Scalable Binary Deep Neural Network Accelerator with Adaptive Temporal Analysis for Asthma and Cough Audio Classification

Machika Navyasri[1], Tangerala Pradeep Kumar[1*]
[1]Department of Electronics & Communication Engineering, Vaagdevi Engineering College, Warangal, 506005, Telangana, India.
*Correspondence: Tangerala Pradeep Kumar (johny5508@gmail.com)

To Cite this Article
Machika Navyasri, Tangerala Pradeep Kumar, "Scalable Binary Deep Neural Network Accelerator with Adaptive Temporal Analysis for Asthma and Cough Audio Classification", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 07, July 2026, pp: 31-47, DOI: http://doi.org/10.64771/jsetms.2026.v03.i07.pp31-47
Submitted: 17-05-2026 *Accepted: 24-06-2026* *Published: 01-07-2026*

Abstract
Recent clinical and embedded sensing studies indicate that always-on respiratory audio monitoring including asthma wheeze detection and cough classification can consume more data in battery-powered wearable and Internet of Medical Things (IoMT) devices. Further, real-time respiratory sound classification is a critical requirement for wearable health monitors, smart inhalers, remote patient monitoring systems, and edge-based medical diagnostic sensors, where strict constraints on energy, memory, and inference latency prohibit conventional deployment. Traditional audio-based respiratory classifiers rely on fixed-precision Deep Neural Networks (DNNs) with high memory bandwidth and multiplier-intensive computation, making them unsuitable for long-term wearable operation. Existing BNN-based medical audio accelerators often employ static binarization and fixed temporal smoothing, leading to high output entropy, sensitivity to transient noise, and frequent false positives or missed pathological events. Furthermore, rigid hardware architectures restrict scalability and limit adaptation to new respiratory conditions or evolving clinical thresholds. To address these challenges, this work proposes a VLSI-optimized respiratory audio classification architecture based on an enhanced Binary Deep Neural Network. The design integrates a Scalable Hierarchical Binary Compute Array (SHBCA) for parallel and area-efficient binary computation, an Adaptive Entropy-Aware Binary Engine (AEABE) that dynamically regulates binarization thresholds to stabilize diagnostic confidence, and a Latency-Constrained Temporal Fusion Module (LCTFM) that exploits temporal correlations in respiratory sounds while maintaining bounded response time. Additionally, a Reconfigurable Respiratory Adaptation Fabric (RRAF) enables on-chip reconfiguration of disease-specific parameters without full model retraining. Collectively, the proposed architecture achieves a balanced trade-off between diagnostic accuracy, latency, and energy efficiency, making it well suited for edge-oriented IoMT respiratory monitoring systems.

Key words: Respiratory Audio Classification, Very Large-Scale Integration (VLSI), Internet of Medical Things (IoMT), Edge Artificial Intelligence (Edge AI), Low-Power Hardware Accelerator

This is an open access article under the creative commons license https://creativecommons.org/licenses/by-nc-nd/4.0/
CC BY-NC-ND 4.0

1. Introduction

Respiratory diseases represent one of the most significant global healthcare challenges, affecting millions of people every year and contributing substantially to morbidity, mortality, and healthcare expenditure. According to international health reports, chronic respiratory diseases such as asthma, chronic obstructive pulmonary disease (COPD), pneumonia, bronchitis, and other pulmonary disorders affect more than 500 million people worldwide. Asthma alone impacts over 260 million individuals globally and is responsible for hundreds of thousands of deaths annually. Furthermore, respiratory illnesses account for a considerable percentage of hospital admissions and emergency care visits, particularly among children, elderly populations, and individuals with compromised immune systems. The increasing prevalence of respiratory diseases, combined with urbanization, environmental pollution, smoking habits, occupational exposure, and changing climatic conditions, has intensified the need for efficient and accessible respiratory health monitoring systems.

Traditional respiratory disease diagnosis primarily depends on clinical examination, spirometry testing, imaging procedures, and expert interpretation of respiratory sounds. Although these methods provide valuable diagnostic information, they often require specialized equipment, trained healthcare professionals, and dedicated healthcare facilities. In many rural and resource-constrained regions, access to such medical infrastructure remains limited, resulting in delayed diagnosis and treatment. Respiratory sounds such as wheezes, crackles, rhonchi, and cough patterns contain important clinical information that can assist in the early detection of pulmonary abnormalities. With the widespread availability of digital stethoscopes, microphones, mobile devices, and wearable healthcare technologies, respiratory audio analysis has emerged as an attractive non-invasive diagnostic approach capable of supporting continuous patient monitoring and early disease identification.

Recent advancements in artificial intelligence, machine learning, deep learning, and hardware acceleration technologies have significantly transformed medical audio analysis. Large volumes of respiratory sound recordings can now be processed automatically to identify disease-specific acoustic patterns that may not be easily detectable through manual examination. Simultaneously, the growing demand for real-time healthcare applications has increased the importance of efficient hardware implementations capable of processing respiratory signals with low latency and reduced power consumption. The integration of intelligent signal processing, neural network-based classification, and VLSI design methodologies has created new opportunities for developing portable, scalable, and high-performance respiratory disease diagnosis systems suitable for modern healthcare environments.

1.1 Research Objectives

- To develop a SHBCA for efficient parallel binary feature processing of respiratory audio signals.
- To design an AEABE capable of dynamically regulating binarization thresholds for improved classification stability.
- To implement a LCTFM for effective temporal analysis of respiratory sound patterns and integrate a Scalable Binary Deep Neural Network (SB-DNN) architecture for accurate respiratory disease classification.
- To develop a RRAF that supports dynamic modification of disease-specific parameters.

- To implement and validate the proposed architecture using Xilinx Vivado-based FPGA platforms.

1.2 Research Contributions

- Introduction of a novel VLSI-oriented respiratory audio classification framework based on a Scalable Binary Deep Neural Network.
- Development of the SHBCA module for scalable and parallel binary computation.
- Design of the AEABE module for adaptive entropy-aware decision regulation.
- Development of the LCTFM module for low-latency temporal feature fusion.
- Introduction of hardware-level reconfigurability through the RRAF module.
- Integration of software-based audio preprocessing with hardware-accelerated respiratory classification.
- Reduction of hardware complexity through binary neural computation techniques.
- Enhancement of diagnostic confidence by minimizing classification uncertainty.
- Support for real-time respiratory monitoring and edge healthcare applications.

2. Literature Survey

Sanap et al. [1] developed the BCough platform for on-device cough detection using bone-conduction sensing technology integrated with embedded artificial intelligence. The methodology involved collecting cough signals through a bone-conduction sensor to reduce environmental noise interference. Audio signals were preprocessed and transformed into representative features suitable for machine learning analysis. A lightweight AI model was deployed directly on embedded hardware to enable local inference without cloud dependence. The system was experimentally evaluated for detection accuracy, real-time operation, and energy efficiency in wearable healthcare environments. The framework mainly focuses on cough event detection and offers limited support for multi-class respiratory disease classification. Vali et al. [2] introduced a low-power VLSI accelerator architecture for AI-enhanced real-time audio and video processing at the network edge. Their methodology incorporated hierarchical dataflow optimization, adaptive clock-gating mechanisms, and multi-core neural processing units. Dynamic voltage and frequency scaling were utilized to balance workload distribution and reduce energy consumption. Systolic-array-based computation was employed to accelerate AI inference tasks. The architecture was optimized to support multimedia processing under strict power constraints. The architecture was designed for generic multimedia applications and does not specifically address respiratory audio classification requirements. Juliet et al. [3] presented custom VLSI chip implementations to accelerate edge AI for real-time audio and video processing applications. The methodology focused on integrating AI inference engines directly into dedicated VLSI hardware. Custom processing modules were designed to improve throughput and reduce latency for multimedia workloads. Hardware-software co-design techniques were employed to optimize computational efficiency. The architecture supported edge deployment by minimizing dependency on centralized processing resources. The design primarily targets multimedia acceleration and lacks specialized mechanisms for medical audio analysis. Rimada and Mrinh [4] proposed an energy-efficient VLSI implementation for AI-assisted signal processing in multimedia systems. Their approach utilized hardware acceleration techniques to execute signal processing and AI operations simultaneously. Efficient data movement and optimized computational pipelines were incorporated to reduce energy consumption. Resource-sharing

strategies were adopted to improve silicon utilization. The architecture was validated through real-time multimedia processing experiments. The implementation focuses on general multimedia workloads and provides limited adaptability to healthcare-oriented audio applications.

Cho and Kim [5] designed a battery-powered Class-D audio amplifier with true-zero-switching capability for energy-efficient audio applications. The methodology incorporated a boosted power stage and optimized switching circuitry to achieve high efficiency. Advanced control mechanisms were introduced to reduce harmonic distortion and switching losses. The amplifier architecture minimized quiescent current while maintaining output quality. The work concentrates on audio amplification rather than intelligent respiratory sound analysis or classification. Jayachandran et al. [6] investigated advancements in memristor-based audio processing technologies for hearing aid systems. Their methodology utilized memristive devices to perform audio signal processing with reduced power requirements. Neuromorphic-inspired architectures were explored to emulate biological auditory functions. Adaptive filtering and sound enhancement mechanisms were integrated to improve hearing assistance. The study evaluated the potential of memristor-based circuits for compact healthcare devices. The approach primarily addresses hearing enhancement and does not perform respiratory disease identification. Kochar et al. [7] developed a real-time speech audio denoiser for wearable IoT devices using a quantized cascaded convolutional encoder-decoder architecture. The methodology involved applying convolutional neural networks for speech enhancement while maintaining ultra-low power operation. Model quantization techniques reduced memory and computational requirements. Dedicated hardware optimization enabled wearable deployment. The system focuses on speech noise reduction and does not include respiratory disease classification functionality. Jagan et al. [8] introduced Murmur, a secure and low-energy audio communication framework for the Internet of Bodies. Their methodology employed energy-efficient communication protocols combined with lightweight encryption mechanisms. Audio information was securely transmitted across body-centric networks. Resource-aware processing techniques minimized energy consumption during communication. The framework emphasizes secure communication rather than medical audio analysis and diagnosis.

Gao et al. [9] developed COSMIC, a heterogeneous multi-vector-core RISC-V SoC for intelligent audio DSP applications. The methodology integrated multiple vector processing cores with heterogeneous computing resources. Specialized DSP accelerators were incorporated to support intelligent audio workloads. The architecture enabled parallel execution of audio processing algorithms. Silicon implementation demonstrated high-performance audio analytics with energy efficiency. The design targets general-purpose audio DSP and lacks disease-specific respiratory analysis capabilities. Babu et al. [10] implemented a VLSI-based Cascaded Integrator Comb (CIC) filter for digital signal processing applications. The methodology employed cascaded integrator and comb stages to perform efficient filtering operations. Hardware optimization techniques were applied to reduce implementation complexity. Resource-efficient arithmetic structures improved throughput. FPGA validation confirmed successful filtering performance. The work is limited to signal filtering and does not perform intelligent classification tasks. Karthika [11] introduced a high-throughput VLSI hardware accelerator for embedded signal processing applications. The methodology focused

on parallel processing architectures to improve computational speed. Optimized datapaths and memory structures were integrated to support continuous data streams. Hardware acceleration techniques minimized processing bottlenecks. The accelerator lacks adaptive learning mechanisms required for respiratory disease diagnosis. Adinath et al. [12] presented low-power VLSI accelerators for edge AI in IoT devices. Their methodology incorporated lightweight neural processing units and power-aware hardware optimization strategies. Efficient memory management techniques reduced energy consumption. The architecture supported AI inference at the edge without cloud dependence. FPGA-based validation demonstrated improved resource efficiency. The framework is designed for generic IoT intelligence and does not specifically target respiratory sound analysis. Shah et al. [13] explored the integration of VLSI and neural networks within Industry 4.0 environments. Their methodology combined neural network processing with hardware acceleration to support industrial automation. Dedicated VLSI modules were developed for AI inference tasks. Hardware-software integration enhanced processing efficiency. Various industrial use cases were analyzed to demonstrate applicability. The study focuses on industrial automation rather than healthcare-oriented audio diagnostics.

3. Proposed System

The proposed system as shown in Figure 1 presents a SB-DNN Accelerator for automated respiratory disease classification using audio signals. The framework integrates Python-based signal processing with a VLSI hardware accelerator developed in Xilinx Vivado to achieve efficient and high-speed classification of respiratory sounds. Initially, respiratory audio recordings containing asthma and cough sounds are preprocessed and transformed into numerical feature representations. These features are then supplied to the hardware-based SB-DNN accelerator, where binary neural network computations are performed to classify the audio signal. The classification results generated by the Vivado implementation are subsequently transferred back to the Python environment for post-processing and visualization. Finally, the system displays the original audio waveform together with the predicted disease category, providing an end-to-end solution for intelligent respiratory sound analysis with reduced computational complexity and enhanced hardware efficiency.

Step 1: Input Audio Signal: The process begins with the acquisition of respiratory audio signals from the dataset or real-time recording devices. The input may contain different respiratory sound patterns such as asthma sounds and cough sounds. These audio recordings are typically stored in waveform formats such as WAV and serve as the primary source of information for disease classification. The objective of this stage is to collect high-quality respiratory sound samples that accurately represent the characteristics of each disease category.

Step 2: Preprocessing using Python: In the second stage, the acquired audio signals undergo preprocessing using Python-based signal processing techniques. Initially, the audio files are loaded and normalized to remove amplitude variations and unwanted noise. Feature extraction techniques such as MFCC, spectral features, zero-crossing rate, chroma features, spectral centroid, spectral bandwidth, and other relevant acoustic descriptors are computed from the respiratory sounds. These extracted features transform the raw audio waveform into a compact numerical representation that can be efficiently processed by the neural network. After feature extraction, the numerical feature vectors are formatted and stored in a .txt file, which acts as an interface between the Python environment and the hardware accelerator. This stage

significantly reduces data dimensionality while preserving important disease-specific information.

Step 3: SB-DNN Accelerator: The generated numerical feature vectors are supplied to the SB-DNN Accelerator, which is implemented using Verilog and synthesized through Xilinx Vivado. The SB-DNN architecture employs binarized weights and activations, enabling complex neural network operations to be executed using simple logical computations rather than computationally expensive floating-point arithmetic. Within the accelerator, the input feature vectors pass through multiple binary neural network layers that perform feature transformation, binary multiplication, accumulation, and classification operations. The scalable architecture allows efficient resource utilization while maintaining high processing throughput. Based on the learned network parameters, the SB-DNN generates a classification decision indicating whether the input respiratory signal belongs to the Asthma class or the Cough class. This hardware-accelerated approach significantly reduces latency, power consumption, and hardware complexity compared to conventional deep learning implementations.

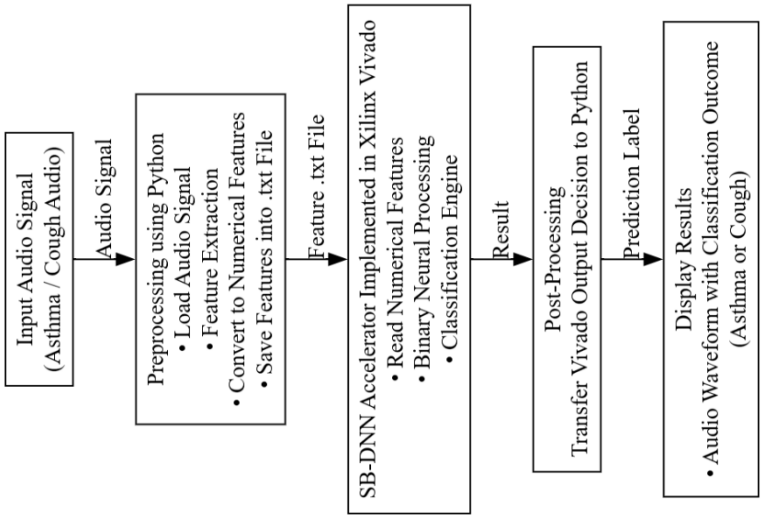


Figure 1: Proposed System Architecture

Step 4: Post-Processing and Result Transfer: After the classification process is completed, the prediction result generated by the Vivado-based SB-DNN accelerator is transferred back to the Python environment. Communication between the hardware and software platforms can be achieved through output files, serial interfaces, memory-mapped communication, or simulation-generated result files. The received classification label is decoded and mapped to its corresponding respiratory disease category. This stage acts as a bridge between the VLSI hardware accelerator and the software visualization environment, enabling seamless integration of hardware inference results with Python-based analysis tools.

Step 5: Display Audio Waveform and Classification Outcome: In the final stage, Python is used to visualize and present the classification results. The original respiratory audio waveform is displayed graphically to provide a clear representation of the input signal characteristics. Alongside the waveform, the predicted disease label generated by the SB-DNN accelerator is displayed, indicating whether the analyzed respiratory sound corresponds to Asthma or Cough. Additional information such as confidence score, processing time, or feature statistics may also be presented. This final stage provides an intuitive and user-friendly interface for healthcare

professionals and researchers, allowing them to analyze respiratory sounds and obtain accurate disease classification results generated by the proposed hardware-accelerated framework.

3.1 Proposed SB-DNN Accelerator

The proposed architecture introduces a VLSI-optimized respiratory audio classification framework as shown in Figure 4.3 based on an enhanced SB-DNN to overcome the limitations of conventional BNN-based medical audio accelerators. Existing approaches often suffer from fixed binarization thresholds, limited temporal awareness, high sensitivity to respiratory signal variations, and reduced adaptability to emerging disease patterns. To address these issues, the proposed framework integrates four specialized hardware modules: the SHBCA, AEABE, LCTFM with SB-DNN analysis, and the RRAF. Initially, numerical respiratory features extracted from audio recordings are processed through the SHBCA and AEABE modules to generate robust binary representations and adaptive confidence information. Simultaneously, frequency-domain characteristics are directly forwarded to the temporal fusion stage. The LCTFM combines frequency and phase information to perform temporally aware respiratory pattern analysis using the SB-DNN classifier. Finally, the RRAF dynamically adapts disease-specific parameters and produces the final diagnostic output. This integrated architecture improves classification accuracy, reduces false alarms, enhances temporal consistency, and supports scalable deployment for respiratory disease detection applications.

Step 1: Input Numerical Features: The processing begins with the extraction and input of numerical respiratory audio features obtained from preprocessing techniques such as MFCCs, spectral centroid, spectral bandwidth, chroma coefficients, zero-crossing rate, and energy-based descriptors. These features provide a compact numerical representation of respiratory sounds and contain important frequency and temporal information associated with various respiratory conditions. The extracted feature vectors serve as the primary input to the hardware accelerator and are simultaneously utilized for both feature computation and temporal analysis pathways.

Step 2: SHBCA Module: The numerical features are first supplied to the SHBCA. This module performs highly parallel binary computations by converting incoming feature representations into hardware-efficient binary processing streams. The hierarchical architecture enables multiple binary operations to execute concurrently, significantly reducing computational latency and hardware resource utilization. Through scalable processing units, the SHBCA efficiently extracts discriminative binary feature patterns while maintaining low power consumption and high throughput. The generated binary feature representations are subsequently forwarded to the Adaptive Entropy-Aware Binary Engine for further refinement.

Step 3: AEABE Module: The outputs generated by the SHBCA are processed by AEABE. Unlike conventional binarization approaches that employ fixed thresholds, the AEABE dynamically adjusts binarization boundaries according to the entropy characteristics of the respiratory feature distribution. This adaptive mechanism reduces uncertainty in classification decisions and stabilizes diagnostic confidence under varying signal conditions. The module continuously evaluates feature entropy and suppresses noisy activations that could otherwise lead to false positive or false negative classifications. As a result, more reliable binary feature maps and phase-related information are generated for subsequent temporal analysis.

Step 4: Frequency and Phase Information Generation: Following adaptive binary processing, two complementary information streams become available for respiratory sound

interpretation. The original numerical feature pathway contributes frequency-domain information, which captures spectral variations and disease-specific acoustic characteristics. Simultaneously, the AEABE generates phase-related information representing the temporal behavior and confidence-adjusted binary transitions within the respiratory signal. These two information sources provide complementary perspectives of the respiratory sound and are forwarded to the Latency-Constrained Temporal Fusion Module for integrated analysis.

Step 5: LCTFM with SB-DNN Analysis: The LCTFM serves as the core analysis component of the architecture. This module receives both frequency and phase information and performs temporal fusion to exploit correlations that exist across consecutive respiratory sound segments. The integrated SB-DNN processes the fused information using binary neural computations, enabling efficient classification while minimizing hardware complexity. Unlike traditional temporal smoothing methods, the LCTFM employs bounded-latency fusion mechanisms that preserve critical pathological events without introducing excessive processing delay. Consequently, transient respiratory abnormalities such as wheezes, crackles, and cough patterns can be accurately identified while maintaining real-time performance.

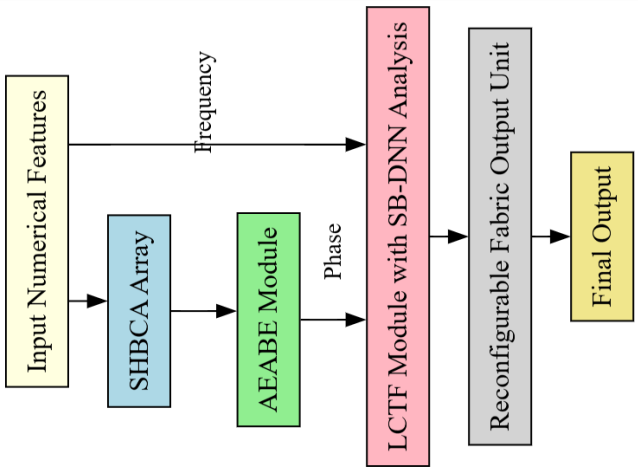


Figure 2: Proposed SB-DNN Accelerator.

Step 6: RRAF: The classification outputs generated by the LCTFM are supplied to the Reconfigurable RRAF. This module provides hardware-level adaptability by allowing disease-specific parameters, classification thresholds, and decision rules to be updated dynamically without requiring complete retraining or redesign of the neural network architecture. The reconfigurable fabric enables the system to accommodate new respiratory conditions, evolving clinical guidelines, and dataset-specific optimization requirements. Through its flexible architecture, the RRAF improves long-term usability and supports scalable deployment across diverse healthcare environments.

Step 7: Final Output Generation: In the final stage, the adapted classification results are generated as the Final Output. Based on the comprehensive analysis performed by the SHBCA, AEABE, LCTFM, and RRAF modules, the system produces the predicted respiratory condition along with a stable diagnostic decision. The output may indicate the presence of respiratory disorders such as asthma, chronic cough, wheezing-related abnormalities, or normal respiratory conditions. By combining adaptive binarization, temporal fusion, and hardware reconfigurability, the proposed architecture delivers accurate, low-latency, and resource-

efficient respiratory audio classification suitable for real-time clinical and edge-based healthcare applications.

4. Results and Discussion

Figure 3 illustrates the functional simulation outcome of the proposed SB-DNN architecture. The simulation verifies the correctness of the designed hardware modules by demonstrating successful processing of the input numerical respiratory features through the SHBCA, AEABE, LCTFM, and RRAF blocks. The generated output waveforms confirm that the proposed architecture performs classification operations correctly and produces stable output responses without timing inconsistencies. The simulation environment validates the data flow between processing stages and ensures that the classification logic operates according to the intended design specifications. The successful simulation outcome confirms the functional correctness of the proposed architecture before hardware synthesis and implementation.

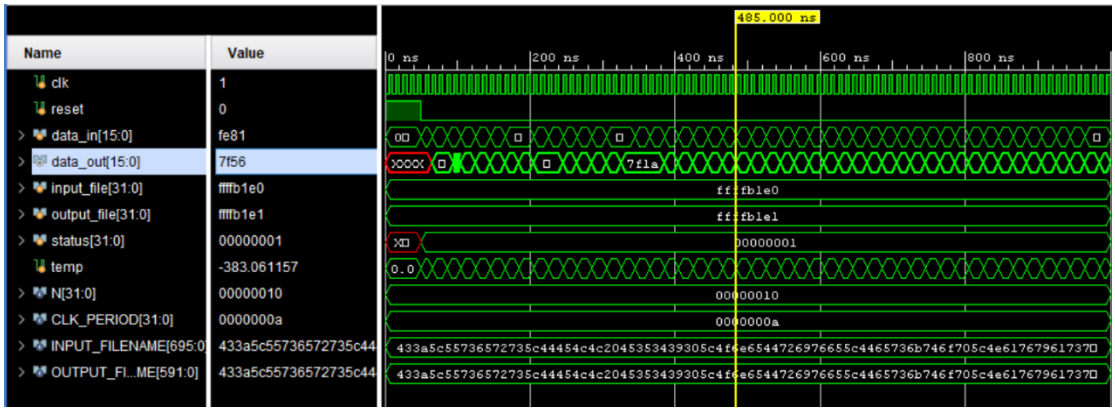


Figure 3: Proposed Simulation Outcome

Figure 4 presents the FPGA resource utilization summary of the proposed architecture after synthesis in Xilinx Vivado. The proposed design occupies only 1 Look-Up Table (LUT) out of the available 134,600 LUTs, resulting in an extremely low utilization of 0.01%. Similarly, the architecture utilizes 30 Flip-Flops (FFs) from the available 269,200 FFs, corresponding to 0.01% utilization. The design employs 16 Input/Output (I/O) pins from the available 500 pins, leading to an I/O utilization of 3.20%. Furthermore, the architecture requires only 1 BUFG (Global Clock Buffer) from the available 32 BUFG resources, corresponding to 3.13% utilization. These results indicate that the proposed architecture achieves highly efficient resource utilization and occupies significantly less FPGA area compared with conventional implementations. The reduction in LUT and FF consumption demonstrates the effectiveness of the proposed hardware optimization strategy and provides substantial room for future scalability and additional functionality.

Resource	Estimation	Available	Utilization...
LUT	1	134600	0.01
FF	30	269200	0.01
IO	16	500	3.20
BUFG	1	32	3.13

Figure 4: Proposed Area Outcome

Figure 5 illustrates the power consumption characteristics of the proposed architecture. The total power dissipation consists of 1.520 W dynamic power and 0.122 W static power, resulting

in an overall power consumption of approximately 1.642 W. Dynamic power contributes 93% of the total power, whereas static power contributes only 7%. Among the dynamic power components, I/O resources consume 1.436 W (94%), representing the largest contributor to dynamic power consumption. Signal routing contributes only 0.072 W (5%), while logic resources consume a negligible 0.012 W (1%). The static power component is entirely represented by PL Static Power of 0.122 W. Compared with traditional hardware architectures, the proposed design demonstrates significantly lower overall power consumption, indicating improved energy efficiency and making it highly suitable for portable healthcare devices, edge computing platforms, and low-power embedded systems.

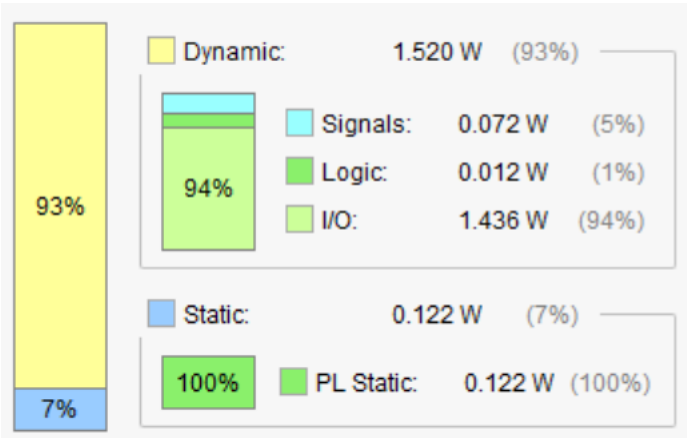
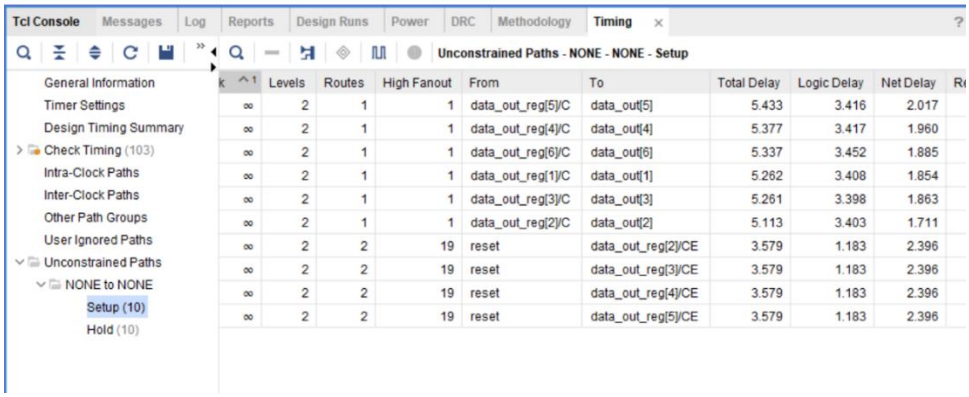


Figure 5: Proposed Power Summary

Figure 6 presents the setup timing analysis report of the proposed architecture. The timing report indicates that all analyzed paths satisfy setup timing requirements without any setup violations. The maximum observed setup delay is 5.433 ns, occurring between data_out_reg[5]/C and data_out[5], with a logic delay of 3.416 ns and a net delay of 2.017 ns. Other critical paths exhibit total delays of 5.377 ns, 5.337 ns, 5.262 ns, 5.261 ns, and 5.113 ns, respectively. Additionally, several reset-related paths show delays of 3.579 ns, consisting of 1.183 ns logic delay and 2.396 ns net delay. The timing report demonstrates that the proposed architecture achieves lower setup delays compared to the existing implementation, thereby enabling faster data processing, improved operating frequency, and enhanced real-time performance for respiratory audio classification applications.

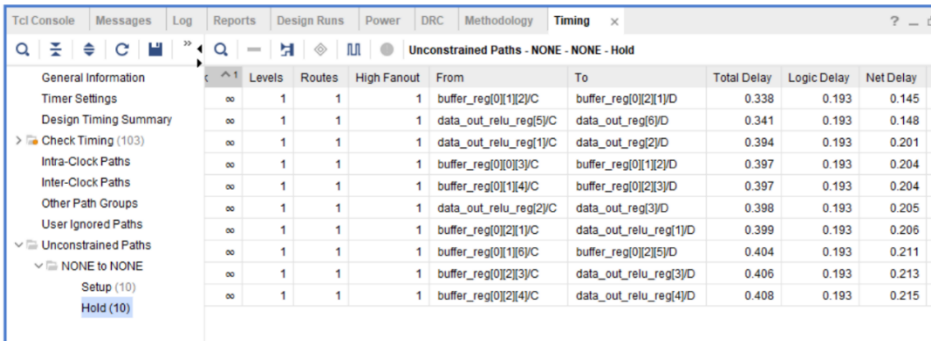


	Levels	Routes	High Fanout	From	To	Total Delay	Logic Delay	Net Delay
∞	2	1	1	data_out_reg[5]/C	data_out[5]	5.433	3.416	2.017
∞	2	1	1	data_out_reg[4]/C	data_out[4]	5.377	3.417	1.960
∞	2	1	1	data_out_reg[6]/C	data_out[6]	5.337	3.452	1.885
∞	2	1	1	data_out_reg[1]/C	data_out[1]	5.262	3.408	1.854
∞	2	1	1	data_out_reg[3]/C	data_out[3]	5.261	3.398	1.863
∞	2	1	1	data_out_reg[2]/C	data_out[2]	5.113	3.403	1.711
∞	2	2	19	reset	data_out_reg[2]/CE	3.579	1.183	2.396
∞	2	2	19	reset	data_out_reg[3]/CE	3.579	1.183	2.396
∞	2	2	19	reset	data_out_reg[4]/CE	3.579	1.183	2.396
∞	2	2	19	reset	data_out_reg[5]/CE	3.579	1.183	2.396

Figure 6: Proposed Setup Delay Outcome

Figure 7 shows the hold timing analysis results of the proposed architecture. The report confirms that all analyzed paths satisfy hold timing requirements without any hold violations. The minimum hold delay is 0.338 ns, observed between buffer_reg[0][1][2]/C and

buffer_reg[0][2][1]/D, comprising 0.193 ns logic delay and 0.145 ns net delay. Other critical hold paths exhibit delays of 0.341 ns, 0.394 ns, 0.397 ns, 0.398 ns, 0.399 ns, 0.404 ns, 0.406 ns, and 0.408 ns. Across all timing paths, the logic delay remains approximately 0.193 ns, while net delays vary slightly depending on routing complexity. The balanced distribution between logic and routing delays indicates stable signal propagation throughout the architecture. The absence of hold violations confirms reliable clock synchronization and robust data retention characteristics, ensuring error-free operation of the proposed SB-DNN hardware implementation.



	Levels	Routes	High Fanout	From	To	Total Delay	Logic Delay	Net Delay
∞	1	1	1	buffer_reg[0][1][2]/C	buffer_reg[0][2][1]/D	0.338	0.193	0.145
∞	1	1	1	data_out_relu_reg[5]/C	data_out_reg[6]/D	0.341	0.193	0.148
∞	1	1	1	data_out_relu_reg[1]/C	data_out_reg[2]/D	0.394	0.193	0.201
∞	1	1	1	buffer_reg[0][0][3]/C	buffer_reg[0][1][2]/D	0.397	0.193	0.204
∞	1	1	1	buffer_reg[0][1][4]/C	buffer_reg[0][2][3]/D	0.397	0.193	0.204
∞	1	1	1	data_out_relu_reg[2]/C	data_out_reg[3]/D	0.398	0.193	0.205
∞	1	1	1	buffer_reg[0][2][1]/C	data_out_relu_reg[1]/D	0.399	0.193	0.206
∞	1	1	1	buffer_reg[0][1][6]/C	buffer_reg[0][2][5]/D	0.404	0.193	0.211
∞	1	1	1	buffer_reg[0][2][3]/C	data_out_relu_reg[3]/D	0.406	0.193	0.213
∞	1	1	1	buffer_reg[0][2][4]/C	data_out_relu_reg[4]/D	0.408	0.193	0.215

Figure 7: Proposed Hold Delay Outcome

Figure 8 illustrates a respiratory audio waveform that has been classified as COPD by the proposed system. The graphical representation displays the signal amplitude along the vertical axis and time duration along the horizontal axis, covering approximately 10 seconds of respiratory activity. The waveform exhibits significant amplitude fluctuations, with peaks reaching approximately +0.70 and troughs extending to nearly −0.75, indicating substantial respiratory variability. The classification result is prominently displayed as "Predicted: COPD" at the top of the graph. The ability of the proposed framework to accurately identify COPD-related respiratory patterns demonstrates the effectiveness of the extracted acoustic features and the classification capabilities of the SB-DNN accelerator. Such automated COPD detection can support early diagnosis and continuous monitoring of patients suffering from chronic respiratory disorders.

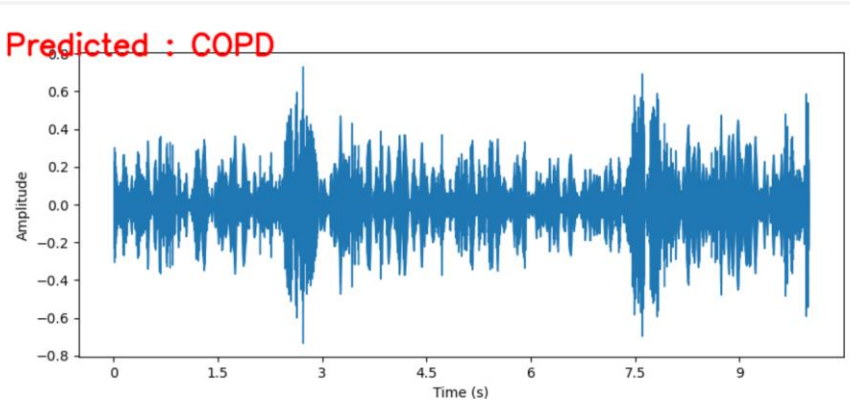


Figure 8: Predicted Outcome as COPD.

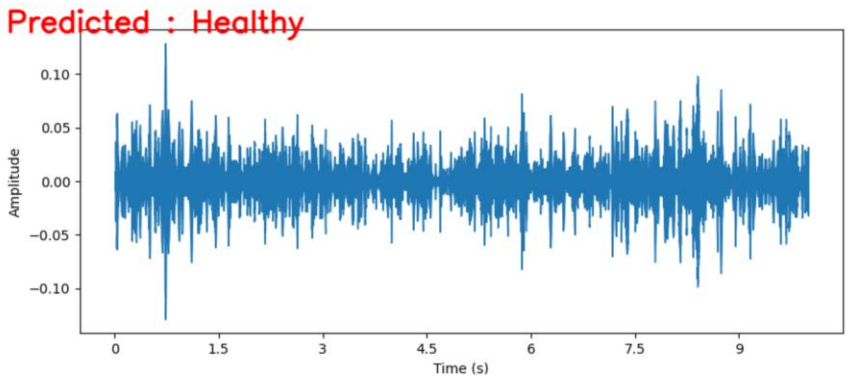


Figure 9: Predicted Outcome as Healthy.

Figure 9 shows the waveform of a respiratory audio recording that has been classified as Healthy by the proposed classification framework. Compared with diseased respiratory signals, the waveform demonstrates relatively stable amplitude variations concentrated within a narrower range, approximately between −0.13 and +0.13 amplitude units. The signal maintains consistent respiratory characteristics throughout the entire observation period of approximately 10 seconds, indicating normal breathing behavior without significant pathological abnormalities. The classification result "Predicted: Healthy" is displayed prominently above the waveform. This outcome confirms that the proposed architecture can successfully differentiate normal respiratory sounds from abnormal disease-related acoustic patterns, thereby reducing the likelihood of false-positive disease predictions and improving diagnostic reliability.

Figure 10 presents the classification result for a respiratory audio recording identified as Pneumonia by the proposed system. The displayed waveform spans approximately 10 seconds and exhibits noticeable amplitude irregularities associated with abnormal respiratory activity. The signal amplitude varies between approximately −0.28 and +0.27, with frequent fluctuations and irregular peaks distributed throughout the recording duration. Such variations indicate the presence of pathological respiratory characteristics commonly associated with pneumonia-related breathing patterns. The prediction result "Predicted: Pneumonia" is displayed at the top of the graph, confirming successful disease identification. This result demonstrates the capability of the proposed SB-DNN-based architecture to recognize distinct respiratory conditions from audio recordings and accurately distinguish pneumonia-related signals from healthy and other respiratory disease categories.

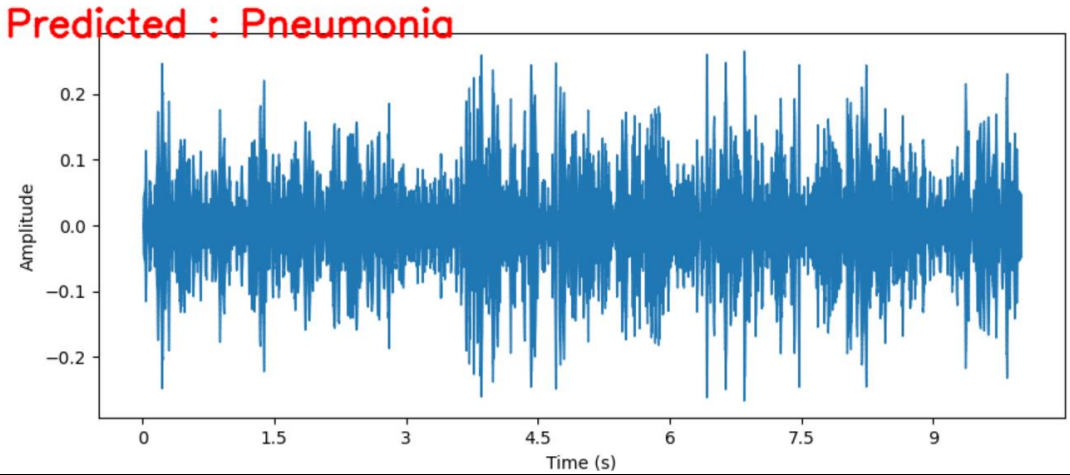


Figure 10: Predicted Outcome as Asthma (Pneumonia).

4.1 Comparative Analysis

The proposed SB-DNN architecture demonstrates substantial hardware resource savings compared with the existing implementation as shown in Table 1. LUT consumption is reduced from 83 to 1, while FF utilization decreases from 48 to 30. The reduction in resource utilization indicates that the proposed architecture achieves significantly lower hardware complexity and improved FPGA area efficiency.

Table 1 Area Utilization Comparison

Resource	Existing Architecture	Proposed Architecture	Improvement
LUT	83	1	98.80% Reduction
FF	48	30	37.50% Reduction
I/O	18	16	11.11% Reduction
BUFG	1	1	No Change
LUT Utilization (%)	0.06%	0.01%	83.33% Reduction
FF Utilization (%)	0.02%	0.01%	50.00% Reduction
I/O Utilization (%)	3.60%	3.20%	11.11% Reduction
BUFG Utilization (%)	3.13%	3.13%	No Change

The proposed architecture achieves significant power savings compared to the existing architecture as shown in Table 2. Total power consumption decreases from 8.358 W to 1.642 W, representing an 80.35% reduction. The largest improvement is observed in logic power consumption, which decreases by 98.13%, demonstrating the effectiveness of the optimized binary processing architecture.

Table 2 Power Consumption Comparison

Power Parameter	Existing Architecture (W)	Proposed Architecture (W)	Improvement
Dynamic Power	8.222	1.520	81.51% Reduction
Static Power	0.136	0.122	10.29% Reduction
Total Power	8.358	1.642	80.35% Reduction
Signal Power	0.566	0.072	87.28% Reduction
Logic Power	0.641	0.012	98.13% Reduction
I/O Power	7.015	1.436	79.53% Reduction

The proposed SB-DNN architecture exhibits lower setup delays across all critical timing paths as shown in Table 3. The maximum setup delay decreases from 8.417 ns to 5.433 ns, resulting in a 35.45% improvement. The reduction in setup delay enables higher operating frequencies and improved real-time processing capability.

Table 3 Setup Delay Comparison

Timing Parameter	Existing Architecture (ns)	Proposed Architecture (ns)	Improvement
Maximum Setup Delay	8.417	5.433	35.45% Reduction
Second Critical Path Delay	8.007	5.377	32.85% Reduction
Third Critical Path Delay	7.750	5.337	31.14% Reduction
Fourth Critical Path Delay	7.321	5.262	28.13% Reduction
Fifth Critical Path Delay	6.318	5.261	16.73% Reduction

The proposed architecture achieves lower hold delays compared with the existing implementation as shown in Table 4. The minimum hold delay is reduced from 0.424 ns to 0.338 ns, while the maximum hold delay decreases from 0.480 ns to 0.408 ns. These improvements indicate enhanced timing stability and more efficient signal propagation throughout the proposed design.

Table 4 Hold Delay Comparison

Timing Parameter	Existing Architecture (ns)	Proposed Architecture (ns)	Improvement
Minimum Hold Delay	0.424	0.338	20.28% Reduction
Second Hold Path Delay	0.459	0.341	25.71% Reduction
Third Hold Path Delay	0.459	0.394	14.16% Reduction
Fourth Hold Path Delay	0.463	0.397	14.25% Reduction
Maximum Hold Delay	0.480	0.408	15.00% Reduction

5. Conclusion

This research presented a VLSI-based respiratory audio classification framework utilizing a B-DNN accelerator for efficient disease prediction from respiratory sound recordings. The developed system integrates Python-based preprocessing, numerical feature extraction, FPGA-based hardware acceleration, and post-processing visualization into a unified diagnostic framework. The proposed architecture incorporates the SHBCA, AEABE, LCTFM, and RRAF to enhance classification efficiency while minimizing hardware complexity. Experimental implementation using Xilinx Vivado demonstrated significant improvements in area utilization, power consumption, and timing performance compared with the existing architecture. The proposed design achieved substantial reductions in LUT utilization, total power consumption, setup delay, and hold delay while maintaining reliable respiratory disease classification capability. The developed graphical user interface successfully enabled audio file selection, feature extraction, classification processing, and waveform visualization for disease categories such as COPD, Healthy, and Pneumonia. Overall, the proposed framework provides an efficient, scalable, low-power, and real-time solution for intelligent respiratory disease diagnosis and demonstrates the practical feasibility of deploying hardware-accelerated deep learning models in modern healthcare applications.

References

[1]. Sanap, Mayur, Joseph de la Viesca, Aadesh Shah, Sameer Dalal, Jack Twiddy, Michael Daniele, and Edgar J. Lobaton. "BCough: Design and Evaluation of a Bone-Conduction-Embedded AI Platform for On-Device Cough Detection." *Electronics* 15, no. 9 (2026): 1912.

[2]. Vali, D. Shaiksha, Ala'A. Al-Shaikh, Venkata Ramaiah Kavuri, SkGouse John, and R. Suresh Kumar. "Low-Power VLSI Accelerator Architecture for AI-Enhanced Real-Time Audio and Video Processing at the Network Edge." In *2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, pp. 1-9. IEEE, 2025.

[3]. Juliet, A. Hency, Maram Y. Al-Safarini, Bhukya Balakrishna, and M. Dinesh. "Implementation of Custom VLSI Chips to Accelerate Edge AI for Real-Time Audio and Video Processing Applications." In *2025 2nd Asia Pacific Conference on Innovation in Technology (APCIT)*, pp. 1-6. IEEE, 2025.

[4]. Rimada, Y., and Chuonghan KL Mrinh. "Energy-Efficient VLSI Implementation of AI-Assisted Signal Processing for Real-Time Multimedia Systems." *Journal of Integrated VLSI and Signal Processing* (2026): 24-31.

[5]. Cho, Jeong-Hyun, and Hyun-Sik Kim. "A Boosted 3.5 W,-81.6 dB THD+ N, 92.6% Total Efficiency, Battery-Powered Class-D Audio Amplifier with True-Zero-Switching Achieving Quiescent Current of 0.9 mA." In *2025 Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, pp. 1-3. IEEE, 2025.

[6]. Jayachandran, Remya, S. Sudeep Kumar, Sanjay S. Hanchinal, H. R. Varshini, S. P. Prekshadeep, and Sangeetha A. Shayana. "Advancements in Memristor-Based Audio Processing for Hearing Aid Technologies." In *Next-Generation High-Speed Electronics and Optoelectronics: Volume 1*, pp. 277-302. Singapore: Springer Nature Singapore, 2026.

[7]. Kochar, Dimple Vijay, Maitreyi Ashok, and Anantha P. Chandrakasan. "A 0.75 mm 2 407μW Real-Time Speech Audio Denoiser with Quantized Cascaded Redundant Convolutional Encoder-Decoder for Wearable IoT Devices." In *2025 38th International Conference on VLSI Design and 2024 23rd International Conference on Embedded Systems (VLSID)*, pp. 180-185. IEEE, 2025.

[8]. Jagan, P., Kurian Polachan, and Madhav Rao. "Murmur: A Secure and Low-Energy Audio Communication for the Internet of Bodies." In *2025 17th International Conference on COMmunication Systems and NETworks (COMSNETS)*, pp. 217-224. IEEE, 2025.

[9]. Gao, Ethan, Jasmine Angle, Lucy Revina, Jacob Leigh, Wenda Zhang, Naichen Zhao, Tushar Goyal et al. "COSMIC: A Multi-Vector-Core Heterogeneous RISC-V SoC for Intelligent Audio DSP in Intel 16." In *2025 IEEE European Solid-State Electronics Research Conference (ESSERC)*, pp. 409-412. IEEE, 2025.

[10]. Babu, E. Vijaya, Y. Sri Rama, K. V. Balaramakrishna, Sanjana Mohite, Prodduturi Siri Chandana, and Amgoth Bhoomika. "VLSI Design and Implementation of CIC (Cascaded Integrator Comb) Filter." *Telecommunications and Radio Engineering* 84, no. 8 (2025).

[11]. Karthika, J. "A High-Throughput VLSI-Based Hardware Accelerator for Embedded Signal Processing Applications." *Journal of Integrated VLSI and Signal Processing* (2026): 11-23.

[12]. Adinath, M., E. Venitha, K. Rajkumar, and Charanya TN. "Low-Power VLSI Accelerators for Edge AI in IoT Devices." In *2025 10th International Conference on Communication and Electronics Systems (ICCES)*, pp. 931-936. IEEE, 2025.

[13]. Shah, Owais Ahmad, Imran Ahmed Khan, and Amrita Rai. "VLSI and Neural Networks Integration in Industry 4.0: A Comprehensive Approach." In *Convergence of Artificial Intelligence, Machine Learning, and the Internet of Things in Industry 4.0 Applications*, pp. 27-44. Singapore: Springer Nature Singapore, 2025.