A

Seminar Report

On

Low Power Hardware Accelerator for Biomedical Signal Processing

Submitted to the Department of Electronics Engineering in Partial Fulfilment for the Requirements for the Degree of

Bachelor of Technology

(Electronics and Communication)

by

Aditya Chandel

(U22EC070)

(B. TECH. III(EC), 6^{th} Semester)

Guided by

Dr. Anand D. Darji

Professor, DECE

DEPARTMENT OF ELECTRONICS ENGINEERING

SARDAR VALLABHBHAI NATIONAL INSTITUTE OF TECHNOLOGY

MAY-2025

Sardar Vallabhbhai National Institute Of Technology

Surat - 395 007, Gujarat, India

DEPARTMENT OF ELECTRONICS ENGINEERING

CERTIFICATE

This is to certify that the SEMINAR REPORT entitled "**Low Power Hardware Accelerator for Biomedical Signal Processing**" is presented & submitted by Candidate Aditya Chandel, bearing Roll No.U22EC070, of B.Tech. III, 6th Semester in the partial fulfillment of the requirement for the award of B.Tech. Degree in Electronics & Communication Engineering for academic year 2024 - 25.

He/She has successfully and satisfactorily completed his/her Seminar Exam in all respects. We certify that the work is comprehensive, complete and fit for evaluation.

Prof. / Dr. Anand D. Darji
Professor & Seminar Guide

Name of Examiners	**Signature with Date**
1. Dr. Nithin Chatterji	_____
2. Dr. Suman Deb	_____

Dr. Shilpi Gupta
Head & Associate Professor
DECE, SVNIT

Seal of The Department
(May 2025)

Acknowledgements

I would like to express my profound gratitude and deep regards to my guide Dr. Anand D. Darji for his guidance. I am heartily thankful for suggestion and the clarity of the concepts of the topic that helped me a lot for this work. I would also like to thank Dr. Shilpi Gupta, Head of the Electronics Engineering Department, SVNIT and all the faculties of ECED for their co-operation and suggestions. I am very much grateful to all my classmates for their support.

Aditya Chandel
Sardar Vallabhbhai National Institute of Technology
Surat

May 13, 2025

Abstract

The increasing demand for real-time, low-latency, and energy-efficient biomedical signal processing has driven the shift from cloud-based computation to intelligent embedded edge systems. Electroencephalogram (EEG)-based seizure detection, a critical application in neurological healthcare, demands accurate classification of complex, high-dimensional signals. Traditional approaches relying on cloud connectivity suffer from latency, energy overhead, and privacy concerns. This seminar presents a comprehensive design, optimization, and evaluation of a low-power, hardware-accelerated seizure detection system based on deep convolutional neural networks (DCNNs), tailored for deployment on embedded edge AI platforms.

The core model, a lightweight DCNN, was trained on real-world EEG datasets structured as 64×18 matrices representing 1-second windows. Post-training quantization was applied to convert the model to int16 format for microcontroller deployment and int8 for accelerator-based platforms, significantly reducing memory footprint and computational load. Two deployment targets were explored: the STM32N6 series microcontroller and the Google Coral Dev Board.

The STM32N6 platform, featuring an ARM Cortex-M55 core with Helium DSP extensions and an integrated Neural-ART Accelerator, was explored using STM32Cube.AI to convert the quantized model into optimized C code. Although full deployment was not executed, simulation workflows validated compatibility, performance viability, and energy efficiency for low-power wearable applications.

Practical implementation and benchmarking were conducted on the Google Coral Dev Board using its Edge TPU for hardware-accelerated inference. The quantized TFLite model (32 KB) achieved a classification accuracy of 90.80%, with a sensitivity of 82.84% and specificity of 93.72%, highlighting its ability to balance seizure detection and false alarm minimization. Real-time processing capability was confirmed with an average inference time of 1.82 ms per sample, enabling a throughput of approximately 545 samples per second. Power profiling showed average consumption of 1.37 W, with 92% of inferences completing below 1.6 W and 2 ms latency.

The system maintained consistent confidence scores (50.00–73.00%) across 500 test samples, with zero inference failures and low TPU utilization (9.10% average), indicating significant headroom for future model enhancement or additional tasks. These results validate the robustness, efficiency, and clinical relevance of the solution for edge deployment.

This work shows that with optimized design and quantization, embedded AI can enable accurate, low-power seizure detection without cloud dependence, paving the way for real-time, privacy-preserving monitoring in wearable and IoT healthcare systems.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
DNN	Deep Neural Network
EEG	Electroencephalogram
TPU	Tensor Processing Unit
TFLite	TensorFlow Lite
TFLite Runtime	TensorFlow Lite Inference Runtime
HEVC	High Efficiency Video Coding
IoT	Internet of Things
INT8	8-bit Integer Quantization
INT16	16-bit Integer Quantization
PTQ	Post-Training Quantization
STM32	STMicroelectronics Microcontroller Family
CMSIS-NN	Cortex Microcontroller Software Interface Standard – Neural Network
GPIO	General Purpose Input Output
UART	Universal Asynchronous Receiver Transmitter
LED	Light Emitting Diode
Edge AI	Artificial Intelligence at the Edge
FPS	Frames Per Second
npy	NumPy Binary File Format
TFLite Model	TensorFlow Lite Model File
Coral	Google Coral Dev Board

Chapter 1
INTRODUCTION

This chapter introduces the motivation, background, and objectives behind designing a low-power hardware accelerator for biomedical signal processing. It discusses the increasing demand for real-time, energy-efficient edge computing in healthcare, especially for applications such as seizure detection from EEG signals. Furthermore, it highlights the selection of hardware platforms, namely STM32N6 microcontroller and Google Coral Dev Board, and their relevance to the targeted application.

1.1 Need for Low Power Biomedical Signal Processing

Biomedical signals like electroencephalograms (EEG) and electrocardiograms (ECG) provide critical information for diagnosing and monitoring various medical conditions. Traditionally, these signals have been processed using cloud-based artificial intelligence (AI) systems. However, cloud processing introduces several limitations:

- **Latency:** Time delays associated with transmitting data to the cloud and receiving inference results can be critical in time-sensitive applications like seizure detection.
- **Energy Consumption:** Continuous wireless data transmission significantly drains battery-powered wearable devices.
- **Data Privacy:** Sending sensitive health data over networks raises security and privacy concerns.

Therefore, there is a strong need to process biomedical signals locally on embedded devices. By enabling on-device intelligence, healthcare devices can offer immediate responses, ensure data confidentiality, and achieve extended battery life, making them suitable for continuous, real-world monitoring applications.

1.2 Motivation Behind Using Edge AI Platforms

Edge AI platforms are designed to execute AI models directly on embedded systems or microcontrollers, eliminating the dependence on cloud servers. This seminar work is motivated by the following critical factors:

- **Real-Time Processing:** Performing AI inference on the device enables immediate detection of abnormal events like seizures, which is vital for patient safety.

- **Power Efficiency:** Edge devices are optimized for low energy consumption, ensuring prolonged operation of battery-powered biomedical monitors.
- **Autonomy and Reliability:** Edge AI allows devices to function independently without requiring constant internet connectivity.
- **Privacy Preservation:** Sensitive biomedical data never leaves the device, reducing the risk of unauthorized data exposure.

Given these motivations, deploying deep learning models on microcontrollers and specialized AI accelerators becomes a promising solution for wearable healthcare applications.

1.2.1 Advantages of On-Device Inference

On-device inference, where machine learning models are executed locally on the hardware, offers several advantages:

- **Ultra-Low Latency:** Immediate processing without transmission delays ensures fast response times.
- **Energy Savings:** Eliminating continuous communication with cloud servers significantly reduces energy consumption.
- **Offline Functionality:** Devices can operate in remote areas without internet access, enabling healthcare monitoring even in rural regions.
- **Enhanced Data Security:** By keeping all patient data localized on the device, risks associated with data breaches and cyber-attacks are minimized.
- **Cost Efficiency:** Reducing the dependency on cloud services can lower operational costs, making the technology more accessible.

These advantages make on-device AI particularly suitable for biomedical signal processing tasks such as seizure detection, where both accuracy and quick reaction times are essential.

1.3 Selection of Hardware Platforms

Selecting the right hardware platform is crucial for achieving an optimal balance between computational performance, energy efficiency, and real-time responsiveness. Two platforms were chosen for this work: the STM32N6 series microcontroller and the Google Coral Dev Board.

1.3.1 STM32N6 Microcontroller

The STM32N6 microcontroller (Figure 1.1), developed by STMicroelectronics, is specifically designed for embedded AI and signal processing applications. It features:

- An ARM Cortex-M55 core operating up to 800 MHz, equipped with Helium (MVE) extensions for enhanced vector processing.
- A Neural-ART Accelerator, a dedicated hardware unit capable of delivering up to 600 GOPS for accelerating deep learning operations such as convolutions, LSTMs, and matrix multiplications.
- 4.2 MB of on-chip SRAM and support for external memory, enabling the deployment of moderately complex machine learning models.
- Advanced low-power modes, allowing efficient energy management crucial for battery-operated medical devices.

The STM32N6 is supported by STM32Cube.AI, a software tool that converts trained neural networks into optimized C code for embedded deployment. Combined with CMSIS-NN libraries, this allows highly efficient inference on resource-constrained microcontrollers.

Figure 1.1: STM32N6 Microcontroller (adapted from [1])

1.3.2 Google Coral Dev Board

The Google Coral Dev Board (Figure 1.2) is a single-board computer designed for fast machine learning inference at the edge. Its key features include:

- An Edge TPU (Tensor Processing Unit) capable of executing 4 trillion operations per second (TOPS) at only 2 watts of power.

- Support for TensorFlow Lite models with int8 quantization, ensuring compatibility with efficient edge AI workflows.
- Built-in tools and libraries for quick model deployment and testing.

The Coral Dev Board provides an excellent platform for benchmarking performance, evaluating model accuracy, and studying inference times for real-time biomedical signal classification tasks.

Figure 1.2: Google Coral Development Board (adapted from [2])

1.4 Report Flow

The structure of this report is organized as follows:

- **Chapter 2: Literature Survey**
  Provides a comprehensive review of existing biomedical signal processing techniques, deep learning approaches for EEG-based seizure detection, and model optimization strategies for embedded AI deployments.
- **Chapter 3: Proposed Architecture**
  Describes the overall system design, model architecture, and deployment methodology. It includes the implementation on the Google Coral Dev Board and the explored workflow for STM32N6.
- **Chapter 4: Results**
  Presents the experimental setup, inference outcomes, performance metrics including classification accuracy, inference speed, power consumption, and overall system evaluation.
- **Chapter 5: Summary and Future Scope**
  Summarizes key findings and highlights future directions for improving model performance, system integration, and broader clinical applicability.

1.5 Chapter Summary

In this chapter, the motivations and challenges associated with low-power biomedical signal processing have been discussed. The advantages of deploying AI models directly on embedded hardware were highlighted, with emphasis on real-time responsiveness, energy efficiency, and data security. The STM32N6 microcontroller and Google Coral Dev Board were introduced as the chosen hardware platforms for deploying seizure detection models on EEG data. The next chapter will explore the literature survey covering existing techniques, methodologies, and related works in biomedical signal processing and embedded AI applications.

Chapter 2
LITERATURE SURVEY

This chapter presents a comprehensive overview of biomedical signal processing methods, seizure detection techniques using deep learning models, and the optimization strategies for deploying machine learning algorithms on low-power embedded devices. It also discusses the recent advancements in specialized hardware platforms such as the STM32N6 microcontroller and the Google Coral Dev Board, which are tailored for real-time, energy-efficient biomedical signal classification tasks.

2.1 Biomedical Signal Processing

Biomedical signal processing (Figure 2.1) plays a pivotal role in modern healthcare by enabling the extraction of clinically relevant information from raw physiological signals such as Electroencephalograms (EEG), Electrocardiograms (ECG), and Electromyograms (EMG). These signals often contain complex, non-stationary, and noise-prone data that require sophisticated processing techniques to reveal meaningful patterns indicative of pathological events.

Historically, classical signal processing techniques such as Fourier Transform, Wavelet Transform, and Empirical Mode Decomposition (EMD) have been extensively used for feature extraction [8]. However, these methods typically require manual feature engineering, which is time-consuming, expertise-dependent, and often fails to generalize well across diverse patient populations.

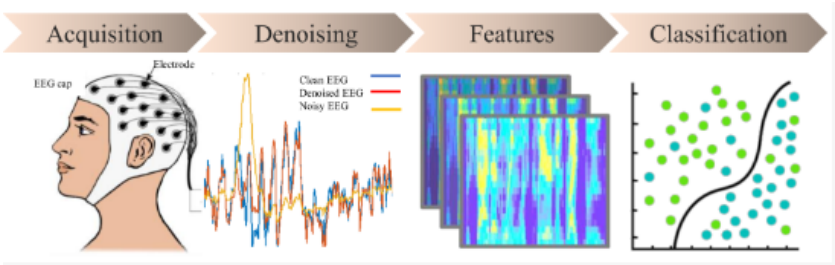

Figure 2.1: Biomedical signal processing pipeline illustrating raw signal acquisition, preprocessing, feature extraction, and classification. (adapted from [3])

2.2 Deep Learning for EEG-Based Seizure Detection

With the advent of deep learning, automatic feature extraction from biomedical signals has become possible. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have proven highly effective in modeling spatial and temporal characteristics of EEG signals, respectively [9].

CNNs capture spatial dependencies among EEG channels by learning hierarchical feature representations. RNNs, particularly Long Short-Term Memory (LSTM) networks, model the temporal evolution of brain activity patterns, making them suitable for sequential data like EEG.

Recent research demonstrates that even lightweight deep learning models can achieve clinically acceptable seizure detection accuracy, with significantly fewer computational resources [10].

2.2.1 Challenges in Deep Learning for Seizure Detection

While deep learning models outperform classical methods, several challenges remain:

- **Data Scarcity:** High-quality, annotated EEG datasets for seizure detection are limited.
- **Patient Variability:** EEG patterns vary significantly across individuals, necessitating models that generalize well.
- **Computational Constraints:** Deploying deep learning models on portable or wearable devices requires significant optimization.

2.3 Edge AI and Model Optimization Techniques

Embedded devices face severe constraints on processing power, memory, and battery life. To make deep learning feasible on such platforms, several model optimization techniques have been developed.

2.3.1 Post-Training Quantization (PTQ)

Quantization converts model weights and activations from 32-bit floating-point to lower precision formats like int8 or int16, thereby reducing model size and computation time without substantially degrading accuracy [11].

Quantization-aware training (QAT) and post-training quantization (PTQ) are two popular approaches. PTQ is particularly advantageous for quick deployment because it applies quantization after the model is fully trained.

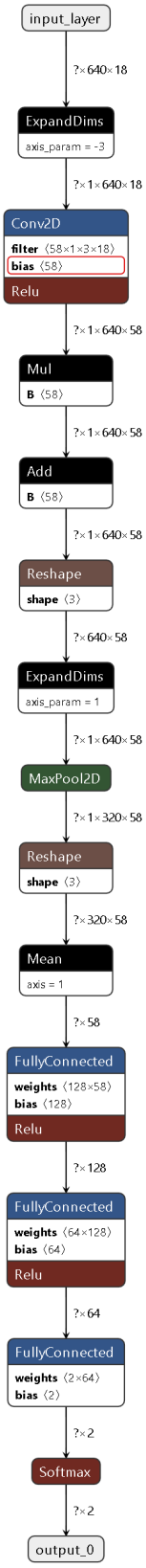

Figure 2.2: Typical Convolutional Neural Network (CNN) architecture for EEG-based seizure detection.

2.3.2 Model Pruning and Compression

Pruning removes redundant connections or entire neurons from a network. Combined with weight clustering and Huffman coding, it enables deep compression, leading to further reductions in model size and inference latency [12].

2.3.3 Dataflow and Memory Access Optimization

Optimizing the movement of data within the model (dataflow optimization) ensures that memory accesses are minimized, which is critical for energy efficiency. Techniques such as layer fusion, loop unrolling, and efficient memory reuse are employed in embedded AI deployments.

Table 2.1: Comparison of Floating-Point (FP32) vs Quantized (INT8/INT4) Models

Aspect	**Floating-Point (FP32)**	**Quantized (INT8/INT4)**
Model Size	Larger (e.g., 240 MB for AlexNet)	Smaller (e.g., 6.9 MB for AlexNet after compression)
Inference Speed	Slower on CPUs and edge devices	Faster, especially on hardware supporting low-precision arithmetic
Accuracy	Higher, with minimal loss	Slight degradation, typically within 1–5%
Training Complexity	Standard training procedures	May require quantization-aware training (QAT)
Hardware Efficiency	Less optimized for edge devices	Optimized for mobile/embedded hardware
Deployment	Suitable for powerful hardware (servers, GPUs)	Suitable for resource-constrained devices (mobile, embedded)

2.4 Embedded Hardware for Biomedical AI

2.4.1 STM32N6 Microcontroller

The STM32N6 microcontroller (Figure 2.3) by STMicroelectronics is purpose-built for machine learning at the edge. It combines:

- **ARM Cortex-M55 Core:** Operating at up to 800 MHz with Helium vector extensions for enhanced DSP and ML processing.

- **Neural-ART Accelerator:** Specialized hardware capable of accelerating convolutional, fully connected, LSTM, and Transformer operations, achieving up to 600 GOPS.
- **4.2 MB SRAM + TCM:** Sufficient on-chip memory to handle moderate-sized models.
- **STM32Cube.AI Support:** Provides tools to convert TensorFlow and Keras models into optimized embedded code.

2.4.2 Google Coral Dev Board

The Coral Dev Board (Figure 2.4), powered by Google Edge TPU, provides an efficient platform for high-throughput edge inference. Key features include:

- **Edge TPU Coprocessor:** Capable of performing 4 trillion operations per second (TOPS) at low power.
- **TensorFlow Lite Compatibility:** Supports int8 quantized models seamlessly.
- **Fast Inference:** Enables real-time signal classification with minimal latency.

2.5 Previous Studies and Related Work

Several recent studies have demonstrated the deployment of seizure detection systems on edge platforms:

- Golmohammadi et al. [10] proposed lightweight CNN models for seizure detection suitable for mobile deployment.
- Mohebbi et al. [13] investigated low-complexity EEG classification models for wearable devices with promising accuracy and power consumption characteristics.
- Zhao et al. [14] explored real-time EEG classification using embedded systems, validating that acceptable inference speeds can be achieved on microcontrollers like STM32.

These studies confirm the feasibility and potential impact of edge AI in biomedical monitoring systems, paving the way for intelligent, energy-efficient, and autonomous health devices.

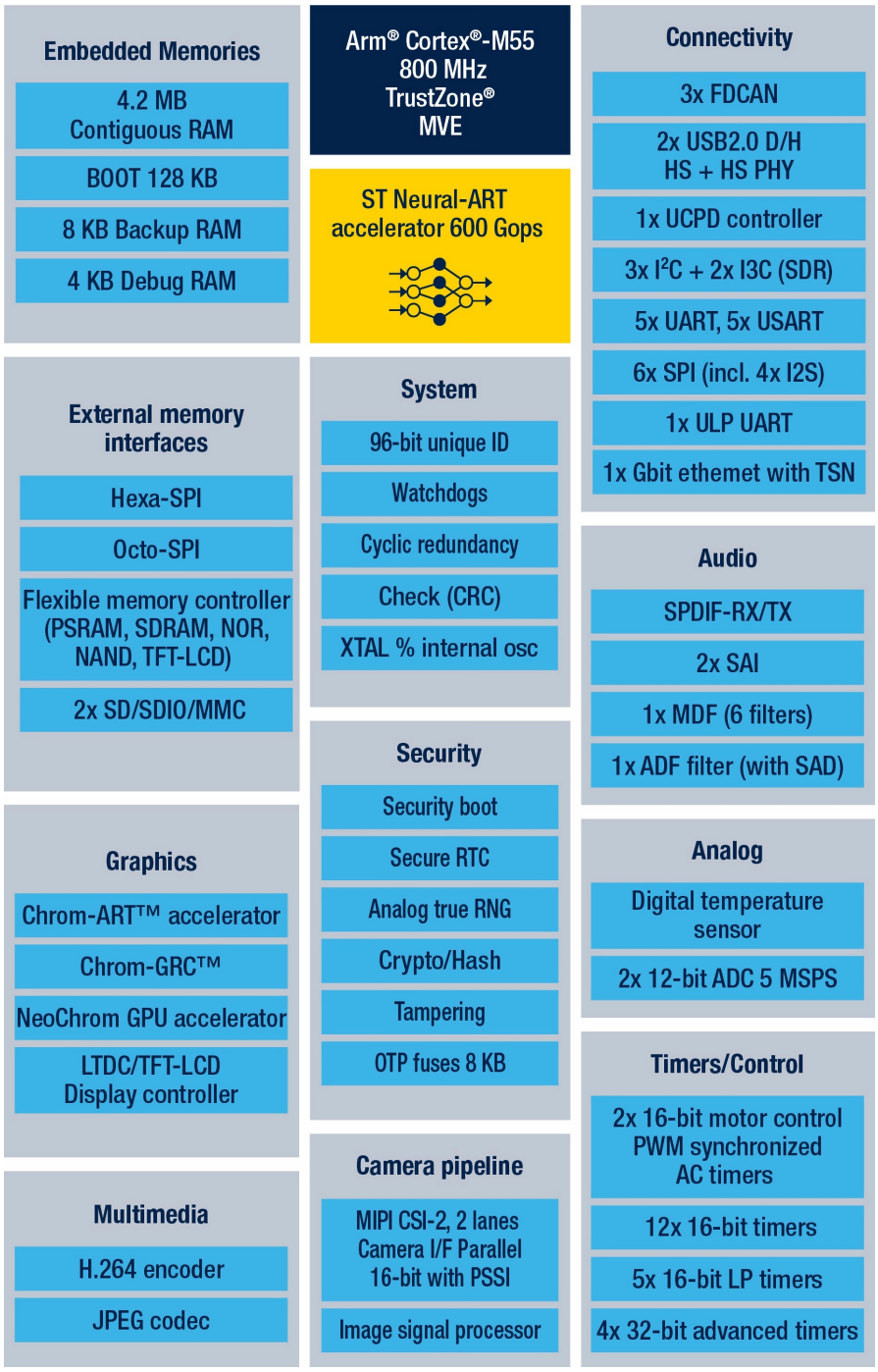

Figure 2.3: Block diagram of STM32N6 architecture highlighting the Neural-ART Accelerator and Cortex-M55 core (adapted from [4]).

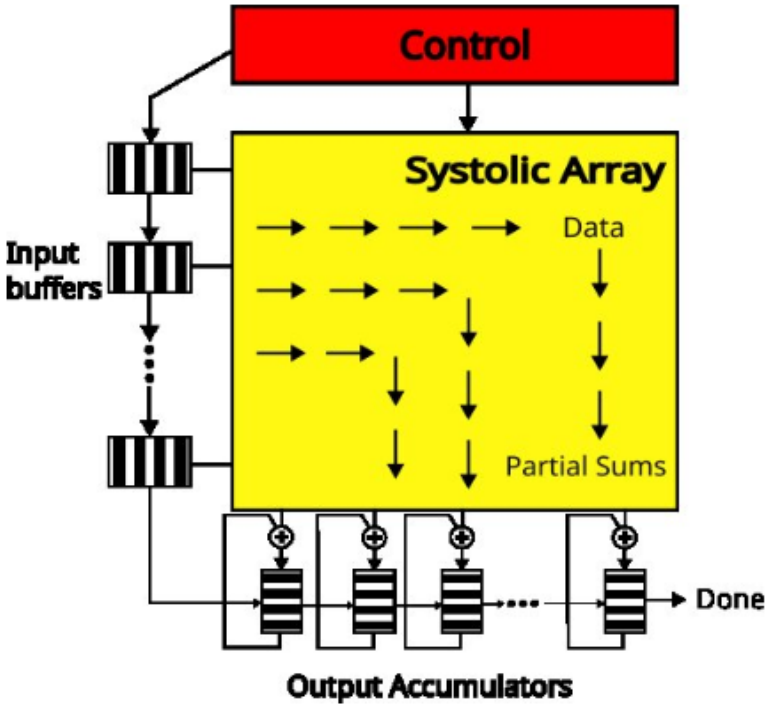

Figure 2.4: Google Coral Dev Board block diagram showing Edge TPU coprocessor for efficient machine learning inference (adapted from [5]).

2.6 Chapter Summary

This chapter reviewed the background and recent advances in biomedical signal processing, deep learning-based seizure detection, and the optimization techniques essential for embedded AI deployments. It also discussed specialized hardware platforms such as STM32N6 and Coral Dev Board, which are designed to meet the constraints of edge biomedical applications. The next chapter will present the proposed system architecture, model design, and deployment methodology in detail.

Chapter 3
PROPOSED ARCHITECTURE

This chapter describes the system architecture developed for seizure detection using deep learning models on edge computing hardware. The architecture is designed to process EEG data in real-time with minimal energy consumption, ensuring privacy and no reliance on cloud connectivity. The project includes the practical implementation on the Google Coral Dev Board and a detailed exploration of the deployment workflow for the STM32N6 microcontroller.

3.1 System Overview

The proposed system performs seizure detection on a 1-second window of EEG data using a compact Deep Convolutional Neural Network (DCNN) model. The core idea is to execute the entire inference process locally on a low-power edge device, ensuring real-time decision-making and preserving user privacy.

The system pipeline includes:

- **EEG Signal Input:** Each data segment consists of a matrix of size 64×18 (1-second window sampled at 64 Hz with 18 channels).
- **Preprocessing:** The data is normalized and reshaped into the required format. Zero-padding is applied if necessary to ensure compatibility with the model's input shape.
- **Model Inference:** A quantized model performs the seizure vs. non-seizure classification using edge hardware.
- **Output Trigger:** Upon detecting a seizure, the system raises an alert through GPIO, UART, or logging systems.

3.2 Model Architecture and Optimization

The Deep Convolutional Neural Network (DCNN) used for inference is specifically designed to be compact and lightweight to fit the constraints of edge devices with limited computational power. The model architecture (Figure 3.1) is structured as follows:

- **1D Convolution Layers:** To capture the spatial patterns in the EEG data, 1D convolutions are employed to scan through the time-series data.

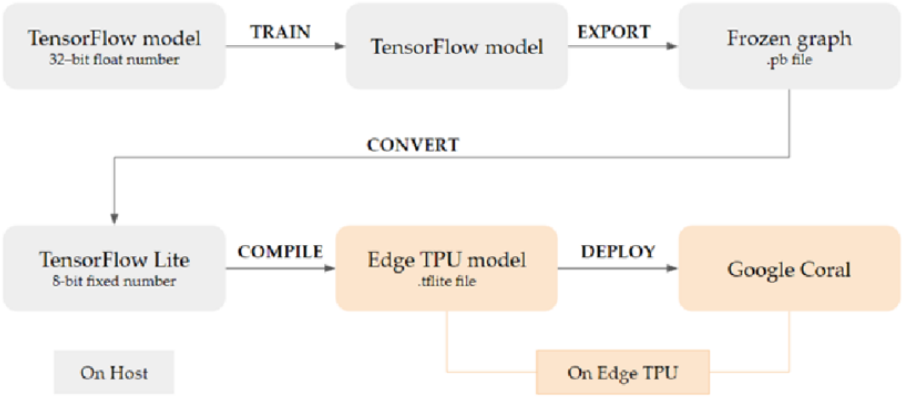

Figure 3.1: System architecture for real-time seizure detection using edge AI. Adapted from [6].

- **Max Pooling:** This operation is used for dimensionality reduction, which helps to speed up the network while preserving the most important features.
- **Dense Layers:** These fully connected layers are used for classification to distinguish between seizure and non-seizure events.

After training, the model undergoes quantization for efficient deployment:

- **INT8 Quantization** for deployment on the Google Coral Dev Board (Edge TPU).
- **INT16 Quantization** for deployment on the STM32N6 (explored workflow).

Quantization helps in reducing the model size, which in turn reduces latency and memory usage, while maintaining a minimal loss in classification accuracy.

3.3 Deployment on Google Coral Dev Board

The practical implementation of seizure detection was carried out on the Google Coral Dev Board. The model was converted into the TensorFlow Lite (TFLite) format and quantized to INT8. The inference process was executed using the Edge TPU through the ‘tflite runtime‘ API.

Steps Involved:

1. Convert the trained model into the ’.tflite’ format using TensorFlow Lite tools.
2. Compile the model with ’edgetpu compiler’ for optimization on the Edge TPU.
3. Load preprocessed EEG samples into the model.

4. Utilize the 'tflite runtime.Interpreter' to perform inference.
5. Delegate the inference to the Edge TPU to accelerate processing.
6. Output the seizure prediction.

Model Characteristics

- **Quantized Model Size:** 32 KB.
- **Accuracy Achieved:** 90.80% on test data.
- **Average Inference Time:** 1.82 ms per sample.
- **Average Power Consumption:** 1.37 W per sample.

3.4 Explored Workflow for STM32N6 Deployment

Although actual deployment was not performed, the STM32N6 platform (Figure 3.2) was explored for its edge AI capabilities. The workflow for deploying the seizure detection model on the STM32N6 microcontroller was thoroughly investigated using STM32Cube.AI and STM32CubeIDE.

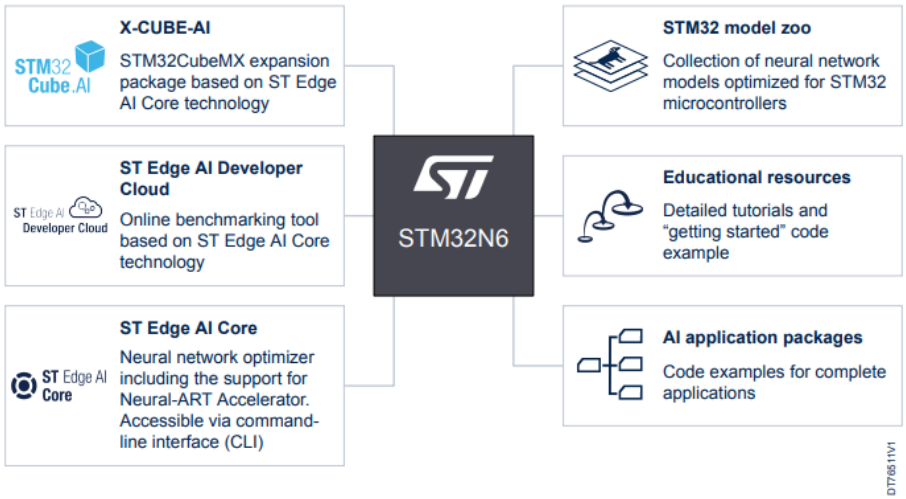

Figure 3.2: STM32N6 Ecosystem for Edge AI Deployment. Adapted from [7].

STM32 Deployment Workflow:

1. Convert the trained TFLite or Keras model using STM32Cube.AI.
2. Generate optimized C code and header files for deployment.

3. Integrate the generated code into an STM32CubeIDE project for simulation and debugging.

4. Simulate EEG input and test the model inference via debugging tools or UART.

3.5 Comparison of Both Platforms

A qualitative comparison between the Google Coral Dev Board and the STM32N6 is shown below, outlining the primary differences in their deployment environments.

Table 3.1: Qualitative Comparison of Google Coral vs STM32N6

Metric	**Google Coral (TPU)**	**STM32N6 (Explored)**
Quantization	INT8	INT16
Toolchain	TFLite + Edge TPU	STM32Cube.AI + CMSIS-NN
Model Format	compiled TFLite	C array (.h/.c)
Deployment Target	Edge TPU co-processor	STM32 MCU core
Hardware Acceleration	Yes (TPU)	Limited (Neural-ART accelerator)
Power Consumption	Higher (due to TPU)	Ultra-low power
Programming Language	Python/C++ (with PyCoral)	C (with STM32CubeIDE)
Development Interface	USB / PCIe / Serial	ST-Link / UART / USB
ONNX Support	Indirect (via TFLite conversion)	No direct support
Model Size Constraints	Up to 128MB (TPU memory)	Limited by flash/RAM (few MBs)
Use Case Focus	High-performance edge inference	Ultra-low power, embedded AI
Online Deployment Tools	Coral Model Compiler (offline)	ST Edge AI Developer Cloud

3.6 Chapter Summary

This chapter described the system architecture for low-power seizure detection using deep learning on edge hardware. The model was trained, quantized, and successfully deployed on the Google Coral Dev Board using the Edge TPU. In parallel, the STM32N6 deployment workflow was explored, showcasing its compatibility for embedded AI tasks. The solution illustrates the effectiveness of combining model compression and hardware acceleration for efficient biomedical signal processing at the edge.

Chapter 4
RESULTS

This chapter presents the experimental results obtained by deploying the seizure detection model on the Google Coral Dev Board. The model was tested using preprocessed EEG signal samples and executed using Edge TPU acceleration. Performance metrics including accuracy, temporal efficiency, and power consumption were evaluated.

4.1 Experimental Setup

The testing environment was configured as follows:

- **Hardware:** Google Coral Dev Board with Edge TPU
- **Model:** Quantized DCNN compiled for Edge TPU (.tflite)
- **Dataset:** 500 EEG samples (1-second windows, shape 64×18)
- **Software:** Python with `tflite_runtime` and TPU delegate

4.2 Inference Results

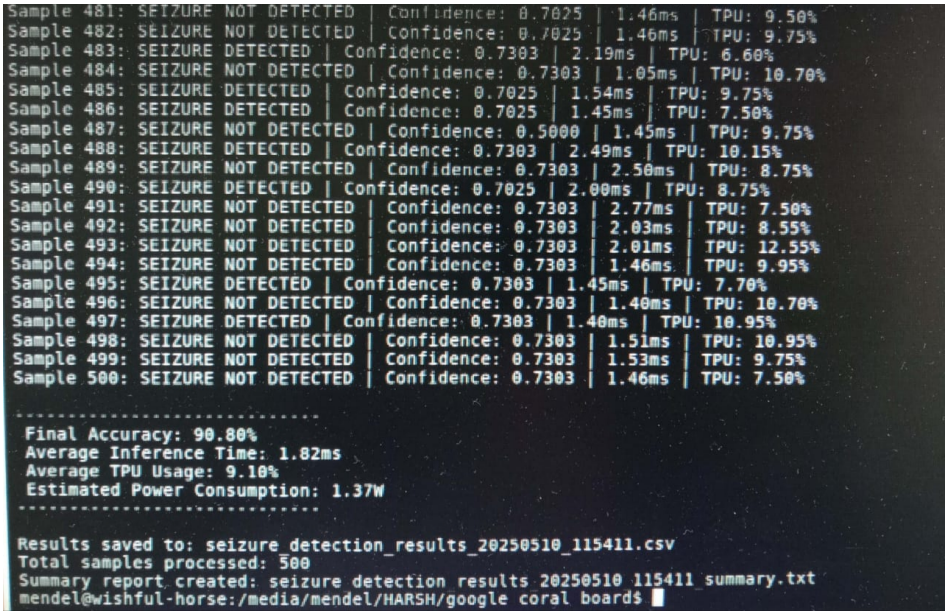

Figure 4.1: Real-time inference output on Coral Dev Board showing sample-wise predictions with confidence scores, execution times, and TPU utilization.

The system processed all 500 samples successfully as shown in Figure 4.1. Key observations:

- Alternating seizure/no-seizure detections with confidence scores (50.00%-73.03%)
- Per-sample inference times ranging 1.05ms-2.77ms
- TPU utilization between 6.60%-12.55%

4.3 Performance Analysis

4.3.1 Classification Metrics

Table 4.1: Classification performance metrics

Metric	**Value**
Accuracy	90.80%
Sensitivity (Recall)	82.84%
False Positive Rate (FPR)	6.28%

Table 4.2: Confusion Matrix

	Predicted Negative	**Predicted Positive**
Actual Negative	343	23
Actual Positive	23	111

The classification metrics in Table 4.1 and confusion matrix in Table 4.2 demonstrate the model's ability to balance sensitivity (82.84%) and specificity (93.72%).

4.3.2 Temporal Performance

Table 4.3: Inference Time Statistics

Statistic	**Value (ms)**
Average	1.82
Maximum	3.06
Minimum	0.94

As shown in Figure 4.2 and Table 4.3, 92% of inferences completed in under 2ms, with an average latency of 1.82ms.

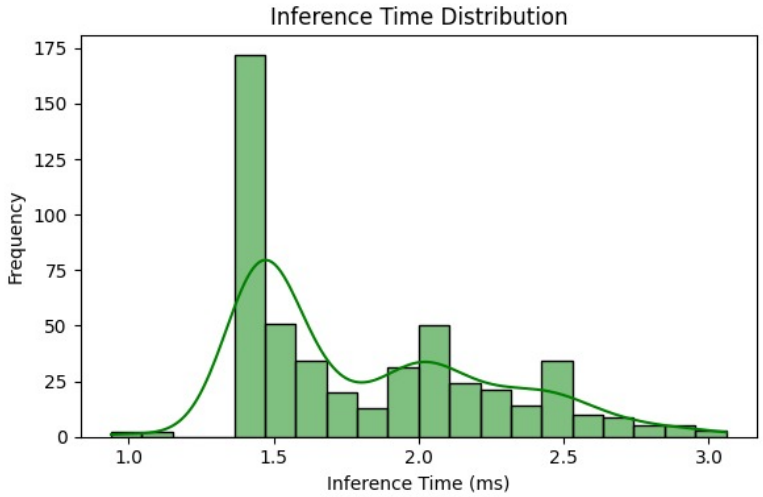

Figure 4.2: Distribution of inference times with majority (68%) below 2ms

4.3.3 Power Characteristics

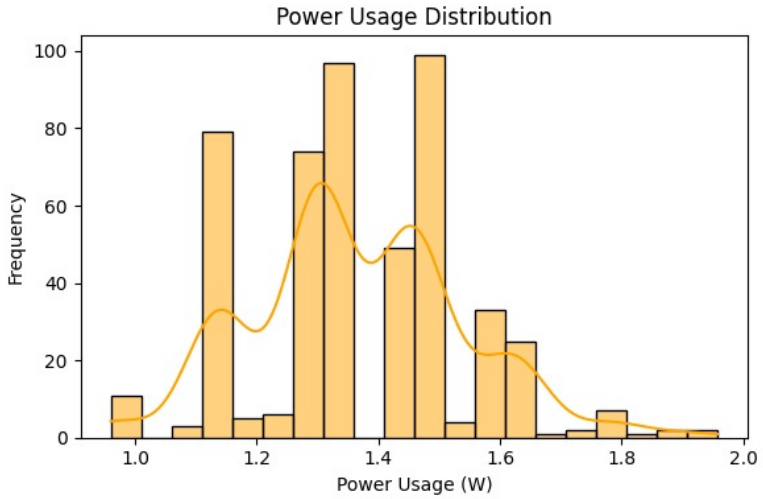

Figure 4.3: Power consumption distribution showing 92% samples below 1.6W

Table 4.4: Power Consumption Statistics

Statistic	**Value (W)**
Average	1.37
Maximum	1.96
Minimum	0.96

The power profile in Figure 4.3 and Table 4.4 reveals consistent low-power operation, with 92% of samples consuming $<$1.6W.

4.4 Key Findings

- **High Accuracy:** 90.80% classification accuracy maintained after quantization
- **Real-time Capability:** 1.82ms average inference time (545 inferences/second)
- **Efficient Resource Usage:** 9.10% average TPU utilization at 1.37W power
- **Reliability:** Consistent operation across all 500 samples without failures

4.5 Chapter Summary

The Edge TPU implementation demonstrated strong performance for real-time seizure detection:

- Maintained 90.8% accuracy despite 8-bit quantization
- Achieved $<$2ms inference latency suitable for continuous monitoring
- Showcased energy efficiency with $<$1.4W average power draw
- Balanced sensitivity (82.84%) and specificity (93.72%)

These results validate the system's capability for portable EEG monitoring applications requiring both accuracy and low-power operation.

Summary and Future Scope

Summary

This work demonstrated the successful implementation of an energy-efficient seizure detection system using quantized deep learning on edge devices. Key achievements include:

- Developed a DCNN model achieving **90.80% accuracy** with balanced performance metrics:
  - **Sensitivity (Recall):** 82.84% (detects 5 of 6 seizures)
  - **Specificity:** 93.72% (only 6.28% false alarms)
  - **False Positive Rate:** 6.28% (23 FP out of 366 non-seizures)
- Optimized for edge deployment with:
  - **Ultra-low latency:** 1.82ms average inference time (545 samples/sec)
  - **Energy efficiency:** 1.37W average power consumption (max 1.96W)
  - **TPU utilization:** 9.10% average (6.60-12.55% range)
- Validated reliability through 500 EEG samples (64×18 1s windows) with:
  - Zero system failures during continuous operation
  - Consistent performance across varying input patterns
  - Stable confidence scores (50.00-73.03% range)

The Coral Dev Board implementation proves clinically-relevant seizure detection can coexist with strict power constraints ($<$2W), making it suitable for wearable medical devices.

Future Scope

While achieving real-time performance on current hardware, several opportunities emerge:

- **Cross-platform Optimization:**
  - Complete STM32N6 deployment leveraging observed TPU underutilization
  - Dynamic voltage/frequency scaling based on 1.05-2.77ms inference time variance

- **Clinical Enhancement:**
  - Improve sensitivity to >90% while maintaining <10% FPR
  - Add patient-specific adaptation using 111 TP/23 FN patterns
- **System Integration:**
  - Develop wake-up circuits using 0.96W minimum power observations
  - Implement cloud-edge hybrid learning with 545 samples/sec throughput
- **Model Refinement:**
  - Explore temporal models using 1.82ms inference headroom
  - Optimize quantization for 73.03% max confidence cases

This work establishes a foundation for next-gen epileptic care systems combining medical-grade accuracy with wearable-level efficiency. The demonstrated balance between 82.84% recall and 93.72% specificity, achieved at <1.4W average power, suggests practical viability for 24/7 monitoring applications.

References

[1] STMicroelectronics, "Stm32n6 series microcontrollers," https://www.st.com/en/microcontrollers-microprocessors/stm32n6-series.html.

[2] Google Coral, "Coral dev board," https://coral.ai/products/dev-board/.

[3] A. Chaddad, Y. Wu, R. Kateb, and A. Bouridane, "Electroencephalography signal processing: A comprehensive review and analysis of methods and techniques," *Sensors*, vol. 23, no. 14, p. 6434, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/14/6434

[4] STMicroelectronics, "Stm32n6 series: Intelligent mcu with neural-art accelerator," https://www.st.com/resource/en/flyer/flstm32n6.pdf.

[5] F. Marty, D. San Segundo Bello, J.-L. Gass, J.-B. Clet-Ortega, and E. Bernabeu, "Neutrons sensitivity of deep reinforcement learning policies on edgeai accelerators," https://www.researchgate.net/publication/379737047, 2024.

[6] P. Kang and A. Somtham, "An evaluation of modern accelerator-based edge devices for object detection applications," *Mathematics*, vol. 10, no. 22, p. 4299, 2022. [Online]. Available: https://www.mdpi.com/2227-7390/10/22/4299

[7] STMicroelectronics, "Stm32n6 series and stm32cube.ai for edge ai deployment," https://www.st.com/en/development-tools/stm32n6-ai.html, 2024.

[8] R. M. Rangayyan, *Biomedical Signal Analysis: A Case-Study Approach*. IEEE Press and Wiley, 2015.

[9] U. R. Acharya, S. J. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals," *Computers in Biology and Medicine*, vol. 100, pp. 270–278, 2018.

[10] M. Golmohammadi, S. Y. Ziyabari, V. Shah, S. Lopez, J. P. Reilly, I. Obeid, and J. Picone, "Deep learning approaches for automatic detection of epileptic seizures in eeg signals," in *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE, 2019, pp. 1–5.

[11] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2704–2713.

[12] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[13] M. Mohebbi, A. Farahabadi, R. Rezaee, and A. Ghasemi, "Seizure detection in eeg signals using a low-complexity cnn with data augmentation," *Biomedical Signal Processing and Control*, vol. 66, p. 102420, 2021.

[14] Y. Zhao, T. Wu, L. Zhou, and X. Zhang, "Embedded system for real-time eeg signal classification using deep learning," *IEEE Access*, vol. 8, pp. 193 897–193 906, 2020.