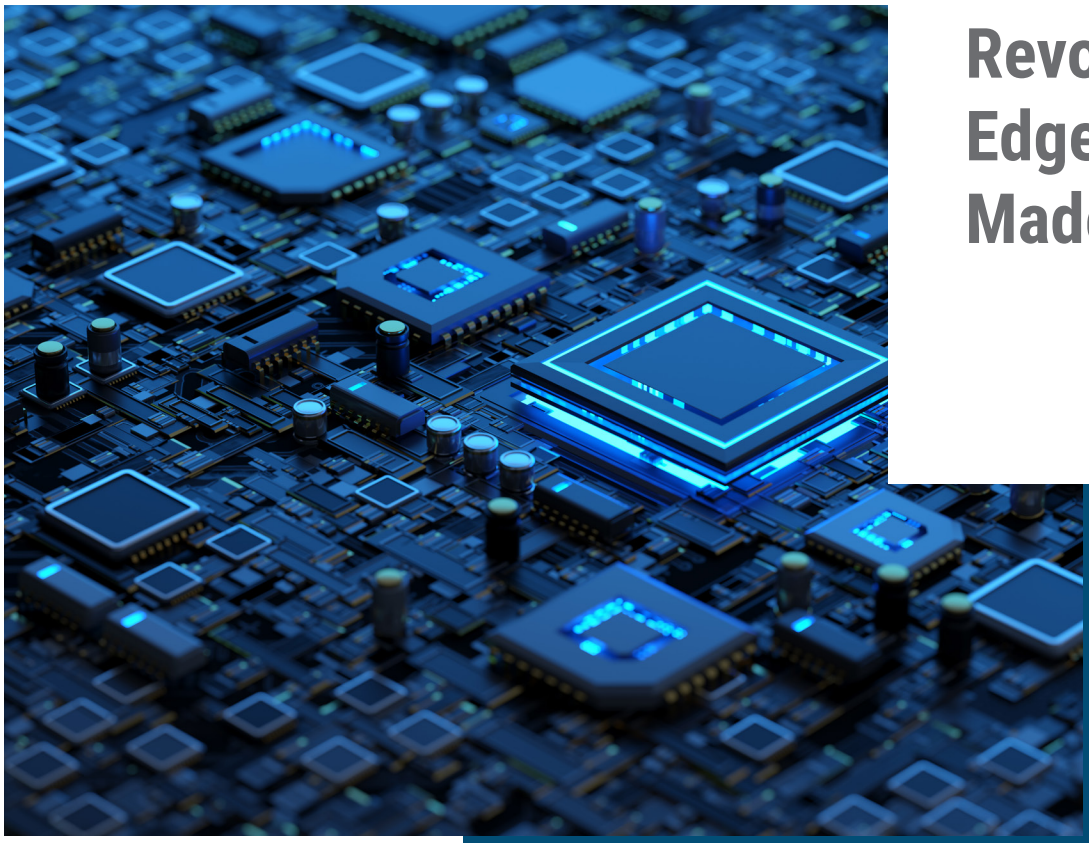
Exploring MemryX AI Accelerator Performance with Datature Vision Models

Datature

memryX



Revolutionary Edge AI Made Easy

About MemryX

MemryX, Inc. (memryx.com) is an AI semiconductor company headquartered in Ann Arbor, Michigan. MemryX delivers AI accelerator chips and software that combine industry-leading performance and accuracy with lower power usage and cost than today's leading solutions. With their power and cost advantages, MemryX solutions are well positioned for Edge AI applications including video security, autonomous driving, robotics, machine vision, and more.

This white paper was authored by Datature in collaboration with MemryX.

Contents

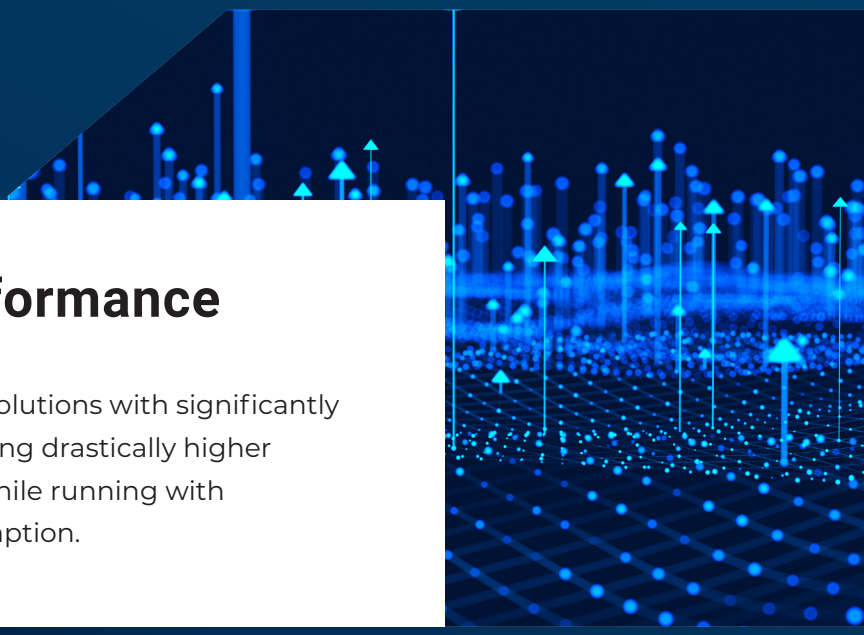
As Edge AI adoption increases in popularity, companies need faster, smaller, and more private solutions rather than bulky, inefficient, cloud-dependent setups. In this article, we showcase the performance of MemryX's MX3 M.2 AI Accelerator using Datature's Nexus platform.

MX3 M.2 Module

Specifications and Benefits

MemryX's MX3 M.2 module is provided in a compact M.2 form factor, delivering up to 24 TFLOPS (6 TFLOPS per MX3 chip) and storing up to 42 million 8-bit parameters while consuming typically 6 to 8 watts of power. The MX3 M.2 module outperforms solutions with significantly higher TFLOPS numbers, showing drastically higher performing inference speeds while running with significantly less power consumption. For example, the MX3 M.2 module is shown to beat the Nvidia Jetson AGX Orin 64GB in inference speed in this video, despite the Jetson consuming much more power and costing nearly 9 times more. Additionally, if even greater performance is required, MX3 chips are designed to cascade natively, allowing them to operate together as a larger and more powerful AI accelerator.

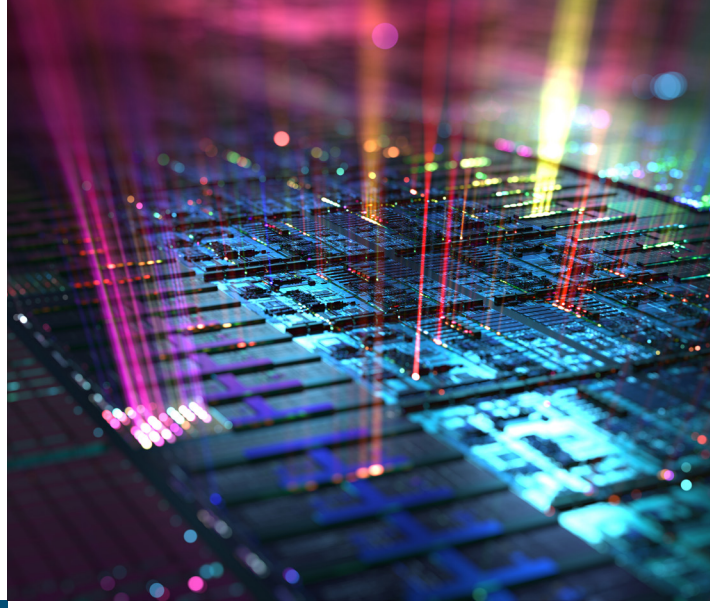
Beyond hardware, MemryX is dedicated to simplifying the complex process of deploying AI at the edge. Its co-designed hardware–software stack compiles trained models



Outstanding Performance

MX3 M.2 module outperforms solutions with significantly higher TFLOPS numbers, showing drastically higher performing inference speeds while running with significantly less power consumption.

The MX3 module is being applied in areas such as smart vision, industrial automation, mobility, and healthcare.



directly for high performance, removing the need for manual tuning or hardware-specific adjustments. Additionally, MemryX also provides comprehensive documentation, a Software Developer Kit (SDK) and hundreds of models and example applications on GitHub, so that developers can easily deploy, run, and benchmark their AI models. For more information, visit the MemryX Developer Hub.

Simplifying the process of deploying AI at the edge.

Overall, despite the compact form factor, the MX3 M.2 module delivers higher performance at a lower cost through its integrated hardware–software stack. This enables versatile integrations and simplified deployment at the edge. The MX3 module is already being applied in areas such as smart vision, industrial automation, mobility, and healthcare. When developing and deploying Edge AI, MemryX and Datature offer a platform that delivers strong cost efficiency, high performance, and scalability compared to other approaches.

Edge Deployment Use Cases

Beyond performance and efficiency, the MX3 M.2 module is designed for edge applications where speed, efficiency, and data privacy are critical.

Because all inference runs locally, there's no need to transmit sensitive data to the cloud. Unlike cloud-powered solutions, this reduces exposure to data breaches, helps meet regulatory standards, and enables reliable edge processing even in environments with limited or no internet connectivity.

The MXA M.2's rapid inference speed makes it exceptionally well-suited for real-time applications, where timely and reliable insights are critical. By delivering timely accurate insights, identifying changes allows you to promptly take corrective action, and optimize performance the moment events occur.

All benchmarks and results were developed and documented by Datature using MemryX hardware.

High bandwidth at-memory computing eliminates memory bottlenecks.

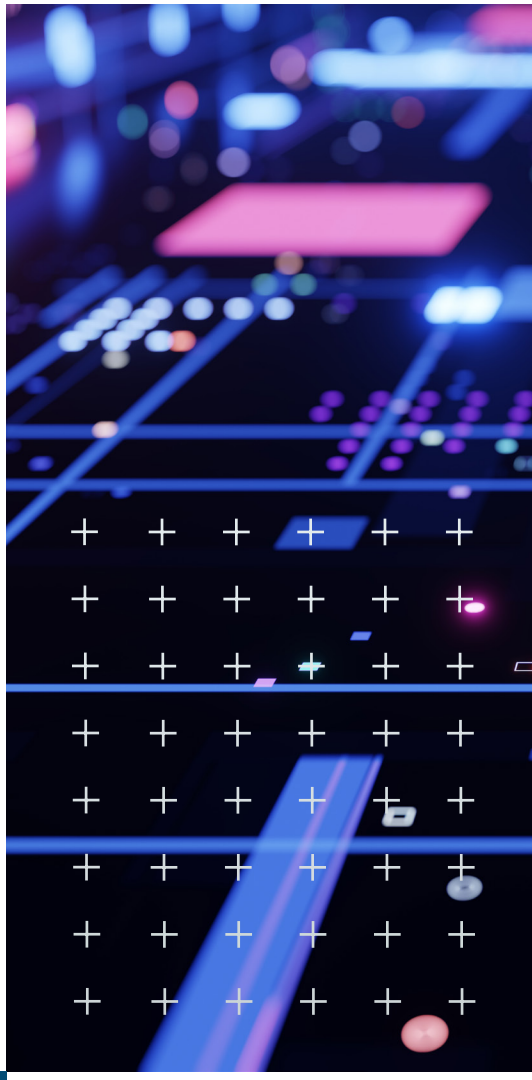
An Ideal Solution

The power efficient MX3 M.2 module typically operates at just 6–8 watts, as opposed to NVIDIA Jetson AGX's estimated 60 watts. This efficiency makes MemryX's solution ideal for applications such as remote surveillance systems and battery-powered robots and drones, where power-hungry hardware isn't an option.

Introducing
Datature's Solutions

MemryX offers a hardware platform to deploy trained models in edge environments that require the performance, efficiency, and privacy advantages of the MX3 M.2 module. However, MemryX only provides a platform to run a model efficiently. MemryX does not develop models or provide model training services. This is where Datature comes in. We enable users to train their own custom vision models on their data, achieving the accuracy you need for any visual insights for your use case, and empowering a seamless deployment of computer vision models.

Datature is an end-to-end MLOps platform built to simplify and accelerate the development of computer vision solutions. From dataset management and annotation to model training and deployment, users can manage their entire model training pipeline in a single, cloud-based platform that easily integrates into your current stack and infrastructure.



Empowering a seamless deployment of computer vision models.

Datature

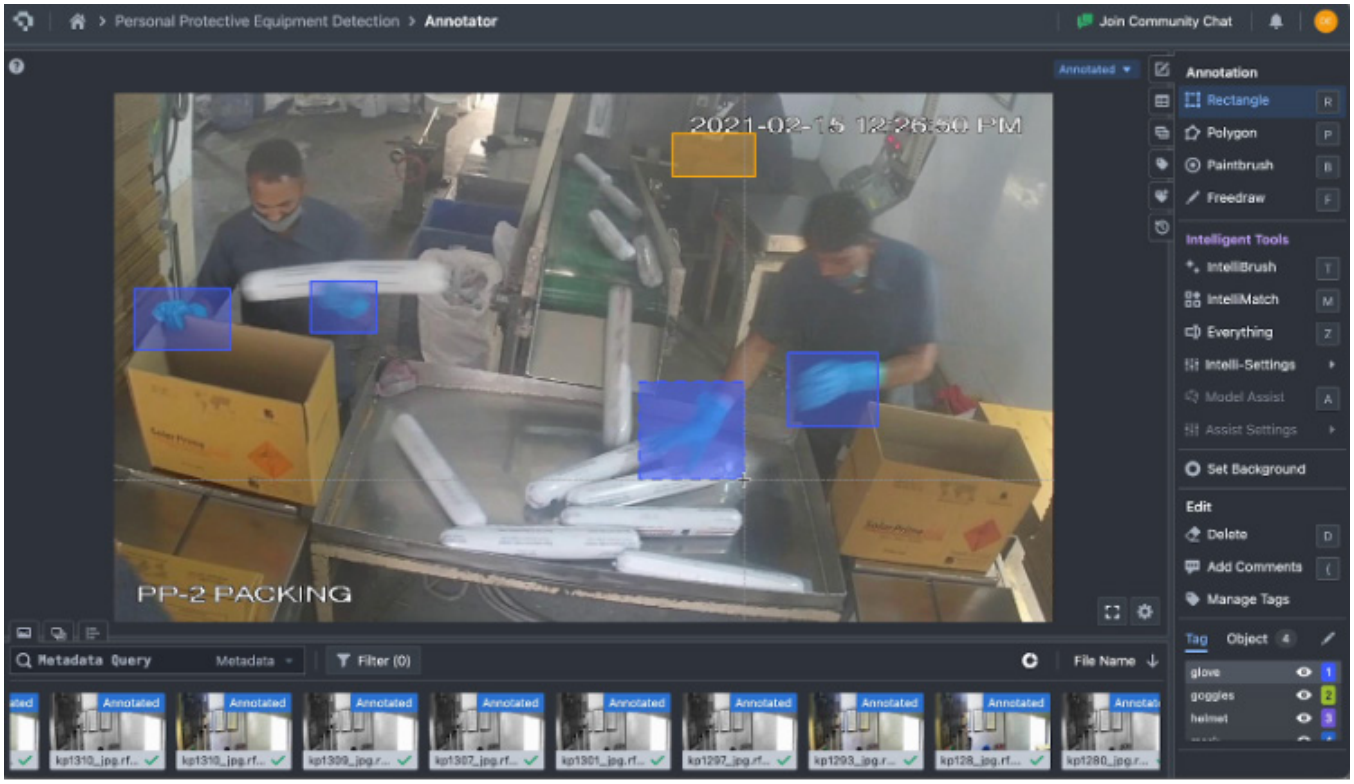
Performance, efficiency, and privacy.

In tandem, Datature and MemryX provide a comprehensive solution that allows you to not only rapidly build custom models from the ground up, but also deploy your solution at the edge with a performant, yet cost-effective solution. Ultimately, this allows you to easily adopt and scale AI solutions in your operations without significant compromise.

Datature's Solutions on MX3 M.2 Modules

Industrial Dataset and Annotations

This workflow begins with a personal protection equipment object detection dataset, using 3508 images sourced from the PPEs Dataset by Roboflow and licensed under CC BY 4.0. The dataset consists of detections of workplace safety compliances and violations, such as missing goggles, gloves, and protective shoes. To learn more about importing your own data and annotations, or annotating your data on the Datature Platform, please refer to our documentation and tutorials.

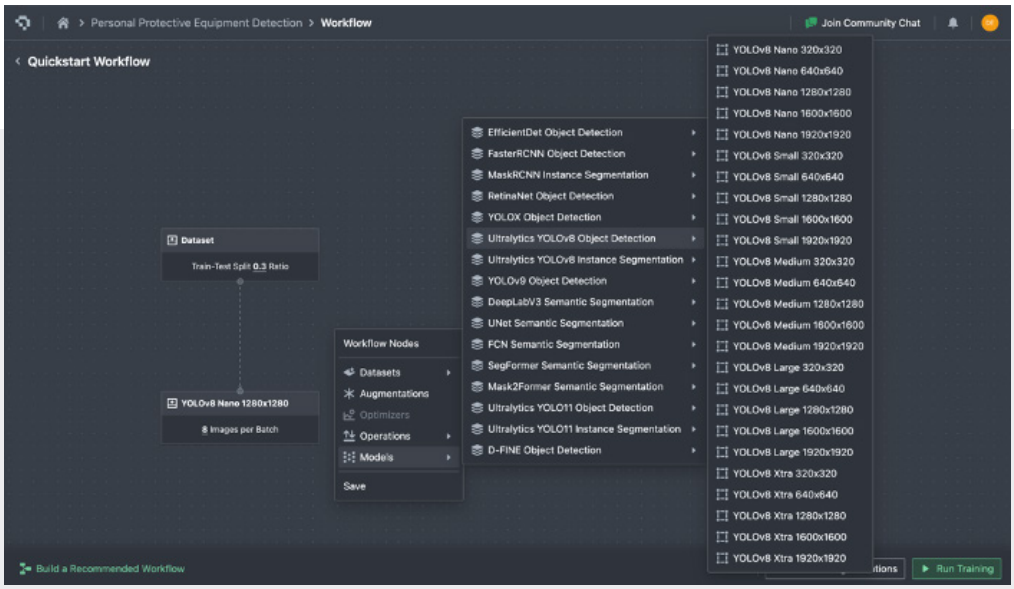


A Sample from the PPE Dataset on Datature's Interactive Annotator

Model Selection

The YOLOv8 Nano model is chosen for this use case as it is small, fast to train, and accurate, perfect for an edge use case with MemryX's accelerator. The model has been trained and tested on this dataset, evaluating the MemryX's MX3 M.2 module's inference speed on a realistic industrial application.

Beyond YOLOv8 Nano, Datature offers a diverse catalog of models for training and iteration, allowing users to select the ones that best align with their specific tasks, ranging from lightweight models optimized for edge computing to larger more powerful models.



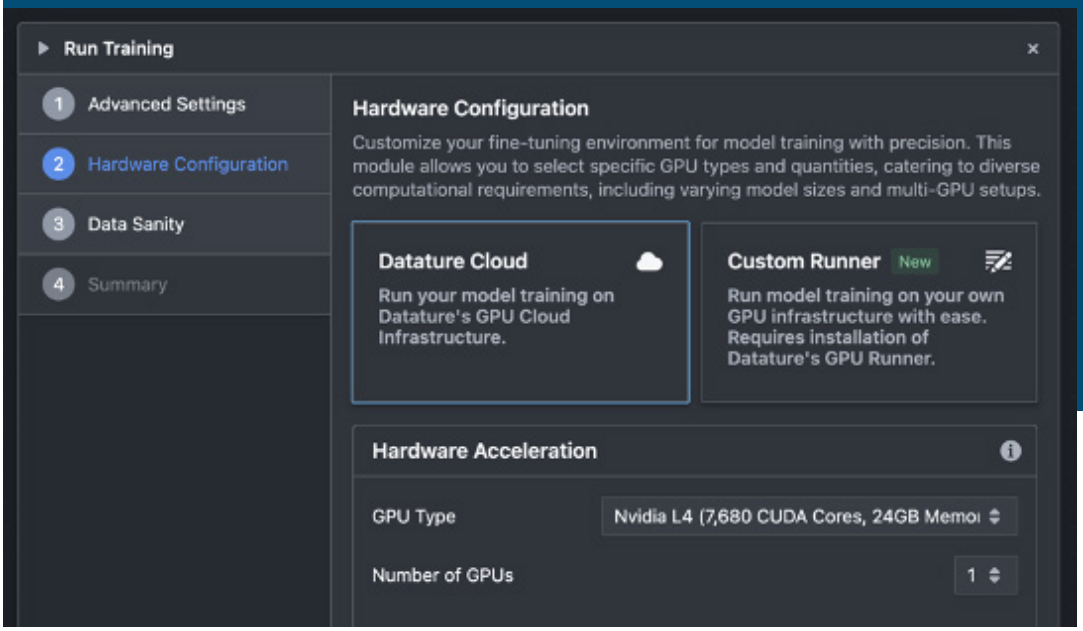
Model Selection Interface for Object Detection Workflow (YOLOv8 Expanded)

Datature offers a diverse catalog of models for training and iteration.

Training Parameters

After the annotation phase, the selected YOLOv8 Nano model begins training through Datature's fast, iterative model finetuning pipeline. The model is trained using its default pretrained weights, with a batch size of 8 for 5,000 steps on a Nvidia L4 GPU using the Adam optimizer with a learning rate of 0.001.

Datature offers access to a variety of dedicated GPUs to suit different training workloads and performance requirements. Users have the option to train models on Datature's dedicated GPU infrastructure or use their own local or cloud-based GPU runners.

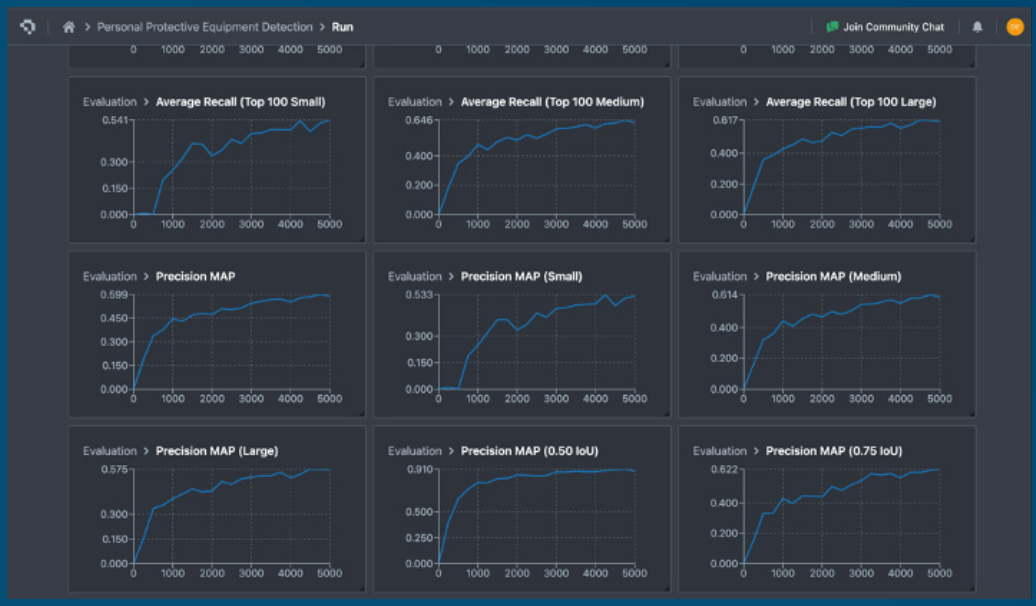


Training Environment Hardware Configuration Interface

For an overview of our training capabilities, available GPU options, and model offerings, take a look at our documentation here: Datature Model Training Documentation.

Training Results and Export

Datature's platform provides real-time visualization of training and validation metrics for evaluation, supporting iterative model development. After training, the model achieves a strong 0.90 mean average precision (mAP) at an IoU threshold of 0.50 and a mAP of 0.62 at a stricter IoU threshold of 0.75. The metrics show the model is robust in detecting safety equipment compliances and workplace hazards in new images.



Datature's Provided Evaluations Metrics on a Held-out Validation Set

Following the completion of training, you can export your trained model through Datature's Artifacts interface, providing users with a straightforward way to export models, or deploy on the cloud through Datature Cloud. Datature and MemryX both support the ONNX model format, a common industry standard. As such, we export our model in ONNX format and transfer the model to our edge device.

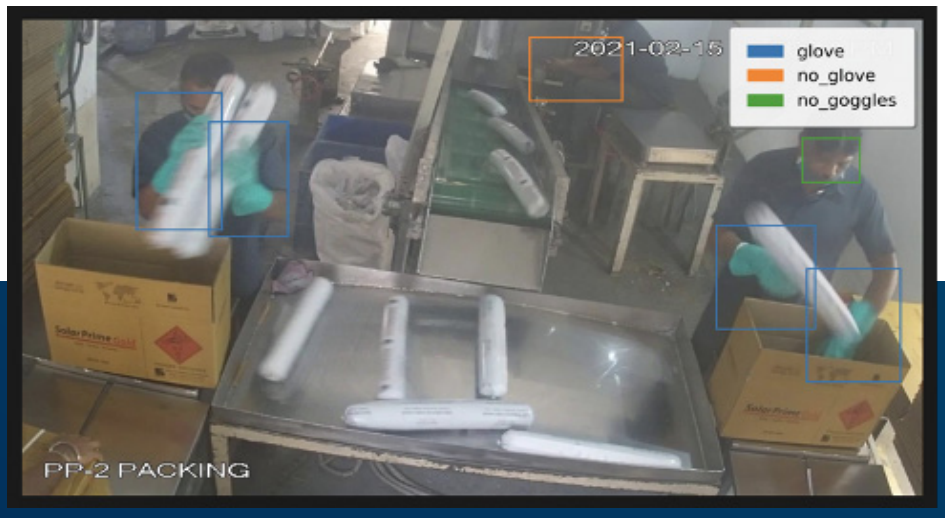
Importantly, all models trained on Datature's platform remain the user's own intellectual property, allowing you to deploy at the edge without worrying about unforeseen inference costs. To learn more about deployment options or IP ownership, please reach out to our sales team here.

Model Deployment and Performance

To deploy the model on MemryX's MX3 module, the ONNX file is compiled into a Dataflow Program (DFP) using the MemryX **Neural Compiler**. The process is as simple as running a single CLI command:

mx_nc -m datature-yolov8n.onnx -c 4 --autocrop

After running the command, the compiler generates a .dfp file and corresponding {model_name}.post files in the same directory as the ONNX model. These files are used to run the optimized model inference. Shown below are sample output images of protective equipment compliances and violations, indicating robust detections with accurate class assignment.



Samples of Inference Output Using Our Trained Model on a Held-out Test Set

After benchmarking the model on multiple industrial images, the difference in performance between CPU and MemryX's MX3 accelerator (labeled MXA in the image) becomes apparent, with the MXA achieving over 18× faster end-to-end inference time and …× higher throughput (FPS).

```
CPU Inference time: 1579.3 msec
MXA Inference time: 87.0 msec
```

Comparison Between CPU and MemryX Accelerator Inference Speed

Conclusion

By combining Datature's all-in-one vision AI platform and MemryX's AI accelerator chips and edge-focused development ecosystem, users can effortlessly train and deploy AI models at the edge, without compromises in accuracy and efficiency. When developing and deploying Edge AI, MemryX and Datature offer a platform that delivers strong cost efficiency, high performance, and scalability when compared to other approaches.

Datature and MemryX are ready to collaborate and develop tailored solutions for your specific use case. Please feel free to reach out to us **on our website**.

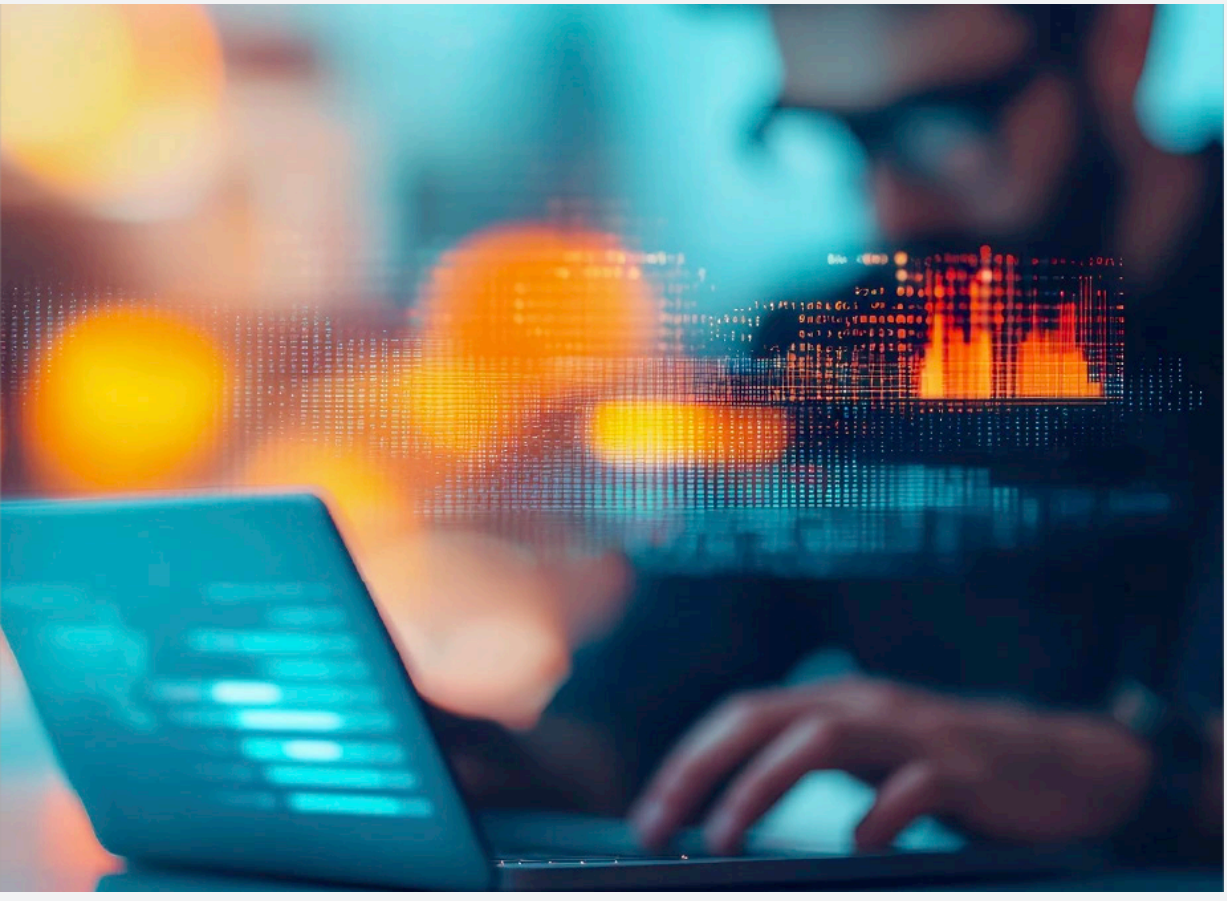
Visit **https://datature.com** for more information.

Our Developer's Roadmap

As the demand for privacy-focused, real-time inference grows, Datature and MemryX are at the forefront of innovative Edge AI deployment. Datature is pleased to partner with MemryX to support the growing adoption and advancement of visual edge computing.

If you have questions, feel free to join our Community Slack to post your questions or contact us if you wish to train and deploy your own vision model on Datature Nexus.

For more detailed information about model functionality, customization options, or answers to any common questions you might have, read more on our Developer Portal.



Datature

535 Mission St
San Francisco, CA 94105
United States
datature.io/

memryX

110 Miller Ave Suite 1
Ann Arbor, MI 48104
+1 734-926-9995
memryx.com