HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Phan Văn Phùng

XÂY DỰNG HỆ THỐNG PHÁT HIỆN LỖI Ổ CỨNG TRONG MẠNG MÁY TÍNH QUÂN SỰ CỦA BINH CHỦNG THÔNG TIN LIÊN LẠC

ĐỀ ÁN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI - 2025

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Phan Văn Phùng

XÂY DỰNG HỆ THỐNG PHÁT HIỆN LỖI Ổ CỨNG TRONG MẠNG MÁY TÍNH QUÂN SỰ CỦA BINH CHỦNG THÔNG TIN LIÊN LẠC

CHUYÊN NGÀNH : HỆ THỐNG THÔNG TIN

MÃ SỐ: 8.48.01.04

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC

PGS. TSKH. HOÀNG ĐĂNG HẢI

HÀ NỘI – 2025

LỜI CAM ĐOAN

Học viên xin khẳng định rằng đây là công trình nghiên cứu và tìm hiểu hoàn toàn của riêng học viên.

Các số liệu và kết quả trong đề án này hoàn toàn chính xác và chưa từng xuất hiện trong bất kỳ nghiên cứu nào trước đây.

Không có sản phẩm/nghiên cứu nào của người khác được sử dụng trong đề án này mà không được trích dẫn theo đúng quy định.

Hà Nội, ngày 29 tháng 7 năm 2025

Tác giả đề án

Phan Văn Phùng

LỜI CẢM ƠN

Trong suốt quá trình học tập và nghiên cứu thực hiện đề án tốt nghiệp thạc sĩ, ngoài nỗ lực của bản thân, tôi đã nhận được sự hướng dẫn nhiệt tình quý báu của quý Thầy Cô, cùng với sự động viên và ủng hộ của gia đình, bạn bè và đồng nghiệp. Với lòng kính trọng và biết ơn sâu sắc, tôi xin gửi lời cảm ơn chân thành tới:

Thầy PGS.TSKH Hoàng Đăng Hải, người đã tận tình hướng dẫn và chia sẻ những lời khuyên quý báu trong suốt quá trình nghiên cứu và thực hiện đề tài.

Em cũng xin bày tỏ lòng biết ơn tới tất cả các thầy cô tại Học viện Công nghệ Bưu chính Viễn thông những người đã giảng dạy và hỗ trợ em trong suốt thời gian học tập tại đây. Sự tận tâm của các thầy cô đã giúp em tích lũy được nhiều kiến thức và kỹ năng quý giá để áp dụng vào đề án này.

Mặc dù em đã nỗ lực rất nhiều để hoàn thiện đề án, em vẫn không thể tránh khỏi một số sai sót. Em rất mong nhận được sự thông cảm cùng với những góp ý từ các thầy cô để giúp cho đề án ngày càng hoàn thiện hơn.

Em xin chân thành cảm ơn!

Hà Nội, ngày 29 tháng 7 năm 2025

Tác giả đề án

Phan Văn Phùng

MỤC LỤC

DANH MỤC CÁC CHỮ CÁI VIẾT TẮT

Từ viết tắt	Tiếng Anh	Tiếng Việt
CNN	Convolution Neural Network	Mạng nơ-ron tích chập
CRC	Cyclic Redundancy Check	Phương pháp kiểm tra và phát hiện lỗi
ECC	Error Correction Code	Thuật toán , kỹ thuật được sử dụng để phát hiện và sửa lỗi trong dữ liệu
GBDT	Gradient Boosted Decision Trees	Phương pháp học máy kết hợp nhiều cây quyết định theo lũy tiến
HDD	Hard Disk Drive	Ổ đĩa cứng
LSTM	Long Short-term Memory	Kiến trúc mạng nơ-ron hồi tiếp
RAID	Redundant Array of Independent Disks	Hình thức ghép nhiều ổ đĩa cứng vật lý thành một hệ thống ổ đĩa cứng
RF	Random Forest	Thuật toán rừng ngẫu nhiên
RNN	Recurrent neural network	Mạng nơ-ron hồi quy
SAS	Serial Attached SCSI	Phương thức kết nối và truyền dữ liệu giữa các thiết bị lưu trữ
SAN	Storage Area Network	Mạng lưu trữ cục bộ
SATA	Serial AT Attachment	Giao diện bus máy tính dùng để kết nối máy chủ tới các thiết bị lưu trữ
SCADA	Supervisory Control And Data Acquisition	Hệ thống điều khiển giám sát và thu thập dữ liệu

Từ viết tắt	Tiếng Anh	Tiếng Việt
SMART	Self-Monitoring, Analysis and Reporting Technology	Công nghệ giám sát chỉ số ổ đĩa cứng
SMOTE	Synthetic Minority Over-sampling Technique	Kỹ thuật lấy mẫu quá mức thiểu số tổng hợp
SSD	Solid State Drive	Ổ đĩa bán dẫn
STW	Sliding Time Window	Thuật toán cửa sổ trượt theo thời gian
SVM	Support Vector Machine	Phương pháp học máy sử dụng thuật toán phân loại nhị phân
XGB	Extreme Gradient Boosting	Thuật toán học máy dùng để giải quyết các bài toán hồi quy và phân loại.

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH VẼ

MỞ ĐẦU

1. Lý do chọn đề tài

Binh chủng Thông tin liên lạc thuộc Bộ Quốc phòng đang là một trong những đơn vị đi đầu về công tác chuyển đổi số, ứng dụng công nghệ thông tin trong quản lý, điều hành hệ thống Thông tin liên lạc quân sự và thực hiện nhiệm vụ quân sự, quốc phòng. Mạng máy tính quân sự của Binh chủng là mạng độc lập, không liên thông kết nối với mạng Internet và các hệ thống thông tin dân sự khác, được triển khai rộng khắp với hơn 2600 máy tính kết nối, 01 trung tâm dữ liệu, 10 phòng máy chủ (gần 100 máy chủ vật lý). Do đặc điểm đòi hỏi độ bảo mật thông tin cao, hệ thống mạng hoàn toàn độc lập nên việc sử dụng dịch vụ lưu trữ của bên thứ ba như Google cloud, iCloud, One drive, … là điều không thể thực hiện được. Dữ liệu của người dùng chủ yếu được lưu trữ trên các máy tính cục bộ. Các giải pháp dự phòng ổ cứng máy chủ (RAID), triển khai hệ thống lưu trữ tập trung chuyên dụng (SAN Storage) đã được triển khai nhằm mục đích trên. Tuy nhiên, việc triển khai các giải pháp lưu trữ chuyên dụng còn gặp nhiều khó khăn do thiếu hệ thống kỹ thuật, thiếu chuyên gia xử lý, v.v. Vẫn còn tồn tại khá nhiều nguy cơ hư hỏng các ổ cứng trên máy chủ, gây mất mát dữ liệu, làm gián đoạn thông tin liên lạc, tốn nhiều thời gian và công sức để phát hiện lỗi và khôi phục hệ thống. Việc theo dõi giám sát tình trạng hoạt động, phát hiện và dự đoán lỗi ổ cứng trong hệ thống mạng của Binh chủng đang là một nhu cầu cấp thiết.

Cho tới nay, việc kiểm tra phát hiện các sự cố hỏng hóc trên đĩa cứng vẫn chủ yếu dựa vào các công cụ tiện ích của các hệ điều hành máy chủ hoặc một số phần mềm giám sát hoạt động máy chủ, máy tính. Hệ điều hành máy chủ đưa ra cảnh báo khi phát hiện cố hỏng hóc trên các thiết bị hoặc giá trị sử dụng vượt ngưỡng được thiết lập, song chủ yếu đưa ra thông báo trực tiếp trên màn hình máy cục bộ. Các phần mềm giám sát đang thường được triển khai như PRTG, Nagios Core, Zabbix, … thu thập, xử lý và hiển thị thông tin, trạng thái hoạt động dưới dạng các giao diện trực quan (Dashboard), các báo cáo, thống kê. Tuy nhiên, các phần mềm này chưa có khả năng phát hiện sớm khả năng sự cố của các đĩa cứng đang hoạt động. Việc phát hiện sớm hỏng hóc có ý nghĩa vô cùng quan trọng đối với người quản trị vận hành hệ

thống, đặc biệt đối với các dữ liệu quan trọng lưu trong các máy chủ.

Trí tuệ nhân tạo, các công nghệ học máy và học sâu đang phát triển mạnh mẽ, có ứng dụng vào nhiều lĩnh vực của cuộc sống. Đã có những đề xuất ứng dụng AI (trí tuệ nhân tạo), học máy và học sâu vào phát hiện và dự báo hỏng hóc của các trang thiết bị nói chung và các đĩa cứng nói riêng. Tuy nhiên, vẫn chưa có một hệ thống như vậy được đưa vào ứng dụng trong thực tế. Một trong những vấn đề là tính đặc thù của hệ thống, ví dụ yêu cầu triển khai trong mạng quân sự, điển hình như tại Binh chủng Thông tin liên lạc cũng như trong Bộ Quốc phòng.

Việc xây dựng hệ thống có khả năng đưa ra cảnh báo và dự báo hỏng hóc không chỉ có ý nghĩa trong việc bảo vệ dữ liệu người dùng, giảm thời gian downtime của hệ thống mà còn có ý nghĩa trong việc xây dựng các kế hoạch bảo dưỡng các trang thiết bị trong đơn vị một cách chính xác, hiệu quả và tiết kiệm, khắc phục được một số hạn chế trong công tác bảo dưỡng các trang thiết bị công nghệ thông tin hiện nay. Dự báo được tình trạng hỏng hóc còn hỗ trợ người quản lý, chỉ huy xây dựng kế hoạch, ra quyết định mua sắm các trang thiết bị dự phòng, thay thế, vật tư bảo đảm kỹ thuật một cách khoa học, hiệu quả.

Từ những lý do trên em đã lựa chọn đề tài "***Xây dựng hệ thống phát hiện lỗi ổ cứng trong mạng máy tính quân sự của Binh chủng Thông tin liên lạc***".

2. Tổng quan nghiên cứu

Trong các hệ thống mạng máy tính hiện đại như trung tâm dữ liệu, hệ thống lưu trữ đám mây và các hệ thống phân tán, ổ cứng đóng vai trò then chốt trong việc lưu trữ và truy xuất dữ liệu. Ổ cứng là thành phần lưu trữ dữ liệu quan trọng, và bất kỳ sự cố nào của nó cũng có thể dẫn đến mất dữ liệu, giảm hiệu năng hệ thống, hoặc gián đoạn dịch vụ.

Phát hiện lỗi ổ cứng là một vấn đề quan trọng trong quản trị hệ thống và bảo trì mạng máy tính, đặc biệt trong các hệ thống lưu trữ lớn như trung tâm dữ liệu (Data Center), hệ thống phân tán (Distributed Systems) hoặc các môi trường điện toán đám mây (Cloud Computing). Việc phát hiện sớm các dấu hiệu lỗi của ổ cứng giúp: 1) Giảm thiểu rủi ro mất dữ liệu. Phát hiện sớm cho phép sao lưu và di chuyển dữ liệu

trước khi ổ cứng hỏng hoàn toàn. 2) Tăng thời gian hoạt động của hệ thống, tránh các sự cố đột ngột gây gián đoạn dịch vụ. 3) Tiết kiệm chi phí bảo trì. Bảo trì dựa trên dự đoán giúp tối ưu hóa lịch trình và nguồn lực.

Những khó khăn, thách thức trong phát hiện lỗi ổ cứng bao gồm:

- Tính phức tạp và đa dạng của lỗi ổ cứng: Ổ cứng có thể bị lỗi do nhiều nguyên nhân như lỗi cơ học, hao mòn thiết bị, lỗi phần mềm, hoặc môi trường hoạt động (nhiệt độ, rung động, điện áp, v.v).

- Khó phát hiện lỗi sớm: Nhiều lỗi ổ cứng không thể phát hiện kịp thời bằng các phương pháp truyền thống cho đến khi dữ liệu đã bị mất. Rất khó dự đoán sớm lỗi ổ cứng. Nhiều lỗi chỉ được phát hiện khi đã xảy ra, không đủ thời gian để phản ứng kịp thời.

- Các phương pháp truyền thống như SMART (Self-Monitoring, Analysis, and Reporting Technology) còn nhiều hạn chế trong dự đoán lỗi. SMART là hệ thống tự giám sát tích hợp trong hầu hết các ổ cứng hiện đại, cung cấp các chỉ số như nhiệt độ, số lần đọc/ghi lỗi, thời gian hoạt động, v.v. Tuy nhiên, hiệu quả dự báo lỗi còn hạn chế, với tỷ lệ phát hiện lỗi chỉ đạt khoảng 3–10% ở ngưỡng báo động thấp [1, 2, 10].

- Một thách thức điển hình là đặc điểm không cân bằng của dữ liệu. Trong các tập dữ liệu, số lượng mẫu lỗi thường rất ít so với mẫu không lỗi, gây khó khăn cho việc huấn luyện mô hình chính xác.

Vì những lý do nêu trên, việc nghiên cứu các phương pháp dự báo lỗi ổ cứng một cách tự động, chính xác và dự đoán sớm đang là một hướng nghiên cứu quan trọng trong theo dõi, giám sát và phát hiện sớm lỗi ổ cứng trong mạng máy tính.

Với sự phát triển của học máy (Machine Learning) và trí tuệ nhân tạo (AI), nhiều nghiên cứu gần đây đã tập trung vào việc nghiên cứu áp dụng các thuật toán để dự đoán, phát hiện sớm lỗi ổ cứng dựa trên dữ liệu lịch sử hoạt động (các chỉ số S.M.A.R.T. metrics) hoặc log hệ thống. Có thể kể đến một số hướng nghiên cứu tiêu biểu như [6, 7, 8, 10, 11, 12]:

- Nghiên cứu áp dụng các mô hình học máy với dữ liệu SMART. Dữ liệu sử dụng là các thông số như: số sector lỗi, nhiệt độ, tốc độ quay, thời gian hoạt động…

Các mô hình học máy thường dùng gồm: RF (Random Forest), SVM (Suport Vector Machine), KNN (K-Nearest Neighbors), XGBoost [10]. Việc sử dụng học máy đã tăng được độ chính xác trong dự đoán lỗi.

- Nghiên cứu áp dụng các mô hình học sâu. Các mô hình học sâu như LSTM (Long Short-term Memory), Convolution Neural Networks (CNN), Recurrent Neural Networks (RNN), Autoencoder (AE) khai thác chuỗi thời gian dữ liệu hoạt động của ổ cứng từ các chỉ số SMART, giúp phát hiện các mẫu thay đổi theo thời gian và tự động phát hiện đặc trưng mà không cần trích chọn thủ công.

- Nghiên cứu áp dụng phương pháp phát hiện bất thường dựa trên phân tích nhật ký (log) hệ thống. Mô hình phổ biến được áp dụng là Autoencoder (AE) thường áp dụng khi thiếu nhãn gán cho lỗi hoặc thiếu dữ liệu lỗi. Khi đó, các mẫu dữ liệu bình thường có thể được sử dụng để phát hiện các hành vi khác biệt, các hoạt động không bình thường. Mẫu dữ liệu bình thường và bất thường có thể thu thập qua phân tích nhật ký hệ thống.

Trong thời gian qua, các phương pháp phát hiện lỗi ổ cứng chủ yếu dựa vào các chỉ số SMART. Các phương pháp truyền thống chỉ dựa vào các tham số SMART có hạn chế về độ chính xác và hiệu quả dự đoán lỗi, tỷ lệ báo động giả cao, khó dự đoán kế hoạch sử dụng. Mặt khác, đặc tính chuỗi thời gian của dữ liệu SMART không được tận dụng để phát hiện các bất thường.

Việc phân tích dữ liệu SMART có thể giúp khắc phục các hạn chế nêu trên. Mặt khác, việc áp dụng các mô hình học máy cho kết quả dự đoán và độ chính xác tốt hơn so với các phương pháp thống kê truyền thống [13, 14].

Qua khảo sát sơ bộ, trong nước vẫn chưa có nghiên cứu nào liên quan đến chủ đề của đề án này được công bố. Trên thế giới đã có các công trình nghiên cứu nội dung này, điển hình như [3, 4, 6 – 8, 10 – 14]. Tuy nhiên, các nghiên cứu cũng đã chỉ ra vẫn chưa có ứng dụng cụ thể nào được triển khai vào thực tế, đặc biệt đáp ứng nhu cầu của môi trường mạng đóng có tính đặc thù như an ninh quốc phòng.

Trong đề án tốt nghiệp này, một hệ thống phát hiện lỗi đĩa cứng trong môi trường mạng đóng sẽ được nghiên cứu, triển khai xây dựng. Hệ thống có nhiệm vụ

thu thập dữ liệu S.M.A.R.T (Self-Monitoring, Analysis and Reporting Technology) chứa các giá trị về trạng thái của các ổ cứng như nhiệt độ tối thiểu/tối đa/trung bình, tổng số lần đọc/ghi, số giờ hoạt động, tần suất mất điện đột xuất đã được ghi lại, … của các ổ cứng chứa dữ liệu trên máy chủ, máy tính. Dữ liệu thu thập theo tần suất khác nhau, sau đó được đóng gói (sử dụng giao thức quản trị mạng SNMP, http, socket, …) gửi về hệ xử lý trung tâm để phân tích, xử lý. Hệ xử lý trung tâm là bộ não hệ thống tiếp nhận thông tin gửi về, tiến hành chuẩn hóa dữ liệu, sử dụng các kỹ thuật học máy để phân tích dữ liệu, đưa ra thông tin cảnh báo về khả năng hỏng hóc của ổ cứng dưới dạng các dashboard, báo cáo, biểu đồ. Hệ thống có khả năng dự báo hỏng hóc ổ cứng theo độ chính xác (phần trăm) theo tần suất hoặc theo tuần, tháng, quý, năm.

Các vấn đề nghiên cứu cụ thể:

- Nghiên cứu một số giải pháp hiện được sử dụng để đưa ra cảnh báo về tình trạng của ổ cứng. Điển hình là các giải pháp được tích hợp trong các hệ điều hành mạng, phần mềm giám sát hiện nay.

- Nghiên cứu một số công trình nghiên cứu về dự đoán hỏng hóc ổ cứng dựa trên học máy, học sâu đã được công bố trên thế giới.

- Nghiên cứu các kỹ thuật để thu thập các giá trị đặc trưng, thuộc tính ổ cứng dựa trên công nghệ S.M.A.R.T.

- Nghiên cứu các kỹ thuật, thuật toán học máy thường được sử dụng đối với bài toán dự báo.

- Thu thập các tập dữ liệu để huấn luyện mô hình. Phân tích đánh giá các đặc điểm của dữ liệu thu thập được, lựa chọn mô hình phù hợp.

Đề án dự kiến đi sâu vào xây dựng một mô hình hệ thống phù hợp với đặc điểm triển khai hệ thống mạng công nghệ thông tin tại đơn vị công tác của học viên (một mạng máy tính quân sự). Hệ thống dựa trên học máy sẽ có khả năng đưa ra các dự báo theo các khoảng thời gian khác nhau (theo tần xuất thu thập dữ liệu, dự báo theo tuần, tháng, quý, năm, …). Đề tài dự kiến kết hợp kết quả phân tích đặc trưng của dữ liệu thu thập, kết quả thực nghiệm trên tập dữ liệu huấn luyện để lựa chọn mô

hình dự báo có độ chính xác cao, triển khai vào hệ thống tại Binh chủng Thông tin liên lạc, Bộ Quốc phòng.

3. Mục đích nghiên cứu

Mục đích của đề tài là nghiên cứu ứng dụng công nghệ học máy vào xây dựng hệ thống phát hiện lỗi ổ cứng trong mạng máy tính quân sự của Binh chủng Thông tin liên lạc.

Các mục tiêu cụ thể của đề tài gồm:

- Về mặt lý luận: Nghiên cứu các cơ chế bảo vệ dữ liệu trên ổ cứng hiện nay, các đặc trưng dữ liệu SMART ổ cứng, cơ chế phát hiện lỗi ổ cứng, các phương pháp học máy trong dự báo hỏng hóc ổ cứng để có thể triển khai áp dụng vào xây dựng một hệ thống phát hiện lỗi ổ cứng.

- Về mặt thực tiễn: Hệ thống được xây dựng hoạt động ổn định trên môi trường mạng đóng (mạng máy tính quân sự), có khả năng phát hiện và đưa ra dự báo hỏng hóc của ổ cứng máy tính, máy chủ co độ chính xác cao, có khả năng ứng dụng tại đơn vị công tác.

4. Đối tượng và phạm vi nghiên cứu

- Đối tượng nghiên cứu: Hệ thống phát hiện lỗi ổ cứng; Các đặc trưng dữ liệu của ổ cứng theo công nghệ SMART; Các kỹ thuật học máy để giải quyết bài toán dự báo hỏng hóc ổ cứng.

- Phạm vi nghiên cứu: Mô hình triển khai hệ thống tại Ban Công nghệ thông tin/Binh chủng Thông tin liên lạc, Bộ Quốc phòng. Bộ dữ liệu thu thập với các thuộc tính đặc trưng theo công nghệ SMART. Các kỹ thuật học máy cho bài toán phân lớp, dự báo trong phát hiện lỗi ổ cứng dựa trên tập dữ liệu thu thập được.

5. Phương pháp nghiên cứu

- Khảo sát các phương pháp, kỹ thuật qua các công trình nghiên cứu đã được công bố trên thế giới.

- Phương pháp thu thập dữ liệu theo công nghệ S.M.A.R.T.

- Phương pháp phân tích, đánh giá dữ liệu thu thập được.

- Phương pháp sử dụng học máy vào bài toán phát hiện, dự báo.

- Phương pháp mô hình hóa hệ thống

- Thử nghiệm đánh giá kết quả trên tập dữ liệu thật thu thập được tại đơn vị công tác.

6. Bố cục của đề án

Nội dung đề án được chia thành 3 chương như sau:

- CHƯƠNG 1. Tổng quan về hệ thống phát hiện lỗi ổ cứng: Trình bày các cơ sở lý thuyết và thực tiễn liên quan đến bài toán giám sát, phát hiện và dự báo lỗi đĩa cứng trong các mạng chuyên dụng như mạng quân sự của Binh chủng Thông tin liên lạc, Bộ Quốc phòng.

- CHƯƠNG 2. Nghiên cứu xây dựng mô hình học máy cho phát hiện lỗi ổ cứng: Nghiên cứu xây dựng mô hình kiến trúc hệ thống và các thành phần hệ thống; nghiên cứu phương pháp phân tích, đánh giá đặc điểm dữ liệu SMART của ổ cứng; lựa chọn mô hình học máy phù hợp.

- CHƯƠNG 3. Triển khai thử nghiệm và đánh giá kết quả: Thu thập bộ dữ liệu lỗi ổ cứng tại đơn vị và thực hiện gán nhãn để huấn luyện. Triển khai cài đặt mô hình học máy dự báo hỏng hóc ổ cứng để thực nghiệm. Triển khai huấn luyện mô hình, đánh giá độ chính xác mô hình với tập dữ liệu huấn luyện đã thu thập.

Chương 1: TỔNG QUAN VỀ HỆ THỐNG PHÁT HIỆN LỖI Ổ CỨNG

Chương này trình bày các cơ sở lý thuyết và thực tiễn liên quan đến bài toán giám sát, phát hiện và dự báo lỗi đĩa cứng trong các mạng chuyên dụng như mạng quân sự của Binh chủng Thông tin liên lạc, Bộ Quốc phòng.

1.1. Bài toán phát hiện và dự báo lỗi đĩa cứng trong các mạng chuyên dụng

Trong các hệ thống mạng chuyên dụng (mạng quân sự, mạng điều khiển công nghiệp SCADA, ...), việc đảm bảo độ tin cậy , tính bảo mật cao và tính sẵn sàng của hệ thống lưu trữ là tối quan trọng. Ổ đĩa cứng (HDD hoặc SSD) là một thành phần chính trong hầu hết các hệ thống lưu trữ dữ liệu hiện đại. Tuy nhiên, ổ đĩa cứng cũng là một trong những thiết bị dễ hỏng nhất theo thời gian và khi phát sinh hỏng hóc sẽ gây thiệt hại lớn nhất, khó khôi phục hệ thống nhất và làm gián đoạn các hoạt động hệ thống lớn nhất so với hỏng hóc các thành phần khác. Vì vậy, việc phát hiện sớm và dự báo lỗi ổ đĩa cứng đóng vai trò quan trọng nhằm giảm thiểu thời gian chết của hệ thống, bảo vệ dữ liệu và hỗ trợ bảo trì chủ động.

Một trong những vấn đề ảnh hưởng lớn đến việc phát hiện và dự đoán hỏng hóc ổ đĩa cứng trong các mạng chuyên dụng là đặc điểm độc lập của mạng chuyên dụng. Đây thường là hệ thống mạng độc lập, không kết nối với các mạng khác (bao gồm mạng internet). Hệ thống mạng chuyên dụng thường được thiết kế để thực hiện một công việc cụ thể, ít có sự thay đổi bổ sung trong suốt quá trình vận hành. Chính vì vậy, việc cập nhật dữ liệu, áp dụng các công nghệ mới để tăng cường tính bảo mật, hoạt động ổn định cho các thiết bị trong mạng (bao gồm ổ đĩa cứng) là rất hạn chế. Do vậy, việc triển khai một hệ thống nhằm phát hiện, cảnh báo sớm hỏng hóc các thiết bị trong mạng chuyên dụng sẽ gặp rất nhiều khó khan, thách thức so với các hệ thống mạng phổ biến khác.

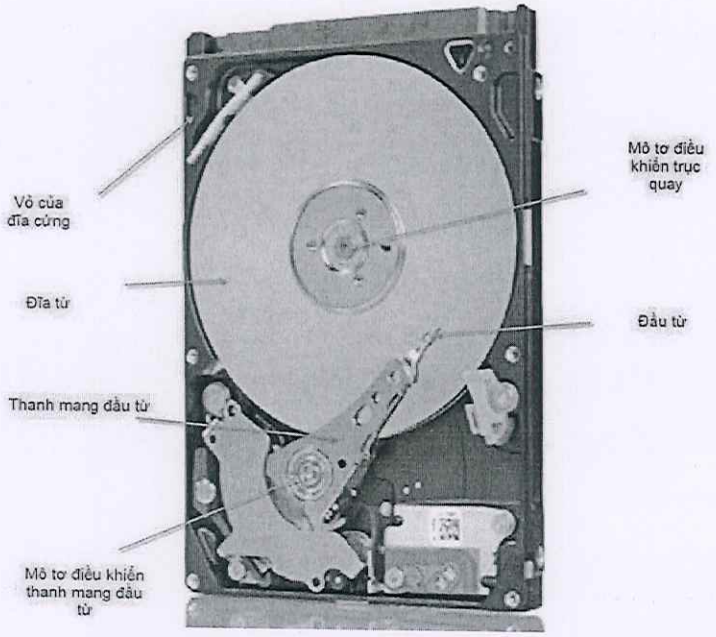
1.2. Đặc điểm cấu trúc ổ cứng, cơ chế bảo vệ dữ liệu ổ cứng

1.2.1. Đặc điểm cấu trúc các ổ cứng phổ biến hiện nay

Trên thị trường hiện nay có nhiều loại ổ đĩa cứng của nhiều hãng khác nhau, một số hãng lớn như Seagate Technology (Mỹ), Western Digital – WD (Mỹ), Toshiba

(Nhật Bản), Samsung (Hàn Quốc), Intel , Itachi, …Đi cùng với đó là nhiều công nghệ sản xuất ổ đĩa cứng khác nhau. Một số công nghệ ổ đĩa cứng phổ biến hiện nay gồm:

Ổ cứng HDD (Hard Disk Drive): Thành phần chính gồm một hay nhiều đĩa từ (platter) được phủ vật liệu từ tính, gắn trên trục quay; đầu đọc/ghi gắn trên cần di chuyển (actuator arm). Khi ghi, từ tính của bề mặt đĩa được thay đổi để lưu thông tin. Khi đọc, đầu từ cảm nhận thay đổi từ tính. Thường được sử dụng phổ biến trong lưu trữ dung lượng lớn, như NAS, máy chủ backup, giám sát camera, …

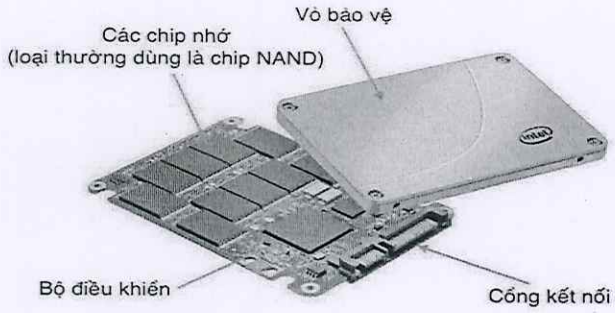


Hình 1.1. Cấu trúc ổ đĩa cứng HDD

(Nguồn https://maytinhdongbo.com/)

Ổ cứng SSD (Solid State Drive): Về cấu trúc không có bộ phận chuyển động; dữ liệu được lưu trữ trên chip flash NAND. Gồm các thành phần: controller (bộ điều khiển), DRAM cache, và nhiều cell NAND chia theo SLC, MLC, TLC, QLC. Được sử dụng phổ biến trong máy chủ ứng dụng, database tốc độ cao, các hệ thống yêu cầu truy cập IOPS cao, máy tính cá nhân và máy trạm chuyên dụng.

Ổ cứng NVMe SSD: Kết nối qua giao tiếp PCIe, sử dụng giao thức NVMe (Non-Volatile Memory express) thay vì SATA.Tốc độ đọc ghi cực nhanh (có thể đạt 3500–7000 MB/s), độ trễ thấp. Ứng dụng cho các máy chủ hiệu năng cao, video editing chuyên nghiệp, gaming.



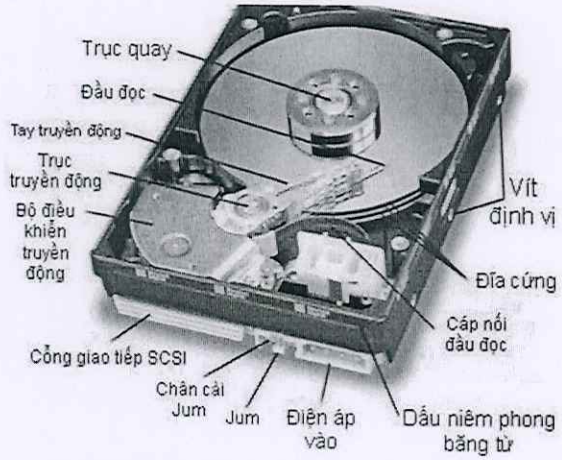
Hình 1.2. Cấu trúc ổ đĩa cứng SSD

(Nguồn https://lagihitech.vn/)

Hình 1.3. Ổ cứng NVMe

(Nguồn https://www.samsung.com/)

Ổ cứng Enterprise: Các ổ cứng dành riêng cho máy chủ (enterprise-grade) được tối ưu với Độ bền cao (MTBF > 2 triệu giờ), Hoạt động 24/7, Có firmware chuyên biệt để tương thích RAID, Giao tiếp SAS, SATA hoặc NVMe chuyên dụng



Hình 1.4. Cấu trúc ổ đĩa cứng Enterprise

(Nguồn https://en-m-wikipedia-org)

1.2.2. Các cơ chế bảo vệ dữ liệu trên ổ cứng

Các ổ đĩa cứng được tích hợp hỗ trợ nhiều giải pháp kỹ thuật để tăng độ bền và bảo vệ dữ liệu. Các cơ chế để bảo vệ dữ liệu phổ biến hiện nay gồm:

- RAID (Redundant Array of Independent Disks): Là kỹ thuật kết hợp nhiều ổ cứng vật lý thành một đơn vị logic để tăng hiệu suất, độ tin cậy hoặc cả hai. Các loại RAID phổ biến được nêu ở Bảng 1.1 [9]

Bảng 1.1. Các loại RAID phổ biến hiện nay

Loại	Mô tả	Ưu điểm	Nhược điểm
RAID 0	Ghép song song không dự phòng	Tăng tốc	Không có bảo vệ dữ liệu
RAID 1	Gương (mirror) 1-1	Dữ liệu an toàn	Tốn dung lượng
RAID 5	Phân mảnh có parity	Cân bằng giữa hiệu suất và bảo vệ	Phức tạp khi rebuild
RAID 6	Như RAID 5 + thêm parity	An toàn cao hơn	Tốc độ ghi giảm
RAID 10	RAID 1 + RAID 0	Hiệu suất + an toàn	Chi phí cao

- Snapshot và Clone: Lưu ảnh chụp nhanh trạng thái dữ liệu tại thời điểm nhất định (snapshot) hoặc tạo bản sao hoàn chỉnh ổ đĩa (clone), được sử dụng trong môi trường ảo hóa (VMware, Hyper-V) hoặc hệ điều hành hỗ trợ snapshot (Windows Volume Shadow Copy).

- ECC và CRC: Các ổ đĩa cứng hiện đại thường được tích hợp sẵn cơ chế tự sửa lỗi ở mức bít ECC (Error Correction Code) và cơ chế kiểm tra dữ liệu ghi/đọc CRC (Cyclic Redundancy Check)

1.3. Khái quát về các kỹ thuật, công nghệ sử dụng để theo dõi giám sát hoạt động của ổ cứng

1.3.1. Công nghệ SMART

SMART (viết tắt của *Self-Monitoring, Analysis and Reporting Technology*) là một công nghệ tích hợp bên trong ổ cứng HDD, SSD hoặc ổ đĩa lai SSHD nhằm tự động giám sát trạng thái hoạt động và dự đoán các lỗi tiềm ẩn có thể xảy ra trong

tương lai. Đây là chuẩn công nghiệp được hỗ trợ bởi hầu hết các nhà sản xuất ổ đĩa hiện nay như Seagate, Western Digital, Toshiba, Samsung, v.v.

SMART hoạt động bằng cách liên tục ghi nhận và cập nhật các chỉ số kỹ thuật trong quá trình ổ đĩa vận hành. Dựa vào các thông số này, người dùng và hệ thống có thể đánh giá tình trạng "sức khỏe" của ổ đĩa, từ đó chủ động sao lưu dữ liệu hoặc thay thế ổ đĩa trước khi xảy ra sự cố nghiêm trọng.

Mỗi ổ cứng có một bảng SMART gồm khoảng 20 - 50 chỉ số (gọi là *attributes*) thể hiện các khía cạnh vận hành khác nhau. Một số chỉ số được nêu Bảng 1.2 [2]

Bảng 1.2. Các thuộc tính chỉ số SMART

ID	Attributes	Chỉ số	Mô tả
1	Raw Read Error Rate	Tỷ lệ lỗi đọc thô	Đã xảy ra lỗi khi đọc dữ liệu thô từ đĩa
			Chỉ ra sự cố với bề mặt đĩa hoặc đầu đọc/ghi.
			Thuộc tính quan trọng
2	Throughput Performance	Hiệu suất thông lượng	Hiệu suất thông lượng chung của ổ cứng
			Chỉ ra sự cố với động cơ, servo hoặc vòng bi.
3	Spin Up Time	Thời gian quay	Thời gian cần thiết để trục chính quay đến vòng tua tối đa
			Chỉ ra sự cố với động cơ hoặc ổ trục.
			Thuộc tính quan trọng
4	Start/Stop Count	Đếm Bắt đầu/Dừng	Số chu kỳ khởi động/dừng của trục chính
			Giá trị này không ảnh hưởng trực tiếp đến tình trạng của ổ đĩa.
5	Reallocated Sector Count (Reallocated Sectors Count)	Số lượng khu vực được phân bổ lại (Số lượng khu vực được phân bổ lại)	Số lượng sector được di chuyển đến vùng dự phòng
			Chỉ ra sự cố với bề mặt đĩa hoặc đầu đọc/ghi.
			Thuộc tính quan trọng
6	Read Channel Margin	Đọc Biên độ Kênh	Khoảng cách của kênh khi đọc dữ liệu
			Chức năng chính xác của thuộc tính này không được chỉ định.
7	Seek Error Rate	Tìm kiếm tỷ lệ lỗi	Tỷ lệ lỗi định vị của đầu đọc/ghi
			Chỉ ra vấn đề với servo, đầu. Nhiệt độ cao cũng có thể gây ra vấn đề này.
			Thuộc tính quan trọng

ID	Attributes	Chỉ số	Mô tả
8	Seek Time Performance	Tìm kiếm hiệu suất thời gian	Thời gian trung bình của các hoạt động tìm kiếm của đầu
			Chỉ ra sự cố với servo.
			Thuộc tính quan trọng
9	Power-On Time Count	Đếm thời gian bật nguồn	Tổng thời gian ổ đĩa được bật
			Đơn vị đo phụ thuộc vào nhà sản xuất.
10	Spin Retry Count	Số lần thử lại vòng quay	Thử lại số lần khởi động quay
			Chỉ ra sự cố với động cơ, ổ trục hoặc nguồn điện.
			Thuộc tính quan trọng
11	Drive Calibration Retry Count	Số lần thử lại hiệu chuẩn ổ đĩa	Số lần thử hiệu chỉnh ổ đĩa
			Chỉ ra sự cố với động cơ, vòng bi hoặc nguồn điện.
12	Drive Power Cycle Count	Số chu kỳ công suất ổ đĩa	Số chu kỳ bật/tắt nguồn hoàn chỉnh
			Giá trị này không ảnh hưởng trực tiếp đến tình trạng của ổ đĩa.
13	Soft Read Error Rate	Tỷ lệ lỗi đọc mềm	Số lỗi đọc phần mềm
			Số lỗi đọc không thể sửa được.
190	Airflow Temperature	Nhiệt độ luồng không khí	Nhiệt độ luồng khí
			Nhiệt độ của luồng khí bên trong vỏ ổ cứng.
191	Mechanical Shock	Sốc cơ học	Số lượng các vấn đề gây ra bởi sốc cơ học
			Tăng tốc (ví dụ như ngã) có thể gây ra sốc cơ học.
192	Power off Retract Cycle	Tắt nguồn Chu kỳ thu hồi	Số chu kỳ tắt nguồn
			Giá trị này không ảnh hưởng trực tiếp đến tình trạng của ổ đĩa.
			Số chu kỳ đầu di chuyển vào vị trí vùng hạ cánh.
194	HDD Temperature	Nhiệt độ ổ cứng	Nhiệt độ đĩa
			Nhiệt độ bên trong vỏ ổ cứng.
			Số lỗi được sửa bằng cơ chế sửa lỗi nội bộ.
196	Reallocation Event Count	Sự kiện phân bổ lại số lượng	Số lượng các hoạt động ánh xạ lại sector
			Số lượng tất cả các hoạt động ánh xạ lại (thành công và thất bại).
			Thuộc tính quan trọng
197	Current Pending	Số lượng ngành đang	Số lượng sector không ổn định
			Các sector đang chờ xử lý này có thể được ánh xạ lại vào vùng dự phòng.

ID	Attributes	Chỉ số	Mô tả
	Sector Count	chờ xử lý hiện tại	*Thuộc tính quan trọng*
198	Off-Line Uncorrectable Sector Count	Số lượng Sector không thể sửa lỗi ngoại tuyến	Số lượng lỗi không thể sửa được khi đọc/ghi
			Chỉ ra sự cố với bề mặt đĩa hoặc đầu đọc/ghi.
			Thuộc tính quan trọng
199	Ultra ATA CRC Error Count	Đếm lỗi CRC Ultra ATA	Số lượng lỗi trong quá trình truyền dữ liệu giữa đĩa và máy chủ
			Chỉ ra sự cố với nguồn điện hoặc cáp dữ liệu.
200	Write Error Rate	Tỷ lệ lỗi ghi	Xảy ra lỗi khi ghi dữ liệu thô từ đĩa
			Chỉ ra sự cố ở bề mặt đĩa hoặc đầu đọc/ghi.
201	Soft Read Error Rate	Tỷ lệ lỗi đọc mềm	Số lỗi đọc phần mềm
			Số lỗi đọc không thể sửa được.
202	Data Address Mark Errors	Lỗi đánh dấu địa chỉ dữ liệu	Số lỗi về dấu địa chỉ dữ liệu
			Số lỗi về dấu địa chỉ không đúng hoặc không hợp lệ.
			Đã phát hiện tổng kiểm tra sửa lỗi không hợp lệ trong quá trình sửa lỗi.
204	Soft ECC Correction	Sửa lỗi ECC mềm	Số lỗi dữ liệu đã được sửa
			Lỗi được sửa bằng cơ chế sửa lỗi nội bộ.
205	Thermal Asperity Rate	Tỷ lệ độ nhám nhiệt	Số lượng vấn đề về nhiệt
			Tổng số vấn đề gây ra do nhiệt độ cao.
206	Flying Height	Chiều cao bay	Chiều cao đầu bay
			Chiều cao của đầu đĩa so với bề mặt đĩa.
207	Spin High Current	Dòng điện xoay chiều cao	Giá trị dòng điện trong quá trình quay
			Dòng điện cần thiết để quay ổ đĩa.
208	Spin Buzz	Quay Buzz	Số chu kỳ cần thiết để quay
			Số lần thử lại trong quá trình quay do dòng điện khả dụng thấp.
209	Offline Seek Performance	Hiệu suất tìm kiếm ngoại tuyến	Kiểm tra hiệu suất ổ đĩa trong quá trình hoạt động ngoại tuyến
			Kiểm tra hiệu suất ổ đĩa trong quá trình tự kiểm tra nội bộ.
220	Disk Shift	Chuyển đĩa	Khoảng cách của đĩa đã dịch chuyển so với trục chính.
			Đĩa quay không chính xác có thể do va chạm cơ học hoặc nhiệt độ cao.
221			Số lỗi cơ học

ID	Attributes	Chỉ số	Mô tả
	G-Sense Error Rate	Tỷ lệ lỗi G-Sense	Số lỗi do va chạm hoặc rung động.
222	Loaded Hours	Giờ đã tải	Số giờ bật nguồn
			Giá trị này liên tục tăng (mỗi giờ một lần).
223	Load/Unload Retry Count	Tải/Gỡ bỏ số lần thử lại	Số lượng hoạt động tải/dỡ
			Số lượng đầu truyền động đi vào/rời khỏi vùng dữ liệu.
224	Load Friction	Tải Ma Sát	Tỷ lệ ma sát cơ học
			Tỷ lệ ma sát giữa các bộ phận cơ khí. Chỉ ra vấn đề với hệ thống cơ khí của ổ đĩa.
226	Load-in Time	Thời gian tải vào	Tổng thời gian đầu đọc được tải
			Thời gian đầu đọc/ghi ở trong vùng dữ liệu.
227	Torque Amplification Count	Đếm khuếch đại mô men xoắn	Tốc độ tăng mô-men xoắn
			Mô-men xoắn tăng trong quá trình quay của ổ cứng.
228	Power-off Retract Count	Số lần thu hồi khi tắt nguồn	Số chu kỳ tắt nguồn
			Số lần đầu được thu vào do mất điện.
230	GMR Head Amplitude	Biên độ đầu GMR	Biên độ định vị đầu
			Khoảng cách di chuyển đầu giữa các thao tác.
231	Hard Disk Temperature	Nhiệt độ ổ cứng	Nhiệt độ đĩa
			Nhiệt độ bên trong vỏ ổ cứng.
240	Head Flying Hours	Giờ bay đầu	Số giờ định vị đầu
			Thời gian dành cho việc định vị đầu truyền động.
250	Read Error Retry Rate	Đọc lỗi Tỷ lệ thử lại	Số lần thử lại trong quá trình đọc
			Số lỗi được tìm thấy trong quá trình đọc một sector từ bề mặt đĩa.

(Nguồn: https://www.hdsentinel.com/smart/smartattr.php)

Mỗi chỉ số SMART có 2 giá trị chính:

- Normalized: giá trị đã chuẩn hóa (thường từ 1–100 hoặc 1–200)
- Raw: giá trị thô đo được (ví dụ: số sector lỗi thực tế)

1.3.2. Công cụ giám sát chuyên dụng

Hệ điều hành Windows: Một số ứng dụng được tích hợp, cài đặt trên hệ điều hành Windows để theo dõi tình trạng, thông số ổ cứng như ứng dụng CrystalDiskInfo,

Hard Disk Sentinel dùng để theo dõi trạng thái SMART, nhiệt độ, tuổi thọ SSD,...

Hệ điều hành Linux: Một số ứng dụng như *smartmontools* (công cụ CLI mạnh mẽ để kiểm tra SMART và tự động gửi cảnh báo), *iostat, hdparm* (Phân tích hiệu năng I/O, thông tin phần cứng ổ đĩa)

1.3.3. Hệ thống giám sát tập trung

Các giải pháp như PRTG, Zabbix, Nagios, Netdata, ... hỗ trợ việc thu thập log từ ổ đĩa thông qua các bộ cảm biến (sensor), phân tích và đưa ra cảnh báo về nhiệt độ, một số bất thường trong đọc/ghi dữ liệu hay cảnh báo về mức độ sử dụng ổ đĩa cứng, ...Các thông tin được hiển thị dưới dạng các biểu đồ trực quan (dashboard), các báo cáo (report) cận thời gian thực.

1.3.4. Công nghệ giám sát mức phần cứng – firmware tích hợp

Các ổ enterprise (Seagate Exos, WD Ultrastar, v.v.) thường được tích hợp sẵn firmware hỗ trợ theo dõi mức độ mài mòn (wear level), block hỏng, ECC error rate, v.v Một số dòng SSD NVMe hỗ trợ NVMe SMART log và telemetry log giúp theo dõi tình trạng ổ đĩa cứng qua các lệnh truy vấn (command line) hoặc qua các giao diện lập trình (API) được các hãng cung cấp.

1.4. Các công trình nghiên cứu liên quan

Qua khảo sát sơ bộ của học viên, cho tới nay vẫn chưa có nghiên cứu nào trong nước liên quan được công bố.

Trong phần sau đây, đề án tốt nghiệp trình bày tóm tắt một số công trình nghiên cứu tiêu biểu trên thế giới, điển hình là các công bố trong các tài liệu [3, 4, 6 – 8, 10 – 16]. Tuy nhiên, các nghiên cứu cũng đã chỉ ra vẫn chưa có ứng dụng cụ thể nào được triển khai vào thực tế, đặc biệt đáp ứng nhu cầu của môi trường mạng đóng có tính đặc thù như an ninh quốc phòng.

1.4.1. Các nghiên cứu theo hướng giám sát, thống kê các chỉ số SMART

Tính xác suất thống kê theo các chỉ số SMART:

Dự đoán lỗi ổ cứng được coi là một phần quan trọng trong kế hoạch bảo trì các hệ thống thông tin. Các nghiên cứu đã chỉ ra lỗi ổ cứng xuất hiện càng nhiều khi

hệ thống lưu trữ được mở rộng, khả năng mất dữ liệu trở lên rất cao [15]. Hầu hết các nghiên cứu đều sử dụng tập dữ liệu theo các chỉ tiêu SMART [2] để dự đoán lỗi ổ cứng [3, 4, 6 – 8]. Tập dữ liệu SMART sử dụng các thuộc tính được truy xuất khi thực hiện các thao tác trên ổ cứng. Trong hầu hết các trường hợp, các thuộc tính SMART có thể thu từ dữ liệu nhật ký truy xuất ổ cứng cho phép phát hiện có lỗi hay không và dự đoán lỗi theo thời gian.

Các tác giả trong [3] đã nghiên cứu xu thế lỗi của một lượng lớn ổ cứng trong trung tâm dữ liệu. Các tác giả đã thực hiện thu thập và phân tích hành vi hỏng hóc của hơn 100.000 ổ đĩa cứng (HDD) trong trung tâm dữ liệu của Google. Phương pháp sử dụng là xác xuất thống kê để phân tích mối tương quan giữa các chỉ số SMART và xác xuất hỏng hóc. Đây là một trong những nghiên cứu đầu tiên với dữ liệu quy mô lớn và theo dõi chi tiết các chỉ số SMART, hoạt động và nhiệt độ của ổ đĩa. Mục tiêu của nghiên cứu này là xác định các yếu tố ảnh hưởng đến độ bền của ổ cứng. Các tác giả đã chỉ ra cách thức Google xây dựng hệ thống thu thập dữ liệu thời gian thực từ các máy chủ, sử dụng Bigtable và MapReduce để phân tích.

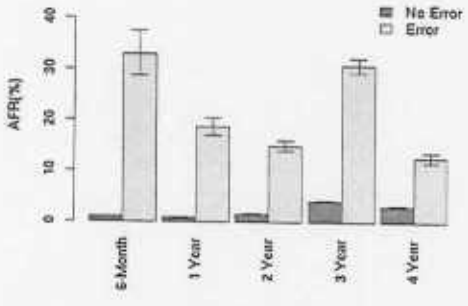


Figure 6: AFR for scan errors.

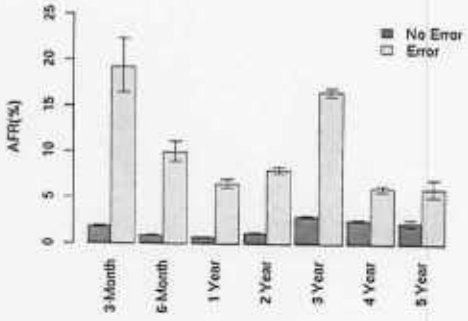
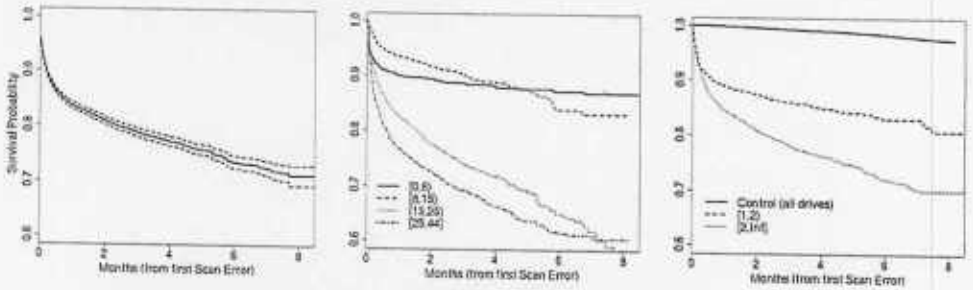


Figure 7: AFR for reallocation counts.



Hình 1.5. Một số kết quả nghiên cứu của Google

Kết quả đạt được của nghiên cứu thể hiện trên Hình 1.5 bao gồm [3, tr.7]:

- Tỷ lệ hỏng hóc trung bình: từ 1.7% năm đầu lên 8.6% vào năm thứ 3.

- Không tìm thấy mối liên hệ mạnh giữa nhiệt độ cao hoặc mức độ sử dụng cao với hỏng hóc.

- Một số thông số SMART có tương quan cao với lỗi, nhưng vẫn không đủ để dự đoán chính xác.

Khác với các nghiên cứu phân loại lỗi ổ cứng theo cách trên, các tác giả trong [13] sử dụng kỹ thuật hồi quy để ước tính trực tiếp thời gian sử dụng còn lại của ổ đĩa cứng. Cách này trái ngược với các phương pháp thông thường để phân loại lỗi ổ đĩa cứng tiến triển theo một khoảng thời gian cụ thể. Tuy các phương pháp Random Forest hay LSTM đã được sử dụng, việc phân tích xác suất các thuộc tính SMART vẫn là trọng tâm nhằm theo dõi sự hư hỏng theo thời gian. Một số phương pháp tương tự sử dụng cách đánh giá dựa trên sai số trung bình tuyệt đối (MAE).

Trong [16], các tác giả cũng đã sử dụng cả mô hình phân loại và hồi quy để dự đoán lỗi ổ cứng bằng các kỹ thuật học máy Random Forest. Cách tiếp cận này cho các kết quả về sai số trung bình tuyệt đối (MAE) và lỗi bình phương căn bậc hai trung bình (RMSE) tốt hơn. Tuy nhiên, kết quả nghiên cứu cho thấy, mức độ tin cậy của dữ liệu đạt được chưa bảo đảm, dữ liệu còn thiếu khâu chọn lọc, độ chính xác chưa cao.

1.4.2. Các nghiên cứu theo hướng sử dụng học máy, học sâu

Nhiều nghiên cứu đã áp dụng các thuật toán học máy để phân tích dữ liệu S.M.A.R.T. nhằm dự đoán lỗi ổ cứng, điển hình là các nghiên cứu trong [4, 6 – 8, 10 – 16].

Nghiên cứu của nhóm tác giả Gargiulo et.al [4] năm 2021 tập trung vào vấn đề dự đoán lỗi ổ cứng thông qua học máy với việc gán nhãn tự động. Nghiên cứu đã sử dụng hơn 65.000 ổ đĩa được thu thập tại trung tâm dữ liệu CERN. Mô hình học máy được sử dụng cho thử nghiệm là Regularized Greedy Forest - RGF (một biến thể của Gradient Boosted Decision Trees). Mô hình có sử dụng thêm tham số tham chiếu (dạng regularization) để tăng độ ổn định, kết hợp chiến lược chọn tham số tối ưu bằng thực nghiệm (Design of Experiment - DOE).

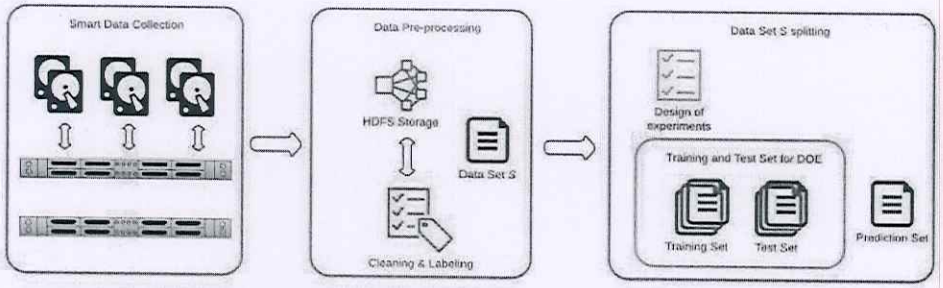
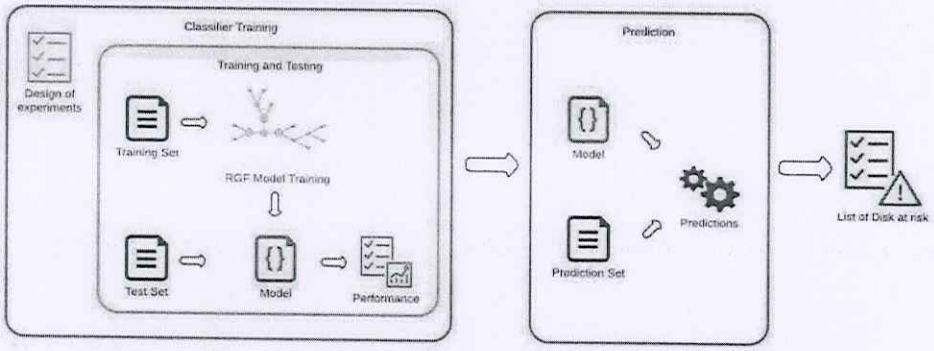


Figure 1. Data Collection and pre-processing phases.



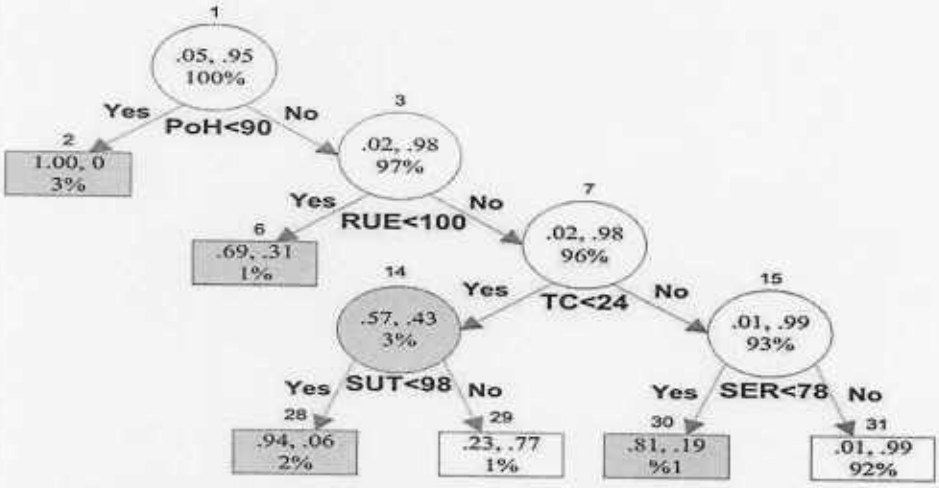
Hình 1.6. Mô hình dự đoán của nghiên cứu của nhóm CERN [4, tr.4]

Kết quả nghiên cứu được nêu ở Bảng 1.4 [4] cho thấy các tỷ lệ Recall, FPR và LR (Recall / FPR) đạt được trong ba loại mô hình sử dụng trong nghiên cứu.

Bảng 1.4. Kết quả nghiên cứu của nhóm CERN

Model	Recall	FPR (False Positive Rate)	LR+ (Recall/FPR)
Ischia	98.4%	0.2%	659
Capri	92.2%	0.8%	115
Ponza	78.1%	0.9%	86

Các tác giả trong [6] đề xuất mô hình dự đoán lỗi ổ cứng sử dụng cây phân loại và hồi quy (Classification & Regression Trees – CART). Dữ liệu thực tế từ trung tâm dữ liệu với 25.792 ổ đĩa, sử dụng 13 đặc trưng SMART gồm: chỉ số lỗi đọc, thời gian bật máy, lỗi không thể sửa, nhiệt độ, tốc độ thay đổi,.v.v. Nhóm nghiên cứu đã áp dụng cây phân loại (CT) để phân biệt ổ tốt / lỗi và cây hồi quy (RT) để đánh giá "mức độ sức khỏe" của ổ cứng.



Hình 1.7. Mô hình phân loại của nghiên cứu CART [6, tr.385]

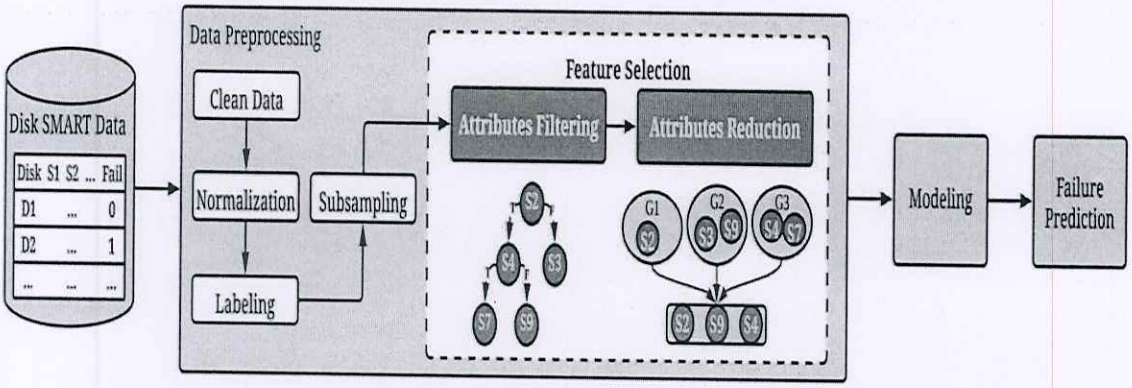
Các ký hiệu biểu thị trên hình gồm: POH (Power on Hours), RUE (Reported Uncorrectable Errors), TC (Temperature Celsius), SUT (Spin Up Time) và SER (Seek Error Rate) đều là các thuộc tính SMART của ổ cứng.

Kết quả nghiên cứu của [6] cho thấy, với mô hình phân loại (CT) có thể dự đoán được >95% lỗi ổ cứng, tỷ lệ cảnh báo sai (FAR) đạt được <0.1%. Mô hình có thể phát hiện lỗi trung bình trước khi xảy ra sự cố 2 tuần. Với mô hình hồi quy (RT), hệ thống cho phép xử lý ưu tiên theo mức độ nguy cơ và điều chỉnh linh hoạt giữa FDR – FAR bằng ngưỡng và đánh giá được mức độ rủi ro của ổ cứng (gần hỏng hay còn ổn) [6, tr.383].

Nghiên cứu của Wang và cộng sự [8] năm 2023 tập trung vào vấn đề tối ưu hiệu quả học máy cho dự đoán lỗi ổ cứng bằng cách sử dụng lựa chọn các đặc trưng dựa vào phân lớp. Mục tiêu của nghiên cứu là cải thiện tốc độ và độ chính xác trong dự đoán lỗi ổ cứng (HDD) sử dụng mô hình học máy. Bài báo tập trung giải quyết vấn đề về thời gian huấn luyện lâu và độ trễ dự đoán cao trong bối cảnh dữ liệu SMART cập nhật liên tục. Các tác giả đề xuất giải pháp chọn đặc trưng 2 lớp (two-layer feature selection) nhằm giảm số lượng đặc trưng đầu vào mà không làm giảm hiệu suất dự đoán, cụ thể như sau:

- Lớp 1: Lọc đặc trưng. Dựa trên entropy và chỉ số Gini (Gini index) của cây phân loại để xác định mức độ quan trọng của từng đặc trưng, kết hợp loại bỏ các thuộc tính không liên quan hoặc làm giảm hiệu quả dự đoán.

- Lớp 2: Rút gọn đặc trưng theo nhóm tương quan. Phân nhóm đặc trưng dựa trên tương quan (Pearson hoặc Spearman). Từ mỗi nhóm, chỉ chọn 1 đặc trưng đại diện để huấn luyện mô hình.



Hình 1.8. Mô hình lọc dữ liệu nghiên cứu của Wang [8, tr.5]

Bảng 1.5. Kết quả nghiên cứu của Wang

		DM000									NM0007								
		Group1			Group2_Prs			Group2_Sprm			Group1			Group2_Prs			Group2_Sprm		
		Pre	Re	F1	Pre	Re	F1	Pre	Re	F1	Pre	Re	F1	Pre	Re	F1	Pre	Re	F1
Original	NB	0.69	0.38	0.49	0.7	0.37	0.49	0.72	0.37	0.48	0.79	0.31	0.45	0.81	0.27	0.41	0.8	0.3	0.43
	RF	0.99	0.98	**0.99**	0.99	**0.99**	**0.99**	1	0.98	**0.99**	1	0.97	0.98	0.99	0.94	0.97	1	0.95	0.97
	SVM	0.91	0.54	0.68	0.89	0.09	0.17	0.94	0.46	0.62	0.84	0.31	0.45	0.83	0.29	0.43	0.77	0.24	0.36
	GBDT	0.97	0.79	0.87	0.95	0.76	0.85	0.97	0.77	0.86	0.93	0.71	0.81	0.92	0.61	0.73	0.91	0.68	0.78
	CNN	0.9	0.62	0.74	0.8	0.54	0.65	0.77	0.54	0.64	0.82	0.35	0.49	0.86	0.36	0.51	0.74	0.33	0.46
	LSTM	0.98	0.9	0.94	0.96	0.95	0.96	0.99	0.91	0.95	0.91	0.92	0.92	0.94	0.92	0.93	0.96	0.81	0.91
Pearson	NB	0.78	0.36	0.49	0.78	0.35	0.48	0.78	0.35	0.48	0.78	0.3	0.43	0.81	0.26	0.4	0.79	0.29	0.42
	RF	0.95	0.46	0.62	0.95	0.46	0.62	0.95	0.47	0.63	0.95	0.72	0.82	0.89	0.57	0.69	0.94	0.72	0.82
	SVM	0.78	0.24	0.36	0.75	0.12	0.21	0.77	0.12	0.21	0.77	0.32	0.46	0.82	0.31	0.45	0.73	0.25	0.37
	GBDT	0.92	0.43	0.59	0.9	0.42	0.57	0.92	0.43	0.59	0.9	0.62	0.74	0.87	0.54	0.67	0.9	0.62	0.74
	CNN	0.72	0.34	0.47	0.77	0.24	0.36	0.8	0.2	0.32	0.72	0.4	0.51	0.77	0.36	0.49	0.68	0.37	0.48
	LSTM	0.94	0.36	0.52	0.87	0.37	0.52	0.87	0.37	0.52	0.95	0.48	0.64	0.88	0.46	0.6	0.88	0.46	0.6
Spearman	NB	0.73	0.35	0.47	0.74	0.34	0.47	0.74	0.34	0.47	0.78	0.3	0.43	0.81	0.26	0.4	0.79	0.29	0.42
	RF	0.99	0.49	0.66	0.98	0.48	0.64	0.96	0.5	0.65	0.96	0.71	0.81	0.91	0.56	0.7	0.95	0.72	0.82
	SVM	0.78	0.23	0.36	0.74	0.12	0.2	0.76	0.11	0.2	0.77	0.32	0.46	0.82	0.31	0.45	0.73	0.25	0.37
	GBDT	0.95	0.42	0.59	0.93	0.41	0.57	0.95	0.42	0.59	0.89	0.62	0.73	0.87	0.54	0.67	0.9	0.62	0.74
	CNN	0.71	0.35	0.47	0.77	0.29	0.42	0.77	0.24	0.36	0.77	0.33	0.46	0.73	0.38	0.5	0.7	0.33	0.45
	LSTM	0.95	0.23	0.37	0.68	0.29	0.41	0.78	0.29	0.42	0.87	0.51	0.65	0.91	0.45	0.61	0.85	0.54	0.66
J-Index	NB	0.62	0.36	0.46	0.63	0.35	0.45	0.71	0.3	0.43	0.77	0.32	0.45	0.77	0.32	0.46	0.75	0.3	0.43
	RF	0.98	0.93	0.96	0.98	0.91	0.95	0.97	0.89	0.93	0.98	0.88	0.93	0.97	0.84	0.9	0.98	0.84	0.91
	SVM	0.81	0.3	0.44	0.77	0.15	0.25	0.79	0.15	0.25	0.88	0.34	0.49	0.88	0.35	0.5	0.8	0.22	0.35
	GBDT	0.88	0.68	0.76	0.87	0.68	0.76	0.87	0.68	0.76	0.89	0.69	0.78	0.87	0.69	0.77	0.9	0.68	0.78
	CNN	0.79	0.47	0.59	0.54	0.65	0.59	0.57	0.62	0.59	0.66	0.54	0.6	0.71	0.44	0.54	0.54	0.5	0.52
	LSTM	0.94	0.83	0.88	0.94	0.81	0.87	0.88	0.75	0.81	0.83	0.73	0.78	0.88	0.73	0.8	0.85	0.67	0.75
Entropy	NB	0.63	0.37	0.47	0.7	0.37	0.49	0.71	0.36	0.47	0.72	0.43	0.54	0.78	0.31	0.45	0.75	0.32	0.44
	RF	1	0.98	**0.99**	0.99	0.98	**0.99**	0.99	**0.99**	**0.99**	1	0.97	0.98	1	0.97	0.98	1	0.97	0.98
	SVM	0.94	0.54	0.69	0.82	0.21	0.33	0.93	0.54	0.68	0.91	0.42	0.57	0.85	0.33	0.47	0.8	0.26	0.4
	GBDT	0.96	0.78	0.86	0.95	0.75	0.84	0.96	0.77	0.85	0.93	0.72	0.81	0.91	0.6	0.72	0.92	0.65	0.76
	CNN	0.78	0.76	0.77	0.86	0.7	0.77	0.86	0.69	0.76	0.85	0.39	0.54	0.36	0.56	0.44	0.63	0.44	0.52
	LSTM	0.99	0.98	**0.99**	0.99	0.96	0.97	0.96	0.85	0.9	0.9	0.92	0.91	0.95	0.88	0.91	0.99	0.94	0.96
Gini	NB	0.7	0.37	0.49	0.76	0.35	0.48	0.77	0.35	0.48	0.75	0.4	0.52	0.78	0.31	0.44	0.75	0.32	0.44
	RF	0.99	**0.99**	**0.99**	1	**0.99**	**0.99**	0.99	0.97	0.98	1	0.97	0.98	1	0.95	0.97	1	0.96	0.98
	SVM	0.94	0.51	0.66	0.92	0.52	0.67	0.92	0.51	0.66	0.89	0.41	0.56	0.84	0.31	0.45	0.78	0.24	0.37
	GBDT	0.96	0.78	0.86	0.96	0.77	0.85	0.97	0.96	0.85	0.91	0.71	0.8	0.92	0.6	0.73	0.92	0.68	0.78
	CNN	0.81	0.71	0.76	0.87	0.73	0.79	0.83	0.78	0.8	0.36	0.72	0.48	0.77	0.42	0.54	0.76	0.36	0.49
	LSTM	1	0.97	0.98	0.99	0.98	0.98	1	0.96	0.98	1	0.98	**0.99**	1	**0.99**	**0.99**	0.98	0.92	0.95

Nghiên cứu sử dụng dữ liệu từ Backblaze (2020) gồm 2 loại ổ cứng ST4000DM000 và ST12000NM0007, thử nghiệm với nhiều mô hình khác nhau (Naïve Bayes, Random Forest (RF), SVM, GBDT, CNN và LSTM) để so sánh kết quả. Kết quả nghiên cứu được nêu ở Bảng 1.5 [8, tr.18].

1.4.3. Nguồn dữ liệu SMART của Backblaze

Backblaze [2] là một công ty cung cấp dịch vụ sao lưu đám mây (cloud backup), sở hữu hàng trăm ngàn ổ đĩa trong trung tâm dữ liệu. Với yêu cầu vận hành liên tục 24/7, Backblaze đặc biệt quan tâm đến việc giám sát tình trạng ổ cứng và tối ưu chi phí bảo trì, thay thế phần cứng. Kể từ năm 2014, họ bắt đầu công khai dữ liệu SMART (Hơn 70 chỉ số ổ cứng bao gồm smart_x_normalized, smart_x_raw, model, serial_number, failure, v.v.) theo thời gian thực với trên 250.000 loại ổ đĩa cứng từ nhiều hãng như Seagate, Western Digital, Toshiba, HGST, v.v.

Backblaze cũng triển khai nhiều nghiên cứu nhằm phát hiện sớm lỗi ổ cứng. Một số kết quả nghiên cứu của Backblaze được nêu ở Bảng 1.3.

Bảng 1.3. Dữ liệu thu thập của Backblaze

Năm	Số ổ đĩa theo dõi	Tổng dung lượng	Tỉ lệ hỏng trung bình (AFR)
2019	~120,000	>600 PB	~1.89%
2022	>200,000	>1.2 EB	~1.37%
2024	~270,000+	>1.6 EB	~1.22%

(Nguồn: https://www.backblaze.com/)

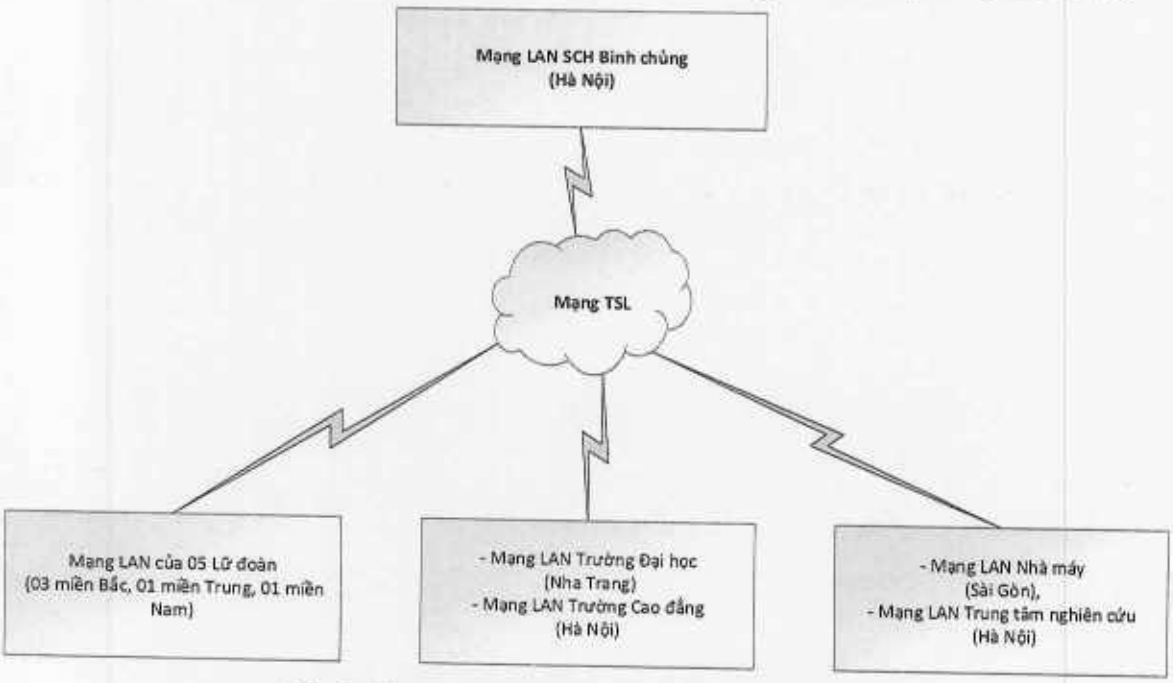
Các kết quả nghiên cứu của Backblaze [2] đã chỉ ra rằng: (1) Không phải tất cả lỗi đều có cảnh báo SMART rõ ràng: Có những ổ đĩa lỗi đột ngột mà không có sự thay đổi rõ rệt về SMART. Điều này thách thức các mô hình truyền thống chỉ dựa trên ngưỡng SMART cố định. (2) SMART_5 (Reallocated Sectors) là chỉ số SMART dẫn đến lỗi ổ đĩa cứng cao nhất. (3) Số giờ hoạt động không tỉ lệ thuận với lỗi.

Để nghiên cứu lỗi ổ cứng, các nguồn dữ liệu SMART rất quan trọng và cần thiết. Do các nguồn dữ liệu ổ cứng không có nhiều, nguồn dữ liệu mở của Backblaze rất quý giá cho các nghiên cứu sử dụng học máy/học sâu.

1.5. Khái quát về mạng máy tính quân sự của Binh chủng Thông tin liên lạc, Bộ Quốc phòng và nhu cầu phát hiện, dự báo lỗi ổ cứng

Mạng máy tính quân sự tại Binh chủng Thông tin liên lạc được tổ chức gồm 10 hệ thống mạng LAN tại các cơ quan, đơn vị kết nối qua hệ thống mạng truyền số liệu, sơ đồ tổng thể như Hình 1.9. Mô hình mạng máy tính quân sự. … Trong hệ thống mạng có hơn 2600 máy tính kết nối, 01 trung tâm dữ liệu, 10 phòng máy chủ với gần 100 máy chủ vật lý. Do đặc điểm về yêu cầu bảo mật thông tin nên hệ thống mạng hoàn toàn độc lập, không kết nối liên thông với bất kỳ hệ thống mạng nào khác (mạng internet, mạng doanh nghiệp, …)

Do đặc thù của quân sự, các dịch vụ trong mạng máy tính quân sự phải bảo đảm thường xuyên, liên tục, thông suốt 24/7. Hạn chế tối đa các sự cố (trong đó có các sự cố do phần cứng máy chủ dịch vụ) làm ảnh hưởng đến các dịch vụ triển khai.



Hình 1.9. Mô hình mạng máy tính quân sự

Cho tới nay, công tác theo dõi, phát hiện lỗi ổ cứng trên mạng vẫn chủ yếu theo hình thức thủ công bằng các công cụ tiện ích của hệ điều hành, chưa có một hệ thống phát hiện lỗi ổ cứng được ứng dụng trong mạng máy tính quân sự của Binh chủng Thông tin liên lạc. Điều này gây khó khăn cho việc sớm phát hiện lỗi, dự đoán

lỗi và lập kế hoạch bảo trì thiết bị. Những vấn đề kỹ thuật đặt ra là: các lỗi ổ cứng rất đa dạng, do nhiều nguyên nhân nên rất khó phát hiện kịp thời, chưa có một giải pháp hiệu quả để phát hiện và dự đoán lỗi, việc sử dụng các phần mềm giám sát thiết bị hiện có còn hạn chế về tính năng dự báo và phát hiện sớm, giải pháp đưa ra cần xem xét đặc thù của mạng máy tính chuyên dụng có tính đóng của quân sự.

Dữ liệu người dùng vẫn chủ yếu được lưu trữ trên ổ đĩa máy tính cá nhân mà không có giải pháp sao lưu dự phòng như sử dụng dịch vụ lưu trữ của bên thứ ba. Nguyên nhân là do yêu cầu bảo mật của hệ thống (mạng độc lập) và yêu cầu bảo mật dữ liệu người dùng (dữ liệu liên quan đến quân sự, quốc phòng). Chính vì vậy vẫn tiềm ẩn rủi ro người dùng bị mất dữ liệu do hỏng hóc ổ cứng gây nên và khi có hỏng hóc xảy ra việc khôi phục, cứu dữ liệu gặp rất nhiều khó khăn (Do đặc thù trong môi trường và bảo mật dữ liệu người dùng).

Vì những lý do trên, việc xây dựng một hệ thống phát hiện lỗi ổ cứng trong mạng máy tính quân sự chuyên dụng của Binh chủng Thông tin liên lạc, Bộ Quốc phòng là một nhu cầu thực tế. Hệ thống xây dựng cần áp dụng những kỹ thuật hiện đại trong lĩnh vực học máy/học sâu để nâng cao độ chính xác, khả năng phát hiện sớm và đưa ra cảnh báo sớm theo các yêu cầu thực tiễn của Binh chủng, giúp phát hiện sớm các lỗi liên quan đến máy chủ cung cấp dịch vụ cũng như máy tính người dùng. Đồng thời, cung cấp số liệu hỗ trợ cho người quản lý, chỉ huy xây dựng kế hoạch, ra quyết định mua sắm các trang thiết bị dự phòng, thay thế, vật tư bảo đảm kỹ thuật một cách khoa học, hiệu quả.

1.6. Kết luận chương

Nội dung của chương tập trung vào các vấn đề nghiên cứu liên quan đến nội dung nghiên cứu của đề án gồm:

- Cơ sở lý thuyết về cấu trúc, nguyên lý hoạt động, cơ chế bảo vệ dữ liệu của các loại ổ cứng hiện nay (điển hình là HDD, SSD, NVMe SSD, Enterprise , RAID) và bài toán phát hiện, dự báo lỗi ổ cứng.

- Trình bày khái quát về công nghệ SMART, một số công cụ theo dõi, giám sát hoạt động ổ cứng như các công cụ tích hợp trên các hệ điều hành (Windows, Linux) và một số giải pháp giám sát tập trung như PRTG, Zambbix, Nagios.

- Phân tích một số nghiên cứu liên quan đến vấn đề phát hiện, dự báo lỗi ổ cứng theo cách truyền thống sử dụng các tham số SMART và theo hướng áp dụng học máy/ học sâu trên thế giới.

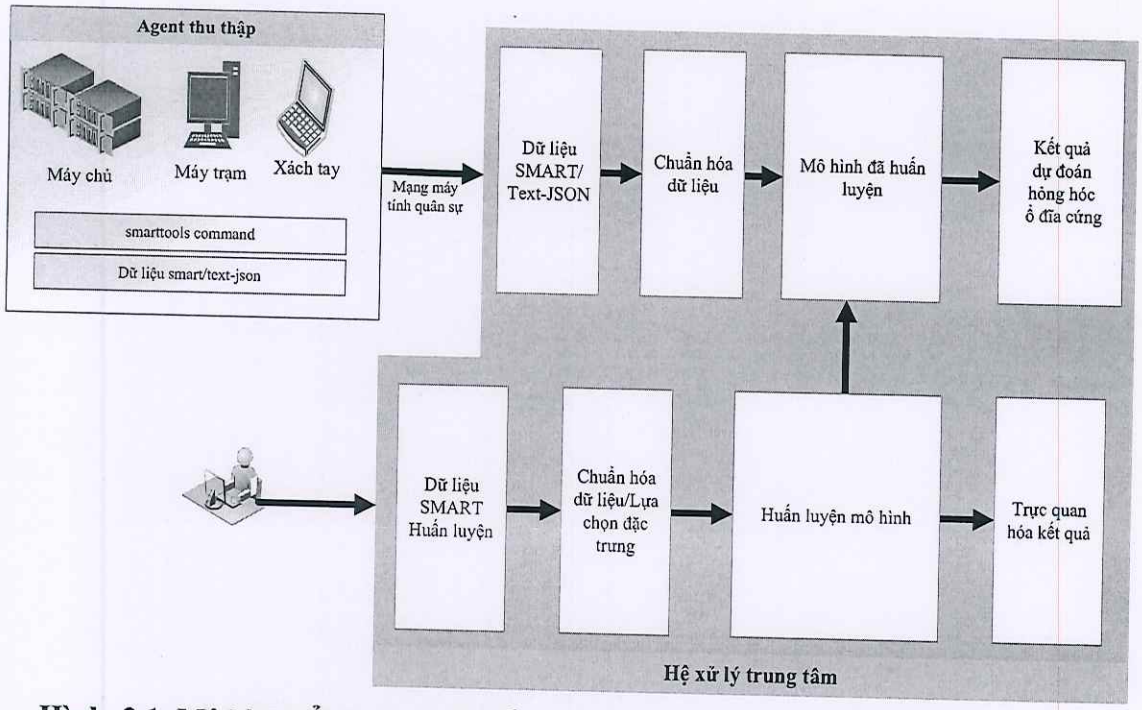
- Giới thiệu mạng máy tính quân sự của Binh chủng Thông tin liên lạc, Bộ Quốc phòng và nhu cầu phát hiện, dự báo lỗi ổ cứng trong mạng.

Chương 2: NGHIÊN CỨU XÂY DỰNG HỆ THỐNG PHÁT HIỆN, DỰ BÁO LỖI Ổ CỨNG VỚI MÔ HÌNH HỌC MÁY

Nội dung chương 2 trình bày về đề xuất mô hình, kiến trúc hệ thống phát hiện, dự báo lỗi ổ cứng trong mạng máy tính quân sự chuyên dụng của Binh chủng Thông tin liên lạc, Bộ Quốc phòng. Các nội dung chính gồm: xây dựng mô hình hệ thống; các thành phần của hệ xử lý trung tâm, tập trung vào các vấn đề về thu thập, phân tích và chuẩn hóa dữ liệu SMART; thực hiện cân bằng dữ liệu; lựa chọn mô hình học máy phù hợp; cách thức huấn luyện và dự báo; phương pháp đánh giá mô hình.

2.1. Xây dựng mô hình hệ thống

Kiến trúc một hệ thống giám sát thiết bị nói chung thường gồm hai hệ chính: hệ Agent thu thập dữ liệu từ thiết bị và hệ xử lý trung tâm thực hiện xử lý, phân tích, đưa ra kết quả cảnh báo. Trên cơ sở đó, mô hình hệ thống phát hiện lỗi ổ cứng trong mạng quân sự được học viên đề xuất như trên Hình 2.1 với 02 thành phần chính: Agent thu thập và hệ xử lý trung tâm.



Hình 2.1. Mô hình tổng quan hệ thống phát hiện hỏng hóc trong mạng quân sự

- Hệ Agent thu thập dữ liệu SMART của ổ cứng: Được triển khai trên máy chủ, máy trạm, máy tính người dùng trong mạng để thu thập các giá trị SMART của ổ đĩa cứng, đóng gói và truyền dữ liệu về máy chủ xử lý.

- Hệ xử lý trung tâm: Thực hiện chuẩn hóa dữ liệu do các Agent đẩy lên, phân tích, dự đoán tình trạng hỏng hóc của ổ đĩa cứng bằng các mô hình đã được huấn luyện, đưa ra kết quả dự đoán (phần trăm tốt/hỏng) của ổ đĩa cứng). Đối với các loại ổ cứng chưa được huấn luyện, người quản trị có thể thu thập dữ liệu (thủ công) và huấn luyện mô hình.

2.2. Hệ Agent thu thập dữ liệu SMART của ổ cứng

Nhiệm vụ của hệ Agent: Thu thập dữ liệu SMART của ổ đĩa cứng trên máy chủ, máy trạm, máy tính người dùng trong mạng máy tính quân sự. Thực hiện đóng gói và đẩy dữ liệu về máy chủ xử lý phân tích qua mạng máy tính quân sự.

Để thu thập dữ liệu SMART, các ổ đĩa cứng phải được tích hợp sẵn công nghệ SMART (công nghệ này được sử dụng trong các ổ cứng ngày nay khá phổ biến). Mỗi hệ điều hành sử dụng (Windows, Linux) sẽ sử dụng các phương thức khác nhau:

+ Hệ điều hành Windows: Sử dụng công cụ smartmontools để thu thập. Chương trình thực thi lấy dữ liệu SMART về đẩy dữ liệu về máy chủ xử lý:

```
@echo off
# Lấy địa chỉ IP của máy
for /f "tokens=2 delims=: " %%A in ('ipconfig ^| findstr /i "IPv4"') do (
   set ip=%%A)
# Loại bỏ khoảng trắng
set ip=%ip: =%
set ip=%ip:.=_%
# Lấy ngày giờ hiện tại
for /f "tokens=1-4 delims=/ " %%a in ("%date%") do (
   set dd=%%a
   set mm=%%b
   set yyyy=%%c )
for /f "tokens=1-2 delims=: " %%a in ("%time%") do (
   set hh=%%a
   set min=%%b )
# Đảm bảo 2 chữ số
if 1%hh% LSS 110 set hh=0%hh%
if 1%min% LSS 110 set min=0%min%
# Tạo tên file
set filename=%ip%_%yyyy%%mm%%dd%_%hh%%min%_smart.txt
# Ghi thông tin SMART
smartctl -a /dev/sda > "\\SERVER_NAME\SharedFolder\%filename%"
```

\\SERVER_NAME được lưu trữ tại thư mục Share_SMART trên máy chủ. File dữ liệu có định dạng *IP_Date_Time_smart.txt*. Sử dụng các công cụ có sẵn trên Windows như *Task Scheduler* để thực thi chương trình tự động theo chu kỳ được thiết lập trước (01 giờ, 01 ngày, ...)

+ Hệ điều hành Linux: Sử dụng ứng dụng smartmontools trên Linux để thực hiện thu thập dữ liệu SMART. Chương trình thực thi như sau:

```
# Lấy địa chỉ IP nội bộ
IP=$(hostname -I | awk '{print $1}' | tr '.' '_')
# Lấy ngày giờ hiện tại
DATE=$(date +%Y%m%d)
TIME=$(date +%H%M)
# Tên file
FILENAME="${IP}_${DATE}_${TIME}_smart.txt"
# Đường dẫn thư mục lưu (ví dụ thư mục chia sẻ NFS hoặc thư mục cục bộ)
OUTPUT_DIR="/mnt/shared_smart_data"
mkdir -p "$OUTPUT_DIR"  # Tạo nếu chưa có
# Lấy dữ liệu SMART (ở /dev/sda) và ghi vào file
smartctl -a /dev/sda > "$OUTPUT_DIR/$FILENAME"
```

Kết quả chương trình đẩy dữ liệu SMART đang thu thập ra thư mục /mnt/shared_smart_data. Thư mục được mount với thư mục \\SERVER_NAME\ Share_SMART trên máy chủ xử lý. File dữ liệu có định dạng *IP_Date_Time_smart.txt*. Tạo crontab để thực thi chương trình tự động theo chu kỳ được thiết lập trước (01 giờ, 01 ngày, ...).

2.3. Hệ xử lý trung tâm

Tập trung các chức năng chính của hệ thống, bao gồm các quá trình chuẩn hóa dữ liệu SMART do các Agent đẩy lên, phân tích dữ liệu, đưa ra kết quả về khả năng hỏng hóc của ổ đĩa cứng. Ngoài ra đối với các model (loại) ổ cứng chưa được huấn luyện, quản trị hệ thống có thể tiến hành huấn luyện bổ sung mô hình.

Tiến trình huấn luyện mô hình cần thực hiện thu thập và tiền xử lý dữ liệu SMART với các bước: Thu thập dữ liệu để huấn luyện, chuẩn hóa dữ liệu SMART, cân bằng dữ liệu SMART. Các bước này cụ thể như sau.

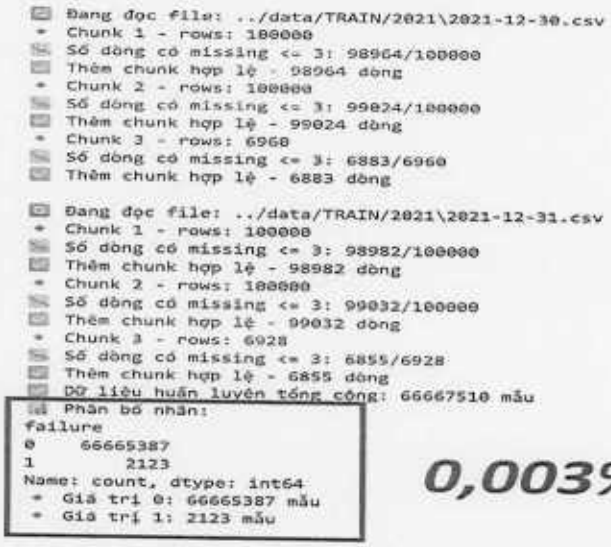
2.3.1. Tiến trình thu thập và tiền xử lý dữ liệu SMART

2.3.1.1. Thu thập dữ liệu để huấn luyện

Để huấn luyện mô hình cho hệ thống, học viên đã sử dụng bộ dữ liệu của Backblaze thu thập được năm 2020 trở lại đây [5]. Các dữ liệu này đã được gán nhãn faild (1) tương đương ổ đĩa hỏng và no_faild (0) ổ đĩa đang hoạt động. Dữ liệu được thu thập lưu trữ dưới dạng các file .csv.

2.3.1.2. Chuẩn hóa dữ liệu SMART

Các dữ liệu SMART của Backblaze thu thập có đặc điểm rất mất cân bằng (tỉ lệ giữa ổ cứng được gán nhãn hỏng - 1 và ổ đĩa không hỏng được gán nhãn - 0 có độ chênh lệch rất lớn). Hình 2.2. thể hiện sự mất cân bằng của dữ liệu thu thập năm 2021, cụ thể: Số lượng ổ cứng không hỏng = 66.665.387 mẫu, trong khi đó số lượng ổ cứng hỏng là 2.123 bản ghi (chỉ bằng 0,003% mẫu hỏng).



Hình 2.2. Tỉ lệ hỏng/không hỏng dữ liệu SMART của Backblaze

Ngoài ra, trong dữ liệu SMART có nhiều trường dữ liệu không hữu ích đối với bài toán phân loại, cần phải được chuẩn hóa, loại bỏ trước khi đưa vào mô hình huấn luyện. Trong nghiên cứu, học viên đã chuẩn hóa các trường dữ liệu sau:

- Loại bỏ các trường mang ý nghĩa cung cấp thông tin như: date (ngày tháng thu thập dữ liệu SMART của ổ đĩa cứng), model (loại ổ cứng), serial_Number, capacity (Dung lượng ổ cứng).

Bảng 2.1. Một số trường dữ liệu SMART thu thập của Backblaze

date	serial_number	model	capacity_by	failure	smart_1_normalized	smart_1_raw	smart_2_normalized	smart_2_raw	smart_3	smart_3	smart_4	smart_4	smart_5	smart_5	smart_7	smart_7	smart_8	smart_8	smart_9
1/2/2021	ZLW0EGG6	ST12000NM001G	1.20001E+13	0	100	2104304			99	0	100	1	100	0	89	722587109			95
1/2/2021	Z305B2QN	ST4000DM000	4.00079E+12	0	118	175898416			91	0	100	19	100	0	87	576938600			50
1/2/2021	ZLW0G6NE	ST12000NM001G	1.20001E+13	0	78	6992144			96	0	100	2	100	0	89	777099064			94
1/2/2021	ZJV0XJQ3	ST12000NM0007	1.20001E+13	0	82	170070192			98	0	100	2	100	0	87	475586677			80
1/2/2021	ZLW18MKT	ST14000NM001G	1.40005E+13	0	80	103967288			95	0	100	4	100	0	81	112739572			99
1/2/2021	Z0H0A0CXF97G	TOSHIBA MG07ACA14TA	1.40005E+13	0	100	0	100	0	100	7790	100	15	100	0	100	0	100	0	89
1/2/2021	ZA1FLE1P	ST8000NM0055	8.00156E+12	0	75	31098384			96	0	100	3	100	0	87	532198811			96
1/2/2021	ZA16NQJR	ST8000NM0055	8.00156E+12	0	80	96039784			89	0	100	8	100	0	86	838762263			64
1/2/2021	ZJV02XWG	ST12000NM0007	1.20001E+13	0	79	71266064			89	0	100	10	100	0	87	465252326			74
1/2/2021	ZA10YBV5	ST8000DM002	8.00156E+12	0	69	8090112			88	0	100	6	100	0	96	3.68E+09			54
1/2/2021	ZJV02XWA	ST12000NM0007	1.20001E+13	0	76	37029920			88	0	100	14	100	0	87	550568698			74
1/2/2021	ZA18CEBS	ST8000NM0055	8.00156E+12	0	84	22697992			94	0	100	4	100	0	96	3.70E+09			67
1/2/2021	Z305DEMG	ST4000DM000	4.00079E+12	0	119	23179920			94	0	100	6	100	0	81	134243168			52
1/2/2021	ZA130TTW	ST8000DM002	8.00156E+12	0	78	59590696			96	0	100	2	100	0	90	1.07E+09			58
1/2/2021	ZLW0EGC5	ST12000NM001G	1.20001E+13	0	82	161414384			98	0	100	2	100	0	88	600382495			95
1/2/2021	ZHZ506CV	ST12000NM0008	1.20001E+13	0	81	136927040			95	0	100	4	100	0	90	567533947			91
1/2/2021	ZJV1CSVV	ST12000NM0007	1.20001E+13	0	79	79912688			96	0	100	3	100	0	88	704210246			80
1/2/2021	ZA18CEBF	ST8000NM0055	8.00156E+12	0	72	15786448			86	0	100	12	100	0	96	3.66E+09			67
1/2/2021	ZJV02XWV	ST12000NM0007	1.20001E+13	0	80	106434000			90	0	100	7	100	0	86	430313266			74
1/2/2021	PL1331LAHD6DZH	HGST HMS5C4040BLE640	4.00079E+12	0	100	0	133	105	100	0	100	4	100	0	100	0	109	44	95
1/2/2021	PL2331LAG9TEEJ	HGST HMS5C4040ALE640	4.00079E+12	0	100	0	135	98	140	548	100	16	100	0	100	0	113	42	96
1/2/2021	Z0H0A0K3F97G	TOSHIBA MG07ACA14TA	1.40005E+13	0	100	0	100	0	100	7589	100	5	100	0	100	0	100	0	97
1/2/2021	PL2331LAH3WYAJ	HGST HMS5C4040BLE640	4.00079E	0	100	0	133	106	129	538	100	10	100	0	100	0	113	42	96
1/2/2021	ZLW0GQ81	ST12000NM001G	1.20001E+13	0	80	107086840			98	0	100	2	100	0	89	739964123			95
1/2/2021	ZLW18MKY	ST14000NM001G	1.40005E+13	0	79	74494312			99	0	100	1	100	0	81	133002040			99
1/2/2021	ZLW0GQ87	ST12000NM001G	1.20001E+13	0	66	3946800			93	0	100	6	100	0	89	781919064			94
1/2/2021	ZLW0GQ86	ST12000NM001G	1.20001E+13	0	83	20761608			96	0	100	4	100	0	83	204079167			98
1/2/2021	ZLW0GQ85	ST12000NM001G	1.20001E+13	0	83	20648920			97	0	100	3	100	0	89	714313210			95
1/2/2021	Z8A0A057F97G	TOSHIBA MG07ACA14TA	1.40005E+13	0	100	0	100	0	100	7928	100	2	100	0	100	0	100	0	75
1/2/2021	ZHZ62XAR	ST12000NM0008	1.20001E+13	0	74	24021112			98	0	100	2	100	0	90	993218615			93
1/2/2021	ZHZ5B3TB	ST12000NM0008	1.20001E+13	0	83	176616184			98	0	100	2	100	0	89	870083871			94
1/2/2021	10B0A0ASF97G	TOSHIBA MG07ACA14TA	1.40005E+13	0	100	0	100	0	100	7802	100	3	100	0	100	0	100	0	89
1/2/2021	ZHZ62XAW	ST12000NM0008	1.20001E+13	0	72	14217680			99	0	100	1	100	0	74	26150855			93
1/2/2021	Z305DEMX	ST4000DM000	4.00079E+12	0	119	221606488			92	0	100	16	100	0	87	533486037			50
1/2/2021	8HH3GHTH	HGST HUH721212ALE604	1.20001E+13	0	100	0	132	95	100	0	100	4	100	0	100	0	128	18	100
1/2/2021	ZA11NWZ6	ST8000DM002	8.00156E+12	0	83	184471568			88	0	100	6	100	0	96	4.28E+09			56
1/2/2021	PL2331LAHDUVVJ	HGST HMS5C4040BLE640	4.00079E+12	0	100	0	134	100	100	0	100	4	100	0	100	0	113	42	95
1/2/2021	K1G953ML	WDC WUH721414ALE6L4	1.40005E+13	0	100	0	138	92	95	231	100	3	100	0	100	0	133	18	100
1/2/2021	Y680A0AFF97G	TOSHIBA MG07ACA14TA	1.40005E+13	0	100	0	100	0	100	7961	100	5	100	0	100	0	100	0	75
1/2/2021	2AGN03VY	HGST HUH721212ALN604	1.20001E+13	0	100	0	132	96	100	0	100	1	100	0	100	0	128	18	99
1/2/2021	ZHZ5YPPW	ST12000NM0008	1.20001E+13	0	84	230965456			98	0	100	2	100	0	84	247099907			98
1/2/2021	8CKSRPKL	HGST HUH721212ALE600	1.20001E+13	0	100	0	132	96	184	366	100	17	100	0	100	0	128	18	100
1/2/2021	ZHZ62XAL	ST12000NM0008	1.20001E+13	0	83	180052696			99	0	100	1	100	0	90	994507775			93

- Dữ liệu SMART của Backblaze mỗi trường giá trị được lưu dưới dạng 02 giá trị Raw (giá trị thô đo được) và Normalized (giá trị đã được chuẩn hóa). Trong nghiên cứu của đề án chỉ xử lý trên các trường giá trị Raw, loại bỏ các cột dữ liệu Normalized.

- Loại bỏ các mẫu dữ liệu không được gán nhãn là hỏng hay tốt (trường failure = NaN), đồng thời loại bỏ các model ổ đĩa cứng chưa hỏng lần nào (tổng failure = 0). Bảng 2.2 thể hiện số lượng mẫu hỏng, không hỏng của một số model ổ cứng được Backblaze thu thập (dữ liệu năm 2022) [5].

Bảng 2.2. Số lượng mẫu hỏng, không hỏng dữ liệu SMART của Backblaze

	count	name	no_failed_count	failed_count
3	1682389	ST4000DM000	1682231	158
2	1826833	ST12000NM0008	1826734	99
0	3513488	TOSHIBA MG07ACA14TA	3513405	83
4	1323908	ST8000NM0055	1323832	76
1	1873090	ST16000NM001G	1873045	45
...	...	...	...	...
48	2576	ST10000NM001G	2576	0
49	2392	ST16000NM005G	2392	0
50	2392	HGST HDS5C4040ALE630	2392	0
51	2298	ST8000DM005	2298	0
70	91	ST1000LM024 HN	91	0

- Đối với model ổ đĩa cứng sau khi thực hiện chuẩn hóa dữ liệu từ các bước ở trên, được lưu vào một file .csv để tiếp tục quá trình chuẩn hóa:

+ Loại bỏ tất cả các cột toàn giá trị NaN.

+ Loại bỏ các cột có độ lệch chuẩn = 0 (giá trị cột không thay đổi)

+ Loại bỏ các mẫu (dòng) chứa giá trị NaN (Giá trị không xác định)

2.3.1.3. Cân bằng dữ liệu SMART

Đối với dữ liệu SMART có tỉ lệ mất cân bằng dữ liệu lớn sẽ ảnh hưởng đến kết quả huấn luyện của mô hình. Để tăng độ chính xác, có thể sử dụng một số thuật

toán cân bằng dữ liệu, trong nghiên cứu học viên sử dụng SMOTE **(Synthetic Minority Over-sampling Technique)** là một kỹ thuật sinh mẫu dữ liệu **giả lập** thuộc lớp thiểu số bằng cách **nội suy tuyến tính** giữa một điểm thiểu số và các điểm lân cận gần nhất trong không gian đặc trưng. Nguyên lý cảu SMOTE như sau:

- Chọn một mẫu thuộc lớp tối thiểu x
- Tìm k láng giềng gần nhất của x (thường k =5)
- Sinh thêm điểm mới bằng công thức: $x\_n = x + \partial * (x\_m - x)$. Trong đó x_n là một hàng xóm ngẫu nhiên và ∂ thuộc [0, 1] là hệ số nội suy ngẫu nhiên.

2.3.2. Mô hình, thuật toán sử dụng cho hệ thống

Trong nghiên cứu này, học viên sử dụng thuật toán Random Forest (RF) và XGBClassifier (XGB) để xây dựng mô hình dự báo, phát hiện hỏng hóc ổ đĩa cứng.

Random Forest: Random Forest là thuật toán kết hợp (ensemble learning), xây dựng từ việc huấn luyện rất nhiều cây quyết định (Decision Tree), sau đó kết hợp kết quả dự đoán của các cây để đưa ra kết quả chính xác hơn và ổn định hơn. Random Forest sử dụng bagging (Bootstrap Aggregating) để huấn luyện song song các cây quyết định. Đối với bài toán phân loại, Random Forest sử dụng kết quả theo hình thức bỏ phiếu đa số (majority voting) từ các cây. Thuật toán Random Forest hoạt động theo các bước:

Bước 1: Bootstrapping dữ liệu: Random Forest chọn ngẫu nhiên một lượng lớn mẫu dữ liệu (sample) từ tập dữ liệu gốc, kích thước các mẫu bằng kích thước dữ liệu ban đầu, nhưng có lấy mẫu lặp lại (một số dữ liệu có thể xuất hiện nhiều lần trong một mẫu).

Bước 2: Xây dựng cây quyết định (Decision Tree):Với mỗi tập dữ liệu con đã tạo, Random Forest xây dựng một cây quyết định (Decision Tree). Điểm đặc biệt của Random Forest: mỗi lần phân chia (split) một nút trong cây quyết định, chỉ có một tập con các đặc trưng (features) được chọn ngẫu nhiên để tìm cách phân chia tốt nhất. Điều này gọi là random feature selection.

Bước 3: Kết hợp kết quả quả dự đoán cuối cùng được tổng hợp từ các cây quyết định con:

- Phân loại: kết quả là lớp (class) được đa số các cây dự đoán.
- Hồi quy: kết quả là trung bình giá trị được dự đoán từ các cây.

Một khó khăn khi áp dụng thuật toán Random Forest là phải tạo lập cấu hình cho các tham số. Thay vì phải thử thủ công từng giá trị tham số cầu hình, thuật toán ***RandomizedSearchCV*** có thể giúp tự động tìm ra cấu hình Random Forest tốt nhất. Điều này rất quan trọng nhằm tối ưu hiệu suất mô hình.

RandomizedSearchCV: Kỹ thuật tìm kiếm siêu tham số tốt nhất cho mô hình. RandomizedSearchCV sẽ chọn ngẫu nhiên n tổ hợp tham số từ không gian của mô hình. Dùng k-fold cross-validation để đánh giá từng tổ hợp. Trả về mô hình với tổ hợp tham số tốt nhất theo tiêu chí đánh giá (F1-score).

Mặt khác, một kỹ thuật có tên gọi là ***Sliding Time Window (STW)*** rất hay được kết hợp với Random Forest cho các tập dữ liệu thời gian (Time-series data), điển hình là các dữ liệu lỗi như lỗi ổ cứng. Random Forest không có khả năng xử lý dữ liệu chuỗi thời gian mà chỉ xử lý dữ liệu dạng bảng với các đặc trưng độc lập. STW sẽ giúp trích xuất các đặc trưng theo cửa sổ thời gian, biến dữ liệu chuỗi thời gian thành tập dữ liệu độc lập để Random Forest có thể xử lý. Ngoài ra, **cơ chế bỏ phiếu chọn lọc (Part-Voting)** có thể giúp nâng cao độ ổn định và tính chính xác trong phát hiện lỗi của hệ thống. Chi tiết về kỹ thuật kết hợp này như sau [7].

Random Forest kết hợp Sliding Time Window (STW) và cơ chế bỏ phiếu chọn lọc (Part-voting): Phương pháp dự đoán xem một ổ cứng có "sắp lỗi" không, bằng cách:

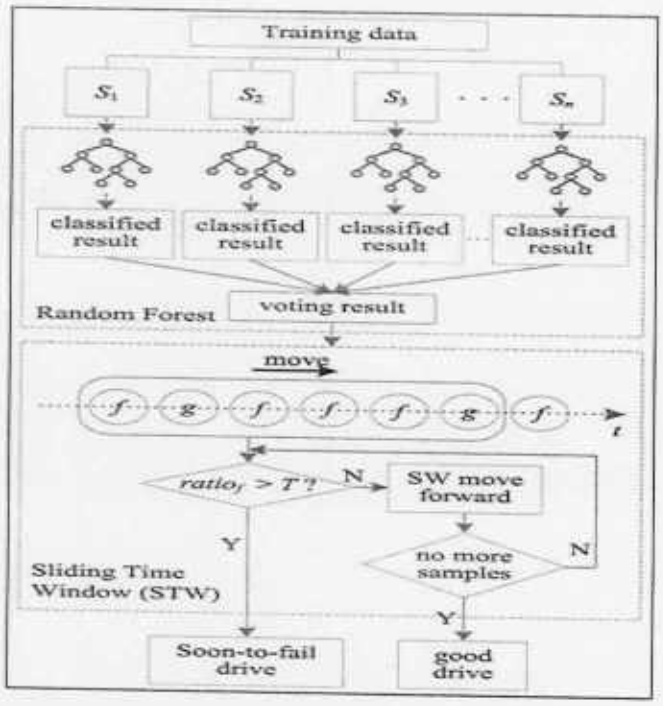
- Dùng nhiều cây quyết định (Random Forest) để phân loại từng mẫu dữ liệu SMART.
- Theo dõi chuỗi kết quả qua một cửa sổ thời gian trượt (Sliding Time Window).
- Nếu tỷ lệ cảnh báo lỗi vượt ngưỡng → kết luận "sắp lỗi".

Thuật toán được mô tả như trong Hình 2.3 *Lưu đồ phương pháp Random Forest kết hợp STW và cơ chế Part-voting.* Trong đó:

- SS: chuỗi mẫu SMART của ổ HDD h (đầu vào)
- Trees: tập các cây trong Random Forest
- ws: kích thước cửa sổ thời gian
- T: ngưỡng tỷ lệ "fail" để kết luận ổ cứng sắp hỏng

Lưu đồ thuật toán:

- Duyệt từng mẫu SMART trong SS.
- Với mỗi mẫu:
  + Dùng các cây trong Trees để phân loại (có chọn lọc theo voting).
  + Đưa kết quả vào STW.
- Tính ratio_f:
  + Nếu ratio_f > T → HDD "soon-to-fail" (Ổ cứng có khả năng hỏng)
  + Ngược lại → HDD "good" (Ổ cứng tốt)
- Di chuyển STW sang bước thời gian tiếp theo và lặp lại.

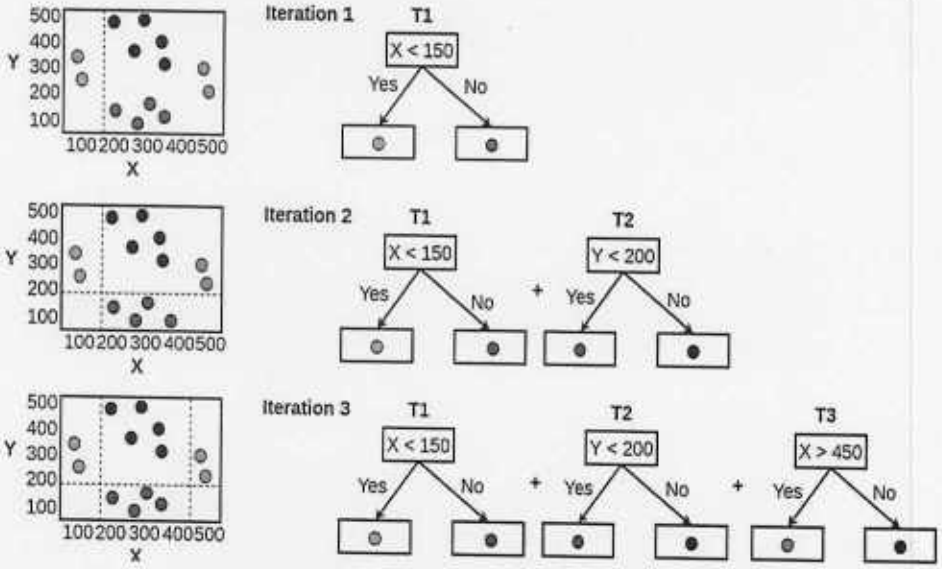


Hình 2.3. Lưu đồ phương pháp Random Forest kết hợp STW và cơ chế Part-voting [7, tr.5]

XGBClassifier: Sử dụng thuật toán Gradient Boosting tối ưu cao để thực hiện bài toán phân loại. XGBClassifier xử lý tốt với bào toán mất cân bằng dữ liệu và kiểm soát overfitting tốt hơn Random Forest.

Gradient Boosting là một thuật toán học máy mạnh mẽ thuộc nhóm học có giám sát, cụ thể là kỹ thuật boosting. *Gradient Boosting* xây dựng mô hình mạnh bằng cách kết hợp nhiều mô hình yếu (thường là cây quyết định nhỏ) một cách tuần tự. Gradient Boosting hoạt động qua các bước:

- Khởi tạo mô hình ban đầu: Mô hình đầu tiên dự đoán giá trị trung bình (hồi quy) hoặc log-odds (phân loại).
- Lặp qua nhiều vòng (rounds):
  + Tính residuals (sai số còn lại): residual = y_true - y_pred
  + Huấn luyện một mô hình mới để dự đoán residuals.
  + Cập nhật mô hình tổng thể bằng cách cộng thêm mô hình mới nhân với learning rate.
- Tổng hợp tất cả các mô hình nhỏ để tạo mô hình cuối cùng.



Hình 2.4. Mô hình hoạt động thuật toán Gradient Boosting [1, tr. 229]

Các hàm tính toán trong Gradient Boosting: Hàm mất mát: L(y, F(x)) và Mô hình tổng: $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$. Trong đó:

- $h_m(x)$ là cây quyết định nhỏ thứ m học theo gradient của hàm mất mát.

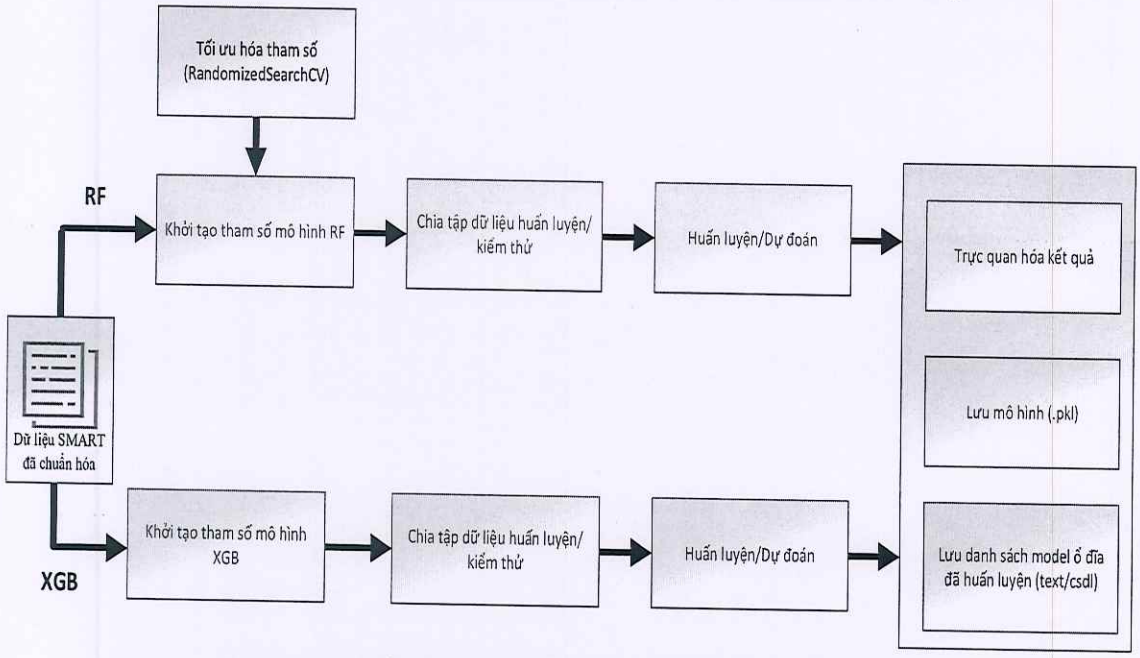
- γ_m là hệ số learning rate.

- $F_0(x)$ là mô hình khởi đầu.

Ta tìm $h_m(x)$ sao cho: $h_m(x) \approx -\nabla_{(F_{m-1})} L(y, F_{m-1}(x))$

2.3.3. Tiến trình huấn luyện mô hình

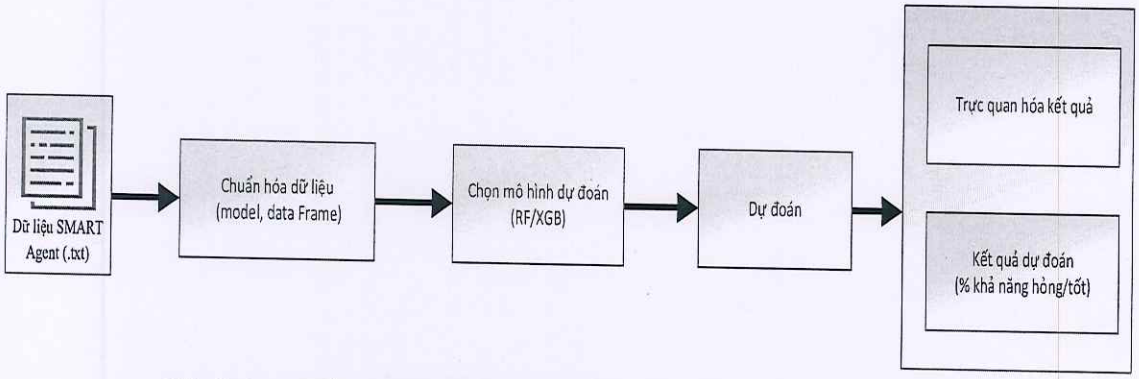
Tiến trình thực hiện được thể hiện ở Hình 2.5. Tiến trình huấn luyện mô hình RF và XGB gồm các bước: Đọc dữ liệu SMART, lựa chọn mô hình (RF hoặc XGB), Chia tập dữ liệu huấn luyện, Huấn luyện, Lưu mô hình, trực quan kết quả.



Hình 2.5. Tiến trình huấn luyện mô hình RF và XGB

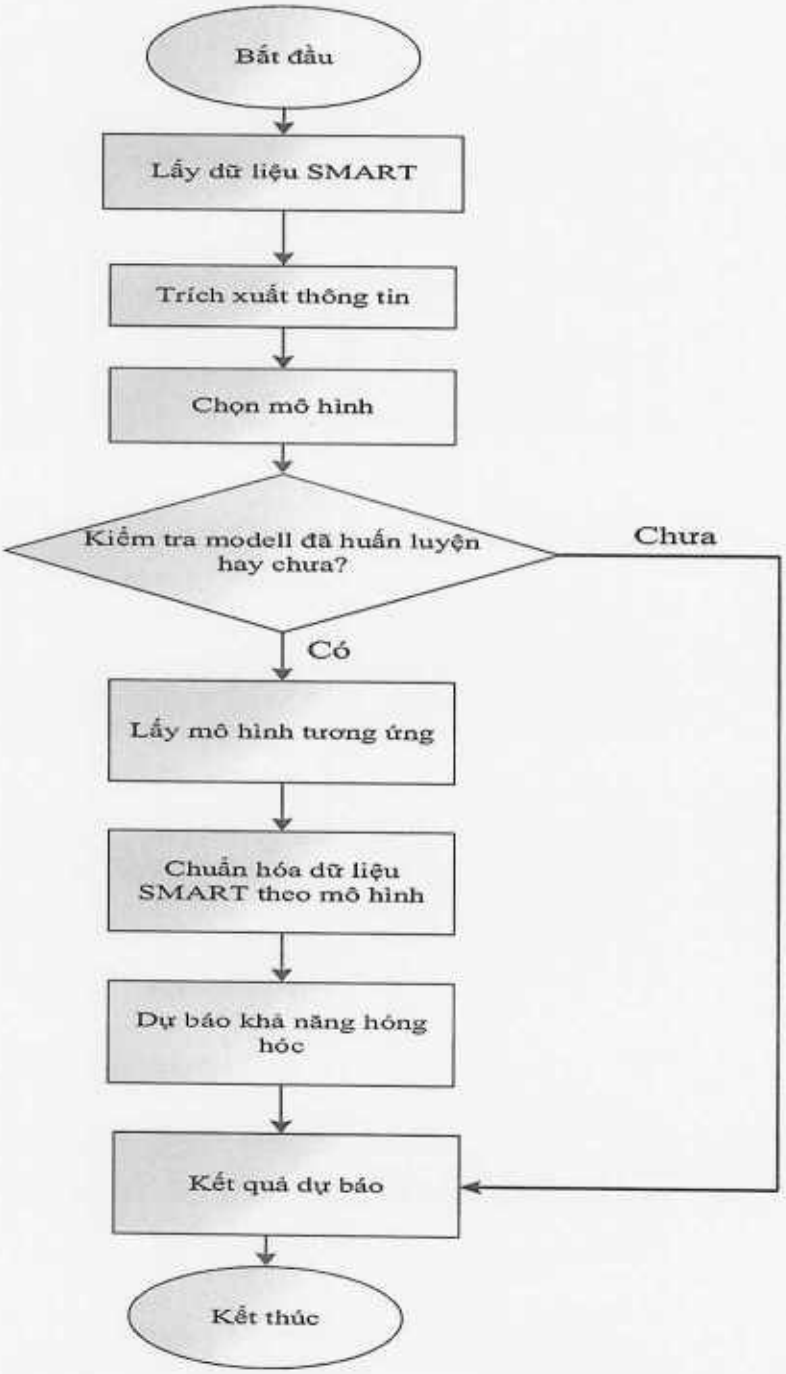
2.3.4. Tiến trình dự đoán, phát hiện hỏng hóc ổ cứng

Tiến trình dự đoán, phát hiện hỏng hóc ổ cứng thể hiện trên Hình 2.6.



Hình 2.6. Tiến trình dự đoán phần trăm hỏng hóc ổ đĩa cứng

Tiến trình này sử dụng mô hình đã huấn luyện để dự đoán kết quả khả năng hỏng hóc của một ổ đĩa với dữ liệu SMART thu thập được trong thực tế. Các bước gồm: Đọc dữ liệu SMART (do Agent thu thập đẩy về, Chuẩn hóa dữ liệu, Chọn mô hình dự báo, Dự báo, Hiển thị kết quả.



Hình 2.7. Lưu đồ thuật toán dự đoán hỏng hóc ổ đĩa cứng

2.3.5. Kết quả dự báo

Sau khi huấn luyện xong, mỗi model ổ đĩa cứng được lưu thành 01 file .pkl tương ứng với mô hình RF hoặc XGB. Với phương pháp tiếp cận, mỗi model ổ cứng sẽ có tập đặc trưng khác nhau:

- ***Model ST12000NM001G***: 18 đặc trưng

['smart_1_raw', 'smart_4_raw', 'smart_5_raw', 'smart_7_raw', 'smart_9_raw', 'smart_12_raw', 'smart_187_raw', 'smart_188_raw', 'smart_190_raw', 'smart_192_raw', 'smart_193_raw', 'smart_194_raw', 'smart_197_raw', 'smart_198_raw', 'smart_199_raw', 'smart_240_raw', 'smart_241_raw', 'smart_242_raw']

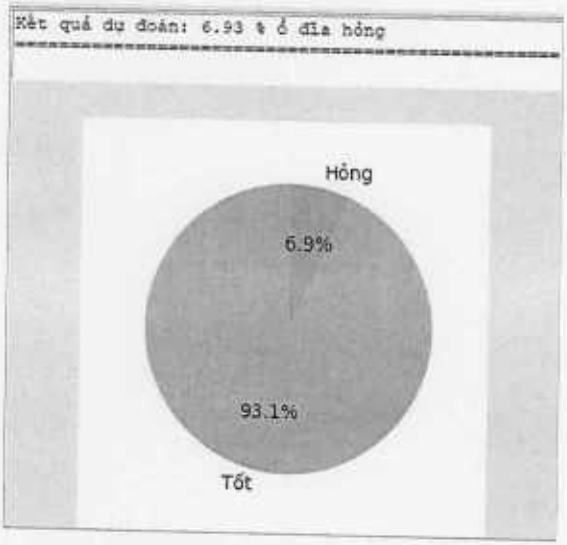
- ***ST8000NM0055***: 20 đặc trưng

['smart_1_raw', 'smart_4_raw', 'smart_5_raw', 'smart_7_raw', 'smart_9_raw', 'smart_12_raw', 'smart_187_raw', 'smart_188_raw', 'smart_190_raw', 'smart_191_raw', 'smart_192_raw', 'smart_193_raw', 'smart_194_raw', 'smart_195_raw', 'smart_197_raw', 'smart_198_raw', 'smart_199_raw', 'smart_240_raw', 'smart_241_raw', 'smart_242_raw']

- ***ST4000DM000:*** *21 đặc trưng*

['smart_1_raw', 'smart_4_raw', 'smart_5_raw', 'smart_7_raw', 'smart_9_raw', 'smart_12_raw', 'smart_183_raw', 'smart_184_raw', 'smart_187_raw', 'smart_188_raw', 'smart_189_raw', 'smart_190_raw', 'smart_192_raw', 'smart_193_raw', 'smart_194_raw', 'smart_197_raw', 'smart_198_raw', 'smart_199_raw', 'smart_240_raw', 'smart_241_raw', 'smart_242_raw']

Kết quả của nghiên cứu dự báo khả năng hỏng của ổ đĩa cứng theo tỉ lệ phần trăm hỏng và tỉ lệ phần trăm không hỏng



Hình 2.8: Kết quả dự báo tỉ lệ hỏng hóc của ổ đĩa cứng

2.4. Phương pháp đánh giá mô hình

Trong nghiên cứu, học viên chia tập dữ liệu SMART thành 02 thành phần:

- Dữ liệu huấn luyện (train): Chiếm tỉ lệ 70% hoặc 80%

- Dữ liệu kiểm thử (test) để đánh giá mô hình: Dữ liệu còn lại (chiếm 20% đến 30%).

Để đánh giá hiệu quả của mô hình, học viên đánh giá trên tập các giá trị sau:

- Tính chính xác (Accuracy) = (Số mẫu dự đoán đúng) / (Tổng số mẫu)

- Độ chính xác (Precision) = Tỉ lệ mẫu dự đoán là "hỏng" thực sự bị hỏng (giảm báo động giả). Precision = TP / (TP+FP). Trong đó:

+ TP: True Positive (dự đoán đúng hỏng)

+ FP: False Positive (báo sai là hỏng → ổ tốt mà báo hỏng)

- Độ bao phủ (Recall): Tỉ lệ ổ hỏng thực sự được mô hình phát hiện đúng (giảm bỏ sót ổ hỏng). Recall = TP/(TP + FN). Trong đó:

+ TP: True Positive (dự đoán đúng hỏng)

+ FN: False Negative (bỏ sót → ổ hỏng mà báo tốt)

- Trung bình điều hòa giữa Precision và Recall (F1-score): F1-score là chỉ số lý tưởng để đánh giá mô hình phân loại mất cân bằng, giúp cân bằng giữa "phát hiện đúng" và "tránh báo động sai".

$$F1\text{-score} = 2 \times ((Precision \times Recall) / (Precision + Recall))$$

Bảng 2.3. Mối liên hệ giữa các giá trị đánh giá mô hình

Tình huống thực tế	Precision	Recall	F1-score	Giải thích
Mô hình báo đúng hỏng và không hỏng	Cao	Cao	Cao	Dự đoán tốt, cân bằng
Mô hình báo sai nhiều ổ tốt là hỏng	Cao	Thấp	Trung bình/thấp	Báo đúng thì đúng, nhưng bỏ sót nhiều ổ hỏng
Mô hình báo tất cả đều hỏng	Thấp	Cao	Trung bình/thấp	Không bỏ sót nhưng báo sai quá nhiều

```
 * RandomForestClassifier

Time to train: 0.5837013999621073  mins

- Results on test set:
Accuracy:  0.9992097984986171
Scores:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2525
           1       0.75      1.00      0.86         6

    accuracy                           1.00      2531
   macro avg       0.88      1.00      0.93      2531
weighted avg       1.00      1.00      1.00      2531
```

Hình 2.9. Kết quả đánh giá mô hình RF

2.5. Kết luận chương

Trong chương 2, đề án đã đề xuất xây dựng mô hình hệ thống phát hiện, dự báo hỏng hóc ổ cứng trong mạng quân sự. Nội dung chương đã trình bày về các vấn đề:

- Xây mô hình hệ thống gồm: Hệ Agent thu thập dữ liệu SMART từ các ổ cứng; Hệ xử lý trung tâm với các thành phần chức năng, luồng xử lý dữ liệu để đưa ra kết quả dự báo.

- Các vấn đề về thu thập dữ liệu huấn luyện, chuẩn hóa dữ liệu với việc loại bỏ các thuộc tính SMART không quan trọng, ảnh hưởng xấu đến kết quả mô hình dự báo, lựa chọn các đặc trưng phù hợp với từng model (loại) ổ đĩa cứng; thực hiện cân bằng dữ liệu.

- Lựa chọn mô hình học máy, cụ thể: Sử dụng mô hình Random Forest và XGBoost cho bài toán dự báo. Lựa chọn siêu tham số cho mô hình, chia tập dữ liệu huấn luyện/kiểm thử. Đánh giá mô hình dựa trên các bộ tham số khác nhau (Accuracy, Precision, Recall, F1-score), lưu mô hình đã được huấn luyện.

- Đưa ra lưu đồ thuật toán dự đoán hỏng hóc ổ cứng, thử nghiệm mô hình với luồng dữ liệu SMART thực tế với các bước: Thu thập; Chuẩn hóa; Lựa chọn mô hình dự báo; Hiển thi kết quả dự báo. dữ liệu dự đoán.

Chương 3: TRIỂN KHAI THỬ NGHIỆM, ĐÁNH GIÁ

Nội dung chương 3 trình bày về việc triển khai thử nghiệm hệ thống phát hiện, dự báo hỏng hóc ổ cứng trong mạng quân sự đã xây dựng. Các nội dung chính gồm: thiết lập môi trường thử nghiệm, hệ Agent thu thập dữ liệu SMART và cài đặt phần mềm; Triển khai thử nghiệm hệ thống với tập dữ liệu Backblaze và tập dữ liệu do học viên thu thập thực tế tại đơn vị công tác. Các kết quả thử nghiệm đạt được có độ chính xác cao, hệ thống hoạt động ổn định.

3.1. Thiết lập môi trường thử nghiệm

3.1.1. Môi trường phát triển phần mềm hệ thống

Trong phạm vi nghiên cứu của đề án, học viên đã thực hiện thiết lập môi trường thực nghiệm với các công cụ gồm:

- Ngôn ngữ và thư viện lập trình: Sử dụng ngôn ngữ Python 3.12.0
- Thư viện học máy gồm: sklearn, xgboost, imbalanced-learn, shap
- Thư viện xử lý dữ liệu: pandas, numpy, ..
- Thư viện trực quan hóa: matplotlib, seaborn, …
- IDE sử dụng: Jupyter Notebook, VSCode

Phần mềm thử nghiệm được tạo với một Project gồm các thành phần chính như sau:

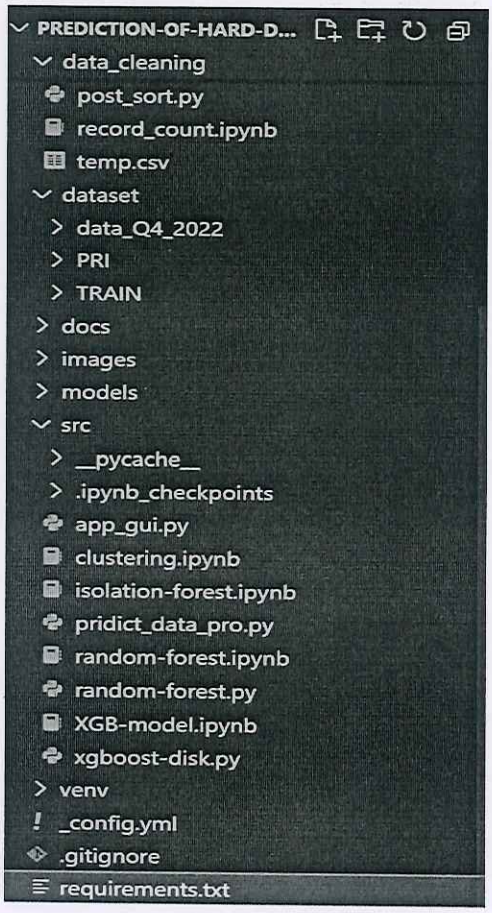
+ data_cleaning: Có chức năng chuẩn hóa dữ liệu SMART trước khi đưa vào mô hình huấn luyện.

+ dataset: chứa 03 tập dữ liệu: Dữ liệu thu thập từ Backblaze (chưa được chuẩn hóa), dữ liệu để huấn luyện mô hình (đã chuẩn hóa), dữ liệu thực tế thu thập được (dữ liệu dự báo).

+ models: Lưu mô hình đã huấn luyện gồm mô hình RF và XGB.

+ src: Chứa các hàm để huấn luyện, dự báo, kết xuất dữ liệu dự báo (giao diện)

+ requiremnets.txt: File chứa các thư viện cần phải cài đặt cho chương trình.



Hình 3.1. Cấu trúc Project đề án

3.1.2. Hệ Agent thu thập dữ liệu SMART

- Dữ liệu huấn luyện: Sử dụng dữ liệu Backblaze từ năm 2022 đến nay. Được tải từ trang chủ của Backblaze [5].

- Dữ liệu thực tế tại đơn vị công tác: Agent tự động lấy dữ liệu SMART qua công cụ smartctl, lưu định kỳ 1 ngày/lần. Thử nghiệm với 02 máy chủ Windows, 01 Linux và 03 máy trạm tại đơn vị công tác.

3.1.3. Triển khai các thuật toán học máy

3.1.3.1. Triển khai thuật toán Random Forest

Thuật toán được khởi tạo với bộ tham số:

- *n_estimators=2000*: Số lượng cây quyết định trong rừng (Nhiều cây hơn giúp giảm overfitting và tăng độ chính xác, nhưng sẽ tăng thời gian huấn luyện).

- *min_samples_split=5:* Một nút sẽ được chia nếu có ít nhất 5 mẫu (Giúp tránh

việc chia quá sớm gây overfitting).

- *min_samples_leaf=4:* Mỗi nút lá phải chứa ít nhất 4 mẫu (giúp cây không quá sâu và ổn định hơn).

- *max_features='sqrt':* Ở mỗi nút chia, chỉ xem xét √(số đặc trưng) → tăng tính ngẫu nhiên, giúp giảm overfittin

-*max_depth=10:* Giới hạn độ sâu của cây để kiểm soát độ phức tạp, giảm overfitting.

- *criterion='entropy'*: Hàm đo mức độ thu được thông tin khi chia (tốt hơn Gini)

- *bootstrap=True:* Lấy mẫu với lặp lại để huấn luyện từng cây

Ngoài ra, sử dụng hàm RandomizedSearchCV() để tối ưu các siêu tham số cho mô hình. Các thông số chính của hàm:

- *n_iter=100*: Thử ngẫu nhiên 100 tổ hợp siêu tham số từ random_grid (giảm thời gian so với GridSearchCV).

- *cv =3*: Chia dữ liệu thành 3 phần (k-fold cross validation) để đánh giá ổn định mỗi cấu hình.

- *scoring=["f1", "accuracy"]*: Đánh giá mô hình theo cả F1-score và Accuracy.

Ngoài ra để tối ưu các bộ tham số sử dụng trong thuật toán Random Forest, đề án đã sử dụng các kỹ thuật ***RandomizedSearchCV*** để tự động tìm ra cấu hình tham số tốt nhất và tối ưu hiệu suất thuật toán Random Forest. Kỹ thuật ***Random Forest kết hợp Sliding Time Window (STW) và cơ chế bỏ phiếu chọn lọc (Part-voting)***: được áp dụng để tăng độ ổn định và tính chính xác trong phát hiện và dự báo lỗi ổ cứng.

3.1.3.2. Triển khai thuật toán XGBoost Classifier

Khởi tạo bộ tham số cho thuật toán XGBoost Classifier như sau:

- *learning_rate=0.2*: Tốc độ học

- *n_estimators=1251*: Số cây được huấn luyện

- *max_depth = 3:* Độ sâu tối đa của mỗi cây – kiểm soát độ phức tạp. Nhỏ để tránh overfitting.

- *min_child_weight =3*: Số lượng mẫu tối thiểu trong một nút để chia tiếp – giá trị cao giúp mô hình tổng quát hơn.

- *gamma = 0.1*: Mức tăng lợi ích tối thiểu để chia cây – kiểm soát việc chia nhỏ không cần thiết.

- *subsample = 0.8*: Tỷ lệ mẫu được chọn ngẫu nhiên để huấn luyện mỗi cây – giảm overfitting.

- *colsample_bytree = 0.8*: Tỷ lệ số đặc trưng được chọn ngẫu nhiên cho mỗi cây – tăng độ đa dạng.

- *objective = 'binary:logistic'*: Bài toán phân loại nhị phân, đầu ra là xác suất.

- *scale_pos_weight = 1*: Trọng số lớp dương – sử dụng khi dữ liệu mất cân bằng.

Ngoài ra, đề án sử dụng hàm RandomizedSearchCV() để tối ưu các siêu tham số cho mô hình XGB.

3.2. Kết quả thử nghiệm trên tập dữ liệu của Backlblaze

3.2.1. Tập dữ liệu SMART của Backblaze

Trong nghiên cứu này, học viên sử dụng dữ liệu năm 2022 [5] để huấn luyện mô hình với các tham số sau:

- Dung lượng (dữ liệu gốc): 26,3 GB (Dữ liệu trong 1 năm, tần suất 01 lần/ngày).

- Model ổ cứng: 71 model ổ cứng được thu thập. Số lượng mẫu thu thập một số model ổ cứng được nêu ở Bảng 3.1

Bảng 3.1.Tổng hợp mẫu dữ liệu một số model ổ cứng của Backblaze

Số lượng mẫu	Model ổ cứng	Số lượng mẫu tốt	Số mẫu hỏng
1682389	ST4000DM000	1682231	158
1826833	ST12000NM0008	1826734	99
3513488	TOSHIBA MG07ACA14TA	3513405	83
1323908	ST8000NM0055	1323832	76
1873090	ST16000NM001G	1873045	45

Số lượng mẫu	Model ổ cứng	Số lượng mẫu tốt	Số mẫu hỏng
878386	ST8000DM002	878345	41
988602	ST14000NM001G	988568	34
1157857	ST12000NM001G	1157831	26
991452	HGST HUH721212ALN604	991429	23
107973	ST10000NM0086	107955	18
1210844	HGST HUH721212ALE604	1210830	14
140485	ST14000NM0138	140471	14
546519	TOSHIBA MG08ACA16TE	546506	13
116481	ST12000NM0007	116471	10
342851	HGST HMS5C4040ALE640	342843	8
22343	TOSHIBA MQ01ABF050	22336	7
23773	TOSHIBA MQ01ABF050M	23766	7
1170975	HGST HMS5C4040BLE640	1170968	7
455051	TOSHIBA MG08ACA16TEY	455045	6
847187	WDC WUH721816ALE6L4	847181	6
20711	ST500LM030	20706	5
345028	TOSHIBA MG08ACA16TA	345023	5

Số lượng mẫu	Model ổ cứng	Số lượng mẫu tốt	Số mẫu hỏng
81512	ST6000DX000	81511	1

3.2.2. Kết quả thử nghiệm với tập dữ liệu Backblaze

Kết quả thử nghiệm cho 7 loại ổ cứng phổ biến được đánh giá dựa theo 04 tiêu chí gồm: Accuracy, Precision, Recall, F1-score. Kết quả thử nghiệm hệ thống với tập dữ liệu Backblaze được thể hiện ở Bảng 3.2 với 2 thuật toán Random Forest và XGBoost.

Bảng 3.2. Kết quả thử nghiệm mô hình với dữ liệu Backblaze

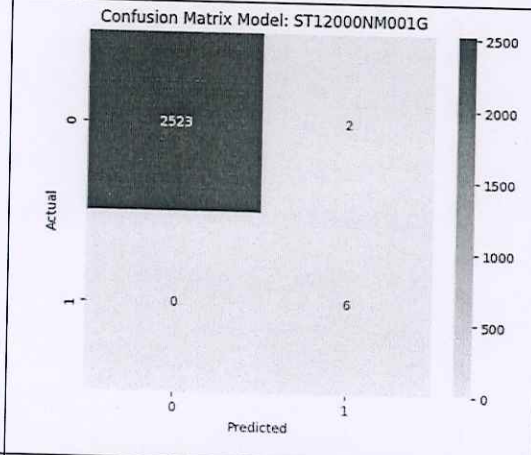
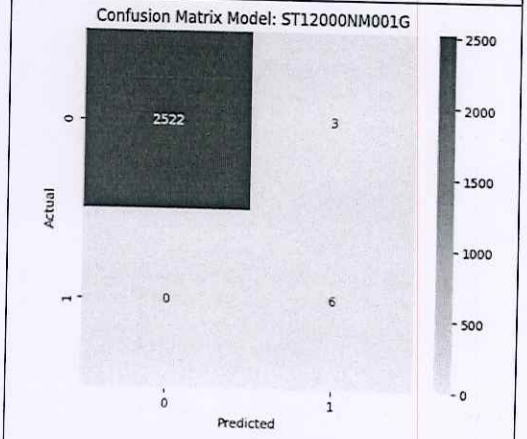
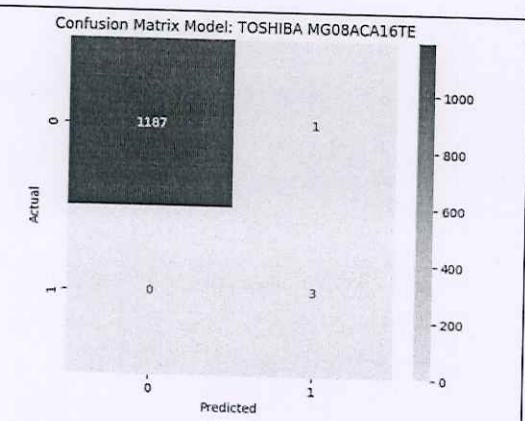
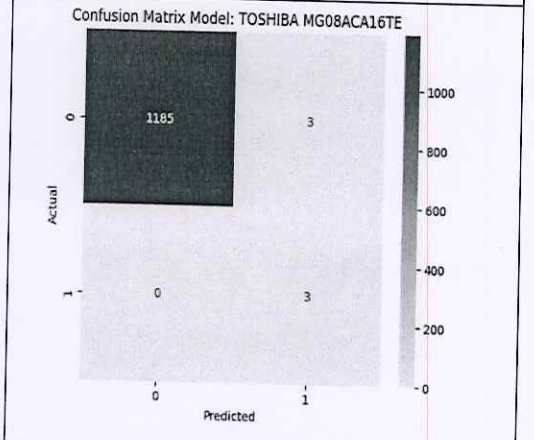
Model	Random Forest	XGBoost *Classifier*
ST12000NM001G	- Results on test set: Accuracy: 0.9992897984986171 Scores: precision recall f1-score support 0 1.00 1.00 1.00 2525 1 0.75 1.00 0.86 6 accuracy 1.00 2531 macro avg 0.88 1.00 0.93 2531 weighted avg 1.00 1.00 1.00 2531	- Results on test set: Accuracy: 0.9988146977479258 Scores: precision recall f1-score support 0 1.00 1.00 1.00 2525 1 0.67 1.00 0.80 6 accuracy 1.00 2531 macro avg 0.83 1.00 0.90 2531 weighted avg 1.00 1.00 1.00 2531
ST12000NM0008	- Results on test set: Accuracy: 0.9959849435382685 Scores: precision recall f1-score support 0 1.00 1.00 1.00 3965 1 0.56 1.00 0.71 20 accuracy 1.00 3985 macro avg 0.78 1.00 0.86 3985 weighted avg 1.00 1.00 1.00 3985	- Results on test set: Accuracy: 0.9962358845671268 Scores: precision recall f1-score support 0 1.00 1.00 1.00 3965 1 0.57 1.00 0.73 20 accuracy 1.00 3985 macro avg 0.79 1.00 0.86 3985 weighted avg 1.00 1.00 1.00 3985

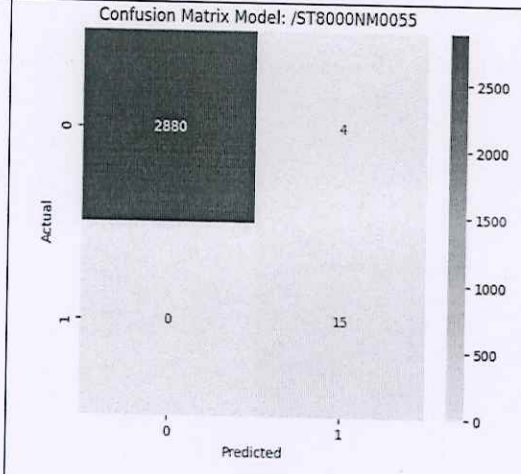
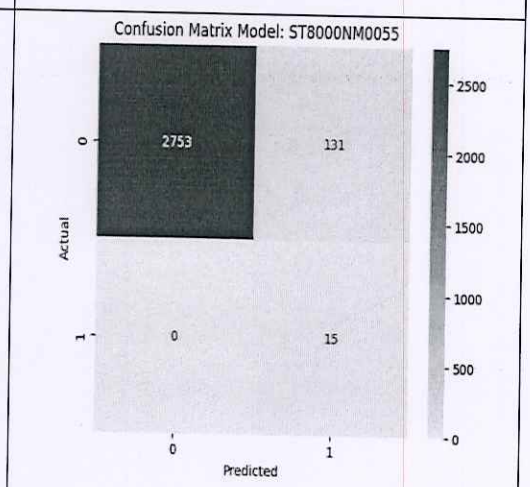
Model	Random Forest	XGBoost *Classifier*
TOSHIBA MG08ACA16TE	- Results on test set: Accuracy: 0.9991603694374476 Scores: precision recall f1-score support 0 1.00 1.00 1.00 1188 1 0.75 1.00 0.86 3 accuracy 1.00 1191 macro avg 0.88 1.00 0.93 1191 weighted avg 1.00 1.00 1.00 1191	- Results on test set: Accuracy: 0.9974811083123426 Scores: precision recall f1-score support 0 1.00 1.00 1.00 1188 1 0.50 1.00 0.67 3 accuracy 1.00 1191 macro avg 0.75 1.00 0.83 1191 weighted avg 1.00 1.00 1.00 1191
ST8000NM0055	- Results on test set: Accuracy: 0.9986202138668506 Scores: precision recall f1-score support 0 1.00 1.00 1.00 2884 1 0.79 1.00 0.88 15 accuracy 1.00 2899 macro avg 0.89 1.00 0.94 2899 weighted avg 1.00 1.00 1.00 2899	- Results on test set: Accuracy: 0.9548120041393584 Scores: precision recall f1-score support 0 1.00 0.95 0.98 2884 1 0.10 1.00 0.19 15 accuracy 0.95 2899 macro avg 0.55 0.98 0.58 2899 weighted avg 1.00 0.95 0.97 2899
ST10000NM0086	- Results on test set: Accuracy: 1.0 Scores: precision recall f1-score support 0 1.00 1.00 1.00 235 1 1.00 1.00 1.00 4 accuracy 1.00 239 macro avg 1.00 1.00 1.00 239 weighted avg 1.00 1.00 1.00 239	- Results on test set: Accuracy: 1.0 Scores: precision recall f1-score support 0 1.00 1.00 1.00 235 1 1.00 1.00 1.00 4 accuracy 1.00 239 macro avg 1.00 1.00 1.00 239 weighted avg 1.00 1.00 1.00 239
ST14000NM0138	- Results on test set: Accuracy: 0.993485342019544 Scores: precision recall f1-score support 0 1.00 0.99 1.00 304 1 0.60 1.00 0.75 3 accuracy 0.99 307 macro avg 0.80 1.00 0.87 307 weighted avg 1.00 0.99 0.99 307	- Results on test set: Accuracy: 0.9804560260586319 Scores: precision recall f1-score support 0 1.00 0.98 0.99 304 1 0.33 1.00 0.50 3 accuracy 0.98 307 macro avg 0.67 0.99 0.75 307 weighted avg 0.99 0.98 0.99 307

Model	Random Forest	XGBoost *Classifier*
ST8000DM002	- Results on test set: Accuracy: 0.9963427377220481 Scores: precision recall f1-score support 0 1.00 1.00 1.00 1905 1 0.56 1.00 0.72 9 accuracy 1.00 1914 macro avg 0.78 1.00 0.86 1914 weighted avg 1.00 1.00 1.00 1914	- Results on test set: Accuracy: 0.9937304075235109 Scores: precision recall f1-score support 0 1.00 0.99 1.00 1905 1 0.43 1.00 0.60 9 accuracy 0.99 1914 macro avg 0.71 1.00 0.80 1914 weighted avg 1.00 0.99 0.99 1914

Bảng 3.3 thể hiện kết quả đánh giá mô hình bằng ma trận Confusion với 2 thuật toán Random Forest và XGBoost cho 3 loại ổ cứng.

Bảng 3.3. Kết quả đánh giá mô hình bằng ma trận Confusion

Model	Random Forest	XGBoost *Classifier*
ST12000NM001G		
TOSHIBA MG08ACA16TE		

Model	Random Forest	XGBoost *Classifier*
ST8000NM0055		

Bảng 3.4. Kết quả đánh giá mô hình dựa trên Precision, Recall, F1-score

Model ổ cứng	Phân lớp ổ cứng	Precision		Recall		F1-score	
		RF	XGB	RF	XGB	RF	XGB
ST12000NM001G	Tốt (0)	1,00	1,00	1,00	1,00	1,00	1,00
	Hỏng (1)	0,75	0,67	1,00	1,00	0,86	0,80
ST12000NM0008	Tốt (0)	1,00	1,00	1,00	1,00	1,00	1,00
	Hỏng (1)	0,56	0,56	1,00	1,00	0,71	0,73
TOSHIBA MG08ACA16TE	Tốt (0)	1,00	1,00	1,00	1,00	1,00	1,00
	Hỏng (1)	0,75	0,50	1,00	1,00	0,86	0,67
ST8000NM0055	Tốt (0)	1,00	1,00	1,00	1,00	0,95	0,98
	Hỏng (1)	0,79	0,10	1,00	1,00	0,88	0,19
ST10000NM0086	Tốt (0)	1,00	1,00	1,00	1,00	1,00	1,00
	Hỏng (1)	1,00	1,00	1,00	1,00	1,00	1,00
ST14000NM0138	Tốt (0)	1,00	1,00	0,99	0,98	1,00	0,99
	Hỏng (1)	0,60	0,33	1,00	1,00	0,75	0,50
ST8000DM002	Tốt (0)	1,00	1,00	1,00	0,99	1,00	1,00
	Hỏng (1)	0,56	0,43	1,00	1,00	0,72	0,60

Từ thử nghiệm trên tập dữ liệu SMART của Backblaze, các kết quả trên Bảng 3.2, 3.3 và Bảng 3.4 cho thấy: Mô hình sử dụng trong nghiên cứu không bỏ sót ổ cứng

lỗi (tỷ lệ recall ~ 100%), tỉ lệ cảnh báo giả ở mức chấp nhận được (ma trận nhầm lẫn) cho hệ thống ưu tiên cảnh báo sớm (không bỏ sót ổ hỏng).

3.3. Thử nghiệm với dữ liệu thu thập thực tế tại đơn vị công tác

Thực nghiệm với dữ liệu SMART thực tế tại đơn vị: Dữ liệu SMART thu thập tại đơn vị được thể hiện ở Bảng 3.4

Bảng 3.5. Dữ liệu SMART thu thập tại đơn vị công tác

Loại máy	Model ổ cứng	Chỉ số SMART	Số mẫu hỏng
Máy trạm	ST4000DM000	19	0
Máy trạm	KBG4AZNV512G KIOXIA	18	0
Máy trạm	SANDISK Z400S 2.5 7MM 256GB	17	0
Máy chủ	HGST HTS545050A7E680	23	0
Máy chủ	ST1000DM010-2EP102	23	0
Máy trạm	TOSHIBA DT01ACA050	17	0

Các thông tin SMART thu thập được (Model ổ đĩa ST4000DM000) được thể hiện ở Hình 3.2.

```
=== START OF INFORMATION SECTION ===
Model Number:                       ST4000DM000
Serial Number:                      AYD7N017811407K2Y
Firmware Version:                   6100114l
PCI Vendor/Subsystem ID:            0x1c5c
IEEE OUI Identifier:                0xace42e
Controller ID:                      1
NVMe Version:                       1.4
Number of Namespaces:               1
Namespace 1 Size/Capacity:          1,024,209,543,168 [1.02 TB]
Namespace 1 Formatted LBA Size:     512
Namespace 1 IEEE EUI-64:            ace42e 004aa44151
Local Time is:                      Mon May 12 14:24:55 2025 SEAST
Firmware Updates (0x16):            3 Slots, no Reset required
Optional Admin Commands (0x0017):   Security Format Frmw_DL Self_Test
Optional NVM Commands (0x00df):     Comp Wr_Unc DS_Mngmt Wr_Zero Sav/Sel_Feat Timestmp Verify
Log Page Attributes (0x1e):         Cmd_Eff_Lg Ext_Get_Lg Telmtry_Lg Pers_Ev_Lg
Maximum Data Transfer Size:         64 Pages
Warning  Comp. Temp. Threshold:     83 Celsius
Critical Comp. Temp. Threshold:     85 Celsius

Supported Power States
St Op     Max   Active     Idle   RL RT WL WT  Ent_Lat  Ex_Lat
 0 +     7.50W       -        -    0  0  0  0        5       30
 1 +     3.90000W    -        -    1  1  1  1       30      330
 2 +     1.50000W    -        -    2  2  2  2      100      400
 3 -     0.05000W    -        -    3  3  3  3      500     1500
 4 -     0.00500W    -        -    4  4  4  4     1000     9000

Supported LBA Sizes (NSID 0x1)
Id Fmt  Data  Metadt  Rel_Perf
 0 +     512       0         0
 1 -    4096       0         0

=== START OF SMART DATA SECTION ===
SMART overall-health self-assessment test result: PASSED

SMART/Health Information (NVMe Log 0x02)
Critical Warning:                   0x00
Temperature:                        29 Celsius
Available Spare:                    100%
Available Spare Threshold:          50%
Percentage Used:                    0%
Data Units Read:                    4,787,659 [2.45 TB]
Data Units Written:                 4,484,998 [2.29 TB]
Host Read Commands:                 34,309,418
Host Write Commands:                28,698,344
Controller Busy Time:               606
Power Cycles:                       137
Power On Hours:                     198
Unsafe Shutdowns:                   59
Media and Data Integrity Errors:    0
Error Information Log Entries:      0
Warning  Comp. Temperature Time:    0
Critical Comp. Temperature Time:    0
Temperature Sensor 1:               29 Celsius
Temperature Sensor 2:               30 Celsius
```

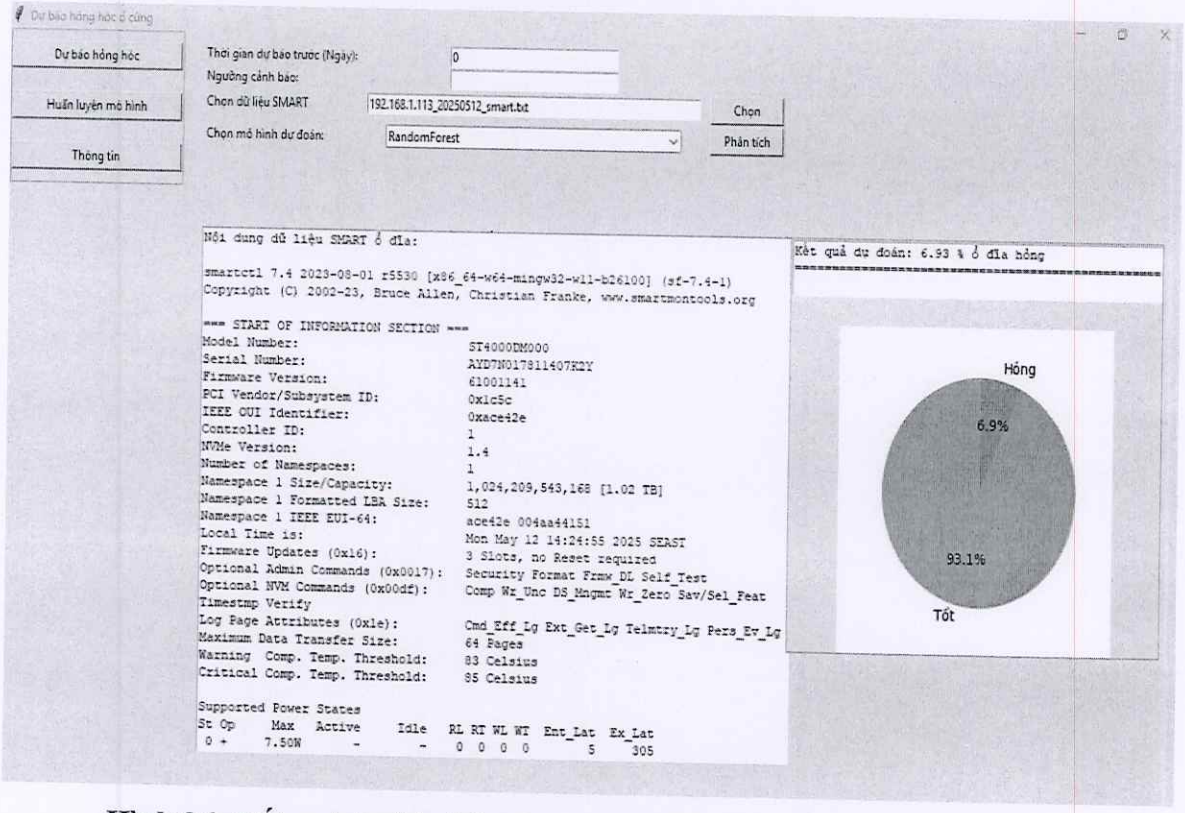
Hình 3.2. Dữ liệu SMART thu thập được tại đơn vị (Model ST4000DM000)

- Kết quả dự báo khả năng hỏng hóc với model ổ đĩa đã được huấn luyện được thể hiện ở Hình 3.2 (Mô hình sử dụng thuật toán Random Forest) và Hình 3.3 (Mô hình sử dụng thuật toán XGB). Model ổ đĩa ST4000DM000 được thu thập từ máy trạm có IP 192.168.1.113. Kết quả dự báo như sau:

Sử dụng thuật toán RF (Hình 3.2):

+ Không thiết lập dự báo khoảng thời gian (STW): Số ngày dự báo trước = 0, Ngưỡng cảnh báo: Không thiết lập.

+ Kết quả dự báo: Hệ thống đưa ra dự báo hỏng của ổ đĩa tại thời điểm thu thập là: 6,93%. (Ổ cứng có khả năng hỏng rất thấp).

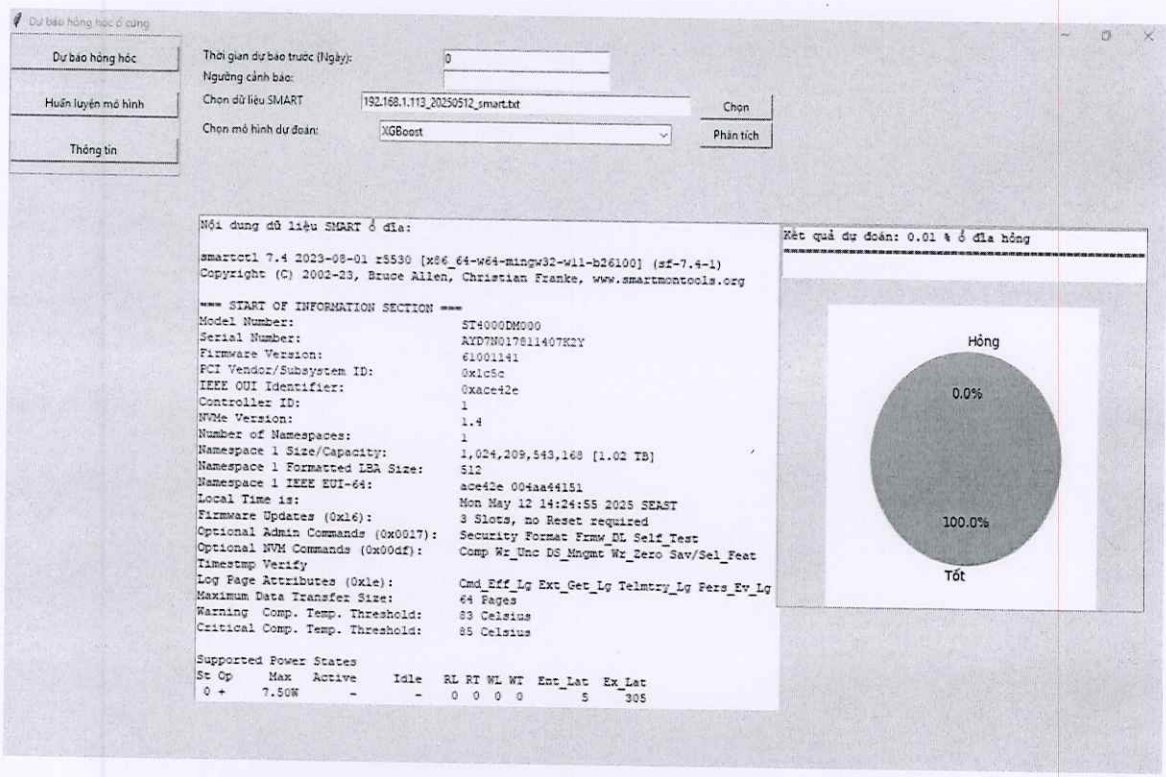


Hình 3.2. Kết quả dự báo hỏng hóc với dữ liệu thực tế với mô hình RF

Sử dụng thuật toán XGB (Hình 3.3):

+ Không thiết lập dự báo khoảng thời gian (STW): Số ngày dự báo trước = 0, Ngưỡng cảnh báo: Không thiết lập.

+ Kết quả dự báo: Hệ thống đưa ra dự báo hỏng của ổ đĩa tại thời điểm thu thập là: 0,01%. (Ổ cứng có khả năng hỏng rất thấp).



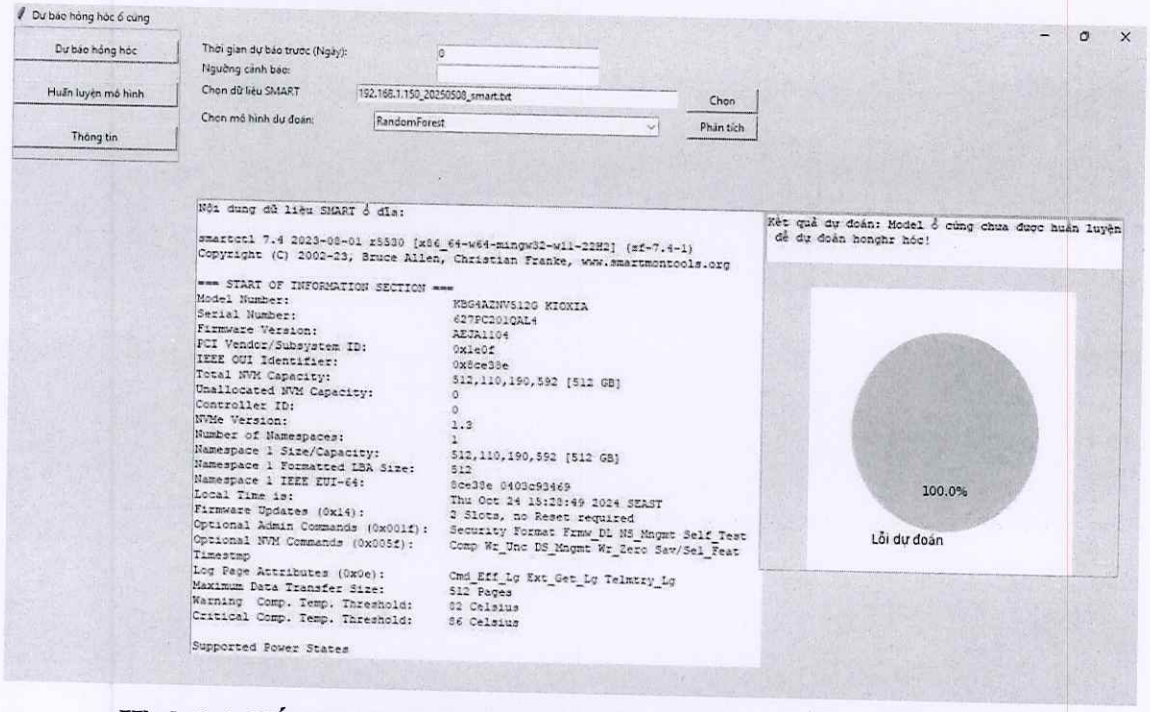
Hình 3.3. Kết quả dự báo hỏng hóc với dữ liệu thực tế với mô hình XGB

- Với các model ổ cứng chưa được huấn luyện (Model ổ đĩa KBG4AZNV512G KIOXIA, thu thập từ máy trạm có IP 192.168.1.150), với các tham số hệ thống được thiết lập như sau (Hình 3.4):

+ Sử dụng thuật toán RF hoặc XGB

- Không thiết lập dự báo khoảng thời gian (STW): Số ngày dự báo trước = 0, Ngưỡng cảnh báo: Không thiết lập.

- Kết quả dự báo (Hình 3.4): Hệ thống đưa ra cảnh báo "*Model ổ cứng chưa được huấn luyện để dự đoán hỏng hóc*".



Hình 3.4. Kết quả dự báo hỏng hóc với model chưa được huấn luyện

3.4. Kết luận chương

Nội dung chương 3 đã trình bày về việc triển khai thử nghiệm hệ thống phát hiện, dự báo hỏng hóc ổ cứng trong mạng quân sự đã xây dựng. Với việc thiết lập môi trường thử nghiệm qua một Project sử dụng các công cụ và thử viện phần mềm, đề án đã thực hiện cài đặt phần mềm và triển khai thử nghiệm hệ thống với tập dữ liệu Backblaze và tập dữ liệu do học viên thu thập thực tế tại đơn vị công tác. Thử nghiệm mô hình trên tập dữ liệu của Backblaze cho kết quả phù hợp, mô hình không bỏ sót ổ cứng lỗi (thông qua tỷ lệ recall xấp xỉ 100%), tỉ lệ cảnh báo giả ở mức chấp nhận được với việc hệ thống ưu tiên cảnh báo sớm (không bỏ sót ổ cứng hỏng). Đối với dữ liệu thực tế thu thập tại đơn vị công tác, mô hình đã cho kết quả chính xác với tỉ lệ phần trăm ổ đĩa cứng có khả năng hỏng hóc thấp đúng với thực tế. Tuy nhiên, do điều kiện về thời gian, học viên vẫn chưa thu thập được nhiều dữ liệu thực tế về các ổ đĩa cứng hỏng nên chưa đánh giá được đầy đủ mô hình trong trường hợp có dữ liệu ổ cứng bị hỏng trong thực tế. Đây cũng là một hướng phát triển tiếp của đề tài.

KẾT LUẬN

1. Kết quả đạt được của đề án

Trong bối cảnh lượng dữ liệu ngày càng nhiều được lưu trữ trên các ổ cứng trên mạng hay qua các dịch vụ lưu trữ đám mây, việc giám sát, phát hiện và dự báo lỗi ổ cứng là một nhu cầu cấp thiết và đang ngày càng được quan tâm hơn. Vấn đề này càng quan trọng đối với các mạng máy tính đóng như mạng chuyên dụng quân sự, điển hình là mạng máy tính của Binh chủng Thông tin liên lạc thuộc Bộ Quốc phòng, nơi học viên đang công tác.

Đã có các giải pháp truyền thống để giám sát và phát hiện lỗi ổ cứng như: sử dụng các công cụ tiện ích của hệ điều hành hoặc một số phần mềm giám sát hoạt động máy tính. Tuy nhiên, các giải pháp này hầu hết vẫn chi thủ công, chưa chính xác, chưa có khả năng phát hiện sớm sự cố lỗi của các đĩa cứng đang hoạt động cũng như khả năng dự báo hỏng hóc theo thời gian. Áp dụng các kỹ thuật học máy, học sâu tạo ra triển vọng mới trong phát hiện và dự báo hỏng hóc của các trang thiết bị nói chung và các đĩa cứng nói riêng, tạo ra khả năng tự động hóa và giúp nâng cao tính ổn định, độ chính xác. Tuy nhiên, vẫn chưa có một hệ thống như vậy được đưa vào ứng dụng trong thực tế, đặc biệt là đối với mạng chuyên dụng của quân sự.

Đề tài của đề án này là nghiên cứu, xây dựng một hệ thống phát hiện và dự báo lỗi ổ cứng hướng tới mục tiêu triển khai trong mạng máy tính quân sự của Binh chủng Thông tin liên lạc. Để xây dựng hệ thống, cần nghiên cứu nguồn dữ liệu lỗi ổ cứng với các thuộc tính SMART, các kỹ thuật học máy/ học sâu phù hợp cho tập dữ liệu, lựa chọn mô hình phù hợp cho mạng chuyên dụng quân sự tại đơn vị. Đây là những nội dung nghiên cứu có tính khoa học, tính cấp thiết và thực tế.

Các kết quả chính đã đạt được của đề án gồm:

- Nghiên cứu cơ sở lý thuyết về ổ cứng, cơ chế bảo vệ dữ liệu, bài toán phát hiện và dự báo lỗi ổ cứng thông qua công nghệ SMART và những vấn đề liên quan đến bài toán, cụ thể về cách thức phát hiện và dự báo lỗi ổ cứng theo kiểu truyền thống và theo hướng áp dụng kỹ thuật học máy/học sâu.

- Nghiên cứu và đề xuất một mô hình hệ thống phát hiện, dự báo hỏng hóc ổ cứng trong mạng quân sự với việc xây dựng và triển khai một hệ Agent thu thập dữ liệu SMART từ các ổ cứng, một hệ xử lý trung tâm áp dụng các thuật toán học máy gồm Random Forest và XGBoost và các kỹ thuật giúp lựa chọn tham số tối ưu như RandomizedSearchCV và kỹ thuật trợ giúp nâng cao độ chính xác như Sliding Time Window (STW) và cơ chế bỏ phiếu chọn lọc (Part-voting). Đưa ra lưu đồ triển khai dự đoán hỏng hóc ổ cứng, đánh giá mô hình qua các bộ tham số Accuracy, Precision, Recall, F1-score, ma trận nhầm lẫn.

- Triển khai thiết lập môi trường thử nghiệm, cài đặt bộ phần mềm thu thập dữ liệu SMART của Backblaze và tập dữ liệu do học viên thu thập thực tế tại đơn vị công tác, triển khai thử nghiệm hệ thống với 2 tập dữ liệu trên. Các kết quả thử nghiệm cho thấy hệ thống với tập dữ liệu của Backblaze và với dữ liệu thực tế thu thập tại đơn vị công tác, Mô hình không bỏ sót ổ cứng lỗi (thông qua tỷ lệ recall xấp xỉ 100%), Tỉ lệ cảnh báo giả ở mức chấp nhận được với việc hệ thống ưu tiên cảnh báo sớm (không bỏ sót ổ cứng hỏng).

Tập dữ liệu thu thập từ Backblaze khá lớn bảo đảm độ tin cậy của thử nghiệm. Các kết quả đạt được của hệ thống với các mô hình sử dụng Random Forest và XGBoost đều cho kết quả tốt và phù hợp với bài toán phát hiện hỏng hóc ổ đĩa cứng trong thực tế.

2. Hướng phát triển tiếp theo

Tuy nhiên trong phạm vi nghiên cứu của đề án, còn một số hạn chế về mặt thời gian thực hiện, cũng như dữ liệu thực tế thu thập được. Để đề án được hoàn thiện và có thể triển khai trong thực tế, cần hoàn thiện một số nội dung sau:

- Nghiên cứu, mở rộng mô hình sang học sâu (deep learning), đặc biệt là các mô hình RNN hoặc Transformer để xử lý chuỗi dữ liệu SMART.

- Kết hợp dữ liệu SMART với log hệ thống (log hệ điều hành, ứng dụng) để tăng khả năng dự báo chính xác hơn.

- Xây dựng hệ thống có khả năng phát triển mô hình học liên tục (online learning) phù hợp với dữ liệu thực tế cập nhật liên tục.

- Hoàn thiện giao diện người dùng: Giao diện cho phép huấn luyện mô hình, giao diện kết xuất dữ liệu dự báo theo thời gian (tuần, tháng, quỹ, năm) phục vụ quản trị hệ thống (trung tâm dữ liệu, phòng máy chủ lớn) trong việc phát hiện sớm hỏng hóc liên quan đến ổ đĩa cứng.

- Triển khai thử nghiệm mô hình trong môi trường mạng quân sự và đánh giá hiệu quả lâu dài.

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

[1] Chhetri Tek Raj, Dehury Chinmaya Kumar, Lind Artjom, Srirama Satish Narayana, Fensel Anna (2021), "A Combined Metrics Approach to Cloud Service Reliability using Artificial Intelligence," *Big Data and Cognitive Computing,* 1(0), 1–19.

[2] Document, S.M.A.R.T Atributes, [Online]. Available: http://ntfs.com/disk-monitor-smart-attributes.htm

[3] Eduardo Pinheiro, Wolf-Dietrich Weber, Luiz Andre´ Barroso (2007), "Failure Trends in a Large Disk Drive Population" *Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST'07)*, 5(1), 17–28.

[4] Gargiulo Federico, Duellmann Dirk, Arpaia Pasquale, Schiano Lo Moriello Rosario (2021), "Predicting Hard Disk Failure by Means of Automatized Labeling and Machine Learning Approach," *Applied Sciences,* 118293(11), 1–16.

[5] Hard Drive Data and Stats, [Online]. Available: https://www.backblaze.com/cloud-storage/resources/hard-drive-test-data

[6] Li Jing, Ji Xinpu, Jia Yuhan, Zhu Bingpeng, Wang Gang, Li Zhongwei, Liu Xiaoguang (2014), "Hard Drive Failure Prediction Using Classification and Regression Trees," *Proceedings of the 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks,* 44(1), 383–394.

[7] Shen Jing, Wan Jian, Lim Se-Jung, Yu Lifeng (2018), "Random-forest-based failure prediction for hard disk drives," *International Journal of Distributed Sensor Networks,* 1411(14), 1–15.

[8] Wang Han, Zhuge Qingfeng, Edwin Hsing-Mean, Xu Rui, Song Yuhong (2023), "Optimizing Efficiency of Machine Learning Based Hard Disk Failure Prediction by Two-Layer Classification-Based Feature Selection," *Applied Sciences,* 137544(13), 1–18.

[9] RAID, [Online]. Available: https://vi.wikipedia.org/wiki/RAID

[10] Wei Li , Haozhou Zhou , Srinivasan Radhakrishnan , Sagar Kamarthi (2025), "Explainable time series features for hard disk drive failure prediction",

Engineering Applications of Artificial Intelligence Volume 152, 15 July 2025, 110674.

[11] Elizabeth Atekoja (2024), "Prediction of Hard Drive Failure using Machine Learning", Global Journal of Computer Science and Technology: Neural & Artificial Intelligence Volume 24, Issue 1, 2024.

[12] Ulinktech (2024), "How Predictive Maintenance Works for Drives", Mar 4, 2024. https://ulinktech.com/how-predictive-maintenance-works-for-drives/

[13] Anantharaman, P., Qiao, M., Jadav, D., (2018), "Large scale predictive analytics for hard disk remaining useful life estimation", IEEE International Congress on Big Data (BigData Congress), 2018, pp. 251–254. http://dx.doi.org/10.1109/BigDataCongress.2018.00044

[14] Shen, J., Wan, J., Lim, S.-J., Yu, L., (2018), "Random-forest-based failure prediction for hard disk drives", Int. J. Distrib. Sens. Netw. 14 (11), 15501477 18806480. http://dx.doi.org/10.1177/1550147718806480

[15] Murray, J.F., Hughes, G.F. and Kreutz-Delgado, K. (2018), "Machine Learning Methods for Predicting Failures in Hard Drives: A Multiple-Instance Application", Journal of Machine Learning Research, 6, 783–816.

[16] Züfle, M. et al. (2020), "To Fail or Not to Fail: Predicting Hard Disk Drive Failure Time Windows", Measurement, Modeling and Evaluation of Computing Systems (MMB 2020). Lecture Notes in Computer Science 12040. Springer, Cham. doi: 10.1007/978-3-030-43024-5_2

BẢN CAM ĐOAN

Tôi xin cam đoan đã thực hiện kiểm tra mức độ tương đồng nội dung đề án qua phần mềm kiểm tra tài liệu một cách trung thực và đạt kết quả mức độ tương đồng 7% toàn bộ nội dung đề án tốt nghiệp. Bản đề án tốt nghiệp kiểm tra qua phần mềm là bản cứng đề án tốt nghiệp đã nộp để bảo vệ trước hội đồng. Nếu sai tôi xin chịu các hình thức kỷ luật theo quy định hiện hành của Học viện.

Hà Nội, ngày 29 tháng 7 năm 2025

HỌC VIÊN CAO HỌC

(Ký và ghi rõ họ tên)

Phan Văn Phùng

KiểmTraTàiLiệu

BÁO CÁO KIỂM TRA TRÙNG LẶP

Thông tin tài liệu

Tên tài liệu:	Xây dựng hệ thống phát hiện lỗi ổ cứng trong mạng máy tính quân sự của Binh chủng Thông tin liên lạc
Tác giả:	Phan Văn Phùng
Điểm trùng lặp:	7
Thời gian tải lên:	22:15 28/07/2025
Thời gian sinh báo cáo:	22:17 28/07/2025
Các trang kiểm tra:	76/76 trang

Kết quả kiểm tra trùng lặp

Có 7% nội dung trùng lặp

Có 93% nội dung không trùng lặp

Có 0% nội dung người dùng loại trừ

Có 0% nội dung hệ thống bỏ qua

Nguồn trùng lặp tiêu biểu

123docz.net www.mdpi.com ieeexplore.ieee.org

Học viên
(Ký và ghi rõ họ tên)

Phan Văn Phùng

Người hướng dẫn khoa học
(Ký và ghi rõ họ tên)

PGS.TSKH. Hoàng Đăng Hải

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BÁO CÁO GIẢI TRÌNH
SỬA CHỮA, HOÀN THIỆN ĐỀ ÁN TỐT NGHIỆP

Họ và tên học viên: Phan Văn Phùng

Chuyên ngành: CNTT

Khóa: 2023 đợt 2

Tên đề tài: Xây dựng hệ thống phát hiện lỗi ổ cứng trong mạng máy tính quân sự của Binh chủng Thông tin liên lạc

Người hướng dẫn khoa học: PGS.TSKH. Hoàng Đăng Hải

Ngày bảo vệ: 19/07/2025

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

TT	Ý kiến hội đồng	Sửa chữa của học viên
1	Cần làm rõ sự cấp thiết đối với mạng quân sự	Học viên đã rà soát, bổ sung nội dung tại mục 1.5. Khái quát về mạng máy tính quân sự của Binh chủng Thông tin liên lạc, Bộ Quốc phòng và nhu cầu phát hiện, dự báo lỗi ổ cứng (trang 23, 24).
2	Chuẩn hóa tài liệu	Tiếp thu rà soát, chuẩn hóa toàn bộ Đề án theo quy định
3	Bổ sung các độ đo phù hợp thay cho Accurary	Học viên đã tiếp thu, chỉnh sửa vào nội dung tại các mục: - 3.2.2. Kết quả thử nghiệm với tập dữ liệu Backblaze: Bổ sung Bảng 3.4. Kết quả đánh giá mô hình dựa trên Precision, Recall, F1-score (trang 49) - 3.4. Kết luận chương (trang 53) - Phần Kết luận, Mục 1. Kết quả đạt được của đề án (trang 54, 55)

Hà Nội, ngày tháng năm 2025

Ký xác nhận của

CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN	THƯ KÝ HỘI ĐỒNG	NGƯỜI HƯỚNG DẪN KHOA HỌC	HỌC VIÊN
PGS.TS.Trần Quang Anh	TS. Đào Thị Thúy Quỳnh	PGS.TSKH. Hoàng Đăng Hải	Phan Văn Phùng

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ
BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BIÊN BẢN
HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 1098/QĐ-HV ngày 26 tháng 06 năm 2025 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi..9....giờ....10phút, ngày 19 tháng 07 năm 2025 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Phan Văn Phùng**

Tên đề án tốt nghiệp: **Xây dựng hệ thống phát hiện lỗi ổ cứng trong mạng máy tính quân sự của Binh chủng Thông tin liên lạc**

Chuyên ngành: **Hệ thống thông tin** Mã số: **8480104**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: .04../ 05

TT	HỌ VÀ TÊN	TRÁCH NHIỆM TRONG HĐ	GHI CHÚ
1	**PGS.TS. Trần Quang Anh**	Chủ tịch	
2	**TS. Đào Thị Thúy Quỳnh**	Thư ký	
3	**TS. Trần Đăng Công**	Phản biện 1	
4	**PGS.TS. Nguyễn Hà Nam**	Phản biện 2	
5	**PGS.TS. Nguyễn Trọng Khánh**	Uỷ viên	

Các nội dung thực hiện:

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên. (có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)
6. Các câu hỏi của thành viên Hội đồng:

- Thời gian dự báo hoàn toàn để nghiên cứu của đề tài có được giải thích rõ hơn?

- Làm rõ mô hình em xử dụng phù hợp với bài toán của em như thế?

7. Trả lời của học viên:

- Đề tài luận văn chưa xét tới "timing" thực hiện dự án

8. Thư ký đọc nhận xét về quá trình thực hiện đề án tốt nghiệp của học viên *(có văn bản kèm theo).*

9. Hội đồng họp riêng:

- Bầu Ban kiểm phiếu:
  1. Trưởng Ban kiểm phiếu: TS. Đào Thị Thúy Quỳnh
  2. Uỷ viên Ban kiểm phiếu: PGS.TS. Nguyễn Hà Nam
  3. Uỷ viên Ban kiểm phiếu: TS. Trần Thị Công
- Hội đồng chấm đề án tốt nghiệp bằng bỏ phiếu kín.
- Ban kiểm phiếu làm việc:
- Trưởng Ban kiểm phiếu báo cáo kết quả kiểm phiếu *(có Biên bản họp Ban kiểm phiếu kèm theo)*
- **Điểm trung bình của đề án tốt nghiệp:** 8.6

Kết luận:

1. Các nội dung cần chỉnh sửa, hoàn thiện sau bảo vệ đề án tốt nghiệp:

- Cần làm rõ sự cấp thiết đề tài mang ý nghĩa gì
- Chuẩn hóa tài liệu
- Đưng các đề đã phù hợp thay cho Accuracy

2. Đề nghị Học viện công nhận (hoặc không) và cấp bằng (hoặc không) thạc sĩ cho học viên:

3. Đề án tốt nghiệp có thể phát triển thành đề tài nghiên cứu cho NCS.....

Buổi làm việc kết thúc vào..... cùng ngày.

Chủ tịch	**Thư ký**
PGS.TS. Trần Quang Anh	TS. Đào Thị Thúy Quỳnh

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài đề án tốt nghiệp: **Xây dựng hệ thống phát hiện lỗi ổ cứng trong trạm máy tính quân sự của binh chủng Thông tin liên lạc**

Chuyên ngành: Hệ thống thông tin

Mã chuyên ngành: 8.48.01.04

Họ và tên học viên: **Phan Văn Phùng**

Họ và tên người nhận xét: Trần Đăng Công

Học hàm, học vị: Tiến sĩ

Chuyên ngành: Khoa học máy tính

Cơ quan công tác: Đại học Đại Nam

Số điện thoại: 0964981451 E-mail: congtd@dainam.edu.vn

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

- Đề tài có tính thực tiễn cao, tác giả đã trình bày được từ cơ sở lý thuyết đến kết quả triển khai bằng mô hình học máy.

- Tác giả đã phân tích các vấn đề liên quan đến hỏng ổ cứng, từ đó sử dụng các loại thông tin và cách thức thu thập.

- Tác giả đã trình bày được phương pháp đánh giá và triển khai giải pháp cho việc phát hiện lỗi ổ cứng.

.....

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

- Tác giả đã trình bày được các bước từ phương pháp thu thập thông tin ổ cứng trên toàn hệ thống đến việc xử lý tạo trung tâm.

- Tác giả đã thử nghiệm với các Agent tại các máy nút trên mạng và triển khai phân tích, phân loại tại trung tâm, từ đó đưa là dự báo.

- Báo cáo gồm 3 chương, 57 trang.

.....

.....

III/ Những vấn đề cần giải thích thêm:

- Hãy đánh giá ảnh hưởng về tốc độ, và bảo mật việc triển khai các Agent thu thập và gửi dữ liệu từ các máy nút về máy chủ trung tâm khi sử dụng phương pháp Sahred Foder tại máy chủ (mục 2.2)?

.....

.....

IV/ Kết luận:

Đề án tốt nghiệp đã được giải quyết với kết quả ứng dụng trong thực tế, phù hợp với điều kiện thực tiễn.

Tuy nhiên, việc phân tích, so sánh kết quả còn chưa kỹ càng.

Đồng ý cho phép học viên bảo về đề án tốt nghiệp.

Ngày 15 tháng 7 năm 2025

NGƯỜI NHẬN XÉT

TS. Trần Đăng Công

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BẢN NHẬN XÉT LUẬN VĂN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài luận văn: Xây dựng hệ thống phát hiện lỗi ổ cứng trong mạng máy tính quân sự của binh chủng thông tin liên lạc
Chuyên ngành: Khoa học máy tính Mã số:8.48.01.01
Tên học viên: Phan Văn Phùng
Họ và tên người nhận xét: Nguyễn Hà Nam.....
Học hàm, học vị: PGS. TS Chuyên ngành: CNTT .....
Cơ quan công tác: Ban Khoa học và Đổi mới sáng tạo, ĐHQGHN.....

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:
Trong hệ thống máy tính, HDD luôn là thành phần rất quan trọng lưu trữ toàn bộ thông tin, dữ liệu đảm bảo vận hành hệ thống. Việc phát hiện sớm lỗi ổ cứng là rất cấp thiết để **giảm thiểu rủi ro gián đoạn thông tin liên lạc, bảo vệ dữ liệu mật và đảm bảo an toàn vận hành**. Về mặt khoa học, đề tài phù hợp với các hướng nghiên cứu về **bảo trì dự đoán (Predictive Maintenance), phát hiện bất thường (Anomaly Detection)**. Đây là một hương nghiên cứu có tính thời sự và thực tiễn cao.

II/ Về nội dung, chất lượng của luận văn, các kết quả đã đạt được (so với đề cương đã được duyệt):
Nội dung của luận văn và các kết quả đạt được cơ bản bám sát theo đề cương đã được phê duyệt.
Luận văn được trình bày trong 3 chương bao gồm giới thiệu tổng quan, nghiên cứu xây dựng hệ thống phát hiện dự báo lỗi HDD với mô hình học máy và cuối cùng thử nghiệm đánh giá mô hinh trên tập dữ liệu BacklBlaze và đã cho ra một số kế quả tốt.

III/ Những vấn đề cần giải thích thêm:
- Bài toán thuộc nhóm bài toán phát hiện bất thường (được nhiều nhóm nghiên cứu quan tâm) tuy nhiên trong luận văn chưa phân tích rõ điểm mạnh, yếu cũng như kết quả đạt được của các phương pháp này, sự khác biệt của đề xuất là gì?
- Các hãng có nhiều giải pháp chẩn đoán khả năng hỏng hóc và quản lý và dung lỗi HDD, giải pháp của tác giả sẽ phù hợp trong tình huống nào?
- Thời gian dữ báo hoặc mức độ nghiêm trọng của dự báo cần được mô tả rõ hơn.

IV/ Kết luận:
Luận văn đáp ứng được yêu cầu cơ bản của luận văn thạc sỹ chuyên ngành KHMT(theo định hướng ứng dụng). Tôi đồng ý để học viên được bảo vệ luận văn trước Hội đồng chấm luận văn thạc sĩ

Ngày 15 tháng 7 năm 2025
NGƯỜI NHẬN XÉT
(Ký và ghi rõ họ tên)

Nguyễn Hà Nam