of MMPF over existing methodologies in identifying a wide range of abnormal medical conditions. Furthermore, the newly curated dataset provides a robust resource for future research in the field, underscoring its critical role in supporting timely medical interventions.

Future work will expand the dataset to reflect real-world healthcare scenarios, integrate clinical information for diagnoses, and apply our approach to chronic disease diagnosis, including early detection of symptoms. These efforts aim to advance both research and practical healthcare applications.

Acknowledgments

This work was supported by Shanghai Municipal Science and Technology Major Project 2021SHZDZX0103 and National Key R&D Program of China 2021ZD0113503.

References

Alam, E.; Sufian, A.; Dutta, P.; and Leo, M. 2022. Vision-based human fall detection systems using deep learning: A review. *Computers in biology and medicine*, 146: 105626.

Baig, M. M.; and Gholamhosseini, H. 2013. Smart health monitoring systems: an overview of design and modeling. *Journal of medical systems*, 37: 1–14.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Challenges and applications in multimodal machine learning. *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, 17–48.

Bargshady, G.; Zhou, X.; Deo, R. C.; Soar, J.; Whittaker, F.; and Wang, H. 2020. Enhanced deep learning algorithm development to detect pain intensity from facial expression images. *Expert systems with applications*, 149: 113305.

Beddiar, D. R.; Nini, B.; Sabokrou, M.; and Hadid, A. 2020. Vision-based human activity recognition: a survey. *Multimedia Tools and Applications*, 79(41): 30509–30555.

Birnbach, B.; Höpner, J.; and Mikolajczyk, R. 2020. Cardiac symptom attribution and knowledge of the symptoms of acute myocardial infarction: a systematic review. *BMC Cardiovascular Disorders*, 20(1): 1–12.

Canal, F. Z.; Müller, T. R.; Matias, J. C.; Scotton, G. G.; de Sa Junior, A. R.; Pozzebon, E.; and Sobieranski, A. C. 2022. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582: 593–617.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Demrozi, F.; Pravadelli, G.; Bihorac, A.; and Rashidi, P. 2020. Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey. *IEEE access*, 8: 210816–210836.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. IEEE.

D'Antonio, E.; Taborri, J.; Palermo, E.; Rossi, S.; and Patane, F. 2020. A markerless system for gait analysis based on OpenPose library. In *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 1–6. IEEE.

Fang, H.-S.; Xie, S.; Tai, Y.-W.; and Lu, C. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, 2334–2343.

Gadzicki, K.; Khamsehashari, R.; and Zetzsche, C. 2020. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd international conference on information fusion (FUSION)*, 1–6. IEEE.

Géron, A. 2022. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.".

Ghazal, S.; Khan, U. S.; Mubasher Saleem, M.; Rashid, N.; and Iqbal, J. 2019. Human activity recognition using 2D skeleton data and supervised machine learning. *IET image processing*, 13(13): 2572–2578.

Goodman, C.; Dening, T.; Gordon, A. L.; Davies, S. L.; Meyer, J.; Martin, F. C.; Gladman, J. R.; Bowman, C.; Victor, C.; Handley, M.; et al. 2016. Effective health care for older people living and dying in care homes: a realist review. *BMC health services research*, 16: 1–14.

Gul, M. A.; Yousaf, M. H.; Nawaz, S.; Ur Rehman, Z.; and Kim, H. 2020. Patient monitoring by abnormal human activity recognition based on CNN architecture. *Electronics*, 9(12): 1993.

Harrou, F.; Zerrouki, N.; Sun, Y.; and Houacine, A. 2017. Vision-based fall detection system for improving safety of elderly people. *IEEE Instrumentation & Measurement Magazine*, 20(6): 49–55.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, W.; Goodkind, D.; Kowal, P. R.; et al. 2016b. An aging world: 2015.

Hu, J.; Shen, L.; Albanie, S.; Sun, G.; and Wu, E. 2019. Squeeze-and-Excitation Networks. arXiv:1709.01507.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Huang, Z.; Liu, Y.; Fang, Y.; and Horn, B. K. 2018. Video-based fall detection for seniors with human pose estimation. In *2018 4th international conference on Universal Village (UV)*, 1–4. IEEE.

Islam, M. M.; Nooruddin, S.; Karray, F.; and Muhammad, G. 2023. Multi-level feature fusion for multimodal human activity recognition in Internet of Healthcare Things. *Information Fusion*, 94: 17–31.

Khraief, C.; Benzarti, F.; and Amiri, H. 2020. Elderly fall detection based on multi-stream deep convolutional networks. *Multimedia Tools and Applications*, 79(27): 19537–19560.

Kim, J.-K.; Lee, K. B.; Kim, J.-C.; and Hong, S. G. 2021. Patient identification based on physical rehabilitation movements using skeleton data. In *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, 1572–1574. IEEE.

Lara, O. D.; and Labrador, M. A. 2012. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials*, 15(3): 1192–1209.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

McCay, K. D.; Ho, E. S.; Shum, H. P.; Fehringer, G.; Marcroft, C.; and Embleton, N. D. 2020. Abnormal infant movements classification with deep learning on pose-based features. *IEEE Access*, 8: 51582–51592.

McLennan, K. M.; Miller, A. L.; Dalla Costa, E.; Stucke, D.; Corke, M. J.; Broom, D. M.; and Leach, M. C. 2019. Conceptual and methodological issues relating to pain assessment in mammals: The development and utilisation of pain facial expression scales. *Applied Animal Behaviour Science*, 217: 1–15.

Mohan, H.; Rao, P.; Kumara, H. S.; and Manasa, S. 2021. Non-invasive technique for real-time myocardial infarction detection using faster R-CNN. *Multimedia Tools and Applications*, 80(17): 26939–26967.

Mshali, H.; Lemlouma, T.; Moloney, M.; and Magoni, D. 2018. A survey on health monitoring systems for health smart homes. *International Journal of Industrial Ergonomics*, 66: 26–56.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked Hourglass Networks for Human Pose Estimation. arXiv:1603.06937.

Noori, F. M.; Wallace, B.; Uddin, M. Z.; and Torresen, J. 2019. A robust human activity recognition approach using openpose, motion features, and deep recurrent neural network. In *Image Analysis: 21st Scandinavian Conference, SCIA 2019, Norrköping, Sweden, June 11–13, 2019, Proceedings*, 299–310. Springer.

Osokin, D. 2018. Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose. arXiv:1811.12004.

Pang, B.; Nijkamp, E.; and Wu, Y. N. 2020. Deep learning with tensorflow: A review. *Journal of Educational and Behavioral Statistics*, 45(2): 227–248.

Prkachin, K. M. 2009. Assessing pain by facial expression: facial expression as nexus. *Pain Research and Management*, 14(1): 53–58.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Rojas-Albarracin, G.; Chaves, M. Á.; Fernandez-Caballero, A.; and Lopez, M. T. 2019. Heart attack detection in colour images using convolutional neural networks. *Applied Sciences*, 9(23): 5065.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. arXiv:1604.02808.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, K. L.; Cameron, P. A.; Meyer, A.; and McNeil, J. J. 2002. Knowledge of heart attack symptoms in a community survey of Victoria. *Emergency Medicine*, 14(3): 255–260.

Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1701–1708.

Tan, M.; and Le, Q. V. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946.

Tome, D.; Russell, C.; and Agapito, L. 2017. Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Weitz, K.; Hassan, T.; Schmid, U.; and Garbas, J.-U. 2019. Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods. *tm-Technisches Messen*, 86(7-8): 404–412.

Yadav, S. K.; Tiwari, K.; Pandey, H. M.; and Akbar, S. A. 2021. A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowledge-Based Systems*, 223: 106970.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10): 1499–1503.

Zhang, Z. 2012. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2): 4–10.

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4490–4499.