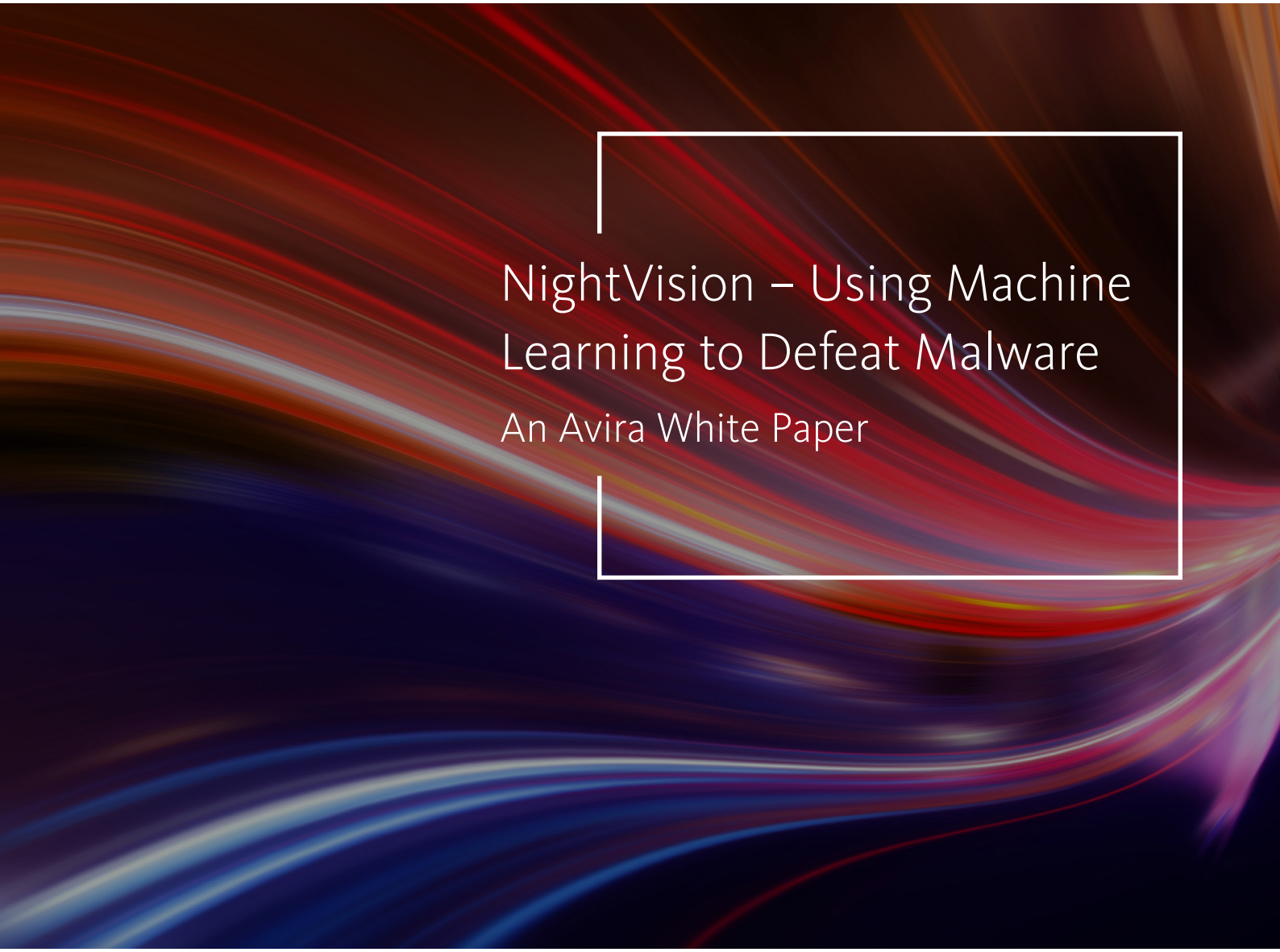
Avira



NightVision – Using Machine Learning to Defeat Malware

An Avira White Paper

www.avira.com

Index

3
Using Artificial Intelligence to Enhance Local Scan Engines
Avira's Approach to Malware Detection

4
NightVision, Part of The Avira Protection Cloud

5
NightVision: A Guide to Avira's AI platform

6
Identifying Malware using NightVision
The Clustering Process

7
Feature Selection
Anomaly Detection

8
Classifier Optimization

9
The Importance of Retraining NightVision

10
What's Next for Machine Learning in Cybersecurity?

Introduction

To a large extent, our online presence today defines who we are. We talk, conference, shop, bank and share personal photos and experiences online – almost everything we do or know leaves a digital footprint. This online presence is a passport to an individual's personal and professional life, often containing the most private of identity information, as well as confidential business data. As we increasingly extend and enrich the virtual data belonging to our lives, the opportunity for cybercriminals to profit by attacking our personal and business presence continues to escalate. We expect to see the pace of cybercrime accelerate indefinitely as it rises to become the predominant criminal threat to our everyday lives.

Using Artificial Intelligence to Enhance Local Scan Engines

The reality for both consumers and businesses using internet-connected devices is that they will need to deploy defenses against increasingly sophisticated forms of criminal activity. Today's first line of defense is often the anti-malware scanning engine on a laptop or mobile device. As one of the world's largest providers of cybersecurity solutions, today, Avira protects nearly 100 million personal users, analyses more than 50 million suspicious files daily, and protects its customers against more than 300,000 malicious files a day – all using a combination of local scanning and cloud-based security services.

However, as the pace of cybercrime intensifies, so too does the burden on threat researchers to analyze and create protection rules to identify malware. Similarly, the frequency and size of the updates being rolled out to each endpoint solution will increase, calling for new techniques to be developed to ensure the local engine remains manageable. As the number and nature of these threats increase, cybersecurity providers are recognizing the need to identify and implement new approaches to successfully defend against these risks. However, the very nature of modern cybercrime dictates that security vendors are always going to be to some extent reactive – after all, a threat must first be created in order for it to be identified and mitigated. Consequently, the cybersecurity industry has started to look beyond the capabilities of local malware scanning engines and heuristics, towards the application of Artificial Intelligence (AI) to scale threat analysis and undertake it at speeds far beyond what it is humanly impossible. We've talked about the different approaches to AI (and how it relates to Machine Learning) in another Avira whitepaper, so here we'll delve further into how Avira has incorporated AI into its malware detection and prevention solutions.

Avira's Approach to Malware Detection

Avira is one of a few vendors that uses its global customer base as a sensor network; a sensor network that identifies threats as they appear in real time, around the world. This sensor network, comprising of malware scanning and local machine learning, works together with Avira's powerful cloud-based security service to create a "Swarm Intelligence"- a network of nearly 100 million sensors.

Each day, 50 million unknown or suspicious files are directly uploaded by customers of Avira's malware detection services to its cloud-based security service: Avira Protection Cloud. This service, based in Germany and accessible

globally, analyzes and detects new malware submitted from Avira's endpoint sensors distributed around the world. This constant flow of new intelligence, allows the service to perpetually evolve as it processes the intelligence from variants of existing and new threats released into the wild by malware authors.

The Avira Protection Cloud holds comprehensive details about how each threat operates, how different malware families and coding patterns interlink and relate to one another, and how tiny obfuscations in code can uniquely alter the behavior and classification of a piece of malware. This type of data is the holy grail for the subset of AI that we use – Machine Learning – which forms the basis for our NightVision AI platform.

Machine Learning is a subcategory of AI that focuses on developing computer programs that can continuously adapt and improve when exposed to new data. In the world of cybersecurity, this form of AI complements the type of manual data analysis undertaken to detect and identify malware threats, and is explained in more detail in the Avira whitepaper 'The Application of AI to CyberSecurity'.

Avira does not rely on a single approach to the problem, but uses an ensemble of different Machine Learning techniques, ranging from linear models such as logistic regression to nonlinear models such as kernelized support vector machines, random forests and, for problems where it is the best choice, Deep Learning techniques such as convolutional neural networks. Those techniques are applied for different detection tasks including malware detection and phishing detection, depending on the needs of the user and the capabilities of the underlying platform.

What makes our approach so successful is being able to combine our expertise in Machine Learning and AI with 30 years of experience in the cybersecurity industry. One of the most essential requirements for any Machine Learning method to work is data. At Avira, we have this data 'in spades' as we maintain databases containing hundreds of millions of malicious files, continually updated with fresh intelligence. It is not as simple as purchasing a database with basic details of every threat; it's about collecting and analyzing every single file to gain a comprehensive understanding of how each one operates, how different malware families and coding patterns interlink and relate to one another, how tiny obfuscations in code can uniquely alter the behavior and classification of a piece of malware. AI cybersecurity solutions can only be as good as the data being fed into the Machine Learning platform, and our data is vast, rich and complex - this is the fuel powering our Machine Learning engine.

NightVision, Part of The Avira Protection Cloud

As cybersecurity vendors augment their security solutions with increasingly sophisticated technology, cyber criminals will be looking to better obfuscate their malicious code to evade the latest detection techniques. One of the most successful techniques harnessed by malware authors is getting their executable files to obscure their true purpose. If a piece of malware can convince a security solution that it is not a malicious file, it stands a chance of remaining unidentified – that means the author must ensure that the file displays as many of the same attributes as that of a legitimate file.

In some cases, a malware author may be able to successfully mirror most file attributes, leaving a question mark over how the file should be classified, or they can encrypt the file, meaning that the scanning technology may not gain full access to the different file attributes required to deliver classification. This is where Avira's approach differs from some other security vendors which, at this stage, will return the file to the endpoint with a 'score' outlining the likelihood of it being malicious. If the file is scored as likely to be malicious, the end-user is blocked from executing

the file, but there's a chance of a false-positive occurring. If the file is identified as safe, the end-user can execute it, but there's still the chance that it may have a malicious payload.

Instead, Avira takes a series of steps to identify the true nature of the file, using sophisticated detection systems deployed within the Avira Protection Cloud. We use powerful cloud-based scanning engines and NightVision, our AI system. We use 'autodumping' to remove layers of obfuscation – such as encryption – to determine whether the file harbors malware. If uncertainty about the nature of the file still remains, we use an analysis sandboxing technology that emulates the end-user client device so that the file can be 'detonated' in a virtual environment. These systems represent a few of the elements that make up the entire Avira Protection Cloud system. The essential issue is that if any malicious behavior is detected, the information is fed back into NightVision to further train it so that if the same behavior is spotted later, the AI will know to flag it as malicious.

Infrequently the automated detection systems within the Avira Protection Cloud may not arrive at a definitive conclusion as to whether a file is malicious. In this case AI and human expertise will come together. At Avira, a malware analyst will conduct a thorough analysis of the malicious file, based on its behavior as well as an assessment of how it managed to obscure its true purpose. The analyst will use this information to create an additional file attribute – a key identifier that marks the piece of malware out from the legitimate file it was mirroring. They can then feed this additional file attribute into NightVision so that it retrains itself to spot the new piece of information, not just within this specific file, but across the 250+ millions of files already contained in Avira's vast database.

NightVision: A Guide to Avira's AI platform

NightVision is a hugely powerful, cloud-based AI technology platform. It is capable of analyzing files against more than 8,000 different attributes (or dimensions), supported by our database of more than 30 years of malware classification data and resides within the Avira Protection Cloud. The NightVision platform trains on a cluster of servers utilizing thousands of CPU cores and many terabytes of RAM. Because the platform does not rely on one

approach, but can use a number of different techniques that can be optimized for the required environment, it can run in the Cloud, but also form part of our distributed network of sensors, forming a 'swarm intelligence'. The result is a highly efficient model with a detection engine that delivers very high zero-day detection rates and exceptionally low false positive rates of less than 0.001%.

Identifying Malware using NightVision

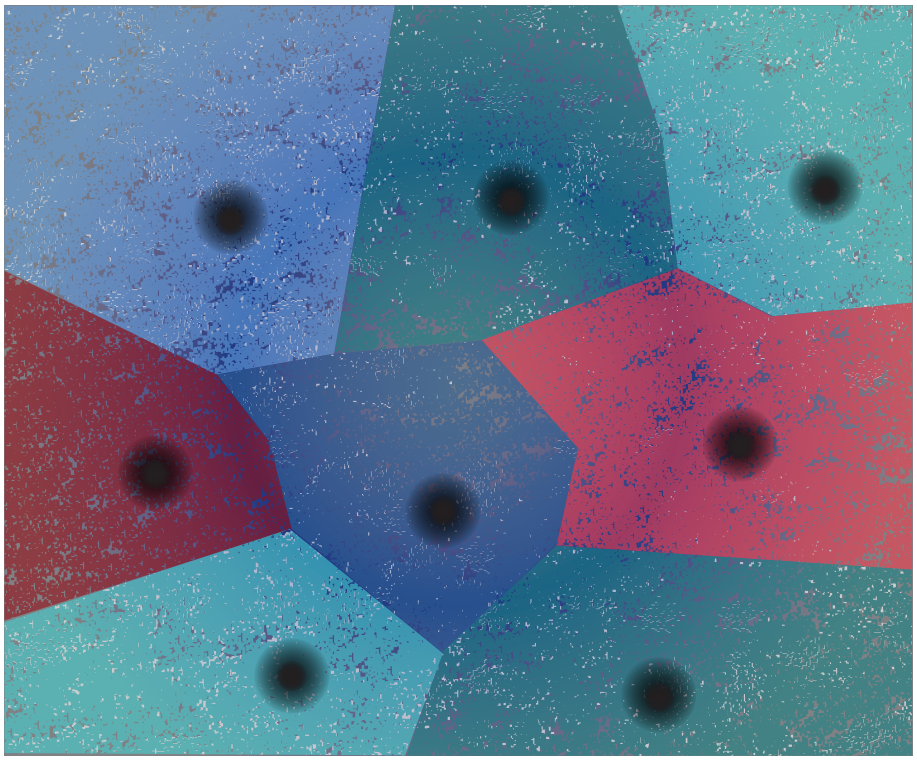
When our global sensor network – and other detection elements in the Avira Protection Cloud – identify a file as potentially malicious, several processes and algorithms are used to classify the file as clean or malicious. NightVision uses the process of both unsupervised and supervised learning. Unsupervised learning is used when NightVision groups all of the files, regardless of type into clusters with similar attributes. Supervised learning then separates the clusters using decision surfaces into clean or malware types.

The Clustering Process

The first step of analysis is to analyze unknown and suspicious files extremely quickly against more than 8,000 different attributes (or dimensions), supported by our database of malware classification data. These attributes comprise everything from the basics – such as file section size or entropy (obfuscation) to those derived from the structure such as anomalies created by intended or unintended modifications to files artificially created by the malware author. Often these are highly specific attributes known to Avira resulting from our decades of cybersecurity experience. They offer a strong indication of which family a particular piece of malware belongs to, helping to make the jump from an individual unknown file to one that can be quickly categorized in relation to thousands of others.

We undertake this process – called clustering - to all the files in our database – some 200 million malicious and clean samples creating 200 million points in a vector space classified by 8000+ attributes. Note that we deliberately apply an approach of using feature extraction based on hand-crafted features since this enables us to incorporate our deep technical experience in malware analysis. This is in contrast to other techniques that try to automatically extract features from the file, where a vendor lacks data, and a database of malware has not been built.

Because the objective of the processing during the clustering phase is to ensure files are dealt with as quickly as possible, at this point NightVision isn't concerned whether the file is clean or malicious, it simply looks to partition the files into clusters based on similarities. To partition the files into clusters we use a variant of a standard technique, k-means, which simply partitions the vector space into k groups of points represented by their mean value. However, what is highly non-standard is the way we implement the algorithm. To speed up the computation of the k-means clusters, we use an in-house developed technique to address the specifics of the algorithm, speeding up the clustering process by a factor of up to 20. This approach was recently published at the International Conference on Machine Learning (ICML), the premier conference in the field, and is the current state-of-the-art in k-means clustering [Bottesch et al, 2016].



The process of clustering files around their mean attributes will identify different 'regions' in the file landscape by clustering together files having similar attributes. The next goal is to learn how to distinguish malicious and clean files within these regions. At this point, the immense scalability of NightVision's processor core is activated and we now look at the clusters in parallel using a number of supervised Machine Learning algorithms.

Clustering of files around their mean

Feature Selection

The mRMR algorithm (minimum Redundancy, maximum Relevance) is used as an independent filter to identify the most important features, and optimize the model for the environment and maximize performance. It selects features which are "far away" from each other (minimum redundancy) while having a large correlation with the target variable i.e. whether the file is clean or malicious (maximum relevance). The selection of good features for classification can be further improved by utilizing clever merging schemes [Bottesch & Palm, 2015].

Anomaly Detection

A different learning technique is used to discover anomalies in the data. This utilizes a currently unpublished technique also developed in-house. In this context, anomalies are considered to be points in highly dense regions of the vector space which disagree in their label, corresponding to a potential mislabeled sample or deliberate attempt of malware authors to poison data. In any case, those anomalies are automatically detected, removed and sent to our malware researchers for manual inspection. The benefit of this technique is that we can address two key issues at once; making our classifier more robust and at the same time detecting new malware threats. In this way, we can combine our Machine Learning technology with our experience in malware research to quickly adapt to new threats.

Classifier Optimization

Finally, a classifier is learned to find a decision boundary enabling us to distinguish between clean and malicious files. Here we have a flexible architecture, where multiple classifiers can be applied, including Logistic Regression, Random Forests and kernelized Support Vector Machines. The choice of the classifier as well as number of features selected in previous steps control the trade-off between detection rate and size of the model as well as detection speed. In the current generation of NightVision running in the Avira Protection Cloud we use the Random Forest algorithm to develop the final classifier for each cluster, which accumulates numerous decision trees on the criteria of the file attributes. Each individual decision tree may not be particularly stable, but together the forest makes the analysis highly robust and accurate. The resulting overall classifier is then a combination of numerous Random Forest models, each providing a nonlinear separation of the corresponding region in the vector space.

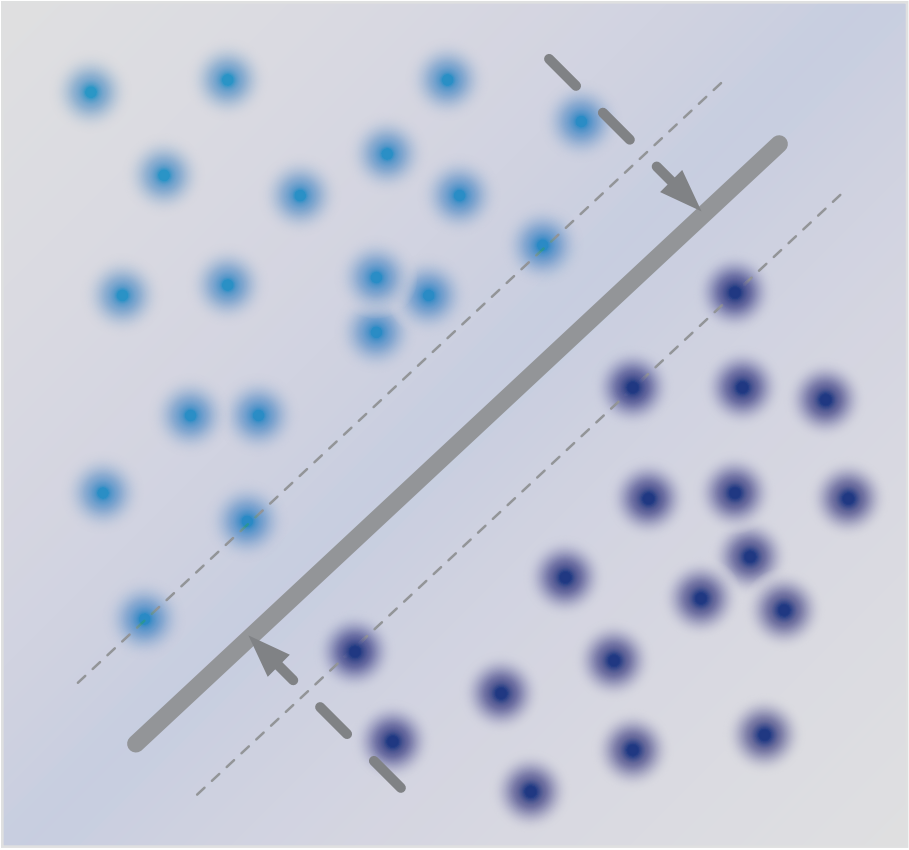


Diagram: Creation of a Decision Boundary: A simplified representation of the process of finding a decision boundary is shown. The algorithms used find a structure in the data, where clean files group together, as do malicious files. This structure can be represented by a decision boundary (which happens to be a straight line in the simplified example above, but in general it would be a 'surface' in three dimensions or a hyper-surface in more dimensions). The term 'decision' boundary stems from the fact that later NightVision can recall this surface and use it to decide whether a file lies on the malicious or clean side of the boundary – NightVision has learnt the underlying concept of the position of the files within the decision surface, not just that dark blue is malicious and light blue is clean.

About NightVision

The NightVision platform is underpinned by ~8TB of RAM, ~1000 CPU cores and is located in Germany, offering compliance to some of the strongest data privacy laws in the world. Currently a third-generation technology, it was originally developed by Avira nearly a decade ago, and has evolved significantly in this time, both in terms of scale, but also in terms of architecture.

The Importance of Retraining NightVision

If AI is being developed to undertake tasks that are usually carried out by people, it stands to reason that, in the same way humans need regular training and development to perform competently, so too will the AI. Machine Learning-based AI platforms rely on learning from data to retrain; thus, these AI platforms can only be as good as the data that is fed into them. In the case of cybersecurity, the AI needs a reliable and substantial dataset to learn how to identify which files are malicious and which are not.

In the case of NightVision, our globally distributed sensor network ensures that our dataset is continually enhanced with emerging malware data. The more data that feeds into the dataset, the more accurate NightVision becomes, as it has more inputs from which to learn. Just as malware authors continue to update their methods of evasion by finding new ways to obfuscate malicious code, the AI platform also needs to retrain itself to identify these new patterns and file attributes.

One advantage of our NightVision pipeline is the ability to continuously adapt to new threats. To achieve this, we have a parallel architecture, where two training paths run in parallel. The first path is full training which performs all the steps in the pipeline described earlier. With the current setup of our compute cluster consisting of thousands of CPU and terabytes of RAM, this process takes around 8 hours. In parallel, we run continuous retraining to be able to quickly adapt to new threats, described below.

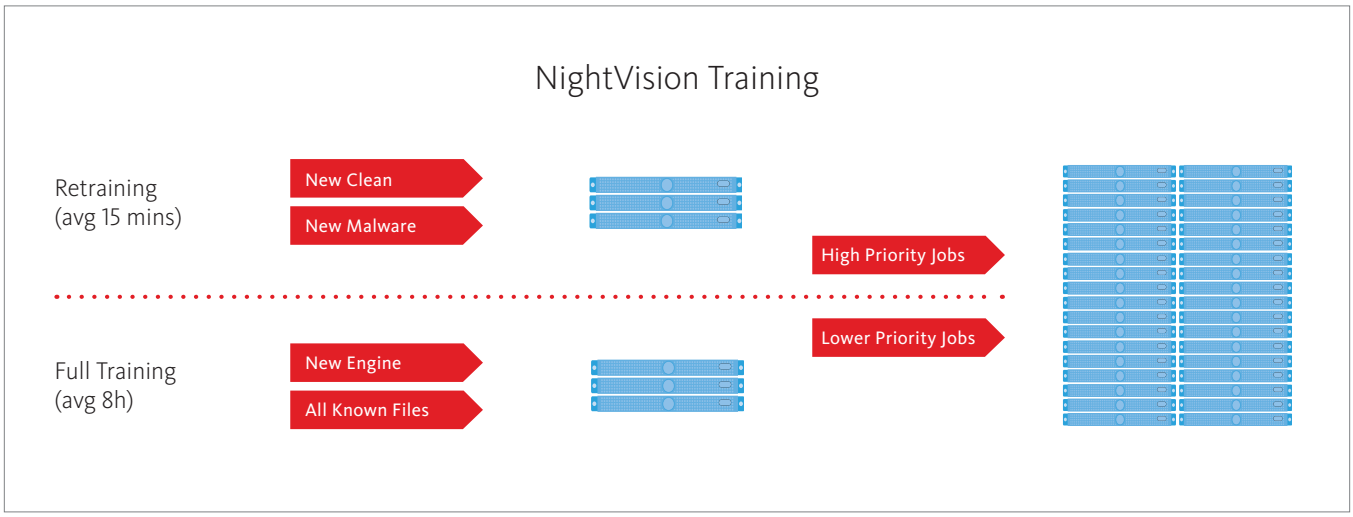


Diagram: NightVision Operates Two Paths of Training

A file may get marked for retraining for a number of reasons: It may be manually reclassified after inspection by an experienced malware researcher, or a different detection engine may flag the file as suspicious. When this occurs, the file is send to the NightVision Machine Learning Engine. Having already computed a k-means clustering model, NightVision can assign the sample to the nearest cluster. This avoids having to retrain the whole classifier and allows us to focus our attention on the particular region within which the file lies. Assigning the file to the cluster means the cluster model recomputation occurs very quickly because of the small size of the region. On average, this whole process takes around 15 min, after which we have updated our classifier and are ready to deploy a new model to our customers. This is how we achieve such exceptional performance and reach such consistent high levels of accuracy in malware detection.

For every new file that is submitted, NightVision will make a prediction as to its nature within a few milliseconds to prevent any file suspected of being malicious from executing. Within a few minutes NightVision will have completed a thorough analysis, constructed a classifier for the file type, and will then need to learn the classifier by retraining itself. NightVision does this automatically many times each day. It takes two to three minutes, and at no stage does the technology go offline, ensuring that customers remain protected at all times.

Retraining is an essential aspect of honing AI platforms, and when an entirely new file attribute is created, a complete relearn will ultimately be required to identify this attribute across the Avira database of more than 250 million files. This complete relearn takes approximately eight hours and will typically occur two or three times per week. Again, at no stage will the system go offline; it will seamlessly switch over to the retrained AI once the relearn is complete.

Why Fast Retraining is Essential

The ability to quickly adapt to new threats is important as there is always a risk that any Machine Learning system may misclassify a file. Should this occur – no system can ever realistically claim to be perfect - slow retraining times (that occur for example in pure Deep Learning approaches) result in users being vulnerable to new malware for extended periods. Fast retraining allows updates to be delivered many times a day, dramatically reducing the vulnerability that potentially exists with systems that only retrain daily or weekly.

What's Next for Machine Learning in Cybersecurity?

Will Machine Learning platforms will ever take over from humans altogether and assume sole responsibility for malware detection and prevention?

At Avira, we believe that analysts will always need to be part of the protection program, to provide a complementary layer of analysis and oversee the continued learning and adaption of machine learning systems. We have already deployed machine learning across a myriad of device types – from endpoints to cloud systems complementing and enhancing other methods of analysis. For now, all we can be 100% sure about is that AI cybersecurity solutions will only ever be as good as the data being fed into the AI platform which is why we place so much significance on our constantly evolving, 30-year old database. Using this to fuel NightVision – as part of the overall Avira Protection Cloud – we have been able to continue reducing the burden on our research team by automatically identifying most of the new threats submitted. Not only that, but as AI can process data far more quickly than humans, we can use NightVision to ensure that our vast customer base continues to benefit from rapid response to all emerging threats, no matter how quickly the pace of threat creation continues to increase.

NightVision rapidly reduces the time between a new malicious file being reported by a single endpoint, to identifying the threat and rolling out protection to our entire user base. Whereas in the past, zero-day threats had the potential

to hit vast numbers of users before they were successfully classified and mitigated, today we can use our AI platform to reduce this number to the bare minimum, ensuring that our user base receives protection against the threat long before it ever comes near them.

Perhaps the biggest positive of all is that, as an AI-platform rather than a manually-updated, human-powered database, NightVision continues to evolve at the same pace as the threat landscape. The more unknown threats that emerge, the more unknown threats get processed and the more NightVision will retrain, thus allowing us to ensure we can continue to afford our customers 100% percent protection against known and unknown threats into the foreseeable future.