TECHNICAL FIRST LOOK

JUNE 2025

# Index Engines CyberSense Validated 99.99% Effective in Detecting Ransomware-induced Corruption

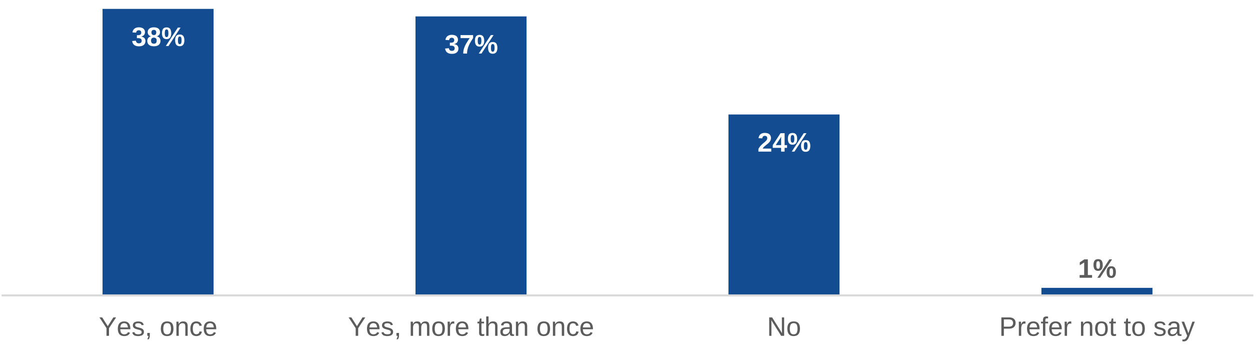
Alex Arcilla, Principal Analyst – Validation Services

## Ransomware Challenges

Ransomware continues to plague organizations, and it’s not slowing down. According to research from Enterprise Strategy Group, 75% of respondents to a recent research survey experienced at least one ransomware attack on a monthly, weekly, or daily basis in the past year.[1] And, unfortunately, 75% of organizations experienced a successful ransomware attack that resulted in operational disruption or, worse, financial impact (see Figure 1).[2]

**Figure 1.** Successful Ransomware Attacks Cause Real Pain for Organizations

**Has your organization been the victim of a successful ransomware attack, meaning an attack that had a negative financial impact or disrupted business operations, within the past 12 months? (Percent of respondents, N=451)**



*Source: Enterprise Strategy Group, now part of Omdia*

To minimize the impact of ransomware attacks and improve cyber resiliency, organizations have been adopting AI and machine learning (ML) into their data and recovery processes. In fact, 53% of respondents cited data backup or snapshots with security as a primary use case that will be in place at their organization within 24 months for AI/ML in backup and recovery.[3]

Recovering data quickly and effectively requires organizations to ensure that copies of data are free of ransomware, malware, and ransomware-induced corruption. One approach is to use AI/ML to detect and verify the presence of corrupted data caused by ransomware in backups or snapshots. However, the effectiveness of using any AI/ML

[1] Source: Enterprise Strategy Group Research Report, [Ransomware Preparedness: Lighting the Way to Readiness and Mitigation](), December 2023.
[2] Ibid.
[3] Source: Enterprise Strategy Group Research Report, [Reinventing Backup and Recovery With AI and ML](), June 2024.

process is dependent on how rigorously and continuously the supporting AI/ML models are trained to detect the most up-to-date and sophisticated ransomware attack patterns with a high degree of accuracy.

## Index Engines CyberSense

Index Engines CyberSense detects ransomware-induced data corruption within backups and snapshots with 99.99% effectiveness. CyberSense analyzes and scans data continuously, while offering alerts for both clean and corrupted data. When an attack or corrupted data is detected, organizations are notified of the attack details and CyberSense creates detailed post-attack forensic reports to support quick and informed recovery efforts. The key to the accuracy in CyberSense is its proprietary AI/ML engine trained with petabytes of data from both real-world and simulated attacks in the Index Engines CyberSense Research Lab, which analyzes actual ransomware in controlled environments to build realistic, reliable data.

CyberSense uses highly trained AI to perform deep, byte-level analysis of files in backups, snapshots, and core infrastructure, including storage platforms and databases. By tracking both content and metadata changes over time, CyberSense detects patterns of ransomware corruption. This solution goes beyond traditional tools that rely solely on metadata or compression thresholds, providing a more reliable indication of cyberattack-driven data loss.

Trusted by well-known storage and backup vendors, organizations both large and small can enable direct scanning of their backups without the need to rehydrate data. CyberSense proves its value to customers based on its experience in the field and its strategic partners to provide a more comprehensive cyber resiliency strategy.

## First Look – Upholding the Standard

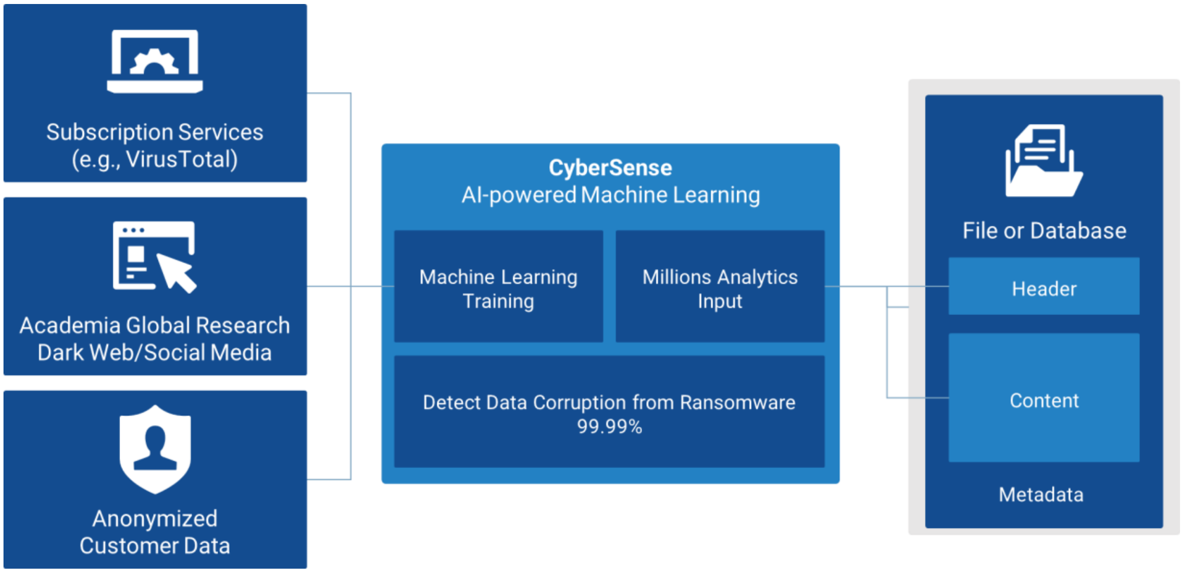
To provide the detection that organizations demand, Index Engines continually and rigorously trains CyberSense's AI/ML engine to recognize current and emerging patterns of ransomware attacks. Recently, Index Engines was awarded U.S. Patent #12248574 for "AI Identification of Computer Resources Subjected to a Ransomware Attack." The patent describes a method to train and validate AI models for ransomware corruption using actual attack data and real-world testing to ensure reliability. It is this method that enables CyberSense's accuracy rate, which Index Engines estimates to be 99.99%.

To further validate that the quoted 99.99% is an accurate, realistic estimate, Enterprise Strategy Group evaluated the process that Index Engines uses to train and test CyberSense's ML engine—specifically, how Index Engines acquires data, generates data sets, trains the supervised ML models, and calculates the accuracy rate.

### Data Acquisition

We first reviewed how raw data is acquired to create the training data set for CyberSense's ML engine. As shown in Figure 2, Index Engines acquires data for creating the training data sets from three main sources: subscription services (e.g., Virustotal.com for ransomware executables, Wayback Machine for examples of clean files and associated changes over time); public sources (e.g., academia, global research from third-party organizations, the dark web, social media); and anonymized Index Engines customer data. A portion of the customer base of varying sizes and verticals opt in to supply data daily through the Index Engines' private cloud.

**Figure 2.** CyberSense's AI-powered Data Analysis and Machine Learning


Subscription Services (e.g., VirusTotal)
Academia Global Research Dark Web/Social Media
Anonymized Customer Data
CyberSense
AI-powered Machine Learning
Machine Learning Training
Millions Analytics Input
Detect Data Corruption from Ransomware 99.99%
File or Database
Header
Content
Metadata

*Source: Enterprise Strategy Group, now part of Omdia*

Enterprise Strategy Group took note of the variety and volume of raw data acquired, as these factors affect the quality and completeness of the training data set. We saw that:

- Index Engines detonates executables, both manually and automatically, via scripting to reveal ransomware attack patterns that the ML engine can use to learn.
- To minimize false negatives, Index Engines adds in its own files from backups of both clean and infected hosts to the training set.
- Index Engines leverages over 200 statistics—such as file properties and number of files added, deleted, or modified over a given period—to characterize how the population of files in all backups change over time.
- Over 7,500 ransomware variants have been identified to train the ML engine, categorized into three behaviors: files in which data has changed with no file name preserved, files in which the file name has changed with known ransomware extension, and files in which the file name has been obfuscated or modified.

## Training Data Set Creation

The effectiveness of CyberSense in detecting ransomware corruption relies on the data set used to train the ML engine. To evaluate how the training set is created, Enterprise Strategy Group evaluated Index Engines' Statistical Analytic Generation (SaGen) process.

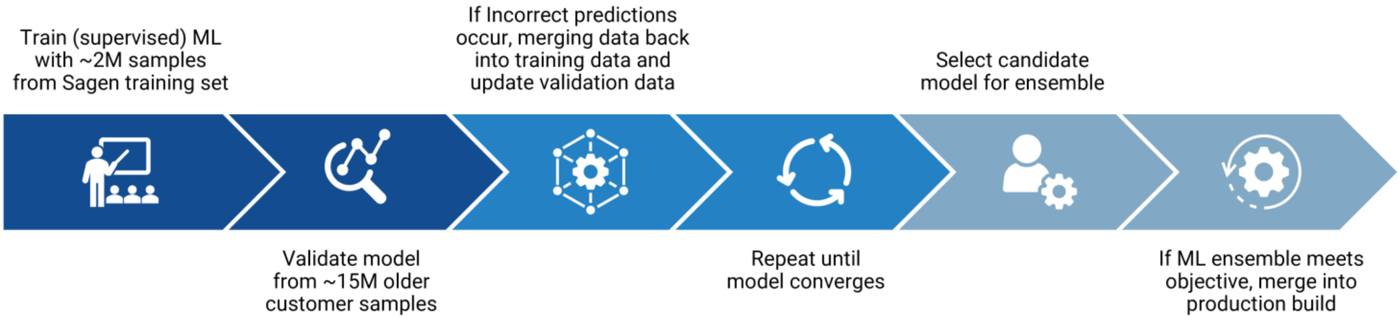
Using Index Engines' native indexing and query capabilities, the ML engine first learns and analyzes how files change from a clean to infected state using clean/detonated backups. Using this knowledge, Index Engines artificially creates millions of different backup scenarios (performing incremental and full backups, varying operating systems and file types). Events that identify unique sequences of changes that can occur within backup files are classified as either normal operations or attacks. Millions of virtual backups representing artificial backup scenarios are generated; up to one million can be created per day. CyberSense then processes and analyzes these backups to create the training data set.

Enterprise Strategy Group noted how comprehensive the SaGen process is for creating a data set for training CyberSense to uncover ransomware corruption in IT production networks.

## ML Engine Training and Validation

To ensure that the ML engine is properly trained, Enterprise Strategy Group noted the thorough and sound approach taken by Index Engines (see Figure 3).

**Figure 3.** Training and Verifying CyberSense's ML Model Ensemble


Train (supervised) ML with ~2M samples from Sagen training set
Validate model from ~15M older customer samples
If Incorrect predictions occur, merging data back into training data and update validation data
Repeat until model converges
Select candidate model for ensemble
If ML ensemble meets objective, merge into production build

*Source: Enterprise Strategy Group, now part of Omdia*

CyberSense's ML engine is an ensemble of 10 AI/ML models, each building off the results of the others during the training process until results converge. ML training is first conducted with approximately 2 million samples generated from the training set. Index Engines then validates the ML engine from approximately 15 million customer samples. Any incorrect predictions are merged back into training data, which will update the validation data. This repeats until the ensemble model converges.

For the ML engine's most recent training, approximately 6,000 samples were merged back into the training data. Subsequent passes quickly reduced it to fewer than 10 samples after three to four iterations.

Once converging, the candidate model group was selected to test on customer scenarios that were deemed difficult (based on specific customer scenarios that were not initially detected on previous releases), then tested on an additional 30 million new customer samples (i.e., actual backups occurring in the field).

Accuracy is currently estimated to be 99.997% based on the fraction of true positives identified from the total true positives and false negatives identified within 160,000 data samples used for model validation.

## Conclusion

Ransomware prevention strategies are critical to implement but are rarely 100% effective. Detecting ransomware-induced data corruption is paramount to minimizing the impact of a ransomware attack when prevention fails. While adopting AI/ML can help to further secure data backups, specifically in detecting ransomware corruption and the presence of ransomware, organizations must ensure that the supporting models are continuously trained with data reflecting existing and emerging ransomware corruption patterns.

Enterprise Strategy Group validated that the approach Index Engines takes to creating the data sets and training the ML models for CyberSense is as complete and thorough as possible, justifying the reported 99.99% accuracy rate. To minimize the impact of ransomware attacks and boost cyber resiliency, Enterprise Strategy Group advises organizations to use well-trained AI/ML to detect data corruption early and recommends CyberSense as solution to consider.

Information contained in this publication has been obtained by sources Informa TechTarget considers to be reliable but is not warranted by Informa TechTarget. This publication may contain opinions of Informa TechTarget, which are subject to change. This publication may include forecasts, projections, and other predictive statements that represent Informa TechTarget's assumptions and expectations in light of currently available information. These forecasts are based on industry trends and involve variables and uncertainties. Consequently, Informa TechTarget makes no warranty as to the accuracy of specific forecasts, projections or predictive statements contained herein.

**About Enterprise Strategy Group**
Enterprise Strategy Group, now part of Omdia, provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

✉ contact@esg-global.com

www.esg-global.com