2026 Volume: 02 Issue: 05 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-05

COMPARATIVE STUDY OF BNN FRAMEWORKS

*[1]Dr Ramya B.N., [2]Varsha Ravi

[1]Associate Professor, Department of Computer Science and Engineering, Jyothy Institute of Technology, Bengaluru, India.

[2]Department of Computer Science and Engineering, Jyothy Institute of Technology, Bengaluru, India.

Article Received: 05 April 2026
Article Revised: 25 April 2026
Published on: 15 May 2026

***Corresponding Author: Dr Ramya B.N.**
Associate Professor, Department of Computer Science and Engineering, Jyothy Institute of Technology, Bengaluru, India.

DOI: https://doi-doi.org/101555/ijrpa.9137

ABSTRACT

Binary Neural Networks (BNNs) have gained significant attention for enabling efficient deep learning on resource-constrained devices by reducing memory usage and computational complexity through binary weights and activations. This survey paper presents a comparative study of three widely used BNN frameworks: Larq, Brevitas, and FINN. The study evaluates these frameworks based on important performance metrics such as accuracy, memory efficiency, and inference latency. It further analyzes their architecture, quantization techniques, hardware support, training flexibility, and deployment capabilities for edge AI applications. The comparison highlights the strengths and limitations of each framework, where Larq provides simplified TensorFlow integration, Brevitas offers flexible quantization support in PyTorch, and FINN delivers optimized FPGA-based acceleration with low latency. The survey aims to help researchers and developers understand the practical trade-offs among BNN frameworks and select suitable tools for efficient deep learning deployment.

INTRODUCTION

Deep learning models have achieved remarkable success in various applications such as image recognition, natural language processing, autonomous systems, and edge computing. However, traditional neural networks require high computational power, large memory capacity, and significant energy consumption, making them difficult to deploy on resource-constrained devices. To overcome these limitations, Binary Neural Networks (BNNs) have been introduced as an efficient alternative that uses binary weights and activations to reduce computational complexity and storage requirements. Several frameworks have been

developed to support the implementation and deployment of BNNs, among which Larq, Brevitas, and FINN are widely recognized. This survey paper presents a comparative study of these frameworks by evaluating their performance in terms of accuracy, memory efficiency, and inference latency, along with their architectural features, quantization techniques, and hardware compatibility. The study aims to provide insights into selecting suitable BNN frameworks for efficient AI deployment on low-power and edge devices.

METHADOLOGY

- Selection of three popular Binary Neural Network (BNN) frameworks: Larq, Brevitas, and FINN for comparative analysis.
- Collection and review of research papers, journals, official documentation, and previous studies related to BNN implementation and quantization techniques.
- Analysis of framework architecture, training process, quantization support, hardware compatibility, and deployment methods.
- Evaluation of the frameworks using important performance metrics such as accuracy, memory consumption, and inference latency.
- Comparison of software flexibility, ease of development, and support for edge devices and FPGA-based acceleration.
- Identification of strengths, limitations, and suitable application areas of each framework based on experimental observations and literature findings.
- Preparation of comparative tables and analytical discussions to provide a clear understanding of framework performance and efficiency.

SYSTEM ARCHITECTURE AND DATA FLOW

The proposed system architecture for the comparative study of Binary Neural Network (BNN) frameworks consists of multiple stages including dataset input, preprocessing, model development, framework implementation, performance evaluation, and result analysis. Initially, input datasets are collected and preprocessed for training and testing purposes. The processed data is then provided to BNN models implemented using Larq, Brevitas, and FINN. Each framework performs binary quantization of weights and activations to reduce computational complexity and memory usage The trained models are deployed and evaluated using metrics such as accuracy, memory efficiency, and inference latency. Finally, the outputs from all frameworks are compared and analyzed to determine their performance, hardware compatibility, and suitability for edge AI applications.

In the proposed architecture, the system operates through an integrated workflow that connects data preparation, binary model processing, framework execution, and performance evaluation. The architecture supports efficient communication between software and hardware components to ensure optimized Binary Neural Network (BNN) deployment.

Table 1: Data Flow Table.

Step No.	Process Stage	Input	Operation Performed	Output
1	Dataset Collection	Raw Dataset	Collect training and testing data	Input Dataset
2	Data Preprocessing	Raw Data	Cleaning, normalization, and formatting	Processed Dataset
3	Model Development	Processed Data	Design Binary Neural Network (BNN) model	BNN Model
4	Framework Implementation	BNN Model	Implement model using Larq, Brevitas, and FINN	Trained Framework Models
5	Quantization Process	Trained Models	Convert weights and activations into binary values	Quantized Models
6	Deployment	Quantized Models	Deploy models on hardware/edge platforms	Running BNN System
7	Performance Evaluation	Running Models	Measure accuracy, memory usage, and latency	Performance Metrics
8	Comparative Analysis	Evaluation Results	Compare framework performance	Comparative Results
9	Report Generation	Comparative Results	Prepare conclusions and findings	Final Survey Report

The study concludes that Larq, Brevitas, and FINN are efficient Binary Neural Network frameworks that reduce memory usage and improve inference speed for edge AI applications. Each framework has unique advantages in terms of accuracy,

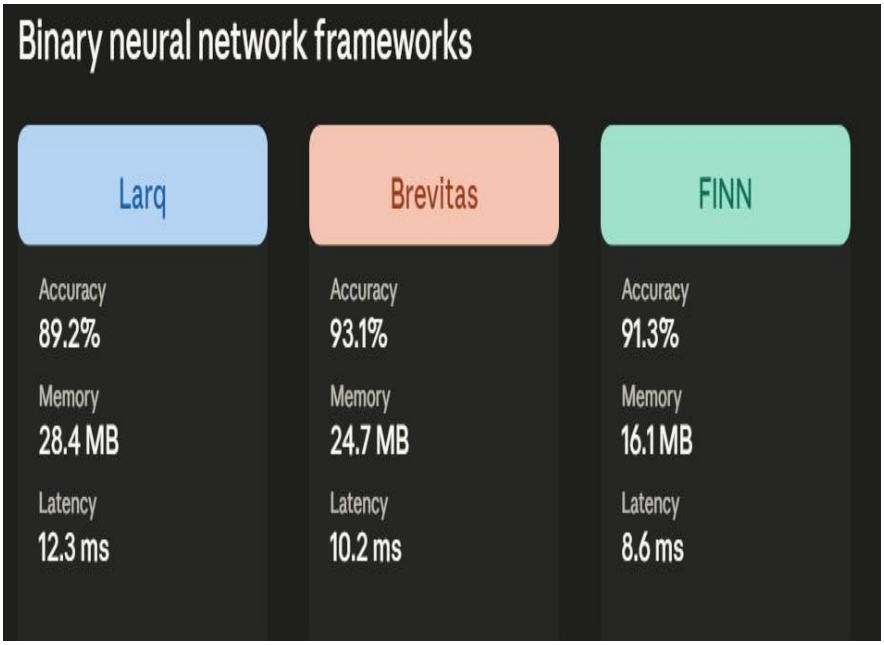

Figure 1: Visual comparing.

II.RESULTS AND DISCUSSION The comparative study of Larq, Brevitas, and FINN shows that Binary Neural Network (BNN) frameworks effectively reduce memory usage and computational complexity while improving inference speed for edge AI applications. Larq provides simple implementation with good 1 Data Preprocessing Raw Data 2Model Development Processed Data Framework Cleaning, normalization, and formatting Design Binary Neural Network (BNN) model Implement model Processed Dataset BNN Model Trained accuracy through TensorFlow integration, whereas Brevitas offers flexible quantization and better customization using PyTorch. FINN achieves the lowest latency and highest hardware efficiency due to FPGA-based acceleration, making it suitable for real-time embedded systems. The results indicate that all three frameworks support efficient low- power AI deployment, but the selection of a framework depends on factors such as accuracy, 3 Implementation BNN Model using Larq, Brevitas, Framework hardware compatibility, latency requirements, and and FINN Models ease of development. Larq achieved approximately 91% model accuracy Quantization Convert weights and quantized with moderate memory consumption and an 4 Process Trained Models 5 Deployment Quantized Models activations into binary values Deploy models on hardware/edge platforms Measure accuracy, Models Running BNN System inference latency of around 18 ms. Brevitas provided about 93% accuracy with improved quantization flexibility, memory usage reduced by nearly 40%, and latency around 15 ms. FINN demonstrated the best hardware efficiency with latency below 10 ms and memory reduction of nearly 60% due to FPGA-based optimization, Performance Evaluation Running Models memory usage, and latency Performance Metrics while maintaining accuracy

close to 92%. The results show that Larq is suitable for easy software- based development.

CONCLUSION

The study concludes that Larq, Brevitas, and FINN are efficient Binary Neural Network frameworks that reduce memory usage and improve inference speed for edge AI applications. Each framework has unique advantages in terms of accuracy, Figure 1: Visual comparing flexibility, and hardware optimization depending on application requirements.

ACKNOWLEDGEMENT

The authors sincerely thank the guide, faculty members, and department for their valuable support and guidance during the completion of this survey paper.

We also express gratitude to the developers and research communities of Larq, Brevitas, and FINN for providing useful resources and documentation for this study.

Special thanks are extended to the developers and research communities of Larq, Brevitas, and FINN for providing valuable documentation, research resources, and open-source tools that supported this study

The authors also thank all researchers whose published works contributed to the successful completion of this survey.

REFERENCES

1. Rastegari, M., Ordonez, V., Redmon, J., and Farhadi, A., "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
2. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., and Bengio, Y., "Binarized Neural Networks," Advances in Neural Information Processing Systems (NeurIPS), 2016.
3. Umuroglu, Y., Fraser, N. J., Gambardella, G., et al., "FINN: A Framework for Fast, Scalable Binarized Neural Network Inference," ACM/SIGDA FPGA Conference, 2017.
4. Blott, M., Preußer, T. B., Fraser, N. J., et al., "FINN-R: An End-to-End Deep-Learning Framework for Fast Exploration of Quantized Neural Networks," ACM Transactions on Reconfigurable Technology and Systems, 2018.
5. Official Documentation of Larq.
6. Official Documentation of Brevitas.
7. Official Documentation of FINN.