




# Multimodal Approach Based on Autistic Child Behavior Analysis for Meltdown Crisis Detection

Marwa Masmoudi<sup>1</sup><sup>a</sup>, Salma Kammoun Jarraya<sup>1,2</sup><sup>b</sup> and Mohamed Hammami<sup>1,3</sup><sup>c</sup>

<sup>1</sup>Mir@cl Laboratory, University of Sfax, Tunisia

<sup>2</sup>CS Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

<sup>3</sup>Department of Computer Science, Faculty of Science, Sfax, Tunisia

**Keywords:** Multimodal Approach, Meltdown Crisis Detection, Autism, Behavior Analysis.

**Abstract:** This paper presents an innovative method for addressing the challenge of recognizing and responding to meltdown crises in autistic children. It focuses on integrating information from emotional and physical modalities, employing multimodal fusion with an emphasis on the early fusion technique. Existing literature outlines three fusion techniques – early, late, and hybrid fusion, each with unique advantages. Due to the distinct nature of datasets representing emotions and physical activities, late and hybrid fusion were considered impractical. Therefore, the paper adopts the early fusion method and introduces a Multi-modal CNN model architecture for efficient meltdown crisis recognition. The architecture comprises three Convolution layers, Max-pooling Layers, a Fully Connected (FC) layer, and Softmax activation for classification. The decision to opt for early fusion is driven by the inconsistent detection of children’s faces in all video frames, resulting in two different output sizes for emotion and physical activity systems. The presented pseudo-code outlines the architecture development steps. The proposed model’s efficiency is highlighted by its outstanding recognition rate and speed, making it the preferred choice for the time-sensitive Smart-AMD (Smart-Autistic Meltdown Detector) System. Beyond technical aspects, the model aims to enhance the well-being of autistic children by promptly recognizing and alerting caregivers to abnormal behaviors during a meltdown crisis. This paper introduces a comprehensive system that integrates advanced technology and a profound understanding of autism, offering timely and effective support to those in need.


## 1 INTRODUCTION


The recognition of abnormal behaviors in autistic children during a meltdown crisis is a crucial aspect of developing effective support systems. This paper introduces a novel approach to address this challenge through the utilization of multimodal fusion techniques in the context of deep learning. Specifically, we focus on the early fusion method as an optimal strategy for combining information from two distinct datasets representing emotional states and physical activities.


The process of multimodal fusion involves the integration of information from various sources, a concept well-defined in literature as demonstrated by (Pandeya and Lee, 2021). Three primary fusion techniques,

namely early fusion, late fusion, and hybrid fusion, have been identified. Early fusion involves the merging of low-level features from each modality using correlations, enhancing task accomplishment. However, it may face challenges related to temporal synchronization problems based on multiple input sources. Late fusion, on the other hand, combines unimodal decision values in the decision phase, allowing flexibility and simplicity in predictions even when certain modalities are missing. Hybrid fusion attempts to combine the advantages of both early and late fusion.

In our case, late and hybrid fusion techniques are deemed unattainable due to distinct datasets and the inability to detect children’s faces in all video frames. This results in two different output sizes for the systems focusing on emotions and physical activities. Consequently, we adopt the early fusion method, utilizing a Multi-modal CNN model architecture for meltdown crisis recognition.

<sup>a</sup> <https://orcid.org/0000-0002-5248-8525>

<sup>b</sup> <https://orcid.org/0000-0002-5248-8525>

<sup>c</sup> <https://orcid.org/0000-0002-5248-8525>

The proposed CNN model consists of three Convolution layers with ReLU function activation, followed by two Max-pooling Layers for mapping extracted features. A Fully Connected (FC) layer with ReLU function is employed, and the extracted features are then concatenated and treated as input to a single Fully Connected Layer. Finally, a Fully Connected layer with Softmax function is utilized for classifying children's states based on two modalities: compound emotions and abnormal-complex physical activities.

The presented pseudo-code outlines the development steps of the architecture, emphasizing the details of the early fusion approach. The efficiency of this model is highlighted, as it not only provides the best recognition rate but also proves to be the fastest. Considering the significance of time in the Smart-AMD System, this efficiency becomes a strong point.

In conclusion, the proposed model represents a comprehensive system designed to recognize abnormal behaviors during a meltdown crisis in autistic children. The goal is to assist caregivers in promptly addressing and preventing harm to the children, emphasizing the role of efficient recognition in ensuring the well-being of autistic individuals during crisis situations.

The remainder of this paper is structured as follows: Section 2 conducts a review of the literature covering emotion, physical activity, and human behavior recognition using multimodal approaches. Then, in Section 3, the "MELTDOWN-CRISIS DATASET" is introduced, followed by a focus on autistic child behavior analysis for crisis detection. In Section 5, the experimental results showcase findings from various experiments, including CNN-model architecture applications and validation processes, ensuring result reliability. Finally, the conclusion summarizes key findings and their significance and suggests future research directions.

## 2 RELATED WORKS

Multimodal fusion is a prominent area in both multimodal and artificial intelligence research, aiming to leverage diverse data types for reliable model classification. This process involves transforming data from various single-mode representations into a compact multimodal form (Zhu et al., 2020). Four distinct data fusion techniques have been identified: data-level fusion, early-level fusion, late fusion, and hybrid fusion. Data fusion involves merging different datasets into a unified database. Early fusion integrates low-level features from each modality through correlation, resulting in enhanced task performance. Late

fusion combines unimodal decision values to derive the final decision, while hybrid fusion seeks to combine the strengths of both early and late fusion in a unified framework (Pandeya and Lee, 2021). In the context of emotion and activity recognition, various works utilize multiple data modalities such as images, videos, audio, and information from wearable sensors. These works explore different fusion levels to enhance recognition accuracy. Despite numerous studies addressing similar problems using data from images and videos with handcrafted features, few have attempted to fuse emotion and physical activity modalities for human behavior recognition. To address this gap and analyze autistic behaviors, establishing a foundation for multimodal approaches is essential. These approaches combine facial expressions and physical activities, with a focus on reviewing relevant literature in this field. In the following subsections, we will present state-of-the-art methods that fuse modalities for recognizing emotions, gestures, activities, and human behavior.

### 2.1 Emotion Recognition Based on Multimodal Approaches

Emotion recognition, a prolific research area, spans various fields and involves diverse data types such as facial expressions from images, videos, trajectories, and speech. Robust frameworks have emerged through the fusion of modalities, including combinations of handcrafted features with deep features, auditory and visual modalities, and various other multimodal approaches. Both classical machine learning and deep learning techniques are employed in these endeavors.

In a study by (Busso et al., 2004), the strengths and weaknesses of facial expression and acoustic emotion classifiers were analyzed. Unimodal systems often encountered misclassifications for certain emotion pairs, but these confusions were mitigated by introducing another modality. Consequently, the bimodal emotion classifier outperformed individual unimodal systems. Two fusion approaches, namely feature-level and decision-level fusion, were compared, yielding similar overall performance. However, specific emotions exhibited notable variations. The feature-level bimodal classifier excelled in recognizing anger and neutral states, while the decision-level bimodal classifier achieved high accuracy in classifying happiness and sadness. Additionally, (Castellano et al., 2007) presented a multimodal approach for recognizing eight emotions. This approach integrated information from facial expressions, body movement, gestures, and speech, demonstrating the

potential of combining multiple modalities for comprehensive emotion recognition. In analyzing a Bayesian classifier model, the authors used a multimodal corpus comprising eight emotions and ten subjects. Initially, individual classifiers were trained for each modality, and subsequent fusion of data at both feature and decision levels resulted in recognition rate improvements exceeding 10% compared to unimodal systems. Notably, feature-level fusion outperformed decision-level fusion. (Kessous et al., 2010) proposed a speech-based multimodal emotion recognition system during interactions. Their dataset featured individuals pronouncing sentences with various emotions during interactions with an agent. Combining facial expressions, gestures, and acoustic speech analysis, a Bayesian classifier was employed for automatic classification of unimodal, bimodal, and multimodal data. Fusion at the feature and results levels significantly enhanced recognition rates, surpassing unimodal systems by over 10%. Investigation into bimodal emotion recognition combinations revealed 'gesture-speech' as the most effective pairing, with a 3.3% improvement over the best bimodal results.

(Psaltis et al., 2019) explored integrating emotion recognition technology into gaming applications to enhance interaction and the gaming experience. They presented an emotion recognition methodology using multimodal fusion analysis to identify players' emotional states during gameplay scenarios. In this context, two mono-modal classifiers were devised for extracting affective state information from facial expression and body motion analysis. To amalgamate modalities, the authors introduced a deep model for determining the player's affective state. Evaluating their approach involved collecting a bimodal dataset using Microsoft's Kinect sensor, incorporating feature vectors from users' facial expressions and body gestures. This method outperformed mono-modal and early-fusion algorithms, achieving a recognition rate of 98.3%. Similarly, (Pandeya and Lee, 2021) proposed a multimodal approach for comprehending human emotions. They constructed a balanced music video emotion dataset, testing it over four unimodal and four multimodal convolutional neural networks (CNNs) for music and video. Evaluation results demonstrated improved performance for multimodal architectures compared to individual unimodal emotion classifiers, with an accuracy of 88.56% achieved by integrating all multimodal structures. In a parallel vein, (Radoi et al., 2021) presented a robust end-to-end architecture incorporating multimodal information for emotion recognition. The Temporally Aggregated Audio-Visual Network (TA-AVN) architecture flexibly merges audio and video data at various

sampling rates across modalities. This approach accommodates an asynchronous combination of temporal multimodal information, achieving competitive results on challenging datasets, with overall accuracies of 84.0% for CREMA-D and 78.7% for RAVDESS.

## 2.2 Physical Activity Recognition Based on Multimodal Approaches

Motion recognition, advancing with diverse sensor applications like wearables, vision-based, and speech sensors, benefits from the integration of multiple modalities for robust performance. In a multimodal approach by (Masurelle et al., 2013), isolated complex human body movements, specifically Salsa dance steps, were recognized. The system utilized motion features from 3D sub-trajectories of dancers' body-joints (extracted from Kinect depth map sequences) through Principal Component Analysis (PCA). Sub-trajectories were obtained from a foot-step impact detection module, utilizing piezoelectric sensors on the dance floor. Two classifiers, Gaussian mixture models and hidden Markov models (HMM), tested on a multimodal Salsa Dataset using HMM classifiers, achieved a 74% F-measure in recognizing gestures among six classes.

(Li et al., 2017) highlighted the performance limitations of individual sensors, especially for categorizing similar activities. They addressed this by fusing information from experimental data collected using different sensors, including a tri-axial accelerometer, a micro-Doppler radar, and a depth camera. The fusion of heterogeneous information improved the overall system performance, leading to a global classification rate increase up to 91.3% based on the combination of accelerometer, radar, and RGB-Depth data. In their work, (Tian et al., 2020) introduced a sample database of RGB-D gesture images, preprocessed the samples, and devised a multimodal, multilevel fusion gesture recognition framework. They designed a convolutional neural network structure with two modes, extracting features at different abstract levels for each mode. To address the challenge of varying feature dimensions in different modes, they proposed a feature mapping model to align features into a common space, creating a unified feature set. (Lin et al., 2020) proposed a data fusion framework to merge data from Microsoft Kinect and wearable sensors, aiming to enhance Human Action Recognition (HAR) accuracy. While Kinect captures body motion characteristics for various activities, its accuracy depends on the viewing angle. The integration of Kinect and wearable sensors compensates for each other's limitations. The authors introduced a novel system utilizing incremental learn-

ing, a decision table, and swarm-based feature selection for quick and accurate HAR based on both sensor data. Experimental results demonstrated a significant improvement in HAR accuracy (from 23.51% to 68.35%) when combining Kinect sensors viewed at a ninety-degree angle with wearable sensors. Human action recognition is pivotal for developing intelligent solutions in home environments, especially in ambient assisted living applications. According to (Franco et al., 2020), automated systems, leveraging the capabilities of Kinect sensors, can significantly enhance human quality of life. By interpreting user needs, recognizing unusual behaviors, and preventing potential hazards, these systems contribute to a safer and more efficient living environment. This study exploits the full potential of the Kinect sensor, combining Skeleton and RGB data streams for a robust activity recognition method. The Skeleton representation tracks body postures, while the RGB images capture the temporal evolution of actions. In the work of (Yu et al., 2020), the authors introduced D3D-LSTM, featuring real-time feature fusion for enhanced discrimination of similar actions. The model includes a high-attention mechanism assigning different weights to frames in real-time. An alternating optimization strategy further refines the model. Evaluating D3D-LSTM on Realset, SBU-Kinect, and MSR-action-3D datasets demonstrated its effectiveness, pushing the average rate of SBU-Kinect to 92.40% and MSR-action-3D to 95.40%.

### 2.3 Human-Behavior Recognition Based on Multimodal Approaches

According to (Ambady and Rosenthal, 1992), humans assess expressive behaviors through both verbal and non-verbal channels. Verbal channels involve speech, while non-verbal channels encompass eye gaze, blink, facial and body expressions, and speech prosody. Various approaches have been proposed to fuse multiple modalities for recognizing human behaviors. (Pimpalkar et al., 2014) explored human-computer interaction to enhance computer awareness of user behaviors, particularly for assisting disabled individuals in expressing themselves. They introduced a multimodal approach for behavior recognition, exemplified by a gesture recognition system using a webcam. The model incorporated facial expression and hand gesture recognition, utilizing the "FABO bimodal database" (Metri et al., 2011) that recorded combined face and body expressions simultaneously. To assess their software, the authors employed the Principle Component Analysis algorithm (PCA) for face recognition and the Cam Shift algo-

rihm for tracking hands and predicting their locations in images.

(Lin et al., 2020) proposed a computational framework for modeling vocal behaviors and body gestures during Autism Diagnostic Observation Schedule interviews. The learnable Interlocutor-Modulated (IM) attention mechanism categorized ASD subgroups considering the subtle and challenging nature of ASD behaviors. The multimodal network comprised speech-IM-aBLSTM and motion-IM-aBLSTM networks, fused to differentiate Autistic Disorder (AD), High-Functioning Autism (HFA), and Asperger Syndrome (AS). The IM attention mechanism tracked non-linear behavioral dependencies between interlocutors, achieving a UAR of 66.8% on a large ADOS collection. (Alban et al., 2021) emphasized technology's utility in detecting and improving therapy for challenging behaviors in autistic children. They explored detecting behaviors using a wearable sensor (Empatica E4 wristband) and machine learning. The annotation approach recorded instances of challenging behaviors and stimuli group interactions with social robots. Features were analyzed using Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Decision Tree (DT) techniques. The model achieved promising results (97% accuracy), suggesting potential efficiency in addressing challenging behaviors. A detection system with wearable sensors could notify parents or caregivers for timely intervention, while social companion robots could mediate and react to alleviate challenging behaviors.

### 2.4 Discussion

All the aforementioned works proposed efficient frameworks for emotion, activity and behavior recognition based on multimodal approaches. These works suggested several types of fusion methods such as data-fusion, feature-level fusion, decision-fusion or hybrid-fusion. Moreover, all these studies have shown that the fusion of several modalities (two or more modalities) increased the performance and robustness of the proposed systems. In addition, in these approaches, authors used either machine learning techniques (SVM, k-Nearest Neighbors (KNN), Naïve Bayes, Hidden Markov Model (HMM), etc.) or deep learning techniques (Convolutional Neural Network or CNN (CNN), Long Short Term Memory (LSTM)) to classify emotions, activities, or behaviors. Models with high accuracy levels are implemented using deep learning techniques. However, none of these works collected multimodal data for autistic children in an uncontrolled environment and during a Melt-



down crisis by using Kinect sensors. Moreover, no work processed facial and body data during a synchronized period to recognize the behaviors of autistic people in normal states or during crises. In our work, we suggest to use the early fusion to merge two different modalities, namely compound emotions and abnormal-complex physical activities, to recognize behaviors and detect meltdown crises. To this endeavor, we tested custom architectures of deep learning to determine the most suitable and efficient model for our case (See Table 1).

### 3 "MELTDOWN-CRISIS DATASET" DESCRIPTION

In this research project, we created a novel dataset called Meltdown Crisis, which contains realistic situations of autistic children in daily activities as well as during a meltdown crisis. This was necessary because there were no publicly available and/or realistic datasets. Making videos is an extremely important, delicate, and serious undertaking. Furthermore, it's crucial to take ethics into account when filming autistic youngsters on camera. In fact, in any society, getting permission to film young kids might be challenging. We were able to register 23 autistic children at the healthcare center "ASSAADA" for autistic children, whose ages range from 6 to 15 years old, because of our extensive study in healthcare centers for autistic children around the world and specifically in Tunisia. Thirteen of the twenty-three autistic youngsters who had the worst meltdown symptoms participated in our study. They were between the ages of five and nine. Using a Kinect V2 camera set to record at 30 frames per second, we watched and documented the behavior of the thirteen youngsters who were chosen over three months in real-world settings. Three rooms are used for video acquisition, with preset parameters and an average video length of one hour. Further diagnosis and description of our "Meltdown Crisis" Dataset can be found in (Masmoudi et al., 2019).

### 4 AUTISTIC CHILD BEHAVIOR ANALYSIS FOR MELTDOWN CRISIS DETECTION

The process of combining information from numerous sources for regression tasks is often defined as multimodal fusion. Multimodal fusion introduces the advantages of using a robust and complementary information gain model and the functional continuity

of the system even in case of failure of one or more modalities ((Ouyang et al., 2017), (Ding et al., 2016), (Zhang et al., 2016)). In other words, early fusion merges low-level features from each modality using correlations for a better task accomplishment. However, it is sometimes difficult to implement due to same temporal synchronization problems based on multiple input sources. The decision phase fusion, however, gains unimodal decision values which are combined to reach the final decision. Although late fusion ignores a few low-level interactions, it permits easy training with more flexibility and simplicity for making predictions when one or more modalities are missing. The hybrid fusion (mid-level) attempts to exploit the advantages of both early and late fusion in a common framework.

In our case, the aforementioned fusions (late and hybrid) are unattainable because the suggested systems of emotions and physical activities were trained, particularly on two distinct datasets. Children's faces cannot be detected at the present time in all video frames. Hence, two different output sizes emerged for both systems. Thus, we adopt the early fusion method. Figure 1 shows the Multimodal CNN-model architecture for meltdown crisis recognition. In this CNN-model, we used three Convolution layers with ReLu function activation. Two Max-pooling Layers are also used for mapping extracted features from the Convolution Layer. Then, a Fully Connected (FC) layer with ReLu function is used. Once these features are extracted, they are concatenated and considered as input to the single Fully Connected Layer. After that, a Fully Connected layer with Softmax function is used to classify the children's states (**0 for Normal state and 1 for Meltdown state**) based on two modalities, compound emotions, and abnormal-complex physical activities. The pseudo-code presented subsequently, reports the details of the architecture development steps:

In this context, this model represents the most efficient model because it is the fastest model that provides us with the best recognition rate. Thus, this is a strong point because time is a significant factor for Smart-AMD System. Consequently, this step allows us to propose a complete system that recognizes abnormal behaviors during a meltdown crisis. Thus, this model should help autistic children avoid harming themselves in case of a meltdown crisis by alerting their caregivers.

Table 1: Overview of emotion and physical activity recognition based on multimodal approaches.

Proposal	Fusion Technique	Key Findings	Recognition Rate
(Busso et al., 2004)	Feature-level and Decision-level fusion	Bimodal emotion classifier outperformed unimodal systems, with feature-level fusion excelling in anger and neutral states recognition.	85.2%
(Castellano et al., 2007)	Feature-level Fusion	Fusion of data at feature and decision levels significantly improved recognition rates, particularly for eight emotions.	91.6%
(Kessous et al., 2010)	Feature and Decision-level Fusion	Gesture-speech fusion showed the most effective improvement in emotion recognition.	89.3%
(Psaltis et al., 2019)	Early Fusion	Deep model for emotion recognition achieved higher accuracy compared to mono-modal and early-fusion algorithms.	92.1%
(Pandeya and Lee, 2021)	Hybrid Fusion	Multimodal CNNs outperformed unimodal classifiers, achieving 88.56% accuracy in emotion recognition.	88.56%
(Radoi et al., 2021)	Late Fusion	Temporally Aggregated Audio-Visual Network (TA-AVN) achieved competitive results, with accuracies of 84.0% and 78.7%.	84.0%
(Masurelle et al., 2013)	Utilized Gaussian mixture models and HMM classifiers	Multimodal approach achieved 74% F-measure in recognizing Salsa dance steps.	74.0%
(Li et al., 2017)	Utilized fusion of accelerometer, radar, and RGB-Depth data	heterogeneous fusion improved overall system performance, leading to a global classification rate increase.	91.3%
(Tian et al., 2020)	Multilevel Fusion	Proposed framework aligned features from different modes into a common space, improving gesture recognition accuracy.	92.5%
(Lin et al., 2020)	Data Fusion	Integration of Kinect and wearable sensors significantly improved Human Action Recognition accuracy.	68.35%
(Franco et al., 2020)	Utilized Skeleton and RGB data fusion	Fusion of Skeleton and RGB data streams enhanced activity recognition in home environments.	
(Yu et al., 2020)	Utilized real-time feature fusion	D3D-LSTM model achieved high accuracy in discriminating similar actions.	95.40%
(Pimpalkar et al., 2014)	Utilized gesture and facial expression recognition	Gesture recognition system using webcam showed potential in enhancing computer awareness of user behaviors.	
(Lin et al., 2020)	Utilized speech and motion-IM-aBLSTM networks	IM attention mechanism differentiated Autistic Disorder (AD), High-Functioning Autism (HFA), and Asperger Syndrome (AS).	66.8%
(Alban et al., 2021)	Utilized Empatica E4 wristband and machine learning	Detection system with wearable sensors achieved promising results in addressing challenging behaviors in autistic children.	97.0%

## 5 EXPERIMENTAL RESULTS OF MELTDOWN CRISIS DETECTION BASED ON AUTISTIC CHILDREN'S BEHAVIOR ANALYSIS

To evaluate and validate our multimodal approach, we carried out a set of experiments by using the proposed deep model architecture with different settings and parameters.

### 5.1 Experimental Results

In this section, an evaluation step is carried out to select the best model that allows to detect a Meltdown crisis state. To this endeavor, we tested three architectures that will be described subsequently.

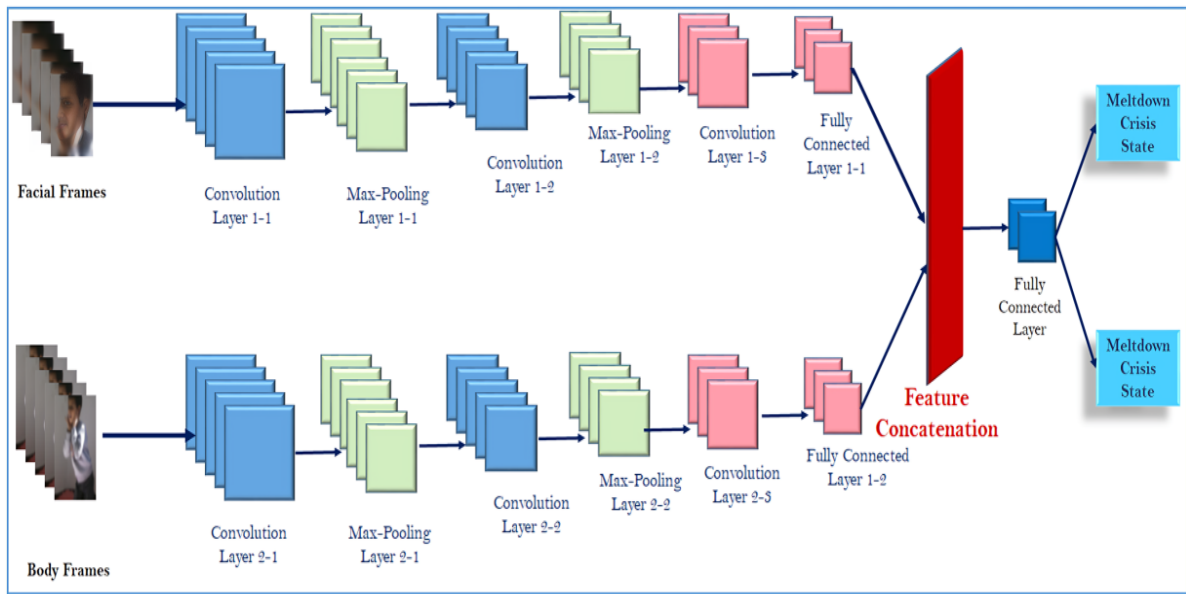


Figure 1: Multi-modal CNN model architecture for meltdown crisis recognition based on compound emotions and abnormal-complex physical activities.

Table 2: Architecture development steps of our Multimodal CNN model.

<b>INPUTS</b>	<p><b>-Input 1:</b> <math>D_f</math> as the labeled data for emotions</p> <p><b>-Input 2:</b> <math>D_b</math> as the labeled data for activities.</p>
<b>PROCESSING</b>	<p>- The layers of the second modality are renamed to remove ambiguity (see Algorithm 1).</p> <p>- Features extracted from the emotion modality are defined as the set of <math>\{X_1^f, X_2^f, X_3^f, \dots, X_n^f\}</math></p> <p>- Features extracted from the activity modality are defined as the set of <math>\{X_1^b, X_2^b, X_3^b, \dots, X_n^b\}</math></p>
<b>OUTPUT</b>	<p>- <math>Y_n^{f+b} = 0</math> for Normal state</p> <p>- <math>Y_n^{f+b} = 1</math> for Meltdown Crisis state</p>

**5.1.1 Experiment 1: CNN-Model Architecture Based on Inception-Resnet-V2 Features**

In this experiment, we suggested to concatenate features extracted from the pre-trained InceptionResnetV2 model. The latter is considered as the best model that achieved the best results for compound emotion and the abnormal-complex physical activity systems (cf. Figure 2). Then we applied the Principal Component Analysis (PCA) method to select relevant features (2286 features). Once these features were selected, they were concatenated and presented as input for machine learning algorithms such as DT, MLP, KNN, NB and SVM. The best results were recorded

(86.9% of Accuracy rate) using DT classifier with Bayesian optimization (See Table 3).

The core of this approach that combine features extracted from each modality generate decisions about the tackle based on multimodal analysis. Based on the experiments, we have obtained a performance about 86.9%. Of course this rate is acceptable because first it is based on multimodal approach. Second, it is considered as the first work analyzing complex and realistic dataset. In fact, based on the experiments of the literature review, the multimodal approach usually improves the results. However, in our case, the results showed a decreasing of performance comparing to the single modalities. At the same time, we cannot consider single modality approach either based on activity or based on emotion because we cannot guarantee that this modality appears in meltdown crisis. So, we should rely on a multimodal approach.

For these reasons, we have tried to enhance our proposed model by proposing other architectures. So, by taking advantage of the early fusion, we proposed a multimodal CNN-model architecture for meltdown crisis behavior recognition. These architectures are described in the following subsections.

**5.1.2 Experiment 2: Customized Multimodal CNN-Model Architecture**

As shown in Figure 3, we proposed a CNN-based architecture with two input streams for detecting facial expressions and for physical activities. Then, we introduced the same layers with the same parameters;

```

Data: Input 1:  $D_f$  //Labeled Data Frame of Emotions
Input 2:  $D_b$ 
// Labeled Data Frame of Activities
Result: Output:  $Y_n^{f+b} = 0$  //for normal state;
 $Y_n^{f+b} = 1$  //for meltdown crisis state
//  $X_n^f$  : representative vector of  $D_f$ 
//  $X_n^b$  : representative vector of  $D_b$ 
//  $Y_n^{f+b}$  : Predicted classes
// numHiddenDimension: dimension of hidden layer
// numClasses=2
// Layers=Layer1 = convolution2dLayer,
Layer2=maxPooling2dLayer,
Layer3=convolution2dLayer,
Layer4=maxPooling2dLayer,
Layer5=convolution2dLayer,
Layer6=fullyConnectedLayer]:layers of the proposed model
//Layer: one of model's layers
Layers = createLayer(
 $X_n^f$ ,numHiddenDimension) //create layers for input1
Layers2 = createLayer( $X_n^b$ ,
,numHiddenDimension) //create layers for input2
//When the two layers are merged, the same name of the layers cannot be used. So, we Use renameLayerfunction to rename the layer name in layers2
For Layer in Layers do
// for each layer in the model Layers2 = renameLayerFunction(Layers2,'-2');
// rename layers of input2 by adding '-2'
end
LayersAdd = concatenationLayer(1, 2, 'Name', 'cat') //add concatenation layer in order to fuse features extracted from our two inputs
FeaturesMap = ConcatenateFeatures{ $X_1^f, X_2^f, X_3^f, \dots, X_n^f + X_1^b, X_2^b, X_3^b, \dots, X_n^b$ }
//The Feature map of the concatenate features extracted from  $D_f$  et  $D_b$ 
LayersAdd = fullyConnectedLayer(numClasses,'Name')
//add fullyConnectedLayer for classification purposes
    
```

Algorithm 1: Rename layers of our CNN-model with two inputs.

- The first layer is allocated to the Convolution layer with kernel - size = 4 and ReLu activation function.
- The second layer is a max-pooling layer with pool - size (2, 2).
- The third layer is allocated to Convolution layer with kernel - size = 4 and ReLu as activation layer.

- The fourth layer is a max-pooling layer with pool - size (2, 2).
- The fifth layer is allocated to Convolution layer with kernel - size = 4 and ReLu as activation layer.
- The sixth layer is a max-pooling layer with pool — size (2, 2).
- The seventh layer is a flatten layer.
- A concatenation Layer is employed to concatenate features extracted from two modalities.
- For classification purposes, three Dense layers are employed; the first layer consists of 128 nodes and ReLu activation function, the second Dense layer is composed by 10 nodes and ReLu activation function.Finally, a Dense Layer is allocated for classification purposes with the Softmax Function.
- To compile this network, the Adam optimizer is utilized with these default settings (Learning rate = 0.001, beta-1 = 0.9, beta-2 = 0.999, epsilon = le-07, amsgrad = False). Moreover, to fit this network, we used 100 epochs, 128 as a batch size. We obtained motivating results with a validation accuracy of **76.80%**.

Table 3: Classification results of the Inception-Resnet-V2 Fused Features.

Algorithm	Parameters	Accuracy
<b>DT</b>	Fine Tree	80.7%
	Medium Tree	85.7%
	Coarse Tree	85.5%
	Optimizable	86.9%
<b>SVM</b>	LinearSVM	69.2%
	Quadratic SVM	74.4%
	Cubic SVM	74.4%
	Fine Gaussian SVM	74.4%
	Medium Gaussian SVM	74.4%
	Coarse Gaussian SVM	60.6%
<b>KNN</b>	FineKNN	73.9%
	MediumKNN	64.0%
	CoarseKNN	39.5%
	CosineKNN	72.2%
	CubicKNN	59.6%
	Weighted KNN	71.2%
<b>MLP</b>	With 2 hidden layers (10 nodes -20 nodes)	85.1%
	With 3 hidden layers (10 nodes -10 nodes -10 nodes)	79.8%
	With 3 hidden layers (20 nodes -10 nodes -10 nodes)	80.0%



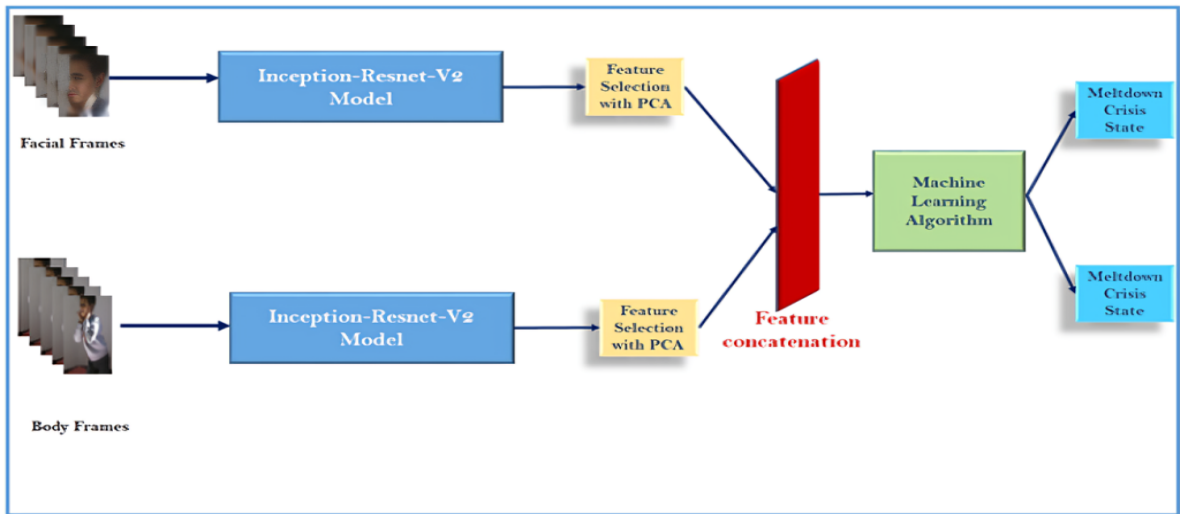


Figure 2: CNN-model architecture based on Features extracted with Inception-Resnet-V2 model.

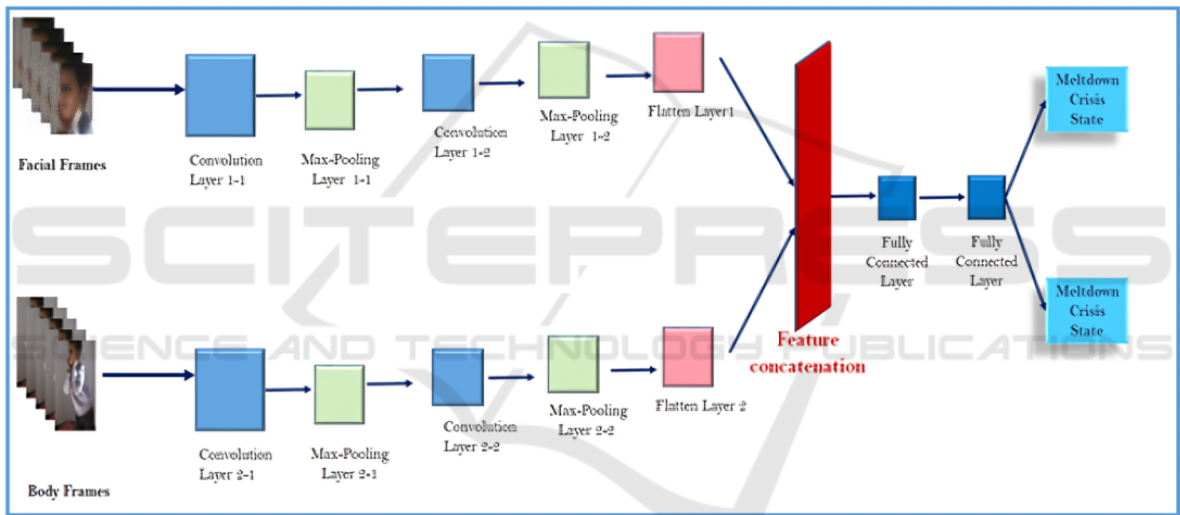


Figure 3: CNN-Model architecture.

### 5.1.3 Experiment 3: Multimodal CNN-Model Architecture for Meltdown Crisis Recognition Based on Compound Emotions and Abnormal-Complex Physical Activities.

As illustrated in Figure 1, we proposed a CNN based architecture with two input streams for detecting facial expressions and for physical activities. Then, we put forward the same layers with the same parameters;

- The first layer is allocated to a Convolution layer with kernel-size= 4 and ReLu as activation function.
- The second layer is a max-pooling layer with pool-size(2, 2).

- The third layer is allocated to a Convolution layer with kernel-size = 4 and ReLu as an activation layer.
- The fourth layer is a max-pooling layer with pool-size(2, 2).
- The fifth layer is allocated to the Convolution layer with kernel-size = 4 and ReLu as the activation layer.
- The sixth layer is allocated to the Fully Connected layer with 128 nodes.
- A concatenation Layer is employed to concatenate the features extracted from two modalities.
- For classification purposes, a Dense layer is employed with the Softmax Function.

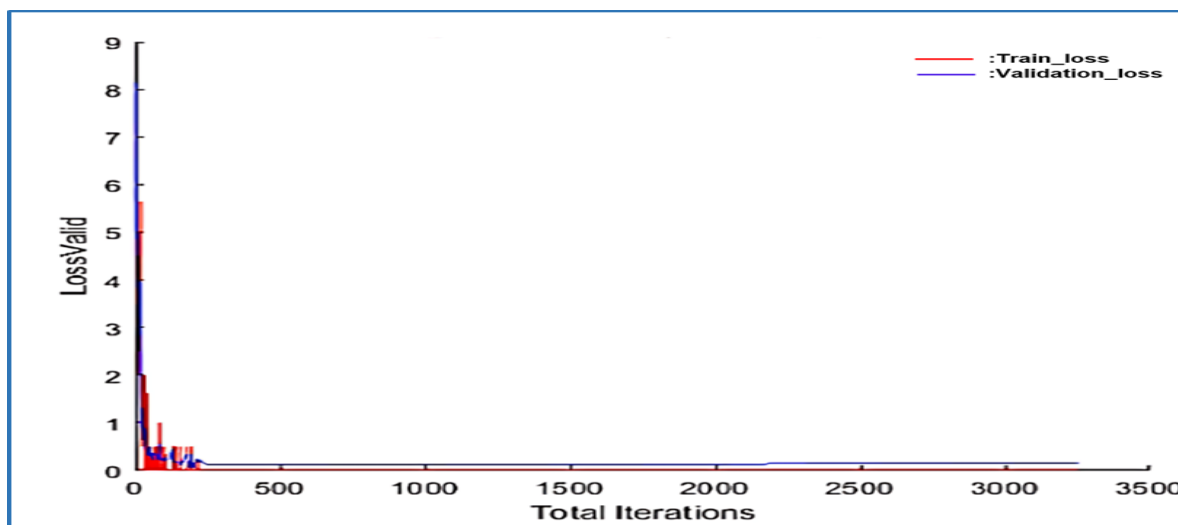


Figure 4: Values for training Loss and validation Loss.

- To compile this network, the Adam optimizer is utilized with these settings (Learning-rate = 0.001, beta-1 = 0.9, beta-2 = 0.999, epsilon = 1e-07, amsgrad = False). Moreover, to fit this network, we used 50 epochs, 32 as a batch-size.

Our "MeltdownCrisis" dataset is composed of Emotions: (15113 Frames); Physical activities (24929 Frames). Because, we run our model on "cpu" environment, by using 3000 frames for each modality. These data are split into 70% for the training dataset, 20% for the testing dataset and 10% for the validation dataset. The recorded result achieved a validation accuracy of **99.50%**. The function Loss of this model is presented in the Figure 4. The red line represents the Loss function of training data and the blue line stands for the Loss function of validation data. So, this model represents the most efficient model because it is the fastest model that provides us with the best recognition rate. Thus, this is a strong point because time is a significant factor for Smart-AMD System.

## 5.2 Validation

To perform a quantitative evaluation of the best recorded result and validate our proposed approach, we measured Recall, Precision and F-measure values. The obtained results are **99.50%** Validation-accuracy, **99.75%** Precision, **98.50%** Recall and **99.62%** F-measure.

### 5.2.1 Discussion

To achieve a more realistic evaluation, we assessed our suggested approach by comparing it with existing literature works. According to Table 4, the work of (Psaltis et al., 2019) permits to recognize of the behaviors of normal people based on their facial and body gesture modalities. In this work, the dataset was acquired using the Kinect camera. The authors proposed a CNN network for feature extraction and SVM for classification purposes. The recorded classification rate was 98.3%. In addition, based on two facial expressions and speech modalities, [94] suggested an Attentional BLSTM model to recognize autistic behaviors and classify autistic children in different groups. The dataset was acquired using High Definition cameras during interviews between the psychiatrist and each autistic child. The obtained classification rate is 68.6%. However, our proposed model proves its effectiveness, compared with both (Psaltis et al., 2019) and (Lin et al., 2020). This can be explained by the fact that these approaches did not address the challenges of recognizing autistic abnormal behaviors during a Meltdown crisis. In addition, they were not conducted in an uncontrolled environment.

## 6 CONCLUSION

In conclusion, this paper presented a comprehensive approach for analyzing autistic children's behavior during a meltdown crisis, with a focus on the detection of such crises. The proposed multimodal fusion method adopted early fusion due to challenges arising from different output sizes of the suggested

Table 4: A Comparative study with state of the art methods.

Paper	Approach	People Category	Modality Type	Camera Type	Dataset	Classification Method	Classification Rate
(Psaltis et al., 2019)	CNN and SVM	Normal	-Facial expression -Body gestures	Kinect camera	Facial expression and body gesture dataset	SVM	98.3%
(Lin et al., 2020)	Attentional BLSTM	Autist	Facial expressions and speech records	High Definition cameras	Audio-video ADOS interview dataset	Dense layer with the Softmax function	68.8%
<b>Our</b>	CNN based model	Autist	Compound emotions Frames and Abnormal-Complex Physical activity Frames	Kinect camera	Meltdown Crisis dataset	Dense layer with the Softmax function	<b>99.5%</b>

emotion and physical activity systems. The resulting Multi-modal CNN model architecture showcased three Convolution layers, two Max-pooling Layers, and Fully Connected layers to efficiently recognize meltdown crisis states based on compound emotions and abnormal-complex physical activities.

The experimental results, detailed in three distinct experiments, demonstrated the effectiveness of the proposed approach. In Experiment 1, using the Inception-Resnet-V2 model features and applying Principal Component Analysis (PCA), the best accuracy of 86.9% was achieved using the Decision Tree (DT) classifier with Bayesian optimization. In Experiment 2, a Customized Multimodal CNN-model Architecture exhibited a validation accuracy of 76.80%, providing a fast and efficient model. Finally, Experiment 3 introduced a Multimodal CNN-model architecture specifically designed for meltdown crisis recognition, achieving an outstanding validation accuracy of **99.50%**.

The validation step further confirmed the robustness of the proposed approach, with a confusion matrix showing high precision (99.75%), recall (98.50%), and F-measure (99.62%) values. The comparison with existing literature highlighted the superiority of the proposed model in addressing the unique challenges of recognizing autistic abnormal behaviors during a meltdown crisis in an uncontrolled environment. Overall, the presented model provides a valuable tool for caregivers to promptly identify and intervene in meltdown crises, potentially preventing harm to autistic children. The approach demonstrates promising results and opens avenues for further re-

search and development in the field of autism behavior analysis and crisis detection. As future works, we will look forward to evaluating Smart-AMD in authentic contexts. Furthermore, we aim to enrich our "MeltdownCrisis" dataset with other videos and different meltdown crisis scenarios taken from other healthcare centers. In addition, we aim to propose a new system that allows us to identify and detect abnormal and stereotyped facial expressions and physical activities during multiple states. These works will be developed using Deep Learning techniques, such as CNN, LSTM, ConvLSTM, etc.

## REFERENCES

- Alban, A. Q., Ayesh, M., Alhaddad, A. Y., Al-Ali, A. K., So, W. C., Connor, O., and Cabibihan, J.-J. (2021). Detection of challenging behaviours of children with autism using wearable sensors during interactions with social robots. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 852–857. IEEE.
- Ambady, N. and Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *The American Psychological Association*, 1111(2):256–274.
- Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., and Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211.
- Castellano, G., Kessous, L., and Caridakis, G. (2007). Multimodal emotion recognition from expressive faces,

- body gestures and speech. *Doctoral Consortium of ACHI, Lisbon*.
- Ding, W., Xu, M., Huang, D., Lin, W., Dong, M., Yu, X., and Li, H. (2016). Audio and face video emotion recognition in the wild using deep neural networks and small datasets. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 506–513.
- Franco, A., Magnani, A., and Maio, D. (2020). A multimodal approach for human activity recognition based on skeleton and rgb data. *Pattern Recognition Letters*, 131:293–299.
- Kessous, L., Castellano, G., and Caridakis, G. (2010). Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, 3(1):33–48.
- Li, H., Shrestha, A., Fioranelli, F., Le Kerneq, J., Heidari, H., Pepa, M., Cippitelli, E., Gambi, E., and Spinsante, S. (2017). Multisensor data fusion for human activities classification and fall detection. In *2017 IEEE SENSORS*, pages 1–3. IEEE.
- Lin, Y.-S., Gau, S. S.-F., and Lee, C.-C. (2020). A multimodal interlocutor-modulated attentional blstm for classifying autism subgroups during clinical interviews. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):299–311.
- Masmoudi, M., Jarraya, S. K., and Hammami, M. (2019). Meltdowncrisis: Dataset of autistic children during meltdown crisis. In *15th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*, pages 239–246. IEEE.
- Masurelle, A., Essid, S., and Richard, G. (2013). Multimodal classification of dance movements using body joint trajectories and step sounds. In *2013 14th international workshop on image analysis for multimedia interactive services (WIAMIS)*, pages 1–4. IEEE.
- Metri, P., Ghorpade, J., and Butalia, A. (2011). Facial emotion recognition using context based multimodal approach.
- Ouyang, X., Kawaai, S., Goh, E. G. H., Shen, S., Ding, W., Ming, H., and Huang, D.-Y. (2017). Audio-visual emotion recognition using deep transfer learning and multiple temporal models. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 577–582.
- Pandeya, Y. R. and Lee, J. (2021). Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 80(2):2887–2905.
- Pimpalkar, A., Nagalkar, C., Waghmare, S., and Ingole, K. (2014). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *International Journal of Computing and Technology (IJCAT)*, 1(2).
- Psaltis, A., Kaza, K., Stefanidis, K., Thermos, S., Apostolakis, K. C., Dimitropoulos, K., and Daras, P. (2019). Multimodal affective state recognition in serious games applications. In *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 435–439. IEEE.
- Radoi, A., Birhala, A., Ristea, N.-C., and Dutu, L.-C. (2021). An end-to-end emotion recognition framework based on temporal aggregation of multimodal information. *IEEE Access*, 9:135559–135570.
- Tian, J., Cheng, W., Sun, Y., Li, G., Jiang, D., Jiang, G., Tao, B., Zhao, H., and Chen, D. (2020). Gesture recognition based on multilevel multimodal feature fusion. *Journal of Intelligent & Fuzzy Systems*, 38(3):2539–2550.
- Yu, J., Gao, H., Yang, W., Jiang, Y., Chin, W., Kubota, N., and Ju, Z. (2020). A discriminative deep model with feature fusion and temporal attention for human action recognition. *IEEE Access*, 8:43243–43255.
- Zhang, S., Zhang, S., Huang, T., and Gao, W. (2016). Multimodal deep convolutional neural network for audio-visual emotion recognition. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 281–284.
- Zhu, H., Wang, Z., Shi, Y., Hua, Y., Xu, G., and Deng, L. (2020). Multimodal fusion method based on self-attention mechanism. *Wireless Communications and Mobile Computing*, 2020.