International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Design of Area and power Efficient MAC architecture using CNN for DSP Applications

Manjula Basavant Bhajantri[1], Dr. Sharanabasaveshwar G Hiremath[2]

Submitted: 06/12/2023 **Revised:** 17/01/2024 **Accepted:** 27/01/2024

Abstract: The planned task aims to develop a highly efficient MAC Architecture utilizing CNN on a Digital Signal Processor through the implementation of Verilog HDL for functional verification, Synthesis, and Physical Design using Cadence Genus and Innovus in the ASIC design flow. This architecture aims to significantly enhance the processor's speed by executing rapid multiplication and addition operations, characteristic of the current MAC unit. With the rapid evolution of technology, digital signal processors have become increasingly potent and resourceful. The cornerstone of this MAC architecture lies in its utilization of CNN, enabling swift operations. Constructing MAC architecture necessitates the integration of various digital blocks within the design. The proposed design achieves an exceptional reduction of 80.13% in both area and power, consequently resulting in a substantial decrease in the architecture's size. Moreover, the proposed design offers the added benefits of optimized power utilization and minimized area requirements.

Keywords: *ASIC, CNN, HDL, MAC.*

1. Introduction

The hardware elements crucial for digital signal processing, such as the Multiply-Accumulate unit (MAC unit), play a vital role in enhancing the speed of digital filters by focusing on multiplication and accumulation unit. These components are integral in various applications like telecommunications, audio and image processing applications, leveraging DSP techniques to manipulate and analyze signals in a digital format. It holds significant importance in executing a range of mathematical tasks, especially those entailing matrix multiplications, filtering, convolution, and similar operations that necessitate combining multiplication and addition. Termed the MAC unit, it typically comprises two dedicated registers depicted in Figure(1), accepting 16-bit data (operands) as input. These inputs undergo with multiplication in the multiplier, with the resultant product accumulated in a 32-bit accumulator. As the accumulator cannot retain data for extended periods, the register normally stores double precision data from the accumulator.

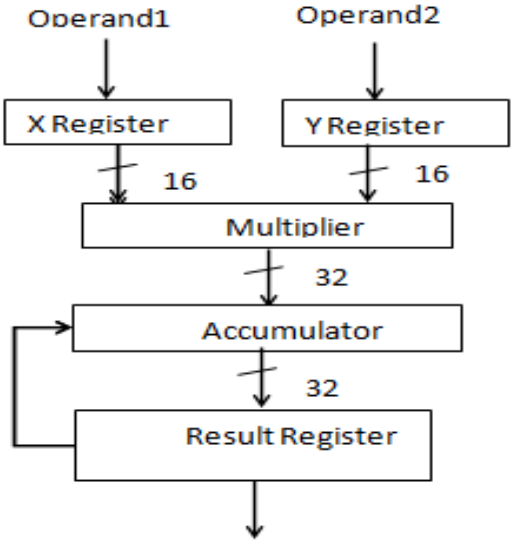

Fig 1: MAC Architecture

In Figure 2, the intricate layers of computation within a CNN are mapped. The convolutional layer specifically hosts the majority of the convolutional computations and operations. For a typical image, numerous convolutional operations are necessary even for a single still image, totaling in the millions. Should the application demand real-time usage of CNNs, such as in video processing, the volume of convolutions required increases significantly.

[1]*Research Scholar,Visvesvaraya Technological University, Belagavi, Karnataka ,Department of Electronics and Communication Engineering, East West Institute of Technology, Bangalore.*
[2]*Professor, Visvesvaraya Technological University, Belagavi, Karnataka.Department of Electronics and Communication Engineering, East West Institute of Technology, Bangalore.*

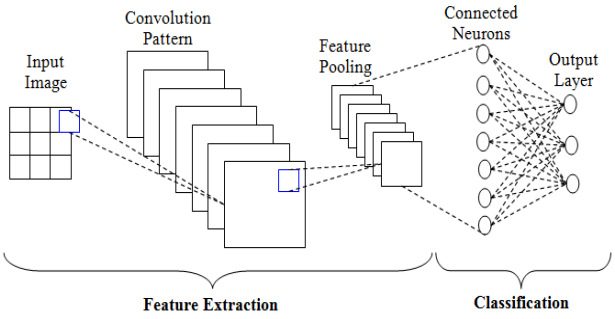

Fig 2: Various layers are involved in a CNN

Tensorflow is the common choice for implementing Convolutional Neural Networks, serving as a robust machine learning tool tailored for image processing. Its extensive set of built-in libraries and functions facilitates the comprehensive software high level implementation of CNN layers through Python scripting language. The approach finds its real application in handling static images. Capitalizing on the benefits of CNNs, their widespread utilization has surged in GPUs (graphic processing units), leveraging their capacity for extensive parallel processing. Notably, recent years have witnessed notable advancements in Convolutional Neural Networks (CNNs) research, driven by the escalating demand for real-time applications in image and video processing [5]. Consequently, there's an increasing need for swift deployment of CNNs to enable accelerated processing in real-time applications. Nonetheless, owing to the computational difficulty inherent in the convolutional layer, CNNs frequently encounter heightened latency, presenting a problem in achieving seamless real-time processing.

The catalyst behind the significant advancements in machine learning stems from the inception of Artificial Neural Networks (ANNs). Within this domain, the aggregation of deep layers within ANNs culminates in the creation of Convolutional Neural Networks (CNNs), which emerges as an exceptional architectural innovation. CNNs find extensive application in complex image recognition tasks, owing to their commendable accuracy and straightforward design, rendering them an ideal entry point for newcomers venturing into the realm of ANNs. At their core, Convolutional Neural Networks (CNNs) heavily rely on the Multiply-Accumulate (MAC) unit. A technique like Truncated Cross Parallel MAC (TCP-MAC) Technique proves effective in mitigating the computational complexity associated with CNNs. Training CNNs demands extensive data and necessitates vast parallel computations involving numerous multiply and accumulate (MAC) units. CNNs exhibit variable channel widths and kernel sizes, catering to the diverse nature of application files. However, the prevalent hardware platforms typically utilize standard optimization methods, leading to inefficient utilization of computational resources. To tackle the challenge posed by current Multiply-Accumulate (MAC) architectures—often characterized by excessive power consumption or extensive space requirements—implementing efficient techniques like Wallace Reduction Trees and Modified Booth Encoders proves instrumental in reducing the number of bits utilized in the MAC operation. This reduction in bit usage results in reduced delay and improved speed accurately. Within digital circuits, the multiplier stands out as the most power-intensive element, handling a majority of arithmetic operations primarily through shift and adds operations. Enhancements in multiplier performance often involve optimizing the adder component, given the multiplier's integral role across numerous modern CPUs. The speed of the multiplier significantly impacts the overall speed of the MAC unit, emphasizing the critical need for a high-speed multiplier. Among various multiplier designs, the Booth Modification technique excels in minimizing incomplete partial products, offering the lowest delay. Simultaneously, the Wallace Tree multiplier architecture significantly achieves high computation by extensively parallelizing partial product additions.

2. Literature Survey

VLSI technology plays a pivotal role in executing digital signal processing applications. DSP chips, reliant on VLSI technology, find extensive applications across audio processing, image processing, telecommunications, and diverse signal processing high level domains. Meanwhile, Deep Learning Neural Networks represent a category of machine learning models, drawing inspiration from the intricate structure and functionalities observed within the human brain. These networks are engineered to autonomously grasp and delineate intricate patterns and associations within data, empowering them to execute tasks such as image recognition, natural language processing, and beyond [12]. Embedding deep learning into Very Large Scale Integration (VLSI) encounters several challenges, stemming from the distinctive attributes of both realms. Addressing power and energy efficiency stands out as a critical hurdle in implementing deep learning within VLSI. Crafting hardware capable of adeptly managing the computational requirements of deep learning while curbing power consumption represents a significant challenge [10]. The natural parallelism ingrained within deep learning accurately models presents a challenge when harnessing it within VLSI designs, constrained by physical layout and interconnect limitations [8]. The recurrent use of multiply and Accumulate operations (MAC) within CNN's convolutional layers signifies an exceptionally high level of computational complexity. Consequently, this elevates the computational workload and energy consumption across various applications [4]. To optimize Power Efficiency is achievable through the utilization of Custom-designed ASICs, capable of attaining superior energy efficiency by eliminating

redundant components and reducing power consumption. This holds particular significance for energy-constrained modern devices and data centers. Tailoring ASICs to minimize area utilization by incorporating only essential components is pivotal for consolidating intricate deep learning models onto a solitary chip. Leveraging ASICs' optimized design for particular computations enables them to achieve remarkable throughput, a valuable trait for applications necessitating real-time or high-speed processing.

The author delves into the efficient development of the MAC unit utilizing deep learning techniques. This implementation employs shifter/adders for the MAC Unit within a Convolutional Neural Network (CNN), creating a modified MAC unit realized in 90nm technology through the Cadence tool. The primary focus lies in amplifying the computational speed of the MAC through a pipelined high level technique while concurrently reducing power consumption. The subsequent sections of the paper are structured as follows: Section 2 encompasses discussions on concept and related works, Section 3 delves into Methodology and implementation, Section 4 presents Experimental results, and finally, Section 5 encapsulates the conclusion and scopes for future exploration.

3. Methodology:

CNNs, a subset of deep learning architectures, are extensively employed for the analysis of image and spatial data. Prior to inputting data into a CNN, ensuring proper preprocessing is crucial.

The design of a CNN's architecture is tailored to capitalize on the hierarchical nature of data, particularly evident in images. Fundamental components of this architecture encompass:

- Convolutional layers, comprising numerous adaptable filters (kernels) that traverse the input data, extracting features via convolutions.

- Following each convolution operation, an activation function (typically ReLU - Rectified Linear Unit) is applied element-wise, introducing non-linearity crucial for the network to discern intricate relationships within the data.

- Pooling layers, responsible for diminishing spatial dimensions while preserving pivotal features [10].

As the network approaches its conclusion, one or multiple fully connected (dense) layers handle the high-level features extracted by earlier layers. These dense layers play a pivotal role in generating final predictions based on the acquired representations. CNNs, a subset of deep learning architectures, are purpose-built for analyzing image and spatial data. Their methodology encompasses several stages: data preprocessing, architecture design, forward propagation, loss computation, back propagation, training, evaluation, and potential fine-tuning. Distinguished for their prowess, CNNs have showcased excellent performance across a broad spectrum of computer vision tasks, spanning image classification, object detection, and image segmentation [8].

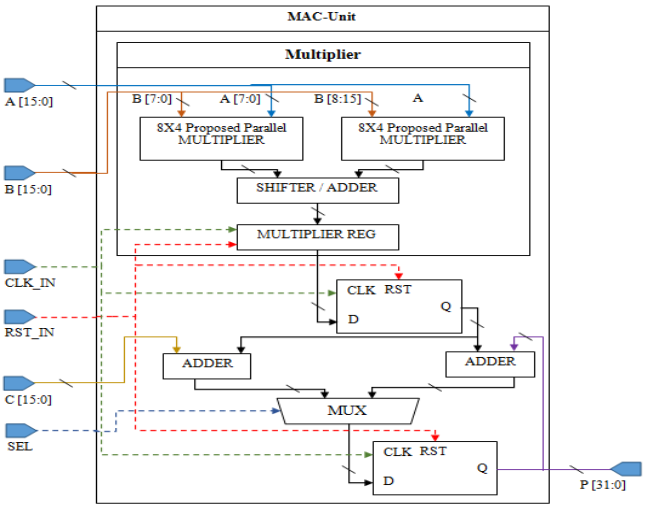

Fig 3: Design flow

4. Results and Discussions

4.1 Synthesis of Designed MAC unit using CNN on Cadence for Area and Power Analysis

Synthesis constitutes the transformation of a hardware design description into a concise gate-level representation. This process relies on a standard cell library composed of essential units like logic gates (e.g., OR, AND, NOR, NAND) and macro cells (including inverters, buffers, logic, and physical cells), which are assembled to form a technology library. The distinguishing factor of the technology library is typically denoted by the transistor size, such as 90nm. The primary objective of the logic synthesizer is to efficiently minimize hardware requirements. The proposed design's logical synthesis involves utilizing inputs such as Verilog HDL files, SDC, and constraints, with the synthesized results visually depicted in the accompanying figures.

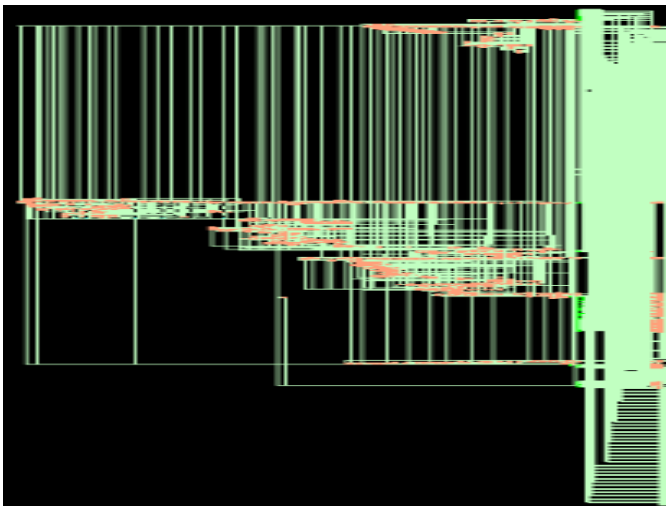

Fig 4: Synthesized View of the MAC unit

```
legacy_genus:/> report_area
============================================================
  Generated by:           Genus(TM) Synthesis Solution 16.21-s018_1
  Generated on:           Dec 09 2023  07:39:23 pm
  Module:                 CNN_RTL
  Technology library:     typical
  Operating conditions:   typical (balanced_tree)
  Wireload mode:          enclosed
  Area mode:              timing library
============================================================

Instance Module  Cells  Cell Area  Net Area   Total Area  Wireload
------------------------------------------------------------------
CNN_RTL          1242       7654         0         7654    <none> (
```

```
============================================================

                    Leakage    Dynamic     Total
  Instance    Cells Power(nW)  Power(nW)  Power(nW)
------------------------------------------------------
CNN_RTL        1242 44960.027 109508.278 154468.306
```

Fig 5: Area and Power report of the Modified Booth Encoder obtained from the Genus Synthesis tool.

Table 1: Comparison between proposed **MAC using CNN** and existing architectures.

Architecture	Area (μm^2)	Power (W)	Delay (ns)	Bit-width
MAC using CNN (Proposed)	7654 (nm)	154468.306 (nw)	6367.776 (fs)	16
BLMAC	2761.517	121.4 µ	9.11	16
Architecture using MBE	8108.755	0.38 m	13.5934	16
Architectur	4398	0.903 m	3048	8
e WRT				
Architecture using Dadda	7904	1.849 m	4213	8
MAC using CNN	29,084	251..832	11582	16

4.2 Physical Design

The physical design process encompasses transforming a circuit's description into a layout dictating the arrangement of standard library cells and the essential pathways for their interconnection. Given its complexity, physical design is usually segmented into multiple stages. Initially, circuit partitioning is required to generate macros/IPs. Subsequently, during the floor planning phase, the strategic placement of standard cells within the layout is determined accurately. Following high level placement, the subsequent step involves the implementation of global routing.

During this phase, approximate routes for interconnections among the IPs are initially outlined. This precedes the detailed routing **important** phase, which meticulously calculates all metallic connections' pathways within the channels interlinking the IPs. The final stage of the physical design process **main** centers on compacting the layout, refining the core's dimensions across all coordinates to augment the chip's die area efficiency. All physical design and synthesis tasks are executed utilizing Cadence Genus and Innovus tools.

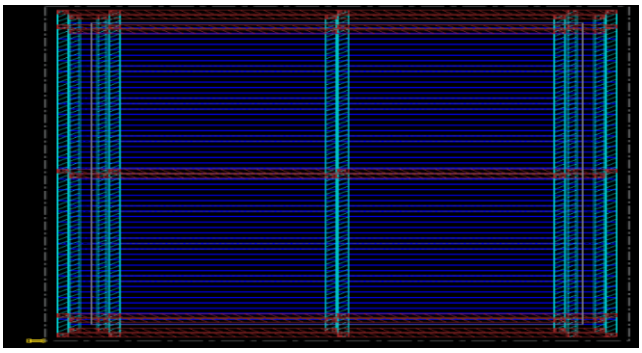

Fig 5: Floorplan View of the Core of MAC unit after PNR

Floor planning is the strategic placement of IPs or macros within the core die area, facilitating the establishment of pathways between these components directly. It calculates optimized parameters like the core's aspect ratio and utilization factor while charting out pathways for interconnecting logic cells sourced from the ASIC Library. Furthermore, it's responsible for delineating power ground (PG) connections and specifying the placement details for I/O pins/pads. A proficient floor plan is contingent upon meeting the following constraints:

- Optimize the core die area,

- Efficient obtained routing (with minimum wire length),
- Enhancing the performance by minimizing the delay in signal.

The proposed design integrates the gate level netlist and design constraints obtained from synthesis as its primary inputs. Clock Tree Synthesis (CTS) serves as a technique employed to incorporate inverters or buffers along the clock paths within the IC design, thereby guaranteeing uniform clock delay across all clock inputs. To construct the proper Floorplan, a compilation of gate-level netlist, Timing Constraints (SDC), partitioning details, .uef and .def files for power intent, IP positioning data, and RC coefficient files is utilized.

The objective of Clock Tree Synthesis (CTS) is to enhance both clock latency and clock skew. Timing constraints encompass parameters such as skew, jitter, latency, maximum capacitance, maximum transition, maximum fan-out, and the count of inverters and buffers, among others. CTS involves the dual aspects of clock tree balancing and clock tree construction. Constructing a clock tree entails utilizing clock tree inverters to ensure optimal switching, whereas clock tree balancing incorporates clock tree buffers (CTB) to meet latency and skew criteria. Complying with power and area constraints demands minimal utilization of clock tree buffers and inverters. The schematic representation following CTS is illustrated in Figure 7.

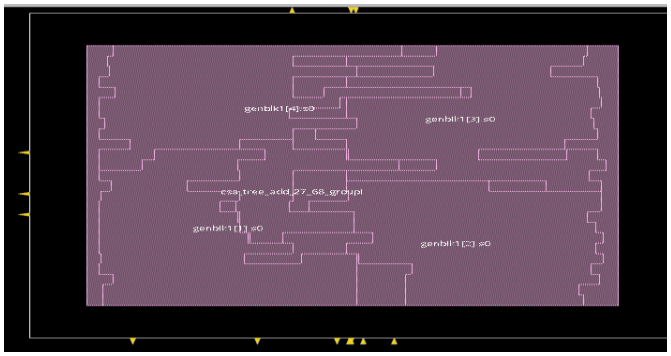

Fig 7: Amoebic View of the Core of MAC unit

Routing entails organizing wires within the allocated routing area, facilitating connections among all IPs and cells delineated in the netlist. Considerations encompass factors like wire width, intersections, and channel capacities. Routing is typically executed through two primary approaches, the first being Global Routing. This method commences with the netlist derived post-Floorplan, incorporating the positions of all fixed and movable IPs and cells, to formulate an initial layout for each net. Each net is assigned a routing area without predefining the exact layout of interconnections. The subsequent approach, Detailed Routing, focuses on delineating precise dimensions for each net within its designated routing section. This initial phase involves establishing the specific paths for the wires based on estimations derived from the global route. The comprehensive physical layout view up to the routing stage is visualized in Figure 8.

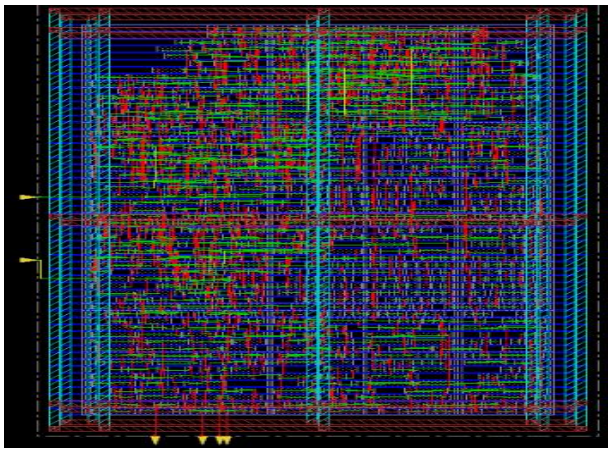

Fig 8: Physical View of the Core of MAC unit after PNR

Assessing the performance of the Proposed Design involves a range of benchmarks customized to meet the specific requirements for swift computations in DSP processors. These benchmarks act as robust benchmarks, providing standards for designing and evaluating MAC architecture utilizing CNNs within DSP applications.

Current Method is physical Design using Innovus Tool (90nm Technology)

Group	Internal Power	Switching Power	Leakage Power	Total Power	Percentage (%)
Sequential	0.04556	0.005555	0.01009	0.0612	37.73
Macro	0	0	0	0	0
IO	0	0	0	0	0
Combinational	0.06955	0.0142	0.01726	0.101	62.27
Clock (Combinational)	0	0	0	0	0
Clock (Sequential)	0	0	0	0	0
Total	0.1151	0.01975	0.02735	0.1622	100

Name of the module instances area

```
innovus 4>
innovus 4> report_area
Depth  Name              #Inst  Area (um^2)
-------------------------------------------
0      CNN_RTL           639    4014.5976
```

Table 2: Power, Area and Timing Results of MAC using CNN

Name of the Module	Power (nw)	Area (μm^2)	Timing(ns)
MAC using CNN	0.1622	4014.5976	77453

5. Conclusion:

This paper delves into the utilization of diverse Conventional and MAC designs within CNN architecture to optimize digital signal processing concerning area, delay, and power consumption. The preliminary stage involved an assessment of various MAC methodologies comprising different logical blocks, both combinational and sequential. Phase two encompassed a software execution involving synthesis using the Genus tool to scrutinize the hardware module's gate-level netlist, generating optimized reports encompassing area, power, and timing. The third phase encompassed the completion of physical design, adhering to standard procedures throughout the entire process. In advanced electronic applications, power, timing, and area stand as crucial concerns, particularly within Digital Signal Processors (DSPs) during physical design. Engaging in custom IC design proves advantageous for analyzing the power, timing, and area aspects of the design, facilitated by industry-standard tools like Cadence-Genus and Innovus. The incorporation of MAC using CNN not only yields reduced delay but also contributes to minimal computation time. Compared to other outlined MAC operations, the proposed high-level design holds the additional advantage of highly optimized power consumption, specifically tailored for DSP processors.

References

[1] Taxonomy and Benchmarking of Precision-Scalable MAC Arrays Under Enhanced DNN Dataflow Representation authors Ehab M. Ibrahim , Linyan Mei , *Graduate Student Member, IEEE,* and Marian Verhelst , *Senior Member, IEEE-2022*

[2] ”The Data Flow and Architectural Optimizations for a Highly Efficient CNN Accelerator Based on the Depthwise Separable Convolution ” authors Hung-Ju Lin1 · Chung-An Shen-2022

[3] ”FPGA-Based Convolutional Neural Network Accelerator with Resource-Optimized Approximate Multiply-Accumulate Unit”Authors Mannhee Cho and Youngmin Kim-2021

[4] ”Area and energe efficient shift and accumulator unit for object detection in IoT applications” Authors Anakhi Hazarika,Sowmyajith poddar,Moustafa M Nasaralla,Hafizur Rehmann-2021

[5] A High-Accuracy Hardware-Efficient Multiply–Accumulate (MAC) Unit Based on Dual-Mode Truncation Error Compensation for CNNs authors are SONG-NIEN TANG , (Member, IEEE), AND YU-SHIN HAN-2020.

[6] High Speed, Approximate Arithmetic Based Convolutional Neural Network Accelerator authors Mohammed E. Elbtity∗,†, Hyun-Wook Son†, Dong-Yeong Lee†, and HyungWon Kim-2020.

[7] Design of Floating-Point MAC Unit for ComputingDNN Applications in PIM,Authors Hun Jae Lee, Chang Hyun Kim, Seon Wook Kim School of Electrical and Computer Engineering Korea University Seoul, Korea-2020

[8] "Very Deep Convolutional Networks for Large-Scale Image Recognition"Authors: Karen Simonyan, Andrew Zisserman,2015.

[9] "Sequence to Sequence Learning with Neural Networks"Authors: I. Sutskever, O. Vinyals, and Q. V. Le Year: 2014

[10] "ImageNet Classification with Deep Convolutional Neural Networks" Authors: A. Krizhevsky, I. Sutskever, and G. E. Hinton,Year: 2012

[11] "Gradient-Based Learning Applied to Document Recognition"Paper: "Gradient-Based Learning Applied to Document Recognition" by Y. LeCun et al. (1998).

Author contributions:

Orcid Id: 0000-0003-2809-7644

Google Scholar Id :manjulabb@gmail.com

Mrs. Manjula Basavant Bhajantru is a Research Scholar in the Electronics and communication Engineering department at East West Institute of Technology Bangalore, affiliated with Visvesvaraya Technological University Belgavi. She completed her bachelor's degree in Electrical and Electronics Engineering at Gogte Institute of Technology, Belagavi , Affiliated to Karnataka university Dharwad, India, and her master's degree in Electronics & Communication Engineering from BMSCE, Bangalore Affiliated to Visvesvaraya Technological University Belgavi, Mrs. Manjula Basavant Bhajantri possesses a strong academic and research background, Particularly in the areas of Vlsi Design ,Embedded System, Digital System Design using Verilog, Analog Electronics and Network Analysis. She has made significant contributions to her field. He has organized various workshops, Seminars and Conferences in the field of Vlsi Design,digital Signal processing and Embedded system Designing.

Google Scholar Id :hirematthsharan123@gmail.com

Dr. Sharanabasaveshwar G Hiremath, is a highly skillful professor with over 20 years of experience in the field of engineering education. He has been awarded Ph D from Anna University in the faculty of ECE. His interest in research areas of Instrumentation, Digital Signal Processing, Neural Networks, Control Engineering, Hardware Design, Electronics Circuit Design, Instrumentation Sensors, Process Controls, PLC s and SCADA, Modeling & Simulation, Neural Networks, Telecom and related Teaching and promotional areas. He is awarded doctorate for his Research in the areas of Bio Medical Signal processing, Electronic circuit design, Chemical sensors, and Neural Network modeling/Simulation. He has more than 30 research publications at peer reviewed international journals and conferences. Five research scholars are doing research under his guidance and two of his scholars were awarded with Ph D degree from Visvesvaraya technological university Belagavi for the contributed works in the field of signal processing and Neural Network modeling. He is one of the main resource persons in the fields of Electronics, Especially in domains like Medical Signal Processing, Embedded System Design, Mathematical modeling and simulation, etc. He has given keynote address in various international and national conferences all over India. He has organized various workshops, Seminars and Conferences in the field of Signal processing and System Designing. At present he is working as Professor in the department of Electronics and Communication Engineering in East West Institute of Technology, Bengaluru, India. He is also consultant for companies in Bengaluru, Pune, Hyderabad, Chennai etc. His project consultancies include in the field of Signal Processing, Neural Networks and Embedded Systems Designing.