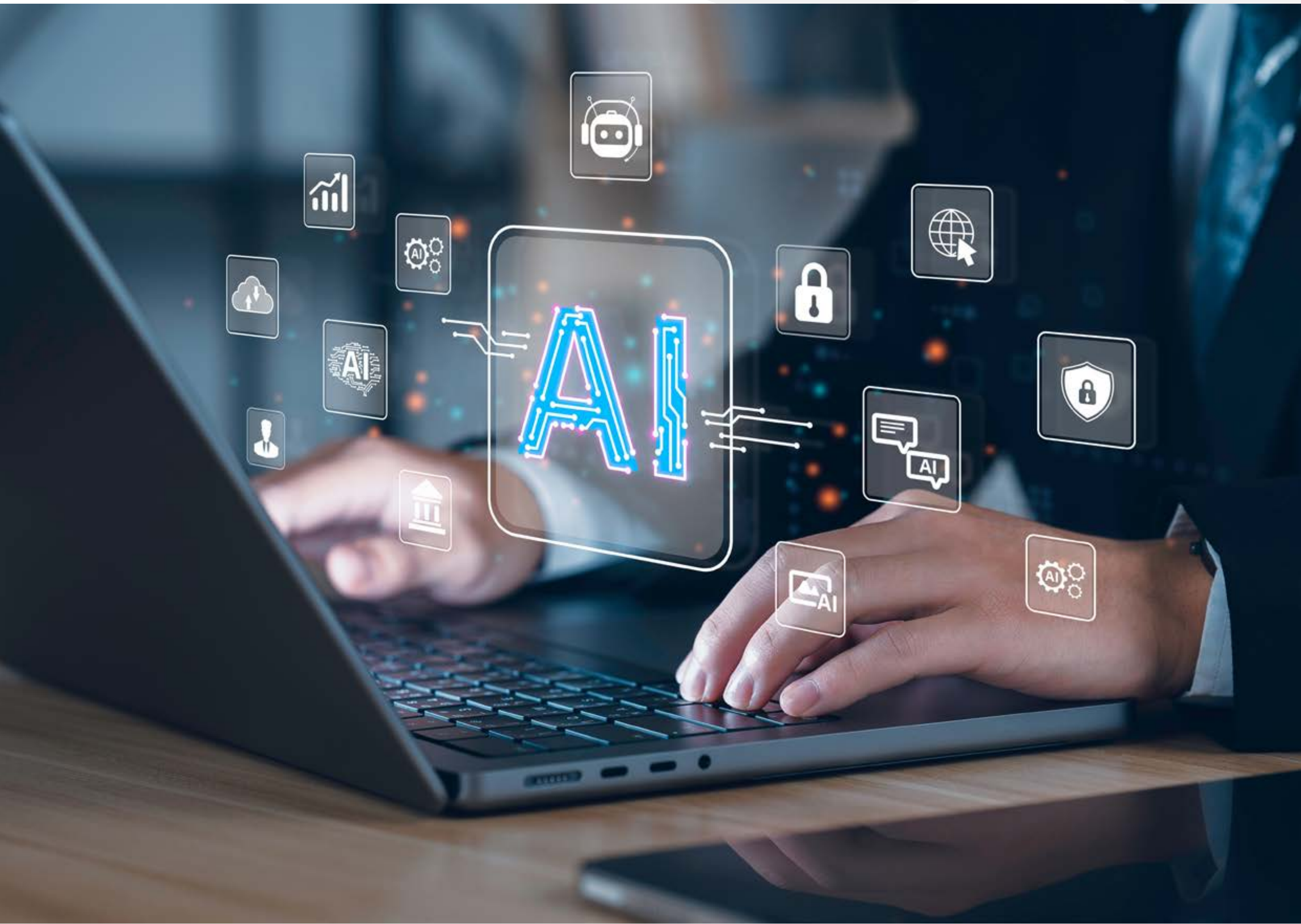
WHITE PAPER

Continuous machine learning: Your AI edge



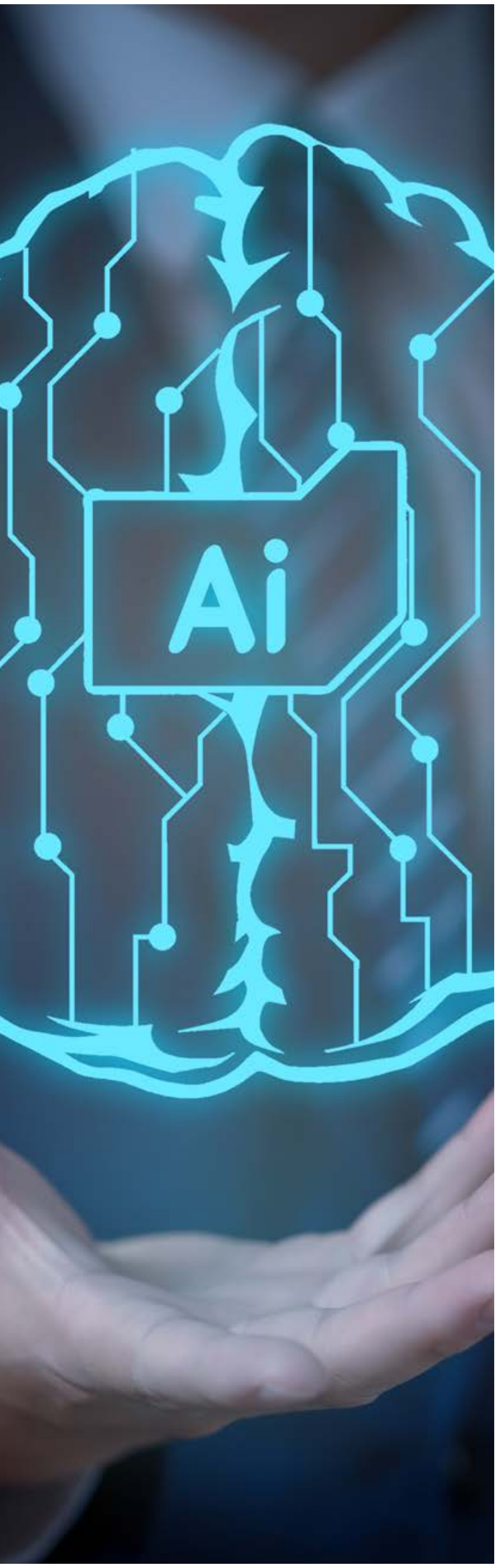
Contents

Executive summary

Digital business documents have replaced many paper documents, and the volume of documentation is expected to soar in the coming years. Furthermore, a large portion of these documents could be unstructured, such as email, images, and videos. Organizations leverage intelligent document processing solutions to help them cope with the digital document deluge. But even today's automated platforms can fall behind. As document content and layouts change over time, systems require costly, time-consuming manual interventions that reduce efficiency and revenue. AI adds efficiency and accuracy to automated capture workflows.

Unfortunately, AI models can also take time and resources to train and calibrate. OpenText information capture products and intelligent document processing (IDP) solutions solve this challenge by embedding continuous machine learning (CML). An AI approach to information capture and data extraction, CML eliminates data staleness through an ongoing refresh as the model self-corrects and relearns. Humans in the loop ensure data accuracy as part of daily production runs – eliminating the need for week- and month-long pauses as data scientists scrub data sets to retrain models.

This position paper shows how OpenText leverages a continuous machine learning (CML) approach that offers flexibility, accuracy, and efficiency while minimizing or eliminating manual model retraining.



Introduction

Every day, AI becomes increasingly ingrained as the go-to approach for adding speed and accuracy to daily activities. Organizations that need to transform content look to intelligent automation to accelerate business processes and power employee productivity. AI enables accurate and efficient automation of content classification, data extraction, and downstream process enablement.

Unfortunately, intelligent automation is not always intelligent or autonomous. Increasingly, documents and other information come in a variety of semi structured and unstructured formats—PDFs, email and attachments, images, digital faxes, SMS texts, and more. In addition, document layouts and content change over time, requiring moderate to significant manual intervention to retrain the AI models to maintain accuracy. These and other factors reduce an organization’s efficiency with time-sensitive processes.

Approaches to intelligent document processing

Automatic data processing has existed in some form for decades. Since the early 2010s, robotic process automation (RPA) has taken repetitive tasks from employees’ hands, adding speed and repeatability to well-worn, tedious administrative tasks. But RPA has limits. It relies on structured, electronic data and follows scripts that users must rewrite or re-record to accommodate data or process changes.

To achieve next-level benefits with automation and high productivity at scale, organizations have turned to machine learning (ML), where an initial data set is required to train one to many models. Data scientists are needed to scrub immense collections of training data, identify outliers, and label the key data elements for the ML model. These specialized skillsets and activities come at a considerable cost to the organization.

With this batch-mode, expensive concierge approach to training and retraining, updates typically occur only periodically and without key knowledge workers’ input. Even after carefully curating a data set, the model results may not represent real-world processing. To implement an update, one or more data scientists must refactor and repeat the training process with net new data, and additional data scrubbing, labeling, and testing cycles.

Although ML speeds document processing, it is data- and processor-hungry. Critically, as document layouts change, accuracy drifts and degrades over time. Keeping models accurate relies on periodic updates requiring the labor-intensive cycle of retraining, sometimes at the code and database level, so everything the model previously learned is retained but adjusted to maximize accuracy based on the latest document set.

However, more accurate and effective ML approaches exist. Many organizations have turned to continuous machine learning (CML) to address content classification and data extraction needs. With CML, your model is retrained and updated on-the-go as it encounters new data and layouts in production. Updates occur in real-time in small batches, which reduces computational time. More importantly, CML reduces the data and human resources required to retrain models.

Continuous machine learning: The OpenText answer

In business processes, the variety of document layouts can change ever so slightly over time. One new document layout may not affect the overall accuracy of the ML model, but multiple, subtle changes in document layouts can add up to a considerable divergence in results—and substantial retraining work.

The OpenText approach to CML relies on methodology embedded in its Information Extraction Engine (IEE). Data and differing layouts can quickly be reinforced with just a few clicks by a knowledge worker using a human-in-the-loop UI. IEE continuously assesses human feedback to reinforce or adjust the model accordingly. Feedback that is subtly different from the established ruleset does not impact the model, ensuring a high level of data integrity.

Organizations have recognized the value associated with no longer having to spend weeks or more to manually reassess, retrain, and examine ML models, only to repeat the process again and again. IEE eliminates the need for a team of data scientists to maintain and retrain ML models. Instead, people can spend time on value-added activities, such as retraining the business's risk models. In addition, depending on your requirements, IEE is effective right out of the box, with limited or no training. There is no need for a long wait after setup to realize its benefits. The OpenText approach to continuous machine learning is an efficient solution to data identification and extraction.

From data deluge to agile AI

In the past, when setting up a document processing platform or creating a prototype, a data scientist often received an overwhelming number of production documents, sometimes up to 10,000 pieces, often filled with sensitive information. Creating as many as 50 templates could take up to 80 days. In one notable example, the proof of concept alone required nearly half a million training documents!

Today, continuous machine learning (CML) has transformed cumbersome deployment and maintenance into an opportunity for agility. Now, companies can launch production without the need for laborious pre-training, with configuration times that can be reduced from 80 days to five.

How does continuous machine learning recognize a document?

To understand how CML works, let's first explore how a platform recognizes a document.

What makes a document?

For classification and extraction purposes, in most cases, business documents have three parts: a header section, a footer section, and a document body. The header usually identifies the business partner, the type of transaction, and high-level details about the particular transaction. The footer often includes totals or amounts, taxes, page numbering, and sometimes provides business partner details. The body is usually a list of records, such as line items, and specifies the unique content for the transaction.

Certain business documents or use cases have a consistent content structure. AI models are built to look for these structures and their content. For example, invoices contain purchase order (PO) numbers, net payment dates, and itemized product descriptions and prices. Human resources (HR) documents include names, addresses, social security numbers, and other information.

Many capture solutions offer a general model that extracts data from a variety of different document types. This approach often sacrifices accuracy to achieve generality. You achieve the highest accuracy using separate models that are trained for each document type.

Semi-structured, structured, and unstructured

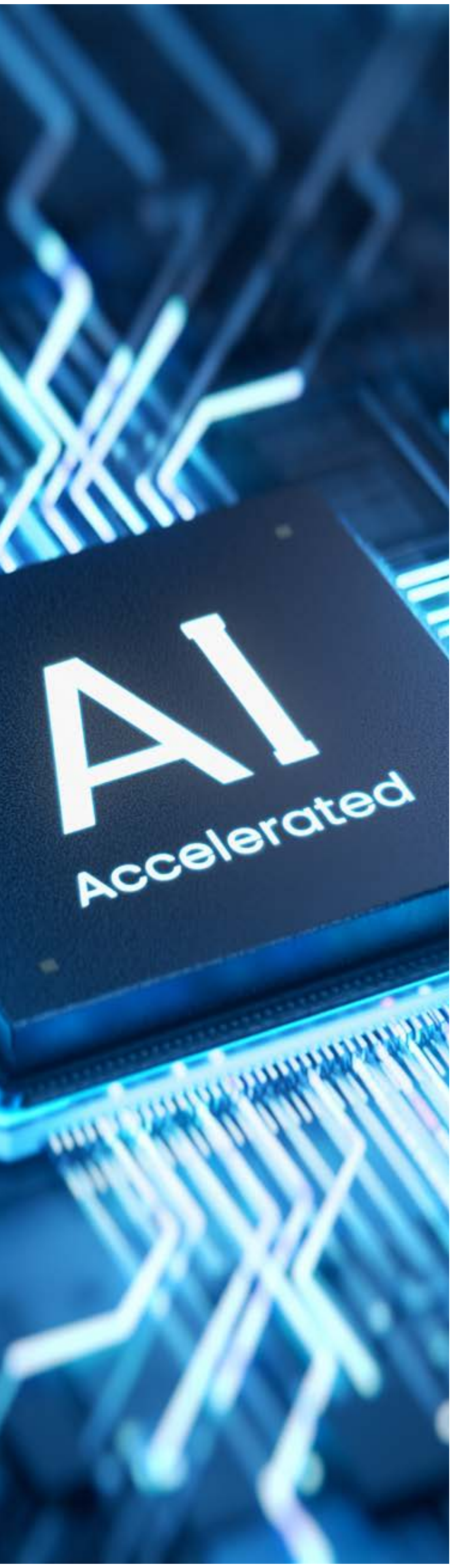
Not all documents are the same. They can be separated into three types: semistructured, structured, or unstructured.

- **Semi-structured documents:** Documents of this type all contain the same information, regardless of customer, but different customers put the information in different locations on the page. Examples include invoices, purchase orders, and delivery notes.
- **Structured documents:** The extracted data is the same for every document and is always found in the same location, such as in tax forms.
- **Unstructured documents:** The types of information contained, and where it could be located, can vary wildly from document to document. Unstructured documents include correspondence and emails.

Classifying and extracting data from these different document types requires different technologies.

Continuous machine learning: What's in the black box

Even to developers, ML can seem like a black box. How much more so for business users? People may think, “With other methods, we had to retrain the model.” They may also experience doubt: “How could something learn that accurately and effortlessly?”



An ensemble of algorithms

At its core, OpenText CML is based on tried and tested architectures--with OpenText's multi-engine-ensemble approach based on several internally developed algorithms. Examples of technologies combined in IEE's proprietary ensemble of algorithms include:

- Natural language processing
- Domain-specific feature-engineered machine learning
- Statistical pattern recognition
- Neural networks
- Deep convolutional neural networks

IEE identifies the document and extracts data using the ensemble of algorithms, with results then evaluated by a voting mechanism.

- Some algorithms are pretrained by OpenText for relevant use cases or document types. So they can work well out of the box.
- Some algorithms start from setup without any training. Algorithms learn the specifics of a customer's requirements during initiation with the first few customer documents.
- OpenText specifically trains other algorithms to perform tasks like scanning headers or identifying documents from a specific sender. Each algorithm learns as it processes documents.
- Algorithms may work independently, or two or three algorithms may work together to complete a task.

What makes OpenText CML the best option is its unique take on classifier voting.

Voting: The deciding factor

"Voting" is a machine learning approach that learns from a set of algorithms and determines which algorithm works best in any given situation. In CML, voting selects the best result for each piece of data to be identified or extracted.

For example, say you want to extract the invoice date field from invoices. Two engines based on different algorithms extract the information in parallel. The engines might deliver contradictory results. For example, one engine might incorrectly extract an order date instead of the invoice date from a particular invoice. In this case, the voting engine decides which result is more appropriate.

Through the CML feedback process, the voting engine learns specific patterns to support the decisions it is making. For instance, if one engine has problems extracting the invoice dates from a vendor's invoices, the voting mechanism recognizes this pattern. It then favors the invoice date results from the other engine for that vendor. Thus, the OpenText CML takes advantage of feedback data both to reinforce the core information extraction models and improve the voting model.

Privacy challenges eliminated

There are growing concerns surrounding privacy, data protection and compliance with organizational guidelines and government privacy regulations. As companies embrace more flexible and efficient approaches to document processing, they must also tread carefully in handling sensitive information. Solutions that require large data sets of live customer or sensitive data to train or retrain models represent a clear and present danger to your business.

With CML, new configurations can be put immediately into production without training. Any training to fine tune them can be completed using just a handful of documents, significantly reducing the exposure risk for confidential information.

Humans are in the loop

When AI was introduced, business users initially believed it was a "magic solution" and then were disappointed when it couldn't do everything perfectly on its own. Now that organizations have experience with AI, they want humans in the loop (HITL) for automated processes in order to validate that all is happening as expected. For information capture and IDP, HITL UIs allow people to validate ifthe results from AI-based document classification and data extraction are accurate. However, managers know that whenever manual tasks are involved, mistakes can happen. Despite its self-learning properties, it's natural to wonder how self-healing CML is and whether, if results start skewing, you can get under the hood to reset your model. What if an employee makes a mistake—or two or three?

If human errors do occur, the model is like a rolling stone. It may indeed gather a little moss, but it rolls through puddles and corrects itself as it goes along. In other words, the algorithm ensemble and voting mechanism automatically and continuously correct the models, removing outdated and old patterns through the ongoing learning process. **The critical thing to note is that a few errors won't derail a hardened model that has processed several thousands of documents.** Having to reinforce a data change minimizes random errors from making unwanted changes. This in-built resilience radically reduces out-of-range outliers that can happen with other machine learning models.

When errors do occur, most retraining usually succeeds after one or two iterations. Hardened data may take several iterations for retraining, but this occurs during normal production and does not require a data scientist or your IT department. In addition, with OpenText CML, humans are always in the loop to review the data identified by AI.

The future of intelligent document processing

What does the future hold? Based on current trends, AI is certain to play a role. For OpenText's IEE, that means:

- Handcrafted optimization of individual customer models is expected to decline.
- Custom optimization for specific use cases will survive or increase.
- Custom optimization based on large generic pretrained models as ML-optimized prompts looks promising.
- Large language models (LLM) such as GPT, Vertex AI, BERT, and PaLM likely will play an increasing role in document parsing and document object detection.
- Increase in individual model accuracy is expected with LLM technology infusion.

What won't change

- CML will continue performing the essentials of automatic data processing by determining document type and performing content extraction.
- CML will continue to use the algorithm ensemble approach to leverage feedback to create self-learning and self-healing in models.

Connect with us

OpenText CEO Mark Barrenechea's blog ›

X (formerly Twitter) ›

LinkedIn ›

Sustain document production with continuous machine learning

As the volume of business and unstructured content grows, organizations need the speed and efficiency of an automated intelligent document processing solution. They also require that solution to be one that they can easily manage. RPA and ML have provided answers to structured document processing demands. However, the human, financial, and time resource overhead required to keep templates and models current and accurate is considerable.

The OpenText approach to CML for content classification and data extraction can seamlessly incorporate AI model updates during daily production runs. Humans in the loop provide oversight to ensure that AI models continue to produce accurate results. OpenText CML offers the low-touch answers for today, with adaptability for new algorithms and technologies for the months, and years, to come.

Next steps

To learn more about OpenText capture and IDP solutions that include continuous machine learning via the Information Extraction Engine (IEE), contact us or visit the OpenText website.

About OpenText

OpenText, The Information Company, enables organizations to gain insight through market leading information management solutions, on premises or in the cloud. For more information about OpenText (NASDAQ: OTEX, TSX: OTEX) visit: opentext.com.

opentext™