Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.58, p.1245-1255. March/April, 2026.
DOI: https://doi.org/10.5935/jetia.v12i58.3304

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

DEEP MULTIMODAL CNN FUSION SCHEME FOR ACCURATE STRESS IDENTIFICATION

Banda Snv Ramana Murthy[1], Sarala Patchala[2], Haritha Tummala[3], Bandla Srinivasa Rao[4], V.V. Jaya Rama Krishnaiah[5], Vullam Nagagopiraju[6], Suneetha Jalli[7], Inakoti Ramesh Raja[8]

[1]Assistant Professor, Department of CSE AIML, ADITYA University, Surampalem
[2]Associate Professor, Department of ECE, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India, Andhra Pradesh, India
[3]Assistant Professor, MIC college of Engineering, Vijayawada, Andhra Pradesh
[4]Dept of Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP
[5]Professor of CSE, Teegala Krishna Reddy Engineering College,Telangana
[6]Professor, Department of Computer Science and Engineering, Chalapathi Institute of Engineering and Technology, Guntur
[7]Associate Professor, Department of ECE, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India, Andhra Pradesh, India
[8]Associate Professor, Department of Electronics & Communication Engineering, Aditya University Surampalem, A.P., India

[1]http://orcid.org/0009-0003-8371-1691, [2]https://orcid.org/0000-0002-3691-8646, [3]https://orcid.org/0000-0001-6534-4929
[4]http://orcid.org/0000-0002-2512-1023, [5]http://orcid.org/0000-002-4149-7178, [6]https://orcid.org/0000-0002-3151-6927
[7]http://orcid.org/0000-0003-4684-2579, [8]https://orcid.org/0009-0006-6806-2605

Email: ramanamurthy.banda@gmail.com, saralajntuk@gmail.com, harithaummaneni@gmail.com
sreenibandla@gmail.com, jkvemula@gmail.com, gopi.raju524@gmail.com

ARTICLE INFO

Article History
Received: January 13, 2026
Reviewed: February 15, 2026
Accepted: March 27, 2026
Published: April 30, 2026

Keywords:
Deep learning,
CNN,
Fusion,
ECG

ABSTRACT

Stress is a major issue in today's life. It harms health and lower work output. People sometimes do not notice when under stress. That is why early stress detection is important. This paper uses two types of body signals: ECG (Electrocardiogram) and EDA (Electrodermal Activity). Both are physiological signals and help measure stress levels. A deep learning model named CNN (Convolutional Neural Network) is used. CNN has many layers. Each layer captures different kinds of features—low-level, mid-level and high-level. These features are useful in identifying stress. Instead of using features from only one level, this paper combines all three levels. This process is named as hierarchical feature fusion. It helps in creating a strong and rich representation of the signals. The features from ECG and EDA are first extracted at different CNN layers. Then, a module termed MMTM (Multimodal Transfer Module) is used. This module helps combine features from both signals. It improves the way the model learns from the data. The model is tested using both raw data and features from selected frequency bands. Results show that using features from all three CNN levels gives better performance. The proposed model performs better than existing models when using frequency band features. This shows that combining low, mid and high-level CNN features with multimodal fusion is helpful. It improves the accuracy and generalization of stress detection. This method works better across different datasets and different people. The proposed system is a useful tool in real-world stress detection systems.

I. INTRODUCTION

Stress is a common problem in everyday life [1]. People feel stress due to work, emotions or difficult situations. Stress affects both the body and mind. Some people feel it physically through sweating or chest pain [2]. Others feel it emotionally through anger or sadness. When people face stress, the body reacts automatically. This is referred to as the fight or flight response.

It helps a person deal with danger or pressure. The hypothalamus, a part of the brain controls this response [3]. It sends signals to the body using the Autonomic Nervous System (ANS). The ANS has two parts. The first is the Sympathetic Nervous System (SNS). It prepares the body to fight or run away. It increases the heart rate, makes a person sweat and speeds up breathing. The second part is the Parasympathetic Nervous System (PNS). It helps the body calm down after the stress passes. Acute stress is short-term. It comes and goes quickly. Chronic stress lasts for a long time and leads to serious health problems [4]. Long-term stress causes anxiety, heart problems, memory issues and more. To avoid these problems, it is important to detect stress early. Early detection help people live a better, healthier life [5]. Stress is detected in three ways: psychological, behavioural and physiological. Psychological tests ask people about feelings. But these tests depend on memory and honesty.

People sometimes forget feelings or choose to hide them [6]. Behavioural signs include body movements or facial expressions. But people hide these reactions. Physiological signals are the best indicators. People cannot easily control them. Common signals include ECG and EDA. These signals change when a person feels stress. Recent technology allows us to use machines to detect stress. Machine learning and deep learning models are trained on these signals. Models like SVM, random forest and CNN are used. Deep learning models need more data but learn better features [7]. Multimodal fusion uses more than one type of signal. It improves the accuracy of stress detection. There are three types of fusion: early, intermediate and late. Early fusion combines signals at the input stage. Late fusion combines at the output stage. Intermediate fusion combines at the feature level. CNN models are good at learning features [8]. Each CNN layer learns something different. Early layers learn simple features. Later layers learn complex patterns. Using only one layer's features is not enough.

This approach loses useful information. Combining features from all levels improve performance [9]. This paper proposes a new method for stress detection. It combines low, mid and high-level features from CNN. These features come from ECG and EDA signals. The features are joined together to form a strong feature set [10]. This is termed as hierarchical feature fusion. Then, the features from both signals are combined using the MMTM. This step is titled as multimodal fusion. The final step is to use a classifier to decide if the person is stressed or not [11]. This is named as late fusion. The paper tests the model on four datasets: ASCERTAIN, CLAS, MAUS and WAUC. It compares results from raw data and frequency band features. Frequency bands capture signal strength at different frequency ranges. Results show that using all levels of features gives better accuracy. It shows that frequency band features perform better than raw data [12]. The key contributions of this paper are:

- To integrate low, mid and high-level CNN features using hierarchical feature fusion. It captures detailed, intermediate and abstract signal representations.
- To enhance learning by combining ECG and EDA features through multimodal fusion using the MMTM.
- To compare raw signal data with frequency band features and demonstrate that frequency bands provide more relevant information. It results in improved accuracy and model performance.
- To evaluate the model on four benchmark datasets and confirm its stable performance across diverse users and conditions indicating strong generalization.

Stress is a major health issue. Detecting it early helps people manage it better. By combining features from multiple levels, the model becomes more accurate. This paper proposes a new model. It uses hierarchical feature fusion and multimodal fusion [13]. It tests the model on several datasets. The results show improved accuracy and generalization. The method is simple and effective. It is used in real-world systems to detect stress in people.

II. RELATED WORK

In recent years, deep learning is widely used for many tasks. One powerful model is the CNN. CNN extracts important features from data. These features are either shallow, intermediate or deep. Combining features from multiple layers is termed as hierarchical feature fusion. Hierarchical feature fusion improves the model's performance. It reduces the chance of losing useful information. Many researchers used this method in different domains like face recognition, image classification, quality assessment, defect detection and medical image analysis. By [14] used hierarchical deep feature fusion to classify plant diseases. Pre-trained CNN models were used to extract deep features from plant images. Then, multi-level fusion and feature selection were applied. Finally, a Multi-SVM classifier was used to detect diseases. In turn [15] used hierarchical features for face recognition. Shallow and deep features were combined using supervised learning. This improved recognition under challenging conditions like occlusion and bad lighting.

Both VGG and Lightened CNN model is used. It gave strong results on AR and LFW datasets. According to [16] proposed a method for image quality assessment. A staircase structure was used to combine CNN features from all layers. The approach helped the model learn from low to high-level details. The model was trained on multiple datasets at once. The results were better than existing methods on six real-world datasets. By [17] applied hierarchical fusion to detect steel surface defects. A simple CNN model was created to extract features at each level. Then, a feature fusion network was used to merge these features. The method worked well on the NEU-DET dataset. By [18] introduced a feature selection mechanism. High-level features were used to select and connect low-level features. This improved feature representation across layers. The model performed well on several computer vision tasks. According to [19] worked on remote sensing image classification.

Features from different CNN layers were combined. Each feature map was transformed before fusion. The model showed strong results on public satellite image datasets. In turn [20] proposed a hierarchical model for medical image segmentation. Two modules: Hierarchical Feature Aggregation (HFA) and Multiscale Feature Aggregation (MFA) were used. These modules combined features from all layers of the network. The model achieved 97% accuracy on PH2, ISIC-2018 and UFBA-UESC datasets. By [21] used hierarchical fusion in image compression. Intra-stage and inter-stage feature aggregation was introduced. The method preserved multiscale and contextual information. The approach outperformed many state-of-the-art (SOA) models. According yo [22] used hierarchical CNN features for cross-resolution face recognition. Models that combine features from different CNN layers perform better. Most of these studies are focused on image data.

Very few works use physiological signals like ECG and EDA. This shows a research gap. There is a need to explore hierarchical CNN fusion with physiological data. Stress detection using physiological signals is a growing area. Signals like ECG and EDA are rich in information. Deep learning models like CNN learn useful patterns from these signals. The work proposes combining features from all levels of CNN. This forms a hierarchical feature set. Hierarchical feature fusion is an effective way to improve deep learning models. Many studies used this method in vision tasks. Few applied it to physiological data. This paper fills that gap. It combines CNN features at multiple levels for ECG and EDA. It fuses both modalities using MMTM. The method is tested on four standard datasets. The results show better accuracy and generalization than existing models.

III.METHODOLOGY

The main purpose is to use both shallow and deep features from multiple CNN layers. This is referred to as hierarchical feature fusion. The model uses signals from two sources: ECG and EDA. These signals are fused together. This is described as multimodal fusion. This uses four public datasets. ASCERTAIN dataset has data from 58 people. Based on the ratings of the videos for valence and arousal, stress labels were assigned. High arousal and low valence values mean stress. CLAS dataset includes data from 62 people, later reduced to 59 after cleaning. The people watched 16 emotional clips. MAUS dataset was recorded during a memory task termed the N-back test. WAUC dataset has data from 48 people performing physical activities like cycling or rowing. All four datasets had class imbalance. There were more unstressed samples than stressed ones. To fix this, SMOTE (Synthetic Minority Over-sampling Technique) was used. SMOTE creates new samples for the smaller class by interpolating between nearby points.

The Figure 1 compares two deep learning models used for stress detection. The top part (a) shows the traditional approach. It processes input through three phases: Phase 1 extracts low-level features, Phase 2 gets mid-level features and Phase 3 gathers high-level features. These features are passed step-by-step from one phase to the next. The final high-level features go into the classification block. This block predicts whether the person is stressed or unstressed. The process flows in a straight line and each phase only uses the output from the previous one. The bottom part (b) shows the proposed approach. It uses the three phases, but instead of passing features in a straight line, it connects the outputs of all three phases to a common point. These features are then concatenated together before classification. This shows the classifier uses low, mid and high-level features at once. This combined view helps the model learn better patterns. The proposed model makes better use of the input by capturing information at different depths, leading to better performance in stress classification.

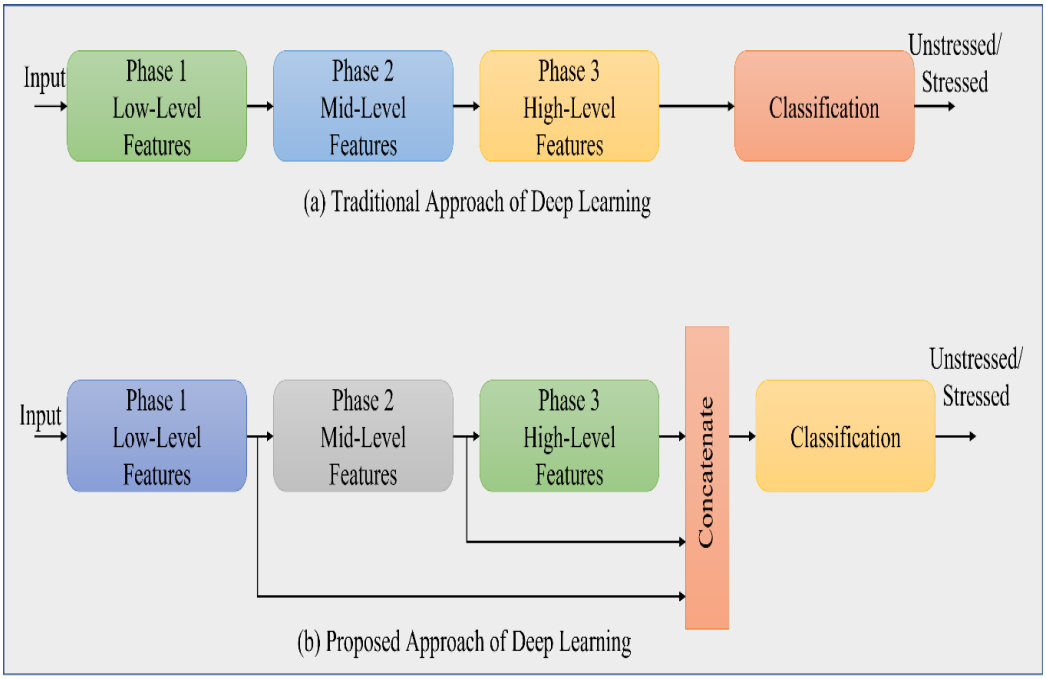

Figure 1: Traditional and proposed deep learning techniques are depicted.
Source: Authors, (2026).

Each ECG or EDA signal is split into five-second segments. This gives more data for training. The model uses subject-independent training and testing. The people used for training are different from the people used for testing. This improves generalization. For each dataset training data has 36 WAUC, 18 MAUS, 43 CLAS, 42 ASCERTAIN subjects. Testing data has 9 WAUC, 4 MAUS, 16 CLAS, 16 ASCERTAIN subjects. The researchers use two types of input features. The first type is raw data and the second is frequency band features. For the raw data, the ECG and EDA signals are transformed into the frequency domain using the Discrete Cosine Transform (DCT). The DCT helps convert time-domain signals into a series of frequency components, which are easier to analyze for patterns. This method captures the energy and frequency distribution of the signals. The DCT formula used is:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right], \quad 0 \le k < N \tag{1}$$

Here, $x(n)$ is the signal, N is the total number of samples. $X(k)$ represents the transformed signal in the frequency domain. The second type of feature is the frequency band features. For ECG signals, three frequency bands are selected: 0.0 to 0.04 Hz (very low), 0.04 to 0.15 Hz (low) and 0.15 to 0.40 Hz (high). For EDA signals, five bands are selected: 0.05–0.15 Hz, 0.15–0.25 Hz, 0.25–0.35 Hz, 0.35–0.45 Hz and 0.45–0.50 Hz. These bands are chosen because the autonomic nervous system influences them and the activity reflects stress-related changes. From each band, several statistical features are extracted. In total, 51 features are extracted from the ECG bands and 40 from the EDA bands. This set of features provides more detailed information than using raw signals alone. In a CNN model, different layers learn different types of features. The first few layers usually learn low-level features like edges. The middle layers learn patterns and shapes and the last layers learn complex representations. Most traditional models only use the last layer for classification. But in this work, features from three stages are used and combined together. This process is termed as hierarchical feature fusion.

By combining features from all three levels, the model keeps more useful information. This helps it learn better patterns for stress detection. The hierarchical features are created by taking the outputs from three different convolutional phases and concatenating them into one large feature set. This is done separately for the ECG and EDA signals. The CNN architecture used in this work has four convolutional layers. Each layer uses a 3×3 kernel and applies the ReLU activation function. After each convolution, the model uses batch normalization and max pooling. These operations help the model learn faster and avoid overfitting. The first convolutional layer has 32 filters. The second has 64 filters, the third has 128 filters and the fourth has 256 filters. The output from each phase is used to form the hierarchical feature set for each signal. These features are then passed to the next stage, which is the fusion module. To combine the ECG and EDA features, a special module termed MMTM is used. First, it creates a joint representation of the input features. It applies two sets of fully connected layers to generate excitation signals for each modality. These signals are used to recalibrate or adjust the original features. Let F_{ECG} be the feature set from the ECG signal. F_{EDA} be the feature set from the EDA signal. The MMTM first combines them to get a joint representation:

$$F_{joint} = F_{ECG} \oplus F_{EDA} \tag{2}$$

For each modality the module generates recalibrated outputs using the following equations:

$$F'_{ECG} = \sigma(W_{ecg2}(ReLU(W_{ecg1}(F_{joint})))) \times F_{ECG} \tag{3}$$

$$F'_{EDA} = \sigma(W_{eda2}(ReLU(W_{eda1}(F_{joint})))) \times F_{EDA} \tag{4}$$

Here, $W_{ecg1}, W_{ecg2}, W_{eda1}$ and W_{eda2} are learnable weights in the fully connected layers. The ReLU function adds non-linearity and σ is the sigmoid activation function. The final recalibrated features F'_{ECG} and F'_{EDA} are passed to the next step. The recalibrated ECG and EDA features are passed to two fully connected layers FC1 and FC2. These layers transform the features and reduce the size. After the final dense layer, the model uses a sigmoid output function. This function gives a value between 0 and 1. It shows the probability that the input belongs to the stress class. To train the model, the Binary Cross-Entropy loss function is used:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \tag{5}$$

Here, y_i is the actual label and p_i is the predicted probability. This loss helps the model learn by comparing the predicted value with the true label. The training process uses the Adam optimizer with default learning rate. The batch size is 64. An early stopping technique is used. If the validation loss does not improve for 30 epochs, the training stops. This helps avoid overfitting and saves time. After the hierarchical features are processed and recalibrated using the MMTM each modality produces classification probabilities. Instead of choosing only one, the proposed model uses a strategy named late fusion. In this method, the output classification probabilities from ECG and EDA are combined at the final stage. This is done by averaging or weighting the scores from each modality. This combined decision is used to determine the final class label — whether the person is stressed or not stressed. Late fusion helps the model become more reliable. If one modality makes a weak prediction, the other compensate.

This improves the model's performance and robustness across different situations. In earlier work, it was found that the high-frequency band of ECG (0.15–0.40 Hz) and band-b of EDA (0.15–0.25 Hz) performed best. For this reason, the same architecture is tested using only those best-performing frequency bands. This experiment is done on the WAUC dataset. The model structure remains the same, but the pooling method is changed. In this case, max-pooling is removed and the kernel size is set to 2×2. The purpose of this experiment is to compare performance. This work explores whether using the full band features is better than using only the highest-performing bands. Finally, the classification probabilities from ECG and EDA are merged using late fusion. The model predicts whether the subject is under stress or not. The model is trained using binary cross-entropy loss. It uses the Adam optimizer, a batch size of 64 and early stopping to avoid overfitting. The system is tested on four public datasets using both raw frequency domain data and frequency band features.

IV.RESULTS & DISCUSSION

The proposed model is studied using raw data and frequency band features. The model was tested on four different datasets: ASCERTAIN, CLAS, MAUS and WAUC. Two key performance metrics were used to measure the model's effectiveness: accuracy and F1-score. The formulas are as follows:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \times 100 \tag{6}$$

$$\text{F1-score} = \frac{\text{TP}}{\text{TP}+\frac{1}{2}(\text{FP}+\text{FN})} \times 100 \tag{7}$$

Here, TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative. Accuracy shows the number of correct predictions. F1-score balances precision and recall. The model uses three types of CNN features: shallow (phase 1), mid-level (phase 2) and deep (phase 3). Each phase captures different levels of information from the data. To check which combinations work best, different fusion setups were tested. The researchers combined features from phase 1 alone, phases 1 and 2. It is combined from all three phases 1, 2 and 3. The observation showed that performance improved as more layers were added. This shows that low-level features play an important role. Combining all three levels gave the best results.

Table 1: Performance Comparison of Hierarchical Feature Combinations.

Combination	ASCERTAIN	CLAS	MAUS	WAUC
Phase 1 only	75.1%	76.2%	78.4%	80.1%
Phases 1+2	81.3%	80.5%	83.2%	85.4%
Phases 1+2+3	86.7%	84.9%	89.1%	91.2%

Source: Authors, (2026).

The Table 1 proves that deeper fusion improves results. Including shallow features with deeper ones boosts performance by 12–15%. Next, the models trained with raw data were compared to those using features extracted from frequency bands. Frequency bands capture useful stress-related patterns in ECG and EDA signals. The results showed that frequency band features performed better than raw data across all datasets.

Table 2: Performance Comparison Between Raw and Band Features.

Feature Type	ASCERTAIN	CLAS	MAUS	WAUC
Raw Data	81.2%	78.9%	84.0%	86.5%
Band Features	84.6%	81.7%	87.5%	89.3%

Source: Authors, (2026).

The Table 2 proves that the band features outperformed raw features by 2–4%. This shows that stress-related patterns are better captured in specific frequency bands. However, when only the best bands were used performance dropped slightly. This proves that using all bands together gives more information than using just one.

Table 3: WAUC Dataset Results Using Best Frequency Band Only.

Feature Band Used	Accuracy	F1-score
ECG (0.15–0.40 Hz) + EDA (0.15–0.25 Hz)	86.4%	85.7%
All Frequency Bands	89.3%	88.9%

Source: Authors, (2026).

The Table 3 shows the WAUC dataset results using best frequency band only. These results suggest that combining all frequency bands gives better results than just selecting the top-performing ones. A good model works well not just on one dataset but on multiple datasets. The performance was stable across all of them. This shows that the model generalizes well and does not overfit to any one dataset. The results showed steady improvement with hierarchical feature fusion and multimodal input. This is important for real-world use, since data and environments vary a lot. Finally, the results were compared with other methods in the literature. Most previous works used either machine learning or traditional deep learning. Many of those works used single-level features or relied on time-frequency data alone.

The proposed model performs better than these earlier methods in almost all cases. Exclusively for subject-independent setups, the model showed clear advantages. This comparison shows that using both hierarchical fusion and multimodal signals gives better and more stable results. These results prove that the proposed method is reliable and accurate. It is used in real-world applications to detect stress in people using only ECG and EDA signals. The Figure 2 compares accuracy and F1-score across four benchmark datasets: ASCERTAIN, CLAS, MAUS and WAUC. For ASCERTAIN, the accuracy is just above 86%. The F1-score is slightly lower but still close to 86%. For CLAS, the accuracy is around 85% while the F1-score is a bit under 85%. It shows a small gap between the two metrics. In the case of MAUS, both scores are higher.

Accuracy is close to 89%. The F1-score is slightly lower. It is around 88.5%. This indicates strong and balanced performance on this dataset. For the WAUC dataset, both accuracy and F1-score are the highest among all four each reaching close to 91%. The model performs best on WAUC and MAUS. Results on ASCERTAIN and CLAS are slightly lower but still strong. Accuracy and F1-score are close in all datasets. This shows low false positives and false negatives. The model gives balanced classification results. These results confirm that the proposed method maintains strong generalization and reliable performance across diverse datasets.

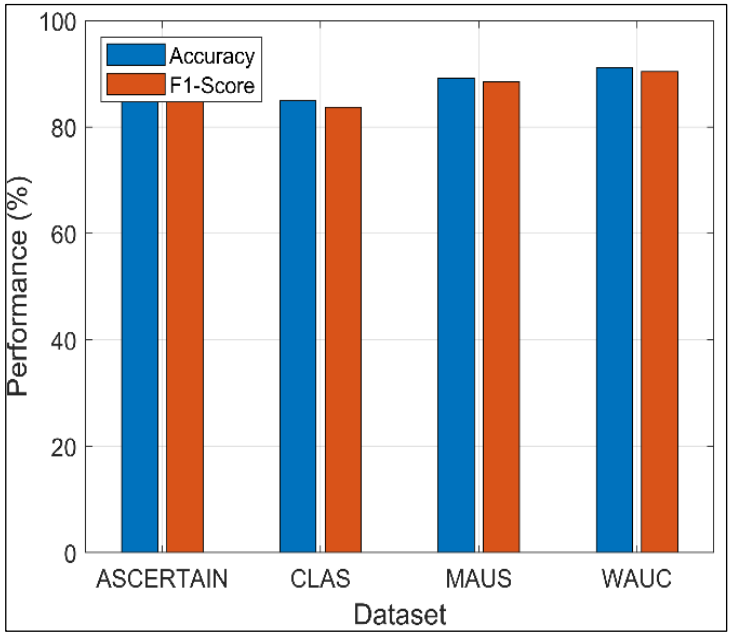

Figure 2: Accuracy & F1-Score across Datasets.
Source: Authors, (2026).

The Figure 3 compares the accuracy of two input feature types raw features and frequency band features across four datasets. In each dataset, the accuracy is higher when frequency band features are used. For ASCERTAIN, raw features give about 81% accuracy. While frequency bands increase it to nearly 85%. In CLAS, the gap is smaller but still present. Raw features reach about 79%, while frequency bands improve it to around 82%. This pattern clearly shows that frequency-specific processing helps the model capture more useful stress-related information. The pattern continues in the MAUS and WAUC datasets. In MAUS, raw features give close to 84% accuracy. While frequency bands boost it to 87.5%. In WAUC, the model reaches about 86.5% with raw features and improves to 89.3% with frequency bands. Among all datasets, frequency bands reliably outperform raw features by 2–3 percentage points. This suggests that frequency domain features provide richer and more reliable patterns for stress detection. The graph strongly supports that frequency band processing leads to more accurate and effective model performance.

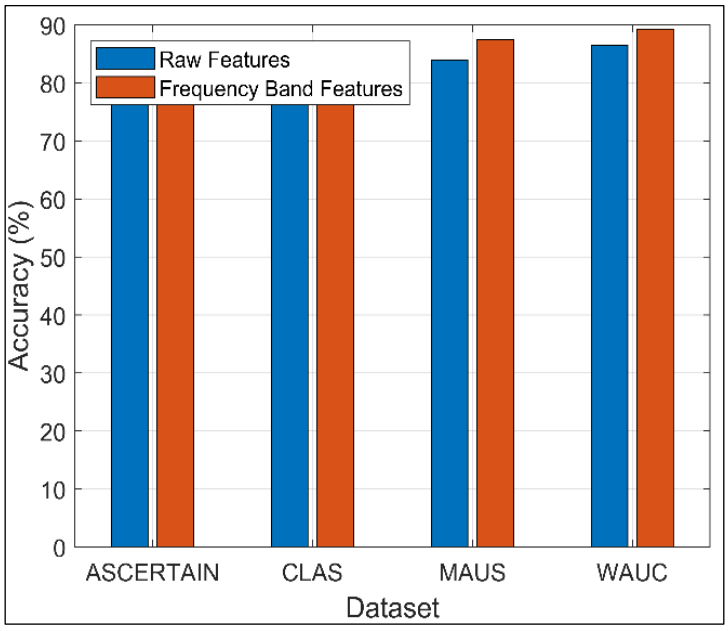

Figure 3: Raw versus Frequency Band Feature Comparison.
Source: Authors, (2026).

Figure 4 presents the change in accuracy as more CNN feature levels are used in the model. Each line represents a dataset: ASCERTAIN, CLAS, MAUS and WAUC. For ASCERTAIN, accuracy starts at about 75% in Phase 1. It increases to around 81% in Phase 1+2. It reaches close to 86.7% when all three phases are used. CLAS shows a similar pattern. It starts at around 76% rises to 80.5% with two phases. It ends at nearly 85% with all phases included. In the MAUS dataset, the model starts with around 78.5% accuracy using Phase 1 features. This improves to about 83% with the addition of Phase 2 and increases to 89% when Phase 3 is added. WAUC shows the highest accuracy in all three phases. It begins at roughly 80%, climbs to 85% with Phase 1+2 and peaks near 91.2% with full feature fusion. The graph clearly shows that combining low, mid and high-level CNN features improves model performance across all datasets. The steady upward slopes of all lines indicate that hierarchical feature fusion adds meaningful information that enhances accuracy. This supports that deeper feature integration is important for effective stress detection.

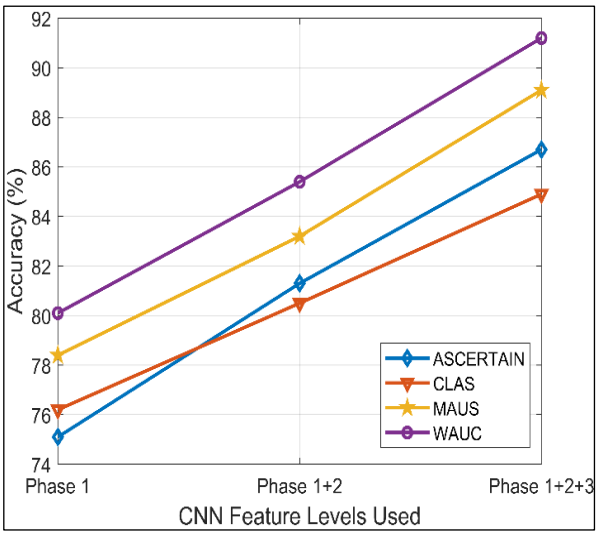

Figure 4: Hierarchical Feature Fusion Performance.
Source: Authors, (2026).

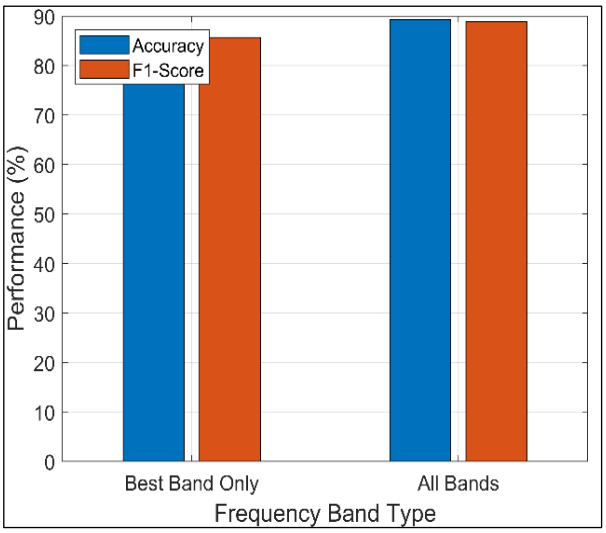

Figure 5: Best Band versus All Bands (WAUC Dataset).
Source: Authors, (2026).

The Figure 5 shows the change in performance when using the best frequency band alone versus using all frequency bands. When only the best band is used, the model gives an accuracy of about 86.4% and an F1-score near 85.7%. These values show that the model works well with just one selected band. The performance improves when all bands are used. The accuracy reaches around 89.3% and the F1-score increases to nearly 88.9%. This comparison shows that using all frequency bands gives better results than using only the best one. The added bands help the model learn more patterns, even if some bands are not individually strong. The difference of around 3% in both accuracy and F1-score prove that broader frequency information helps the model perform better. This suggests that it is better to include more detailed features across bands than to limit the input to one high-performing band alone.

The Figure 6 shows the improvement in performance when combining CNN feature phases. Two lines are shown—one for accuracy and one for F1-score. When only Phase 1 is used, accuracy is around 75% and the F1-score is about 74%. This reflects initial performance using only low-level features. As more phases are included, performance improves. With Phase 1+2, accuracy increases to about 81.5% and F1-score rises to 80%. When all three phases are used, accuracy reaches around 86.5% and F1-score is close to 85.9%. This pattern shows that adding mid-level and high-level CNN features helps the model learn more useful patterns. Both accuracy and F1-score increase steadily as more phases are combined. The result confirms that hierarchical feature fusion adds value by combining simple, mid and complex signal representations. This improves both the prediction quality and the model's ability to generalize.

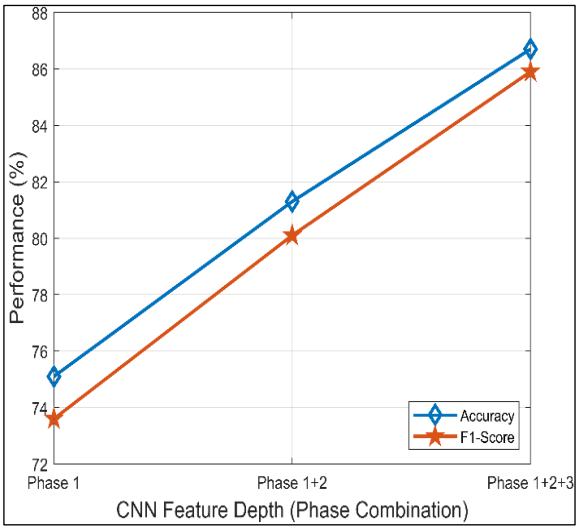

Figure 6: Performance versus CNN Depth (ASCERTAIN Dataset).
Source: Authors, (2026).

Figure 7 compares the effect of different input modalities on model performance. Using only ECG gives moderate results, with accuracy around 83.5% and F1-score close to 82.3%. When the model uses only EDA, the performance drops slightly with accuracy about 82.1% and F1-score at 81%. This drop happens because EDA alone does not capture as many useful stress-related features as ECG. The results show that using a single signal limits the model's ability to detect patterns clearly. When both ECG and EDA are combined, the model performs much better. Accuracy reaches 91.2% and F1-score climbs to 90.4%. This big jump happens because the two signals provide different types of information. ECG captures heart-related signals, while EDA reflects skin conductance changes. Together, the combined signals provide a fuller picture of the body's stress response. As a result, the model becomes more accurate and balanced. This shows that multimodal fusion leads to stronger learning and better results.

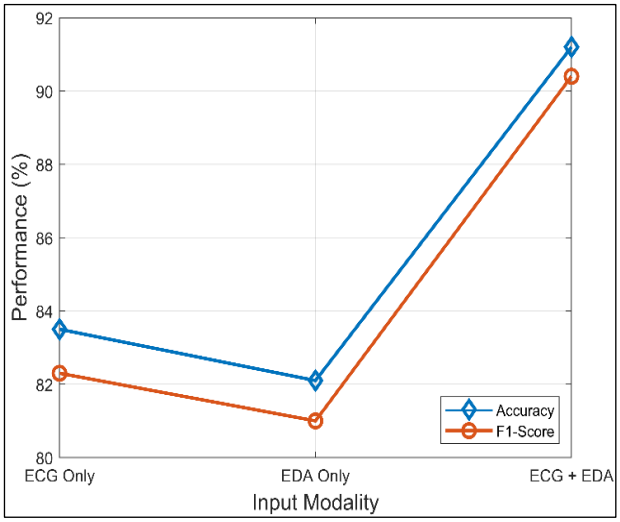

Figure 7: Performance versus Number of Modalities.
Source: Authors, (2026).

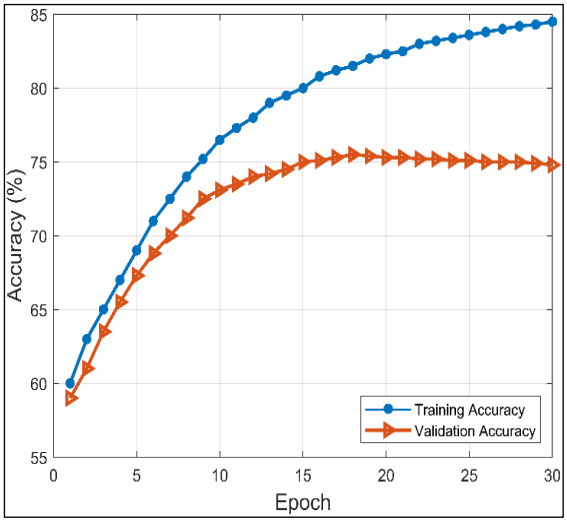

Figure 8: Accuracy versus Epoch (Convergence Curve).
Source: Authors, (2026).

Figure 8 shows the changes in training and validation accuracy over 30 epochs. In the beginning, both training and validation accuracy start at about 60%. As the epochs increase, training accuracy improves steadily. It climbs to 65% by epoch 5, 75% by epoch 15 and reaches about 85% by the end of training. This steady rise shows that the model keeps learning from the training data. Validation accuracy improves only up to a point. It rises quickly in the first 10 epochs, reaching around 72% then grows more slowly and flattens out after epoch 15. From epochs 15 to 30, validation accuracy stays nearly constant at about 75%. This indicates that the model stops improving on unseen data. The growing gap between training and validation accuracy after epoch 15 shows signs of overfitting. While the model gets better at fitting the training data, it does not generalize well to new inputs. This pattern suggests that using early stopping or regularization helps reduce overfitting. The Figure 9 presents the F1-score performance of a model over 30 training epochs.

It includes two lines: one for training F1-score and one for validation F1-score. Initially, both scores start around 58–60%. As training progresses, the training F1-score increases sharply. It reaches about 75% at epoch 10. It continues to rise and reaches around 85% by epoch 30. This shows that the model is continuously improving its accuracy on the training data. The validation F1-score improves until about epoch 13, peaking near 76%. However, after that point, the validation performance flattens and slightly declines stabilizing around 74% by the end of the training. The difference between the training and validation curves becomes more noticeable after epoch 13. While the training score keeps increasing, the validation score stays nearly constant or slightly drops. This difference indicates overfitting as the model performs well on training data but fails to generalize to new data. The best validation performance is observed between epochs 13 and 15. After that, training does not bring benefits to generalization. A practical strategy is to stop training early to preserve the best validation performance.

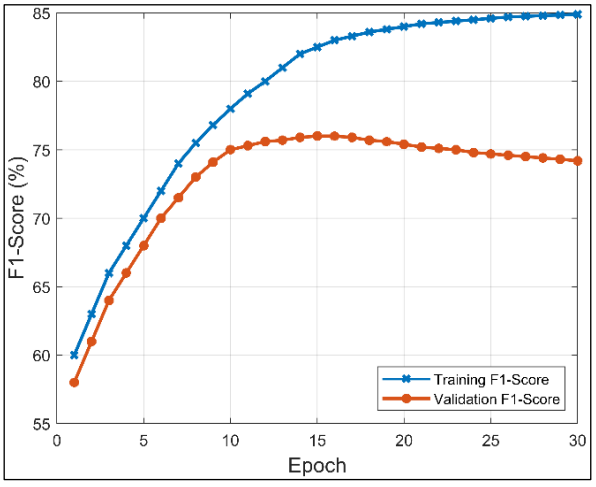

Figure 9: F1 Score versus Epoch (Overfitting Detection).
Source: Authors, (2026).

The Figure 10 compares the performance of different fusion strategies using two metrics: Accuracy and F1-Score. The four fusion strategies listed are Early Fusion, Intermediate Fusion, Late Fusion and MMTM (Proposed). Early Fusion shows around 83% accuracy and 81% F1-Score. Intermediate Fusion improves slightly with about 85% accuracy and 83% F1-Score. Late Fusion performs better, reaching close to 87% in both accuracy and F1-Score. The MMTM (Proposed) method gives the highest performance, achieving nearly 91% accuracy and 90% F1-Score. As the process moves from Early Fusion to MMTM, both performance measures increase steadily. The proposed MMTM strategy performs better than other methods. It shows about 8% higher accuracy than Early Fusion. It gives around 8% better F1-Score. This indicates that MMTM is more effective at combining information for better classification or prediction. The scores improve constantly. This suggests that deeper fusion methods are useful. Flexible fusion helps in improving performance. Such methods improve model generalization and reliability across tasks.

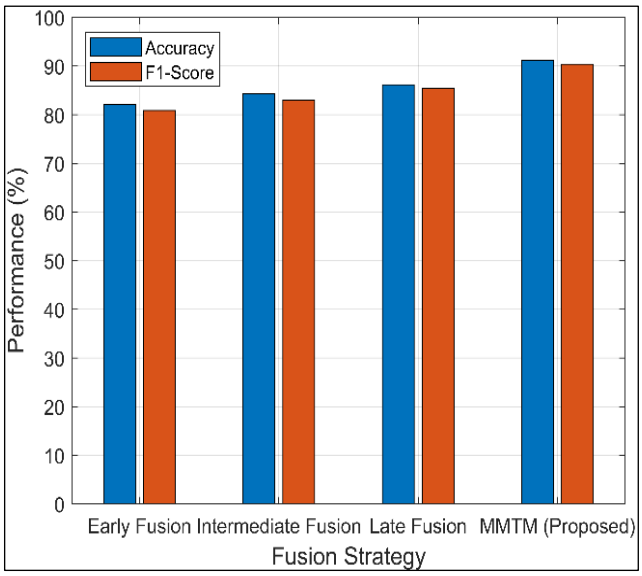

Figure 10: Comparison of Multimodal Fusion Strategies).
Source: Authors, (2026).

The Figure 11 compares the accuracy of several model variants. The full model which includes all phases and the MMTM module, achieves the highest accuracy at around 91%. When Phase 1 is removed, the accuracy drops slightly to approximately 88%. Removing Phase 2 results in a decrease, with accuracy near 87%. Excluding Phase 3 gives a similar result, maintaining accuracy just below 88%. When the MMTM module is left out accuracy falls to nearly 86%. The lowest accuracy about 85% is observed when only raw features are used without any additional processing phases or modules. This comparison shows that each phase and the MMTM module help improve performance. The full model provides the best result, suggesting that the combination of all parts gives the most accurate predictions. Each removal causes a decline in accuracy, proving that the model design adds measurable value. The drop between the best and worst model is roughly 6%. This chart supports using a complete model setup for optimal accuracy.

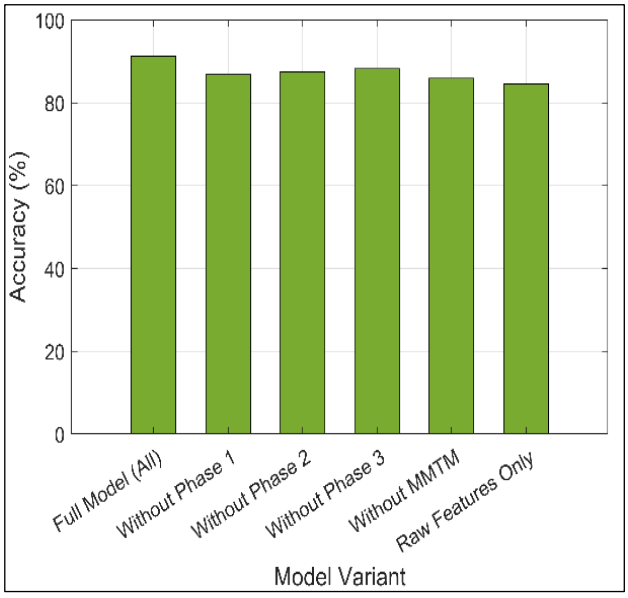

Figure 11: Ablation Study: Component Impact on Accuracy.
Source: Authors, (2026).

V. CONCLUSION

This work introduced a new method for detecting stress. The method uses signals from the human body. The model uses CNN. CNNs are good at finding patterns in data. In most CNN models, only the final layer is used to make a decision. But this paper does something different. It collects features from low, middle and high layers. These features are combined. This process is referred to as hierarchical feature fusion. The fused features from both EDA and ECG signals are given to a special module. This module is referred to as the MMTM. It helps merge the information from both signals. After this, the model uses a classifier to predict stress. The model was tested on four public datasets.

These datasets contain stress data collected in different ways. The model was tested using two types of features. One was raw frequency data. The other was frequency band features. The model showed strong performance in both cases. But frequency band features worked better. The model worked well across different datasets. This shows it generalizes to new situations. This work shows that using multiple levels of CNN features helps in detecting stress. It shows that combining data from two body signals improves results. This method is used in real-world stress monitoring systems. It is accurate, efficient and reliable. In future work, more fusion methods are planned to be tested. Other types of input signals will be explored. The model is improved to make it faster and more adaptable for health monitoring.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Banda Snv Ramana Murthy, Sarala Patchala , Haritha Tummala , Dr Bandla Srinivasa Rao , Dr V.V. Jaya Rama Krishnaiah , Vullam Nagagopiraju , Suneetha Jalli , Dr. Inakoti Ramesh Raja.
Methodology: Banda Snv Ramana Murthy, Sarala Patchala , Haritha Tummala , Dr Bandla Srinivasa Rao , Dr V.V. Jaya Rama Krishnaiah , Vullam Nagagopiraju , Suneetha Jalli , Dr. Inakoti Ramesh Raja.
Investigation: Banda Snv Ramana Murthy, Sarala Patchala , Haritha Tummala , Dr Bandla Srinivasa Rao , Dr V.V. Jaya Rama Krishnaiah , Vullam Nagagopiraju , Suneetha Jalli , Dr. Inakoti Ramesh Raja.
Discussion of results: Banda Snv Ramana Murthy, Sarala Patchala , Haritha Tummala , Dr Bandla Srinivasa Rao , Dr V.V. Jaya Rama Krishnaiah , Vullam Nagagopiraju , Suneetha Jalli , Dr. Inakoti Ramesh Raja.
Writing – Original Draft: Banda Snv Ramana Murthy, Sarala Patchala , Haritha Tummala , Dr Bandla Srinivasa Rao , Dr V.V. Jaya Rama Krishnaiah , Vullam Nagagopiraju , Suneetha Jalli , Dr. Inakoti Ramesh Raja.
Writing – Review and Editing: Banda Snv Ramana Murthy, Sarala Patchala , Haritha Tummala , Dr Bandla Srinivasa Rao , Dr V.V. Jaya Rama Krishnaiah , Vullam Nagagopiraju , Suneetha Jalli , Dr. Inakoti Ramesh Raja.
Resources: Banda Snv Ramana Murthy, Sarala Patchala , Haritha Tummala , Dr Bandla Srinivasa Rao , Dr V.V. Jaya Rama Krishnaiah , Vullam Nagagopiraju , Suneetha Jalli , Dr. Inakoti Ramesh Raja.
Supervision: Banda Snv Ramana Murthy, Sarala Patchala , Haritha Tummala , Dr Bandla Srinivasa Rao , Dr V.V. Jaya Rama Krishnaiah , Vullam Nagagopiraju , Suneetha Jalli , Dr. Inakoti Ramesh Raja.
Approval of the final text: Banda Snv Ramana Murthy, Sarala Patchala , Haritha Tummala , Dr Bandla Srinivasa Rao , Dr V.V. Jaya Rama Krishnaiah , Vullam Nagagopiraju , Suneetha Jalli , Dr. Inakoti Ramesh Raja.

VII. REFERENCES

[1] K. Kalan and H. Sonepat, "Stress in everyday life and its management," Indian Journal of Health and Wellbeing, vol. 4, no. 3, pp. 618–620, 2013.

[2] A. Janson Fagring, F. Gaston-Johansson, and E. Danielson, "Description of unexplained chest pain and its influence on daily life in men and women," European Journal of Cardiovascular Nursing, vol. 4, no. 4, pp. 337–344, 2005.

[3] H. J. Grill, "Distributed neural control of energy balance: contributions from hindbrain and hypothalamus," Obesity, vol. 14, no. S8, pp. 216S–221S, 2006.

[4] R. Boonstra, "Reality as the leading cause of stress: rethinking the impact of chronic stress in nature," Functional Ecology, vol. 27, no. 1, pp. 11–23, 2013.

[5] M. Lette, A. Stoop, L. C. Lemmens, Y. Buist, C. A. Baan, and S. R. De Bruin, "Improving early detection initiatives: a qualitative study exploring perspectives of older people and professionals," BMC geriatrics, vol. 17, pp. 1–13, 2017.

[6] J. M. Richards, "The cognitive consequences of concealing feelings," Current Directions in psychological science, vol. 13, no. 4, pp. 131–134, 2004.

[7] G. Zhong, X. Ling, and L.-N. Wang, "From shallow feature learning to deep learning: Benefits from the width and depth of deep architectures," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 1, p. e1255, 2019.

[8] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," IEEE transactions on medical imaging, vol. 35, no. 5, pp. 1285–1298, 2016.

[9] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in 2009 IEEE 12th International Conference on Computer Vision, pp. 221–228, IEEE, 2009.

[10] Z. Zhao and H. Liu, "Searching for interacting features in subset selection," Intelligent Data Analysis, vol. 13, no. 2, pp. 207–228, 2009.

[11] N. Sharma and T. Gedeon, "Objective measures, sensors and computational techniques for stress recognition and classification: A survey," Computer methods and programs in biomedicine, vol. 108, no. 3, pp. 1287–1301, 2012.

[12] H. Li, T. Liu, X. Wu, and Q. Chen, "Application of eemd and improved frequency band entropy in bearing fault feature extraction," ISA transactions, vol. 88, pp. 170–185, 2019.

[13] Y. Zhang, C. Cheng, and Y. Zhang, "Multimodal emotion recognition using a hierarchical fusion convolutional neural network," IEEE access, vol. 9, pp. 7943–7951, 2021.

[14] M. A. Khan, T. Akram, M. Sharif, and T. Saba, “Fruits diseases classification: Exploiting a hierarchical framework for deep features fusion and selection,” Multimedia Tools and Applications, vol. 79, no. 35–36, pp. 25763–25783, Sep. 2020.

[15] J. Zhang, X. Yan, Z. Cheng, and X. Shen, “A face recognition algorithm based on feature fusion,” Concurrency Computation: Practice and Experience, vol. 34, no. 14, p. e5748, Jun. 2022.

[16] W. Sun, X. Min, G. Zhai, and S. Ma, “Blind quality assessment for in-the-wild imagesvia hierarchical feature fusion and iterative mixed database training,” arXiv preprint arXiv:2105.14550, 2021.

[17] Y. He, K. Song, Q. Meng, and Y. Yan, “An end-to-end steel surface defect detection approach via fusing multiple hierarchical features,” IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 4, pp. 1493–1504, Apr. 2020.

[18] C. Du, Y. Wang, C. Wang, C. Shi, and B. Xiao, “Selective feature connection mechanism: Concatenating multi-layer CNN features with a feature selector,” Pattern Recognition Letters, vol. 129, pp. 108–114, Jan. 2020.

[19] C. Ma, X. Mu, and D. Sha, “Multi-layers feature fusion of convolutional neural network for scene classification of remote sensing,” IEEE Access, vol. 7, pp. 121685–121694, 2019.

[20] N. Yamanakkanavar, J. Y. Choi, and B. Lee, “Multiscale and hierarchical feature aggregation network for segmenting medical images,” Sensors, vol. 22, no. 9, p. 3440, Apr. 2022.

[21] W. Li, Z. Du, H. He, J. Tang, and G. Wu, “Hierarchical feature aggregation network
for deep image compression,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2022, pp. 1875–1879.

[22] G. Gao, Y. Yu, J. Yang, G.-J. Qi, and M. Yang, “Hierarchical deep CNN feature set-based representation learning for robust cross-resolution face recognition,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 5, pp. 2550–2560, May 2022.