Reliable and Energy-efficient Diabetic Retinopathy Screening using Memristor-based Neural Networks

Diware, Sumit; Chilakala, Koteswararao; Joshi, Rajiv V.; Hamdioui, Said; Bishnoi, Rajendra

DOI
10.1109/ACCESS.2024.3383014

Publication date
2024

Document Version
Final published version

Published in
IEEE Access

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 March 2024, accepted 25 March 2024, date of publication 29 March 2024, date of current version 4 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3383014

RESEARCH ARTICLE

Reliable and Energy-Efficient Diabetic Retinopathy Screening Using Memristor-Based Neural Networks

SUMIT DIWARE[1], KOTESWARARAO CHILAKALA[2], RAJIV V. JOSHI[3], (Life Fellow, IEEE), SAID HAMDIOUI[1], (Senior Member, IEEE), AND RAJENDRA BISHNOI[1]

[1]Computer Engineering Laboratory, Delft University of Technology, 2628 CD Delft, The Netherlands
[2]Capgemini Engineering, 5652 AA Eindhoven, The Netherlands
[3]IBM Thomas J. Watson Research Centre, Yorktown Heights, NY 10598, USA

Corresponding author: Sumit Diware (S.S.Diware@tudelft.nl)

This work was supported in part by the European Union, Distributed Artificial Intelligent Systems (DAIS), under Grant 101007273; in part by CONVOLVE under Grant 101070374; and in part by NEUROKIT2E under Grant 101112268.

ABSTRACT Diabetic retinopathy (DR) is a leading cause of permanent vision loss worldwide. It refers to irreversible retinal damage caused due to elevated glucose levels and blood pressure. Regular screening for DR can facilitate its early detection and timely treatment. Neural network-based DR classifiers can be leveraged to achieve such screening in a convenient and automated manner. However, these classifiers suffer from reliability issue where they exhibit strong performance during development but degraded performance after deployment. Moreover, they do not provide supplementary information about the prediction outcome, which severely limits their widespread adoption. Furthermore, energy-efficient deployment of these classifiers on edge devices remains unaddressed, which is crucial to enhance their global accessibility. In this paper, we present a reliable and energy-efficient hardware for DR detection, suitable for deployment on edge devices. We first develop a DR classification model using custom training data that incorporates diverse image quality and image sources along with improved class balance. This enables our model to effectively handle both on-field variations in retinal images and minority DR classes, enhancing its post-deployment reliability. We then propose a pseudo-binary classification scheme to further improve the model performance and provide supplementary information about the model prediction. Additionally, we present an energy-efficient hardware design for our model using memristor-based computation-in-memory, to facilitate its deployment on edge devices. Our proposed approach achieves reliable DR classification with three orders of magnitude reduction in energy consumption over state-of-the-art hardware platforms.

INDEX TERMS Diabetic retinopathy, neural networks, computation-in-memory, resistive random access memory, RRAM, memristor, edge computing, reliability.

I. INTRODUCTION

Diabetic retinopathy (DR) refers to a condition where elevated glucose levels and blood pressure lead to irreversible retinal damage. It is a leading cause of permanent vision impairment across the globe, and the number of affected people is expected to reach 70 million by 2045 [1]. Moreover, every diabetic person is susceptible to the development of DR [2]. As the vision loss caused by DR is irreversible, detecting it at an early stage is crucial for timely treatment to prevent further retinal damage. Regular screening for DR is essential for such early detection. Recent advancements in artificial intelligence have paved the way for developing automated systems to provide fast, efficient, and convenient

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues.

DR screening. These systems employ neural network-based DR classifiers to categorize retinal images into distinct screening classes, capitalizing on the inherent proficiency of neural networks in classification tasks.

The publicly available DR datasets exhibit inherent image inconsistencies to pose a tougher classification challenge than private ones, resulting in more robust and adaptive models. Moreover, their wider accessibility is valuable for driving further innovation in automated DR classification. Hence, we focus on DR classification literature based on publicly available datasets. Such works are susceptible to reliability issues, where the model performs well during development but exhibits poor performance upon deployment. This can arise due to several factors such as small training data size [2], [3], [4], [5], [6], [7], [8], absence of external test data [9], [10], [11], [12], [13], and lack of diversity in training data [14], [15]. Furthermore, the inherent class imbalance in public datasets can bias the model performance towards majority classes. This can hinder the identification of minority DR classes (indicating retinal damage), further aggravating the reliability concern. Additionally, supplementary information about model prediction is crucial for widespread adoption of automated DR classification [13], [16]. For instance, when a DR model assists human specialists in double reading [17], supplementary information bolsters specialist's confidence when human diagnosis matches the model prediction, and helps resolve conflicts when these two differ. However, none of the aforementioned works provide supplementary information about model prediction. Lastly, deploying automated DR classification on portable edge devices can address the global scarcity of DR screening facilities [18], [19]. This requires energy-efficient hardware design of DR classifiers to achieve uninterrupted operation despite limited energy resources, enabling large-scale screening programs even in remote regions. However, the aforementioned literature only focuses on software model development while neglecting hardware design considerations. Hence, there exists a pressing need for hardware solution that facilitates reliable and energy-efficient DR screening at the edge.

In this paper, we present a reliable and energy-efficient DR screening hardware targeting deployment on edge devices. We first develop a reliable DR classification model via training on a newly created custom dataset. This dataset encompasses image quality inconsistencies, diverse image sources, and reduced class imbalance. This enables our trained model to effectively handle real-world retinal images and perform well on minority classes, ensuring post-deployment reliability. Furthermore, we introduce a pseudo-binary classification scheme that internally uses multiclass classification to achieve binary screening. This enhances our model's classification performance and also provides supplementary information to aid its wider adoption. We then present energy-efficient hardware design of our model based on computation-in-memory (CIM) paradigm. It uses emerging memory devices known as memristors, to perform computations directly within the memory. This eliminates the data transfer bottleneck and provides superior energy efficiency suitable for edge device deployment. The key contributions of this paper are:

- We develop a reliable DR classification model by using inconsistent quality images collected from diverse sources and addressing the class imbalance problem.
- We propose a pseudo-binary classification scheme to improve the classification performance and provide a more informative classification output.
- We present an energy-efficient hardware design for our DR classification model using memristor-based CIM to facilitate its deployment on edge devices.

Simulation results show that we achieve reliable DR classification while consuming three orders of magnitude less energy compared to the state-of-the-art hardware platforms.

The rest of the paper is structured as follows: Section II presents the basics of diabetic retinopathy and memristor-based computation-in-memory, while Section III provides a review of related existing literature. The proposed methodology is described in Section IV, followed by simulation setup details in Section V and simulation results in Section VI. Finally, Section VII concludes the paper.

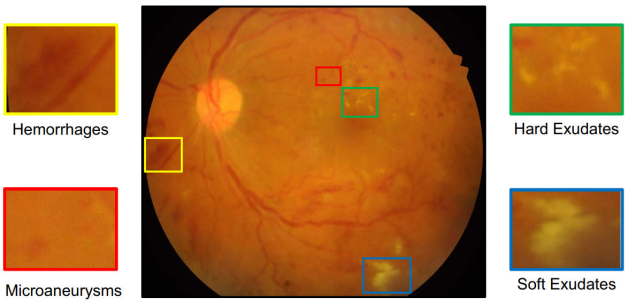


FIGURE 1. Retinal image annotated with the four most common lesions: microaneurysms, hemorrhages, hard exudates, and soft exudates [20].

II. BACKGROUND

A. DIABETIC RETINOPATHY (DR)

1) BASICS

Diabetic retinopathy is an irreversible condition arising from elevated glucose levels and hypertension. It damages blood vessels in the retina and can potentially cause permanent vision impairment. As per International standards [21], there are five severity levels (classes) of DR as follows:

- No DR (DR-0)
- Mild non-proliferative (DR-1)
- Moderate non-proliferative (DR-2)
- Severe non-proliferative (DR-3)
- Proliferative (DR-4)

These classes are distinguished from each other based on certain features present in the retina, known as lesions. Figure 1 depicts the four most common lesions: microaneurysms, hemorrhages, hard exudates, and soft exudates. They can be described as follows [2], [20]:

TABLE 1. Lesion-based diagnosis of DR classes.

DR Severity Level	DR Class	Lesion-based Diagnosis
No DR	DR-0	No lesions
Mild non-proliferative	DR-1	Only MA present
Moderate non-proliferative	DR-2	MA and other lesions present Less prominence than DR-3
Severe non-proliferative	DR-3	At least one of the following present: HM (>20 in each quadrant), Blood spillage (>2 quadrants), No indicators of DR-4
Proliferative	DR-4	Vitreous/preretinal HM and/or NV

- Microaneurysms (MA): These are the earliest visible signs of retinal damage. They manifest as tiny red dots arising due to capillary dilation.
- Hemorrhages (HM): These are red spots with irregular margins and/or uneven density. They are bigger than MA and occur due to leakage of weak capillaries.
- Hard exudates (HE): These are yellow-white deposits in outer layers of retina caused by leakage of plasma.
- Soft exudates (SE): These are greyish oval or round-shaped patches arising due to the swelling of the nerve fiber. They are also called cotton wool spots.
- Neovascularization (NV): It refers to the abnormal growth of new blood vessels on the inner surface of the retina. Such blood vessels often bleed into the vitreous cavity and lead to obscured vision.

Table 1 shows how various DR classes are diagnosed based on the composition of these lesions.

2) DETECTION

Conventional DR detection typically begins with pre-capture medical procedures on patients to enlarge their pupils and facilitate better coverage of the retinal area during image capture. Skilled operators then employ specialized fundus cameras and meticulously adjust settings such as focus, exposure, alignment, etc., to capture high-quality retinal images. Subsequently, the severity of DR is evaluated through visual inspection of various lesions within the captured retinal image. The advancements in artificial intelligence have opened avenues to employ automated systems to achieve DR identification from retinal images. They leverage neural networks, which inherently excel at extracting crucial lesion information from retinal images and autonomously categorize them into distinct DR classes. Thus, neural network-based DR classification systems offer an effective and efficient approach to DR detection.

B. MEMRISTOR-BASED COMPUTATION-IN-MEMORY (CIM) FOR NEURAL NETWORKS

1) CIM PARADIGM

Traditionally, neural networks are implemented on hardware platforms like CPUs [22], GPUs [23], and specialized ASICs like TPUs [24], which adhere to the von-Neumann architecture and employ CMOS technology. The von Neumann architecture involves separated memory and computation units, leading to numerous data transfers for executing vector-matrix multiplication (VMM) in neural networks. This results in high energy consumption as VMM operations constitute a large portion of neural network computations [25]. Furthermore, CMOS technology suffers from challenges such as excessive sub-threshold leakage and scalability issues [26]. *Computation-in-Memory* (CIM) paradigm leverages emerging memory technologies, such as resistive random access memories (RRAMs), also called memristors [27], [28], [29], to address these limitations. It performs computations directly within the memory, eliminating the data transfer bottleneck. Furthermore, memristors are non-volatile, offer high scalability, and are compact in size. Hence, CIM emerges as a highly promising alternative to conventional hardware for neural network implementation [30], [31], [32].

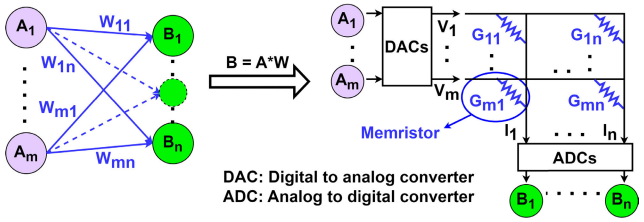


FIGURE 2. Mapping of neural network layers to CIM architecture.

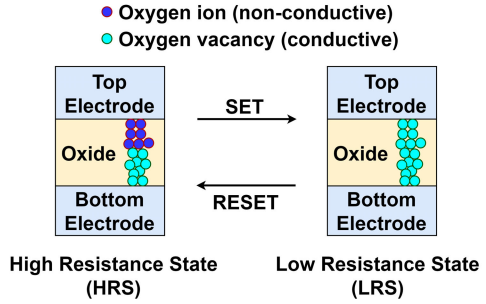


FIGURE 3. Memristor device technology.

2) CIM ARCHITECTURE

The mapping of VMM operation between two layers of a neural network onto CIM hardware is illustrated in Figure 2. It employs a mesh-like structure of memristors known as the crossbar. The crossbar carries out computations in the analog domain and communicates with other digital system components through data converters like digital-to-analog converters (DACs) and analog-to-digital converters (ADCs). Weights are translated into memristor conductances (G's) within the crossbar, while inputs are applied as voltages (V's) using DACs. The resulting current through each conductance is equivalent to the element-wise multiplication of the voltage and conductance. The accumulation of currents within each column yields the accumulation of element-wise products as currents (I's). Thus, CIM executes a multiply-and-accumulate operation in the analog domain for each column. The collective multiply-and-accumulate operations across all the columns constitute

TABLE 2. Summary of the related works on neural network-based DR classification.

Paper	Large Training Data	Diverse Training Data	External Testing	Supplementary Information	Hardware Design Consideration
Feng et al. [3]	×	×	×	×	×
Khan et al. [4]	×	×	×	×	×
Wong et al. [5]	×	×	×	×	×
Islam et al. [6]	×	×	×	×	×
Mohammadi et al. [7]	×	×	×	×	×
Zhou et al. [2]	×	×	×	×	×
Sikder et al. [8]	×	×	×	×	×
Poranki et al. [9]	✓	×	×	×	×
Sadeghzadeh et al. [10]	✓	×	×	×	×
Sait et al. [11]	✓	×	×	×	×
Alyoubi et al. [12]	✓	×	×	×	×
Li et al. [13]	✓	×	×	×	×
Korot et al. [14]	✓	×	✓	×	×
Ludwig et al. [15]	✓	×	✓	×	×
This Work	✓	✓	✓	✓	✓

a VMM operation. Subsequently, ADCs convert the column currents into digital outputs, and the digitized VMM result is transmitted to other parts of the system for further processing.

3) MEMRISTOR DEVICE TECHNOLOGY

Memristor device comprises of an oxide material sandwiched between two metal electrodes [33], as shown in Figure 3. It possesses two distinct states: a high-resistance state (HRS) and a low-resistance state (LRS), which serve as data storage equivalent to 0 and 1. The transition from HRS to LRS is referred to as "SET", while the reverse process of transitioning from LRS to HRS is termed "RESET". In the SET process, applying a set voltage (V_{SET}) to a memristor in HRS creates a conductive path known as a filament. This enhances the oxide layer's conductivity, leading to a change in state from HRS to LRS. Conversely, applying a reset voltage (V_{RESET}) to a memristor in LRS results in the rupture of the conductive filament. This reduces the oxide layer's conductivity and changes the state from LRS to HRS. To read data from a memristor, its resistance state is detected by applying a very low voltage, denoted as V_{READ} (where $V_{READ} \ll |V_{SET}|$ and $V_{READ} \ll |V_{RESET}|$), and measuring the resulting output current. Moreover, a single memristor can exhibit multiple conductance states by controlling the extent of filament creation or rupture, known as multi-level cell (MLC) operation [34], [35].

III. RELATED WORK

The publicly available DR datasets play a pivotal role in the advancement of neural network-based DR classification. They contain real-world image inconsistencies arising from varying equipment quality, operator expertise etc. This makes their classification more challenging than private datasets obtained under controlled conditions, leading to robust and adaptive models. Additionally, the widespread accessibility of these datasets to research community can accelerate the development of innovative solutions. Hence, we focus on DR classification literature based on publicly available datasets. Such works suffer from several challenges like model reliability issue, lack of supplementary information and scarcity of DR screening facilities. These challenges are discussed in detail next, with a summary provided in Table 2.

Model reliability issue refers to the situation where the model performs well during development phase but struggles after deployment. Existing works that are susceptible to reliability issue can be organized into three categories. The first category involves works using small datasets for model development. This hinders model's generalization ability leading to reliability concern. Datasets like APTOS [36], Messidor-2 [37], [38], [39], HRF [40], small-scale merger of EyePACS with Messidor-2 [41], and FGADR [2] are small-sized. Several studies exemplify this issue. For instance, Feng et al. [3] present a cascaded convolutional and graph neural networks using APTOS and Messidor-2 datasets. Khan et al. [4] perform transfer learning using HRF and APTOS datasets. Wong et al. [5] propose transfer learning with optimized feature weights on APTOS dataset and small-scale merger of EyePACS and Messidor-2 datasets. Islam et al. [6] evaluate supervised contrastive learning on APTOS and Messidor-2 datasets. Mohammadi et al. [7] use Messidor-2 dataset to explore automated machine learning platforms. Zhou et al. [2] introduce FGADR dataset and its benchmarking. Sikder et al. [8] present decision tree-based ensemble learning using APTOS dataset.

The second category involves works that use large training datasets like EyePACS [42] or DDR [13] but do not use an

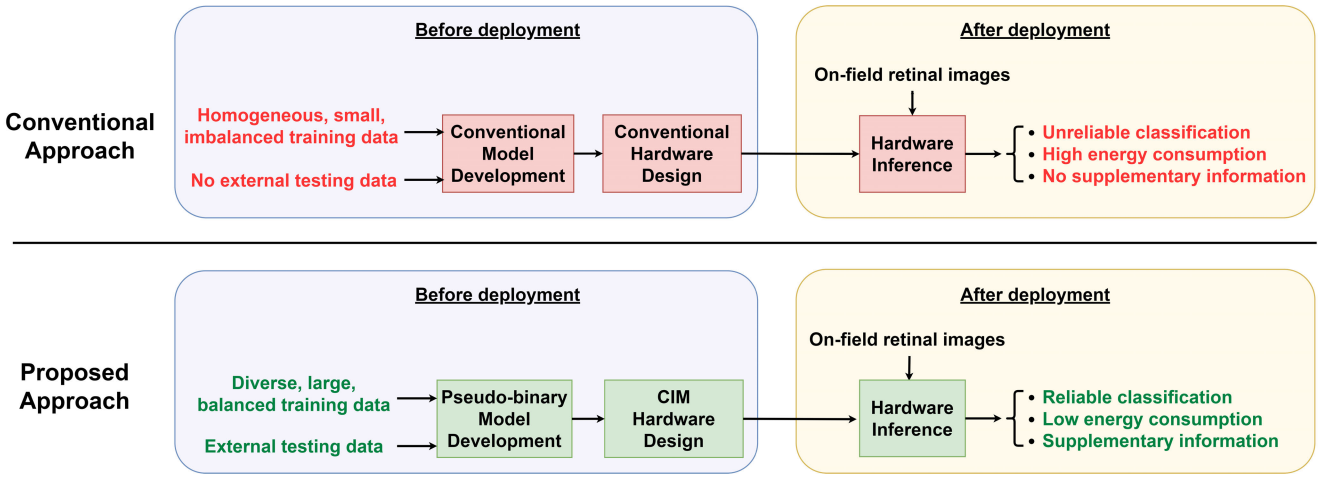


FIGURE 4. Overview of the conventional and proposed approaches for neural network-based DR screening.

external test data. External test data with a different origin is necessary to expose the model to truly unseen characteristics, mimicking real-world deployment. The internal test data is not enough as it shares common origin and characteristics with training/validation data. As a result, generalization ability of the model is not properly assessed leading to reliability issue. Many works suffer from this challenge. Poranki et al. [9] use EyePACS dataset to develop hybrid learning model combining graph learning and deep learning. Sadeghzadeh et al. [10] propose a hybrid of convolutional and transformer network evaluated on EyePACS dataset. Sait [11] train MobileNet V3-based network on EyePACS dataset. Alyoubi et al. [12] use DDR dataset for convolutional neural network-based classification. Li et al. [13] present DDR dataset and its benchmarking.

The third category includes works that incorporate both large training dataset and external test dataset but lack diversity in training data. This affects the reliability by hindering the model's adaptability to post-deployment data. For instance, Korot et al. [14] use EyePACS datasets for model development and IDRiD dataset [43] for external validation. However, relying on a single training dataset restricts the data diversity. Similarly, Ludwig et al. [15] use EyePACS and APTOS datasets for model development while employing Messidor-2 dataset for external validation. However, a significantly larger number of EyePACS images compared to APTOS ones dominate the training data, thus restricting its diversity. Moreover, the inherent class imbalance in all public datasets can lead to the model excelling at handling majority classes while struggling with minority DR classes. This can aggravate the reliability issue by reducing model's effectiveness in identifying instances of retinal damage.

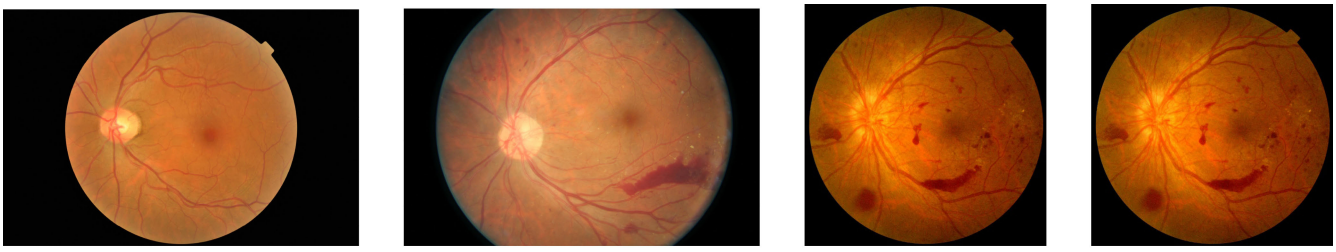
Double reading is a process where two specialists (readers) independently analyze the same recording (e.g. retinal image) and exchange their insights to provide a unified diagnosis [44], [45]. A neural network model can streamline double reading-based DR screening by replacing one of the human specialists [17]. However, the remaining human specialist has no insight into the rationale behind model's diagnosis due to lack of supplementary information. This black-box nature of model predictions hinders collaborative reasoning and building trust with the specialist, reducing effectiveness of its assistance [16]. Moreover, supplementary information is also crucial for instilling confidence in patients [13]. Despite such importance, none of the aforementioned works incorporate supplementary information in their output.

There exists a global scarcity of DR screening facilities due shortage of specialists, inadequate medical infrastructure, and economic constraints [18]. Deploying automated DR classification on portable edge devices offers a promising solution to this challenge. These devices can integrate portable imaging technology [19] with a dedicated DR classification chip. This chip must exhibit energy-efficiency to ensure prolonged operation, even with limited battery resources and unstable power grids. Sharing such edge device among multiple regions with similar healthcare resource limitations can significantly expand the screening outreach and serve a larger population over time. Hence, there exists a pressing need for a hardware solution that can facilitate reliable and energy-efficient DR screening at the edge.

IV. PROPOSED METHODOLOGY

A. OVERVIEW

An overview of both conventional and proposed approaches for developing neural network-based DR screening solutions is shown in Figure 4. They both involve two phases: 1) pre-deployment phase, where the model is trained and hardware is designed for the trained model, and 2) post-deployment phase, where the hardware performs inference using on-field images. Models developed using conventional approach are susceptible to reliability issue, where they perform well during development but fail after deployment. This is a consequence of model's poor generalization ability arising from small-sized, non-diverse and imbalanced training data, coupled with absence of external test data. Moreover, they



(a) DR-0 (EyePACS) (b) DR-4 (EyePACS) (c) DR-4 (DDR) (d) DR-4 (APTOS)

FIGURE 5. Sample images across various public datasets (EyePACS [42], DDR [13] and APTOS [36]) that are merged for use in this work.

TABLE 3. Overview of the original public datasets and newly proposed datasets. Imbalance is the ratio of the sizes of the largest and smallest class.

	Original Datasets			Proposed Datasets		
	APTOS [36]	**DDR** [13]	**EyePACS** [42]	**Small (S)**: APTOS	**Medium (M)**: APTOS & DDR	**Large (L)**: APTOS, DDR & EyePACS
DR-0	1805	6266	65343	1798	3000	10000
DR-1	370	630	6205	365	991	7180
DR-2	999	4477	13153	991	3000	10000
DR-3	193	236	2087	188	424	2504
DR-4	295	913	1914	292	1204	3117
Total	3662	12522	88702	3634	8619	32801
Imbalance	9.4×	26.5×	34.1×	9.6×	7.1×	4.0×

do not provide supplementary information about the model prediction. This severely limits their widespread adoption by both specialists and patients. Additionally, there has been almost no effort directed towards energy efficient hardware design for DR classification models targeting edge device deployment, which is critical to improve their global accessibility.

Our proposed approach overcomes all of these challenges. We first create a large, diverse and balanced custom dataset by combining data from multiple sources and taking measures to reduce class imbalance. We then train our DR classification model with this dataset and also assess the model reliability with external test data. Moreover, we propose a pseudo-binary classification scheme that improves the model performance and also provides supplementary information to facilitate its widespread adoption. Furthermore, we present energy-efficient hardware design for our model using memristor-based CIM, to facilitate its deployment on edge devices for improved accessibility. Thus, we provide a solution that offers reliable DR classification, supplementary information and energy efficient edge deployment. We will now delve into the details of our approach in the next subsections.

B. MODEL DEVELOPMENT

1) DATASET CREATION

The DR classification models often encounter inconsistent quality retinal images post-deployment. This arises due to various factors such as improper exposure, misalignment, incomplete retinal coverage etc. Furthermore, the distribution of post-deployment data can diverge substantially from the data used during model development. Hence, the model must be trained using data that encompasses these inconsistencies and reflects diversity of post-deployment data to ensure reliability. We build such a comprehensive training dataset by leveraging the following publicly available datasets, whose samples are shown in Figure 5:

- EyePACS dataset [42]: It is provided by EyePACS Inc. for DR detection competition sponsored by California Healthcare Foundation in 2015. It contains 88,702 images collected from different parts of the USA.
- DDR dataset [13]: It contains 13,673 images collected across 147 hospitals in China from 2016 to 2018. The dataset actually has 12,522 usable images as 1,151 images are deemed ungradable.
- APTOS dataset [36]: It is a part of DR detection competition organized by Asia Pacific Tele-Ophthalmology Society in 2019. It contains 3662 retinal images provided by Aravind Eye Hospital in India.

After acquiring these datasets, we filter out corrupt images and merge them in varying proportions to create three merged datasets. We then undersample the majority classes in each merged dataset to limit class imbalance to 10× or less. This is because 10× or less imbalance suffices for neural networks perform well on minority classes [46], and achieving perfect class balance is impractical due to huge number of healthy retina images. As a result, we end up with three new proposed datasets: Small (S), Medium (M), and Large (L).

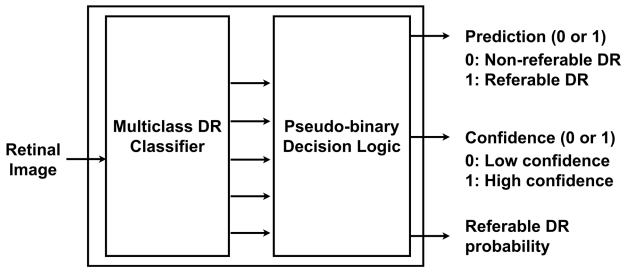


FIGURE 6. Pseudo-binary classification concept.

The classwise distributions of the original and new datasets is shown in Table 3. It can be seen that the new M and L datasets exhibit better class balance than original EyePACS and DDR datasets. Moreover, merging enhances the data diversity in M and L datasets compared to EyePACS and DDR datasets. As a result, models developed using M and L datasets can potentially exhibit better reliability in handling both on-field image inconsistencies and minority DR classes. Furthermore, the three datasets can be used to obtain crucial insights into how dataset size influences classification performance.

2) PSEUDO-BINARY CLASSIFICATION

The recommended management guidelines for various DR classes are as follows [47]:

- Annual screening for DR-0 or DR-1.
- A follow-up every six months for DR-2.
- Referral to an ophthalmologist for DR-3 or DR-4.

Thus, the recommended DR management approach shifts from annual screening to more frequent monitoring as the severity reaches DR-2. Consequently, grouping the five DR classes into the following two categories suffices for screening [15]: non-referable DR (consisting of DR-0 and DR-1) and referable DR (consisting of DR-2, DR-3, and DR-4). Thus, DR screening becomes a binary classification task involving referable DR and non-referable DR as its two classes.

A straightforward approach to binary screening involves relabeling the five original classes into these two broad categories and developing a binary classification model. However, this leads to a hard decision between the two categories which is more susceptible to misclassifications. It also weakens the interpretability by hindering the derivation of supplementary information. To alleviate this problem, we introduce an approach called pseudo-binary classification. It internally employs a multiclass DR classifier and uses additional decision-making logic to ultimately produce a binary classification outcome, as shown in Figure 6. It capitalizes on cumulative probabilities within non-referable (0) and referable (1) categories instead of hinging on a single maximum probability for decision-making, reducing susceptibility to misclassifications. Moreover, it presents the outcome as a tuple containing prediction, confidence level, and referable DR probability, providing better interpretability.

Algorithm 1 Pseudo-Binary Classification Algorithm

input : Confidence threshold C_{th}, retinal image I
output: Prediction tuple P

```
softmax ← multiclass_inference(I);
non_ref_score ← softmax(DR-0) + softmax(DR-1);
ref_score ← softmax(DR-2) + softmax(DR-3) +
  softmax(DR-4);
Δ ← ref_score - non_ref_score;
if Δ > 0 then
    prediction ← 1;
    if Δ > C_th then
        confidence ← H;
    else
        confidence ← L;
else
    prediction ← 0;
    if |Δ| > C_th then
        confidence ← H;
    else
        confidence ← L;
P ← (prediction, confidence, ref_score);
return P;
```

TABLE 4. Interpretation of various prediction tuples of pseudo-binary classification. Here, S indicates the probability of referable DR in all cases.

Prediction Tuple	Interpretation
(0, H, S)	Healthy (no DR).
(0, L, S)	DR developing, checkup recommended.
(1, L, S)	DR found, seek medical help soon.
(1, H, S)	DR found, seek medical help immediately.

Algorithm 1 describes the pseudo-binary classification process. It begins with multiclass classification to obtain prediction probabilities for the five original DR classes. Subsequently, it calculates a score for the non-referable (0) class by adding the probabilities of DR-0 and DR-1. Similarly, a score for the referable (1) class is computed by adding the probabilities of DR-2, DR-3, and DR-4. The broad class (referable 0 or non-referable 1) with the higher score is selected as the prediction value. Furthermore, if the scores differ by more than a predefined confidence threshold, we indicate high confidence ('H'); otherwise low confidence ('L'). Additionally, the referable class score indicates the probability of referable DR. For example, consider a scenario with confidence threshold 0.25 and softmax probabilities as [0.20 (DR-0), 0.33 (DR-1), 0.4 (DR-2), 0.03 (DR-3), 0.04 (DR-4)]. This leads to a pseudo-binary prediction tuple as (0, L, 47%), indicating that the patient has non-referable DR (class 0), detected with low confidence (L) and a 47% likelihood of referable DR. Thus, the patient appears to be developing DR-1 and is recommended to have a checkup in

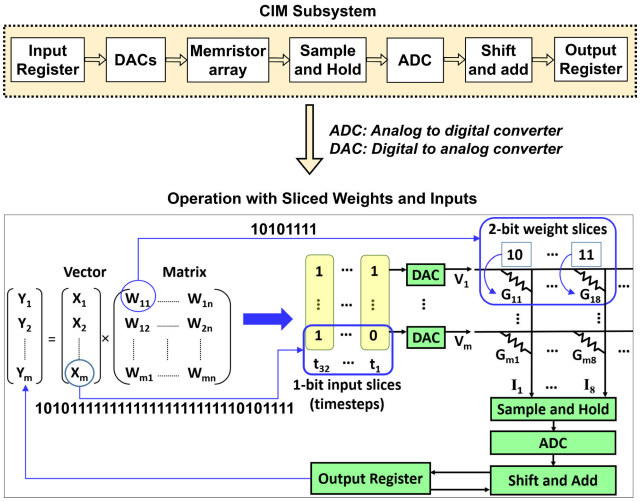


FIGURE 7. CIM hardware architecture for DR model implementation.

the near future. The interpretation of various output tuples of pseudo-binary classification is summarized in Table 4.

C. MEMRISTOR-BASED ENERGY-EFFICIENT HARDWARE DESIGN

1) PRUNING AND QUANTIZATION

Before mapping our trained pseudo-binary DR classification model to CIM hardware, we perform pruning and quantization to reduce its hardware resource requirements. Pruning refers to selectively removing a user-defined portion of low-magnitude weights from each layer. The reduction in hardware resource requirements due to pruning often comes at the cost of accuracy degradation. To counter this, we adopt pruning followed by retraining to recover lost accuracy. An essential consideration for such post-pruning retraining is the selection of hyperparameters, particularly the learning rate. An excessively low learning rate can hinder the network's adaptability to recover the pruning-induced accuracy loss. Hence, we dynamically adjust the learning rate within a narrow range centered around its original value during the retraining process. This iterative cycle continues until we achieve the desired level of pruning while preserving the network's original accuracy. We then quantize the weights of the pruned model to further reduce hardware resource requirements. However, an aggressive quantization can lead to high quantization error and degraded classification performance. Hence, we adopt a design space exploration approach to minimize bit-sizes for weights while ensuring minimal impact on classification performance.

2) MAPPING TO CIM ARCHITECTURE

We map our pruned and quantized pseudo-binary DR classification model to the memristor-based CIM architecture described in [50]. The fundamental building block of this architecture is depicted in Figure 7. It divides the full-precision neural network weights and inputs into smaller slices. This is because memristors have limited bit capacity

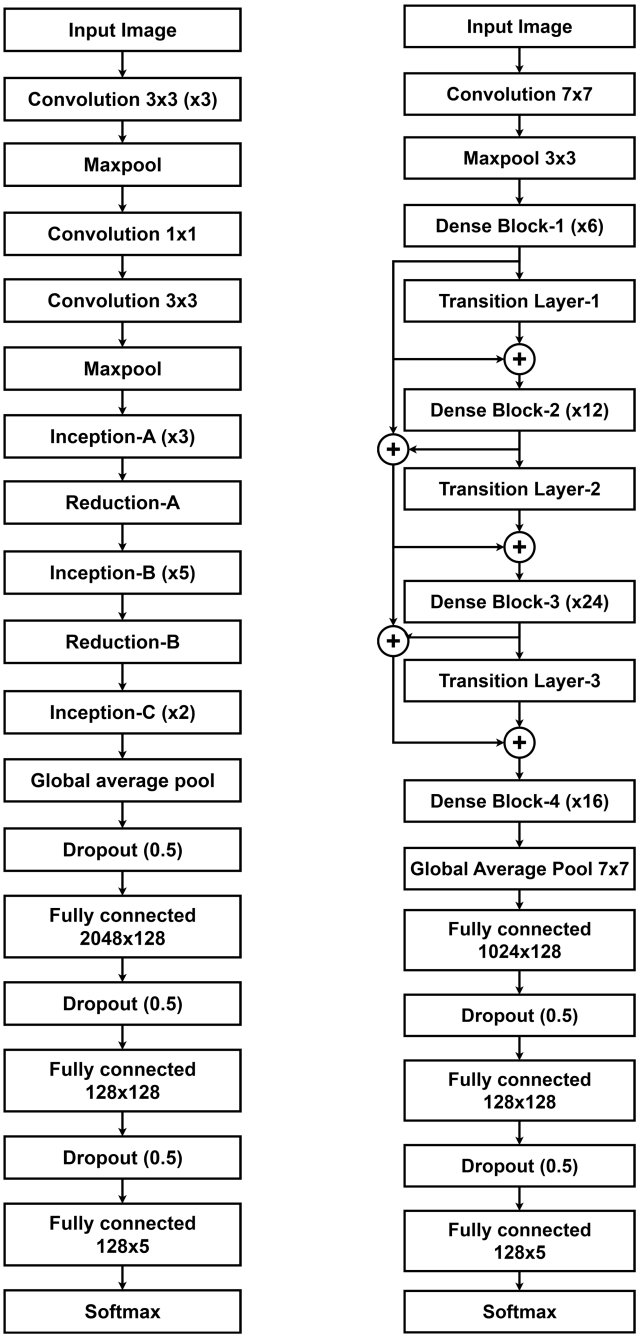


FIGURE 8. Modified inception-V3 (IV3) and densenet-121 (DN121) neural network architectures used to develop our DR classification model. Inception and reduction layers of IV3 are described in [48], while the dense blocks and transition layers of DN121 are covered in [49].

and high-resolution data converters (DACs and ADCs) consume significant energy and area. We transform 2-bit slices of the weights into conductance values, which are then mapped onto distinct columns within the memristor crossbar. We also convert 1-bit slices of the inputs into voltages that are applied to the crossbar at different timesteps. For instance, with 1-bit DACs for 32-bit digital inputs, the DACs are fed with 1-bit at a time across 32 timesteps. The DACs

TABLE 5. Training-validation-test split for the three newly proposed datasets: Small (S), Medium (M) and Large (L).

Proposed Dataset	Training Set					Validation Set					Test Set				
	DR-0	DR-1	DR-2	DR-3	DR-4	DR-0	DR-1	DR-2	DR-3	DR-4	DR-0	DR-1	DR-2	DR-3	DR-4
Small (S)	1059	229	589	112	191	382	67	197	33	48	357	69	205	43	53
Medium (M)	1775	598	1808	257	733	620	202	565	87	250	605	191	627	80	221
Large (L)	6021	4314	5992	1465	1887	2034	1432	1975	517	603	1945	1434	2033	522	627

convert the bits at each timestep into voltages, generating a current in each column of the crossbar. These currents are captured by sample and hold circuits (S&H) and then converted into digital outputs by ADCs. To account for the slicing of weights across crossbar columns, a shift and add operation is performed across the columns for the ADC outputs. Furthermore, an additional round of shift and add operations is performed to merge such partial outputs from various timesteps to produce the final full-precision digital output.

V. SIMULATION SETUP

A. SIMULATION PLATFORM

We use TensorFlow [51] framework for developing our DR classification model. Inception-v3 (IV3) [48] and DenseNet121 (DN121) [49] neural network architectures are selected for exploration, as they have demonstrated remarkable performance on complex image datasets [52]. We adapt these architectures for DR classification by introducing three new fully-connected layers (IV3: 2048 × 128, 128 × 128, 128 × 5 and DN121: 1024 × 128, 128 × 128, and 128 × 5) and few dropout layers (probability 0.5), as depicted in Figure 8. We train these networks with each of our new S/M/L datasets, following the train-validation-test split shown in Table 5. Employing transfer learning, we only train the newly added fully connected layers while freezing the pre-trained weights from the ImageNet dataset for all the other layers. During this training phase, we perform grid search followed by manual fine-tuning to establish optimal values for the hyperparameters. The post-training model performance is evaluated with the corresponding S/M/L test set.

To evaluate the reliability of the trained models, we employ publicly accessible Messidor-2 dataset [37], [38], [39] as an external training set. It contains 1748 images where 1058 images are provided by the Messidor program partners [37] and the remaining are collected at Brest University Hospital in France between 2009 and 2010. The labels for Messidor-2 dataset are sourced from [53], following the study in [54]. This labeling process has deemed four images ungradable, leaving 1744 usable images with classwise distribution as follows - DR-0: 1017, DR-1: 270, DR-2: 347, DR-3: 75, and DR-4: 35. As Messidor-2 embodies a data characteristics distinct from S/M/L train-validation-test sets, model performance on Messidor-2 serves as an indicator of its reliability.

To map the trained reliable model onto the memristor-based CIM hardware, we employ the architecture described in Section IV-C2. We leverage power consumption and latency data as presented in [50] to assess CIM energy consumption. This energy consumption is then compared against three conventional state-of-the-art hardware platforms: CPU (Intel Core i7-9750H [55]), GPU (NVIDIA GeForce GTX 1650 [56]), and mTPU (Google Edge TPU on Coral development board [57]). CPU and GPU represent general-purpose conventional hardware, while the mTPU embodies AI-optimized conventional hardware. To quantify energy consumption across these conventional hardware platforms, we first measure their latency and power consumption, and then calculate energy consumption as a product of these two values. The latency for both CPU and GPU is obtained via TensorFlow profiler [58], while that for mTPU is measured using Python datetime package [59]. The power consumption of the CPU is recorded using s-tui [60], while nvidia-smi [61] is used to record GPU power consumption. We use mTPU's datasheet to obtain its power consumption [62]. All of these details are summarized in Table 6.

TABLE 6. Simulation platform details.

Component	Specification
Deep learning framework	TensorFlow [51]
Network architectures	Modified inception-v3 (IV3) Modified densenet121 (DN121)
Datasets	New merged datasets (in Table 3): Small (S), Medium (M), Large (L)
CIM hardware	ISAAC [50]
Conventional hardware	CPU: Intel Core i7-9750H [55] GPU: NVIDIA GeForce GTX 1650 [56] mTPU: Google Edge TPU on Coral dev board [57]
Power profiling tools	CPU: s-tui [60] GPU: nvidia-smi [61] mTPU: datasheet [62] CIM: data provided in [50]
Latency profiling tools	CPU: Tensorflow profiler [58] GPU: Tensorflow profiler [58] mTPU: Python datetime package [59] CIM: data provided in [50]

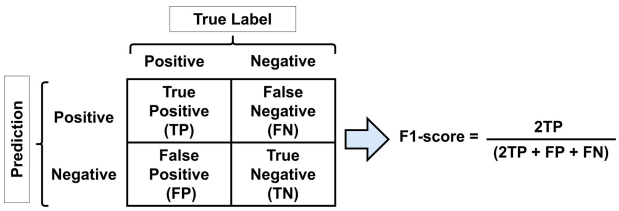
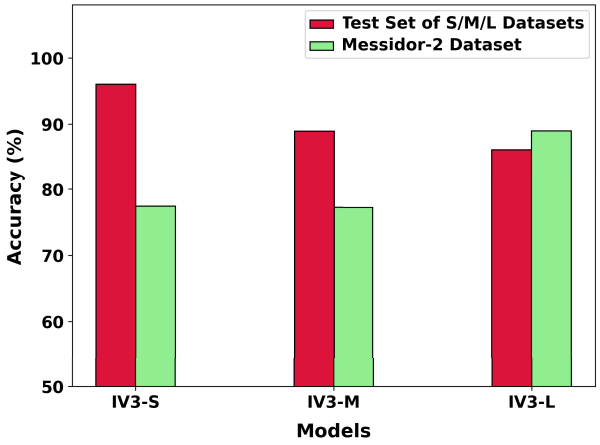
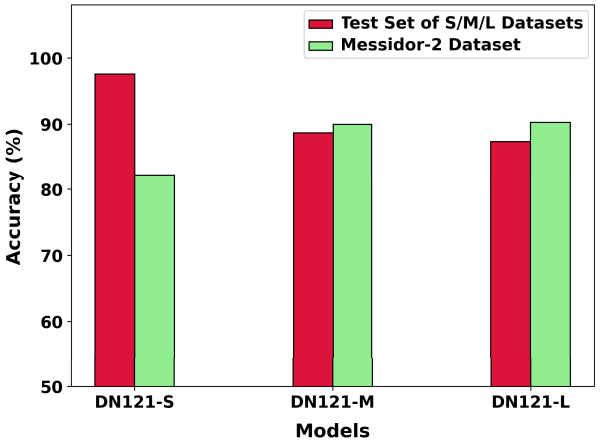


FIGURE 9. F1-score calculation from confusion matrix representation.

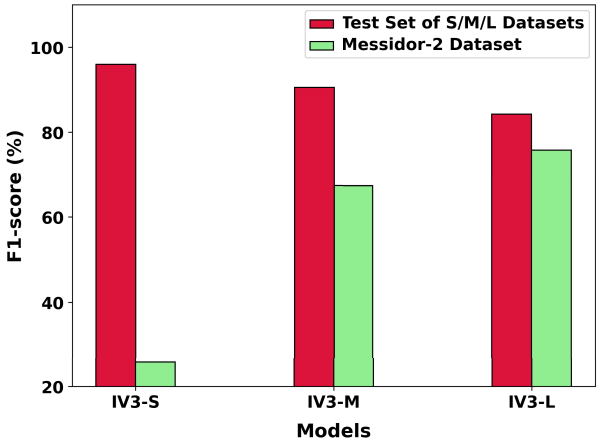


(a) IV3 accuracy.

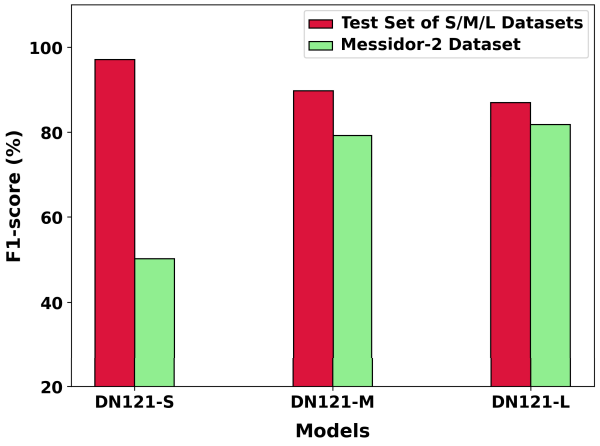


(b) DN121 accuracy.

FIGURE 10. Reliability assessment of IV3 and DN121 models in terms of accuracy.



(a) IV3 F1-score



(b) DN121 F1-score.

FIGURE 11. Reliability assessment of IV3 and DN121 models in terms of F1-score.

B. PERFORMANCE METRICS

The key performance metrics for the evaluation of the proposed DR classification include accuracy, F1-score, and energy consumption. These are described in detail as follows:

- **Accuracy**: It is defined as the ratio of the number of correctly classified retinal images to the total number of input retinal images, expressed as a percentage.
- **F1-score**: While accuracy is a valuable indicator of overall classification performance, there is a need for metrics that delve deeper into the model's behavior. The F1-score is one such metric that reflects the model's ability to make correct predictions while keeping false alarms to a minimum. It can be calculated using a table called the confusion matrix with true labels as column headers and predicted labels as shown in Figure 9.
- **Energy consumption**: Deployment of automated DR screening on portable edge devices can significantly improve its global accessibility, even in remote areas. To achieve this, such devices must be able to operate with limited and interrupted energy availability. Hence, energy consumed by a DR classification hardware is an important performance metric.

C. EXPERIMENTS PERFORMED

1) MODEL RELIABILITY ASSESSMENT

This experiment assesses the reliability of modified DN121 and modified IV3 models using Messidor-2 as external test set. The model with better reliability is selected for deployment and also compared with other works from literature.

2) MEMRISTOR-BASED CIM DESIGN OF RELIABLE MODEL

This experiment presents CIM-based hardware design of the selected reliable model (out of modified DN121 and modified IV3 models). The model is subjected to pruning and quantization to reduce the hardware resource requirement, followed by energy consumption assessment on CIM hardware and comparison with state-of-the-art hardware platforms.

TABLE 7. Comparison with other works using Messidor-2 as external testing dataset. 'N/A' indicates values not available from the paper.

Paper	Test Images	Accuracy (%)	F1-score (%)
This Work (DN121-L)	1744	90.3	81.8
Blair et al. [63]	1054	93.5	N/A
Ludwig et al. [15]	1058	N/A	83

VI. SIMULATION RESULTS

A. MODEL RELIABILITY ASSESSMENT

The reliability assessment of our modified IV3 and DN121 networks is shown in Figures 10 and 11. We train these networks across our three new datasets (S, M, and L in Table 3) by employing the pseudo-binary classification approach. This yields six distinct models: IV3-S (IV3 network trained on S dataset), IV3-M (IV3 network trained on M dataset), IV3-L (IV3 network trained on L dataset), DN121-S (DN121 network trained on S dataset), DN121-M (DN121 network trained on M dataset), and DN121-L (DN121 network trained on L dataset). The development phase performance of these models is assessed as their accuracy and F1-score on the corresponding S/M/L test set. To emulate post-deployment scenarios, we evaluate their accuracy and F1-score on Messidor-2 as an external test dataset. A reliable neural network model should exhibit consistent accuracy and F1-score, both during the development phase (on S/M/L test sets) and in post-deployment situations (on Messidor-2 dataset). This criterion forms the basis to assess the reliability of these six models in Figures 10 and 11.

Models developed using S dataset (IV3-S and DN121-S) show commendable performance on S/M/L test sets but struggle on Messidor-2, indicating low reliability. The other models trained with M and L datasets exhibit a marked improvement in reliability, consistently maintaining robust performance on both S/M/L test sets and Messidor-2 data. Also, model reliability improves as we transition from the M to the L dataset. This can be observed as IV3-L and DN121-L models outperform their M dataset counterparts IV3-M and DN121-M. This also highlights the pivotal role of large datasets in ensuring model reliability. As DN121-L exhibits better reliability than IV3-L in terms of both accuracy and F1-score, it becomes our final choice.

We now compare the reliability of our DN121-L model with other works from the literature. While several works [3], [6], [7], [9] have conducted evaluations on Messidor-2 dataset, they incorporate it within the training data rather than exclusively reserving it for testing. Consequently, a fair comparison with such studies is not possible. Additionally, works like [14] use datasets other than Messidor-2 for external testing and cannot be directly compared with our work. Therefore, we compare our approach with [15] and [63] which use Messidor-2 dataset only for external testing. As shown in Table 7, DN121-L achieves competitive accuracy and F1-score despite testing on 65% more Messidor-2 images compared to [15] and [63], validating its reliability.

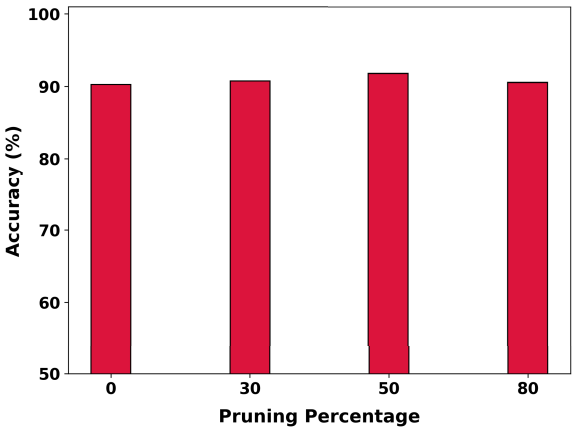


FIGURE 12. Impact of pruning percentages on accuracy of DN121-L model.

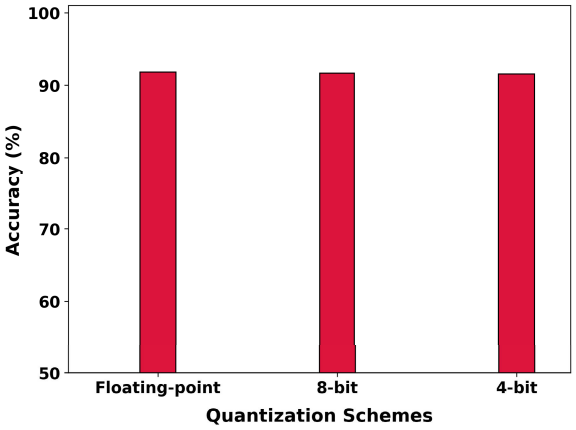


FIGURE 13. Impact of quantization on accuracy of DN121-L-P50 model.

We will discuss CIM hardware design for this reliable DN121-L model next.

B. MEMRISTOR-BASED CIM DESIGN OF RELIABLE MODEL

In this subsection, we first optimize the reliable DN121-L model for hardware design through pruning and quantization, and then assess its energy consumption on CIM hardware. We will now delve into the details of these steps.

1) PRUNING AND QUANTIZATION

Figure 12 shows the impact of pruning on DN121-L model. The pruning percentage indicates the fraction of model parameters with low magnitudes that undergo removal during the pruning process. Following the removal of the specified fraction of parameters, we conduct retraining for a few epochs to recover the accuracy loss incurred during pruning. Observing the classification performance across various pruning percentages, it becomes evident that the 50% pruned version of DN121-L exhibits the best classification performance. We select this version, denote it as DN121-L-P50 and subject it to weight quantization. Figure 13 shows that we can use 4-bit weights with almost no accuracy loss. Hence, we select the 4-bit quantized version of DN2121-L-P50 and denote it

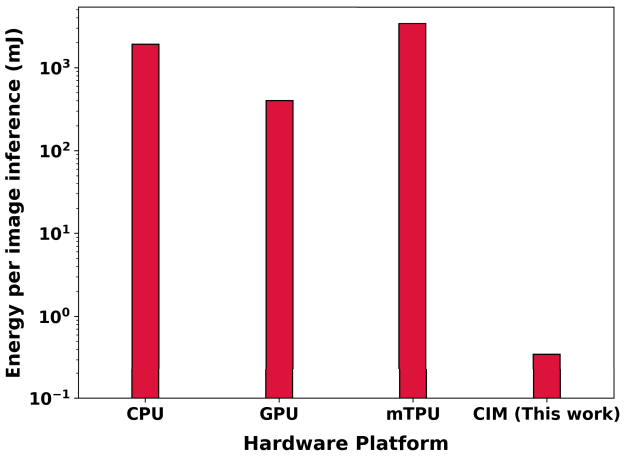


FIGURE 14. Energy per image inference for various hardware platforms.

as CIM-DN121. We will analyze its energy consumption on CIM hardware next.

2) ENERGY EFFICIENCY ASSESSMENT

The energy consumption for CIM-DN121 model, along with that of DN121 on state-of-the-art conventional hardware platforms like CPU, GPU, and edge TPU (mTPU) is depicted in Figure 14. It quantifies the energy required for executing inference on a single retinal image. The mTPU turns out to be the least energy-efficient despite being designed for AI applications. This is because mTPU's efficiency is limited by other resources on Coral dev board which handle tasks such as code context management and input/output data processing. For a large neural network model like DN121, these resources become the bottleneck, leading to significantly longer execution latency and increased energy consumption compared to CPUs or GPUs. Therefore, mTPU dev board may not be the best choice for energy-efficient execution of large neural network models. On the other hand, CIM-DN121 demonstrates 5441× reduction in energy consumption compared to CPU. Furthermore, it consumes 1144× and 9686× less energy compared to GPU and mTPU respectively. This highlights the tremendous potential of memristor-based CIM for developing energy-efficient hardware for DR screening.

VII. CONCLUSION

This paper has presented a reliable and energy-efficient hardware design for DR screening. We have achieved reliable classification by training the model with diverse and inconsistent quality data, while addressing class imbalance issue. We have then proposed a pseudo-binary classification technique to further improve the model performance and provide supplementary information. Furthermore, we have explored energy-efficient hardware design for our reliable DR model targeting deployment on edge devices for enhanced healthcare accessibility. Our final DR screening solution, based on DenseNet121 model, provides reliable classification with three orders of magnitude less energy consumption compared to the state-of-the-art hardware platforms. Thus, this work has laid the groundwork for reliable and accessible healthcare through the intersection of technology and medical science. In the future, extending this work for multiclass classification has a significant potential to increase the utility of AI-based DR diagnostic. Also, use of a larger and diverse external test data can provide better assessment of model reliability.

REFERENCES

[1] *Sight for All*. Accessed: Mar. 1, 2024. [Online]. Available: https://sightforall.org/diabetic-retinopathy-initiative/
[2] Y. Zhou, B. Wang, L. Huang, S. Cui, and L. Shao, "A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 818–828, Mar. 2021.
[3] M. Feng, J. Wang, K. Wen, and J. Sun, "Grading of diabetic retinopathy images based on graph neural network," *IEEE Access*, vol. 11, pp. 98391–98401, 2023.
[4] M. B. Khan, M. Ahmad, S. B. Yaakob, R. Shahrior, M. A. Rashid, and H. Higa, "Automated diagnosis of diabetic retinopathy using deep learning: On the search of segmented retinal blood vessel images for better performance," *Bioengineering*, vol. 10, no. 4, p. 413, Mar. 2023.
[5] W. K. Wong, F. H. Juwono, and C. Apriono, "Diabetic retinopathy detection and grading: A transfer learning approach using simultaneous parameter optimization and feature-weighted ECOC ensemble," *IEEE Access*, vol. 11, pp. 83004–83016, 2023.
[6] M. R. Islam, L. F. Abdulrazak, M. Nahiduzzaman, M. O. F. Goni, M. S. Anower, M. Ahsan, J. Haider, and M. Kowalski, "Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105602.
[7] S. S. Mohammadi and Q. D. Nguyen, "A user-friendly approach for the diagnosis of diabetic retinopathy using ChatGPT and automated machine learning," *Ophthalmol. Sci.*, Feb. 2024, Art. no. 100495.
[8] N. Sikder, M. Masud, A. K. Bairagi, A. S. M. Arif, A.-A. Nahid, and H. A. Alhumyani, "Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images," *Symmetry*, vol. 13, no. 4, p. 670, Apr. 2021.
[9] V. K. R. Poranki and B. Srinivasarao, "Computer-aided diagnosis-based grading classification of diabetic retinopathy using deep graph correlation network with IRF," *SN Comput. Sci.*, vol. 5, no. 2, p. 228, 2024.
[10] A. Sadeghzadeh, M. S. Junayed, T. Aydin, and M. B. Islam, "Hybrid CNN+transformer for diabetic retinopathy recognition and grading," in *Proc. Innov. Intell. Syst. Appl. Conf. (ASYU)*, Oct. 2023, pp. 1–6.
[11] A. R. Wahab Sait, "A lightweight diabetic retinopathy detection model using a deep-learning technique," *Diagnostics*, vol. 13, no. 19, p. 3120, Oct. 2023.
[12] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, "Diabetic retinopathy fundus image classification and lesions localization system using deep learning," *Sensors*, vol. 21, no. 11, p. 3704, May 2021.
[13] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Inf. Sci.*, vol. 501, pp. 511–522, Oct. 2019.
[14] E. Korot, Z. Guan, D. Ferraz, S. K. Wagner, G. Zhang, X. Liu, L. Faes, N. Pontikos, S. G. Finlayson, H. Khalid, G. Moraes, K. Balaskas, A. K. Denniston, and P. A. Keane, "Code-free deep learning for multimodality medical image classification," *Nature Mach. Intell.*, vol. 3, no. 4, pp. 288–298, Mar. 2021.
[15] C. A. Ludwig, C. Perera, D. Myung, M. A. Greven, S. J. Smith, R. T. Chang, and T. Leng, "Automatic identification of referral-warranted diabetic retinopathy using deep learning on mobile phone images," *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, p. 60, 2020.
[16] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, "Deep image mining for diabetic retinopathy screening," *Med. Image Anal.*, vol. 39, pp. 178–193, Jul. 2017.
[17] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, S. Xu, S. Barb, A. Joseph, M. Shumski, J. Smith, A. B. Sood, G. S. Corrado, L. Peng, and D. R. Webster, "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy," *Ophthalmology*, vol. 126, no. 4, pp. 552–564, Apr. 2019.

[18] Z. L. Teo, Y. C. Tham, M. Yu, M. L. Chee, T. H. Rim, N. Cheung, M. M. Bikbov, Y. X. Wang, Y. Tang, Y. Lu, I. Y. Wong, D. S. W. Ting, G. S. W. Tan, J. B. Jonas, C. Sabanayagam, T. Y. Wong, and C.-Y. Cheng, "Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis," *Ophthalmology*, vol. 128, no. 11, pp. 1580–1591, 2021.

[19] N. Panwar, P. Huang, J. Lee, P. A. Keane, T. S. Chuan, A. Richhariya, S. Teoh, T. H. Lim, and R. Agrawal, "Fundus photography in the 21st century—A review of recent technological advances and their implications for worldwide healthcare," *Telemed. e-Health*, vol. 22, no. 3, pp. 198–208, 2016.

[20] L. Lin, M. Li, Y. Huang, P. Cheng, H. Xia, K. Wang, J. Yuan, and X. Tang, "The SUSTech-SYSU dataset for automated exudate detection and diabetic retinopathy grading," *Sci. Data*, vol. 7, no. 1, p. 409, Nov. 2020.

[21] S. D. Solomon and M. F. Goldberg, "ETDRS grading of diabetic retinopathy: Still the gold standard?" *Ophthalmic Res.*, vol. 62, no. 4, pp. 190–195, 2019.

[22] E. Rotem, A. Yoaz, L. Rappoport, S. J. Robinson, J. Y. Mandelblat, A. Gihon, E. Weissmann, R. Chabukswar, V. Basin, R. Fenger, M. Gupta, and A. Yasin, "Intel alder lake CPU architectures," *IEEE Micro*, vol. 42, no. 3, pp. 13–19, May 2022.

[23] J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, "NVIDIA A100 tensor core GPU: Performance and innovation," *IEEE Micro*, vol. 41, no. 2, pp. 29–35, Mar. 2021.

[24] N. Jouppi, G. Kurian, S. Li, P. Ma, R. Nagarajan, L. Nai, N. Patil, S. Subramanian, A. Swing, B. Towles, C. Young, X. Zhou, Z. Zhou, and D. A. Patterson, "TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings," in *Proc. 50th Annu. Int. Symp. Comput. Archit.*, 2023, pp. 1–4.

[25] S. Jain, A. Sengupta, K. Roy, and A. Raghunathan, "RxNN: A framework for evaluating deep neural networks on resistive crossbars," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 40, no. 2, pp. 326–338, Feb. 2021.

[26] *International Roadmap for Devices and Systems*. Accessed: Mar. 1, 2024. [Online]. Available: https://standards.ieee.org/develop/indconn/irds/index.html

[27] S. Yu, W. Shim, X. Peng, and Y. Luo, "RRAM for compute-in-memory: From inference to training," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 7, pp. 2753–2765, Jul. 2021.

[28] M. A. Yaldagard, S. Diware, R. V. Joshi, S. Hamdioui, and R. Bishnoi, "Read-disturb detection methodology for RRAM-based computation-in-memory architecture," in *Proc. IEEE 5th Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, Jun. 2023, pp. 1–5.

[29] A. Singh, R. Bishnoi, A. Kaichouhi, S. Diware, R. V. Joshi, and S. Hamdioui, "A 115.1 TOPS/W, 12.1 TOPS/mm^2 computation-in-memory using ring-oscillator based ADC for edge AI," in *Proc. IEEE 5th Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, Jun. 2023, pp. 1–5.

[30] X. Yang, B. Taylor, A. Wu, Y. Chen, and L. O. Chua, "Research progress on memristor: From synapses to computing systems," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 69, no. 5, pp. 1845–1857, May 2022.

[31] R. Bishnoi, S. Diware, A. Gebregiorgis, S. Thomann, S. Mannaa, B. Deveautour, C. Marchand, A. Bosio, A. Bosio, I. O'Connor, H. Amrouch, and S. Hamdioui, "Energy-efficient computation-in-memory architecture using emerging technologies," in *Proc. Int. Conf. Microelectron. (ICM)*, Dec. 2023, pp. 325–334.

[32] A. Gebregiorgis, A. Singh, S. Diware, R. Bishnoi, and S. Hamdioui, "Dealing with non-idealities in memristor based computation-in-memory designs," in *Proc. IFIP/IEEE 30th Int. Conf. Very Large Scale Integr. (VLSI-SoC)*, Oct. 2022, pp. 1–6.

[33] L. Batina, R. Cammarota, N. Mentens, A.-R. Sadeghi, J. Sepúlveda, and S. Zeitouni, "Invited: Security beyond bulk silicon: Opportunities and challenges of emerging devices," in *Proc. 58th ACM/IEEE Design Autom. Conf. (DAC)*, Dec. 2021, pp. 1–4.

[34] Y. Luo, X. Han, Z. Ye, H. Barnaby, J.-S. Seo, and S. Yu, "Array-level programming of 3-bit per cell resistive memory and its application for deep neural network inference," *IEEE Trans. Electron Devices*, vol. 67, no. 11, pp. 4621–4625, Nov. 2020.

[35] S. Diware, A. Singh, A. Gebregiorgis, R. V. Joshi, S. Hamdioui, and R. Bishnoi, "Accurate and energy-efficient bit-slicing for RRAM-based neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 1, pp. 164–177, Feb. 2023.

[36] *APTOS 2019 Blindness Detection*. Accessed: Mar. 1, 2024. [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-detection/data

[37] *Messidor*. Accessed: Mar. 1, 2024. [Online]. Available: https://www.adcis.net/en/third-party/messidor/

[38] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, "Feedback on a publicly distributed image database: The Messidor database," *Image Anal. Stereol.*, vol. 33, no. 3, pp. 231–234, 2014.

[39] M. D. Abràmoff, J. C. Folk, D. P. Han, J. D. Walker, D. F. Williams, S. R. Russell, and P. Massin, "Automated analysis of retinal images for detection of referable diabetic retinopathy," *JAMA Ophthalmol.*, vol. 131, no. 3, pp. 351–357, 2013.

[40] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson, "Robust vessel segmentation in fundus images," *Int. J. Biomed. Imag.*, vol. 2013, Dec. 2013, Art. no. 154860.

[41] *Diabetic Retinopathy Messidor EyePac Pre Processed*. Accessed: Mar. 1, 2024. [Online]. Available: https://www.kaggle.com/datasets/mohammadasimbluemoon/

[42] *Diabetic Retinopathy Detection, Kaggle*. Accessed: Mar. 1, 2024. [Online]. Available: https://www.kaggle.com/competitions/diabetic-retinopathy-detection/overview

[43] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, p. 25, Jul. 2018.

[44] P. Taylor and H. W. W. Potts, "Computer aids and human second reading as interventions in screening mammography: Two systematic reviews to compare effects on cancer detection and recall rate," *Eur. J. Cancer*, vol. 44, no. 6, pp. 798–807, Apr. 2008.

[45] H. Geijer and M. Geijer, "Added value of double reading in diagnostic radiology,a systematic review," *Insights Imag.*, vol. 9, no. 3, pp. 287–301, 2018.

[46] *Determining Whether a Dataset is Imbalanced or Not*. Accessed: Mar. 1, 2024. [Online]. Available: https://datascience.stackexchange.com/questions/122571/determining-whether-a-dataset-is-imbalanced-or-not

[47] R. Chakrabarti, C. A. Harper, and J. E. Keeffe, "Diabetic retinopathy management guidelines," *Expert Rev. Ophthalmol.*, vol. 7, no. 5, pp. 417–439, 2012.

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[49] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[50] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, "ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2016, pp. 14–26.

[51] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement.*, 2016, pp. 265–283.

[52] *ImageNet Large Scale Visual Recognition Challenge*. Accessed: Mar. 1, 2024. [Online]. Available: https://www.image-net.org/challenges/LSVRC/

[53] *MESSIDOR-2 DR Grades*. Accessed: Mar. 1, 2024. [Online]. Available: https://www.kaggle.com/datasets/google-brain/messidor2-dr-grades

[54] J. Krause, V. Gulshan, E. Rahimy, P. Karth, K. Widner, G. S. Corrado, L. Peng, and D. R. Webster, "Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy," *Ophthalmology*, vol. 125, no. 8, pp. 1264–1272, 2018.

[55] *Intel Core i7-9750H Processor*. Accessed: Mar. 1, 2024. [Online]. Available: https://ark.intel.com/content/www/us/en/ark/products/191045/intel-core-i7-9750h-processor-12m-cache-up-to-4-50-ghz.html

[56] *GTX 1650 Gaming Laptops*. Accessed: Mar. 1, 2024. [Online]. Available: https://www.nvidia.com/en-eu/geforce/gaming-laptops/gtx-1650/

[57] *Dev Board*. Accessed: Mar. 1, 2024. [Online]. Available: https://coral.ai/products/dev-board/

[58] *Optimize TensorFlow Performance Using the Profiler*. Accessed: Mar. 1, 2024. [Online]. Available: https://www.tensorflow.org/guide/profiler

[59] *Datetime—Basic Date and Time Types*. Accessed: Mar. 1, 2024. [Online]. Available: https://docs.python.org/3/library/datetime.html

[60] *The Stress Terminal UI: S-TUI*. Accessed: Mar. 1, 2024. [Online]. Available: https://github.com/amanusk/s-tui

[61] *System Management Interface SMI*. Accessed: Mar. 1, 2024. [Online]. Available: https://developer.nvidia.com/nvidia-system-management-interface
[62] *Coral Dev Board Datasheet*. Accessed: Mar. 1, 2024. [Online]. Available: https://coral.ai/docs/dev-board/datasheet/
[63] J. P. M. Blair, J. N. Rodriguez, R. M. L. Vitar, M. A. Stadelmann, R. Abreu-González, J. Donate, C. Ciller, S. Apostolopoulos, C. Bermudez, and S. De Zanet, "Development of LuxIA, a cloud-based AI diabetic retinopathy screening tool using a single color fundus image," *Transl. Vis. Sci. Technol.*, vol. 12, no. 11, p. 38, 2023.

SUMIT DIWARE received the master's degree in VLSI design tools and technology from Indian Institute of Technology (IIT) Delhi, in 2018. He is currently pursuing the Ph.D. degree with the Computer Engineering Laboratory, Delft University of Technology, The Netherlands. His research interests include artificial intelligence, computation-in-memory, and neuromorphic architectures.

KOTESWARARAO CHILAKALA received the master's degree in embedded systems (computer architecture) from Delft University of Technology (TU Delft), The Netherlands, in 2021. He is currently a Consultant with Capgemini Engineering, The Netherlands. He has worked in both the medical and semiconductor industries, focusing on developing new-generation technologies. His expertise encompasses embedded system development, computer vision, and hardware-oriented optimization.

RAJIV V. JOSHI (Life Fellow, IEEE) received the B.Tech. degree from IIT Bombay, India, the M.S. degree from MIT, and the Dr.Eng.Sc. degree from Columbia University. He is currently a Research Staff Member and a Key Technical Lead with the IBM T. J. Watson Research Center. He has led successfully predictive failure analytic techniques for yield prediction and also the technology-driven SRAM at the IBM Server Group. His statistical techniques are tailored for machine learning and AI. He developed novel memory designs, which are universally accepted. He commercialized these techniques. He holds 60 invention plateaus and has over 250 U.S. patents and over 400, including international patents. He has authored and coauthored over 200 articles. He has given over 45 invited/keynote talks and given several seminars. He is a member of the IBM Academy of Technology and the Master Inventor. He is an ISQED Fellow and a World Technology Network Fellow and a Distinguished Alumnus of IIT Bombay. He received three outstanding technical achievements (OTAs) and the three highest corporate patent portfolio awards for licensing contributions. He received the NY IP Law Association "Inventor of the Year" Award, in February 2020. He received the Prestigious IEEE Daniel Noble Award, in 2018. He received the Best Editor Award from IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS journal. He was a recipient of the 2015 BMM Award. He was inducted into the New Jersey Inventor Hall of Fame, in August 2014. He served as a Distinguished Lecturer for the IEEE CAS and EDS Society. He is currently a Distinguished Lecturer of CEDA.

SAID HAMDIOUI (Senior Member, IEEE) received the M.S.E.E. and Ph.D. degrees (Hons.) from TUDelft. He is currently a Chair Professor of dependable and emerging computer technologies with Delft University of Technology, The Netherlands. He is also the Co-Founder and the CEO of Cognitive-IC, a start-up focusing on hardware dependability solutions. Prior to joining TUDelft as a Professor, he spent about seven years within the industry, including the Microprocessor Products Group, Intel Corporation, CA, USA; the IP and Yield Group, Philips Semiconductors Research and Development, Crolles, France; and the DSP Design Group, Philips/NXP Semiconductors, Nijmegen, The Netherlands. He has consulted for many companies (such as Intel, ST, Altera, Atmel, and Renesas) in the area of memory testing and has collaborated with many industry/research partners in the field of dependable nano-computing and emerging technologies. He is currently involved in different national and EU projects. He owns two patents, has published one book and contributed to the other two, and coauthored over 200 conference papers and journal articles. His research interests include dependable CMOS nano-computing (including testability, reliability, and hardware security) and emerging technologies and computing paradigms (including memristors for logic and storage and in-memory-computing for big-data applications). He is strongly involved in the international community as a member of organizing committees or a member of the technical program committees of the leading conferences. He delivered dozens of keynote speeches, distinguished lectures, and invited presentations and tutorials at major international forums/conferences/schools and leading semiconductor companies. He is a member of the AENEAS/ENIAC Scientific Committee Council (AENEAS—Association for European NanoElectronics Activities). He is an Associate Editor of IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS. He serves on the editorial board of *IEEE Design and Test*, *Microelectronic Reliability* (Elsevier), and *Journal of Electronic Testing: Theory and Applications*.

RAJENDRA BISHNOI received the Ph.D. degree in computer science from Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, in 2017. He is currently an Assistant Professor with the Computer Engineering Laboratory, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology (TU-Delft). Before joining TU-Delft, he was a Research Leader with the MRAM Group, Chair of Dependable Nano Computing, KIT, for more than two years. From 2006 to 2012, he was a Design Engineer with Freescale (NXP), where he was a part of the Technical Solution Group in memory and SoC flow. His current research interests include hardware AI, computation-in-memory, and emerging technologies. He was a recipient of the EDAA Outstanding Dissertation Award, in 2017.