“Healthcare delivery in India is now uniquely poised to undergo a change at all its stages – prevention, diagnosis, and treatment. No single entity in the healthcare sector can work in isolation. The evolution of the sector calls for involvement from all stakeholders and the use of innovation to bridge intent and execution.”

- Nayan Kalnad, Chief Business Officer, Marai Health

In a country in which healthcare practitioners are concentrated in urban areas, delivering effective radiological screening to rural populations is a challenge that can be addressed very well with AI advances.

Challenge
Providing quality health services and screening to rural populations in a nation as large as India can be extremely challenging. For example, India has only three accredited radiologists per million people. Using AI technology to provide more extensive, effective radiological screening has the potential for saving lives and providing overall improvements to health across the country.

Solution
A unified approach to handling diverse medical images that span modalities presents a distinct challenge to researchers and developers, one requiring a compute-intensive processing platform and an innovative approach to the deep neural network model.

Background and Project History
A research article in The Journal of Global Radiology titled “Implementing Diagnostic Imaging Services in a Rural Setting of Extreme Poverty” outlined the obstacles faced by health practitioners to provide wide-ranging imaging services. Assessing the difficulty in obtaining consults with experts in radiology, the authors noted, “When doctors feel an X-ray or ultrasound is beyond their skill level to diagnose, limited options exist for consultation. This is of particular concern with ultrasound at hospitals like BH, where doctors estimate learning about 95% of diagnostic procedures from more specialized providers.”

To help address screening radiology challenges in rural India, Debdoot Sheet launched Project MIRIAD, exploring innovative AI-guided screening techniques. Based at the Indian Institute of Technology Kharagpur, Debdoot is the Assistant Professor in the Department of Electrical Engineering and Principal Investigator in the Kharagpur Learning, Imaging and Visualization Group. A presentation describing his objectives, Project MIRIAD: Exploring 3K+ CNNs Beyond ImageNet for Screening Radiology, summarized the scope of the challenge with some eye-opening statistics:

- 67 percent of the Indian population resides in rural areas.
- 90 percent of medical imaging facilities are in cities.
- There are only 3 radiologists per 1,000,000 people in the country.
With a background that includes an undergraduate degree in electronics and communication engineering with specialization in digital signal processing, a master’s degree in computer vision for medical imaging, and a PhD in computational medical imaging and machine learning, Debdoot is well positioned to devise and apply new AI technologies for addressing the challenge and devise methods that help unify a diverse range of medical image types.

“During my research,” Debdoot said, “I’ve focused on compute optimizations for training learning-based models in settings with limited compute and data resources. Leading my research group as an assistant professor since 2014, we have been focusing on efficient computing platforms and hybrid computing-based learning systems for developing competing solutions in the medical image analysis space under significant computing power limitations.”

“We do not necessarily rely only on the community dominant graphics processing units (GPUs) for our deep neural networks (DNNs) to work,” he continued. “In fact, since we often work on introducing new layer definitions to accelerate learning and promote domain adaptation, CPUs are a better matched solution than GPUs. GPUs typically require a longer time for integrating the custom layer definitions into compliant libraries, and they sometimes do not permit certain classes of functions to be optimally implemented on platforms other than CPUs.”

**Adopting AI-Guided Radiology Screening**

In India—as in similar countries with large rural populations—there are generally not more than three accredited radiologists per million people. Health and wellbeing programs have successfully raised awareness of the importance of medical screening, but this, in turn, has increased the workloads for the few available radiologists, who are already overburdened with their screening responsibilities. As a result, radiologists are finding less time to perform detailed diagnoses of critical cases.

While AI deployments in physiology screening devices—including blood pressure (BP) monitors and glucometers—have significantly reduced the screening workloads for pathology labs, the power of AI-enabled screening has not yet been extended to radiological screening practices.

Project MIRIAD team members set out to explore the ways in which DNNs can enhance AI-enabled radiological screening techniques. The team evaluated the existing collections of ImageNet scale DNNs to be domain adapted with limited corpus of radiological images for screening of X-rays. The specific focus areas included breast cancer screening using bilateral mammograms and ultrasonography, lung cancer and disease screening, including tuberculosis, using chest X-rays, and traumatic bone injury detection.

![Figure 1. Indian doctor shares X-ray results with a patient.](image1)

Figure 2 shows some examples of medical images as captured by different types of diagnostic equipment.
Datasets
Debdoot notes that there are many publicly available datasets for medical image analysis that are made accessible through the National Institutes of Health (NIH)—The Cancer Imaging Archive (TCIA) and Grand Challenges in Biomedical Image Analysis. To train such large-sized networks for medical image analysis to achieve a precision comparable to what ImageNet provides for object recognition in natural images, each dataset would require some 14 million training images.

“This poses both a compute challenge and data challenge for these networks,” Debdoot said. “With a limited amount of domain- and disease-specific data at roughly 1,000 images per task, researchers have begun employing ImageNet pre-trained models—adapting them to the target domain for these new kinds of challenges. With the ever-growing collections of newly emerging DNNs, and the large number of challenges to be solved, as well as the multi-view nature of these diagnostic images (which need a multi-view extension to ImageNet class of DNNs), their deployment is a compute challenge at scale, requiring PFlops of compute power.”

“In the last 9 months,” Debdoot said, “we have completed some of the initial phases in the project. The first was the development of a deep neural compression engine for mammograms, which was published at CVPR2018. We are in advanced stages with benchmarking existing DNNs for extension to mammography and chest X-ray screening. This research spans 3 years starting in September 2017 and we have achieved a reasonable good performance with the first objective of having learning-based compression.”

Methods
Early works on medical image compression date back to the 1980’s with the impetus on deployment of teleradiology systems for high-resolution digital X-ray detectors. Commercially deployed systems during the period could compress 4,096 × 4,096 sized images digitally acquired at 12 bits per pixel (bpp) to store them in compressed form with 2 bpp using lossless arithmetic coding.

“Over the years,” Debdoot said, “JPEG and JPEG2000 were ingested reaching up to 0.1 bpp. Inspired by the reprise of deep learning based compression for natural images over the last two years, we designed and trained a fully convolutional autoencoder like model for diagnostically relevant feature preserving lossy compression of mammograms.”

“This is followed,” Debdoot continued, “by leveraging arithmetic coding for encapsulating high amount of spatial redundancy in features for further high-density code packing leading to variable bit length. We demonstrate performance on two different publicly available digital mammography datasets using peak signal-to-noise ratio (pSNR), structural similarity (SSIM) index and domain adaptability tests between datasets.”

At high density compression factors of >300× (0.04 bpp), the approach developed by Debdoot’s team rivals JPEG and JPEG2000 as evaluated through a Radiologist’s visual Turing test (See Figure 3).

**Figure 3.** Overview of the method for training the deep neural compression engine and deploying it.
Figure 4 shows the details of the architecture of the deep neural network used for high density compression of mammograms.

![Diagram of the architecture of the deep neural network for high density compression of mammograms.]

**Figure 4.** Technique for high-density compression of mammograms.

Figure 5 shows a magnified view comparison of the same region of interest in a mammogram under different types of compression techniques. While all have comparable quality factor measured in terms of SSIM and pSNR scores, it can be noted that the newly developed CAE consumes the least bits per pixel for storage favoring higher compression ratio than JPEG2000. The complete collection of quantitative comparison of quality factors under varying compression factors up to high density compression greater than 300x can be seen in this paper presented at CVPR2018.

![Magnified view comparison of identical areas of a mammogram under different compression techniques.]

**Figure 5.** Effect of compression on identical areas of a mammogram.
Discoveries and Outcomes

The discoveries and outcomes of the MIRIAD project look promising in terms of direct commercial applications. Companies developing medical imaging equipment and system manufacturers of computer-aided detection and diagnostic (CADx) equipment can take advantage of the architectures and methods for learning-based radiological image compression and efficient radiological image screening.

“These developments and the solution space,” Debdoot noted, “address the needs of hospitals and healthcare service providers, as well as government healthcare-providing agencies that mandate mass screening. Also, our work and the advanced AI techniques we developed can have a positive impact on medical imaging device manufacturers, system integrators, and vendors. CADx system developers, integrators, and vendors can also benefit from the techniques.”

Enabling Technologies

For training the deep neural networks, the team took advantage of the latest generation of scalable series of processors from Intel, including the Intel® Xeon® Platinum 8160 and Intel® Xeon® Gold 6128. The networks were implemented on PyTorch with Intel MKL and MKL-DNN bindings and Intel Python 3.5.

“We train our DNNs,” Debdoot said, “on the Intel AI DevCloud and the resources for accelerating PyTorch with MKL, MKL-DNN with Intel distribution of Python available through Intel AI Developer Program have helped us significantly.”

The team relied on the current generation of scalable processors, and communications with Intel engineers, to achieve maximum optimization, enabling the accelerated training of DNNs in mixed-precision datasets. Available software resources were also an essential factor in enabling the accelerated training, as well as access to a supportive network of developers willing to share expertise and provide guidance.

Offering advice to developers embarking on projects like Project MIRIAD, Debdoot encouraged them to explore the full range of Intel AI developer resources, Intel optimized software tools, and strengthen their knowledge of Intel reference platforms. “Reading more on the compute platform details and benchmarks that have been achieved,” Debdoot suggested, “can also be very helpful before building your own hardware platform.”

“AI today is really a set of tools. It allows us to sift through data in a much scalable way. It is scaling our intelligence up. We want our machines to personalize and change and adapt. The way we shop, the way we interact with each other, is already vastly being changed. It is kind of happening under the hood. Intel has a broad portfolio of products for AI. We start with the Intel Xeon Scalable processor, which is a general-purpose computing platform. It also has hooks in it to do very efficient inference for deep learning. In addition to that, we recently announced the neural network processor (or NNP) for maximum performance on training neural networks.

To build the best products, we need to have the best research. The neuromorphic computing is an example of this. This is basically taking much more close inspiration from the brain and how neurons communicate with one another. Concepts from that research can make its way into products and lead to the next product breakthrough.”

- Naveen Rao, Vice President and General Manager, Artificial Intelligence Products Group, Intel

Forward-Looking Development Perspectives

The insights gained so far during the ongoing Project MIRIAD work have led to recommendations that Debdoot and his team offer to the developer community engaging in similar projects. Optimizing performance, given the massive volumes of images needed to achieve effective results, is a major consideration. Debdoot noted that software optimization of high density tensor computations—such as those associated with the learning performed through deep neural networks—can be performance tuned with an in-depth understanding of the target compute platforms. To significantly increase performance and take advantage of parallelism, understand the processor cache size to tune the batch size, binding cores to memory channels, and correctly populating and distributing the available physical memory modules to match the number of available memory channels.
“We have observed,” Debdoot said, “that with these customizations, it is possible to achieve the highest throughput—close to the theoretical throughput limit of the CPUs.”

Understanding the hardware features is also important—including those available on the CPU being used on the host system’s motherboard—to achieve the best performance in high-density tensor computations.

“Having known earlier of the significant acceleration that can be achieved through core-binding, we would have procured CPU nodes with a DIMM configuration to harvest the maximum channel occupancy on the motherboard,” Debdoot said. “This way we could have saved significant time in getting started.”

### AI is Expanding the Boundaries of Computer Vision

Through the design and development of specialized chips, sponsored research, educational outreach, and industry partnerships, Intel is accelerating the progress of AI to solve difficult challenges in medicine, manufacturing, agriculture, scientific research, and other industry sectors. Intel works closely with policymakers, standards bodies, educational institutions, and enterprises of all kinds to uncover and advance solutions that address major challenges in the sciences.

### AI on the Edge

The OpenVINO™ toolkit helps accelerate visual computing solutions, providing another means for optimizing the performance of neural networks. The OpenVINO toolkit includes the Deep Learning Deployment Toolkit, which makes it possible to take full advantage of the Intel® architecture platform when developing deep learning solutions. The built-in model optimizer can import trained models from Caffe®, TensorFlow®, and MXNet®, converting and optimizing them for enhanced performance on the target hardware platform. The high-level API for the inference engine supports dynamically loaded plugins—supporting processors, graphic processing units, field programmable gate arrays, and Intel® Movidius™ Myriad™ vision processing units (VPUs)—to ensure the best performance without the need to maintain separate streams of code. In scenarios where network access may not be continually available, the OpenVINO toolkit streamlines design of vision solutions that can work effectively at the network edge.

Complemented by binary neural network architectures (BNNs) and DNN architectures, developers and system architects using the OpenVINO toolkit have a means for constructing small footprint solutions that can perform sophisticated AI functions in environments where the compute resources are limited.

“Developers are now using OpenVINO toolkit and other Intel® vision products to easily port computer vision and deep learning inference solutions from a wide range of common software frameworks, such as TensorFlow, MXNet, and Caffe, to Intel processor and accelerator technologies, including Intel CPUs, Intel integrated graphics, Intel® field programmable gate arrays (Intel® FPGAs), and Intel® Movidius™ Vision Processing Units (VPUs).”

- Adam Burns, Intel

The Intel® AI portfolio includes:

- **Intel Xeon Scalable processors**: Tackle AI challenges with a compute architecture optimized for a broad range of AI workloads, including deep learning.
- **Framework Optimization**: Achieve faster training of deep neural networks on a robust scalable infrastructure.
- **Intel Movidius Myriad VPU**: Create and deploy on-device neural networks and computer vision applications.
- **Intel AI DevCloud**: Offers a free cloud compute platform for machine learning and deep learning training and inference.
- **OpenVINO Toolkit**: Gives developers an accelerated method, including pretrained models and code samples, for implementing deep learning inference solutions using computer vision at the network edge.
- **Intel® Distribution for Python**: Supercharge applications and speed up core computational packages with this performance-oriented distribution.
- **Intel® Data Analytics Acceleration Library (Intel® DAAL)**: Boost machine learning and data analytics performance with this easy-to-use library.
- **Intel® Math Kernel Library (Intel® MKL)**: Accelerate math processing routines, increase application performance, and reduce development time.

For more information, visit the portfolio page.
SUCCESS STORY | Privacy Rights in the Ambient Computing Era

RESOURCES

Intel® AI Developer Program https://software.intel.com/en-us/ai

Inside Artificial Intelligence – Next-level computing powered by Intel AI ai.intel.com/


Intel® Software Innovator Program – Supports innovative, independent developers software.intel.com/intel-software-innovators/overview

OpenVINO™ toolkit – Develop multiplatform computer vision solutions software.intel.com/en-us/openvino-toolkit


Intel® Math Kernel Library software.intel.com/mkl

---


Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. Check with your system manufacturer or retailer or learn more at intel.com. Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked “reserved” or “undefined.” Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order. Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or by visiting Intel's Web site at www.intel.com.

Copyright © 2018 Intel Corporation. All rights reserved. Intel, the Intel logo, Movidius, Myriad, and Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

© 2018 Intel Corporation Printed in USA 0918/KP/MESH/PDF 337324-001US