CYIENT

TINYML OPTIMIZATION APPROACHES FOR DEPLOYMENT ON EDGE DEVICES

Achieving optimal accuracy and model size reduction in deep learning models for edge devices

CONTENTS

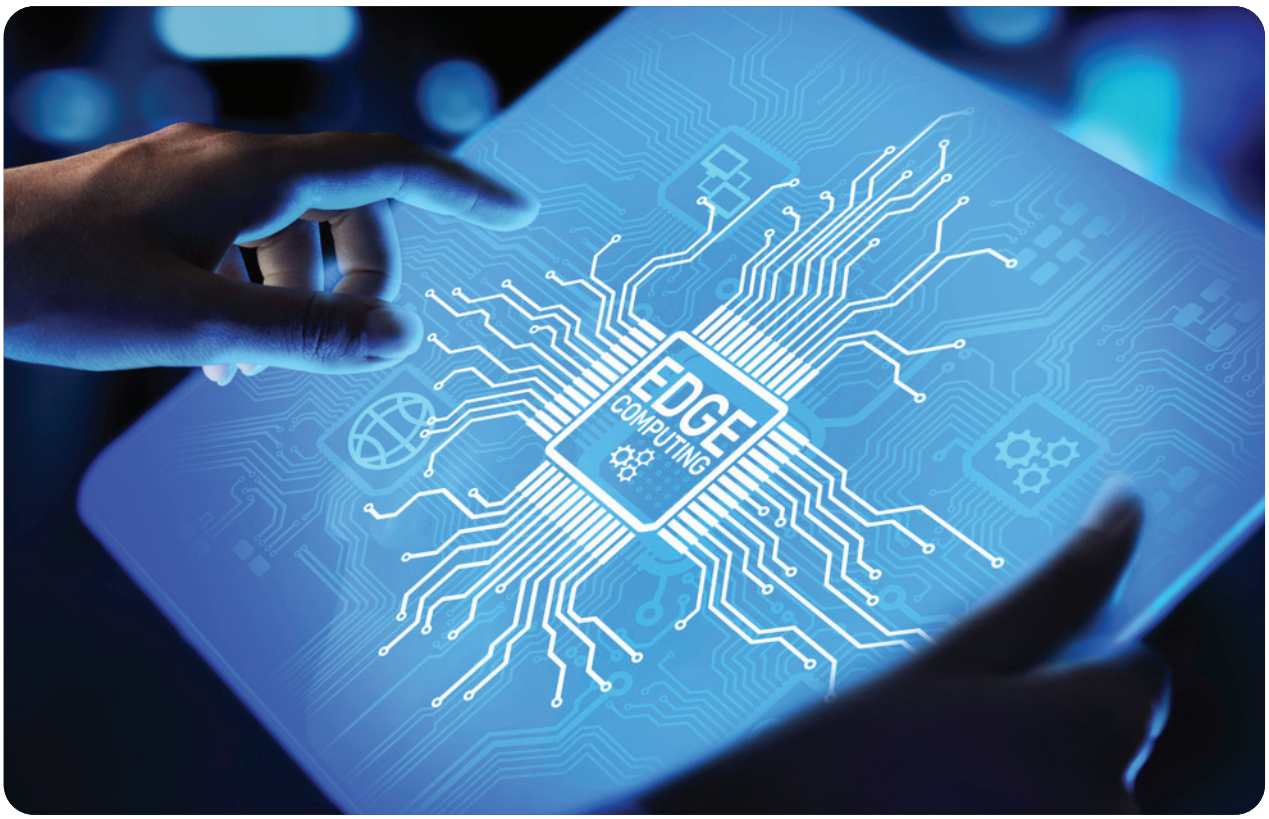
Abstract	1
Introduction	2
What is TinyML?	3
Model Compression for Efficient Deployment	5
Model Pre-processing and Training	6
Optimization Techniques	7
Data Set	7
Results	11
Conclusion	12
Key Insights	12
About the Authors	13
About Cyient	15

ABSTRACT

The demand for applications deploying deep learning models in resource-constrained environments is on the rise today, driven by the need for low latency and real-time inference. However, many deep learning models are too large and complex to perform effectively on such devices, posing challenges for scaling and model deployment.

Therefore, striking a compromise between maintaining high accuracy and reducing inference time and model size becomes essential. In the study presented in this white paper, three different models—Custom, VGG16[2], MobileNET[3]—are compressed using tiny machine learning or TinyML, a framework for model optimization and compression. The primary goal is to preserve optimal accuracy while significantly reducing inference time and model size.

The study will assess the trade-offs between accuracy, size reduction, and inference time by comparing the compressed models by tailoring and comparing the performance with the original models. Additionally, the study intends to explore TinyML's potential to enhance user experience and enable edge computing in medical applications.



I. Introduction

In recent years, the deployment of deep learning models on resource-constrained edge devices, such as smartphones, wearable devices, IoT devices, edge servers, and embedded systems has increased exponentially, posing challenges due to their limited computational power and memory space.

The study presented here, aimed to employ TinyML (tiny machine learning) techniques for compressing Custom, VGG16, and MobileNet models across datasets taken from the fashion, radiology, and dermatology fields. It prioritizes the following three features: achieving optimal accuracy, reduced inference time, and minimized model size suitable for deployment on resource-constrained edge devices.

The main compression techniques applied here are quantization and pruning. Quantization makes numbers in the model much less precise, while pruning selectively eliminates redundant or unnecessary connections within the neural network architecture without compromising the model too much. This helps in reducing the computational cost and memory usage for optimal and quick deployment and is adjusted to the requirements of TinyML.

For our study, which involved meticulous testing of different permutations and combinations on the datasets, our focus was to tweak the models and parameters to increase accuracy, inference time, and model size. Further, this endeavor leverages quantization techniques from TensorFlow Lite compression to investigate how TinyML might facilitate edge computing and improve user experience. It makes it easier to implement effective and lightweight models, allowing for real-time inference on edge devices with limited resources.

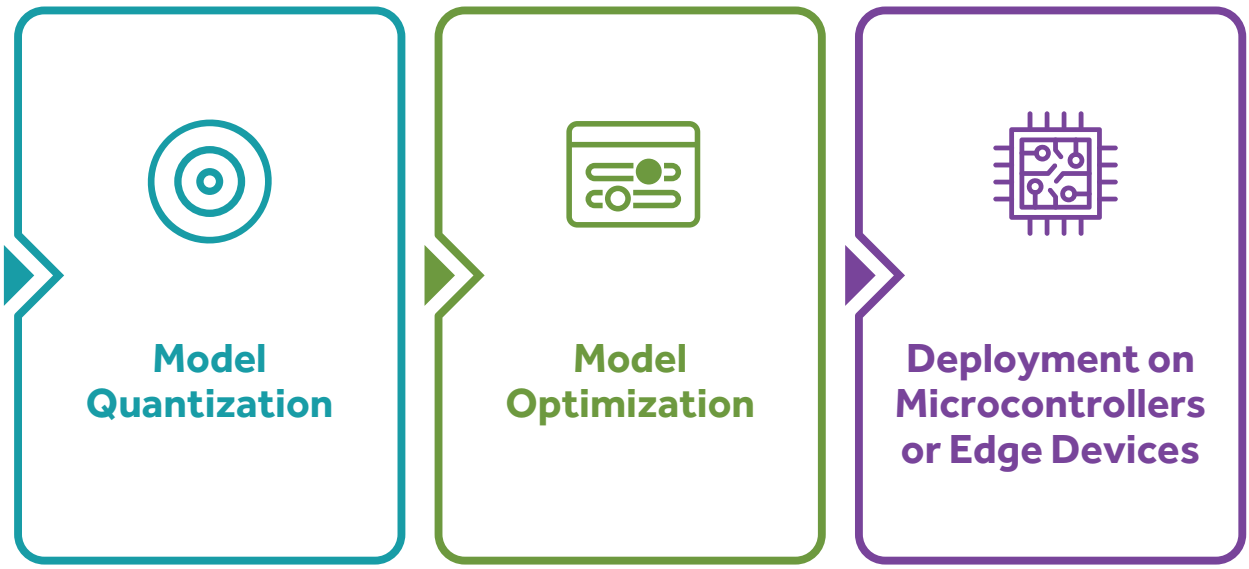
What is TinyML?

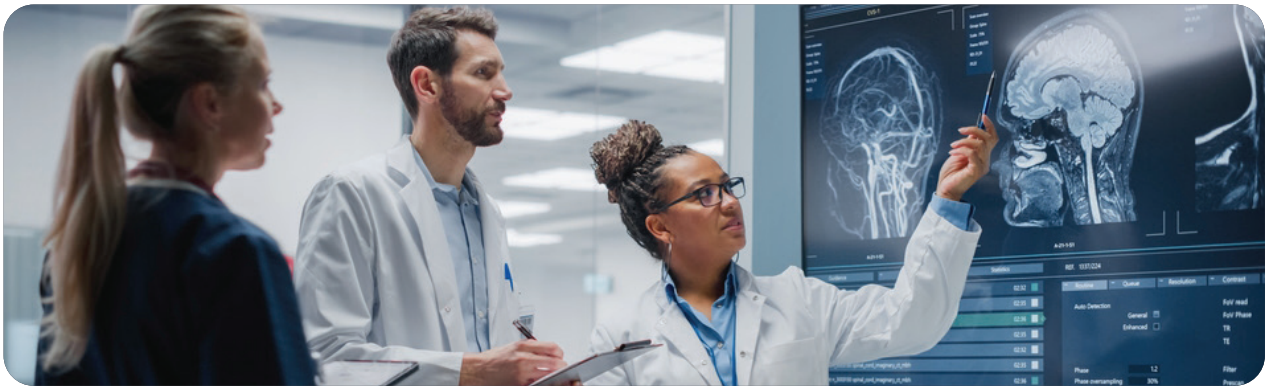
Tiny machine learning or TinyML is a decentralized technique that allows us to deploy machine learning models and algorithms on extremely low-power and small-footprint devices such as microcontrollers and IoT devices. With TinyML, microcontroller-based embedded devices can quickly respond in real time to machine-learning tasks.

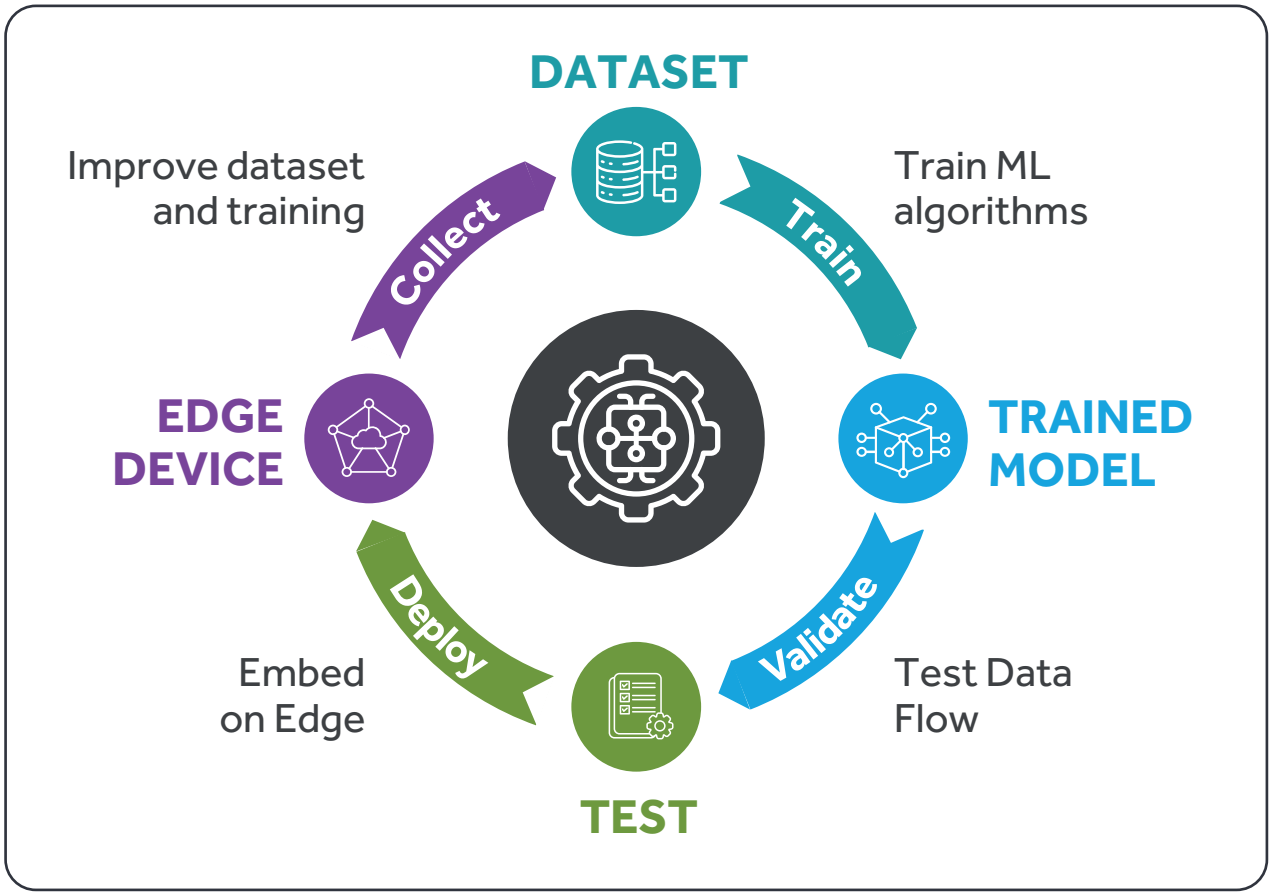
Its applications span multiple disciplines of technology such as hardware, algorithms, and software capable of processing on-device sensor data.

The most widely adopted ecosystem for TinyML development is TensorFlow Lite. TensorFlow Lite provides a Python-based environment with an extensive collection of pre-installed libraries and toolkits that offer developers ways to effectively implement machine learning models.

The general process followed for deployment on edge devices is shown below:







Applications of TinyML

Personalized Healthcare

Industry IoT

Smart Homes

Autonomous Driving

II - Problem Definition:

Model Compression for Efficient Deployment

We focused on the following elements for this study:

- Implementation of TinyML techniques to compress Custom, VGG16, and MobileNet models on three different datasets across various industry disciplines (fashion, radiology, and dermatology).
- Prioritizing a low footprint for deployment while ensuring the best possible accuracy, reduced inference time, and minimized model size.

Solution Details

The principal objective was to create an accelerator methodology for deep learning model compression that is tailored for TinyML edge device deployment. Initially, we trained SSD-MobileNet and EfficientDet models on a custom dataset using the TensorFlow 2 Object Detection API. These models were then transformed into the TensorFlow Lite format using TinyML compression techniques to ensure their integration with edge devices while keeping the three primary metrics of good accuracy, low inference time, and reduced model size.

Overall, our solution offers a comprehensive and effective approach to deploying compressed deep learning models optimized for TinyML on edge devices, addressing the growing demand for efficient and lightweight machine learning solutions in various industry sectors.

Business Benefits/ Best Practices

The global edge computing market size was valued at $16.45 billion in 2023 and is expected to grow at a CAGR of 37.9% from 2023 to 2030, so optimizing deep learning models for edge devices emerges as a pivotal strategy. [1]

In the rapidly changing field of healthcare edge computing, the need for real-time patient data analysis, enabling remote diagnostic and treatment decisions without the need for continuous Wi-Fi connectivity, is imperative. The model needs to be scalable, customizable, and tailored to the needs of the healthcare industry. For example, wearable health monitoring systems with low-power microcontrollers can accurately detect irregularities in vital signs, promptly alerting medical personnel and improving patient outcomes.

The TinyML approach facilitates businesses with a competitive advantage by delivering substantial cost savings and performance enhancements, increasing customer satisfaction. These systems can be deployed in environments without Wi-Fi networks and in remote or inaccessible locations, seamlessly executing required tasks.

III - Model Pre-processing and Training:

Image Resizing

- Resized images to specific dimensions to showcase model adaptability

Pre-Processing

- Organized datasets by conducting essential pre-processing steps

Model Assessment and Training

- Utilized pre-trained models such as VGG16 and MobileNet, along with Custom models
- Evaluated model effectiveness and performance through various configurations and optimization techniques

IV-Tensorflow Lite Quantization Techniques

- TensorFlow Lite Float16
- TensorFlow Lite Quant by Size
- TensorFlow Lite Quant by Latency
- TensorFlow Lite Exp. Sparsity
- TensorFlow Lite int16
- TensorFlow Lite int8

V- Data Sets

Data Set Optimization Domains: Fashion, Radiology, and Dermatology

Optimization algorithms were implemented using datasets from a variety of fields such as fashion, radiology, and dermatology.

Utilizing diverse datasets provides important insights into the effectiveness and versatility of optimization methods. Each dataset has different challenges, traits, and subtleties that call for investigating a large range of optimization strategies. Through the integration of diverse datasets, we can acquire a thorough understanding of the performance of the different optimization approaches.

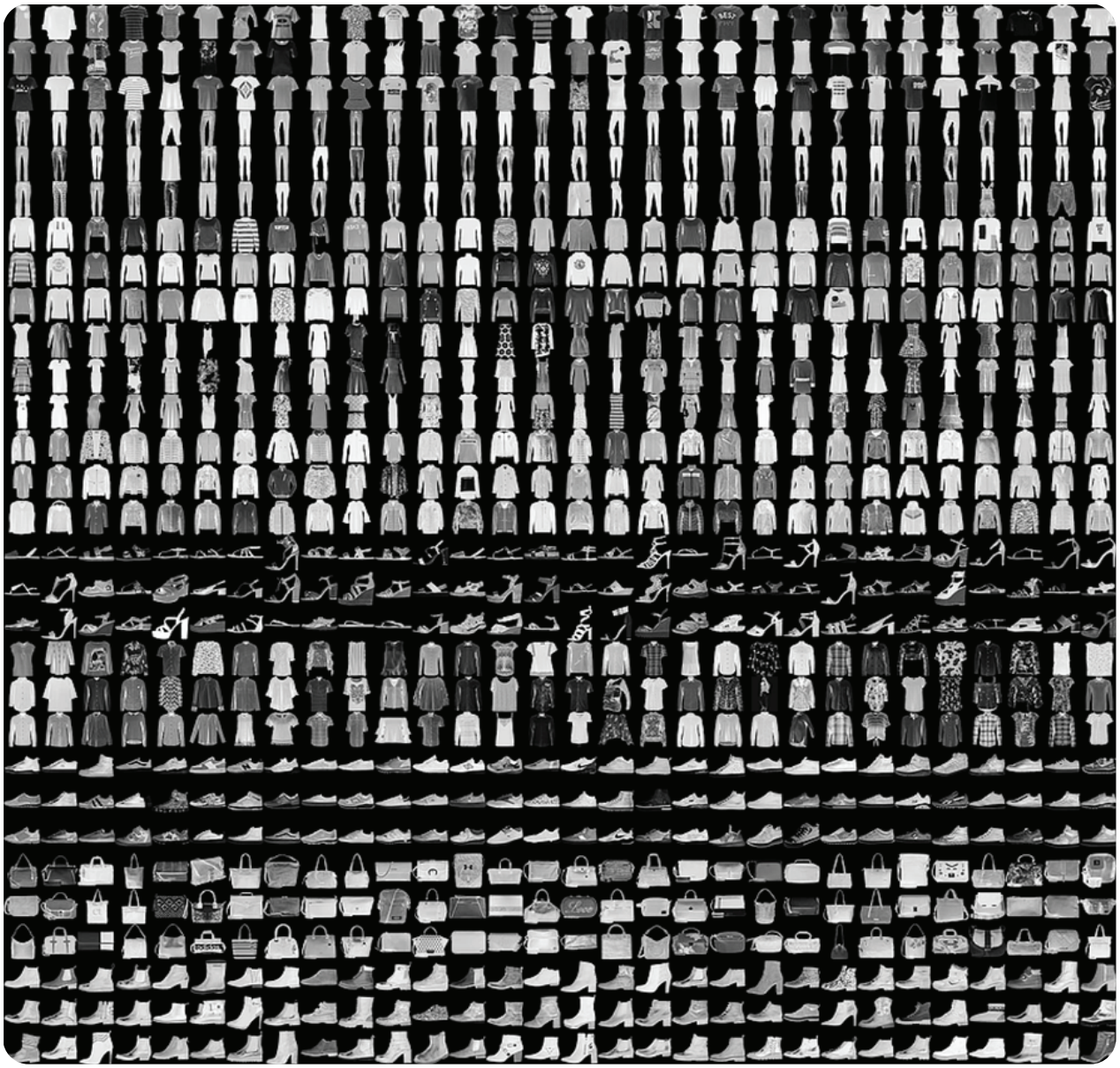
To optimize models, we first executed a script on the data of the respective dataset to gain familiarity with its characteristics. Using TensorFlow Lite, quantization and pruning techniques were employed to meet our primary imperatives of good accuracy, low inference time, and reduced model size.

A. Fashion: FashionMNIST optimization

For the fashion domain, we optimized the models trained on the FashionMNIST dataset.

Fashion datasets contain images of clothing items, accessories, and fashion products.

FashionMNIST Dataset Specifications	Dataset Information
Dataset	Contains 70,000 grayscale images in 10 categories. The images show individual articles of clothing at low resolution (28 by 28 pixels).
Train	60,000 images
Test	10,000 images
Classes	0-9
Class names (labels)	'T-shirt/top', 'Trouser', 'Pullover', 'Dress', 'Coat', 'Sandal', 'Shirt', 'Sneaker', 'Bag', 'Ankle boot'

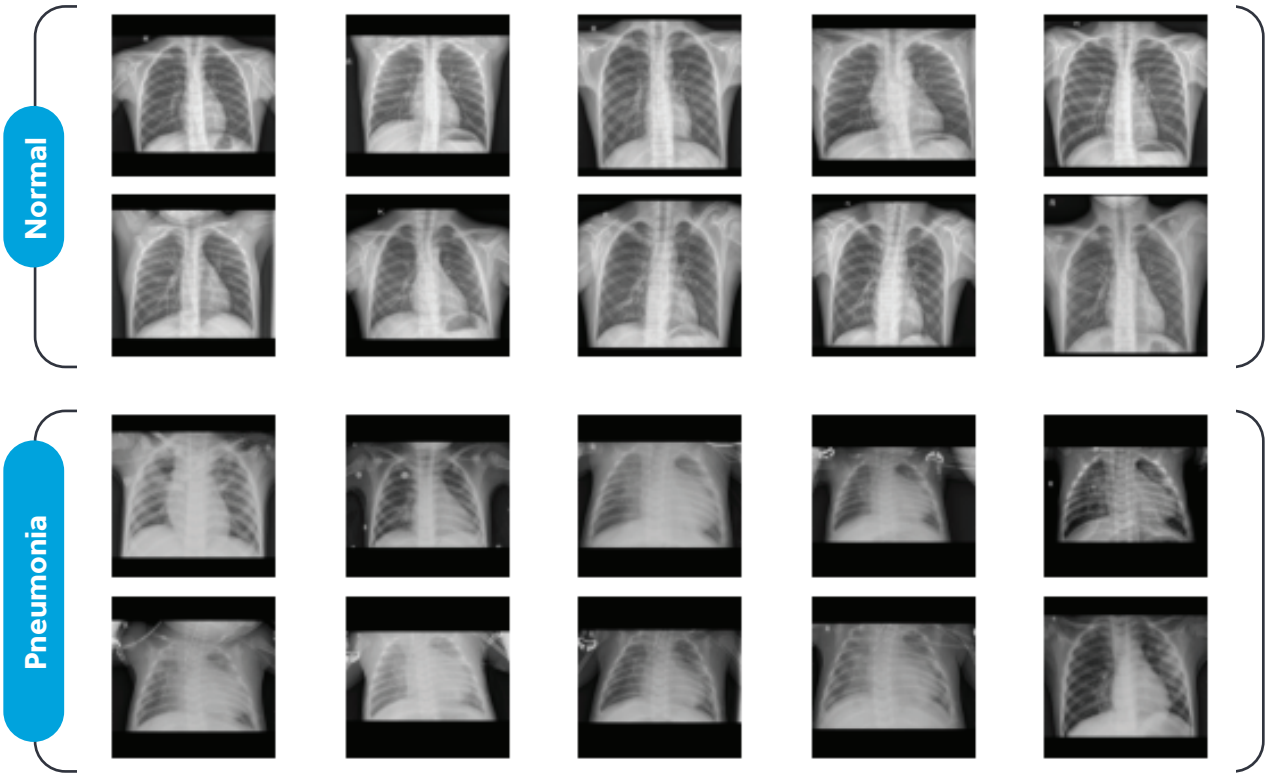


B. Radiology: Pneumonia dataset optimization

Radiology datasets comprise medical images such as chest X-rays, CT scans, and radiology function test results. For example, machine learning models trained on radiology datasets can analyze chest X-ray images to detect abnormalities such as pneumonia, tuberculosis, or lung cancer, assisting radiologists in accurate diagnosis and treatment planning.

For this dataset, we first organized and prepared a dataset comprising the chest X-ray images of two types of patients: normal and with pneumonia. Then, we resized images to different dimensions (150x150, 1000x1000, and 224x224) to show how well models can adjust. After that, we trained models using both custom and pre-trained approaches, noting how different model setups performed.

Radio logy Dataset Specifications	Dataset Information
Dataset	Chest X-ray images
Train	5216 images (Normal: 1341, Pneumonia: 3875)
Test	624 images (Normal: 234, Pneumonia: 390)
Validation	16 images (Normal: 8, Pneumonia: 8)
Classes	0-1
Class names (labels)	'Normal', 'Pneumonia'

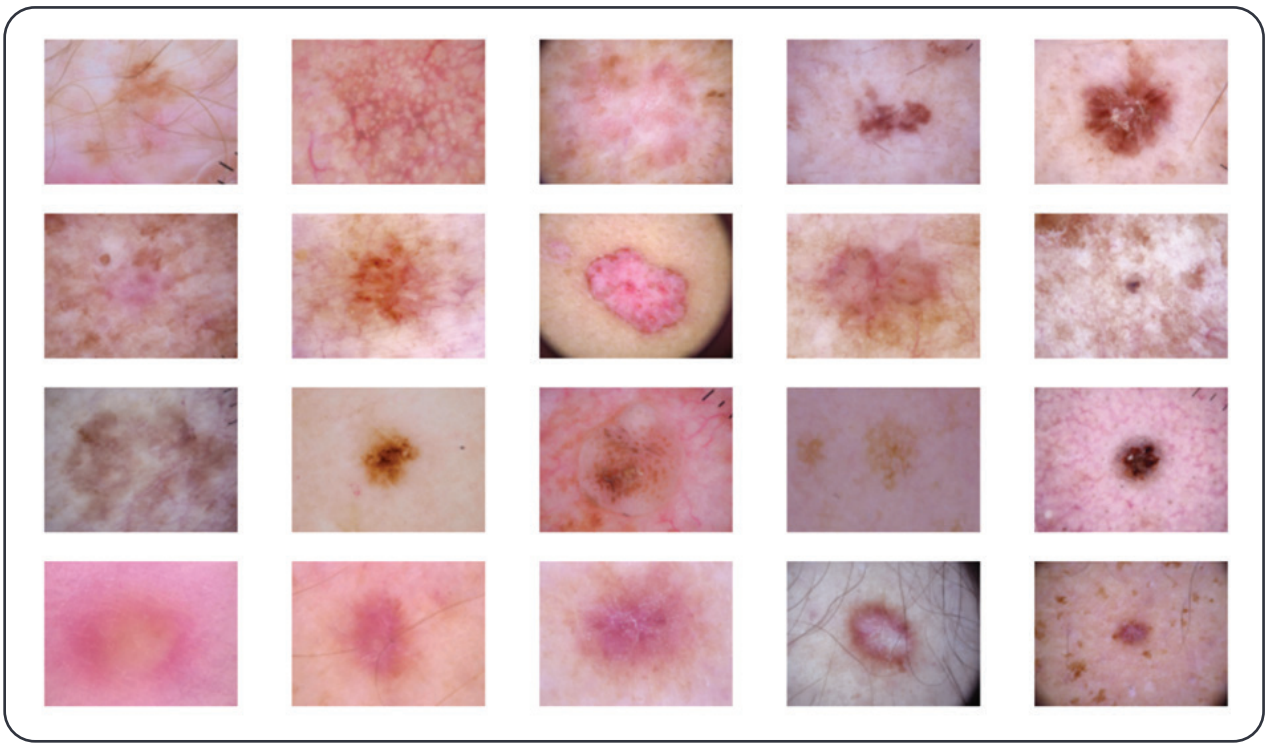


C. Dermatology: Skin lesion dataset optimization

Dermatology: Dermatology datasets typically consist of images of skin conditions, lesions, and diseases.

We began by sampling a skin lesion dataset, which included multi-source dermatoscopic images of pigmented lesions. We organized and prepared the dataset through pre-processing to ensure readiness for analysis. Like other datasets, we resized the images to 224x224, demonstrating the adaptability of the models to specific image dimensions. For training, we utilized pre-trained models such as MobileNetV2, and various configurations were assessed to gauge performance metrics.

Skin Lesion Dataset Specifications	Dataset Information
Dataset	Dermatoscopic images
Train	6585 images
Test	840 images
Classes	0-6
Class names (labels):	(AKIEC) Actinic Keratosis', 'Basal Cell Carcinoma (BCC)', 'Benign Keratosis (BKL)', 'Dermatofibroma (DF)', 'Melanocytic nevus (NV)', 'Melanoma (MEL)', 'Vascular lesion (VASC)



Models Utilized Across Diverse Domains:

Domain	Dataset	Models Used
Fashion	FashionMNIST dataset	Custom, VGG16, MobileNET
Radiology	Pneumonia dataset	Custom, VGG16, MobileNET
Dermatology	Skin lesion dataset	Custom, VGG16, MobileNET

VI - Results:

	Accuracy (%)								Size Reduction (%)								Inference Time (Seconds)							
Dataset/ Model	Tensor Flow	Tensor Flow Lite	TF Lite Float16	TF Lite Quant by Size	TF Lite Quant by Latency	TF Lite Exp. Sparsity	TF Lite int16	TF Lite int8	Tensor Flow	TensorFlow Lite	TF Lite Float16	TF Lite Quant by Size	TF Lite Quant by Latency	TF Lite Exp. Sparsity	TF Lite int16	TF Lite int8	Tensor Flow	TensorFlow Lite	TF Lite Float16	TF Lite Quant by Size	TF Lite Quant by Latency	TF Lite Exp. Sparsity	TF Lite int16	TF Lite int8
Fashion MNIST (28x28 - Custom)	88.23	88.23	88.23	88.24	88.24	88.23	*	*	N.A.	67.21	83.49	91.65	91.65	67.21	*	*	1.22	0.11	0.09	0.25	0.15	0.094	*	*
Fashion MNIST (32 x32 - VGG16)	84.47	84.47	84.47	*	*	84.47	*	84.09	N.A.	52.95	76.47	*	*	52.95	*	88.18	8.19	152.71	110.83	*	*	110.98	*	56.93
Fashion MNIST (96 x 96- MobileNet)	88.84	88.85	88.93	84.87	84.87	88.85	84.87	10	N.A.	6.43	52.81	73.55	73.55	6.43	73.55	71.41	9.01	20.98	52.81	3697.43	3697.43	21.44	3655.64	19.84
Pneumonia (150x150 - Custom)	89.58	37.5	37.5	37.5	37.5	37.5	*	*	N.A.	66.83	83.36	91.54	91.54	66.83	*	*	7.95	5.25	3.87	19.06	11.28	4.04	*	*
Pneumonia (224*224 - Mobilenet)	81.25	37.5	37.5	37.5	37.5	37.5	37.5	37.5	N.A.	11.44	55.42	75.08	75.08	11.44	75.08	73.11	5	7.3	7.27	1466.15	1466.15	8.57	1157.9	7.06
Pneumonia (224x224 - VGG16)	96.91	91.51	91.51	*	*	91.51	*	62.5	N.A.	0.8	50.38	*	*	0.8	75.08	74.99	15	279.91	255.43	*	*	293.63	*	192.72
Skin Lesion (224 x 224- VGG16)	78.82	19.88	19.88	*	*	19.88	*	*	N.A.	50.02	75	87.41	87.41	50.02	*	87.36	26	328.14	330.23	*	*	327.63	*	*
Skin Lesion (224 x 224- MobileNet)	76.17	66.82	66.82	66.59	66.59	66.82	66.59	63.53	N.A.	51.86	75.72	86.38	86.38	51.86	86.38	85.28	19	56.63	52.46	1948.54	1881.23	46.32	1832.91	1967.95

CONCLUSION

Our study shows that TinyML is a powerful framework for deploying deep learning models on edge devices. It offers a balance between efficiency and accuracy making it ideal for diverse applications in computer vision, natural language processing, and beyond.

1. **Accuracy vs Inference time vs Model size trade-off:**
   - TensorFlow consistently outperforms TensorFlow Lite optimizations across all models.
   - VGG16 shows the highest accuracy among the models optimized by TensorFlow Lite.

2. **Inference time assessment:**
   - While some TensorFlow Lite optimizations increase inference time, it remains generally acceptable.
   - Specific optimization techniques on certain models lead to prolonged inference times.

3. **Quantization for model size reduction:**
   - Custom and MobileNet models achieve significant size reduction with TensorFlow Lite.
   - VGG16 exhibits notable size reduction with INT16 and INT8 optimizations.

Key Insights:

1. **TinyML optimization effectiveness:** Our research shows that TinyML is effective in significantly reducing size by quantization without compromising model fidelity. Among the TinyML framework's accuracy performers, VGG16 was the better-performing model but only if the parameters were tweaked to support it.

2. **Effect on inference time:** Overall, TinyML optimizations provide good inference speeds, even though some of them result in longer inference times. This emphasizes the significance of selecting optimization strategies carefully.

3. **Quantization for size reduction:** TinyML applications demonstrate impressive model size reductions, especially with Custom and MobileNet models. Additionally, using INT16 and INT8 optimizations significantly shrinks the size of VGG16 models, highlighting the value of quantization methods in TinyML.

4. **Achieving balanced trade-offs for ideal implementation:** For TinyML models to be deployed as effectively as possible, the ideal balance between accuracy, inference time, and model size must be navigated. To successfully traverse and optimize trade-offs and ensure optimal performance across a variety of deployment circumstances, ongoing customization based on unique model requirements continues to be essential.

ABOUT THE AUTHORS

Srinivas has over 25 years of experience which spans across Consumer Electronics, Biomedical Instrumentation and Medical Imaging. He has led research and development teams, focused on end-to-end 3D/4D quantification applications and released several "concept to research to market" solutions. He also led a cross functional team to drive applied research, product development, human factors team, clinical research, external collaboration and innovation. He has garnered diverse sets of skill sets and problem challenges. and has over 25 Patent filings and 12 Patent Grants across varied domains, mentored over 30+ student projects, been a guide for over 10+ master thesis students, peer reviewer for papers and an IEEE Senior Member (2007).

Malavika is an Electronics and Instrumentation engineer with experience in biomedical Instrumentation. She also brings expertise in business analysis within supply chain management incorporating the tools in data analytics to devise solutions to optimize and track KPIs. She predominantly contributes to shaping impactful solutions for customer-facing endeavors in technology and business domains.

Key Contributors:

- Amol Gharpure: amol.gharpure@cyient.com
- Akanksha Jarwal: akanksha.jarwal@cyient.com
- Manshi Mithra: manshimithra.t@cyient.com

References

[1] https://www.grandviewresearch.com/industry-analysis/edge-computing-market#:~:text=The%20global%20edge%20computing%20market,36.9%25%20from%202024%20to%202030

[2] https://keras.io/api/applications/vgg/

[3]https://keras.io/api/applications/mobilenet/

For more information on this Whitepaper, please feel free to get in touch with Srinivas Rao Kudavelly at srinivasrao.kudavelly@cyient.com

ABOUT CYIENT

Cyient (Estd: 1991, NSE: CYIENT) partners with over 300 customers, including 40% of the top 100 global innovators of 2023, to deliver intelligent engineering and technology solutions for creating a digital, autonomous, and sustainable future. As a company, Cyient is committed to designing a culturally inclusive, socially responsible, and environmentally sustainable Tomorrow Together with our stakeholders.

For more information, please visit www.cyient.com

Follow us on:

CONTACT US

North America Headquarters
Cyient, Inc.
99 East River Drive
5th Floor
East Hartford, CT 06108
USA
T: +1 860 528 5430
F: +1 860 528 5873

Europe, Middle East, and Africa Headquarters
Cyient Europe Limited
Apex, Forbury Road,
Reading
RG1 1AX
UK
T: +44 118 3043720

Asia Pacific Headquarters
Cyient Limited
Level 1, 350 Collins Street
Melbourne, Victoria, 3000
Australia
T: +61 3 8605 4815
F: +61 3 8601 1180

Global Headquarters
Cyient Limited
Plot No. 11
Software Units Layout
Infocity, Madhapur
Hyderabad - 500081
India
T: +91 40 6764 1000
F: +91 40 2311 0352