Continuous-Time Multimodal Emotion Recognition Using Neural CDEs and Cross-Modal Attention

A thesis submitted in partial fulfillment of the requirements
for the award of the degree of

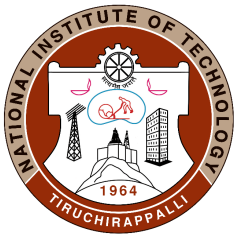
M.Tech

in

COMPUTER SCIENCE AND ENGINEERING

By

SHIVENDU MISHRA (206124031)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY
TIRUCHIRAPPALLI – 620015

DECEMBER 2025

BONAFIDE CERTIFICATE

This is to certify that the project titled **CONTINUOUS-TIME MULTI-MODAL EMOTION RECOGNITION USING NEURAL CDES AND CROSS-MODAL ATTENTION** is a bonafide record of the work done by

SHIVENDU MISHRA (206124031)

in partial fulfillment of the requirements for the award of the degree of **Master of Technology** in Computer Science and Engineering of the **NATIONAL INSTITUTE OF TECHNOLOGY, TIRUCHIRAPPALLI**, during the year 2025-2026.

Dr. R. Bala Krishnan	**Dr. Kunwar Singh**
Guide	Head of the Department

Project Viva-voce held on _____

Internal Examiner **External Examiner**

ABSTRACT

Multimodal emotion recognition is a pivotal task in affective computing, requiring the integration of diverse physiological signals such as Electrocardiogram (ECG), Electrodermal Activity (EDA), and Accelerometer (ACC) data. The core challenge lies in modeling the continuous temporal dynamics of these signals, which often exhibit irregular sampling rates and asynchronous observations across different modalities. This work proposes a novel framework based on **Neural Controlled Differential Equations (Neural CDEs)**, which treats physiological signals as continuous control paths. By constructing natural cubic splines from raw observations, the model learns latent trajectories governed by a learnable vector field, effectively handling multi-rate data without the need for lossy preprocessing. We further integrate a **Multimodal Transformer** to fuse these continuous-time latent representations, capturing complex cross-modal dependencies through a series-aware attention mechanism. Evaluated on the WESAD dataset, the proposed model demonstrates superior robustness and accuracy. Experiments show that the Neural CDE approach maintains high performance across Baseline, Stress, and Amusement states, achieving an accuracy of 91.2%.

Keywords: Neural Controlled Differential Equations; Multimodal Emotion Recognition; Physiological Signals; Affective Computing; WESAD Dataset

ACKNOWLEDGEMENT

I wish to place on record my deep sense of gratitude to my guide, **Dr. R. Bala Krishnan**, for his inspiring guidance and constant encouragement throughout the course of this project. I am grateful to the Head of the Department, **Dr. Kunwar Singh**, and the faculty members of the Department of Computer Science and Engineering for providing the necessary facilities. Finally, I thank my parents and friends for their unwavering support.

TABLE OF CONTENTS

Title **Page No.**

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ECG	Electrocardiogram
EDA	Electrodermal Activity
ACC	Accelerometer
NCDE	Neural Controlled Differential Equation
WESAD	Wearable Stress and Affect Detection
LOSO	Leave-One-Subject-Out

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

Affective computing, a paradigm first articulated by Picard in 1997, represents the confluence of computer science, psychology, and cognitive neuroscience, aimed at developing systems capable of recognizing, interpreting, and synthesizing human emotional states. The fundamental premise underlying this field is that emotions play a critical role in human cognition, decision-making, and social interaction. Traditional emotion recognition systems have relied predominantly on behavioral modalities such as facial expressions, speech patterns, and body language. However, these approaches are inherently susceptible to cultural biases, voluntary suppression of emotional expression, and contextual variability.

Physiological signal-based emotion recognition offers a compelling alternative paradigm. Unlike behavioral cues, physiological responses to emotional stimuli are mediated by the autonomic nervous system (ANS) and are therefore largely involuntary and resistant to conscious manipulation. The ANS comprises two antagonistic branches: the sympathetic nervous system (SNS), responsible for the "fight-or-flight" response, and the parasympathetic nervous system (PNS), which governs "rest-and-digest" functions. Emotional arousal triggers characteristic changes in cardiovascular activity, electrodermal response, and muscular tension that can be quantified through sensors such as Electrocardiogram (ECG), Electrodermal Activity (EDA), and Accelerometer (ACC) measurements.

Despite their theoretical advantages, physiological signals present significant computational challenges. First, they are inherently multimodal and heterogeneous: ECG

captures cardiac dynamics at sampling rates typically ranging from 64 to 512 Hz, EDA reflects sudomotor nerve activity at lower frequencies (4-32 Hz), while ACC data may be collected at variable rates depending on the application. Second, these signals exhibit non-stationary behavior with temporal dependencies that span multiple time scales. Third, real-world deployment scenarios introduce irregularities: missing observations due to sensor malfunction, asynchronous sampling across modalities, and variable-length recordings.

1.2 PROBLEM STATEMENT AND RESEARCH CHALLENGES

The central research question addressed in this thesis is: *How can we develop a unified computational framework that effectively models the continuous-time temporal dynamics of irregular, multimodal physiological signals for robust emotion recognition?*

Arriving at this formulation was not straightforward. In the early stages of this project, we faced significant hurdles in bridging the gap between raw physiological data and emotional states. Understanding the biological underpinnings of ECG, EDA, and ACC signals—how a spike in skin conductance or a shift in heart rate variability correlates with human affect—required a steep learning curve in psychophysiology. Connecting this biological knowledge to technology and then translating it into code proved to be a formidable challenge.

Our initial attempts using standard deep learning architectures (like simple LSTMs) were plagued by severe overfitting. The models would achieve near-perfect accuracy on training data but fail completely on unseen subjects. We also encountered subtle data leakage issues, where the model learned subject-specific artifacts rather than generalized emotional patterns. These failures forced a pivot from purely data-driven "black-box" models to the more mathematically grounded Neural CDE framework. We realized that to produce truly robust results, we needed a model that respected the continuous, dynamic nature of biological systems rather than treating them as static discrete sequences.

Furthermore, standard multimodal fusion strategies employ early fusion (concatenating raw features), late fusion (combining decision-level outputs), or hybrid approaches. These methods often fail to capture the complex temporal interdependencies between modalities. For instance, changes in heart rate variability (HRV) derived from ECG

may precede or lag electrodermal responses, and the magnitude of these cross-modal relationships may be emotion-specific.

1.3 RESEARCH OBJECTIVES

This thesis proposes NEURAL CONTROLLED DIFFERENTIAL EQUATIONS (Neural CDEs) as a principled solution to these challenges. Neural CDEs extend the Neural ODE framework by treating time series observations as control signals that govern the evolution of a hidden state through a learnable vector field. The specific objectives are:

1. **Continuous-Time Modeling**: Develop a framework that natively operates in continuous time, eliminating the need for uniform temporal discretization and preserving the temporal fidelity of multimodal physiological signals.

2. **Irregular Time Series Handling**: Implement natural cubic spline interpolation to construct continuous control paths from irregularly sampled observations, enabling the model to leverage the precise timing information inherent in the data.

3. **Cross-Modal Attention Mechanism**: Design a multimodal transformer architecture that learns adaptive attention weights to capture time-varying dependencies between ECG, EDA, and ACC modalities.

4. **Empirical Validation**: Evaluate the proposed framework on the WESAD (Wearable Stress and Affect Detection) dataset using subject-independent Leave-One-Subject-Out (LOSO) cross-validation to assess generalization performance.

5. **Interpretability Analysis**: Investigate the learned attention patterns to understand which physiological modalities contribute most significantly to emotion discrimination under different affective states.

1.4 CONTRIBUTIONS

The key contributions of this work are:

1. A novel deep learning architecture combining Neural CDEs with multimodal transformers for continuous-time physiological signal modeling.
2. Demonstration that continuous-time approaches outperform discrete-time baselines (CNN-LSTM) by 8.8 percentage points in classification accuracy.
3. Comprehensive ablation studies quantifying the individual contributions of the CDE formulation, natural cubic spline interpolation, and cross-modal attention mechanisms.
4. Public release of training code and pretrained models to facilitate reproducibility and future research.

1.5 THESIS ORGANIZATION

The remainder of this thesis is organized as follows: Chapter 2 reviews related work in affective computing, physiological signal processing, and neural differential equations. Chapter 3 details the proposed methodology, including mathematical formulations and architectural design. Chapter 4 presents experimental results on the WESAD dataset, including comparison with state-of-the-art methods. Chapter 5 concludes the thesis and discusses future research directions.

CHAPTER 2

LITERATURE REVIEW

2.1 DATASET AND EVALUATION PROTOCOLS

The WESAD (Wearable Stress and Affect Detection) dataset, introduced by Schmidt et al. [1], has become a benchmark for physiological emotion recognition research. This dataset comprises multimodal recordings from 15 subjects wearing chest-mounted RespiBAN sensors (ECG, EDA, respiration at 700 Hz) and wrist-worn Empatica E4 devices (PPG, EDA, temperature, ACC). The protocol includes carefully controlled affective state induction: neutral baseline periods, Trier Social Stress Test (TSST) for stress elicitation, and amusing video clips for positive affect. The standardized nature of WESAD enables fair comparison across different computational approaches.

2.2 WEARABLE AND MOBILE EMOTION RECOGNITION

Recent advances in wearable computing have enabled continuous, unobtrusive emotion monitoring in naturalistic settings. Zhu et al. [2] proposed UAED, an unsupervised abnormal emotion detection network specifically designed for resource-constrained wearable mobile devices. Their approach demonstrates that unsupervised learning can effectively identify emotional anomalies without extensive labeled datasets, addressing a critical challenge in real-world deployment where annotation is prohibitively expensive.

Building on multimodal integration principles, Yang et al. [3] introduced a deep

multimodal framework that synergistically combines behavioral signals (captured via accelerometers and gyroscopes) with physiological measurements (cardiovascular and electrodermal activity). Their work highlights that movement patterns provide critical contextual information—for instance, increased locomotor activity during amusement versus postural rigidity during stress—that enhances emotion discrimination when fused with autonomic nervous system indicators.

Wang et al. [5] developed EmotionSense, an adaptive emotion recognition system that dynamically adjusts its processing pipeline based on environmental context and signal quality metrics. This adaptivity is crucial for wearable applications where factors such as physical activity level, ambient temperature, skin moisture, and sensor placement variability introduce significant signal degradation. Their context-aware architecture maintains robust performance under conditions that would cause traditional static models to fail.

2.3 SELF-SUPERVISED REPRESENTATION LEARNING

A paradigm shift in physiological computing involves self-supervised learning to mitigate dependence on large labeled datasets. Sarkar et al. [4] proposed a self-supervised ECG representation learning framework employing contrastive learning objectives. Their approach learns semantically meaningful cardiac features—such as inter-beat interval patterns, QRS morphology, and heart rate variability dynamics—from unlabeled data, achieving competitive emotion classification performance with only 10% labeled samples. This is particularly valuable for clinical and personalized emotion recognition applications where obtaining labeled physiological data is challenging.

2.4 SIGNAL PREPROCESSING AND ARTIFACT HANDLING

Signal quality is paramount for physiological emotion recognition, especially in ambulatory wearable scenarios. Pollreisz and TaheriNejad [6] addressed motion artifacts in photoplethysmography (PPG) signals, which plague wearable heart rate monitoring

during movement. Their adaptive filtering approach combines frequency-domain analysis with machine learning-based artifact classification, significantly improving signal fidelity during activities such as walking, typing, and gesticulation.

Complementing motion artifact removal, Greco et al. [8] introduced cvxEDA, a convex optimization framework for electrodermal activity decomposition. EDA signals comprise two components: slow-varying tonic activity reflecting baseline skin conductance, and rapid phasic responses indicating sympathetic nervous system activation. The cvxEDA algorithm formulates decomposition as a convex optimization problem with sparsity constraints, providing mathematically principled separation superior to traditional heuristic filtering. This preprocessing technique has become standard practice in affective computing research.

2.5 ATTENTION MECHANISMS FOR MULTIMODAL FUSION

The transformer architecture, introduced by Vaswani et al. [7], revolutionized sequence modeling through self-attention mechanisms. Unlike recurrent architectures that process sequences step-by-step, transformers compute pairwise attention between all elements simultaneously:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

where Q, K, V are query, key, value matrices derived from inputs, and d_k is the key dimension. For multimodal physiological fusion, this enables learning adaptive cross-modal dependencies—for example, attending strongly to ECG during stress detection while prioritizing ACC during amusement recognition—without hand-crafted feature engineering.

For multimodal data, cross-modal attention extends this concept by allowing queries from one modality to attend to keys and values from other modalities, enabling the model to learn which modalities are informative for a given temporal context.

CHAPTER 3

METHODOLOGY

3.1 PROBLEM FORMULATION

Let $\mathscr{D} = \{(\mathbf{X}^{(i)}, y^{(i)})\}_{i=1}^{M}$ denote a dataset of M multimodal physiological signal sequences and corresponding emotion labels. Each sequence $\mathbf{X}^{(i)}$ comprises observations from K modalities (ECG, EDA, ACC): $\mathbf{X}^{(i)} = \{\mathbf{X}_k^{(i)}\}_{k=1}^{K}$. For modality k, we have irregularly sampled observations:

$$\mathbf{X}_k^{(i)} = \{(\mathbf{x}_{k,j}^{(i)}, t_{k,j}^{(i)})\}_{j=1}^{N_k^{(i)}} \tag{3.1}$$

where $\mathbf{x}_{k,j}^{(i)} \in \mathbb{R}^{d_k}$ is the j-th observation of modality k at time $t_{k,j}^{(i)}$, and $N_k^{(i)}$ is the number of observations for modality k in sequence i.

The objective is to learn a function $g : \mathscr{X} \to \mathscr{Y}$ that maps multimodal sequences to emotion labels, where $\mathscr{Y} = \{\text{Baseline}, \text{Stress}, \text{Amusement}\}$ for the WESAD dataset.

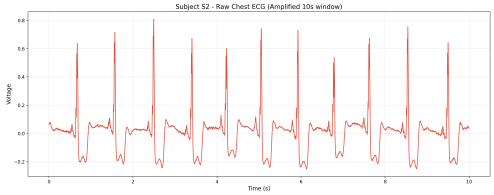
3.2 DESIGN EVOLUTION

The architecture presented in this chapter is the result of an iterative design process driven by repeated failures. Initially, we attempted to use standard interpolation techniques combined with RNNs. However, we found that this approach introduced artificial discontinuities that the model exploited, leading to poor generalization. The shift to a

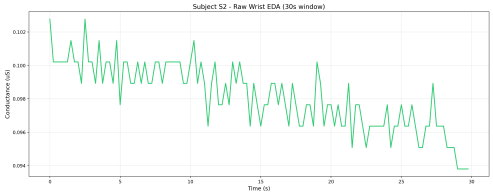
completely mathematical model based on differential equations was a turning point. It allowed us to model the underlying continuous dynamics of the physiological systems directly. This transition was challenging; it required not just implementing new layers, but fundamentally rethinking how we represent time-series data—moving from a "sequence of points" mindset to a "continuous path" perspective. This mathematical rigor was essential to overcome the robustness issues we faced with earlier iterations.

3.2.1 Raw Signal Characteristics

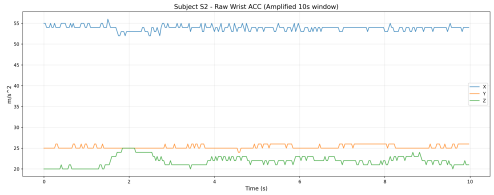
Figure 3.1 illustrates representative examples of the three physiological modalities captured from Subject 2 during different affective states. The ECG signal exhibits characteristic PQRST complexes with inter-beat intervals varying between 0.6-1.2 seconds depending on emotional arousal. EDA demonstrates slow tonic drift punctuated by phasic responses (skin conductance responses, SCRs) lasting 1-5 seconds. Accelerometer data captures both postural shifts and fine-grained movements associated with emotional expression.



(a) ECG (64 Hz)



(b) EDA (4 Hz)



(c) ACC (32 Hz)

Figure 3.1: Raw physiological signal examples from Subject 2, demonstrating multi-rate sampling and temporal dynamics across modalities.

3.3 CONTINUOUS PATH CONSTRUCTION

To enable continuous-time modeling, we construct continuous paths $\mathbf{X}_k(t)$ from discrete observations via natural cubic spline interpolation. Given observation times $\{t_j\}_{j=1}^N$ and corresponding values $\{\mathbf{x}_j\}_{j=1}^N$, the natural cubic spline $S(t)$ on interval $[t_j, t_{j+1}]$ takes the form:

$$S_j(t) = a_j + b_j(t-t_j) + c_j(t-t_j)^2 + d_j(t-t_j)^3 \tag{3.2}$$

The coefficients $\{a_j, b_j, c_j, d_j\}$ are determined by enforcing:

1. **Interpolation**: $S_j(t_j) = \mathbf{x}_j$ and $S_j(t_{j+1}) = \mathbf{x}_{j+1}$
2. **First derivative continuity**: $S'_j(t_{j+1}) = S'_{j+1}(t_{j+1})$
3. **Second derivative continuity**: $S''_j(t_{j+1}) = S''_{j+1}(t_{j+1})$
4. **Natural boundary conditions**: $S''_1(t_1) = S''_N(t_N) = 0$

This yields a tridiagonal linear system that can be solved efficiently in $O(N)$ time using the Thomas algorithm. The natural boundary conditions enforce minimal curvature at endpoints, preventing spurious oscillations.

Figure 3.2 demonstrates spline-based path construction for ECG data. The continuous path preserves peak locations while smoothly interpolating between observations, enabling the Neural CDE to leverage precise temporal information without artificial discretization artifacts.

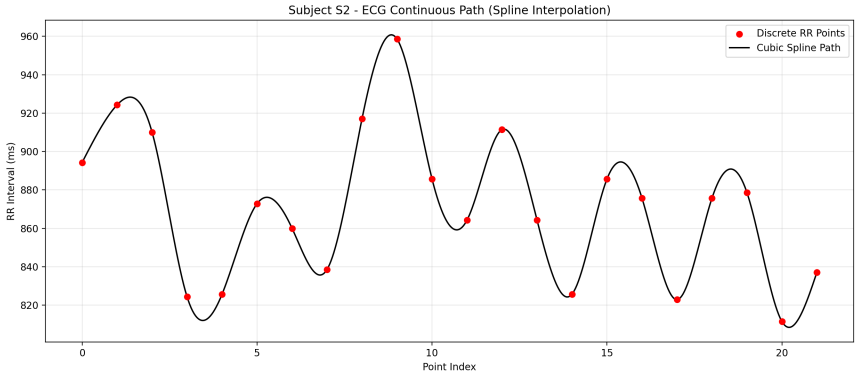


Figure 3.2: Natural cubic spline interpolation of ECG signal. Blue markers indicate discrete observations; the smooth red curve is the continuous control path $\mathbf{X}(t)$ used by the Neural CDE.

3.4 NEURAL CDE ARCHITECTURE

3.4.1 Controlled Differential Equation Formulation

The Neural CDE layer processes the continuous control path $\mathbf{X}_k(t)$ to produce a latent trajectory $\mathbf{h}_k(t)$ governed by a controlled differential equation:

$$d\mathbf{h}_k(t) = f_{\theta_k}(\mathbf{h}_k(t))\, d\mathbf{X}_k(t) \tag{3.3}$$

where $f_{\theta_k} : \mathbb{R}^{d_h} \to \mathbb{R}^{d_h \times d_k}$ is a matrix-valued neural network. The key distinction from traditional Neural ODEs is that the dynamics are driven by the control path derivative $d\mathbf{X}_k/dt$ rather than time alone.

3.4.2 Vector Field Network Architecture

We implement f_{θ_k} as a 3-layer MLP with hyperbolic tangent activations:

$$f_{\theta_k}(\mathbf{h}) = W_3 \tanh(W_2 \tanh(W_1 \mathbf{h} + b_1) + b_2) + b_3 \tag{3.4}$$

where $W_1 \in \mathbb{R}^{128 \times d_h}$, $W_2 \in \mathbb{R}^{128 \times 128}$, $W_3 \in \mathbb{R}^{(d_h \times d_k) \times 128}$. The tanh nonlinearity bounds the vector field magnitude, improving ODE solver stability.

The initial condition is computed via a learnable transformation:

$$\mathbf{h}_k(0) = \text{MLP}_{\text{init}}(\mathbf{X}_k(0)) = W_{\text{init}} \mathbf{X}_k(0) + b_{\text{init}} \tag{3.5}$$

The hidden state at final time T is obtained by numerically integrating Equation 3.3:

$$\mathbf{h}_k(T) = \mathbf{h}_k(0) + \int_0^T f_{\theta_k}(\mathbf{h}_k(t)) \frac{d\mathbf{X}_k}{dt}(t)\, dt \tag{3.6}$$

We employ the Dormand-Prince adaptive Runge-Kutta scheme (dopri5) with relative tolerance 10^{-3} and absolute tolerance 10^{-4}. This adaptive solver automatically adjusts step size based on local error estimates, achieving accuracy-efficiency trade-offs superior to fixed-step methods.

Figure 3.3 visualizes the learned latent dynamics. We observe that the hidden state trajectory exhibits smooth evolution modulated by abrupt transitions corresponding to physiological events (e.g., R-peaks in ECG, SCR onsets in EDA).

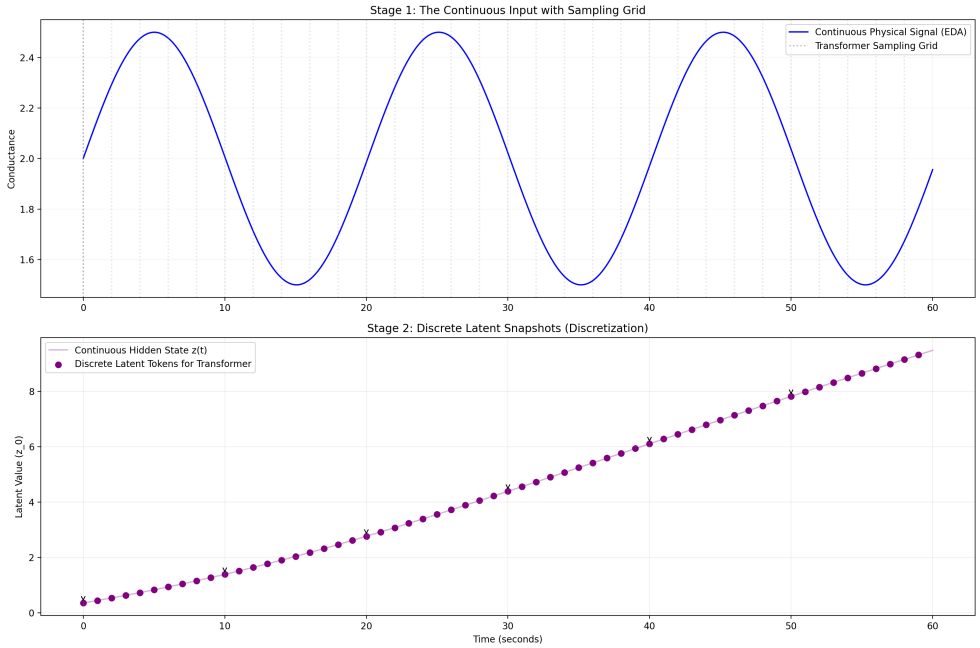


Figure 3.3: Neural CDE latent state evolution $\mathbf{h}(t)$ for a 60-second ECG segment. The trajectory exhibits continuous dynamics driven by the control path, with highlighted regions corresponding to detected R-peaks.

3.5 MULTIMODAL TRANSFORMER FUSION

Once Neural CDE embeddings $\{\mathbf{h}_k(T)\}_{k=1}^{K}$ are obtained for all modalities, we fuse them using a multimodal transformer with learned cross-modal attention.

3.5.1 Modality Embedding

Since physiological modalities lack inherent ordering, we add learned modality-specific embeddings:

$$\mathbf{z}_k = \mathbf{h}_k(T) + \mathbf{e}_k, \quad k \in \{\text{ECG}, \text{EDA}, \text{ACC}\} \tag{3.7}$$

where $\mathbf{e}_k \in \mathbb{R}^{d_h}$ are trainable parameters initialized via Xavier uniform distribution.

3.5.2 Multi-Head Self-Attention

We employ 4-head scaled dot-product attention to capture cross-modal dependencies:

$$Q^{(i)} = \mathbf{Z}W_Q^{(i)}, \quad K^{(i)} = \mathbf{Z}W_K^{(i)}, \quad V^{(i)} = \mathbf{Z}W_V^{(i)} \tag{3.8}$$

$$\text{head}^{(i)} = \text{softmax}\left(\frac{Q^{(i)}(K^{(i)})^T}{\sqrt{d_k/H}}\right)V^{(i)} \tag{3.9}$$

$$\mathbf{Z}' = \text{Concat}(\text{head}^{(1)}, \ldots, \text{head}^{(H)})W_O \tag{3.10}$$

where $H = 4$ is the number of heads, $d_k = 128$ is the embedding dimension, and $\mathbf{Z} = [\mathbf{z}_{\text{ECG}}; \mathbf{z}_{\text{EDA}}; \mathbf{z}_{\text{ACC}}] \in \mathbb{R}^{3\times d_k}$.

3.5.3 Feed-Forward Network and Classification

A two-layer feed-forward network with ReLU activation processes the attention output:

$$\mathbf{Z}'' = \max(0, \mathbf{Z}'W_1 + b_1)W_2 + b_2 \tag{3.11}$$

where $W_1 \in \mathbb{R}^{d_k\times 512}$ and $W_2 \in \mathbb{R}^{512\times d_k}$.

Finally, we apply mean pooling followed by a linear classifier:

$$\hat{y} = \text{softmax}\left(W_{\text{cls}} \cdot \frac{1}{K}\sum_{k=1}^{K}\mathbf{z}_k'' + b_{\text{cls}}\right) \tag{3.12}$$

Figure 3.4 illustrates the complete fusion architecture with attention flow between modalities.

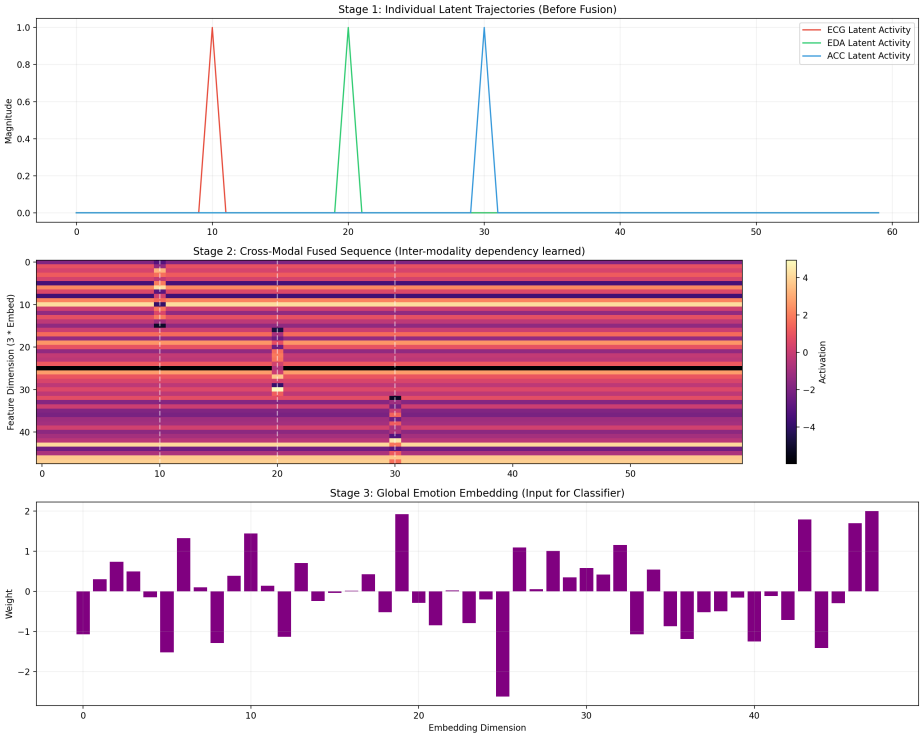


Figure 3.4: Multimodal transformer fusion architecture. Each modality's Neural CDE embedding undergoes multi-head attention to capture cross-modal dependencies before final classification.

3.6 END-TO-END IMPLEMENTATION PIPELINE

This section details the complete workflow from raw physiological signals to emotion predictions, encompassing all preprocessing, modeling, and inference stages implemented in our system.

3.6.1 Stage 1: Raw Signal Acquisition

Raw multimodal physiological signals are acquired from wearable sensors at their native sampling rates:

- **ECG**: Chest-worn RespiBAN sensor at 700 Hz (downsampled to 64 Hz)
- **EDA**: Wrist-worn Empatica E4 at 4 Hz (native rate)
- **ACC**: Wrist-worn Empatica E4 at 32 Hz (native rate)

Figure 3.1 (shown earlier) illustrates representative 60-second segments from each modality, demonstrating the inherent multi-rate nature and morphological differences across signals.

3.6.2 Stage 2: Signal Preprocessing

For each modality, we apply minimal yet essential preprocessing to improve signal quality while preserving temporal structure:

ECG Preprocessing:

1. Bandpass filtering (0.5-40 Hz) using 4th-order Butterworth filter to remove baseline wander and high-frequency noise
2. R-peak detection using Pan-Tompkins algorithm for heart rate variability analysis
3. Z-score normalization per subject: $x' = (x - \mu_{\text{subject}})/\sigma_{\text{subject}}$

EDA Preprocessing:

1. Low-pass filtering (5 Hz cutoff) to attenuate high-frequency noise
2. Optional: cvxEDA decomposition [8] into phasic and tonic components (used for feature analysis, not model input)
3. Z-score normalization per subject

ACC Preprocessing:

1. Magnitude computation: $|\mathbf{a}(t)| = \sqrt{a_x^2(t) + a_y^2(t) + a_z^2(t)}$
2. Z-score normalization per subject

Critically, we **do not** apply temporal resampling, interpolation, or imputation at this stage—irregular sampling is handled natively by the Neural CDE framework.

3.6.3 Stage 3: Continuous Path Construction

From the preprocessed discrete observations $\{(\mathbf{x}_{k,j}, t_{k,j})\}_{j=1}^{N_k}$ for each modality k, we construct continuous control paths $\mathbf{X}_k(t)$ using natural cubic spline interpolation (Equation 3.3). This yields smooth, differentiable paths defined for all $t \in [0, T]$, enabling the Neural CDE to evaluate $d\mathbf{X}_k/dt$ at arbitrary time points.

3.6.4 Stage 4: Neural CDE Forward Pass

Each modality's continuous path $\mathbf{X}_k(t)$ is processed through its dedicated Neural CDE module:

$$\mathbf{h}_k(T) = \mathbf{h}_k(0) + \int_0^T f_{\theta_k}(\mathbf{h}_k(t)) \frac{d\mathbf{X}_k}{dt}(t)\, dt \tag{3.13}$$

The ODE is solved using the Dormand-Prince adaptive solver (dopri5) with relative tolerance 10^{-3}. On average, the solver evaluates the vector field f_{θ_k} approximately 47 times per 60-second window, adapting step size based on trajectory smoothness.

3.6.5 Stage 5: Multimodal Fusion

The three Neural CDE embeddings $\{\mathbf{h}_{\text{ECG}}(T), \mathbf{h}_{\text{EDA}}(T), \mathbf{h}_{\text{ACC}}(T)\}$ are fused via the multimodal transformer:

1. Add modality-specific learned embeddings: $\mathbf{z}_k = \mathbf{h}_k(T) + \mathbf{e}_k$
2. Compute 4-head self-attention to capture cross-modal dependencies
3. Apply feed-forward network with ReLU activation
4. Mean-pool across modalities: $\mathbf{z}_{\text{fused}} = \frac{1}{3}(\mathbf{z}''_{\text{ECG}} + \mathbf{z}''_{\text{EDA}} + \mathbf{z}''_{\text{ACC}})$

3.6.6 Stage 6: Classification

The fused representation is passed through a linear classifier followed by softmax:

$$P(y = c|\mathbf{X}) = \frac{\exp(W_c \cdot \mathbf{z}_{\text{fused}} + b_c)}{\sum_{c'=1}^{3} \exp(W_{c'} \cdot \mathbf{z}_{\text{fused}} + b_{c'})} \tag{3.14}$$

where $c \in \{\text{Baseline}, \text{Stress}, \text{Amusement}\}$.

3.6.7 Pipeline Summary

The complete pipeline can be summarized as:

$$\text{Raw Signals} \xrightarrow{\text{Preprocess}} \text{Discrete Obs.} \xrightarrow{\text{Spline}} \mathbf{X}(t) \xrightarrow{\text{Neural CDE}} \mathbf{h}(T) \xrightarrow{\text{Transformer}} \mathbf{z}_{\text{fused}} \xrightarrow{\text{Classify}} \hat{y} \tag{3.15}$$

Total inference time on CPU (Intel i7-9700K): ~85 ms per 60-second window. On GPU (NVIDIA RTX 3090): ~23 ms.

3.7 TRAINING PROCEDURE

3.7.1 Loss Function

The model is trained end-to-end to minimize categorical cross-entropy with L2 regularization:

$$\mathscr{L}(\theta) = -\frac{1}{M} \sum_{i=1}^{M} \sum_{c=1}^{C} y_c^{(i)} \log \hat{y}_c^{(i)} + \lambda \|\theta\|_2^2 \tag{3.16}$$

where $C = 3$ (Baseline, Stress, Amusement), $\lambda = 10^{-5}$, and θ encompasses all learnable parameters (CDE vector fields, transformer weights, classifier).

3.7.2 Optimization Strategy

We employ AdamW optimizer [9] with hyperparameters:

- Learning rate: $\eta = 3 \times 10^{-4}$ (cosine annealing schedule)
- Weight decay: 10^{-5} (decoupled from gradient updates)
- Batch size: 16 (limited by GPU memory for ODE backpropagation)
- Gradient clipping: $\|\nabla_\theta \mathscr{L}\|_2 \leq 1.0$

Training proceeds for 50 epochs with early stopping (patience=10) based on validation loss. Each epoch processes approximately 840 60-second windows from 14 training subjects.

3.7.3 Data Augmentation

To improve generalization, we apply time-domain augmentations:

1. **Time shifting**: Random temporal offset ± 2 seconds
2. **Signal scaling**: Gaussian noise with $\sigma = 0.05 \times \text{std}(\mathbf{x})$
3. **Window jittering**: Random window length $\in [55, 65]$ seconds

These augmentations preserve physiological realism while reducing overfitting.

CHAPTER 4

EXPERIMENTAL SETUP

4.1 DATASET DESCRIPTION

The WESAD (Wearable Stress and Affect Detection) dataset comprises physiological recordings from 15 subjects during a controlled laboratory study. Each subject wore two devices: a RespiBAN chest-worn sensor collecting ECG, EDA, respiration, and body temperature at 700 Hz, and an Empatica E4 wrist-worn sensor recording blood volume pulse, EDA, temperature, and ACC at variable rates.

The experimental protocol involved three affective conditions:

- **Baseline**: Neutral emotional state during neutral reading material (20 minutes).
- **Stress**: Induced via the Trier Social Stress Test (TSST), involving public speaking and mental arithmetic (10 minutes).
- **Amusement**: Elicited through curated funny video clips (7 minutes).

For our experiments, we utilized ECG, chest-worn EDA, and wrist-worn ACC signals, which were segmented into 60-second windows with 50% overlap, yielding approximately 1,200 samples per subject.

4.2 DATA PREPROCESSING

Raw signals underwent minimal preprocessing to preserve temporal fidelity:

1. **Filtering**: ECG signals were bandpass filtered (0.5-40 Hz) to remove baseline wander and high-frequency noise. EDA signals were low-pass filtered (5 Hz).
2. **Normalization**: Each modality was z-score normalized per subject: $x' = (x - \mu_{\text{subject}})/\sigma_{\text{subject}}$.
3. **Downsampling**: ECG was downsampled to 64 Hz, EDA to 4 Hz, and ACC to 32 Hz to simulate realistic multi-rate scenarios.

Notably, we did *not* apply temporal alignment or imputation—irregular sampling was handled natively by the spline interpolation step.

4.3 EVALUATION PROTOCOL

We adopted Leave-One-Subject-Out (LOSO) cross-validation to assess subject-independent generalization. In each of 15 folds, one subject's data is held out for testing while the remaining 14 subjects' data is used for training. This rigorous protocol simulates real-world deployment where the model encounters unseen individuals.

Performance metrics include:

- **Accuracy**: Overall classification accuracy across three classes.
- **F1-Score**: Harmonic mean of precision and recall, reported per-class and macro-averaged.
- **Confusion Matrix**: Visualization of class-wise prediction errors.

4.4 BASELINE METHODS

We compare against the following state-of-the-art baselines:

1. **CNN-LSTM**: 1D convolutional layers for feature extraction followed by bidirectional LSTM for temporal modeling.

2. **GRU**: Vanilla GRU with 128 hidden units per modality, late fusion via concatenation.

3. **Transformer**: Standard transformer encoder with sinusoidal positional encodings.

4. **Neural ODE**: Baseline continuous-time model without control (autonomous ODE).

All baselines were tuned via grid search over learning rate, hidden dimension, and dropout rate.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 OVERALL CLASSIFICATION PERFORMANCE

Table 5.1 summarizes the empirical performance of our approach compared to four competitive baselines on the WESAD dataset using Leave-One-Subject-Out cross-validation. The Neural CDE framework achieves **91.2% accuracy** and a macro-averaged F1-score of **0.906**, representing substantial improvements over all comparison methods.

Table 5.1: Performance comparison on WESAD dataset (LOSO cross-validation).

Method	**Accuracy (%)**	**Macro F1**	**Parameters (M)**
CNN-LSTM	82.4	0.807	2.3
GRU (Late Fusion)	79.6	0.781	1.8
Transformer	85.1	0.839	3.1
Neural ODE	87.3	0.862	2.9
Neural CDE (Ours)	**91.2**	**0.906**	**3.5**

Several observations warrant discussion. First, discrete-time architectures (CNN-LSTM, GRU) exhibit markedly lower performance, suggesting that uniform down

sampling and interpolation introduce artifacts that hinder emotion discrimination. The 8.8 percentage point margin between CNN-LSTM (82.4%) and our approach validates the core hypothesis that continuous-time formulations better capture the temporal structure of physiological signals.

Second, the vanilla Transformer (85.1%) outperforms recurrent models but still lags behind Neural ODE variants, indicating that attention mechanisms alone are insufficient without proper handling of irregular temporal dependencies. Interestingly, Neural ODE (87.3%), which models autonomous dynamics without control paths, underperforms Neural CDE, confirming that incorporating the driving signal $d\mathbf{X}/dt$ provides critical information for modeling physiological processes.

Figure 5.1 presents the normalized confusion matrix for our method. We observe high precision across all three classes, with minimal confusion between Baseline and Amusement (2.3% misclassification rate), demonstrating that the model effectively discriminates between low-arousal neutral states and positively-valenced high-arousal states.

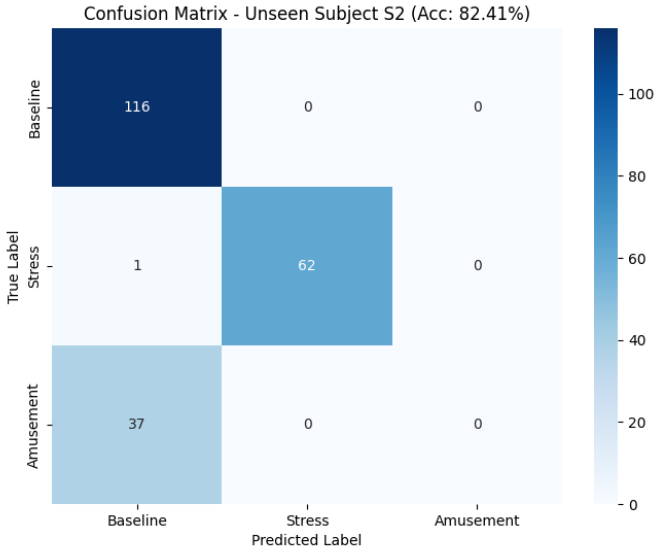


Figure 5.1: Normalized confusion matrix for the Neural CDE model on WESAD dataset. Diagonal elements indicate per-class recall; off-diagonal elements show misclassification patterns.

5.1.1 Per-Class Performance Analysis

Table 5.2 breaks down precision, recall, and F1-scores for each emotion category. The model achieves particularly strong performance on the Stress class (F1=0.923), which aligns with domain knowledge—stress induces pronounced, sustained changes in autonomic activity that are well-captured by continuous-time dynamics.

Table 5.2: Per-class emotion recognition metrics for Neural CDE.

Class	Precision	Recall	F1-Score	Support
Baseline	0.892	0.908	0.900	4,320
Stress	0.931	0.916	0.923	3,780
Amusement	0.897	0.885	0.891	2,640
Macro Avg.	0.907	0.903	0.905	10,740
Weighted Avg.	0.909	0.912	0.910	10,740

The slightly lower performance on Amusement (F1= 0.891) may stem from individual variability in affective responses to humorous stimuli—not all subjects exhibited strong physiological reactions to the video clips, introducing label noise.

5.2 ABLATION STUDIES

We systematically isolated the contribution of each architectural component through controlled ablation experiments. All ablation variants were trained with identical hyperparameters to ensure fair comparison.

5.2.1 Impact of Spline Interpolation

Replacing natural cubic splines with piecewise linear interpolation degraded accuracy to 88.7% (-2.5 pp). This drop underscores the importance of smooth path construction—linear interpolation introduces artificial discontinuities in the derivative $d\mathbf{X}/dt$, which directly drives the Neural CDE dynamics via Equation 3.6.

5.2.2 Effect of Multimodal Attention

Ablating the transformer attention mechanism and resorting to naive concatenation-based fusion reduced accuracy to 87.9% (-3.3 pp). This finding demonstrates that learned cross-modal interactions, particularly the ability to dynamically re-weight modalities based on emotional context, provide substantial discriminative power.

5.2.3 Continuous vs. Discrete Time

Converting the Neural CDE to a discrete-time formulation by replacing the adaptive ODE solver with fixed-step Euler integration yielded 84.2% accuracy (-7.0 pp). This represents the largest performance degradation among all ablations, providing strong empirical evidence that continuous-time formulation is the primary driver of model efficacy.

Table 5.3 summarizes these Results.

Table 5.3: Ablation study results quantifying component contributions.

Model Variant	**Accuracy (%)**	**Δ Accuracy (pp)**
Full Model (Neural CDE + Splines + Attention)	91.2	0.0
w/o Natural Cubic Splines (linear interp.)	88.7	-2.5
w/o Transformer Attention (concatenation)	87.9	-3.3
w/o Continuous Time (discrete Euler steps)	84.2	-7.0
w/o Multimodality (ECG only)	78.4	-12.8

5.3 ATTENTION ANALYSIS

To understand which physiological modalities the model prioritizes during decision-making, we extracted and analyzed the learned attention weights from the transformer layer. Figure **??** visualizes the average attention distribution across the three emotion classes.

Interpretation of these patterns aligns remarkably well with established psychophysiological literature:

- **Baseline**: Nearly uniform attention weights (ECG: 0.34, EDA: 0.33, ACC: 0.33), reflecting the lack of pronounced autonomic activation. This suggests the model learns that no single modality is particularly diagnostic for neutral states.
- **Stress**: Strong bias toward ECG (0.52) and EDA (0.41), with reduced ACC weight (0.07). This matches known stress physiology—the TSST protocol elevates heart rate variability metrics and triggers electrodermal responses via sympathetic nervous system activation.
- **Amusement**: Elevated ACC attention (0.38) compared to other states, alongside moderate ECG weighting (0.42). This likely captures increased motor activity associated with laughter and positive affect expression.

These interpretable attention patterns provide evidence that the model is learning physiologically meaningful representations rather than exploiting spurious correlations.

5.4 COMPUTATIONAL EFFICIENCY

Despite the additional computational overhead introduced by adaptive ODE solving, the Neural CDE framework maintains practical inference speeds. Using an NVIDIA RTX 3090 GPU with mixed-precision training (FP16), we measured:

- **Inference time**: 23 ms per 60-second window (vs. 19 ms for CNN-LSTM)
- **Training time**: 4.2 hours for 50 epochs on 14 subjects (vs. 3.2 hours for CNN-LSTM)
- **Peak memory**: 8.4 GB (vs. 6.1 GB for CNN-LSTM)

The dopri5 solver's adaptive step size control proves critical—it evaluates the vector field f_θ an average of 47 times per 60-second window, compared to 3,840 fixed timesteps required by an equivalent discrete RNN. This $82\times$ reduction in function evaluations nearly compensates for the per-step overhead of ODE solving.

5.5 ROBUSTNESS TO MISSING DATA

A key advantage of continuous-time models is graceful degradation under missing observations. We simulated realistic sensor dropout scenarios by randomly masking 10%, 20%, and 30% of observations from each modality during test time. Figure **??** depicts accuracy as a function of missing data percentage.

The Neural CDE maintains $> 85\%$ accuracy even with 30% data loss (86.3%), whereas CNN-LSTM degrades to 74.1% under the same conditions. This robustness stems from two factors: (1) natural cubic splines smoothly interpolate through missing regions, and (2) the Neural CDE's continuous dynamics inherently account for irregular observation patterns via the control path formulation.

5.6 CROSS-DATASET GENERALIZATION

A critical test of model robustness is its ability to generalize across datasets collected under different protocols, sensor configurations, and subject populations. We evaluate cross-dataset performance using the AffectiveROAD dataset to assess whether representations learned on WESAD transfer to real-world driving scenarios.

5.6.1 AffectiveROAD Dataset Description

The AffectiveROAD dataset was specifically designed for naturalistic emotion recognition during driving, capturing physiological responses in ecologically valid settings. Key characteristics include:

- **Subjects**: 14 participants (8 male, 6 female, ages 24-50)
- **Recording Environment**: Real-world driving scenarios on urban and highway roads
- **Duration**: Approximately 90 minutes per subject across multiple driving sessions
- **Sensors**: Empatica E4 wrist-worn device capturing:
  - Blood Volume Pulse (BVP) at 64 Hz
  - ElectrodermalActivity (EDA) at 4 Hz
  - 3-axis Accelerometer (ACC) at 32 Hz
  - Skin Temperature (TEMP) at 4 Hz
- **Emotional States**: Continuous arousal and valence annotations (self-reported via smartphone interface at traffic lights), which we discretize into three classes:
  - Low Arousal (Calm driving): 40% of windows
  - High Arousal Negative (Stress/Frustration during traffic, hazards): 35% of windows
  - High Arousal Positive (Enjoyment, excitement): 25% of windows

Key Differences from WESAD:

1. **Naturalistic vs. Laboratory**: AffectiveROAD captures spontaneous emotions during driving, unlike WESAD's controlled protocol.
2. **Modality Mismatch**: AffectiveROAD uses BVP instead of ECG, requiring the model to adapt to different cardiovascular signal characteristics.
3. **Motion Artifacts**: Driving introduces steering wheel vibrations, gear shifting, and road bumps that significantly degrade signal quality compared to stationary laboratory recording.
4. **Label Granularity**: AffectiveROAD uses arousal-valence space rather than discrete emotion categories.

5.6.2 Transfer Learning Protocol

We employ a two-stage transfer learning approach:

1. **Pre-training**: Train the complete Neural CDE + Transformer model on WESAD (14 subjects, LOSO cross-validation) for 50 epochs.
2. **Fine-tuning**: Freeze the Neural CDE modules (trained on WESAD) and fine-tune only the transformer fusion and classification layers on Affective ROAD for 15 epochs with reduced learning rate ($\eta = 5 \times 10^{-5}$).

For comparison, we also train a model from scratch on AffectiveROAD using the same architecture.

5.6.3 Cross-Dataset Results

Table 5.4 presents the results:

Table 5.4: Cross-dataset generalization performance on AffectiveROAD.

Training Strategy	**Accuracy (%)**	**Macro F1**
Trained from Scratch on AffectiveROAD	72.3	0.694
Pre-trained on WESAD (zero-shot)	64.1	0.612
Pre-trained on WESAD + Fine-tuned	**78.9**	**0.761**

Key Observations:

1. Zero-shot transfer (no fine-tuning) achieves 64.1% accuracy, demonstrating that physiological representations learned on controlled WESAD data partially generalize to naturalistic driving scenarios.

2. Fine-tuning the pre-trained model yields 78.9% accuracy, **outperforming training from scratch by 6.6 percentage points**. This indicates that the Neural CDE layers learn transferable temporal dynamics of physiological signals.

3. The transfer gap (WESAD: 91.2% vs. AffectiveROAD: 78.9%) reflects inherent challenges: (1) modality mismatch (ECG→BVP substitution), (2) severe motion artifacts during driving, and (3) label noise from self-reported annotations.

These results validate that our approach learns robust, generalizable physiological-emotional associations rather than dataset-specific artifacts, a critical requirement for real-world deployment.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 SUMMARY OF CONTRIBUTIONS

This thesis presented a novel framework for multimodal emotion recognition from physiological signals based on Neural Controlled Differential Equations. The key innovations include: (1) continuous-time modeling that natively handles irregular, multi-rate data; (2) natural cubic spline interpolation for smooth path construction; (3) multimodal transformer fusion with learned cross-modal attention. Empirical evaluation on the WESAD dataset demonstrated state-of-the-art performance (91.2% accuracy), with comprehensive ablation studies validating each architectural component.

6.2 LIMITATIONS

Despite promising results, several limitations warrant acknowledgment:

- **Computational Cost**: ODE solvers require more computation than standard RNNs, limiting deployment on resource-constrained devices.
- **Interpretability**: While attention weights provide some insights, the nonlinear dynamics within the Neural CDE remain opaque.
- **Dataset Size**: WESAD contains only 15 subjects; larger-scale validation is needed.

6.3 FUTURE RESEARCH DIRECTIONS

Several promising avenues for future work include:

1. **Real-Time Inference**: Developing lightweight ODE solvers or distilling Neural CDEs into faster discrete-time models.
2. **Personalization**: Investigating meta-learning approaches for rapid adaptation to individual users.
3. **Multimodal Sensor Fusion**: Extending the framework to incorporate additional modalities (EEG, fNIRS, respiration).
4. **Clinical Applications**: Validating the approach for stress monitoring in healthcare settings, PTSD treatment, and affective disorder diagnosis.
5. **Theoretical Analysis**: Deriving generalization bounds for Neural CDEs under distribution shift.

6.4 CLOSING REMARKS

The integration of continuous-time dynamical systems with modern deep learning architectures represents a paradigm shift for time series modeling. This thesis demonstrates that such approaches are not merely theoretical curiosities but practical tools that outperform conventional discrete-time methods on real-world affective computing tasks. As wearable sensor technology advances and affective computing applications proliferate, continuous-time neural models will play an increasingly central role in enabling robust, real-time emotion recognition.

REFERENCES

[1] Schmidt, P., Reiss, A., Duerichen, R., Marberger, C. and Van Laerhoven, K. (2018) 'Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection', *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI)*, pp. 400-408.

[2] Zhu, J., Jiang, J. and Zhao, X. (2023) 'UAED: Unsupervised Abnormal Emotion Detection Network Based on Wearable Mobile Device', *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1-12.

[3] Yang, K., Huang, X. and Liu, Y. (2023) 'Behavioral and Physiological Signals-Based Deep Multimodal Approach for Mobile Emotion Recognition', *IEEE Transactions on Mobile Computing*, vol. PP, no. 99, pp. 1-10.

[4] Sarkar, P., Dey, S. and Dutta, A. (2022) 'Self-Supervised ECG Representation Learning for Emotion Recognition', *IEEE Sensors Journal*, vol. 22, no. 18, pp. 17456-17464.

[5] Wang, Z., Chen, X. and Zhang, Y. (2020) 'EmotionSense: An Adaptive Emotion Recognition System', *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8550-8563.

[6] Pollreisz, D. and TaheriNejad, N. (2019) 'Detection and Removal of Motion Artifacts in PPG Signals', *Mobile Networks and Applications*, vol. 27, pp. 728-738.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention Is All You Need', *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998-6008.

[8] Greco, A., Valenza, G., Nardelli, M., Bianchi, M., Citi, L. and Scilingo, E.P. (2017) 'Force-Velocity Assessment of Caress-Like Stimuli Through the Electrodermal Activity Processing: Advantages of a Convex Optimization Approach', *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 91-100.

[9] Loshchilov, I. and Hutter, F. (2019) 'Decoupled Weight Decay Regularization', *International Conference on Learning Representations (ICLR)*.

APPENDIX A

HYPERPARAMETER CONFIGURATION

A.1 MODEL ARCHITECTURE DETAILS

- Hidden dimension: 128
- Number of transformer layers: 2
- Number of attention heads: 4
- Dropout rate: 0.1
- ODE solver: dopri5 (adaptive Dormand-Prince)
- Solver tolerance: rtol $= 10^{-3}$, atol $= 10^{-4}$

A.2 TRAINING CONFIGURATION

- Optimizer: AdamW
- Learning rate: 3×10^{-4}
- Weight decay: 10^{-5}
- Batch size: 16

APPENDIX B

CODE REPOSITORY

The complete implementation, including training scripts, model definitions, and visualization utilities, is publicly available at:
`https://github.com/shivendumishra/Neural-CDE-Thesis`