# Using Machine Learning to Detect Ransomware Attacks on Electronic Health Records (EHRs)

*Author: Miguel Duran Zamora*
*Advisor: Dr. Jeffrey Duffany*
*Master in Computer Science*
*Graduate Project EXPO, October 2025*

## Abstract

Ransomware attacks have always been a burden for many industries, and nowadays, with the advancement of technology and computers, they have taken over everything. One crucial industry is the health industry, which has been a victim of these attacks for a while. Could we explore ways to detect and mitigate these attacks in critical areas, such as health? That's why in this project, we will discuss how we can use machine learning to detect ransomware attacks on electronic health records.

## Introduction

Healthcare attacks have increased throughout the years, not only because of the advancement of technology but also because of ways to protect and defend against them. According to the NIH, the PIH data breaches went up from 4% to 81%, and surprisingly enough, from 2010 to 2024, ransomware attacks went down from 31% to 11% [1]. You might think the decrease means it is safe, but as stated before, when attacks increase, protection increases, but there is still more we can do to detect and mitigate these attacks even more. With the increase in the use of machine learning (ML) in all industries, we can make sure we catch these early ransomware attacks against healthcare infrastructures and create systems that automatically fight against them without human intervention.

## Background

The reason I picked this project was because, as a cybersecurity student and having spoken with family who are in the healthcare industry, I learned the serious need for more security that protects the integrity and confidentiality of patients. Also, diving more into the world of AI and ML, I felt there is more we can do with these to make the security of EHRs even stronger than before. I strongly believe that we must use new technology and tools that meet industry standards. Also, the automation of processes is key because when we have things like generative AI where you can write a prompt and it generates an answer, imagine that but for ransomware attacks where you type the ransomware specification and automatically create and execute the ransomware without human intervention. That will be a game changer in the field of cybersecurity and that's why it's crucial for us to prepare for those days.

## Problem

A problem I found that inspired me to do this project is the lack of automation when it comes to security threats like ransomware, especially in the health industry. I believe that we should implement, with the use of ML, more autonomous detection and mitigation systems, and with the use of ML now, it's possible. ML will help us detect ransomware patterns in these EHRs, and then we can use these ML models in artificial intelligence to create systems that counterattack these attacks autonomously. Not only will this protect the integrity of the patient's 1 data, but it will also save the industry money and resources.

## Methodology

For this article, we started by creating a VM in a Windows 10 laptop where we simulated ransomware attacks. The sample used for this project was Yashma ransomware, which is a variant of the Chaos ransomware. This ransomware was the sample selected due to its fast encryption and ease of use. The specifications of the test machine are:

| | |
|---|---|
| CPU | Intel Core i7-8550U 1.80GHz |
| RAM | 16GM |
| Network | Intel Dual band Wireless AC-7265 |
| SSD | 512GB |
| Operating System | Windows 10 |

After I ran the ransomware in the secure VM then Sysmon will start capturing the activities or events. In this case, the events used were:

| | |
|---|---|
| 1 | Process Created |
| 2 | Process changes the file creation time |
| 11 | File Creation |
| 12 | Modifies registry keys and values |
| 23 | Detect file deletion and archive it in C:\Sysmon |
| 26 | Same as the previous event, but doesn't archive |

Then using a python script, I extracted some features which are:

| Feature | Description |
|---|---|
| File Deleted | Detects deleted files and backup files (EvendId: 26). |
| File Created | This is the most common because ransomware when they encrypt a file is categorized as file creation (EventID: 11). |
| File Create Time Changed | This happens when a process changes the creation time of a file (EventId: 2). |
| Process Create | This happens when a new process is created (EventId:1). |
| Suspicious Path | This checks if the Image or path from where the process is coming contains known suspicious keywords. |
| System Executable | Checks if the process is being executed from a known system path or Image as c:\windows\system32. |
| Path Length | Get the length of the path. |
| Directory Depth | Check the directory and see how deep in the directory the process is. |
| Process Length | The length of the process. |
| Extension Similarity | Checks if extensions of the target files are similar. For example, some ransomwares encrypt the file into .encrypted. |
| Filename Entropy | This calculates the Shannon entropy of the base name of the path. In other words, it checks the randomness of that path. |
| EHR Related | It checks if it is EHR related depending on the keywords within the path of the target file. |

For this project, I used a dataset containing 7,365 entries, where 3,067 are benign and 4,298 are ransomware. This dataset was collected by running the ransomware repeatedly to collect the ransomware activity and doing normal operations to collect the benign ones.

## Results and Discussion

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.96 | 0.98 | 968 |
| 1 | 0.97 | 1.00 | 0.98 | 1242 |
| accuracy | | | 0.98 | 2210 |
| macro avg | 0.98 | 0.98 | 0.98 | 2210 |
| weighted avg | 0.98 | 0.98 | 0.98 | 2210 |

Random Forest (RF)

Classification Report:

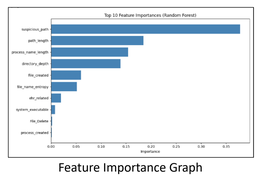
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 0.98 | 0.99 | 969 |
| 1 | 0.98 | 1.00 | 0.99 | 1241 |
| accuracy | | | 0.99 | 2210 |
| macro avg | 0.99 | 0.99 | 0.99 | 2210 |
| weighted avg | 0.99 | 0.99 | 0.99 | 2210 |

K-Nearest Neighbor (KNN)

Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.9858 | 0.9721 | 0.9789 | 645 |
| 1 | 0.9785 | 0.9891 | 0.9838 | 828 |
| accuracy | | | 0.9817 | 1473 |
| macro avg | 0.9822 | 0.9806 | 0.9814 | 1473 |
| weighted avg | 0.9817 | 0.9817 | 0.9817 | 1473 |

Support Vector Classifier (SVC)




Feature Importance Graph

Entropy:

$$-\sum p(x)\ log_2(p(x)) \quad (1)$$

Precision:

$$\frac{TP}{TP+FP} \quad (2)$$

Recall:

$$\frac{TP}{TP+FN} \quad (3)$$

Accuracy:

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

F1:

$$\frac{2*Precision*Recall}{Precision+Recall} \quad (5)$$

## Conclusions

After running the dataset in each of the three algorithms (random forest, k-nearest neighbor, and support vector classifier), we can see that the best one to perform overall was k-nearest neighbor with an 0.99 accuracy. Also, we can see that the most favorable feature was suspicious path, which happens when files are created in sensitive or suspicious directories such as \\appdata\\ or \\temp\\. When analyzing ransomware with ML, we must target these heuristic characteristics as part of the conclusion to check if it's ransomware because that's how the model does the predictions and learns from those patterns.

## Future Work

There is a lot more we can do with ML, Ransomware, and EHRs. For future work, I would like to expand on this topic by making the model better by adding more data and creating a more robust feature extractor. Also, I would like to create software or applications where you can connect to your EHR systems and automatically detect ransomware attack patterns using the ML models created. Ransomware is growing, and so is ML. Using ML to automate things and detect early signs of ransomware in different systems will strengthen the security of crucial industries, such as the healthcare industry. Finally add a signature-based approach where it help detect known attacks in the future.

## Acknowledgements

First, I want to thank Dr. Duffany for his guidance, help and recommendations that helped me and motivated me to finish this project. When it comes to funding for this research, I used personal savings and resources.

## References

[1] J. X. Jiang, J. S. Ross, and G. Bai. (2025, May 14). Ransomware attacks and data breaches in US Health Care Systems [Online]. Available:https://pmc.ncbi.nlm.nih.gov/articles/PMC12079295/#:~:text=Results,decreased%20(Figure%2C%20A). [Accessed: Oct. 3, 2025].
[2] NHI Pragmatic Trials Collaboratory. (n. d.). Rethinking Clinical Trials [Online]. Available: https://rethinkingclinicaltrials.org/chapters/conduct/acquiring-real-world-data/ data-formats/. [Accessed: Oct. 3, 2025].
[3] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan, "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record," in Journal of the American Medical Informatics Association, vol. 25, no. 3, pp. 230–238, March 2018. Available: https://doi.org/10. 1093/jamia/ocx079.
[4] ISO. (n. d.). Electronic Health Records Explained [Online]. Available: https://www.iso.org/healthcare/electronic-health - records. [Accessed: Oct. 3, 2025].
[5] J. Reed. (n. d.). When ransomware kills: Attacks on Healthcare Facilities[Online].Available: https://www.ibm.com/think/insights/when-ransomware kills-attacks-on-healthcare-facilities. [Accessed: October 1, 2025].
[6] T. Pham, "Ethical and legal considerations in Healthcare AI: Innovation and policy for safe and Fair use," in Royal Society Open Science, vol. 12, no. 5, May 2025. Doi:10.1098/rsos.241873.
[7] M. Hirano, R. Hodota, and R. Kobayashi, "RANSAP: An open dataset of ransomware storage access patterns for training machine learning models," in Forensic Science International: Digital Investigation, vol. 40, pp. 301314, Mar. 2022. Doi: 10.1016/j.fsidi.2021.301314.
[8] R. Islam (Creator), "CSU-Ransomware-Data," in GitHub, Nov. 2024.