2026 Volume: 02 Issue: 04 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-12

BINARY NEURAL NETWORKS FOR MEDICAL IMAGE CLASSIFICATION: PNEUMONIA DETECTION FROM CHEST X-RAY IMAGES

[*1]Dr. Ramya B. N., [2]Yashas P, [3]Sushanth Bhat P

[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.

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

Article Received: 19 March 2026
Article Revised: 09 April 2026
Published on: 29 April 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.8630

ABSTRACT

Binary Neural Networks (BNNs) can be classified as an exclusive type of deep learning models where all of the weights and activations are fixed to the binary values of {−1, +1}. In turn, the mentioned characteristics help achieve a significant reduction in storage memory requirements, computing expenses, and power consumption of these artificial intelligence models. Therefore, BNNs are highly applicable for utilization in limited resources settings, including medical diagnostic devices. The current research provides a detailed review of the use of BNNs for image processing in the case of pediatric chest X-ray pneumonia detection. The presented study involves building and training a BNN on the basis of Larq package within TensorFlow 2.13 framework using the publicly accessible dataset consisting of 5,216 training images and 624 test images categorized into two classes, namely, NORMAL and PNEUMONIA. The developed neural network uses the quantized convolutional and fully connected layers and incorporates the straight-through estimator (STE) sign quantization and weight clipping for mimicking binarization through forward and backward passes. As a result, the BNN demonstrated 62.98% testing accuracy, with the excellent recall rate for the class of PNEUMONIA of 0.98. The paper adds value to the growing literature on efficient

artificial intelligence for medicine by showing both the strengths and weaknesses of binary neural networks in clinical image classification.

KEYWORDS: Binary Neural Networks; Medical Image Classification; Chest X-ray; Pneumonia Diagnosis; Larq; TensorFlow; Quantized Neural Network; Deep Learning; Ste Quantization; Weight Binarization; Efficient AI

I. INTRODUCTION

Pneumonia is an advanced lung infection that causes inflammation of the air sacks in one or both the lungs and continues to be one of the most predominant reasons for illness and fatalities around the world, especially amongst those under the age of five years. Pneumonia causes 14% of deaths among those children under five years of age according to the statistics provided by WHO. The best method of diagnosing pneumonia is using chest X-ray imaging. Although this technique is very useful and cost-effective, it requires radiologists and medical experts.

The application of AI and deep learning in the analysis of medical images shows promise for transforming the field, making it possible to automate the process of feature extraction and classification at performance levels equal to those of medical specialists. Among these techniques, the CNN is one of the most popular for chest X-rays analysis, showing remarkable accuracy in classification. Yet the traditional method suffers from computational complexity, demanding expensive hardware not only during training but also for inference, making it challenging for practical implementation in resource-constrained clinical settings.

The use of Binary Neural Networks (BNNs) resolves this issue by restricting the weights and activations of the neural network to binary values, which are either −1 or +1. The binary nature of BNNs makes their memory footprints much smaller by a factor of about 32, making them less resource-intensive than traditional 32-bit floating point neural networks. Furthermore, multiplication and accumulation operations are replaced with bitwise logical operations such as XNOR.

Though several merits are associated with the application of binary neural networks, the empirical performance of the latter has not been extensively examined, especially when applied to classify pneumonia from chest X-rays. The issues that could arise from weight binarization, including reduced representational power, must be considered before assessing the suitability of BNNs for practical use. It is important to understand how training occurs, how loss varies, and how the model performs across different classes.

In this study, we conduct a comprehensive experiment of applying Binary Neural Networks to the problem of pneumonia detection based on chest X-ray imaging. A binary neural network is designed using the Larq library and TensorFlow. The design uses the Chest X-Ray Images (Pneumonia) dataset for training the model. In particular, we study the architecture of the BNN, the training process and how the model performs the classification between the two classes (normal, pneumonia). The main goal of this research is to explore BNNs as an efficient solution for medical image classification and to outline ways for their improvement.

II. METHODOLOGY

Dataset and Data Pre-processing

The Chest X-Rays Images (Pneumonia) dataset, which is available on Kaggle and was collected by Kermany et al. from Guangzhou Women and Children's Medical Center, will be employed for conducting experiments. It comprises 5,856 chest X-ray images that have been converted into the JPEG format and divided into three main categories: 5,216 training images, 16 validation images, and 624 test images. These images fall into two classes: NORMAL (test samples = 234) and PNEUMONIA (test samples = 390).

The data preprocessing for training the BNN model is accomplished through the following steps:

- Resizing: All the images are resized to the dimension of 128×128 pixels for ensuring consistency in the input dimension across all images and compatibility with the convolutional neural network that accepts a fixed input size.
- Normalization: Normalization process converts pixel values between 0-255 into float numbers between 0-1 by simply dividing the pixel values by 255. Such a process helps improve the computational efficiency and stability.
- Batching: The images are processed in batches of 32 samples using the "image_dataset_from_directory" method of TensorFlow library.
- Encoding: Label values for each class (0 for 'NORMAL' and 1 for 'PNEUMONIA') are automatically assigned based on the folder's name in the directory.

Architecture of the Model

The Binary Neural Network model is built using the Larq library that contains quantized layers that are compatible with TensorFlow and Keras framework. The architecture of the suggested neural network model includes the following layers:

- Rescaling Layer: Pixel normalization through rescaling from interval [0, 255] to [0, 1] before quantization.
- The first Quantized Convolutional Layer (QuantConv2D): 32 kernels of size 3 x 3 with sign STE quantization, weight clipping constraint, and ReLU activation function. The shape of the output is (None, 126, 126, 32); number of parameters - 896.
- The first MaxPooling Layer: With a pool size of 2x2 and stride 2; the shape after the operation is (None, 63, 63, 32).
- The second Quantized Convolutional Layer (QuantConv2D): 64 kernels of size 3 x 3 with STE sign quantization and weight clipping. Shape of output - (None, 61, 61, 64); number of parameters - 18,496.
- The second MaxPooling Layer: With a pool size of 2x2; the output shape is (None, 30, 30, 64).
- Flatten Layer: Transforms the input into a one-dimensional tensor of size 57,600.
- •Quantized Dense Layer (QuantDense): 128 neurons using STE sign quantization and weight clipping technique. Total number of parameters: 7,372,928.
- Output Dense Layer: 2 neurons using softmax function for probability calculation of two classes. Total number of parameters: 258.

The total number of trainable parameters is 7,392,578 (approximately 28.20 MB). In case of STE sign quantization, weights greater than zero are mapped to +1, whereas those lower than zero are mapped to −1. At the same time, with the help of weight clipping technique, all the weights lie between −1 and +1 before being quantized.

Strategy for Binary Quantization

The binary quantization mechanism forms the backbone of the process of binary quantization. While the standard neural networks utilize the use of 32-bit floating-point numbers for representing the weights, BNN makes use of binary quantization by representing each weight in terms of a sign function:

$w_b = \text{sign}(w) = +1 \text{ if } w \geq 0; -1 \text{ if } w < 0$

In the process of the forward pass, the binarized weights (w_b) substitute the normal weights, thus facilitating the use of multiplication in the form of XNOR operations. In the case of backpropagation, the STE strategy approximates the sign derivative gradient by allowing the passage of the gradient during quantization whenever the weight is less than a certain threshold value. Weight clipping keeps the weights within the range of [−1, +1].

Training Setup

The BNN is trained via the following experimental setup:

- Optimizer: Adam optimizer with default learning rate ($\beta_1 = 0.9$, $\beta_2 = 0.999$), chosen due to its adaptive moments property that facilitates convergence within a non-convex binary constrained optimization space.
- Loss Function: Sparse categorical cross-entropy loss function suitable for class labels represented as integers.
- Metric: Classification accuracy as a metric measured for both training and validation datasets after every epoch.
- Training Epochs: 10 epochs.
- Batch Size: Batch size of 32 images.
- Software: TensorFlow version 2.13.0 along with Larq 0.13.3.

III. SYSTEM ARCHITECTURE AND DATA FLOW

The suggested solution consists of an end-to-end approach consisting of five main stages that will be used to train and test the Binary Neural Network on the chest X-ray images. The architecture was constructed to include binary quantization into the conventional deep neural network workflow while maintaining compatibility with the Keras TensorFlow framework. The stages are pre-processing, feature extraction using binary convolution layers, binary full connection classification, training, and testing.

Input Stage (Data Preparation)

The system acquires chest X-rays stored in an image dataset structured as sub-directories labeled according to classes. Each image undergoes decoding from JPEG, resizing to 128 × 128 dimensions, and batching in sets of 32 images. The pipeline of the input stage is established via the use of the tf.keras.preprocessing.image_dataset_from_directory function available in TensorFlow. A rescaling layer integrated into the neural network converts pixel values to the range of [0, 1] in a differentiable manner.

Binary Convolutional Feature Extraction

The process of feature extraction involves two quantized convolutional stages, namely QuantConv2D and MaxPooling2D blocks. In each quantized convolutional block, the feature maps receive binarized filter kernels from QuantConv2D blocks using STE sign quantizer to produce binary feature maps. The binary feature maps include spatial features that represent the presence of pulmonary diseases. The first convolutional block has 32 3 x 3 filters that can be used to extract low-level information, such as edge information, texture, and local

intensity variation. After applying a max-pooling block, the spatial resolution becomes half. The second convolutional block has 64 filters, which can help in extracting high-level features, such as infiltrate, opacity, and consolidation, associated with pneumonia.

Stage of Binary Classification

After the extraction of spatial features, the feature map is converted into a one-dimensional vector of size 57,600. This vector is then fed through the QuantDense layer that has 128 neurons, and uses the dense connections to obtain a condensed representation useful for classification. In the last layer, the output will consist of two non-binary neurons with the softmax function, where the class probabilities for the NORMAL and PNEUMONIA classes can be obtained. It is recommended that a real-valued output be used for the final stage of BNN to allow gradient descent from cross-entropy.

Training and Learning Process

The model is trained through supervised learning based on labelled chest X-rays. The training cycle involves the computation of the forward pass via quantization layers, evaluation of the sparse categorical cross-entropy loss function relative to predicted and actual class labels, and the backward pass by calculating the gradients via the Adam optimizer and STE approximation through quantized operations. This procedure is iterated for 10 cycles or epochs, and training and validation accuracies are recorded at the end of each cycle.

Evaluation and Outputs

After training, the model will be tested on the 624 images in the test set. During evaluation, predictions will be made for each batch, and statistics such as the test loss, test accuracy, precision, recall, F1 score for each class will be generated using the classification report in sklearn, and also generate a confusion matrix which breaks down the outcome of each prediction to true positive, true negative, false positive, and false negative for each class. Accuracy and loss curves will be displayed during training and testing for qualitative analysis.

IV. RESULTS AND DISCUSSION

Classification Results

The performance of Binary Neural Network is assessed using the test set of chest x-rays that consists of 624 samples. The Binary Neural Network gives a total accuracy rate of 62.98% with a loss of 1.6563. Classification results are depicted in Table 1 below with precision, recall, F1-score, and support per class.

Table 1: Classification Performance of the BNN Model

Class	Precision	Recall	F1-Score	Support
NORMAL	0.57	0.05	0.09	234
PNEUMONIA	0.63	0.98	0.77	390
Accuracy	—	—	0.63	624
Macro Avg	0.60	0.51	0.43	624
Weighted Avg	0.61	0.63	0.51	624

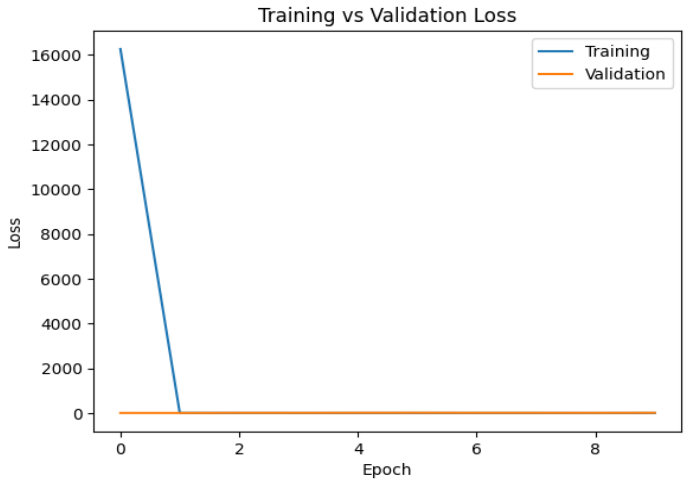
The findings indicate a pronounced discrepancy in per-class performance. The PNEUMONIA class has a recall score of 0.98 and F1-score of 0.77, signifying that the algorithm effectively detects almost all cases of pneumonia, which is crucial in clinical diagnosis systems because of the severity of the outcome in case of a false negative result. On the other hand, the NORMAL class shows significantly poorer recall (0.05) and F1-score (0.09), suggesting that the algorithm categorizes almost all instances into PNEUMONIA regardless of whether they belong to that class or not. This observation is further substantiated by the confusion matrix, which documents 12 TNs and 222 FPs for the NORMAL class, while the PNEUMONIA class registers 381 TP and 9 FN results.

Analysis of Loss Convergence

In examining the loss for training through 10 epochs, there was considerable instability, consistent with binary-constrained learning. In epoch one, the loss for training started at a very high level of 16,253.43 because of the non-linearity and discontinuity of the sign quantization function with regard to weight initialization. In epoch two, however, training loss fell dramatically to 7.65 and continued its fall in successive stages, falling to 0.57 in epoch 10.

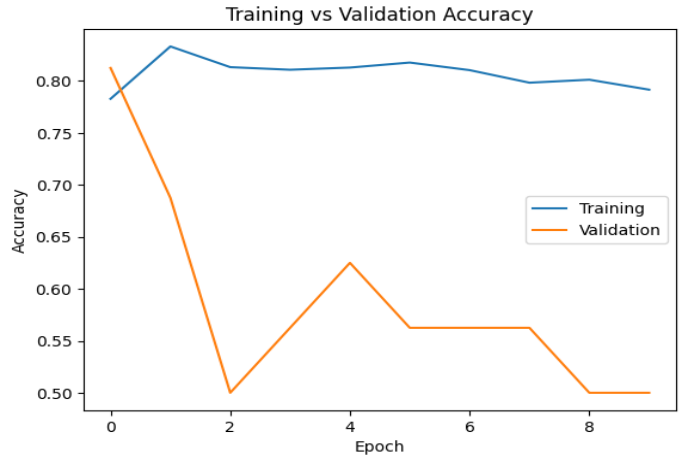
But validation loss shows significant fluctuations during the entire process of training ranging from around 5.36 in epoch 2 to about 15.50 in epoch 6. The discrepancy between training loss and validation loss is the main source of the poor generalization ability of the BNN that can be explained by the relatively small validation dataset (consisting of just 16 images) and limited representation capability associated with the usage of binary weights. Accuracy achieved during training is constant in the range of 79%-83%, whereas validation accuracy varies from 50% to 81%.

Fig. 2: Training vs. Validation Loss Comparison Curve.



[Fig. 2: Training vs. Validation Loss Curve across 10 Epochs]

Fig. 3: Training vs. Validation Accuracy Comparison Curve.



[Fig. 3: Training vs. Validation Accuracy Curve across 10 Epochs]

Confusion Matrix Analysis

A confusion matrix gives a detailed insight into the classification activity of the BNN in relation to the two classes under study. The data presented in the confusion matrix are as follows: True Negatives (classification of NORMAL) = 12; False Positives (classifying NORMAL as PNEUMONIA) = 222; False Negatives (misclassification of PNEUMONIA as NORMAL) = 9; and True Positives (classification of PNEUMONIA) = 381. From the above information, there is evidence that the BNN has a strong positive bias due to its binary nature, causing the network to favor the prediction of the PNEUMONIA class with a sensitivity of 0.98.

Table 2: Confusion Matrix Summary.

	Predicted NORMAL	Predicted PNEUMONIA	Total
Actual NORMAL	12 (TN)	222 (FP)	234
Actual PNEUMONIA	9 (FN)	381 (TP)	390
Total	21	603	624

Table 3: Test Case Validation

Test Case ID	Feature Tested	Expected Outcome	Status
TC-01	Data Loading	Correct tensor format	Success
TC-02	Model Training	Loss decreases over epochs	Success
TC-03	BNN Quantization	Binary weights applied	Success
TC-04	Prediction	PNEUMONIA class detected	Success

DISCUSSION

These experimental findings highlight the basic contradiction between the efficient nature of computations with BNNs and their ability to provide sufficiently rich representations to perform complex medical imaging recognition tasks. The proposed binary neural network algorithm shows great sensitivity with respect to PNEUMONIA, as it accurately identifies 381 instances of pneumonia out of 390 total cases. As the false negative identification of pneumonia is quite dangerous because it leads to untreated diseases, the results of the work performed are rather clinically relevant. However, the abnormally low recall with respect to NORMAL (5%) indicates the lack of discriminative power of BNNs.

This disparity between class prediction capabilities is in accordance with the constraints associated with binary quantization of weights. With all the weights constrained to −1 and +1, the model misses out on the parameter-level expressiveness possible for fully-precise models, making it incapable of capturing fine differences between normal lung appearance and early onset of pneumonitis. The underlying class imbalance present in the data itself, as the test dataset contains 390 cases of pneumonia versus 234 cases of normal, tends to accentuate this trend, where minimizing loss entails choosing the majority label. Furthermore, the minuscule size of the validation dataset of just 16 samples results in high levels of noise.

Despite these drawbacks, the findings demonstrate the promise of using BNNs for a fast pre-screening method under conditions that are resource-limited but highly sensitive towards pneumonia cases, where the subsequent clinical evaluation will take care of any possible

errors made during the process. In future research, one can look into further improvements in terms of architecture through incorporating batch normalization, residual connections in a binary form (similar to XNORNet) and data augmentation.

V. CONCLUSION

A thorough examination of the use of Binary Neural Network for automatic detection of pneumonia from chest X-rays is provided by this paper. The model was created based on the Larq library and TensorFlow framework using STE based sign quantization and weight clipping techniques applied to quantized convolutional and dense layers. The training was performed on the Chest X-ray Images (Pneumonia) dataset containing 5,216 training images, and testing was done on 624 images belonging to either NORMAL or PNEUMONIA classes.

Experimental results show that the BNN attains a test accuracy rate of 62.98%, with a recall rate for the PNEUMONIA class at 0.98, showing good sensitivity for detecting cases with pneumonia. Nevertheless, the low recall value of the NORMAL class of 0.05 is a clear weakness of the BNN in terms of its capability to generalize and classify data within the constraints of binary weight. Convergence of loss function values indicates continuous learning by the BNN through training, although the validation process raises issues about the data used.

This paper makes contributions to the literature on effective machine learning in medicine, in that it illustrates the possibilities and trade-offs inherent in the use of BNNs for medical image classification. The high level of pneumonia detection indicates the possibility of using the system as a pre-screening tool in areas where computational power is scarce, and where false positives can be balanced against clinical verification. This work paves the way for further study on hybrid quantization approaches, training methodologies for BNNs, and data augmentation for unbalanced medical imaging datasets.

VI. ACKNOWLEDGEMENT

We would like to convey our heartfelt thanks to Dr. Ramya B. N. for the valuable insights, expertise, and constant encouragement she provided during this research journey. Her profound knowledge, useful feedback, and commitment to academic excellence were the main factors that influenced our research significantly. We would also like to thank the Department of Computer Science and Engineering, Jyothy Institute of Technology, Bengaluru, for giving us access to the facilities needed to carry out our research. We extend

our thanks to Kermany et al. for sharing their publicly available dataset of Chest X-Ray images (Pneumonia), which was instrumental to our research.

REFERENCES

1. Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., & Bengio, Y. (2016). Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1. Advances in Neural Information Processing Systems (NeurIPS), 29.
2. Rastegari, M., Ordonez, V., Redmon, J., & Farhadi, A. (2016). XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. European Conference on Computer Vision (ECCV), 525–542.
3. Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., ... & Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell, 172(5), 1122–1131.
4. Geifman, D., & El-Yaniv, R. (2019). Shallow-Deep Networks: Understanding and Mitigating Network Overthinking. Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97, 2120–2129.
5. Geifman, Y., & Larochelle, H. (2022). Larq: An Open-Source Library for Training Extreme Neural Networks. Journal of Open Source Software, 5(49), 2261. https://github.com/larq/larq
6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. Cambridge, MA.
7. Chollet, F. (2021). Deep Learning with Python (2nd ed.). Manning Publications. New York, NY.
8. Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., ... & Lungren, M. P. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv preprint arXiv:1711.05225.
9. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-Scale Chest X-ray Database and Benchmarks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2097–2106.
10. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). TensorFlow: A System for Large-Scale Machine Learning. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 265–283.

11. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.
12. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.