TLP: WHITE

ENJOY SAFER TECHNOLOGY™

ESET ADVANCED MACHINE LEARNING

Authors:
Ondrej Kubovič, ESET Security Awareness Specialist

with contributions from

Juraj Jánošík, ESET Head of AI/ML team
Peter Košinár, ESET Technical Fellow
Előd Kironský, Head of Core Technology Development

CONTENTS

Authors:
Ondrej Kubovič, ESET Security Awareness Specialist

with contributions from

Juraj Jánošík, ESET Head of AI/ML team
Peter Košinár, ESET Technical Fellow
Előd Kironský, Head of Core Technology Development

INTRODUCTION

Machine learning technology is influencing, and possibly even reshaping, our everyday lives. Some of its real-world applications serve the "average Joe", like Google's search engine or Netflix's algorithm for suggesting what to watch next. Others are less obvious, hidden in the background, helping researchers to develop new solutions to existing problems, and businesses in their fight against large-scale fraud and cybercrime.

At ESET, our engineers are old acquaintances of machine learning. We recognized its potential early on and employed it to help detect malware over 20 years ago. To this day, this symbiosis continues, with neural networks, deep learning, and classification algorithms being integral parts of the protective layers in ESET products and services.

This white paper introduces the reader to decades of ESET experience with machine learning, emphasizing how the latest applications of this technology blend into ESET's current home security solutions.

There are many misconceptions – and outright misrepresentations – about ML and AI; we appreciate having the opportunity to address these, as well as to introduce you to how they benefit our software offerings. Thanks for reading!

GLOSSARY:

For the purposes of this white paper, we take Artificial Intelligence, Machine Learning and Deep Learning to mean the following:

Artificial Intelligence (AI)

AI represents the as-yet-unachievable ideal of a generally intelligent and self-sustainable machine that can operate, make decisions, and learn independently, based solely on inputs from the environment – all without human involvement.

Machine Learning (ML)

Machine learning means data processing algorithms that allow computer systems to perform chosen tasks by identifying patterns and anomalies in vast amounts of data, and transforming it into a compact representation, known as a "model". Machine learning is considered one of the possible stepping stones on the way to true AI.

Deep Learning (DL)

Deep learning is a subset of machine learning models, inspired by the human brain, that has proven effective in processing massive sets of sequential data. Deep Learning has allowed significant improvements in the cybersecurity field. Its contribution to malware detection capabilities can be compared to the difference between the viewer's experience when looking at a grainy still photo and a high-definition video recording.

ESET'S 20 YEARS OF MACHINE LEARNING

Machine learning is a field of computer science with deep roots. Its inception dates back to the 1950s and despite many technical and performance limits, it saw its first real-world applications in computer security solutions even before the year 2000.

One of those instances was ESET's detection engine. In the mid 1990s, the internet was booming and so were new threats. Macro viruses, perhaps the greatest of these, were often spread via emails and their attachments. ESET engineers realized that one type of machine-learning algorithms - neural networks - had the potential to help detect these rapidly evolving threats, so they started experimenting.

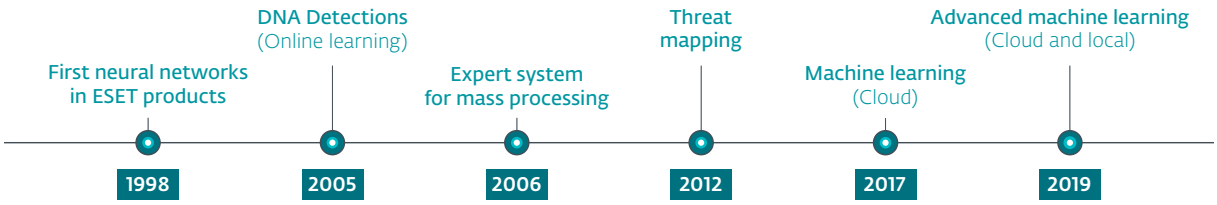


Figure 1 // Timeline of incorporation of machine learning into ESET solutions and services

The results were not only promising, but led to a completely new approach to malware detection – including heuristics and behavioral detection – that also utilized neural networks. Despite being an entirely "under the hood" change that customers would not notice in the product itself, from 1998 they greatly benefited from the improved detection and protection levels this early adoption of neural networks provided.

ESET engineers didn't rest on their laurels, though, and in 2005 announced another – and this time clearly visible – machine-learning-based technology that we named "DNA Detection". This protective layer converts the analyzed sample on the endpoint into a form more suitable for matching and detection and, by extracting precisely selected features, in essence describing "genes" for samples both malicious and benign.

Over the years, ESET's collection of DNA detections has grown to represent a complex model, splitting cyberspace into malicious and clean binaries with grey areas of "potentially unwanted". DNA detections are created either by experienced ESET experts or by automated systems leveraging machine learning. Since its inception, this regularly updated model serves as ESET's "online machine-learning model".

Inspired by the effectiveness of DNA Detections against known and emerging threats, a series of internal projects focusing on machine learning followed. At their end stood a new backend expert system designed to do the mass processing of hundreds of thousands of samples every day, as well as a series of internal ML-based tools that help ESET researchers with threat mapping.

The big changes, however, were yet to come: in the 2010s, a new paradigm started to emerge.

Tech giants started to invest heavily in machine learning as well as related technologies, allowing the whole ML field to take great leaps forward.

Big data and cheaper hardware provided the information and the infrastructure necessary to build affordable and applicable machine learning algorithms in various fields. While some of those areas instantly gained massive attention – such as driverless cars – others were quietly moving forward only to become prominent a few years later – such as ML-enabled threat detection in cybersecurity.

The growing popularity of machine-learning algorithms also led to a surge of investment into academic as well as practical research. This allowed ML technologies to become widely available and easy to deploy, and fueled more and more real-life applications of ML.

In early 2016 ESET used its years of experience with machine learning and started to shape its new and exceptionally robust detection engine in the cloud – having one significant advantage – its highly organized malware collection. After three decades of fighting black hats, ESET had what could be described as a modern-day "Library of Alexandria" of benign and malicious software. It consists of millions of extracted features and DNA genes that form high-quality training material for the machine-learning detection algorithm.

Combining these technological trends, wealth of information and human expertise, ESET has created its ML-based detection engine, which today consists of two parts:

- **ESET Advanced Machine Learning in the cloud**
- **ESET Advanced Machine Learning on the endpoint**

ESET ADVANCED MACHINE LEARNING (CLOUD)

The boom in ML has also created new challenges: With the wide array of new machine-learning algorithms, ESET experts had to test extensively and hand-pick the best-performing ones, as not all were equally suited for highly specific cybersecurity purposes. In the end, ESET settled for a mix of two approaches:

- **Processing of samples via various deep-learning methods**
- **Multi-model processing (combining an array of supervised learning methods)**

This design makes the ESET machine-learning cloud engine very robust and resilient against malicious attempts to tamper with and manipulate its decisions.

So, how does ESET Advanced Machine Learning in the cloud work? (See also Figure 2)

1. Every sample entering ESET Advanced Machine Learning in the cloud is subjected to static analysis. The engine extracts the features of the sample, collecting information that is then fed to deep-learning algorithms.
2. The sample is also emulated as a part of dynamic analysis, producing DNA genes. These are fed to a series of precisely chosen classification models and another deep-learning algorithm.
3. The sample is then executed in a sandbox[1] and subjected to advanced memory analysis. Results are then compared with a set of previously known, periodically reviewed, and automatically updated clean and malicious samples.
4. The results from the previous steps are consolidated either via a neural network or other forms of evaluation and used to produce a final decision, labeling the sample as:
   a. clean
   b. potentially unwanted/unsafe application (PUA/PUSA)
   c. malicious
5. The information is then distributed to all ESET clients either via regular update or via ESET *LiveGrid®*[2].

1 A separated and highly controlled environment where untrusted programs can be run without risking harm to the host device or operating system.
2 ESET LiveGrid® transmits newfound infiltrations and their statistical information from user's device to experts at ESET, who then analyze and process the information and send the results of their analysis to the endpoints within minutes. ESET LiveGrid® results in faster reactions to malware and a greater awareness of emerging threats.

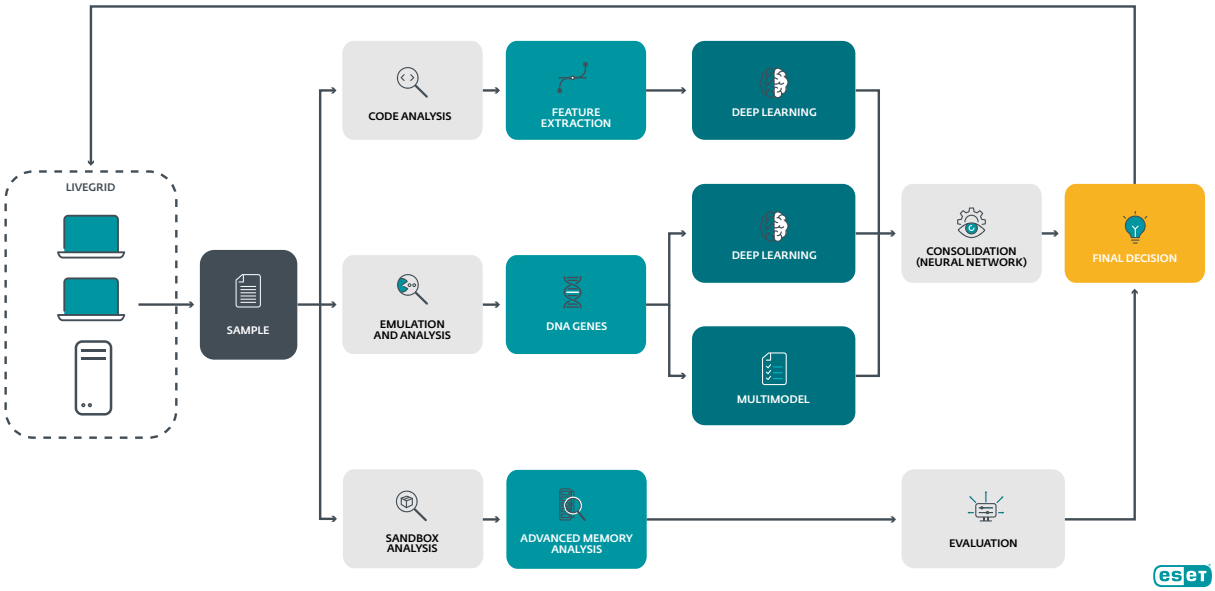


Figure 2 // Scheme detailing how ESET Advanced Machine Learning in the cloud handles a sample

It is important to note that as part of sample processing, unlike with some of the *post-truth security vendor products*, ESET also utilizes unpacking and behavioral analysis, as well as sample emulation. These steps are crucial to extracting sufficient, relevant sample features, before they can be fed to the ML engine.

Analyzing compressed or encrypted samples with no further processing is attempting to classify noise, producing meaningless results. This approach can be compared to picking a winner of a singing contest solely by looking at photos of the candidates, without giving them a chance to perform.

ADVANTAGES OF ESET ADVANCED MACHINE LEARNING (CLOUD):

Processing samples via ESET Advanced Machine Learning in the cloud has multiple advantages:

- The cloud's processing power is much greater than that of an endpoint, allowing ESET Advanced Machine Learning in the cloud to process larger quantities of data and run far more demanding computations.
- The resources of ESET Advanced Machine Learning in the cloud are flexible, and their capacity can be adjusted quickly if necessary.
- A bigger pool of resources means ESET Advanced Machine Learning in the cloud can use a small army of machine learning models simultaneously, while locally-run machine learning has only the resources of the hardware (e.g. notebook, mobile device, etc.) and hence uses only a limited number of models.
- Experienced ESET engineers constantly update, improve and monitor the performance of the machine-learning models in the cloud, thus protecting them from degradation.

Despite all the positives, even this approach has its limitations. Analysis occurs solely in the cloud, which means that any suspicious sample must be sent for processing and only after being declared good or bad can the final assessment be distributed via ESET LiveGrid®.

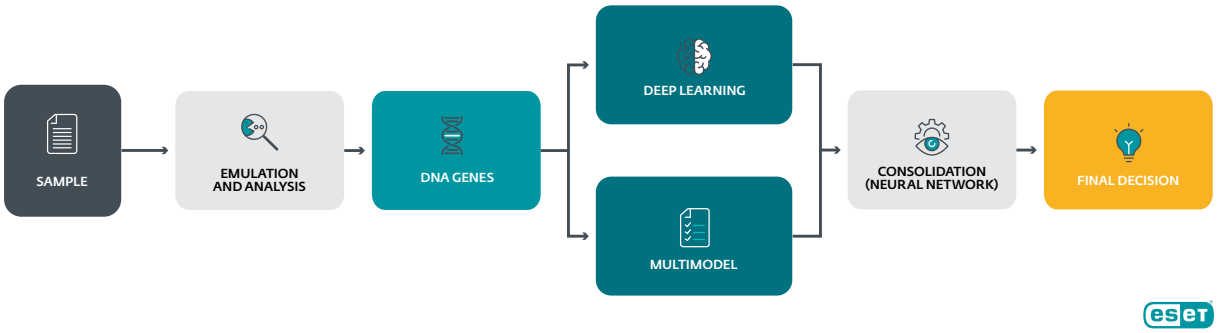
To sum it up, the cloud method yields excellent results but takes several minutes to run its course – making the method reactive – which extends the time for emerging threats to wreak havoc in the wild.

ESET ADVANCED MACHINE LEARNING (ENDPOINT)

ESET Advanced Machine Learning (AML) on the endpoint is an additional detection layer that proactively protects our users from previously unknown threats. It expands ESET detection capabilities by analyzing all suspicious samples on the user's device at the instant they are encountered.

How ESET Advanced Machine Learning on the endpoint processes the samples:

1. Local security solution encounters an unknown yet suspicious sample and scans it with help of AML on the endpoint.
2. AML on the endpoint runs static analysis, producing basic characteristics of the analyzed sample without executing it.
3. AML on the endpoint runs a dynamic analysis and extracts DNA genes describing some of sample's active features and behaviors, uncovering malicious characteristics even in packed or obfuscated items.
4. Information extracted in steps 2 and 3 is further processed by several carefully-chosen classification models and a deep learning algorithm.
5. Outputs of the ESET Advanced Machine Learning on the endpoint algorithms are then consolidated via simplified, yet still powerful, methods used by ESET Advanced Machine Learning in the cloud.



By processing samples locally, the time needed to declare the suspicious sample to be malware, potentially unwanted/unsafe, or clean is significantly shortened, but at the price of a less in-depth analysis than the one possible in the cloud. Proactive detection translates into increased protection from never-before-seen threats.

ADVANTAGES OF ESET ADVANCED MACHINE LEARNING (ENDPOINT):

ESET Advanced Machine Learning in the cloud is a demanding beast, requiring "heavy machinery" that is not available in regular user devices. Therefore, ESET engineers designed ESET Advanced Machine Learning on the endpoint as a lightweight solution, allowing it to run directly on the user's device.

This engine analyses samples locally, with machine-learning models and the consolidation of the verdict happening offline. This makes the results available to the user in real time and translates into proactive protection from unknown threats even if the user has no or limited internet connectivity.

Unlike some *post-truth vendors* who claim their models can withstand degradation without being updated for a long time, we update ESET Advanced Machine Learning models regularly to ensure the models are always trained with the latest threat landscape in mind. This ensures that our customers always enjoy the best protection possible. If ESET Advanced Machine Learning on the endpoint lacks the regional or global context necessary to make a final decision, it will consider the sample suspicious and submit it[3] – together with all the previously obtained information – for further analysis to its more powerful ESET Advanced Machine Learning in the cloud counterpart.

3 To find out more about what information is and is not sent, visit *this ESET Knowledgebase site*.

This is similar to a local bookstore that doesn't have in stock the book requested by the client but can order the title from a much bigger and better-supplied central storage. Yet, there is one significant difference if compared to ESET's solutions – the delivery time. A bookstore needs hours or days to have the book delivered, while ESET machine learning engines cooperate via ESET LiveGrid® in minutes.

WHEN CAN ESET ADVANCED MACHINE LEARNING (ENDPOINT) PROTECT YOU?

If the previous sections were too theoretical for you, don't worry. The following are real-life scenarios where this ESET Advanced Machine Learning layer on the endpoint would come into play and stop threats the from wreaking havoc.

The main problem is that, even in the internet era, users sometimes end up in a situation or location where there's no or very limited internet connectivity. However, an internet security solution must continue to protect the user, despite having little to no data about the latest threats. And believe us, being offline does not guarantee that users won't encounter new items or files that out to be malicious:

Private LAN party scenario:

Just imagine a group of gamers that organizes a private LAN party in a mountain cabin. Of course, with no internet connectivity. In preparation, one of the gamers downloads and copies a game file on a USB drive, unfortunately choosing a version with a brand-new ransomware strain injected. The compromised USB is then passed to all attendees, spreading a threat previously unknown to their security solutions, potentially ending the party before it even begins.

However, if the users in the cabin had ESET home solution with ESET Advanced Machine Learning on the endpoint installed, a quick analysis of the game file would be run. This would show that the file behaves maliciously, putting it under quarantine or deleting it. A small inconvenience for the gamers, who can easily move to another game, without having to cope with the malware fallout.

Traveler's email scenario:

Let's imagine a typical salesperson. While on the road, the internet connection is often slow and unreliable. This makes downloading emails on the notebook and working through them offline a very convenient option. But what if one of those emails is a compromised message from attackers who attached a new and previously unknown threat? The salesperson runs the file, unleashing a minor Armageddon and destroying weeks of work.

While the threat might already be known to the cloud by the time it is executed, without internet connectivity, there is no way to inform the security solution on the notebook. That is where ESET Advanced Machine Learning on the endpoint steps in. By unpacking the attachment and running static and dynamic analysis and processing the data via machine learning algorithms, it can uncover malicious intent and send the file into quarantine or delete it.

Visitor threat scenario:

ESET Advanced Machine Learning on the endpoint can be of much use to the user who runs multiple devices in a home network with close-to-zero security. An old friend comes by and wants to share his/her vacation photos, which are stored on a compromised external hard drive. The threat is activated shortly after the disk is plugged in and tries to disconnect the device from the internet and spread an unknown threat to all connected devices.

Despite poor security measures, the user is lucky enough to be running the latest version of ESET home solution with the Advanced Machine Learning on the endpoint on the original machine. Despite the malware trying to cut the machine off the internet, ESET Advanced Machine Learning on the endpoint works as intended and scans the contents of the drive for suspicious files. The culprit is quickly found and isolated, neutralizing its activities as well as its distribution mechanisms.

Home license in a small company scenario:

A hypothetical small business tries to "save a buck" and, instead of buying a business license, opts for a consumer-oriented package[4]. Its network consists of multiple devices running Microsoft Windows, including an air-gapped backup server, which doesn't have a direct internet connection. Such a machine receives no (or very sporadic) updates of the detection engine, yet it encounters new – potentially dangerous – content every day. Backups, files downloaded to the server from USB or from an external hard drive: all of those can be the source of problems.

By running the newest version of ESET home solutions on this air-gapped system, Advanced Machine Learning on the endpoint can perform scans of uploaded content. This way it can identify possible threats and act accordingly, moving suspicious items into quarantine or removing them altogether, thereby alerting business owners and protecting their valuable data.

LIMITS OF MACHINE LEARNING

At ESET we have been experimenting with various forms of machine learning since early versions of the product in the 1990s. In that process, our experts also came to learn limitations of the technology:

LIMIT: TRAINING SET

To use machine learning effectively for cybersecurity purposes, a vast number of correctly labeled samples are needed, divided into categories – malicious, clean, and potentially unsafe/unwanted applications (PUSA/PUA).

ESET's training material is a carefully chosen subset of hundreds of millions of samples collected over more than 30 years. However, even when an algorithm has been fed a large quantity of data, there is still no guarantee that it can identify all new items correctly. Thus, even the latest and most powerful machine-learning models need human expertise and constant monitoring to avoid degradation.

LIMIT: INTELLIGENT AND ADAPTIVE ADVERSARY

Another limitation to machine-learning applications in cybersecurity is **the intelligent adversary**. Sure, machines have gotten smart enough to *defeat humans at chess* and *Go*; however, these games have binding rules. In cybersecurity, the attackers do not hesitate to bend or break rules, often changing the entire playing field without a warning. Again, our human experts help mitigate these threats by constantly monitoring our machine-learning algorithms and tweaking them as needed.

LIMIT: MACHINE LEARNING IS NOT ENOUGH

Some *post-truth cybersecurity vendors* present machine learning as the silver bullet that solves all cybersecurity-related issues. After 30 years in the field and more than 20 years of experience with machine learning, ESET experts know the dangers of a security approach that relies solely on one technology – even if it is the newest machine-learning algorithm.

Only a fine-tuned blend of multiple security layers – including machine learning and human expertise – can offer the highest detection rates in combination with lowest false positive rates, providing the very best in protection.

4 While this approach strips the security solution of some aspects necessary for a basic level of business security, it offers a reasonable level of protection.

CONCLUSION

In the past 30 years, ESET engineers have been fighting cybercriminals on many fronts. In ESET's striving to create a safer digital world, these experts developed an array of effective detection technologies that comprise ESET's multi layered protection.

In 2019, with the new line of ESET home solutions, there is another addition to those layers - an improved version of our machine-learning-based model called ESET Advanced Machine Learning. This newly introduced technology consists of two parts:

1. Cloud engine, running on ESET's infrastructure
2. Local engine, protecting ESET users directly on their own devices.

This separation of tasks offers multiple advantages, with the main ones being:

- While the part of the engine running on the device protects the user proactively by analyzing and detecting emerging threats when encountered, the cloud machine-learning engine offers context and power that helps to identify even sophisticated and difficult-to-spot attacks.
- Also, the local engine protects the user from any never-before-seen threats even when the internet connection is unreliable or non-existent.

As indicated by the name of the technology, it is built on an array of modern machine-learning algorithms, offering users a combination of best possible detection results and a robust solution able to withstand external attacks. To prevent these machine-learning models from degradation, our vigilant engineers are always monitoring the models' performance and correcting any deviations that might arise. This provides ESET solutions with an ideal mix of human expertise and machine performance necessary for rapid and reliable protection.

With all the benefits of the ESET Advanced Machine Learning engine in mind, we just need to add one more thing. There is no silver bullet in cybersecurity and even the best protective technology can fail, if it isn't backed up by experienced human analysts. In the more than three decades of its existence, ESET has found and built it isn't backed up by multiple other layers and experienced human analysts.

ABOUT ESET

For 30 years, *ESET®* has been developing industry-leading IT security software and services for businesses and consumers worldwide. With solutions ranging from endpoint and mobile security, to encryption and two-factor authentication, ESET's high-performing, easy-to-use products give consumers and businesses the peace of mind to enjoy the full potential of their technology. ESET unobtrusively protects and monitors 24/7, updating defenses in real time to keep users safe and businesses running without interruption. Evolving threats require an evolving IT security company. Backed by R&D centers worldwide, ESET becomes the first IT security company to earn *100 Virus Bulletin VB100* awards, identifying every single "in-the-wild" malware without interruption since 2003. For more information, visit www.eset.com or follow us on *LinkedIn*, *Facebook* and *Twitter*.

ENJOY SAFER TECHNOLOGY™