SCENIC: An Area and Energy-Efficient CNN-based Hardware Accelerator for Discernable Classification of Brain Pathologies using MRI

Bodepu Sai Tirumala Naidu*‡, Shreya Biswas†‡, Rounak Chatterjee†‡, Sayak Mandal†,
Srijan Pratihar†, Ayan Chatterjee**, Arnab Raha§, Amitava Mukherjee¶, Janet Paluh¶
*SandLogic Technologies, Bangalore, India.
†Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata, India.
**Network Science Institute, Northeastern University, Boston, USA.
§Advanced Architecture Research, Intel Edge.AI, Intel Corporation, Santa Clara, USA.
¶College of Nanoscale Science and Engineering, Nanobioscience, SUNY Polytechnic Institute, Albany, New York, USA.
Corresponding author email: *rounakchatterjee007@gmail.com*

***Abstract*—Biomedical brain imaging lies at the interface of visual, and spatial neuropathology and neurosurgical intervention. Future treatments are expected to require greater feature detail in imaging as well as underlying mechanisms that will be synergistically advanced. This will be made more accessible via low power embedded devices and cloud platforms through applied deep learning software and hardware analysis. Towards that end, we develop a hardware-software co-design technique referred to as SCENIC (or, Separable Convolution Enabled Non-Invasive Classification) for the identification and classification of glioma brain tumors, using physical tissue features reflected in parameter weighted MRI scan types such as- T1-w, T1-ce, T2-w and FLAIR. The high performance hardware exceeds current accuracy, resource efficiency and time consumption parameters. The proposed SCENIC-CNN Accelerator is synthesized on 45 nm process technology and it can operate at a minimum frequency of 1 GHz while maintaining low-power consumption of only 0.36 W and a low chip-area size of 0.431 mm^2. Our classification accuracy achieves 98.3% in detection of the presence of a tumor pathology and 99.62% within classification of imaging modalities that relate to tissue parameters such as fat content, blood or CSF flow and tissue density. Compared to prominent state-of-the-art Convolutional Neural Network (CNN) models being designed for biomedical imaging, SCENIC is competitive versus XceptionNet, InceptionV3, ResNet-50, and VGG-16. With model compression techniques, SCENIC requires a memory space of less than 0.265 MB. We discuss design methodology as applied to future goals to meet challenging needs to distinguish tumor origins, such as glioma and metastasized tumors, along with other neuropathologies that may be TBI, vascular or developmental. SCENIC will aid impactful, cost-effective, rapid and accurate neurosurgical intervention and treatments.**

***Index Terms*—Brain tumor, biomedical image, discrete wavelet transform, hardware accelerator, neural networks, MRI, pathology classification**

I. Introduction

Brain tumors have histological complexity and heterogeneity, challenging epidemiology and limited treatments that result in high morbidity and mortality. Ranking among the most fatal of tumors in the United States, 18,600 adults will die from primary cancerous brain and CNS tumors in 2021 [1]. High grade gliomas (HGGs) are most common, although brain metastases occur in 20-30% of patients with cancers [2]. Despite maximal surgical resection with concurrent radiotherapy and alkylating chemotherapy the survival rate is abysmal at 3% in adults age 65 and older and 27% in youth to middle aged 20 to 39 years. A clear primary source of the tumor is not always evident, requiring both biomedical magnetic resonance imaging (MRI) and clinical diagnostics to determine the origin and appropriate management strategy [2]. However, inconsistencies remain and 89% of non-metastatic lesions identified by MRI are revealed to be malignant on biopsy [3]. The development of artificial intelligence platforms and tools are expanding to generate the next generation deterministic capabilities in parallel with medical imaging advancements.

To improve the diagnostic capabilities of MRI for identifying tumors and establishing a Deep Learning (DL) platform for future distinction of other brain pathologies, we establish a DL platform with edge based hardware. The latter enables image processing, management and maintenance when large databases are analyzed via content-based image retrieval (CBIR). The CNN incorporates the information in MRI to identify pathology and differentiate anatomical differences through parameter weighting. We chose T1-weighted (T1-w), T2-weighted (T2-w), Fluid-Attenuated Inversion Recovery (FLAIR) contrasts, and T1-ce weighted scans as initial ideal techniques for monitoring new lesions (T1-w; T1-ce) versus older inactive lesions (T2-w), and relevant to fat content, cerebrospinal fluid, edema and even further distinguishing neural tissue changes, such as demyelination (FLAIR). The CNN architecture developed by us for brain tumor classifications is evaluated against several pre-trained networks to provide an effective decision-support tool with near instantaneous results to benefit diagnostic radiology for actionable neurosurgery.

A machine learning algorithm that has achieved substantial results in image segmentation and classification is a CNN. However, the power consumption of a CNN has an exponential relationship with the accuracy it offers. Increasing the accuracy of models becomes a resource exhaustive affair that introduces impractical barriers to applications. Therefore, it is important to achieve high accuracy while consuming a reasonably low amount of power. The proposed hardware accelerator is a fully automatic CNN architecture for brain tumor classification of different types of tumors. The developed network was optimized to include a reduced parameter set than already-existing pre-trained networks [4]–[7] while offering a 99.62% accuracy, making it optimum for designing the hardware. With its optimized generalization capability and execution speed, the new CNN architecture is being developed as an effective decision-support tool for radiologists in medical diagnostics which provides almost instantaneous actionable results.

II. SCENIC framework Design

To achieve our main goals of image classification in software and integration into a low power platform, we implement

‡*These authors contributed equally to this work.*

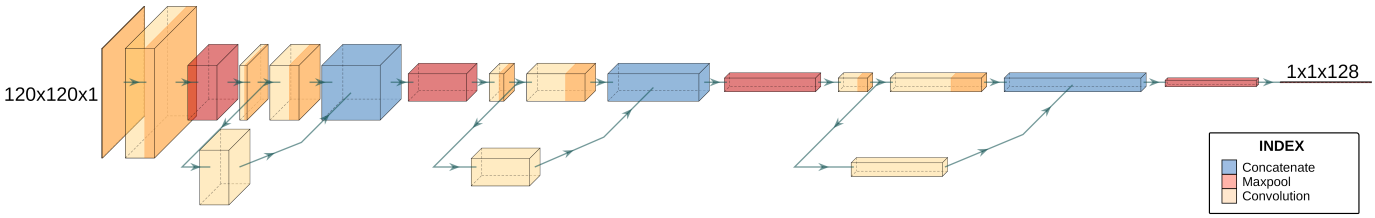

Fig. 1. SCENIC Architecture

a DL architecture and optimize its performance through pre-processing and layer compression techniques for biomedical tumor image classification.

A. CNN Architecture accuracy for MR image classification

To achieve our main goals of image classification in a low power platform, we designed a simple CNN architecture for the proposed SCENIC. The model was initialized with sequential layers and following this, we flattened the data and added two additional Dense layers juxtaposed with a Dropout of 0.5. The final activation was with Softmax. The peak accuracy demonstrated by this rudimentary architecture for the detection of tumor in MR images is 84.3%. The MR image data is first pre-processed using a Discrete Wavelet Transform (DWT). This mitigates the impact of noise in training as well as reduces the image dimensions to half of their original size. The impact of DWT and additional use of Spark architecture [8] are described in the following sections. The use of DWT compression on model accuracy was unsatisfactory, instead an algorithmic simplicity approach to achieve successful and efficient implementation in our hardware was applied.

B. Layer compression using Spark architecture

One of the most important aspects of SCENIC is to capture the complexity of tumor classification while having algorithmic simplicity for successful and efficient implementation of the hardware accelerator. For this, we incorporate the Spark module in SCENIC architecture at various stages. A Spark module is obtained from the Fire module [9], a building block for convolutional neural networks, by compressing it via depth-wise separable convolutions. In the Squeeze layer, the input is first passed through a batch normalization layer followed by a convolution operation. Then an element-wise Rectified-linear-non-Linearity (ReLU) activation function is applied. The Squeeze layer is followed by the expand layer, where it is subjected to two parallel convolution operations using 1×1 kernels and 3×3 kernels. Similar to the Squeeze layer, each convolution operation is preceded by a batch normalization layer and is followed by a ReLU operation.

C. MR image pre-processing

The use of WT has broad applications in the analysis of stationary and non-stationary signals. These applications include the removal of electrical noise from the signals, detection of abrupt discontinuities, and compression of large amounts of data. The discretized WT, DWT, is key to eliminating noise and compressing the images to the desired size [10].

All images were resized into 240x240 in order to apply the same DWT function for noise reduction and image compression. After performing the DWT operation on each image, the resultant pre-processed images are resized to 120x120 to coincide with memory requirements of the hardware implementation.

D. SCENIC - a Deep Learning classifier

A DL image classification model is resource intensive for higher accuracies such that a hardware design for the same would be impractical. In order to efficiently translate software to hardware, we need to reduce the complexity of the DL model such that it can be embedded with limited hardware resources, which SCENIC achieves by reducing the number of parameters in a DL model. This enables us to perform complex operations in low power without losing accuracy to distinguish neuropathologies.

The input to the first convolution layer is a fixed size 120x120 grayscale image. The convolution layer is followed by a max-pooling 2D layer. The image is then passed through a stack of three identical blocks consisting of - a convolutional layer with 3×3 filter size, a spark layer, and a max-pooling layer. The size of the convolutional layer corresponds to the minimum layer size required to capture the notion of left/right, up/down and center. All max-pooling 2D layers have a 2×2 window size and a stride of 2. Figure 1 shows the schematic diagram of SCENIC.

E. SCENIC Deep Learning architecture for detection of Tumors in MR Images

To classify MR images with or without pathology (tumors), we use SCENIC architecture and add additional layers after the last max-pooling 2D layer. This layer is followed by an average-pool 2D layer - also with a window size of 2×2 and a stride of 2, a flatten layer, and a dense layer with a dropout. The final layer is another dense layer with two channels for the two output classes - either pathology is present or there is no pathology detected in the MRI scan. All hidden layers are equipped with the ReLU activation function.

F. SCENIC Deep Learning Architecture for differentiating anatomical structures

MRI weighted parameters allow detection of different features, that are for example relevant for newer versus older lesions or anatomical damage and edema. We next incorporated the ability to differentiate anatomical and pathology features based on four classes of MRI scans- T1, T1-ce, T2-w and FLAIR. We again use SCENIC architecture, and the last max-pooling 2D layer is followed by an average-pool 2D layer - again with a window size of 2×2 and a stride of 2, and a flatten layer. The final layer is another dense layer with four channels for the respective four classes. All hidden layers are equipped with the ReLU activation function.

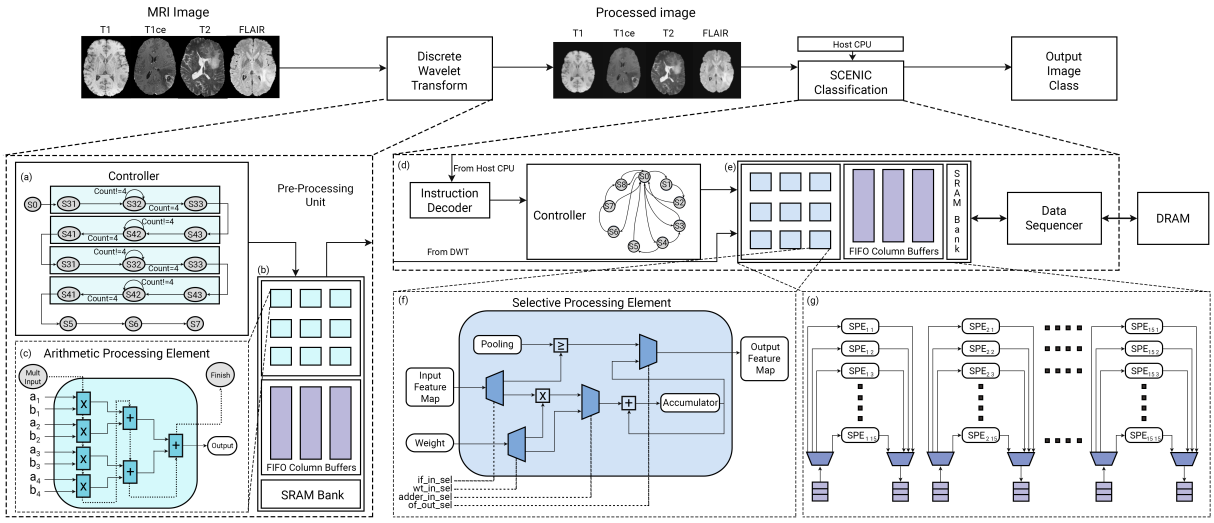

Fig. 2. MR Images are first compressed from 240×240 to 120×120 in DWT Pre-processing Unit using (c) Arithmetic Processing Unit controlled by a FSM Controller (a), the data-flow for pre-processing the images is performed as show in (b). The DWT pre-processed MR Images and CNN model are fed into the Memory Hierarchy of SCENIC-CNN Accelerator (e) which are processed by Selective Processing Elements (f) arranged in 15×15 grid array as show in (g)

III. SCENIC Hardware Architecture Design

SCENIC hardware Accelerator provides a low cost end-to-end platform that can provide high throughput and energy efficiency while realizing high accuracy of the CNN. Towards the end, SCENIC is implemented on a low-cost and low power Application Specific Integrated Circuit (ASIC) based CNN Accelerator that is also able to obtain comparable performance to that of the GPUs and customised processors. Lots of research has been done on the efficient processing of CNNs in hardware [11], [12]. SCENIC hardware accelerator leverages key architectural techniques and ideas laid down for efficient CNNs processing and implements them for biomedical applications at the edge.

A. DWT Pre-processing Block

The input images forwarded to the accelerator were pre-processed first using DWT involving Daubechies wavelets [13] to generate a compressed image. The pre-processing block computes the matrix multiplication of the Daubechies forward transform matrix with the matrix of image pixels and subsequently generates a compressed version of the original image. We next describe the core components of the DWT pre-processing block that includes the arithmetic processing element, data flow and DWT controller.

1) Arithmetic Processing Engine: The Arithmetic Processing Engine (APE) computes the result of a cell of the output matrix. Inputs to the processing element are a row of the first matrix and a column of the second matrix each having four elements. Coefficients of Daubechies wavelet function are of floating-point precision. The APE employs floating-point adders and multipliers to compute the corresponding result as shown in Figure 2(c). In our case, we have used four multipliers and three adders. Each multiplier performs an element-wise multiplication of the elements in the rows and columns. The results obtained from multipliers are added using three adders. We have used a counter to keep track of the number of times the APE is reused.

2) Data flow: The input image and the Daubechies forward transform matrix are at first broken up into several 4×4 blocks and then using a sequencer we generate the rows of the transform matrix and columns of the data matrix that are to be fed to the APE. The rows and columns are at first fed into FIFO buffer banks as shown in Figure 2(b). Each FIFO is of depth four and each bank consists of four such FIFOs. There are four sets of column buffer banks, each set of which contains the data of the columns of the input image and a row buffer bank which contains the data of the rows of the transform matrix. In addition to these there is an output FIFO buffer bank of dimension 4×4 which stores the results computed from the APE.

3) DWT controller: FSM based controller with sixteen states is used to control the loading and reading operations from the FIFO buffer banks and to compute results of matrix multiplication using the APE as shown in Figure 2(a). The first state is used to load inputs into the row and column FIFO buffer banks. The next three states read the first row from the row FIFO buffer bank and the entire first set of column FIFO buffer banks, *i.e.*, all the four columns of the input image data, concerned, computes the results for the first row of the output matrix, and loads the results in the output FIFO buffer bank. Similarly, in the next nine states, we have loaded the second, third and fourth rows from the row buffer bank and the corresponding set of column buffer banks, computed the results for the rest of the rows in the output matrix and stored the results in the output FIFO buffer bank. In the next two states we read from the output FIFO buffer bank and provide corresponding outputs. The final state stops all computations.

B. SCENIC Hardware Accelerator

1) Selective Processing Engine: Each Selective Processing Engine (SPE) has two inputs for input feature map and filters, and an output for output feature map. Based on the select line input it can be used to compute multiply-and-accumulate (MAC) for convolution, comparison for max-pooling and addition for average-pooling and bias-addition as shown in Figure 2(f). The multiplexers across the arithmetic blocks are enabled by these select lines inside each SPE block. They reuse the same SPEs for different recurring operations like pooling,

depth-wise separable convolutions, and fully-connected. Such a consolidated approach enables us to optimise the area of the accelerator considerably. Multiple instances of SPEs are instantiated in a grid array of 15×15 as shown in Figure 2(g). Input feature maps are distributed from IF-FIFO banks of each SPE column array and similarly, output feature maps are collected from each SPE column array to OF-FIFO banks. All SPE blocks in the array load the filter data from the common interface since at any point in convolution, all SPEs are sent the same weight data.

2) Memory Hierarchy: We implemented various levels of memory hierarchy to decrease the latency and high energy consumption of DRAM accesses as shown in Figure 2(e),(f) and (g). The input feature map and weights are first transferred into the accelerator from internal SRAM and then fed into a tensor sequencer which generates input tensor data to the SPE by reusing the data present in the SRAM. After the feature map and weight data are sequenced according to the layer properties, the data is fed into FIFO column buffer banks. To maximize parallel processing, another layer of memory blocks are introduced called FIFO buffer banks which contain the final tensor data that will be loaded and unloaded after the respective operation is computed. This approach is mainly useful for large input channel convolution layers in SCENIC so that the next channel will be readily sequenced and stored in the FIFO banks until the previous channel is computed.

3) Accelerator Controller: We introduce a FSM based controller that is configured as per the instructions received from the host CPU as shown in Figure 2(d). Once the instructions are decoded, various counters and registers are configured to control the FSM. FSM has nine states: one IDLE state for resetting various blocks that acts as a bridge between transition to next states, four states to load weights and IF data from SRAM to FIFO banks to RF, one state for computation, four states to unload OF data from RF to FIFO banks to SRAM. Functionality of each state in FSM is described in Table I.

TABLE I
DIFFERENT STATES OF CONTROLLER FSM

State	Description
S0	Idle
S1	Load weight from SRAM to PE
S2	Enqueue IF to FIFO banks from SRAM
S3	Dequeue IF from FIFO
S4	Write IF to PE RF
S5	Compute
S6	Read IF from PE RF
S7	Enqueue OF to FIFO banks from PE RF
S8	Dequeue OF from FIFO banks
S9	Read OF to SRAM

IV. DATA AND MODELING ANALYSIS

A. Dataset

The BraTS 2017 de-identified data set [14]–[16] created exclusively from low- and high-grade gliomas for brain tumor classification and/or segmentation. These multi-modal scans describe native T1-w, T2-w, T1-ce, and FLAIR volumes, with each scan of dimensions 240 × 240 × 155. These images were acquired using MR scanners from different vendors and with different field strengths (1.5T and 3T) and implementations of the imaging sequences.

In SCENIC, to classify the presence of tumors, we used a dataset [17] containing 1.5T MRI images from three classes of tumors: Gliomas, Meningioma and Pituitary, and a class with no tumor. In our model, the first three classes were combined to form a super-class, labelled "Tumor" and the last class was labelled "No tumor". The distribution of the above-mentioned datasets has been shown in Table II.

TABLE II
DATASET DISTRIBUTION

	Data division for classifying brain pathologies				**Data division for tumor detection**	
	FLAIR	**T1-w**	**T1-ce**	**T2-w**	**Tumor**	**No tumor**
Train Set	189	206	184	190	2050	1803
Test Set	40	40	40	40	100	100

B. Model Mapping based on Hardware Software Co-design

Initially, we embedded SCENIC directly onto our accelerator and observed that the translation of feature maps was inefficient and led to low SPE utilization as shown in the Table III. As with other previous accelerators with their own efficient mapping principles [11], our mapping is based on hardware-software co-design. The model is optimized to have OF dimensions symmetrical to the SPE array size ensuring the highest SPE utilization at any point of inference, increasing the overall throughput. Hence, all the OF dimensions of Spark modules were made multiples of SPE array size so that OF can be computed by symmetrically dividing it into multiple batches avoiding hardware padding for the leftover OF neurons. Here, we discuss the mapping of 3D and depth-wise separable convolutions which contribute to around 94% to total operations of SCENIC. The mapping principles of pooling and dense layers will be mostly based on these techniques itself.

TABLE III
SPE UTILIZATION OF CNN LAYERS OF SCENIC AND HARDWARE-OPTIMIZED SCENIC

	SCENIC	**SPE Utilization**	**Hardware-Optimized SCENIC**	**SPE Utilization**
Conv-1	238x238x32	75.11%	240x240x32	100%
Spark-1	238x238x64	75.11%	240x240x64	100%
Spark-2	119x119x128	87.12%	120x120x128	100%
Spark-3	59x59x256	87.12%	60x60x256	100%
Spark-4	29x29x512	87.12%	30x30x512	100%
Spark-5	14x14x512	87.12%	15x15x512	100%

During the computation of the convolution layer, each output feature map is allocated to a single SPE element of the spatial array till the computation is complete. A new input feature map is loaded into an SPE after every MAC operation and a new filter element is loaded after every output neuron computation is done. Using this kind of output stationary data-flow, a single output point is assigned inside one SPE element while IF neurons and weight elements are continuously updated. This decreases the high memory bandwidth of transferring the output OF. The tensor division and mapping flow in the accelerator is performed as shown in Figure 3. Each input feature map is partitioned into multiple blocks using SPE array dimensions as the divisor. As discussed earlier, the output feature map dimensions of each convolution layer of Spark modules can be symmetrically partitioned and mapped on the array. Based on the convolution properties such as dimensions, stride and padding, tensors are sequenced and loaded to IF-FIFO buffers banks and unloaded to OF-FIFO buffer banks once the computation is done.

V. RESULTS

A. Simulation Results

1) Advanced performance in SCENIC by inclusion of DWT and Spark: There are various metrics, we have used to test

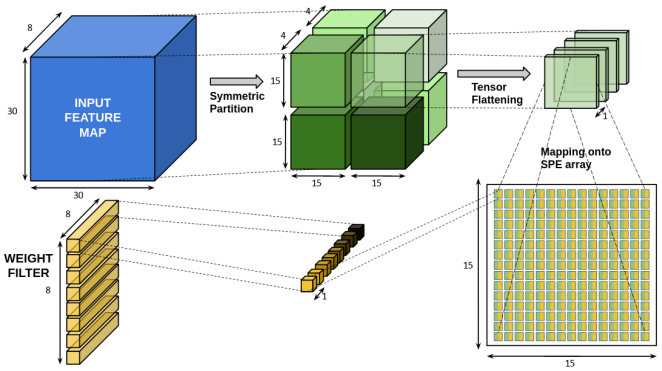

Fig. 3. Tensor mapping flow of SCENIC layers

the performance of SCENIC: Accuracy, AUC score, Specificity, Recall, Precision, and F1-Score. Table IV shows the overall results we achieved with our experimental models in Python 3.7, and summarizes SCENIC performance in terms of accuracy and size.

TABLE IV
RESULTS OBTAINED ON DIFFERENT STAGES OF SCENIC

	Accuracy	Average Runtime (s)	Memory Size (MB)
Simple CNN	84.3%	0.19	55
CNN + DWT	94.37%	0.09	6
Spark model	95.74%	0.052	2
SCENIC	**99.62%**	**0.065**	**0.265**

Table V shows the results obtained by our two software model components along with those obtained by the past models, tested on our dataset. An example of the actual classification by anatomical features is shown in Figure 5. SCENIC is able to detect and categorize based on anatomical features. Figure 5 provides a snapshot of training, tested categorized, and alternately categorized data. Thus SCENIC recognizes when feature details in a T1ce image reflect similar features as T1-w or when a FLAIR manually assigned image is classified as T2-w. In future iterations, we will further test and train SCENIC to maximize its discerning potential in more challenging comparisons of HGG and metastasized tumors to benefit neurosurgery. This current platform allows us to move forward with clinical samples in collaborative efforts.

We compared SCENIC and its design iterations with other classical pre-trained models evaluating the same datasets. SCENIC and Xception are the best performers, with SCENIC achieving the highest scoring. Figure 4 demonstrates the ability of SCENIC to discern features with high accuracy.

TABLE V
A COMPARISON OF MODEL CHARACTERISTICS WITH EXISTING MODELS

Model Name	Accuracy	F1-Score	AUC-Score
Simple CNN architecture	84.3%	0.814	0.887
ResNet50 [5]	90.00%	0.942	0.990
Inception-V3 [6]	92.13%	0.881	0.994
CNN + DWT	94.37%	0.907	0.934
VGG-16 [4]	95.20%	0.932	0.9711
Spark model	95.74%	0.965	0.988
Xception [7]	98.31%	0.913	0.993
SCENIC for detecting Brain Tumors in MR Images	98.3%	0.959	0.961
SCENIC for detecting Brain Pathology in MR Images	99.62%	0.981	0.998

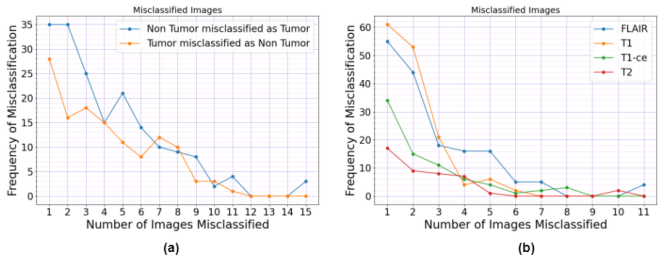

Fig. 4. The trend of different modalities over different sets of data over a 100 runs- (a) For detecting the Presence of Tumor in the MR Images (b) For detecting the Brain Pathology in the MR Images

2) SCENIC performs strongly on multiple distinct datasets: To ensure SCENIC's ability to perform with high accuracy on a distinct MRI dataset we performed additional analysis. Without changes to SCENIC architecture, we achieve reduced but still high accuracy, of 96.7%, in the RSNA-MICCAI Brain Tumor Radiogenomic Classification Dataset [18].

B. Hardware Synthesis Results

SCENIC Hardware Accelerator is written in System Verilog and simulated in ModelSim Simulator by simulation scripts to test both SCENIC for detection of Tumors in MR Images and SCENIC for differentiating anatomical structures. SCENIC model parameters and input MR images after pre-processing by Python scripts to a binary format are processed by simulator test-benches. The simulation results are then verified with golden-reference output generated by a python-based custom simulator. The functionally verified RTL is then mapped and synthesized using Synopsys Design Compiler on 45 nm OpenNangate technology. The power consumption is derived using Synopsys Power Compiler.

The operating frequency of 1041 MHz is met for SCENIC hardware Accelerator and 671 MHz for DWT Pre-processing Unit. The number of gates in the synthesized hardware is about 579,949. Table VI shows the operational parameters of SCENIC hardware accelerator versus other state-of-the-art accelerators. SCENIC hardware accelerator also achieved higher throughput of CNN with the average inference runtime per image of 6.8 ms (excluding DWT pre-processing time) and around 145 frames per second.

TABLE VI
OPERATIONAL PARAMETERS OF SCENIC HARDWARE ACCELERATOR AND OTHER STATE-OF-THE-ART ACCELERATORS [19]

CNN Accelerator	Process (nm)	Area (mm^2)	Power (W)
DianNao	65	0.485	3.02
DaDianNao	28	15.970	67.70
ShiDianNao	65	0.320	4.86
PuDianNao	65	0.596	3.51
SCENIC CNN Accelerator and DWT Unit	**45**	**0.431**	**0.360**

VI. DISCUSSION AND CONCLUSION

There is a critical need to be able to optimize image information from multiple MRI modalities to aid diagnosis and treatment strategies for brain pathologies. SCENIC is able to compress parameter data while excelling at feature classification, enabling implementation into low powered embedded devices and flexible communication platforms. The different MR image modality groups evaluated in SCENIC reflect features of tissues, CSF, and pathology. Briefly, in T1-w MRI fatty tissues are brighter in contrast to water, whereas in T2-w MRI this contrast is opposite and helpful

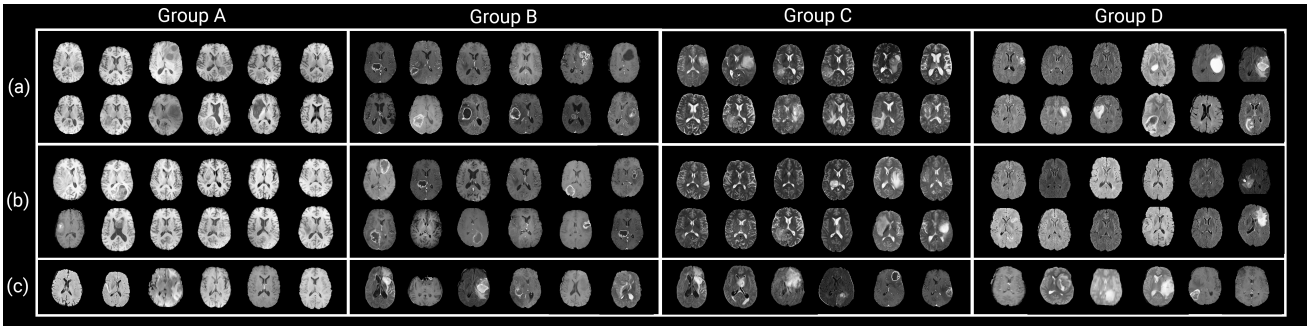

Fig. 5. Anatomical Feature Classification of MRI images into four different groups based on different MRI modalities. Row (a) shows the sample type of images from each group SCENIC was trained on. Row (b) and (c) constitute of the test images classified by SCENIC. Row (c) represents images alternatively classified when compared to their manually assigned classes, whereas Row (b) represents images which are in accordance with the manually assigned classes

for tracking edema in damaged tissues. The contrast agent enabled T1-ce provides enhanced visualization of anatomy, including pathologies affecting the blood-brain-barrier and inflammation while FLAIR, highlights endamatous tissues and facilitates evaluating demyelinating pathologies. SCENIC performs with 99.62% accuracy with low power consumption versus manual classification on the BRATS dataset [14]-[16], which consisted of exclusively horizontal section MR images. Importantly, those images that are classified differently by SCENIC compared to manual classification are by visual inspection 'reasonable' categorizations. The value of interrogating MRI scans from multiple weighted parameters is to train SCENIC to recognize anatomical and pathology variations that can help to discriminate pathology types, tumor origins, and variations from and within normal tissues. The accuracy for classification of tumour types in MR images of the mid-sagittal section [17] was also explored and is 93% in both SCENIC and VGG-16 [4], where the MR images consist of both the brain tissues and the surrounding sternocleidomastoid muscle. Segmentation will allow us to isolate the brain in such MR images and enable feature extraction more specific to the brain, positively impacting the learning efficiency and accuracy. Continued analysis of larger datasets with SCENIC, as well as those that include different subtypes of pathologies such as HGG tumors and metastasized tumors are underway and will enable us to continually refine SCENIC.

Advancements in DL models applied to feature extraction will have the ability to normalise variations, detect subtle features and integrate rapidly with higher power Tesla MRI technologies. SCENIC has robust energy efficiency versus existing DL models for similar applications, due to its more efficient utilisation of memory. SCENIC can process a greater amount of images in a given time interval via multithreading, demonstrating a higher classification rate. SCENIC Hardware Accelerator's low power and area optimized features makes it possible to run CNNs from smart devices. In consideration of future iterations of SCENIC, the use of DWT for noise reduction may not capture sufficient feature information. Hence MR images where even higher feature capture is needed, such as staging disease progression or distinguishing features within HGGs or between HGGs and metastasized tumors may require advancements such as with the memory interface. Iterative designs will continue to balance optimal compression parameters, scaling and single or multiple image access needs with maintaining low latencies in operation, particularly for smart devices. In conclusion, SCENIC is a powerful new tool for advancing discerning MRI analysis of brain pathologies.

VII. Acknowledgements

The views, opinions, assumptions and conclusions or any other information set out in this article are solely those of the authors and not necessarily of the authors' employers, organizations, committees or other groups or individuals.

References

[1] Porter KR et. al Prevalence estimates for primary brain tumors in the United States by age, gender, behavior, and histology. Neuro-Oncology 12(6):520-527, 2010
[2] Morgenstern, Jason D. and Michael D. Staudt. "Brain metastasis masquerading as glioblastoma multiforme and lymphoma." (2016).
[3] Patchell, R A et al. "A randomized trial of surgery in the treatment of single metastases to the brain." The New England journal of medicine vol. 322,8 (1990): 494-500
[4] Simonyan, Karen & Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.
[5] He, Kaiming & Zhang, et al. (2016). Deep Residual Learning for Image Recognition. 770-778.
[6] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1-9
[7] Chollet, François. "Xception: Deep Learning with Depthwise Separable Convolutions." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 1800-1807.
[8] Ghosh, S., Pal, A., Jaiswal, S. et al. SegFast-V2: Semantic image segmentation with less parameters in deep learning for autonomous driving. Int. J. Mach. Learn. & Cyber. 10, 3145–3154 (2019).
[9] Iandola FN, Han S, et al. (2016) Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. arXiv preprint arXiv:1602.07360
[10] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," IEEE Transactions on Image Processing, vol. 9, no. 9, pp. 1532–1546, 2000.
[11] Sze, Vivienne, et al. "Efficient processing of deep neural networks: A tutorial and survey." Proceedings of the IEEE 105.12 (2017): 2295-2329.
[12] A. Raha et al., "Design Considerations for Edge Neural Network Accelerators: An Industry Perspective," 2021 34th International Conference on VLSI Design and 2021 20th International Conference on Embedded Systems (VLSID), 2021, pp. 328-333.
[13] S. K. Madishetty, A. Madanayake, et al. "VLSI Architectures for the 4-Tap and 6-Tap 2-D Daubechies Wavelet Filters Using Algebraic Integers," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 60, no. 6, pp. 1455-1468, June 2013
[14] B. H. Menze, A. Jakab, et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)", IEEE Transactions on Medical Imaging 34(10), 1993-2024 (2015)
[15] S. Bakas, H. Akbari, et al., "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features", Nature Scientific Data, 4:170117 (2017)
[16] S. Bakas, M. Reyes, et al., "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge", arXiv preprint arXiv:1811.02629 (2018)
[17] Sartaj Bhuvaji, Ankita Kadam, et al "Brain Tumor Classification (MRI)." Kaggle, 2020.
[18] U.Baid, et al., "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification", arXiv:2107.02314, 2021.
[19] Chen, Yiran, et al. "A survey of accelerator architectures for deep neural networks." Engineering 6.3 (2020): 264-274.