

AHEAD

How to Choose:

An AI Inference Platform for
Your Medical Application



While artificial intelligence is likely to become a foundational technology over the next decade, there are still many challenges that medical technology developers face when deploying AI. One obstacle is that AI technologies are evolving at a rapid pace, so it's difficult to design and build AI applications before they become obsolete.

Choosing the right machine learning platform can enable you to deliver an AI-enabled medical application more quickly and cost-efficiently to market. However, this requires a deep understanding of the hardware and software requirements for your specific use case.

In this article, we'll discuss some key considerations for efficiently designing an AI-powered medical solution. We'll also give an overview of three popular edge AI hardware platforms: NVIDIA Clara, Google TensorFlow Coral, and Intel OpenVINO.

What Are Your Medical Application Requirements?

Since balancing performance and cost is crucial when developing a complex, AI-powered medical application, it's important to choose a machine learning platform based on the overall software and hardware requirements.



HERE ARE SOME KEY CONSIDERATIONS WHEN DESIGNING AN AI-POWERED MEDICAL DEVICE:

Use case: Are there specific requirements for your solution that pre-trained models could be used as software building blocks to streamline development? For example, for patient monitoring applications, there are body positioning and gesture recognition models available.

Certification requirements: What are the certification requirements for the device's specific use case?

Machine learning models: Can you use pre-trained models and is there a particular framework you plan to use, such as NVIDIA Clara? How many models will you use and what size are they?

Inference processing: Can the application leverage onboard AI processing or is an accelerator card required?

Environment: Is the inference component being racked or installed directly within a medical device? Does the device need to be regularly disinfected or wiped down after subjecting it to operating room use or bodily fluids?

Life cycle: How long does the product need to be on the market before it is refreshed with next-generation hardware? There are long-life components available for extended-life platforms.

Top Machine Learning Inference Platforms to Consider

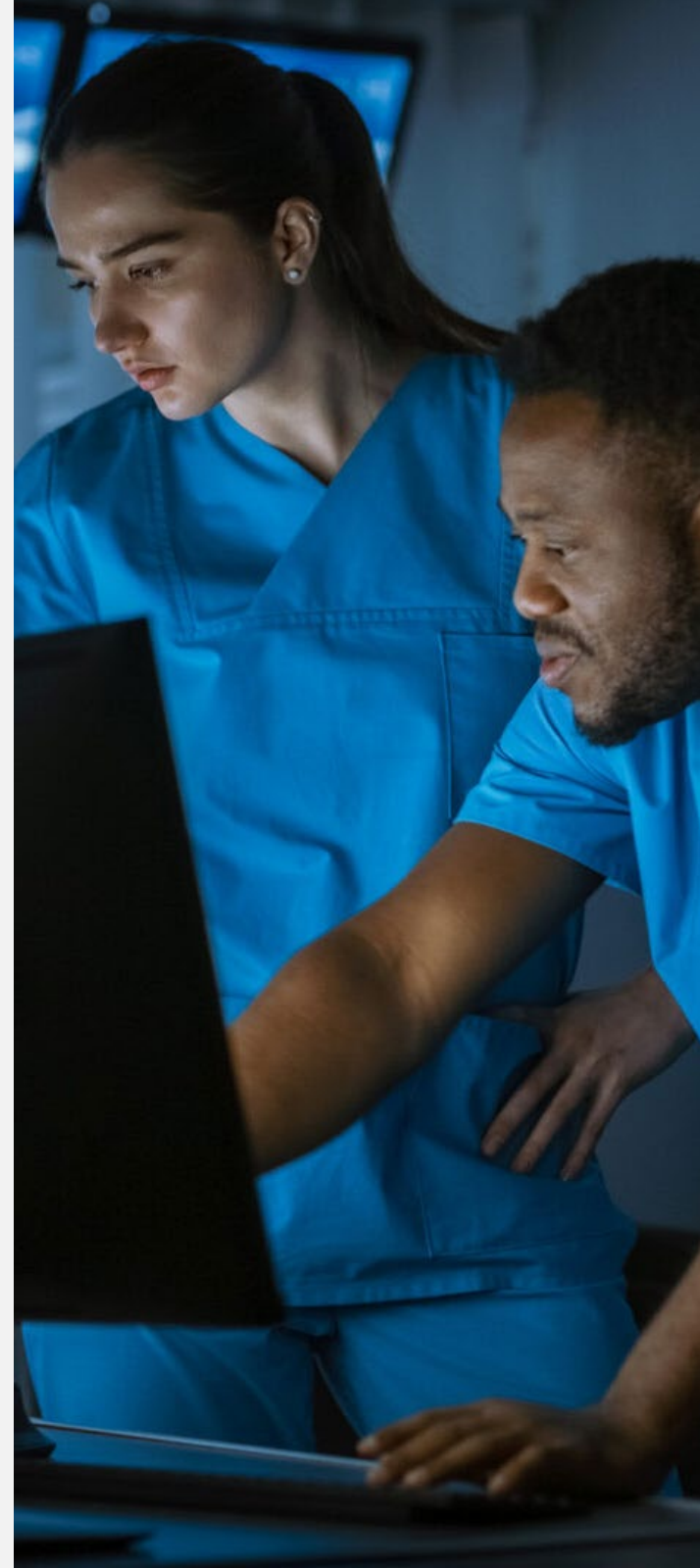
Once you've determined the preliminary considerations for developing your medical solution, you can choose a compatible machine learning platform. Here's an overview of three popular machine learning platforms that can be used to deploy AI-powered medical devices.

NVIDIA CLARA

[NVIDIA Clara](#) is an AI framework designed specifically for healthcare and life sciences. The framework includes the Clara Train SDK for training ML models, the Clara Holoscan SDK for building AI applications, and NVIDIA Clara SDKs for deploying advanced computations.

In addition, NVIDIA provides predefined building blocks to accelerate the launch of AI-powered medical devices. These specific deployment options for healthcare—Clara Holoscan, Clara Imaging, and Clara Guardian—make it easier to deliver certain use cases to market without designing and integrating the hardware and software from scratch.

Technology developers that plan to use NVIDIA graphics cards or deploy an embedded system with Jetson should use NVIDIA Clara. Pre-trained models and predefined hardware specifically for healthcare use cases also make NVIDIA Clara a great choice for quickly delivering medical devices to market.



AI Optimized Hardware with AHEAD

As you can see, building a medical application using a particular AI platform is a critical, yet complicated decision. It's imperative to identify the right hardware componentry based on application requirements and then determine which platform will perform optimally on that target device.

Partnering with a trusted hardware specialist like AHEAD can help you choose the right hardware and software combination for your unique AI-powered solution. This expert guidance will enable you to deliver innovative AI applications faster and cheaper than navigating the evolving machine learning technology landscape on your own.

Learn more about delivering AI-powered medical devices in our latest whitepaper, "The Shift to Smart Hospitals: How You Can Deliver AI-Driven Devices Effectively Using Pre-Engineered Platforms."

GOOGLE TENSORFLOW CORAL

[Google TensorFlow and Coral](#) is an edge hardware and software platform for ML inferencing. Coral is based on the TensorFlow machine learning framework and allows you to easily execute models on the Edge TPU coprocessor. There are also [pre-compiled models](#) that are trained and ready to deploy to Edge TPU devices.

The Coral platform works best for AI applications that do not require a lot of processing power because the

Edge TPU coprocessor is designed to be low power and cost efficient. Since Coral runs right atop the accelerator, it can be much cheaper than using a graphics card. For more advanced AI applications it might be better to choose a different edge hardware solution.

INTEL OPENVINO

[Intel OpenVINO](#) is an open-source machine learning platform optimized for edge inferencing on [Intel Movidius processors](#). Along with pretrained models from the [Open Model Zoo](#), the OpenVINO toolkit includes a model optimizer and inference engine.

You should choose OpenVINO if you want to deploy an AI application on Intel hardware. OpenVINO supports multi-device execution, which enables simultaneous inferencing of the same model on several Intel devices in parallel. For example, the model could be inference against an Intel GPU, CPU and CPU accelerator.



AHEAD builds and manages digital platforms that power the most successful organizations in the world. Our consultative approach, unmatched engineering, and innovative solutions combine to accelerate the impact of technology for every client we serve.

National Hubs

CHICAGO

401 Michigan Ave.
#3400
Chicago, IL 60611

NEW YORK

500 Fifth Avenue
Suite 1500
New York, NY 10110

ATLANTA

1117 Perimeter Center
W406
Atlanta, GA 30338

SAN FRANCISCO

2000 Crow Canyon Place
Suite 250
San Ramon, CA 94583