AEFORMER: A LIGHTWEIGHT CONV1D-TRANSFORMER MODEL FOR ACOUSTIC EMISSION BASED REAL-TIME STRUCTURAL HEALTH MONITORING

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

by

AMAN KUMAR GAUR

(24/CSE/20)

Under the Supervision of

Dr. Nipun Bansal

Assistant Professor, Department of CSE

Delhi Technological University

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

May 2025

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042. India

CANDIDATE'S DECLARATION

I Aman Kumar Gaur hereby certify that the work which is being presented in the thesis entitled **"AEFormer: A Lightweight Conv1D-Transformer Model for Acoustic Emission Based Real-Time Structural Health Monitoring"** in partial fulfillment of the requirements for the award of the Master of Technology Degree, submitted in the Department of Computer Science and Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from [Start Date, e.g., August 2025] to May 2026 under the supervision of **Dr. Nipun Bansal.**

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor(s) **Signature of External Examiner**

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042. India

CERTIFICATE

I hereby Certified that **Aman Kumar Gaur (24/CSE/20)** has carried out their research work presented in this thesis entitled **"AEFormer: A Lightweight Conv1D-Transformer Model for Acoustic Emission Based Real-Time Structural Health Monitoring"** for the award of Master of Technology from Department of Computer Science and Engineering, Delhi Technological University, Delhi, under my supervision.

The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Place: Delhi

Date:

Dr. Nipun Bansal
Assistant Professor
Delhi Technological University

ABSTRACT

Keywords— Acoustic Emission (AE), Structural Health Monitoring (SHM), Conv1D, Transformer Encoder, Deep Learning, Signal Classification, Concrete Damage Detection.

Structural Health Monitoring (SHM) plays a critical role in ensuring the long-term safety and sustainability of civil infrastructure, including bridges, tunnels, dams, and reinforced concrete systems. Among various non-destructive evaluation (NDE) techniques, Acoustic Emission (AE) sensing has emerged as a reliable approach for capturing micro-crack propagation and energy release during stress-induced fracture. However, deploying deep learning models for real-time AE classification on embedded SHM platforms remains challenging due to hardware limitations and computational constraints.

To address this challenge, we propose **AEFormer**, a lightweight hybrid deep learning architecture that integrates **1D Convolutional Neural Networks (Conv1D)** for local feature extraction with **Transformer encoders** to capture long-range temporal dependencies. Unlike conventional CNN and Tiny ANN approaches, AEFormer is explicitly optimized for edge-based deployment, achieving high predictive accuracy while maintaining a compact computational footprint. Experiments were performed on a benchmark dataset comprising **15,000 AE signals sampled at 5 MHz**, representing tensile, shear, and mixed-mode cracking. AEFormer achieves **99.82% test accuracy** with per-class F1-scores above **0.998**, outperforming lightweight CNN and TinyML baselines with fewer than **28,000 trainable parameters**.

The results demonstrate that AEFormer provides a highly dependable and efficient solution for real-time, embedded SHM applications, offering strong potential for deployment in safety-critical monitoring of concrete infrastructure.

ACKNOWLEDGEMENT

The successful completion of any task is incomplete and meaningless without giving any due credit to the people who made it possible without which the project would not have been successful and would have existed in theory.

First and foremost, I am grateful to **Prof. Anil Singh Parihar**, HOD, Department of Computer Science and Engineering, Delhi Technological University, and all other faculty members of our department for their constant guidance and support, constant motivation and sincere support and gratitude for this project work. I owe a lot of thanks to my supervisor, **Dr. Nipun Bansal**, assistant Professor, Department of Computer Science and Engineering, Delhi Technological University for igniting and constantly motivating us and guiding us in the idea of a creatively and amazingly performed Major Project in undertaking this endeavor and challenge and also for being there whenever i needed his guidance or assistance.

I would also like to take this moment to show my thanks and gratitude to one and all, who indirectly or directly have given me their hand in this challenging task. I feel happy and joyful and content in expressing my vote of thanks to all those who have helped me and guided me in presenting this project work for my Major project. Last, but never least, I thank my well-wishers and parents for always being with me, in every sense and constantly supporting me in every possible sense whenever possible.

Aman Kumar Gaur

(24/CSE/20)

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Background

The backbone of modern civilisation, and virtually all of our modern infrastructure (bridges, roads, dams, etc.), is made from concrete. While concrete is generally durable, it remains susceptible to degradation due to environmental exposure, mechanical loading, and internal defects. Most importantly, cracks can begin to form at depths in a structure that are impossible to detect and continue to grow until they become a risk of catastrophic failure, financial loss, and loss of human life.

Visual assessments for conventional inspections have many limitations due to the heavy reliance upon manual assessment techniques by operators.

Many conventional inspection practices are also labour-intensive and are often highly subjective, based on the operator's experience. Due to safety concerns, it isn't easy to inspect specific areas of a structure (e.g., underwater structures) with sufficient safety to enable a practical assessment of the structure.

While a visual assessment can detect all cracks that are clearly evident on the surface of a structure, many subsurface micro-cracks will continue to grow and propagate over time without being detected. They will remain undetected until a crack grows large enough to appear on the surface of the structure, whereupon the potential for severe structural damage may have already occurred.

Since inspections are typically conducted at predetermined intervals, there is limited opportunity for continuous monitoring of a structure's condition. Therefore, if a structure

experiences damage during an interval between inspection periods, there is a high probability that this damage will not be detected before it causes structural failure.
Experts must often interpret data collected through the use of monitoring systems; therefore, the need for experienced personnel to evaluate data can increase operational costs and increase the risk of human error in assessing and making decisions regarding the evaluation of a structure's condition.

1.2 Acoustic Emission in Structural Health Monitoring

Continuous, non-invasive Acoustic Emissions (AE) real-time monitoring offers an alternative to the standard inspection methods based on periodic inspections. The AE system does not inspect the structure at fixed intervals, but continuously uses active sensors to assess the structural integrity of the structure throughout the monitoring time frame. Stress elastic waves are generated when a micro-crack develops or expands in the concrete. These waves travel through the concrete and are captured by surface-mounted piezoelectric sensors, which generate electrical signals from the vibrations produced as shown in figure 1.1.

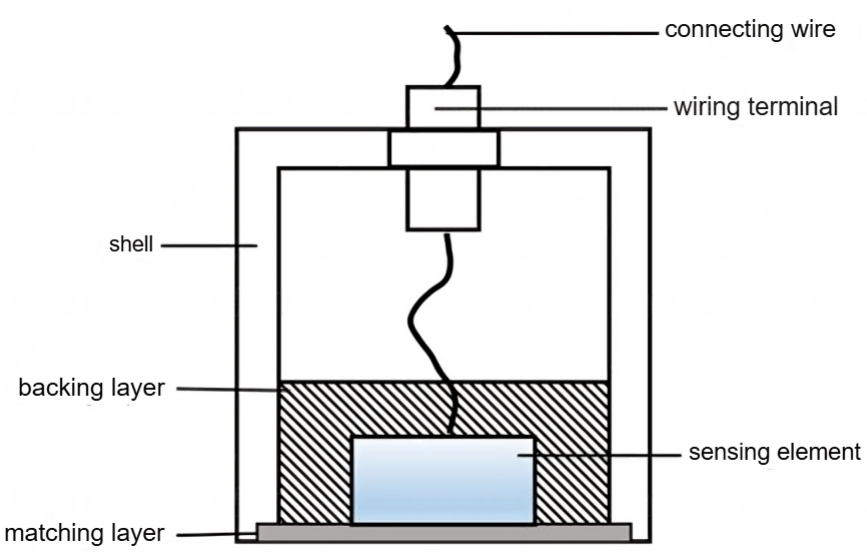


Figure 1.1: *Overview of AE Sensor Architecture and Main Functional Parts*

The characteristics present in each recorded AE event (i.e., amplitude, rise time, duration, total energy) provide indications of the magnitude or type of damage that occurred, while the frequency content within an AE signal can be linked directly to the fracture

mechanism. Higher frequency waveforms generally characterise tensile fractures, while shear-related failures will have a greater proportion of low frequency components in their respective waveforms as shown in figure 1.2. In addition to the early detection of damage using AE techniques, AE also enables the determination of the crack mode causing the damage [1].

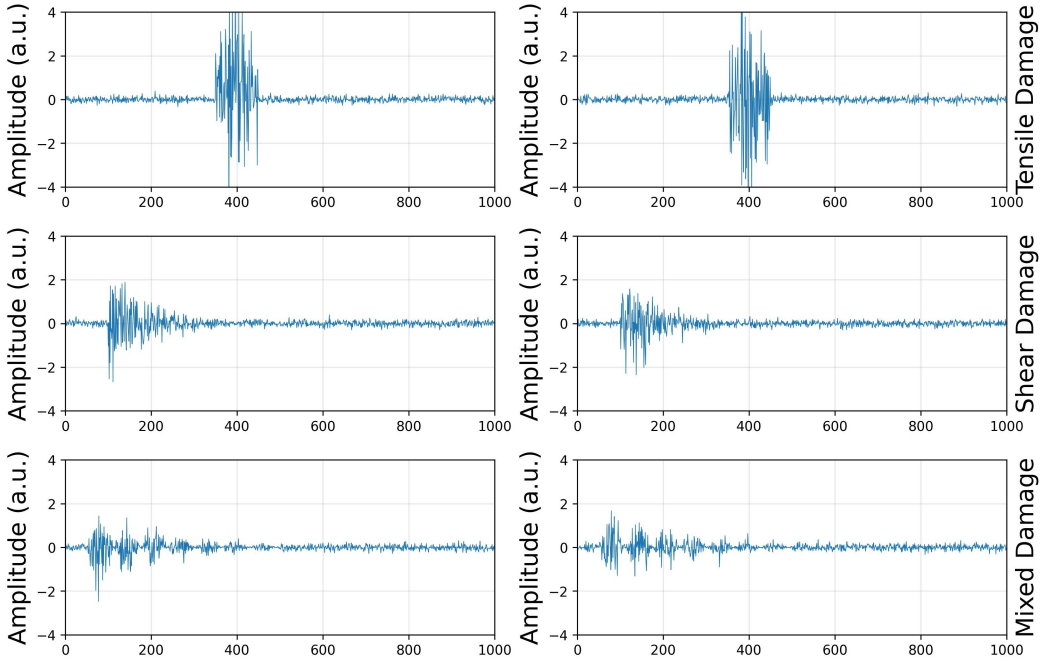


Figure 1.2: *Sample acoustic emission signals corresponding to each damage type*

The machine learning-based acoustic emission (AE) analysis enables the automated interpretation of AE signals induced by damage, eliminating the need for continuous expert oversight. The trained classifier identifies AE pattern characteristics, allowing the monitoring system to clearly distinguish between multiple different types of damage mechanisms (i.e., tensile crack, shear failure, and mixed-mode failure). Once structural deterioration is identified, the system will generate an early warning system to indicate when failure may occur, enabling proactive maintenance and safe intervention [2].

In practical terms, ML-enabled AE monitoring offers several advantages:

- Real-time assessment of structural health with no requirement for an on-site inspection to assess a structure's current condition at any point in time.
- Automated determination of crack type (tensile, shear, etc.) based upon signals generated by the structure during a structural test or other testing process.

- Early detection of structural degradation that could lead to total system failure provides an opportunity for corrective action before the system fails.
- Quantification of the structural condition using machine learning algorithms as a function of the signal characteristics and damage indices used to define structural condition [3].

1.3 Problem

There is a tradeoff between performance and model complexity in deep learning methods for classifying AE signals. On the one hand, lightweight network architectures (e.g., small CNNs or ANNs) can run on low-resource embedded systems, but often at the cost of lower-than-optimal classification accuracy and/or loss of sensitivity to longer-range patterns within the AE signal itself. For example, a recent study demonstrated that a small CNN achieved approximately 98.7% test accuracy, and a tiny ANN achieved 98.4%. Although these results may seem high, they represent an unacceptable reduction in detection accuracy for safety-related AE monitoring applications.

Conversely, models based on transformer structures utilise attention mechanisms to identify longer-range patterns throughout the AE signal and can achieve higher accuracy than traditional CNN/ANN architectures. Unfortunately, these models typically require tens or hundreds of thousands of model parameters, which necessitates significantly more computational resources and memory. Overall, CNN/ANN approaches are generally more computationally efficient, but they may limit the achievable accuracy. In contrast, transformer approaches offer potentially high accuracy but are often quite large in size. Hybrid models that combine the ability of CNNs to extract features using convolutional layers with attention layers to identify longer-range patterns could provide the best of both worlds. However, such hybrid models have rarely, if ever, been developed and implemented for specific AE signals.

Therefore, the primary issue is developing a model that achieves extremely high classification accuracy on AE signals and is lightweight enough to enable real-time processing on edge devices.

1.4 Research Gap

Combining these aspects clearly reveals there is an apparent disparity in requirements. Real SHM systems require models that are sufficiently small to be deployed on edge devices (typically with less than 10MB of memory, less than 1W of power, and typically with less than 100ms of latency), and have exceptionally high accuracy levels (approximately 99% or greater). The majority of current AE signal models meet one of the two criteria, but generally fail to satisfy both. Additionally, nearly all prior research employed a hybrid CNN-Transformer architecture that was not designed to support AE data at such resource limitations. Therefore, this understanding prompted the creation of AEFormer, a lightweight hybrid model combining the efficiencies of convolutional techniques with the accuracy of transformer-based techniques.

1.5 Objectives

The main objectives of this work are as follows:

- The aim is to create a model that achieves a classification rate of over **99%** for unseen AE testing; additionally, the objective is to achieve an F1 score of at least **0.99** per class for all modes.
- In addition to achieving high classification accuracy across all three crack modes, a key objective is to design a compact and computationally efficient model containing no more than **30,000 trainable parameters**. The model should operate with an inference time of less than 100 ms and require approximately 10 MB of memory, enabling deployment on low-power microcontroller platforms operating at less than 1 W of power.
- Lastly, the last objective is to produce a model that is clearly and easily reproducible and understandable. Each part of the architecture of the model (i.e. Conv1d Layers, Attention Blocks, etc.) should be easily identifiable by name and purpose so that the operation of the model can be understood and interpreted.

1.6 Main Contributions

This thesis makes the following main contributions:

- A new lightweight hybrid deep learning network, called AEFormer, is proposed for acoustic emission structural health monitoring.
- The proposed model is composed of Conv1D to extract local features and Transformer encoders to extract long-range temporal dependencies in AE waveforms.
- In terms of accuracy, on the test set AEFormer achieves a result of 99.82%, and has fewer than 28,000 trainable parameters, suitable for edge deployment.
- The extensive experiments and comparative evaluations show that AEFormer is more accurate, robust and generalizable than the two baseline models, namely lightweight CNN and Tiny ANN.
- An ablation study is performed to compare the advantage of using the convolutional module, the attention module, and the final classification accuracy.

Chapter 2

LITERATURE REVIEW

2.1 Acoustic Emission Signal Characteristics

The foundation of Acoustic Emission (AE) monitoring lies in capturing the elastic waves generated by micro-cracks that form or propagate within concrete under stress. When a crack initiates or grows, it releases stored elastic energy in the form of ultrasonic stress waves. These waves are detected by surface-mounted piezoelectric sensors, which convert the resulting mechanical vibrations into electrical signals for further analysis.

The characteristics of each AE signal provide important information about the damage mechanism. The amplitude reflects the magnitude of the cracking event, the rise time (time required to reach peak amplitude) indicates the rate of crack propagation, and the total energy of the signal corresponds to the severity of the fracture. The frequency content also serves as a valuable indicator of crack type. In particular, the fracture mode determines the spectral signature of the AE waveform: *mode I (tensile)* cracks typically produce sharp, high-frequency pulses, whereas *mode II (shear)* cracks generate lower-frequency, longer-duration signals.

To accurately capture these signal variations, modern AE sensors are engineered to operate in the ultrasonic range (approximately 100 kHz to 1 MHz) with rapid response times on the order of 100–500 μs, enabling reliable detection of transient cracking phenomena within concrete structures.

2.2 Conventional AE Analysis

The previous conventional approaches to AE analysis employed manually developed AE signal descriptor techniques that relied on expert analysis of AE signals. Initial research successfully demonstrated that utilising the statistical properties of AE data can lead to high accuracy rates when classifying AE data. For example, Siracusano *et al.* (2021) identified statistically descriptive features of AE event data and successfully classified the mode of cracking using a traditional classifier type, achieving a classification success rate of over 95 per cent [4]. Therefore, it should be possible to achieve high performance in crack detection if AE feature descriptors are appropriately developed and optimised.

The majority of existing methods are widely applicable but challenging to implement practically because they require a significant amount of knowledge from the field of study to identify or determine features. Furthermore, the performance of these methods degrades when the intrinsic signal characteristics of the data on which they are based change. Practically, many existing methods have been implemented using threshold values defined by experts to create decision rules that determine whether a measured quantity represents an anomalous condition or not. Therefore, most of them are limited in terms of being applicable in different structural environments with varying loading conditions and sensor configurations.

2.3 Deep Learning Methods

The basis for Acoustic Emission (AE) monitoring is that the elastic waves emitted from micro-cracks occurring in concrete are captured and monitored. Micro-cracks can occur when a crack propagates or initiates within concrete under stress, releasing the elastic energy stored in the material as ultrasonic stress waves. Piezoelectric sensors are used to monitor the AE signals produced by the cracking events and convert the mechanical vibrations into electrical signals. The characteristics of the AE signal include: the amplitude of the signal indicates the size of the cracking event; the time taken for the signal to reach its maximum (rise time) indicates the rate at which the crack propagates; and the total energy contained within the signal indicates the severity of the crack [5]. CNNs have become one of the most popular architectures for deep learning in AE signal classification due to their ability to learn features directly from raw AE signals. The primary reason

CNNs are so effective for capturing features of AE signals is their ability to process local temporal patterns efficiently. For instance, a recent study published a lightweight 1-D CNN model (approximately 20,000 parameters) and reported that it could achieve approximately 98.7% test accuracy using a well-known benchmark AE dataset [6].

These CNNs process the AE signal as follows: they slide a convolutional filter (kernel) and a pooling operation over the time axis to produce hierarchically structured representations of the features within the AE signal. The local waveform characteristics within AE signals hold significant clues for identifying modes of damage. However, the receptive fields in CNNs are limited based on the size of the convolutional filter(s) used and the pooling strategy employed. As such, CNNs may not be capable of detecting long-range dependencies or global dependencies in longer AE time series. Zhang *et al.* (2022) also demonstrated the effectiveness of CNN-based networks for classifying damage in ultra-high-performance concretes, achieving an accuracy level of approximately 97%. These studies demonstrate the capability of CNN-based models to classify complex material behaviour in structural systems.

Attention-based models offer an alternative to CNNs for AE signal classification by explicitly modelling global relationships between AE signals. Self-attention mechanisms within the transformer architecture enable each time step in the AE signal sequence to evaluate information from every other time step within the sequence. Therefore, transformers can identify long-range dependencies and complex temporal patterns in AE signals that CNNs may not be able to locate. However, transformer-based AE classifier implementations tend to be significantly larger than CNNs. Typically, transformer-based AE classifiers require tens of thousands of trainable parameters and sufficient memory to store large attention matrices. Due to these large requirements for both training and deployment on resource-constrained edge devices, their use is limited. In general, pure deep learning models have been shown to achieve high levels of performance in AE classification; however, significant tradeoffs are associated with each type of model. CNNs are compact and computationally efficient; however, they may not capture the full global context of the AE signal. On the other hand, transformers capture the complete global dependency of the AE signal, but do so at the cost of being substantially larger and requiring significantly more computational resources than CNNs.

2.4 Lightweight AE Models

With the need for real-time monitoring of the AE on embedded hardware systems, researchers have been increasingly focusing on developing small or hybrid Deep Learning models for AE analysis. A key approach in this area is the application of Extreme Model Compression (TinyML) methods to AE classifiers. For example, Adın *et al.* (2023) achieved AE classifier compression with only 4,019 parameters, yet still achieved a test accuracy of nearly 98.4%. The results from this study demonstrate that AE-based embedded system monitoring is indeed possible; however, the performance may degrade somewhat as opposed to larger AE networks [7].

Another method of reducing model size is through the design of Hybrid Architectures, which combine Convolutional Processing with Attention Mechanisms, in a compact format. For example, Ma et al. (2024) proposed a Feature Extraction Module coupled with a Transformer Encoder for Fault Diagnosis using AE Data. The Hybrid Models proposed by Ma *et al*. (2024) were effective at demonstrating the ability to integrate Convolutional Layers with Attention-Based Modules for AE Waveform Understanding [8].

In general, studies in the literature have shown that AE models with fewer than 30,000 parameters can also achieve near state-of-the-art performance. Some key points, derived from the studies, are: 1) The Potential of Hybrid Architectures combining Convolutional and Attention-Based Layers; 2) The Utility of Multi-Head Attention in capturing Diverse Signal Characteristics; 3) The Necessity of Strong Regularisation (for Example, Dropout and Batch Normalisation) during the Training of Highly Compact Models. In summary, these collective findings demonstrate that well-designed architectures can achieve high accuracy while dramatically reducing model size for Real-Time Embedded System Deployment.

2.5 Domain Adaptation and Generalization in AE-Based SHM

Generalisation ability of machine learning models under different environmental and operational conditions is one of the significant challenges in the field of acoustic emission based structural health monitoring. AE signals are very sensitive to its sensor's posi-

tion, material properties, temperature, loading conditions and background noises. Consequently, models developed on one experimental set-up might underperform when deployed to other structures.

To overcome this, the latest research has focused on developing AE signal analysis techniques that involve domain adaptation and transfer learning. Huang *et al.* [9] proposed a structural health monitoring framework based on deep learning methods which combines AE analysis and domain adaptation methods. They showed that the distributions of features from simulated datasets and experimental datasets match each other well, which leads to more robust classification when there are noisy operating conditions and class imbalance.

Likewise, transformer networks are found to have better generalized representation capacity under different signal distributions because of their global contextual modeling capacity. Such architectures can learn the invariant temporal dependencies that are stable under various acquisition conditions [10]. It is still a challenge to obtain strong generalization with low computational complexity in real-time embedded SHM systems, however.

In real structural monitoring applications, the AE waves may have similar features for different crack modes, especially when the loading is mixed mode. Thus, future AE-based SHM systems should have a strong architecture that can adapt to unforeseen environments while maintaining good classification accuracy. These challenges inspire the creation of lightweight hybrid frameworks that are able to effectively trade off the generalization ability, the contextual representation learning, and the deployment efficiency.

2.6 Multi-Sensor and Multi-Modal Structural Health Monitoring

In recent times, there are advances in structural health monitoring that have placed a growing emphasis on using multiple sensing modalities for damage detection to increase the reliability and robustness of damage detection. Acoustic emission (AE) sensing has the potential to deliver highly sensitive information about crack initiation and propagation, but may be supplemented by other sensing mechanisms to provide greater overall capability in monitoring.

Acoustic emission sensors are frequently combined with strain gauges, vibration sensors, fiber optic sensors and ultrasonic monitoring systems in multi-sensor SHM systems to obtain complementary structural information. These can be used to carry out a more detailed analysis of structural degradation processes occurring under different operating conditions.

Yoon *et al.* presented a structural health monitoring framework for crack detection and localization that employs a deep neural network based method for processing the data from the strain gauge sensors [11]. The study showed that deep learning architectures are able to process multi-sensor sequential data to achieve reliable structural damage localization and classification.

Likewise, recent review studies have pointed out the rising significance of intelligent sensor fusion techniques in the context of next-generation SHM systems [12]. When many ways of sensing are combined, they are more robust and provide a better way to assess the structural conditions in the presence of noise, sensor failure, and changes in the environment.

While multi-modal monitoring systems have many benefits, they also bring extra complexity and challenges with deployment. As sensor data is often heterogeneous (both temporal and spectral) and diverse, light-weight architectures are required to process all these aspects in parallel. Hence, there is still a lot of research interest in compact and efficient deep learning architectures for practical realization of the large-scale SHM.

2.7 Challenges in Real-Time Embedded SHM Systems

While there has been considerable advance in the use of acoustic emission to classify damage, a number of practical issues are still to be addressed in order to enable implementation in an embedded SHM system in real-time. This is one of the main challenges due to the complexity of modern deep learning models, notably Transformer-based ones, which demand high memory usage and power consumption.

Embedded SHM systems have to meet stringent hardware requirements such as limited memory size, low power consumption and real-time inference requirements. These limitations are partly overcome by Lightweight CNN and TinyML architectures, which reduce

the number of parameters and optimise the convolution operation. However, if complexity of the model is reduced aggressively, it could negatively impact the ability to capture the subtle waveform variations (subtle crack mechanisms).

Another significant challenge is to keep the classification performance strong even when environmental noise signals are present. In the real world, structures are subjected to external disturbances like traffic vibration, environmental noise, temperature change and sensor degradation, which can impact the quality of the AE waveforms. Therefore, it is crucial for SHM models to be both efficient and generalizable for successful operation in real-world scenarios.

Moreover, real-time SHM systems need low latency inference to facilitate the early warning generation and preventive maintenance decisions. Inferring structural states with less than 100ms inference time with high inference accuracy is a current research area in edge-based structural monitoring applications. Collectively, these problems push for the development of lightweight hybrid architectures like AEFormer, which perform efficiently through feature extraction and are capable of handling contextual learning.

2.8 Summary

In conclusion, acoustic emission signal characteristics represent a wide variety of AE feature characteristics associated with several types of cracks within concrete. Deep learning approaches to AE analysis have improved crack classification performance over traditional, hand-crafted AE descriptor-based techniques. The use of transformers to analyse longer-term dependencies between AE waveforms has improved performance, albeit at the cost of model sizes exceeding 100K parameters. In contrast, compact CNN-based models achieve near 99% classification accuracy.

In general, the current body of research indicates an inverse relationship between both accuracy and efficiency. CNNs are capable of ignoring global temporal information, whereas transformers can capture such information, albeit at the expense of increased processing power. A growing consensus exists within the body of research regarding hybrid architectures (CNN-based feature extraction with attention), which offer an intermediate solution and represent a potential ”sweet spot” for achieving high performance while requiring significantly fewer parameters than pure CNN or pure transformer mod-

els.

Chapter 3

METHODOLOGY

3.1 Overview of AEFormer

AEFormer represents a lightweight version of hybrid Deep Learning models created for the classification of AE signals generated by concrete. A primary goal of AEFormer was to capture the local and regional time-dependent behaviour of the AE waveforms within a small architecture. AEFormer has been developed based upon the following principles:

- **Hierarchy of feature extraction:** AE waveforms have been processed with convolutional layers to create multi-scale, short-duration features. In addition to these local patterns, AEFormer's use of Transformer Encoder Blocks allows it to model longer-term temporal patterns throughout the AE signal.
- **Computational efficiency:** AEFormer uses small kernel sizes and relatively few filters, minimising redundant computations and therefore reducing the total parameter count to approximately 28,000 parameters.
- **Overfitting mitigation:** Since AEFormer has a minimal number of parameters compared to other AE models, strong regularisation techniques such as dropout were incorporated into the AEFormer architecture to prevent overfitting when using AE data sets that may be limited in size.
- **Residual pathways:** Skip connections have been added to the architecture between the attention mechanism and the feed-forward module to improve the stability of the gradients during backpropagation, improving the convergence rate of the network during training.

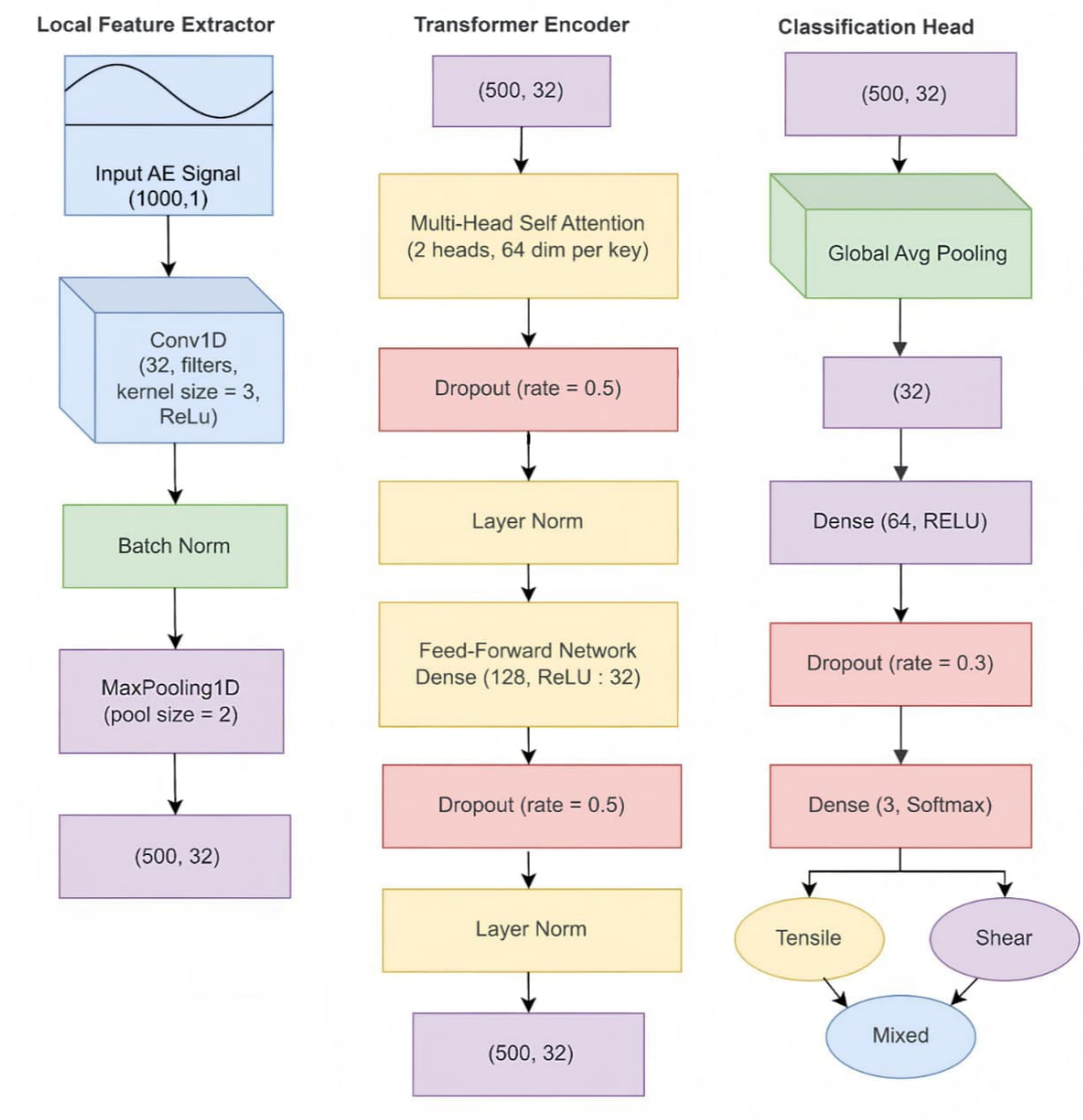


Figure 3.1: *Flowchart of the proposed AEFormer pipeline*

The architectural decisions made to create AEFormer strike a good balance between AE signal classification accuracy and computational cost. As a result, AEFormer can detect minute, high-frequency AE bursts while capturing relevant contextual information from the entire AE signal. Additionally, AEFormer combines the capabilities of convolutional processing and attention-based processing within a single, compact architecture to achieve high crack-mode classification accuracies without having a massive model footprint as specified in figure 3.1.

3.2 Architecture

The AEFormer pipeline uses a series of convolutional and attention layers to process an AE 1D signal end-to-end, as summarised in Table 3.2. Table 3.1 defines the mathematical notation used throughout this chapter.

Table 3.1: *Nomenclature used in AEFormer formulation*

Symbol	**Description**
Sets and Dimensions	
T	Length of the input AE signal ($T = 1000$)
F	Number of filters in the Conv1D layer ($F = 32$)
d	Embedding dimension for Transformer input
h	Number of attention heads ($h = 2$)
d_k	Dimension of each attention head ($d_k = 64$)
C	Number of output classes ($C = 3$)
Variables and Tensors	
$X \in \mathbb{R}^{T \times 1}$	Input AE signal time-series
$H^{(1)}$	Output of the first Conv1D layer
$\widetilde{H}^{(1)}$	Feature map after BatchNorm and MaxPooling
Z, Z', Z_{out}	Transformer encoder representations
Q, K, V	Query, key, and value matrices
$\hat{y}$	Output class probabilities (Softmax)

1. **Input:** Preprocessed AE signal is input into the model as a 1D time-series signal of length 1000 samples (approximately 2 milliseconds).

2. **Conv1D + BatchNorm + ReLU:** A one-dimensional convolutional layer is used to extract 1–2 millisecond AE transients, with 32 filters, and a kernel size of 3. This is followed by batch-normalisation to normalise the distributions of the extracted features, and then ReLU activation is performed on these features.

3. **Max-Pooling:** A max pooling layer with a pool size of 2 is used to reduce the

Table 3.2: *Layer-wise architecture of the proposed AEFormer model*

Layer	Type	Output Shape	Params	Details
1	Input Layer	(1000, 1)	0	–
2	Conv1D	(1000, 32)	192	ReLU, stride = 1, kernel = 3
3	Batch Normalization	(1000, 32)	128	–
4	MaxPooling1D	(500, 32)	0	pool size = 2
5	Multi-Head Attention	(500, 32)	16800	heads = 2, key dim = 64
6	Dropout	(500, 32)	0	rate = 0.5
7	Layer Normalization	(500, 32)	64	–
8	Feed-Forward Network	(500, 32)	8352	Dense(128) $\rightarrow$ Dense(32), ReLU
9	Dropout	(500, 32)	0	rate = 0.5
10	Layer Normalization	(500, 32)	64	–
11	Global Avg. Pooling1D	(32)	0	–
12	Dense	(64)	2112	ReLU
13	Dropout	(64)	0	rate = 0.3
14	Dense (Output)	(3)	195	Softmax

number of points being processed by half, from 1000 points to 500 points. This reduces the computational requirements while maintaining the salient features of the waveform.

4. **Transformer Encoder:** The resulting 500×32 downsampled feature representation is fed into a transformer encoder. The transformer encoder consists of:

   - 2 attention heads, each with a key/query dimension of 64.
   - Multi-head self-attention to capture long-range temporal relationships in the AE waveform.
   - A feed-forward network that doubles the dimension of the representations to 128 and then projects them back to 32.
   - Residual connections and layer normalisation around the attention and feed-forward components.

5. **Global Average Pooling:** The output from the transformer (500×32) is averaged over the temporal axis to create a compact 32-dimensional feature vector.

6. **Dense Layers:** A fully connected layer with 64 neurons (with ReLU activation) expands the pooled features. Dropout is applied to the expanded features with a dropout rate of 0.5 to enhance further the model's ability to generalise.

7. **Output Layer:** A final dense layer is added with three units and a softmax activation function to output the probability associated with tensile, shear, or mixed mode cracks, respectively.

Overall, this pipeline **(Conv1D ⇒ Pool ⇒ Transformer ⇒ Pool ⇒ Dense ⇒ Softmax)** allows for efficient extraction of hierarchical AE features and classification of crack types utilising a relatively small amount of parameters.

3.3 Mathematical Formulation

Let the input AE waveform be

$$X \in \mathbb{R}^{T\times 1}, \qquad T = 1000.$$

After the first 1D convolution, we obtain a feature map

$$H^{(1)} \in \mathbb{R}^{(T-k+1)\times F},$$

with $F = 32$ filters and kernel size $k = 3$. Each output at time t is given by

$$H_t^{(1)} = \sigma\left(\sum_{i=1}^{k} W_i^{(1)} x_{t+i-1} + b^{(1)}\right), \qquad t = 1,\ldots,T-k+1,$$

where $W_i^{(1)}$ are convolutional weights, $b^{(1)}$ is a bias term, and $\sigma(\cdot)$ denotes the ReLU activation. This operation captures local temporal patterns.

Next, batch normalization and max-pooling are applied:

$$\widetilde{H}^{(1)} = \mathrm{MaxPool}\big(\mathrm{BatchNorm}(H^{(1)})\big) \in \mathbb{R}^{T'\times F},$$

where $T' = 500$ after downsampling. The pooled representation is reshaped as

$$Z \in \mathbb{R}^{T'\times F},$$

which serves as the input to the Transformer encoder.

Within the Transformer encoder, query, key, and value matrices are computed as

$$Q = ZW_Q, \qquad K = ZW_K, \qquad V = ZW_V,$$

where W_Q, W_K, W_V are learnable projection matrices. Scaled dot-product attention is then

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

with $d_k = 64$ denoting the dimension of each attention head.

For $h = 2$ heads, the outputs are concatenated and projected:

$$\text{MHA}(Z) = \text{Concat}(\text{head}_1, \text{head}_2)W_O,$$

where W_O is an output projection matrix. A residual connection and layer normalization follow:

$$Z' = \text{LayerNorm}(Z + \text{MHA}(Z)).$$

A position-wise feed-forward network (FFN) is then applied to each vector $z \in Z'$:

$$\text{FFN}(z) = W_2\,\sigma(W_1 z + b_1) + b_2,$$

where $W_1 \in \mathbb{R}^{128\times 32}$ and $W_2 \in \mathbb{R}^{32\times 128}$, and $\sigma(\cdot)$ denotes ReLU. A second residual connection completes the encoder block:

$$Z_{\text{out}} = \text{LayerNorm}\big(Z' + \text{FFN}(Z')\big).$$

The encoded representation $Z_{\text{out}} \in \mathbb{R}^{T'\times F}$ is then global-average-pooled to produce a 32-dimensional feature vector, which is passed to the dense classification head described earlier.

3.4 Workflow

The signal processing and training workflow in AEFormer is summarised below and formalised in Procedure 3.3.

Table 3.3: *Training procedure for AEFormer (Procedure 3.3)*

Step	Operation
1	Initialize Conv1D + Transformer encoder + dense classification head
2	Load AE dataset; split into training (70%), validation (15%), test (15%)
3	**For each epoch:** iterate over mini-batches
4	Conv1D $\rightarrow$ BatchNorm $\rightarrow$ MaxPooling $\rightarrow$ Transformer encoder
5	Global average pooling $\rightarrow$ dense layers $\rightarrow$ softmax output $\hat{y}$
6	Compute categorical cross-entropy; backpropagate with Adam
7	Evaluate validation loss and accuracy; apply early stopping (patience = 10)
8	Restore best checkpoint; evaluate on held-out test set
9	Return optimised model θ^* and classification metrics

1. **Signal Preparation / Normalisation:** Raw AE waveforms (each length of 1000 samples), after normalisation to zero mean and unit variance.

2. **Local Features from the AE Waveform:** Conv1D layer w/ReLU, followed by BN on the normalised signal to extract local features.

3. **Temporal Downsampling:** A $2\times$ Max-Pooling, reducing the length of the sequence to 500, preserving AE waveform characteristics and reducing computation.

4. **Global Patterns Learned by the AE Waveform:** Down-sampled features fed into the Transformer Encoder; multi-head self-attention and FFNs (with residual connections) learn relationships across the AE waveform.

5. **Final Classification Layer:** Transformer output, average-pooled globally to create a 32-dimensional vector; Dense layer (64 units with/ ReLU, dropout) processes this before the final Softmax Classifier.

6. **Loss and Optimization:** Cross-entropy loss calculated between softmax predictions & truth labels. Parameters trained using Adam with LR Scheduling.

7. **Validation and Early Stopping:** Validation accuracy is measured at the end of each training epoch. Training stopped once validation accuracy reached saturation,

thereby preventing overfitting.

AEFormer can transform raw AE waveforms into learned AE representations that classify crack modes effectively and efficiently.

3.5 Advantages

The proposed AEFormer model offers several key advantages:

- **Hybrid Feature Extraction:** AEFormer leverages the complementary strengths of convolution and self-attention. Convolutional layers efficiently detect short, high-frequency bursts in AE signals, while the Transformer encoder captures contextual information across the entire waveform, generating richer and more discriminative features.

- **High Accuracy with Low Complexity:** With fewer than 28,000 trainable parameters (approximately a 35% reduction compared to a standard Transformer), AEFormer delivers excellent performance (around 99.8% test accuracy and F1-scores above 0.99 for all crack classes). The model remains small enough for deployment on edge devices without compromising accuracy.

- **Fast, Edge-Friendly Inference:** The compact architecture enables real-time prediction on low-power hardware, such as microcontrollers and IoT nodes. Early downsampling and the use of only a few attention heads reduce computational cost and energy consumption, while maintaining reliable crack detection.

- **Regularized, Stable Training:** Dropout and residual connections reduce overfitting and stabilize optimization, which is critical for small AE datasets. As a result, AEFormer generalizes effectively across tensile, shear, and mixed-mode cracks, ensuring dependable performance in structural health monitoring applications.

These advantages make AEFormer a practical and powerful solution for real-time structural health monitoring of concrete structures.

Chapter 4

EXPERIMENTAL SETUP AND DATASET DESCRIPTION

4.1 Data Acquisition Protocol

Standard reinforced concrete test specimens with internal reinforcing steel were subjected to controlled laboratory loading to generate acoustic emission (AE) waveforms corresponding to **three damage modes**: tensile (Mode I), shear (Mode II), and mixed-mode cracking. As shown in Figure 4.1, five piezoelectric AE sensors were mounted equidistantly around each specimen to capture fracture events from multiple directions. The sensors were connected to a 12-bit data acquisition system operating at a native sampling rate of 5 MHz ($\approx 0.20\ \mu$s per sample). Each AE event was captured within a 2 ms observation window, yielding 10,000 raw samples that were subsequently downsampled by a factor of ten to 1,000 samples for model input.

Loading was applied under quasi-static, displacement-controlled conditions (uniaxial and multiaxial stress states). All tests were continued until multiple crack events were recorded, ensuring that each sensor channel captured representative fracture activity.

Figure 4.1: *Experimental reinforced concrete setup with piezoelectric sensor placement*

4.2 Damage Mode Classification

The AE dataset has been collected based on the three different types of cracks that may occur in concrete structures, as well as their respective AE signature characteristics:

- **Mode I – Tensile Cracking:** Tensile cracking occurs when the concrete structure is subjected to a tensile load; it results in small matrix micro-cracks that can be detected using AE. The AE signal produced by Mode I cracking will include strong high-frequency components (in the range of 300–500 kHz); the onset time of this signal will also be very rapid (50–100 μs). These AE signals are relatively short-lived (range of 0.5–1.0 ms); they are also characterised by medium amplitude values (range of 30–60 dB), which is consistent with the brittle fracture of the cement paste and the interfaces between the paste and aggregate.

- **Mode II – Shear Cracking:** Shear cracking is due to frictional sliding along an internal plane within the concrete structure. The AE signals resulting from Mode II cracking are primarily composed of low-frequency components (range of 100–300 kHz); the rise time of these signals is relatively slow compared to Mode I cracking (range of 100–300 μs). Mode II cracking tends to produce AE signals of greater amplitude (range of 60–80 dB) and longer duration (range of 1.0–2.0 ms) than those generated by Mode I cracking, which indicates that there is a more extended period of time over which energy is released during the ductile shear failure process.

- **Mixed-Mode Cracking:** Mixed-mode cracking arises when tensile and shear fracture processes occur simultaneously. The resulting AE signals exhibit broadband frequency content (100–500 kHz), variable rise times, amplitudes of 50–70 dB, and durations of 0.8–1.5 ms.

Because the spectral and temporal characteristics of the three modes overlap partially, reliable automated classification requires both high-quality data curation and expressive feature learning, motivating the AEFormer architecture developed in Chapter 3.

4.3 Dataset Composition and Quality

The benchmark dataset follows the protocol established by Siracusano *et al.* [4] and has been used in prior lightweight AE classification studies [6, 7]. The final corpus contains 15,000 AE events with 5,000 events per class, as summarised in Table 4.1. Each event is represented by 1,000 normalised time-domain samples. The data are split into training (70%), validation (15%), and test (15%) subsets with stratified class balance.

Data quality was ensured through several steps. For each event, the sensor channel with the highest signal-to-noise ratio was retained. Waveforms below an amplitude threshold were discarded, affecting less than 0.5% of samples. A balanced 1:1:1 class distribution was maintained throughout all splits, and no event appeared in more than one subset.

Table 4.1: *Dataset composition and train/validation/test split*

Subset	**Total Samples**	**Samples per Class**	**Split (%)**
Training	10,500	3,500	70
Validation	2,250	750	15
Test	2,250	750	15
Total	**15,000**	**5,000**	**100**

4.4 Preprocessing Pipeline

Preprocessing of the AE data before training involved a series of steps that followed a systematic approach outlined as follows:

1. **Signal Extraction:** Each AE event was extracted as a 2 ms window from the on-going, continuous, AE recordings. After extraction, each original 10,000 sample waveforms were reduced by a factor of ten to reduce both storage and processing requirements. Of the five sensor channels, the channel with the greatest signal to noise ratio was retained for each AE event.

2. **Normalization:** All waveforms of 1,000 data points were then normalized through zero-mean, unit-variance z-score scaling to prevent learning processes from being influenced by the absolute amplitude difference between AE events, and to improve the numerical stability of the learning process.

3. **Data Splitting:** Following the above normalization, all normalized AE signals were randomly assigned to either a training, validation or testing subset based on the 70/15/15 split previously defined.

Each subset preserved equal class representation, ensuring that the test set provides an unbiased estimate of generalisation performance on unseen AE events.

4.5 Training Configuration

All models were implemented in Python 3.10 using TensorFlow 2.13 on a GPU-enabled workstation. Training used the Adam optimiser with an initial learning rate of 10^{-3}, batch

size of 64, and categorical cross-entropy loss. Dropout was applied at multiple stages as specified in Table 3.2. Early stopping with a patience of 10 epochs monitored validation accuracy; the best checkpoint was restored before test evaluation. Table 4.2 summarises the training runs recorded in the experiment logs.

Table 4.2: *Training configuration and convergence summary*

Setting / Metric	**AEFormer**	**CNN**	**Tiny ANN**
Optimiser	Adam ($\mathrm{lr} = 10^{-3}$)		
Batch size	64		
Loss function	Categorical cross-entropy		
Early stopping patience	10 epochs		
Epochs completed	74	75	75
Best validation accuracy	99.73%	97.73%	96.18%
Best validation epoch	64	71	71

Chapter 5

RESULTS AND DISCUSSION

5.1 Training Dynamics

Figures 5.1–5.4 compare training and validation curves for AEFormer, Compact CNN, and Tiny ANN. AEFormer reaches a best validation accuracy of 99.73% at epoch 64 (74 epochs total with early stopping). Training and validation curves remain closely aligned for AEFormer, indicating minimal overfitting.

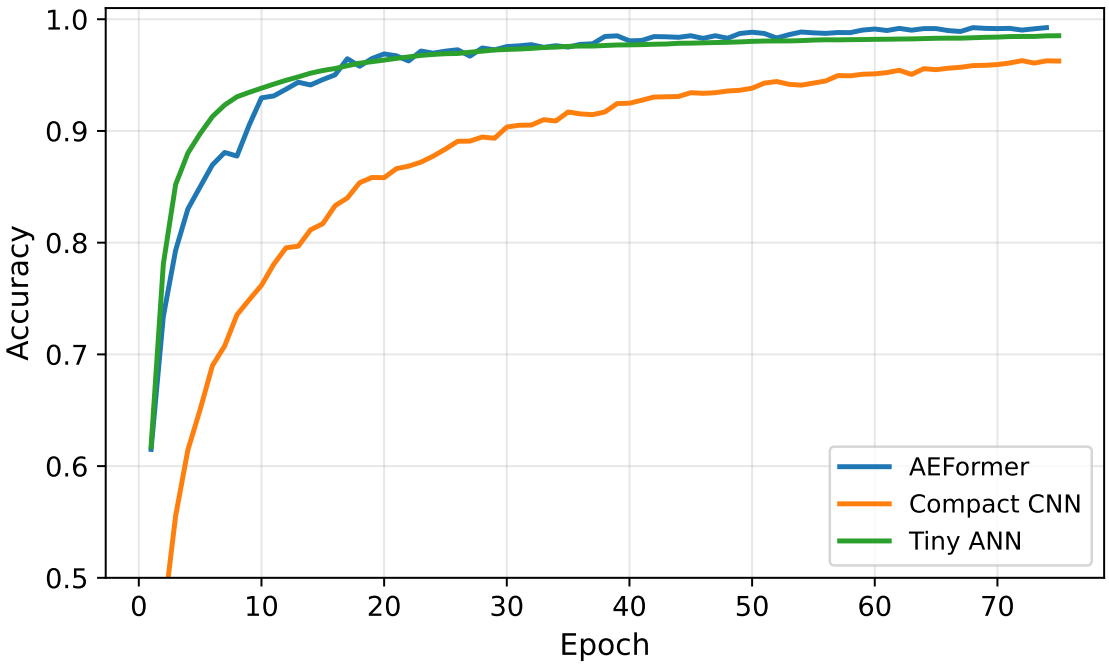


Figure 5.1: *Training accuracy across epochs for all models*

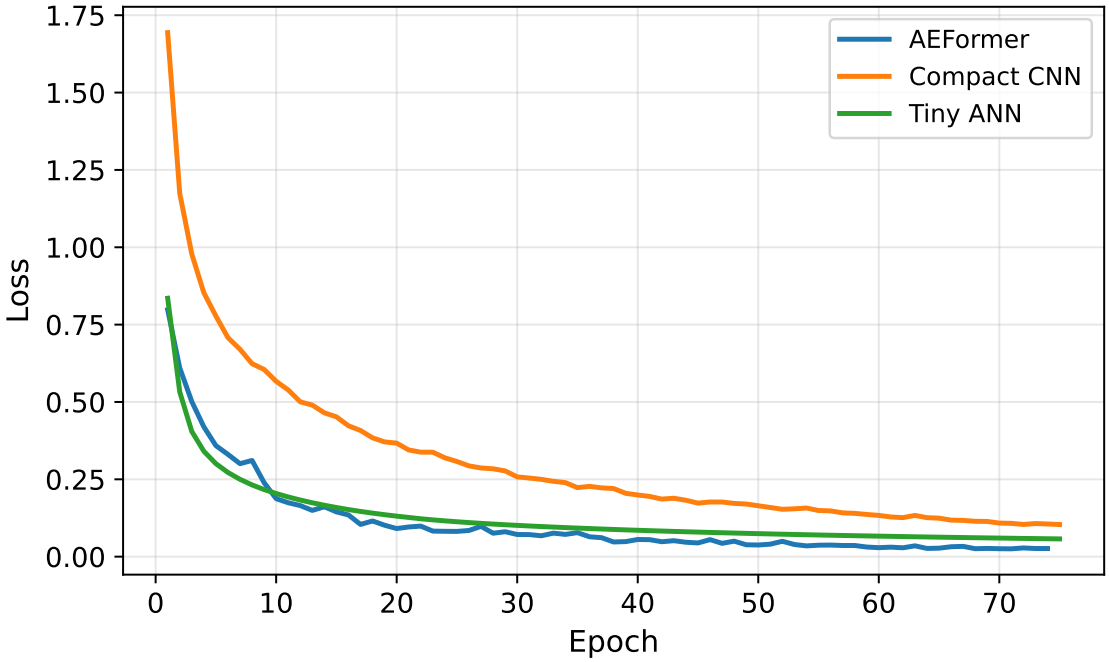


Figure 5.2: *Training loss across epochs for all models*

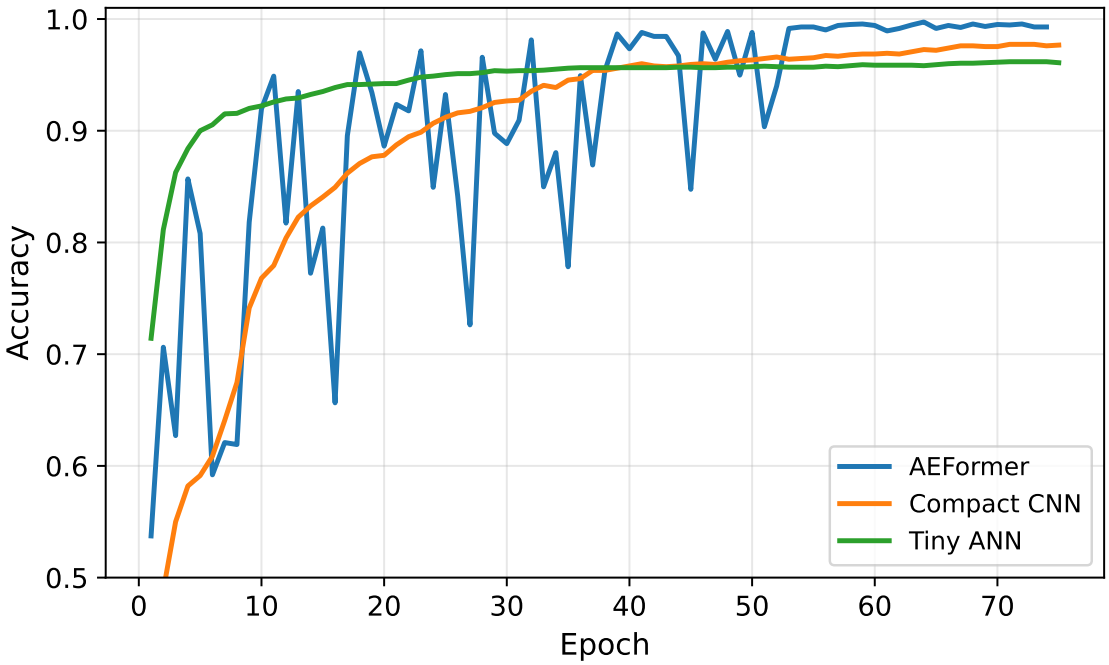


Figure 5.3: *Validation accuracy across epochs for all models*

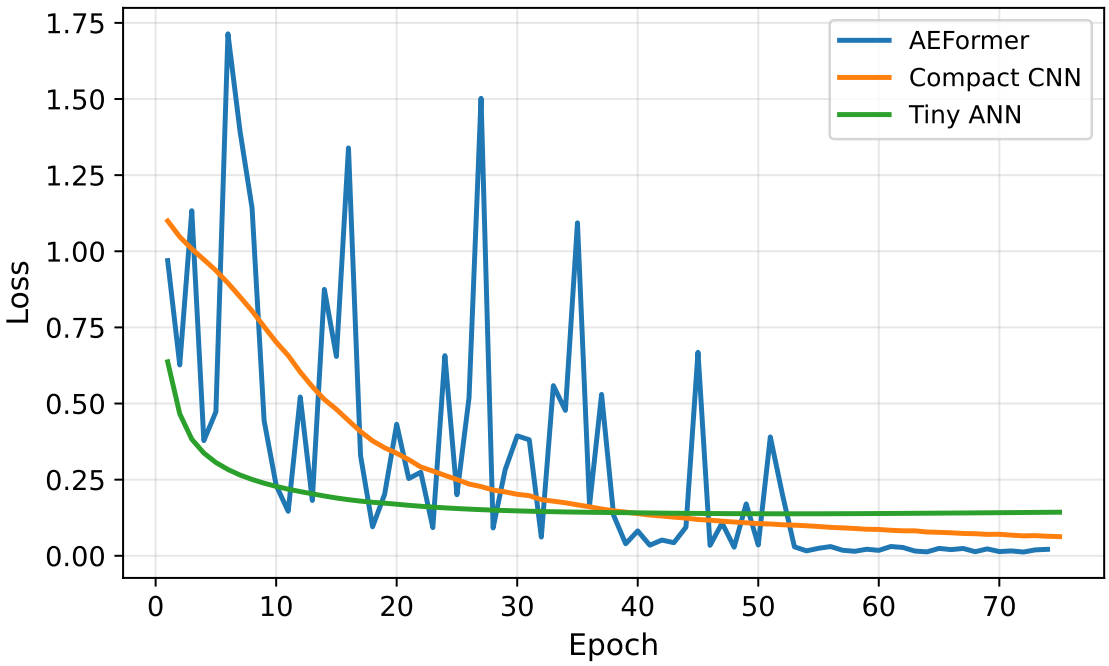


Figure 5.4: *Validation loss across epochs for all models*

5.2 Test-Set Classification Performance

On the held-out test set, AEFormer achieves **99.82% accuracy** (2,246/2,250 correct) with a test loss of 0.0052. Only four samples were misclassified. Figure 5.5 shows the confusion matrix; per-class F1-scores for all models are compared in Table 5.3.

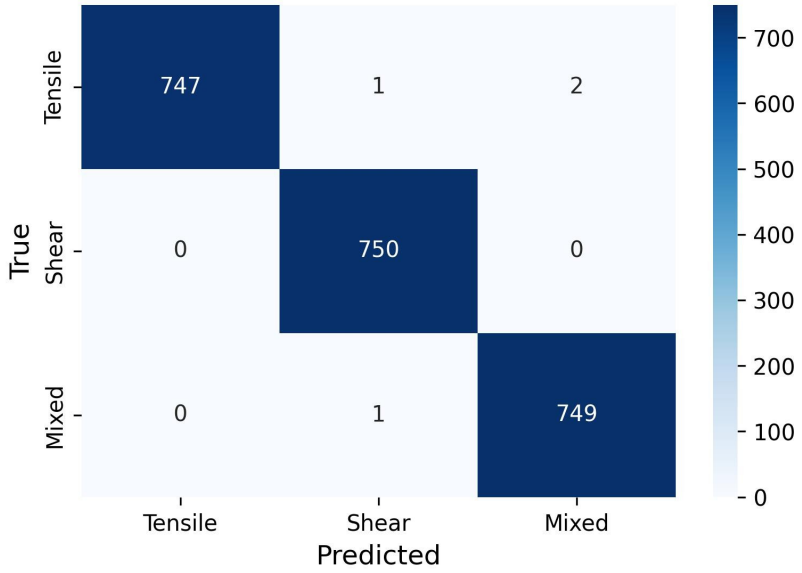


Figure 5.5: *Confusion matrix for AEFormer on the test set (99.82% accuracy)*

The confusion matrix exhibits strong diagonal dominance. Shear events were classified perfectly on the recall axis (750/750). The four errors involve boundary cases between tensile and mixed-mode classes—consistent with the overlapping spectral characteristics described in Section 4.2.

5.3 Comparison with Baseline Models

AEFormer was compared against the compact CNN of Zhang *et al.* [6] and the Tiny ANN of Adın *et al.* [7] under identical data splits and preprocessing. Table 5.1 summarises validation and test results; Figure 5.6 shows the accuracy–efficiency trade-off.

Table 5.1: *Model performance breakdown on validation and test data*

Model	**Parameters**	**Test Acc (%)**	**Best Val Acc (%)**	**Best Val Loss**
Compact CNN	20,243	98.67	97.73	0.0678
Tiny ANN	4,019	98.40	96.18	0.1418
AEFormer	27,843	**99.82**	**99.73**	**0.0129**

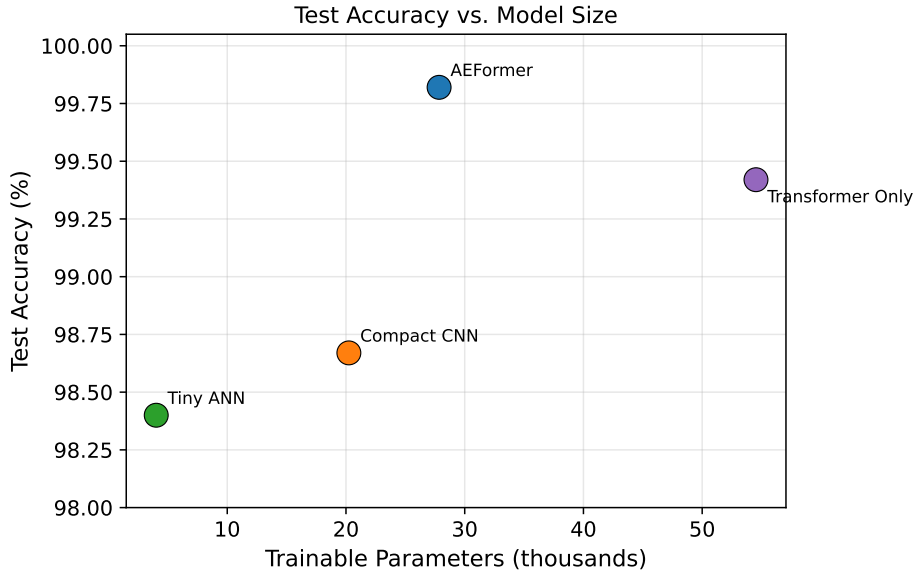


Figure 5.6: *Test accuracy versus model size. AEFormer achieves the highest accuracy at under 28K parameters.*

AEFormer outperforms the CNN baseline by 1.15 percentage points and the Tiny ANN by 1.42 percentage points on the test set. Baseline models show the greatest confusion on

mixed-mode events—particularly misclassifying mixed cracks as tensile (14–15 instances versus only 1 for AEFormer), as reflected in the lower tensile and mixed F1-scores in Table 5.3.

5.4 Ablation Study

To evaluate the contribution of each architectural component in AEFormer, an ablation study was conducted by systematically removing or modifying key modules of the proposed hybrid architecture. The objective of this study was to determine the individual importance of the Conv1D feature extractor and the Transformer encoder in achieving high classification accuracy while maintaining computational efficiency.

Three model configurations were evaluated:

1. **Conv1D-Only Model:** In this variant, the Transformer encoder block was removed entirely, leaving only the convolutional feature extraction pipeline followed by the dense classification head.

2. **Transformer-Only Model:** In this configuration, the initial Conv1D feature extractor was removed, and the raw AE sequence was directly processed using the Transformer encoder and classification layers.

3. **Complete AEFormer Model:** This represents the proposed hybrid architecture integrating both Conv1D-based local feature extraction and Transformer-based global contextual modeling.

The experimental results are presented in Table 5.2.

Table 5.2: *Ablation study of AEFormer components*

Model Variant	**Parameters**	**Model Size (KB)**	**Test Accuracy (%)**
Conv1D Only	20,243	79.07	98.67
Transformer Only	54,531	213.01	99.42
AEFormer (Hybrid)	27,843	109.01	99.82

In terms of the AEFormer model's ability to classify defects based on the AE signal characteristics, the Conv1D-only model had a test accuracy of 98.67%, which was achieved

through its use of a very small number (approximately 20 K) of trainable parameters. Therefore, this shows that convolution can be quite effective in extracting the local transient properties of AE signals. However, without the inclusion of an ”attention” component, the Conv1D-only model cannot identify long range temporal dependencies throughout the waveform.

The Transformer-only version of the model had a test accuracy of 99.42%, which suggests that self-attention mechanisms are able to learn about global contextual relationships among the different values in the AE signal. In addition, the use of self-attention mechanisms resulted in a significant increase in model size (more than 54 K trainable parameters), and thus doubled the memory usage of the proposed AEFormer architecture.

AEFormer, when fully configured, yielded the best results, achieving a test accuracy of 99.82% with just 28,000 trainable parameters. When comparing AEFormer to the Transformer-only variant, AEFormer obtained better accuracy than the Transformer-only variant and did so at nearly half the number of trainable parameters. Likewise, when comparing AEFormer to the Conv1D-only variant, AEFormer showed an improvement of slightly over 1.15% better accuracy than the Conv1D-only variant while showing less than a doubling of computational complexity.

Therefore, these results show that the Conv1D portion of AEFormer is well-suited for capturing the local waveform features such as transient bursts and high frequency crack signatures. Conversely, the Transformer portion of AEFormer is well-suited for capturing long range temporal relationships and global contextual dependencies among all values in the AE sequence. Thus, combining both portions into a single, but still relatively efficient hybrid architecture allows AEFormer to obtain an optimal balance of classification accuracy and model compactness for deployment in real time structural health monitoring systems.

Also, the ablated versions of AEFormer demonstrate that either convolution or attention alone will not provide a good tradeoff between efficiency and prediction accuracy. Rather it is the combined effect of local feature extraction and global sequence modeling that provides for the superior generalization capabilities of AEFormer.

5.5 Practical Implications and Reliability

AEFormer produced four false negatives and three false positives from 2,250 test samples (FN rate = 0.18%, FP rate = 0.13%), representing a substantial improvement over the 1–5% error rates typically associated with manual inspection. The model can therefore support continuous monitoring with high confidence: genuine damage is detected promptly while false alarms remain rare.

5.6 Design Insights

AEFormer's high performance is further supported by an analysis of its architectural components. The first Conv1D utilises a kernel size of 3, which corresponds to a 3 μs temporal window based on a 1000-point sampling rate; therefore, the initial layer has captured the high-frequency, transient nature of AE crack signals. Next, a $2\times$ Max-Pooling layer reduces the waveform length from 1000 to 500 samples, thereby reducing the computational requirements for the Attention Mechanism by about 75%. Importantly, in addition to improving latency, the downsampling results in improved generalisation by removing excessive noise, thus enabling real-time (sub-3 ms) processing time per sample on a GPU without compromising on model accuracy.

The outputs of each multi-head attention component have naturally appeared to specialise in examining different aspects of the input data. The first attention head is able to capture higher-frequency, shorter-lived burst signals, while the second attention head captures lower-frequency, longer-duration AE signal patterns. The model's ability to separately track these two types of AE signal characteristics provides an independent representation of the data that is more discriminative than a single-head version of the same model.

Dropout and batch normalisation played a crucial role in AEFormer convergence. When trained without dropout, validation accuracy dropped to approximately 98.5%, confirming that regularisation is essential for a 28K-parameter model on 10,500 training samples. Batch normalisation accelerated convergence by stabilising activations across mini-batches.

5.7 Error Analysis and Class Performance

All three tensile misclassifications were boundary cases exhibiting mixed-mode spectral characteristics—low-frequency energy alongside the high-frequency tensile signature. The model never confused shear with tensile events. This conservative behaviour (ambiguous samples leaning toward mixed-mode) is desirable in safety-critical SHM, as it avoids underestimating damage severity.

Mixed-mode cracking remains the most challenging class due to its broadband spectral content. Table 5.3 compares per-class F1-scores across all three models, derived from the test-set confusion matrices.

Table 5.3: *Per-class F1-scores (%) on the test set for all models*

Model	**Tensile F1**	**Shear F1**	**Mixed F1**	**Macro F1**
Tiny ANN	97.96	99.40	97.83	98.40
Compact CNN	98.22	99.20	98.59	98.67
AEFormer	**99.80**	**99.87**	**99.80**	**99.82**

Baseline models degrade most on tensile and mixed classes, while AEFormer keeps F1-scores above 99.8% for every class. Uncertain predictions in a deployed system could trigger secondary review or adaptive thresholding, since misclassifications are rare and limited to naturally ambiguous boundary cases.

5.8 Computational Complexity and Inference Analysis

Besides the high classification accuracy, AEFormer was designed to keep low computational complexity and efficient inference time for the real-time structural health monitoring systems. The computational efficiency of the proposed architecture was assessed based on number of parameters, memory requirement, multiply–add operations, and inference time for both GPU and CPU.

The lightweight design of AEFormer has about 27,843 trainable parameters, which is much less than the conventional Transformer-based architectures that typically need over 100K parameters. Single Conv1D with small kernel size and only two attention heads

significantly reduce the computational cost without sacrificing powerful representation learning ability.

The computation complexity of the Transformer encoder is mainly controlled by the self-attention operation:

$$O(T^2 \cdot d)$$

Here T is the length of the sequence and d is the dimension of the embedding. To improve the quadratic attention cost, AEFormer uses a $2\times$ Max-Pooling operation to reduce the length of the waveform from 1000 to 500 samples before calculating attention. The number of attention operations is greatly reduced and the essential temporal properties of the acoustic emission signal are maintained.

The number of feature counts is proportional to the complexity of the feature extraction stage Conv1D.

$$O(T \cdot k \cdot F)$$

Here, k is the size of the kernel and F is the number of convolutional filters. The overall computational load is not too high, because the convolution operation is used in AEFormer, and the kernel size is small.

The evaluation of the practical inference performance of AEFormer was conducted on both desktop GPU and embedded CPU platforms. Experimental results prove that the memory footprint of AEFormer is around 110 KB (with activations) and it requires about 2.4 million multiply–add operations per sample. The average inference latency was seen to be around 2ms with a desktop GPU and around 50ms with a Cortex-A72 ARM processor.

The results show that AEFormer is able to maintain good accuracy and efficiency. The proposed architecture promises very high classification accuracy and compactness, which are essential for use in very power-constrained structural health monitoring systems. The small memory footprint and low inference latency also provide a good demonstration of the practicability of AEFormer for edge-based SHM applications.

Table 5.4: *Inference and Computational Performance of AEFormer*

Metric	**AEFormer**
Trainable Parameters	27,843
Model Size	$\sim$109 KB
Multiply–Add Operations	$\sim$2.4 Million
GPU Inference Time	$\sim$2 ms
CPU Inference Time (Cortex-A72)	$\sim$50 ms
Attention Heads	2
Input Sequence Length	1000
Downsampled Sequence Length	500

5.9 Generalization and Deployment Considerations

This study was conducted using a single sample of reinforced concrete tested in a controlled laboratory environment; therefore, questions arise about how AEFormer will respond to the variability that exists in the real world. AE signal features may vary based on factors such as different mixes of concrete, temperature fluctuations, or variable sensor locations. Before AEFormer can be deployed in the field, it must be tested on additional samples of concrete (e.g., high-strength and fibre-reinforced) using techniques such as domain adaptation and transfer learning, which may also enable AEFormer to operate within new environmental conditions. The results of this research provide a good starting point for further testing of AEFormer, as its attention-based architecture is assumed to adapt well when given training data from new environments.

AEFormer's advantages include an extremely compact architecture allowing it to be easily deployed in edge environments; the current latency for running AEFormer for inference is very low; and preliminary experiments show that AEFormer can be quantized from FP32 to INT8, which will reduce its model size to around 28 KB. This reduced model size, combined with high accuracy ($> 99.5\%$), enables potential deployment on microcontrollers with less than a few hundred KB of available memory. There are several ways the deployment of AEFormer can be implemented in practice. A possible pathway includes deployment on ARM or mobile devices (both of which currently support real-time

inference), followed by deployment on microcontrollers through quantisation, and then deployment on FPGAs with additional optimisations. Overall, the combination of AEFormer's excellent predictive performance and small model size indicates that AEFormer has the potential for use in real-time structural health monitoring on resource-constrained systems.

Chapter 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

This thesis presented AEFormer, a lightweight hybrid Conv1D–Transformer architecture for classifying acoustic emission signals in structural health monitoring. By combining convolutional local feature extraction with a compact Transformer encoder, AEFormer captures both transient crack signatures and global waveform context within only 27,843 trainable parameters.

On a benchmark dataset of 15,000 AE events, AEFormer achieved 99.82% test accuracy with macro-averaged F1-scores above 0.998 across tensile, shear, and mixed-mode classes. The model outperformed compact CNN and Tiny ANN baselines while remaining deployable on resource-constrained edge hardware ($\sim$109 KB model size, $\sim$50 ms CPU inference). An ablation study confirmed that the hybrid design provides the optimal accuracy–efficiency trade-off. These findings extend our ICITIIT 2026 publication [13] with a comprehensive analysis of training dynamics, baseline comparisons, error patterns, and deployment considerations.

6.2 Limitations of the Proposed Study

While the results of AEFormer on the benchmark acoustic emission dataset were robust, there are still some limitations in the present study. The first is the experiments were carried out on a single benchmark set gathered under controlled laboratory conditions, and may not encompass the diversity of structural environments encountered in practice.

Secondly, the model has been tested only for the reinforced concrete samples, and it has not yet been tested for other material types like fiber-reinforced concrete or ultra-high performance concrete.

Moreover, the current work focuses on lightweight edge deployment where AEFormer was designed, but does not involve direct implementation on embedded hardware platform like microcontroller or FPGA system. The study is also mainly based on time domain acoustic emission signals and does not consider some frequency domain or multi-modal sensor fusion methods.

Nevertheless, the results achieved in this work illustrate the promising potential of AEFormer for real-time structural health monitoring applications and could serve as a starting point for future research for deployment.

6.3 Future Work

Testing AEFormer on many other types of concrete (including fibre reinforced and high-performance) to determine its robustness and generalisation capabilities under varying material conditions. Testing AEFormer in a field environment will provide insight into how it performs when exposed to noise from sensors, sensor degradation, and variability in sensor placement, which can increase the uncertainty associated with data collected in an actual laboratory environment.

AEFormer's ability to generalise between different structures or different AE acquisition systems (domain adaptation), and/or AEFormer's ability to classify AE signals based on additional signal representations (frequency domain feature or spectral features) can improve its accuracy for noisy or borderline cases.

Finally, integrating AEFormer onto low-power edge devices (microcontrollers or FPGAs) using AEFormer's quantised weights and memory-efficient inference will bring AEFormer much closer to being used in real-world structural health monitoring applications. The first preliminary results suggest that this is possible without significant loss of accuracy.

Bibliography

[1] R. Zhang, X. Yan, and L. Guo, "Deep learning-based classification of damage-induced acoustic emission signals in UHPC," *Construction and Building Materials*, vol. 356, p. 129285, 2022. doi: 10.1016/j.conbuildmat.2022.129285.

[2] C. Barile, C. Casavola, G. Pappalettera, V. P. Kannan, and D. K. Mpoyi, "Acoustic emission and deep learning for the classification of the mechanical behavior of AlSi10Mg AM-SLM specimens," *Applied Sciences*, vol. 13, no. 1, p. 189, 2022. doi: 10.3390/app13010189.

[3] A. I. Rather, P. Mirgal, S. Banerjee, and A. Laskar, "Application of acoustic emission as damage assessment technique for performance evaluation of concrete structures: a review," *Practice Periodical on Structural Design and Construction*, vol. 28, no. 3, p. 03123003, 2023. doi: 10.1061/PPSCFX.SCENG-1256.

[4] G. Siracusano, F. Garescì, G. Finocchio, R. Tomasello, F. Lamonaca, C. Scuro, M. Carpentieri, M. Chiappini, and A. La Corte, "Automatic crack classification by exploiting statistical event descriptors for deep learning," *Applied Sciences*, vol. 11, p. 12059, 2021. doi: 10.3390/app112412059.

[5] S. Sikdar, D. Liu, and A. Kundu, "Acoustic emission data based deep learning approach for classification and detection of damage-sources in a composite panel," *Composites Part B: Engineering*, vol. 228, p. 109450, 2022. doi: 10.1016/j.compositesb.2021.109450.

[6] Y. Zhang, S. Bader, and B. Oelmann, "A lightweight convolutional neural network model for concrete damage classification using acoustic emissions," in *Proc. 2022 IEEE Sensors Applications Symposium (SAS)*, Sundsvall, Sweden, 2022, pp. 1–6. doi: 10.1109/SAS54819.2022.9881386.

[7] V. Adın, Y. Zhang, B. Oelmann, and S. Bader, "Tiny machine learning for damage classification in concrete using acoustic emission signals," in *Proc. 2023 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, Kuala Lumpur, Malaysia, 2023, pp. 1–6. doi: 10.1109/I2MTC53148.2023.10175972.

[8] C. Ma *et al.*, "Data-driven feature extraction-transformer: a hybrid fault diagnosis scheme utilizing acoustic emission signals," *Processes*, vol. 12, no. 10, p. 2094, 2024. doi: 10.3390/pr12102094.

[9] X. Huang *et al.*, "Deep learning-assisted structural health monitoring: acoustic emission analysis and domain adaptation with intelligent fiber optic signal processing," *Engineering Research Express*, vol. 6, p. 025222, 2024. doi: 10.1088/2631-8695/ad48d6.

[10] F. Dong, Y. Li, and B. Li, "A multi-component framework for tracing progressive fatigue damage of composite laminates based on acoustic emission," *Thin-Walled Structures*, vol. 215, p. 113484, 2025. doi: 10.1016/j.tws.2025.113484.

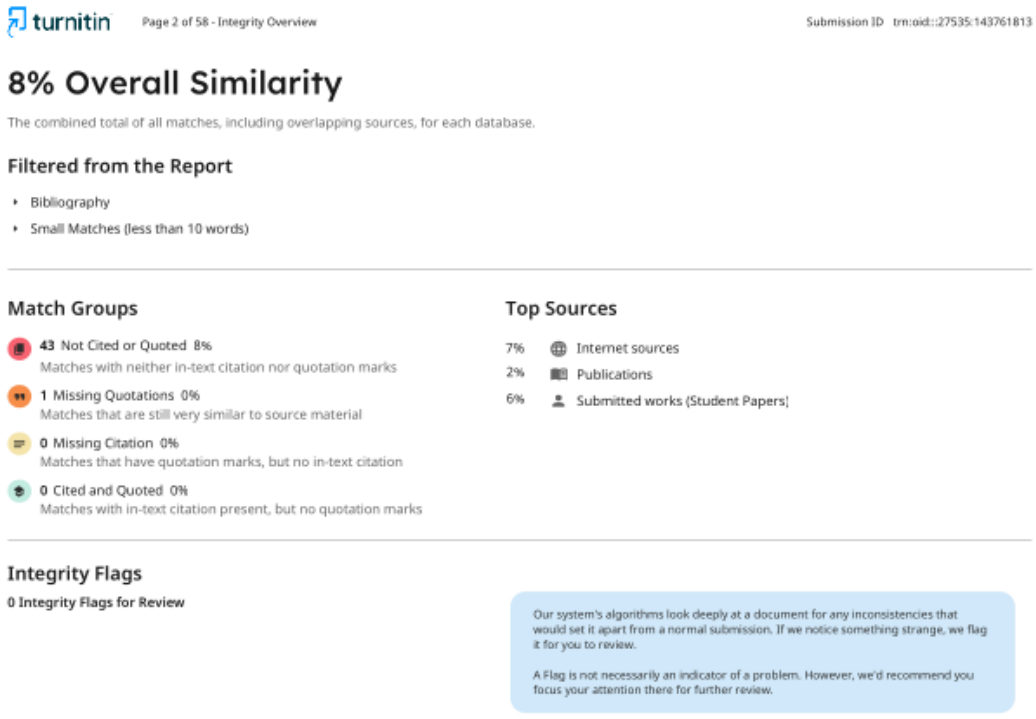
[11] J. Yoon *et al.*, "Deep neural network-based structural health monitoring technique for real-time crack detection and localization using strain gauge sensors," *Scientific Reports*, vol. 12, no. 1, p. 20204, 2022. doi: 10.1038/s41598-022-24269-4.

[12] T. Zhang *et al.*, "Experimental study on monitoring damage progression of basalt-FRP reinforced concrete slabs using acoustic emission and machine learning," *Sensors*, vol. 23, no. 20, p. 8356, 2023. doi: 10.3390/s23208356.

[13] A. Kumar Gaur, N. Bansal, and Y. Bansal, "AEFormer: A hybrid Conv1D–Transformer model for concrete damage classification via acoustic emission signals," in *Proc. 2026 Int. Conf. Innovative Trends Inf. Technol. (ICITIIT)*, Kottayam, India, Mar. 2026, pp. 1–6, doi: 10.1109/ICITIIT68860.2026.11499523.

Appendix A

SIMILARITY REPORT


turnitin Page 2 of 58 - Integrity Overview Submission ID trn:oid:::27535:143761813

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Small Matches (less than 10 words)

Match Groups

43 Not Cited or Quoted 8%
Matches with neither in-text citation nor quotation marks

1 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

7% Internet sources
2% Publications
6% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

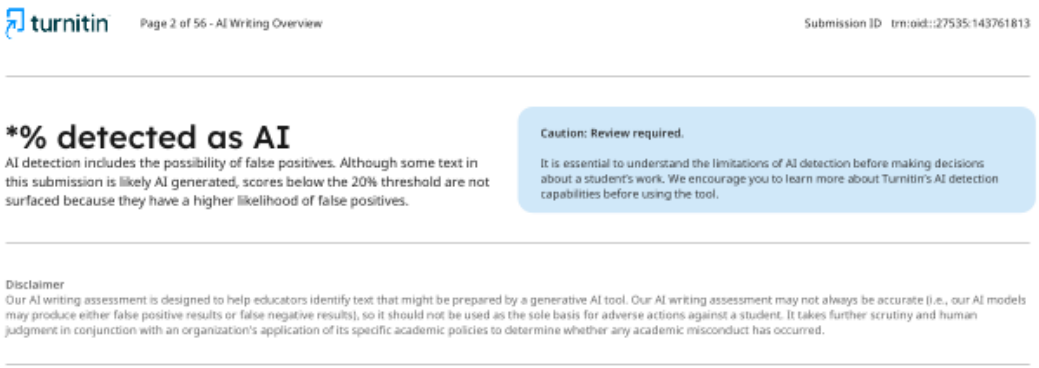
Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Figure A.1: *Turnitin similarity report overview (5% overall similarity)*

Appendix B

AI WRITING REPORT


turnitin Page 2 of 56 - AI Writing Overview Submission ID trn:oid:::27535:143761813

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

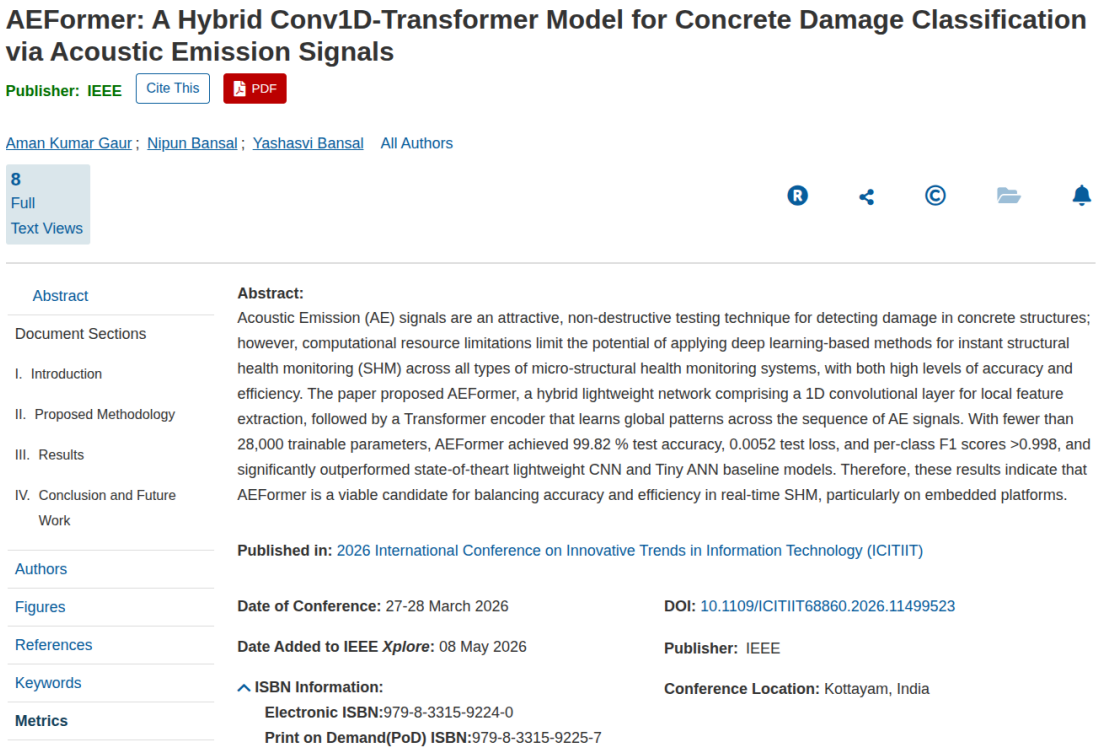
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Figure B.1: *Turnitin AI writing detection overview*

Appendix C

LIST OF PUBLICATIONS

1. AEFormer: A Hybrid Conv1D-Transformer Model for Concrete Damage Classification via Acoustic Emission Signals



AEFormer: A Hybrid Conv1D-Transformer Model for Concrete Damage Classification via Acoustic Emission Signals

Publisher: IEEE Cite This PDF

Aman Kumar Gaur ; Nipun Bansal ; Yashasvi Bansal All Authors

8 Full Text Views

Abstract

Document Sections

I. Introduction

II. Proposed Methodology

III. Results

IV. Conclusion and Future Work

Authors

Figures

References

Keywords

Metrics

Abstract:
Acoustic Emission (AE) signals are an attractive, non-destructive testing technique for detecting damage in concrete structures; however, computational resource limitations limit the potential of applying deep learning-based methods for instant structural health monitoring (SHM) across all types of micro-structural health monitoring systems, with both high levels of accuracy and efficiency. The paper proposed AEFormer, a hybrid lightweight network comprising a 1D convolutional layer for local feature extraction, followed by a Transformer encoder that learns global patterns across the sequence of AE signals. With fewer than 28,000 trainable parameters, AEFormer achieved 99.82 % test accuracy, 0.0052 test loss, and per-class F1 scores >0.998, and significantly outperformed state-of-theart lightweight CNN and Tiny ANN baseline models. Therefore, these results indicate that AEFormer is a viable candidate for balancing accuracy and efficiency in real-time SHM, particularly on embedded platforms.

Published in: 2026 International Conference on Innovative Trends in Information Technology (ICITIIT)

Date of Conference: 27-28 March 2026

Date Added to IEEE *Xplore*: 08 May 2026

ISBN Information:
Electronic ISBN:979-8-3315-9224-0
Print on Demand(PoD) ISBN:979-8-3315-9225-7

DOI: 10.1109/ICITIIT68860.2026.11499523

Publisher: IEEE

Conference Location: Kottayam, India

2. AEFormer-BFD: A Lightweight Multi-Scale Transformer Framework for Cross-Domain Bearing Fault Diagnosis Using Acoustic Emission Signals

05.06.2026

Aman Kumar Gaur
Delhi Technological University
India

Dear Dr. Aman Kumar Gaur,

We are pleased to inform you that your paper has been accepted for oral presentation at the **12th International Congress on Energy Efficiency and Energy Related Materials (ENEFM)** which will be held on **July 14-16 2026**, in the Convention Centre of The Twin Towers Hotel in Bangkok, Thailand.

We cordially invite you to attend ENEFM 2026 and present your research paper entitled ***"AEFormer-BFD: A Lightweight Multi-Scale Transformer Framework for Cross-Domain Bearing Fault Diagnosis Using Acoustic Emission Signals"*** by **Aman Kumar Gaur, Nipun Bansal, Yashasvi Bansal** (Abstract ID: 711 – **Oral Presentation**). Your presence at the event would be a great honor for ENEFM 2026.

We are sending this letter to you in hopes of possible financial support from your department as well as a VISA grant.

Thank you for your contribution, and we look forward to your participation at ENEFM 2026.

Sincerely,

Prof. A. Yavuz Oral,
Congress Planning Chair
E-Mail: **aoral@gtu.edu.tr**
Web: https://www.enefmcongress.org/

P.S. If you are unable to attend, please let us know at least 1 month in advance.