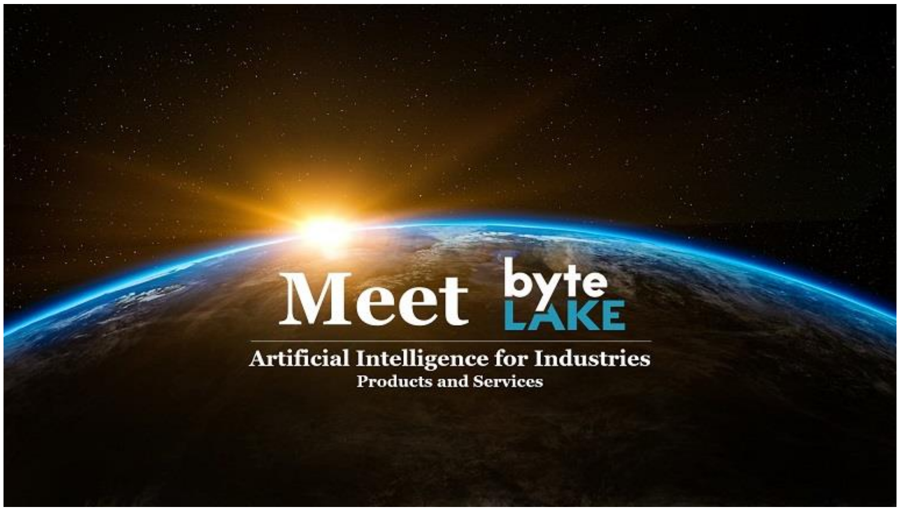


byteLAKE’s Cognitive Services

MAXIMUM PERFORMANCE

byteLAKE's Cognitive Services have been optimized for Edge AI deployments, providing clients with maximum performance and cost-efficiency. We are pleased to present a comprehensive report that summarizes the exceptional results achieved and highlights the numerous benefits that our Cognitive Services product delivers.

Artificial Intelligence

AI-accelerated CFD Simulations

AI-assisted Visual Inspection

AI-powered Microphones

AI Products for Manufacturing, Automotive, Restaurants, Paper, and Chemical Industries.

Complex Tasks Automation

byteLAKE

Europe & USA

+48 508 091 885
+48 505 322 282
+1 650 735 2063

Foreword

It is our utmost pleasure to present to you this comprehensive report on the exceptional results achieved with our Cognitive Services product at byteLAKE. Our team has worked tirelessly to ensure that we deliver maximum performance to our clients, and we are proud to say that our efforts have paid off.

At byteLAKE, we understand the importance of innovation and staying ahead of the curve in the ever-evolving world of technology. That is why we have optimized all our AI models within Cognitive Services to the latest Edge AI technologies. By doing so, we not only deliver maximum performance, but also cost-efficiency for our clients.

This report summarizes the outstanding results we have achieved, highlighting the numerous benefits that our Cognitive Services product delivers. From improved accuracy to faster processing times, we are confident that our clients will be impressed with the results we have achieved.

As startup founders, we are always looking for ways to innovate and improve our products, and our Cognitive Services offering is no exception. We will continue to push the boundaries of what is possible, ensuring that our clients receive the best possible service, results and products.

So, without further ado, we invite you to enjoy this report and learn more about the exceptional results we have achieved with our Cognitive Services product at byteLAKE across industries.

Sincerely,

Marcin Rojek, Mariusz Kolanko, byteLAKE’s founders.

Executive Summary

byteLAKE, in collaboration with Intel® and Lenovo, offers a scalable and readily deployable reference configuration comprising both hardware and software for byteLAKE's Cognitive Services. This cutting-edge collection of AI pre-trained models is designed to streamline quality inspection and data analytics, enabling industries to make better decisions, reduce downtime, and increase profitability.

When integrated with cameras, **byteLAKE's Cognitive Services can inspect products and processes, automatically identifying any quality issues, such as scratches, dents, wrong or missing labels**, in metal components, boxes, or any products. In the paper industry, byteLAKE's Cognitive Services can **monitor the papermaking process**, and in the automotive industry, these services can **analyze the sound produced by car engines, bearings, or suspension, delivering valuable quality assessments**.

By connecting Cognitive Services to sensor data, industries can leverage the technology to **predict maintenance requirements, optimize processes, and make better decisions by analyzing historic and live data from various sources**. For instance, byteLAKE's Cognitive Services can help manufacturers **detect early signs of wear and tear in machinery or equipment, reducing the risk of unplanned downtime**.

Through its deployment in restaurants, byteLAKE's Cognitive Services use cameras to recognize items selected by clients and transmit the list directly to the point of sale or cashier's machine, improving the quality of service and reducing waiting times. The software also automates table reservations and assists the kitchen staff by sending reminders when a table needs cleaning, notifying who needs a check, or assisting with drink refills or orders.

byteLAKE's Cognitive Services are now **available as a complete solution**, easily deployable in environments ranging from restaurants to manufacturing facilities, where edge implementation challenges such as long-life reliability, enhanced security, and manageability features are crucial. The reference configuration includes a PC, **Lenovo's ThinkEdge SE50, equipped with Intel® Core™ i7 processor and byteLAKE's Cognitive Services optimized to make the most of the hardware for the client's selected scenario**.

By leveraging this solution, clients can benefit from the ease of installation, long-term reliability, and enhanced security and manageability features, resulting in improved operational efficiency, reduced downtime, and increased profitability. Furthermore, optimizations summarized in this report translate into significantly improved data processing speeds, resulting in faster decision-making and overall increased productivity. Also, byteLAKE’s software optimizations maximize the utilization of embedded hardware components, resulting in reduced overall hardware costs.

Technical Appendix to the Executive Summary

byteLAKE's Cognitive Services are designed to deliver optimal results and maximize performance, cost efficiency, and highly performing self-learning AI models. To achieve this, byteLAKE has carefully selected the Lenovo ThinkEdge SE50 with 12th Gen Intel® Core™ i7 or newer and 8, 16 GB, or higher RAM as the ideal reference hardware configuration for Cognitive Services deployments. This configuration helps meet the challenges of edge implementations in enterprises by providing extended temperature support, long-life reliability, as well as enhanced security and manageability features.

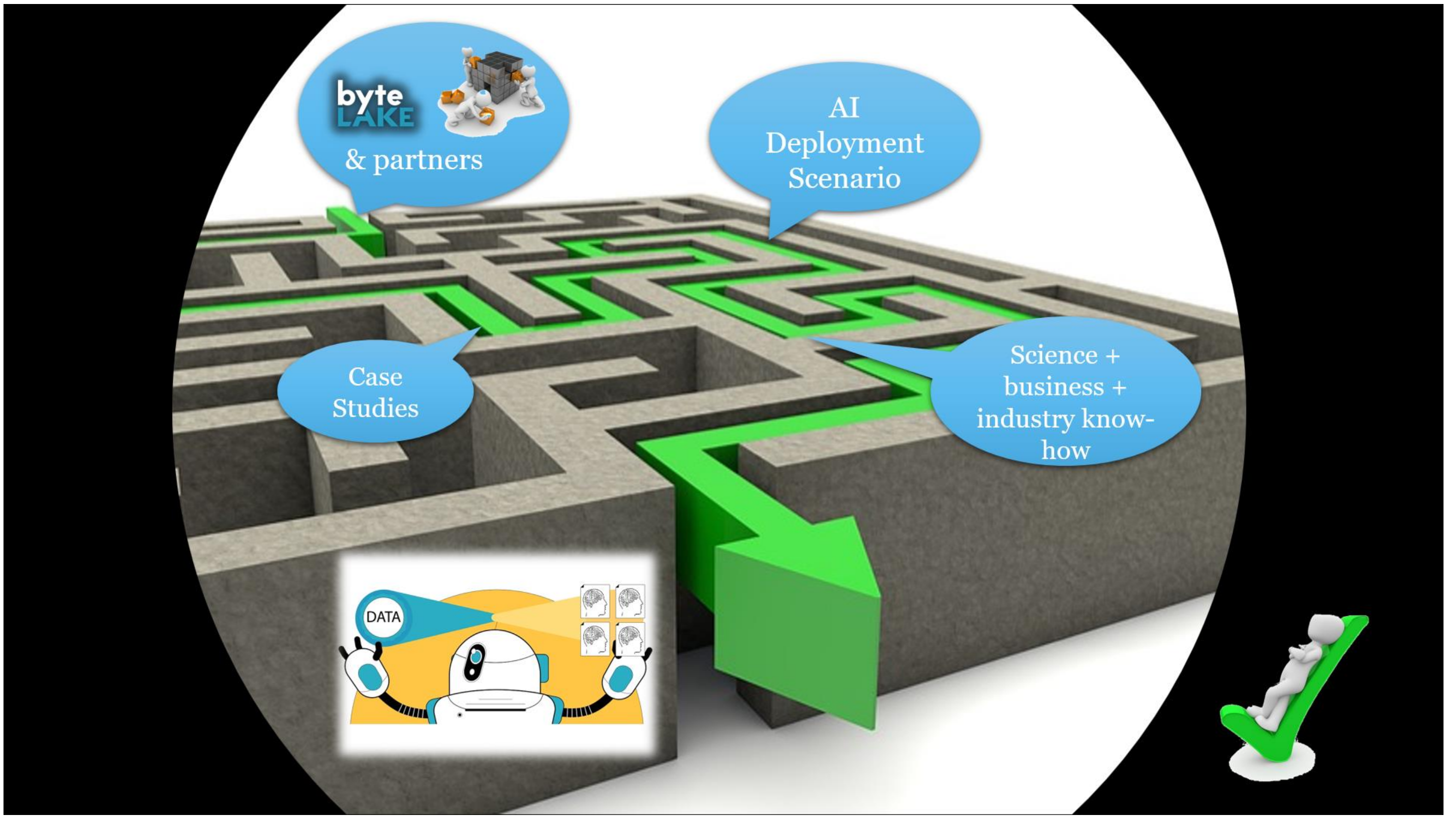
Furthermore, byteLAKE has optimized their Cognitive Services to ensure that all embedded hardware components are fully utilized, resulting in AI-assisted Visual Inspection related algorithms being accelerated by 25 times and AI-powered Big Data Analytics related algorithms being accelerated up to 22 times, depending on the scenario and use case.

By leveraging this reference configuration, clients of byteLAKE's Cognitive Services can benefit from maximum performance and cost efficiency while ensuring optimal results and self-learning of AI models.

byteLAKE’s Cognitive Services

byteLAKE's Cognitive Services represent a comprehensive collection of Artificial Intelligence (AI) models, designed to address the needs of Industry 4.0 and Restaurants. Our AI models have been expertly designed and pre-trained to focus on specific tasks, resulting in maximum accuracy and efficiency.

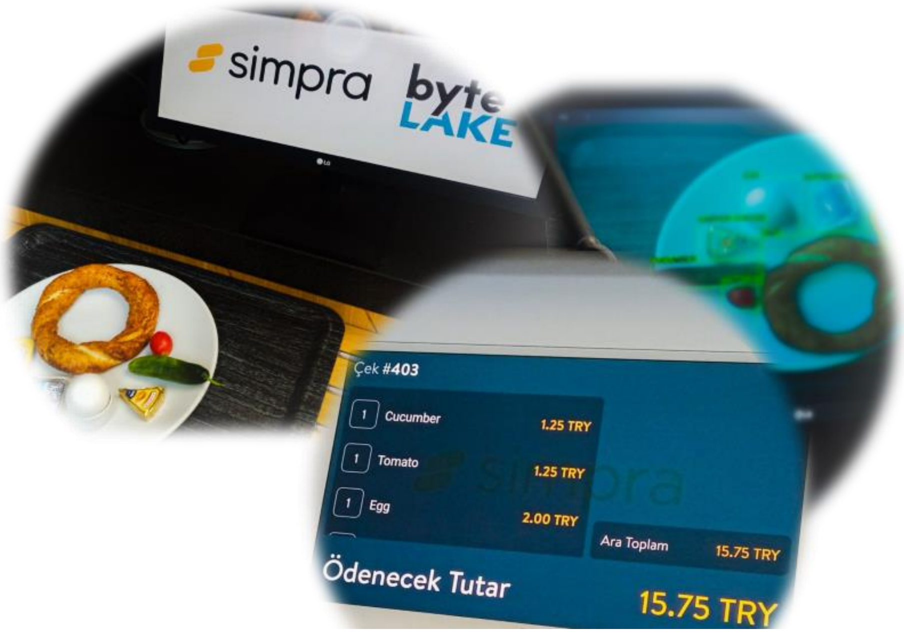
At byteLAKE, we understand the importance of collaboration and bringing together the best minds in the industry. That is why we have strategically partnered with leading industry experts and talented researchers to combine human knowledge, expertise, and know-how with the latest AI algorithms and technologies.



Our Cognitive Services offering includes AI models for **AI-assisted Visual Inspection in manufacturing**, enabling real-time quality monitoring at assembly lines and production.

Trigger an alarm

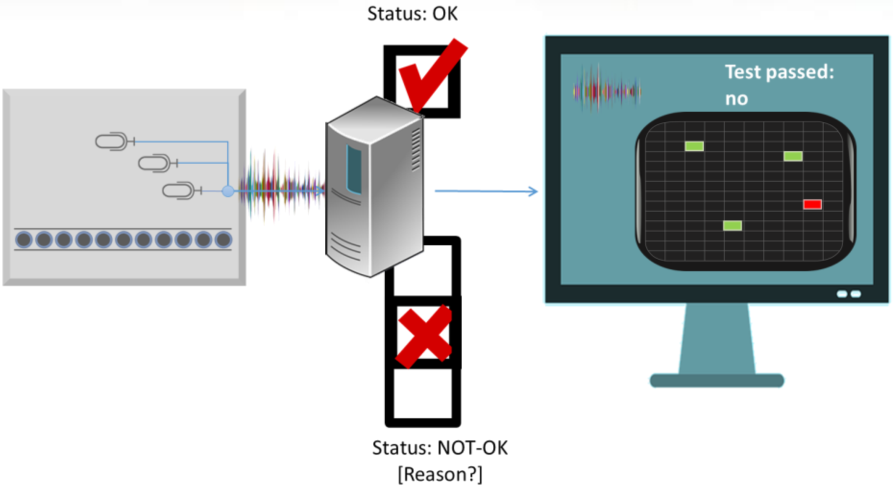
For Restaurants, our AI models facilitate self-check-out services and provide advanced analytics, such as notifying the kitchen staff when a customer needs a drink refill or automating table reservations by analyzing occupancy.



Çek #403

1	Cucumber	1.25 TRY
1	Tomato	1.25 TRY
1	Egg	2.00 TRY
	Ara Toplam	15.75 TRY
	Ödenecek Tutar	15.75 TRY

Furthermore, our Cognitive Services product includes **AI-powered Big Data analytics, which is especially valuable in the automotive industry**. By utilizing microphones and other IoT sensors to record sound produced by car engines, bearings, and suspension, our solution can **effectively monitor quality, detect issues, and offer predictive maintenance suggestions**. Our AI-powered analytics also support decision-making processes and help **prevent downtime in production**.



Our Cognitive Services have also been successfully implemented in the pulp and paper industry. Through the utilization of cameras and pre-trained AI models, we are able to monitor the papermaking process and prevent unforeseen downtimes and catastrophic events that can result in substantial costs and production delays. One example of our technology in action is our Wet Line Detector, which automates the process of Waterline management. By supplying crucial parameters such as the wet line's position and size, as well as information regarding its presence, our technology can be integrated with other IoT sensors and time series data analytics to improve overall efficiency.

Trigger an alarm
Waterline
END OF LINE
8000MM
BEGINNING OF LINE
6000MM

We take pride in our commitment to continued improvement and innovation, and we are thrilled to offer this cutting-edge technology to our clients. Our AI models are designed to combine the best of both worlds, leveraging the strengths of both humans and machines to deliver exceptional results.

In summary, byteLAKE's Cognitive Services offering is an indispensable tool for businesses seeking to improve efficiency and productivity. With our innovative AI models, manufacturing clients can achieve accurate real-time monitoring, predictive maintenance, and advanced analytics, leading to **improved decision-making, better quality inspection at lower cost, avoid downtime in production** and greater success. For restaurant owners byteLAKE's AI delivers a range of benefits, including self-checkout capabilities, reduced waiting times, and improved customer assistance.

Edge AI

At byteLAKE, we place Edge AI, or AI at the Edge, at the center of all our products. This means that our software solutions are specifically designed to **process data where it is produced**, without the need to transfer it to third-party solutions. Our products are entirely independent of Cloud services, internet connection, and internet bandwidth capabilities, and can function offline. This design avoids any budgetary surprises and allows our clients to manage costs effectively.

Edge AI deployment also brings numerous benefits, including scalability, which decentralizes AI services, making it **easier to expand IoT ecosystems**. Modern, low-power, high-performance, small form factor accelerators enable real-time AI experiences, while deploying AI directly on the device eliminates round-trip latencies, enabling faster responses.

Intermittent connectivity-related issues are also resolved as data does not need to be sent from the device to external AI services and wait for results. This approach also helps reduce total cost of ownership as AI-enabled devices pre-process data and send only results to external services instead of raw data.

Another advantage of Edge AI is that data can stay locally on the device, providing a selective option for sending data to external storages. Overall, deploying AI at the Edge helps optimize operations, reduce costs, and enhance the user experience, making it a compelling solution for businesses seeking innovative and efficient ways to manage data.

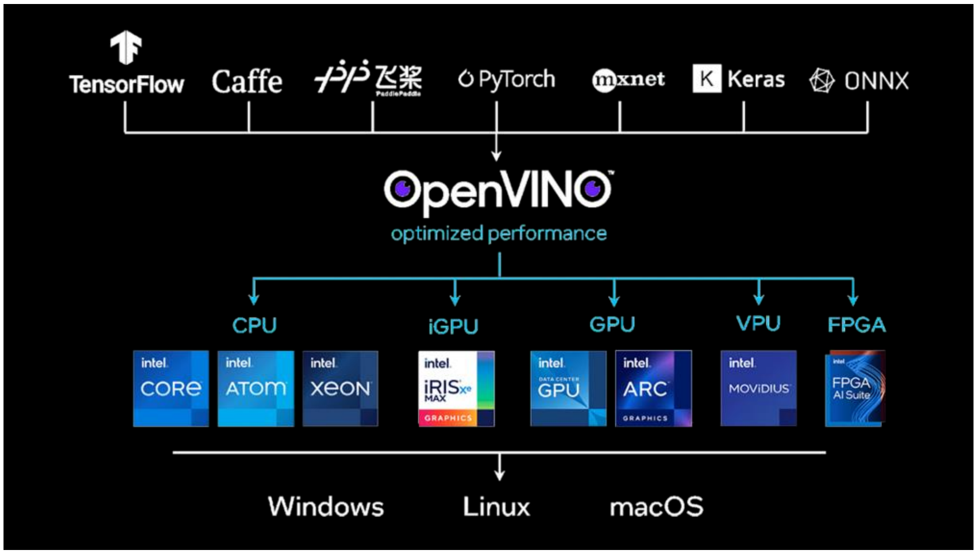
Innovating for the Future: byteLAKE's Commitment to Continued Improvement

At byteLAKE, we have established strategic partnerships with top technology providers such as Lenovo and Intel® to ensure that our software products are always up-to-date and optimized for the latest technologies. These partnerships also enable us to offer a comprehensive range of hardware components, allowing us to address the needs of various clients.

Our clients can expect a wide range of benefits. These include access to the latest technologies, enhanced performance and efficiency, and a reduced total cost of ownership. Our clients can also expect unparalleled support, ensuring that they always receive the assistance they need when using our software solutions. By partnering with industry-leading providers, we can continue to deliver innovative and efficient solutions that drive growth and success for our clients.

Edge AI powered by Intel® Software

Intel® OpenVINO™ toolkit helps run inference on a range of compute devices. This toolkit is designed to accelerate the development of machine learning solutions. Based on convolutional neural networks (CNNs), the Intel® Distribution of OpenVINO™ toolkit shares workloads across Intel® hardware (including accelerators) to maximize performance.



Besides OpenVINO™, Intel® also offers a number of optimized software solutions that help accelerate time-to-insight with machine learning and deep learning, using Intel® hardware.

Some of these include:

- Intel® Distribution for Python (with Intel® optimized scikit-learn, XGBoost and more)
- Intel® Distribution of Modin
- Intel® Optimization for TensorFlow
- Intel® Optimization for PyTorch
- Intel® Low Precision Optimization Tool
- Model Zoo for Intel® Architecture

Learn more: https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/overview.html and https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI/Intel-Optimized-AI-Software/post/1335731.

Intel® DL Boost Vector Neural Network Instructions

Based on Intel® Advanced Vector Extensions 512 (Intel® AVX-512), the Intel® DL Boost Vector Neural Network Instructions (VNNI) delivers a significant performance improvement by combining three instructions into one—thereby maximizing the use of compute resources, utilizing the cache better, and avoiding potential bandwidth bottlenecks and further unleashes the performance of computations and significantly speeds up the inferencing with INT8 models.

Learn more: https://www.intel.com/content/www/us/en/developer/articles/technical/introduction-to-intel-deep-learning-boost-on-second-generation-intel-xeon-scalable.html and https://www.intel.com/content/www/us/en/developer/articles/guide/deep-learning-with-avx512-and-dl-boost.html.

Edge AI enabled by Lenovo ThinkEdge solutions

The Lenovo ThinkEdge portfolio offers a full range of client and server edge infrastructure solutions, software and services. Lenovo's ThinkEdge portfolio has purpose-built devices designed to be networked on-premises. Or embedded in solutions, to give you the advantage in performance, security, and scalability.

ThinkEdge SE10

ThinkEdge SE450

ThinkSystem SE350

ThinkEdge SE70

ThinkEdge SE50

ThinkEdge SE30

Learn more: https://news.lenovo.com/pressroom/press-releases/lenovo-ushers-in-new-era-of-edge-automation-at-scale/ and https://www.lenovo.com/us/en/servers-storage/solutions/edge-computing/.

Benchmark

At byteLAKE, our top priority is ensuring that our Cognitive Services are optimized to the latest technologies. This enables our clients to choose hardware that meets their specific needs and budget, giving them the flexibility they require for their business.

As of Q2 2023, we will start offering a hardware reference design that we will recommend for the majority of our deployments. Configuration details are described further in this report as part of the benchmarked use cases. Our optimizations ensure that the Cognitive Services can perform to their full potential, eliminating the need for industrial clients to invest in additional hardware accelerators that would increase hardware costs.

We optimized our Cognitive Services for the Paper Industry last year, and the details are available in a separate report. This report outlines the results of our optimization efforts across all industries where byteLAKE's Cognitive Services are offered. The optimization results bring new features and improved performance, demonstrating our commitment to continually improving our products and services.

Overall, our benchmarking and optimization efforts have resulted in **better performance, increased efficiency, and lower costs for our clients, making it easier for them to deploy our Cognitive Services in their businesses**. Our partnerships with leading technology providers like Intel® and Lenovo are critical to our optimization efforts, and we will continue to leverage these partnerships to deliver innovative and efficient solutions to our clients.

The benchmark was conducted on Lenovo ThinkPad P16s G1 Intel® Core™ i7-1260P | 32GB RAM equipped with the 12th Gen Intel® Core™ i7-1280P (microarchitecture Alder Lake). This processor is a high-performance CPU designed for use in a variety of applications, including machine learning, gaming, and content creation. The CPU supports Intel® Deep Learning Boost technology, which provides acceleration for certain machine learning workloads. With its powerful performance and advanced features, the 12th Gen Intel® Core™ i7-1280P is an excellent option for demanding applications that require significant processing power. More information about the compute platform can be found below.

Also, the products have been optimized to Lenovo's ThinkEdge SE50 which is byteLAKE's recommended hardware reference design for industrial deployments of byteLAKE's Cognitive Services.

AI-assisted Visual Inspection (benchmark)

Use case: images/videos analytics in restaurants. More: www.byteLAKE.com/en/AI4Restaurants.

Applicable to: all images/videos analytics scenarios available within byteLAKE's Cognitive Services.

Specification of the computing platform

Software:

- Windows 11 PRO, release 22H2
- Darknet – base framework for YOLO
- DarkFlow – Tensorflow implementation of Darknet
- OpenVINO™ 2022.3

Hardware: Lenovo ThinkPad P16s G1 / Lenovo ThinkEdge SE50

- CPU: Intel® Core™ i7-1280P with 14 physical cores and 20 logical cores
- RAM: 32 GB

Benchmarked Use Case

Computer Vision is a technique that focuses on interpreting and understanding visual information from the world around us. It includes several algorithms and techniques that allow for the analysis and interpretation of images in order to recognize and classify objects or identify patterns. Typically, such data is used to make further decisions.

Computer vision is utilized to create products in various industries, including healthcare, manufacturing, automotive, and others. The goal is to relieve people of manual work and to use machine capabilities to analyze and interpret data in ways similar to how humans would. With the aid of machine learning algorithms, many processes can be automated, resulting in faster and more accurate performance.

One of the key features of the Cognitive Services for Restaurants software is object detection. This computer vision technique is used to locate products on the customer's tray, recognize and identify meals, and then send a list of them to the cashier's machine, ultimately reducing queues and overall waiting time.

In this benchmark, our focus is on reducing the time required for object detection within the inference process. To achieve this goal, we have investigated the Intel® OpenVINO™ (Open Visual Inference and Neural Network Optimization) toolkit to enhance the performance of our product.

Porting Cognitive Services for Restaurants to OpenVINO™

Intel® OpenVINO™ is a toolkit that provides developers with tools and libraries for building and deploying computer vision applications. OpenVINO™ enables programs to optimize and accelerate AI applications developed using popular machine learning frameworks across a variety of Intel® computing platforms, including CPUs, GPUs, and FPGAs.

To optimize performance, OpenVINO™ utilizes the Model Optimizer (MO) command-line tool. The MO tool allows programmers to move models between training and deployment environments, perform static model analysis, and adapt models for optimal execution on target devices. In practice, this tool converts pre-trained models to the OpenVINO™ Intermediate Representation (IR) format, which can be used later to infer with OpenVINO™ Runtime.

As mentioned earlier, the CS for Restaurants software is based on object detection, and our product uses YOLO (You Only Look Once), a state-of-the-art object detection algorithm. YOLO stands out from other algorithms in that it operates on the entire image at once and uses a single neural network to predict bounding boxes and class probabilities for each object in the image.

Our product uses the Darknet open-source framework, a native framework for YOLO networks written in C and CUDA that allows CPU and GPU computations (source: source https://github.com/AlexeyAB/darknet). Since it is not possible to directly pass YOLO's *.cfg and *.weight files to OpenVINO™ Model Optimizer, we used DarkFlow to translate Darknet implementation of YOLO to its Tensorflow counterpart. Finally, the resulting Tensorflow-based YOLO can be passed through OpenVINO™ Model Optimizer, resulting in a pair of files that describe the model (IR):

- .xml file with network topology;
- .bin file containing the weights and biases binary data.

In addition to the MO tool, OpenVINO™ provides a command-line Post-training Optimization Tool (POT) that supports the uniform integer quantization method. It allows moving from floating-point precision (FP32 or FP16) to INT8 integer precision for weights and activations during inference time. We used this tool to quantize the model to take advantage of the VNNI instruction set available in the Intel® Core™ i7-1280P processor.

Benchmarking procedure

The benchmarking methodology is designed to align with the functioning of byteLAKE’s Cognitive Services for Restaurants software. Initially, we load 35 images into the RAM memory and perform inference. Subsequently, we calculate the duration taken to process all the images and present it in seconds and the FPS (Frames per second) factor. We analyze the images one-by-one, without batch processing, to stick to the real-life scenario. The measurements comprise resizing and preparing the image for Intel® OpenVINO™, inferencing, and filtering the results. The YOLO model utilized in the benchmark can detect 13 classes, representing artificial products. To optimize the detection accuracy, we set the input image size for the model as 608x608, assuming RGB images.

Performance results

Table 1 displays the performance results attained for Darknet and two OpenVINO™ implementations. For OpenVINO™, we consider two model precisions: FP32 and INT8, respectively. Along with the execution time and FPS factor, the table also illustrates the speedup value achieved against the Darknet implementations.

Table 1. Performance results achieved for Darknet and two OpenVINO™ implementations

Code implementation	Execution time [s]	FPS factor	Speedup
Darknet FP32	58.3	0.6	---
OpenVINO™ FP32	6.6	5.3	**8.85x**
OpenVINO™ INT8	2.3	15.2	**25.4x**

The performance results indicate that the FP32 implementation of OpenVINO™ outperforms the Darknet implementation, executing the inference 8.85 times faster. Furthermore, the INT8 model quantization reduces the inferencing time by up to 2.9 times compared to FP32 precision. It is worth noting that the model quantization was performed using the *DefaultQuantization* algorithm provided by the Post-training Optimization Tool (POT) (see paragraph: "Porting Cognitive Services for Restaurants to OpenVINO™" for more details). Finally, the INT8 model enhances the performance up to 25.4 times compared to the Darknet implementation.

To optimize the performance of the OpenVINO™ code, we assessed various runtime parameter setups such as *PERFORMANCE_HINT*, *AFFINITY*, and *INFERENCE_NUM_THREADS*. We discovered that the default settings of OpenVINO™ runtime produced the best performance, which is comparable to the LATENCY setup of the *PERFORMANCE_HINT* parameter. Furthermore, the shortest execution time is achieved when only 6 logical CPU cores are used.

During our inference benchmark, we discovered that all cores were not fully utilized. OpenVINO™ provides the Benchmark Python Tool, which allows for estimating the inference performance of a model on a given device. The performance of the model can be measured in two modes: synchronous (latency-oriented) and asynchronous (throughput-oriented). We used this tool to measure the performance of the model in the throughput-oriented mode, which maximized the number of parallel inference requests to utilize all threads of the device.

Figure 1 summarizes the FPS metrics achieved for all OpenVINO™ experiments mentioned above. The FP32 and INT8 bars refer to the FPS values achieved from our benchmark (see Table 1). The bars show a boost of about 2.87 times. The FP32 streams and INT8 streams bars correspond to the FPS results from the Python Benchmark Tool. We can see that the superiority of INT8 over FP32 increases to 3.35 times, as a result of better utilization of CPU logical cores.

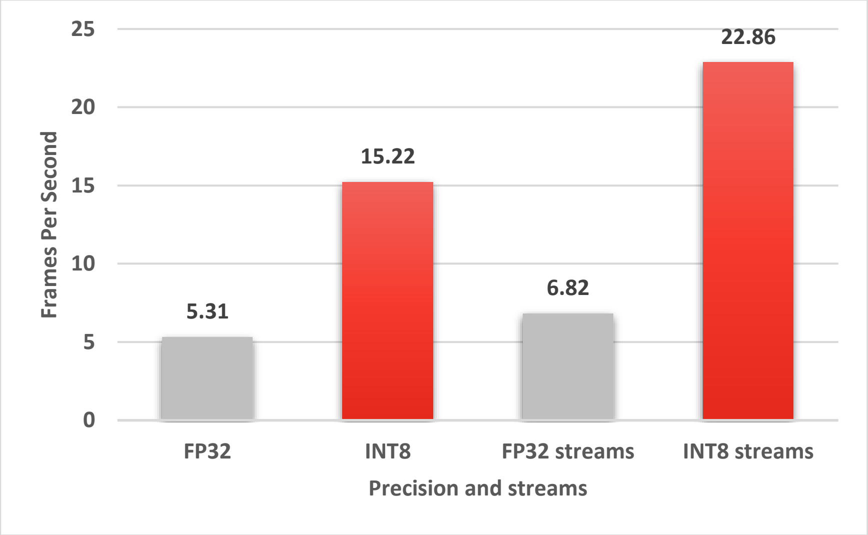


Figure 1. FP32 vs INT8 vs FP32 streams vs INT8 Streams

Quick note about the chart: the higher the bar, the better the speedup and overall performance. It's important to keep in mind that 'precision' in this context refers specifically to the data types used (FP32, INT8), not to the accuracy achieved by the trained AI model.

The performance improvements obtained by using INT8 over FP32 are consistent with the benchmark conducted by Intel® Corporation on their server with two Intel® Xeon® 8280 CPUs, as described in this article: docs.openvino.ai/...openvino_inference_engine_tools_benchmark.... In Intel® study, they observed a boost of approximately 3.42 times when comparing the performance of FP32 Streams to INT8 Stream.

It is obviously of interest to investigate the detection accuracy of the INT8 model. Our analysis shows that for the studied model, the detection results achieved by all OpenVINO™ models are consistent with the detections achieved by the native Darknet framework, with the only difference being the size of the bounding box assigned to the objects. Table 2 below and Figures 2, 3, and 4 present the detection results achieved, respectively, for the Darknet framework, OpenVINO™ FP32 model, and OpenVINO™ INT8 model.

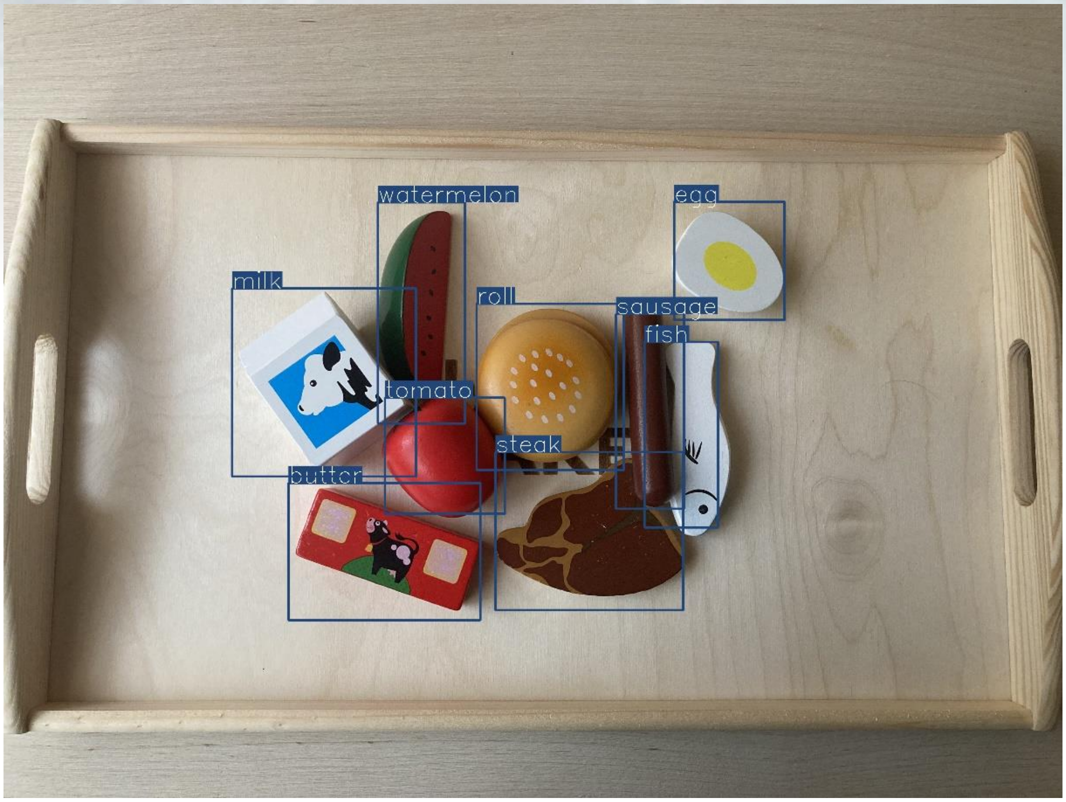


Figure 2. Detection results achieved for Darknet implementation

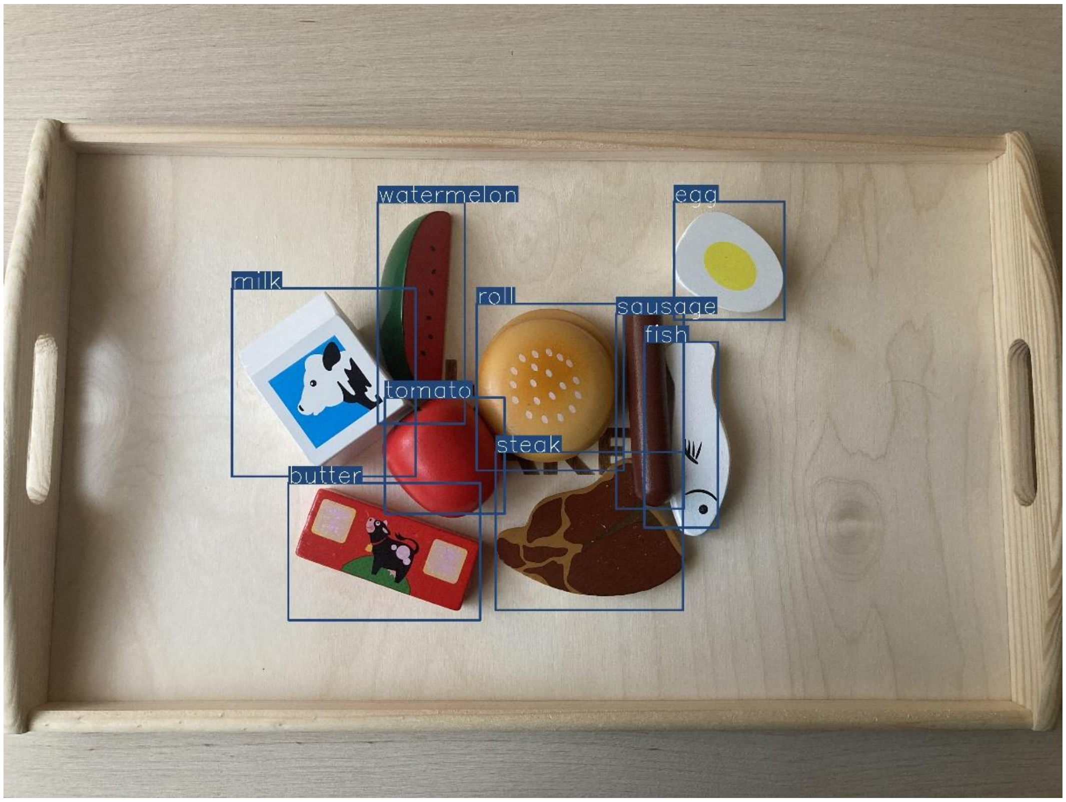


Figure 3. Detection results achieved for OpenVINO™ FP32 model

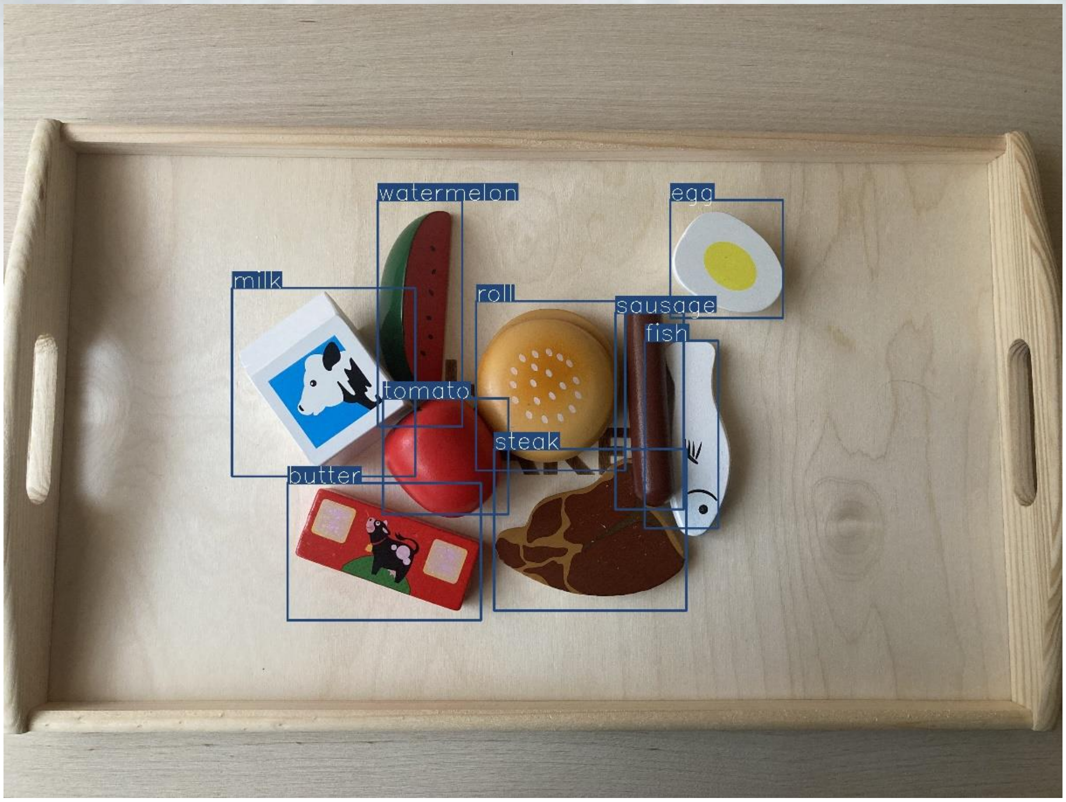


Figure 4. Detection results achieved for OpenVINO INT8 model

Table 2. Accuracy analysis

	Darknet	**OpenVINO FP32**	**OpenVINO INT8**
Objects detected	100%, all *(*)*	100%, all *(*)*	100%, all *(*)*
Average probability per object *(**)*	92.998	92.996	92.858
Median probability per object *(**)*	93.756	93.733	93.621

(*) byteLAKE's Cognitive Services for Restaurants detected and properly recognized all the objects (items on tray) in the benchmarked scenario.

(**) probability indicates the level of confidence that the detection/recognition is correct, calculated by byteLAKE's Cognitive Services for Restaurants for each detected and recognized object (item on tray).

Conclusions

In conclusion, OpenVINO™ is a powerful toolkit for optimizing the performance of AI model inference, offering easy-to-use tools that streamline the process of porting existing models. For the YOLO model used in our benchmark, OpenVINO™ accelerated object detection by about 8.85 times compared to native Darknet implementation for FP32 precision, and the usage of INT8 model permits further performance improvements. Our benchmark demonstrates that inferencing can be performed up to 2.87 times faster compared to FP32. **Finally, the detection process** with 8-bit precision allows for reducing object detection up to **25.4 times faster against Darknet implementation**.

In addition to performance gains, accuracy of detection is noteworthy. Our evaluation of detection results obtained for the OpenVINO™ INT8 model shows that the **quantization of the testbed model does not affect accuracy**. In practice, the INT8 model returns results consistent with the FP32 model.

In summary, the utilization of OpenVINO™ provides a powerful toolset for improving the inferencing process of machine learning algorithms, which can bring significant benefits to end clients in terms of faster and more accurate AI-powered applications.

AI-Powered Big Data Analytics (benchmark)

Use case: sound analytics in automotive. More: www.byteLAKE.com/en/automotive.

Applicable to: all sound/IoT analytics scenarios available within byteLAKE's Cognitive Services.

Specification of the computing platform

Software:

- Windows 11 Pro
- Python ver. 3.9.13
- scikit-learn ver. 1.2.1
- scikit-learn-intelex ver. 2023.0.1

Hardware: Lenovo ThinkPad P16s G1 / Lenovo ThinkEdge SE50

- CPU: 12th Gen Intel® Core™ i7-1280P clocked 1.80 GHz, 14 physical cores and 20 logical cores
- RAM: 32 GB

Benchmarked Use Case

Regular maintenance is crucial for keeping car engines functioning properly, and one important aspect of maintenance is detecting faults or malfunctions early on. Sound analysis is a useful technique for identifying engine issues. The sound produced by a car engine can provide valuable information about its condition, and a trained technician can use their ears and specialized equipment to analyze the sound and identify any anomalies. Common faults that can be detected through sound analysis include misfires, knocking or pinging, whistling or hissing, and grinding or rattling. In addition to these specific faults, sound analysis can also be used to detect more general issues with the engine's overall performance.

At byteLAKE, we have developed a tool that uses machine learning / artificial intelligence (AI) algorithms to automatically detect faults in car engines. These algorithms enable technicians to automate the fault detection process, making it faster and more accurate. To improve the performance of our tool, we investigated an Intel® extension during the self-learning process (being part of the Cognitive Services), in which a set of algorithms are trained under many different configurations. We also provide a benchmark of the inference part to explore ways to further improve the performance of real-time analytics.

Algorithms used in our framework that are supported by scikit-learn-intelex ver. 2023.0.1:

Please note that byteLAKE's Cognitive Services offers a vast array of algorithms that cater to different quality inspection scenarios in manufacturing, including those in the automotive industry. The present study showcases only a subset of these algorithms, while noting that the performance enhancements using the approaches outlined in this report and other methodologies were consistent across the entire set of algorithms. This report mostly focuses on the machine learning part of the Cognitive Services.

Scikit-learn is a widely used open-source machine learning library for Python that offers a broad range of tools for numerous tasks, including data preprocessing, model selection, and performance evaluation. It comprises a wide range of machine learning algorithms, including DBSCAN, KMeans, PCA, and Nearest Neighbors, which were chosen for this benchmark.

In addition to the standard scikit-learn library, there is an optimized version of scikit-learn called scikit-learn-intelex (SKL-IE). SKL-IE is an extension of scikit-learn that utilizes Intel® Math Kernel Library (MKL) to accelerate various parts of the machine learning pipeline, including linear algebra operations, distance calculations, and model fitting.

Using SKL-IE can result in significant speed improvements when training and inferring machine learning models. This is particularly advantageous when working with large datasets or complex models that require substantial computation. When applied to car engine fault detection, SKL-IE can expedite the training and inference of machine learning models that utilize sound analysis. Consequently, this can result in faster and more accurate fault detection, which is essential for guaranteeing the safety and dependability of vehicles.

Overall, scikit-learn and its optimized extension, SKL-IE, offer potent tools for addressing machine learning problems, including the detection of car engine faults via sound analysis. The library enjoys widespread use in both industry and academia, and its user-friendly interface and comprehensive documentation have made it a popular choice among machine learning professionals.

We have selected a group of algorithms that are supported by SKL-IE and that we use in order to benchmark our tools. This group of algorithms includes:

a. DBSCAN (training only)
b. KMeans (training only)
c. PCA
d. NN – NearestNeighbors

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm commonly used in machine learning for data analysis. It groups together points that are close to each other, based on their proximity and density, while ignoring outliers or noise points. The algorithm works by defining a neighborhood around each point, and then grouping points that are within a certain distance and density threshold. This makes DBSCAN particularly useful for identifying clusters in data with varying densities and shapes.

KMeans is a popular clustering algorithm that partitions data points into k clusters based on their similarity. The algorithm starts by selecting k random points as the initial centroids and then iteratively assigning each data point to the closest centroid and updating the centroids based on the mean of the points in each cluster. This process continues until the centroids no longer change significantly or a maximum number of iterations is reached. KMeans is commonly used in data analysis and image segmentation.

Principal Component Analysis (PCA) is a technique used for dimensionality reduction and feature extraction. It works by finding the principal components of the data, which are the directions with the highest variance. PCA then projects the data onto these components, creating a lower-dimensional representation of the original data. This can be useful for visualizing high-dimensional data, reducing computational complexity, and removing noise or irrelevant features.

Nearest Neighbors is a machine learning algorithm used for classification and regression. It works by finding the k-nearest points to a given data point, where k is a user-defined parameter, and using their labels or values to predict the label or value of the given point. This algorithm is particularly useful when there is a strong correlation between the target variable and its closest neighbors.

Overall, these algorithms are commonly used in machine learning and data analysis for various tasks, such as clustering, classification, regression, and dimensionality reduction. Each algorithm has its strengths and weaknesses, and the choice of which algorithm to use depends on the specific problem and data being analyzed.

Dataset includes a set of sound files. The technical details of the dataset are:

- Number of samples: 548, 1918, 3014, 4384 – for 4 different groups of benchmarks
- Type of samples: wav
- Sampling rate: 22050 Hz
- Duration of sample: 1 second
- PCA: input number of features is 22050, but output is from 8 to 256 – depending on benchmark
- Number of features (other algorithms): 8, 64, 128, 256 – depending on benchmark

Performance results

Speedups of Intel® Extension for Scikit-learn over the original Scikit-learn (training)

Table 3 (Figure 5) presents the performance results, which include the algorithm, its configuration, and the number of samples and features of each sample. The final column displays the speedup achieved using SKL-IE over the standard SKL library. These results were obtained during the system training phase.

Table 3. Speedups of Intel® Extension for Scikit-learn over the original Scikit-learn (training)

Training			
Algorithm	Samples	Features	Speedup
DBSCAN	548	8	3,75
DBSCAN	1918	64	9,36
DBSCAN	3014	128	12,68
DBSCAN	4384	256	22,14
KMeans	548	8	4,35
KMeans	1918	64	11,81
KMeans	3014	128	16,32
KMeans	4384	256	23,31

PCA	548	8	1,12
PCA	1918	64	1,65
PCA	3014	128	1,92
PCA	4384	256	2,74
NN	548	8	2,51
NN	1918	64	6,06
NN	3014	128	9,05
NN	4384	256	16,93

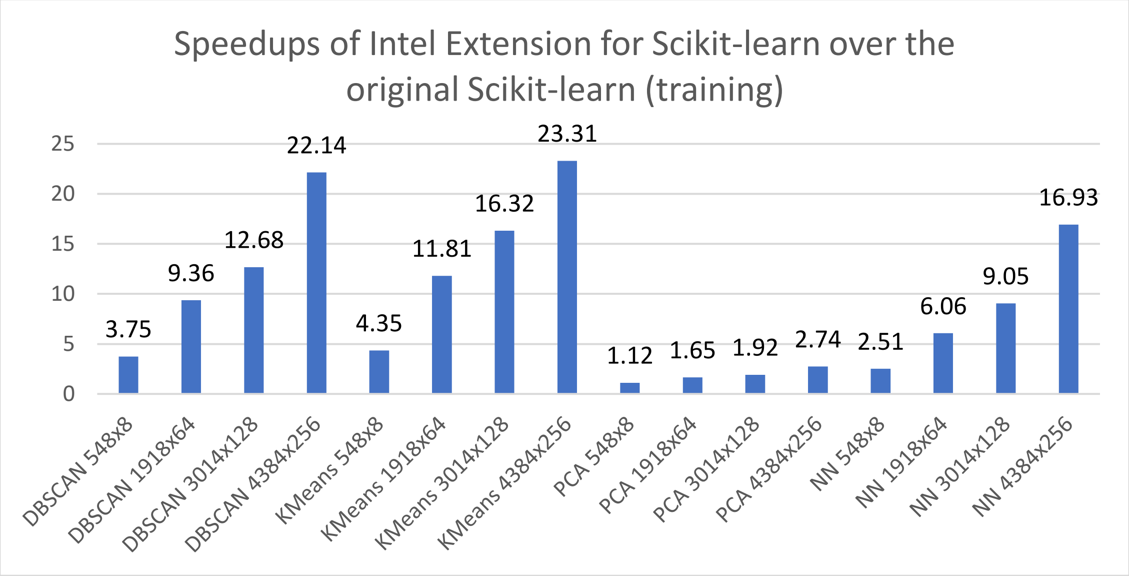


Figure 5. Speedups of Intel® Extension for Scikit-learn over the original Scikit-learn (training)

Quick note about the chart: the higher the bar, the better the speedup and overall performance.

Speedups of Intel® Extension for Scikit-learn over the original Scikit-learn (inferencing)

In Table 4 (Figure 6), we included the performance results from the inferencing phase of our tool.

Table 4. Speedups of Intel® Extension for Scikit-learn over the original Scikit-learn (inferencing)

Inferencing			
Algorithm	Samples	Features	Speedup
PCA	548	8	1,64
PCA	1918	64	1,91
PCA	3014	128	2,13
PCA	4384	256	3,55
NN	548	8	4,17
NN	1918	64	8,09
NN	3014	128	11,15
NN	4384	256	21,17

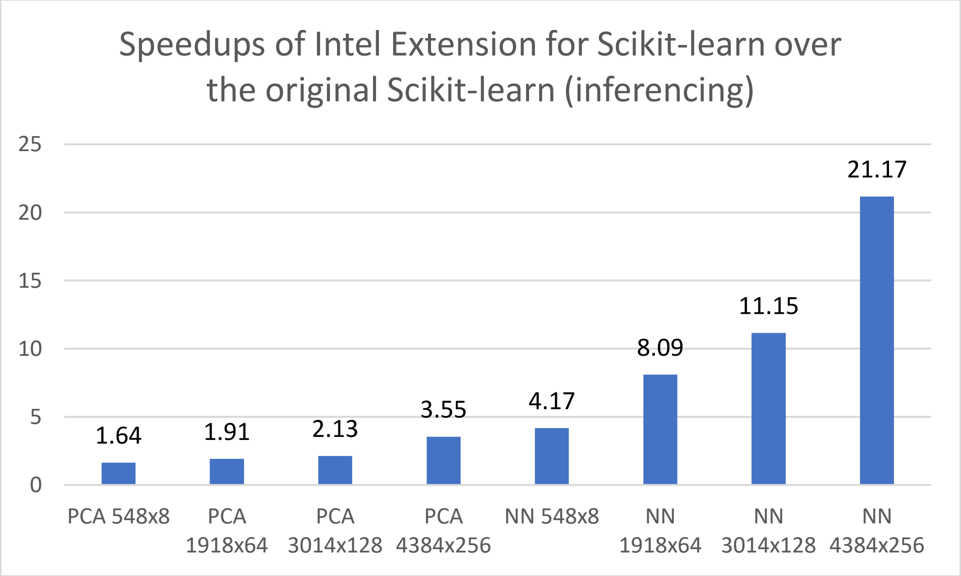


Figure 6. Speedups of Intel® Extension for Scikit-learn over the original Scikit-learn (inferencing)

Quick note about the chart: the higher the bar, the better the speedup and overall performance. It's important to keep in mind that DBSCAN and KMeans are not applicable to the inference process. Therefore they appear only in the AI training related benchmarks.

Accuracy analysis

For the benchmarked scenario, the following results have been achieved:

Table 5. Accuracy analysis

Samples	Good engines	Damaged engines	False-negative	False-positive	**False-negative [%]**	False-positive [%]	Error [%]
5000	4920	80	0	137	**0**	2.78	1.6

Where:

Samples	-	number of engines tested in the experiment
Good engines	-	engines with no fault
Damaged engines	-	engines with fault
False-negative	-	damaged engines recognized as good by the system
False-positive	-	good engines recognized as bad by the system (false alarms)
Error	-	the ratio between false recognitions (False-negative+False-positive) and Samples

Conclusions

In summary, SKL-IE is a powerful machine learning library that provides a wide range of tools and algorithms, making it an ideal solution for various tasks, including car engine fault detection based on sound analysis. SKL-IE not only provides an easy-to-use interface for developing and deploying machine learning models, but it also offers significant speed improvements for both training and inference phases.

With **speedups ranging from 1.12 for PCA to 22.14 for DBSCAN during training and up to 21.17 during inference**, SKL-IE's optimized performance can be particularly beneficial for dealing with large datasets or complex models that require substantial computation. Machine learning practitioners can leverage these speed improvements to achieve faster and more accurate results, leading to better performance and efficiency in various applications.

In essence, the combination of scikit-learn and its optimized extension, SKL-IE, provides a powerful and efficient toolset for solving machine learning problems, including car engine fault detection based on sound analysis. By using SKL-IE, clients can benefit from faster and more accurate results, ultimately leading to improved performance and efficiency in their applications.

byteLAKE’s Recommended Hardware Reference Design for byteLAKE’s Cognitive Services Deployments

byteLAKE's Cognitive Services have been optimized for efficient edge computing on a range of hardware platforms, offering end clients maximum performance and minimum cost. To achieve optimal results for AI-assisted visual inspection, production line monitoring, big data analytics, and predictive maintenance, byteLAKE recommends the following hardware options that have been benchmarked and optimized by byteLAKE.

Edge Computing platform

- Device: Lenovo’s ThinkEdge SE50 (link: PSREF ThinkEdge ThinkEdge SE50 (lenovo.com))
- Configuration:
  CPU: 12th Gen Intel® Core™ i7-1280P clocked 1.80 GHz, or newer
  RAM: 32 GB
- Recommended deployment type: 1 device per production line (manufacturing) / point of sale (restaurants)

Edge Server

- Device: Lenovo's ThinkEdge SE450 (link: ThinkEdge SE450 Datasheet > Lenovo Press)
- Configuration:
  CPU: 2* Intel® Xeon® Silver 4314 (16c) @ 2.40GHz
  GPU: Intel® Data Center GPU Flex Series 170 *(*)*
  RAM: 256Gb DDR4 @ 3200MHz

byteLAKE's consultancy services offer clients the expertise and guidance needed to identify the ideal configurations for their needs, ensuring optimal performance and benefit. Learn more at: https://www.bytelake.com/en/request-bytelake-service/.

(*) benchmarks related to Intel® GPU Flex series will be released by byteLAKE soon.

Key Takeaways

byteLAKE offers a comprehensive collection of Artificial Intelligence (AI) models as part of their Cognitive Services portfolio. These AI models have been expertly designed and pre-trained to focus on specific tasks resulting in maximum accuracy and efficiency. byteLAKE has strategically partnered with leading industry experts and talented researchers to combine human knowledge, expertise, and know-how with the latest AI algorithms and technologies.

The Cognitive Services offering includes AI models for AI-assisted Visual Inspection in manufacturing, facilitating self-check-out services, and providing advanced analytics in the restaurants. In the automotive industry, the solution utilizes microphones and other IoT sensors to monitor quality, detect issues, and offer predictive maintenance suggestions. The technology has also been successfully implemented in the pulp and paper industry, where it can monitor the papermaking process and prevent unforeseen downtimes and catastrophic events.

Clients of byteLAKE can expect benefits such as access to the latest technologies, enhanced performance and efficiency, and a reduced total cost of ownership. The partnership with top technology providers such as Lenovo and Intel® ensures that their software products are always up-to-date and optimized for the latest technologies.

byteLAKE's Cognitive Services are optimized for Edge AI deployment, which offers numerous benefits, including scalability, faster responses, and reduced total cost of ownership. Data can stay locally on the device, providing a selective option for sending data to external storage, optimizing operations and enhancing the user experience.

The optimization results have demonstrated better performance, increased efficiency, and lower costs for clients, making it easier for them to deploy byteLAKE's Cognitive Services in their businesses. For deployments, byteLAKE recommends Lenovo's ThinkEdge SE50, equipped with the 12th Gen Intel® Core™ i7-1280P (microarchitecture Alder Lake), which supports Intel® Deep Learning Boost technology, providing acceleration for certain artificial intelligence workloads.

Our innovative AI models enable clients to benefit from accurate real-time monitoring, predictive maintenance, and advanced analytics. These features help clients **improve** their **decision-making, enhance their quality inspection at lower cost, prevent downtime in production and achieve greater success.**

Way forward

byteLAKE is continuously expanding the features and use cases of the Cognitive Services software. The software is optimized for upcoming and new hardware platforms, ensuring the best possible performance. To stay updated on our latest advancements, follow byteLAKE on social media channels. byteLAKE is also a pioneer in benchmarking and optimizing the Cognitive Services to the Intel® Data Center GPU Flex Series, which has yielded impressive results. More details on these achievements will be released soon. Intel® GPU Flex series have been benchmarked in a real-life scenario by byteLAKE using Lenovo's ThinkEdge SE450 Edge Server equipped in 2* Intel® Data Center GPU Flex Series 170, 2* Intel® Xeon® Silver 4314 (16c) @ 2.40GHz and 256Gb DDR4 @ 3200MHz.

Be the first to catch a glimpse of our latest updates from the Euroshop 2023 conference! Our video includes early commentary on byteLAKE's product optimization for the Intel® GPU Flex series.

To learn more go to www.byteLAKE.com or follow us on:

- LinkedIn: https://www.linkedin.com/company/byteLAKE
- Twitter: https://twitter.com/byteLAKEGlobal
- Facebook: https://www.facebook.com/byteLAKE/
- Medium: https://marcrojek.medium.com/

Let’s continue our discussion!

Marcin Rojek
byte LAKE

mrojek@byteLAKE.com
+48 508 091 885

Mariusz Kolanko
byte LAKE

mkolanko@byteLAKE.com
+48 505 322 282

byteLAKE

Artificial Intelligence for Industries.

AI Products for Manufacturing, Automotive, Restaurants, Paper, and Chemical Industries.

About byteLAKE

We are a software company, focused on building Artificial Intelligence products for various industries. byteLAKE's CFD Suite leverages AI to reduce CFD (Computational Fluid Dynamics) chemical mixing simulations' time from hours to minutes. byteLAKE's Cognitive Services offer a collection of pre-trained AI models for Industry 4.0 and Restaurants. In manufacturing, these are used to visually inspect processes, parts, components, or products. In automotive, Cognitive Services leverages microphones to assess the quality of car engines. In the paper industry, Cognitive Services leverages cameras to monitor the papermaking process and detect, measure, and analyze the wet line. In restaurants, Cognitive Services typically acts as an add-on software that recognizes meals and sends a list of these to the cashier's machine, ultimately reducing lines and the overall waiting time. The company also offers custom AI software development for real-time analytics of images, videos, sounds, and time-series data. To learn more about byteLAKE's innovations, go to www.byteLAKE.com.