XÂY DỰNG MÔ ĐUN ĐỒNG BỘ HÓA DỮ LIỆU HỆ THỐNG PHỤC VỤ GIÁM SÁT VÀ CẢNH BÁO SỚM SỰ CỐ TRÊN ĐƯỜNG DÂY TẢI ĐIỆN 110 kV

BUILDING SYSTEM DATA SYNCHRONIZATION MODULE FOR FAULT MONITORING AND EARLY WARNING ON 110 kV POWER LINE

Vũ Thị Thu Nga, Nguyễn Thị Thanh Tân*

Trường Đại học Điện lực

Ngày nhận bài: 22/08/2023, Ngày chấp nhận đăng: 29/12/2023, Phản biện: PGS.TS. Trịnh Trọng Chưởng

Tóm tắt:

Trong quá trình làm việc của hệ thống truyền tải điện, các sự cố trên đường dây thường xảy ra tương đối phức tạp do nhiều nguyên nhân khác nhau. Để có thể thực hiện việc giám sát vận hành, cảnh báo sớm sự cố dễ dàng trên đường dây ở các địa hình khác nhau, công nghệ thông tin, tín hiệu số, công nghệ trí tuệ nhân tạo (AI), sử dụng máy bay không người lái (UAV) đã được áp dụng và ngày càng phổ biến trên thế giới trong thời gian gần đây. Do cơ sở dữ liệu giám sát hệ thống theo thời gian thực thu được không đồng nhất từ nhiều hình thức khác nhau nên để một lược đồ duy nhất và có thể truy vấn, cung cấp cho người dùng một cái nhìn thống nhất về chúng là không dễ dàng. Như vậy, quá trình tích hợp và đồng bộ hóa dữ liệu đường dây từ các nguồn khác nhau là vô cùng cần thiết nhằm dự báo chính xác loại sự cố, vị trí sự cố và từ đó đưa ra được phương án xử lý kịp thời. Trong nghiên cứu này, nhóm nghiên cứu đã thực hiện thiết kế, lập trình xây dựng module đồng bộ hóa dữ liệu cho hệ thống từ các nguồn dữ liệu khác nhau của đường dây truyền tải 110 kV phục vụ cho quá trình giám sát và cảnh báo sớm sự cố trên đường dây truyền tải điện 110 kV.

Từ khóa:

Tích hợp và đồng bộ hóa dữ liệu, cảnh báo sớm sự cố, AI, giám sát hệ thống.

Abstract:

During the operation of the power transmission system, the faults on the line still often occur relatively complicated due to many different reasons. To be able to easily perform operational monitoring and early warning of problems on lines in different terrains, information technology, digital signals, artificial intelligence (AI) technology, using unmanned aircraft have been applied and increasingly popular in the world during recent times. Because real-time system monitoring databases are obtained heterogeneously from many different forms, leaving a single and queryable schema that provides users with a unified view of them is not easy. So, the process of synchronizing data on the line is extremely necessary to accurately forecast the type of fault, the location of the fault, and from there to come up with a timely treatment plan. In this study, the research team designed, programmed, and built a data synchronization module for the system from different data sources of the 110 kV line to serve the monitoring and early fault warning on 110 kV power transmission line.

Keywords:

Data integration and synchronization, early failure warning, AI, system monitoring.

1. GIỚI THIỆU

Vận hành hệ thống điện (HTĐ) là tập hợp các thao tác nhằm duy trì chế độ làm việc bình thường của hệ thống điện đáp ứng các yêu cầu chất lượng, tin cậy và kinh tế. Hệ thống điện bao gồm các phần tử có mối liên hệ chặt chẽ với nhau, sự làm việc tin cậy và kinh tế của hệ thống xuất phát từ sự tin cậy và chế độ làm việc kinh tế của từng phần tử. Cùng với sự ra đời của các thiết bị công nghệ mới, những yêu cầu về vận hành hệ thống điện nói chung và các thiết bị điện nói riêng ngày càng trở nên nghiêm ngặt nhằm đảm bảo hiệu quả kinh tế cao và đảm bảo chất lượng điện, độ tin cậy cung cấp điện liên tục, tính linh hoạt và đáp ứng đồ thị phụ tải [1]. Trên thực tế quá trình vận hành HTĐ trên thế giới và Việt Nam cho thấy, mặc dù chế độ vận hành HTĐ được tính toán và phân tích kỹ lưỡng trong quá trình lập quy hoạch, báo cáo khả thi, thiết kế kỹ thuật, lập kế hoạch và xây dựng cho phương thức vận hành hệ thống điện, các sự cố về đường dây và trạm vẫn xảy ra, thậm chí tương đối phức tạp có tính chất ngày càng gia tăng theo sự phát triển của hệ thống [2], [3], [4]. Do vậy, các giải pháp giám sát tự động, cảnh báo sớm sự cố trên hệ thống điện không tốn kém về nhân lực, có thể thực hiện dễ dàng ở các địa hình khác nhau, ứng dụng công nghệ thông tin, tín hiệu số, công nghệ AI, sử dụng máy bay không người lái là thực sự cần thiết, đã và đang được nghiên cứu phát triển trên thế giới và ở Việt Nam [5], [6], [7], [8].

Từ các kết quả khảo sát cũng như kinh nghiệm triển khai thực tế, nghiên cứu xác định để đảm bảo được độ tin cậy cho các thuật toán ứng dụng trí tuệ nhân tạo và xử lý dữ liệu lớn phục vụ giám sát và cảnh báo sớm sự cố cần phải kết hợp dữ liệu từ nhiều nguồn khác nhau, do vậy quá trình tích hợp và đồng bộ hóa dữ liệu từ các nguồn khác nhau là vô cùng quan trọng để kết hợp dữ liệu không đồng nhất trong các nguồn khác nhau vào một lược đồ duy nhất và có thể truy vấn, cung cấp cho người dùng một cái nhìn thống nhất về chúng [9], [10], [11]. Nghiên cứu này tập trung vào thiết kế và lập trình xây dựng module đồng bộ hóa dữ liệu cho hệ thống được trình bày chi tiết ở các phần sau của bài báo.

2. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP THỰC HIỆN

2.1. Đối tượng nghiên cứu

Sự cố có thể xuất hiện trên bất kỳ phần tử nào trong hệ thống điện, tuy nhiên theo thống kê xác suất xảy ra sự cố nhiều nhất là trên các hệ thống truyền tải điện do các nguyên nhân như vi phạm hành lang an toàn lưới điện, do thiên tai hoặc do chuyên môn nghiệp vụ. Do vậy, cảnh báo sớm sự cố trên đường dây sẽ giảm thiểu được xác suất xảy ra sự cố và gián đoạn cung cấp điện trên hệ thống.

Trên hệ thống truyền tải điện nói chung và hệ thống đường dây 110 kV nói riêng, các loại sự cố, hỏng hóc xảy ra trên các thiết bị rất đa dạng, tập trung ở các nhóm

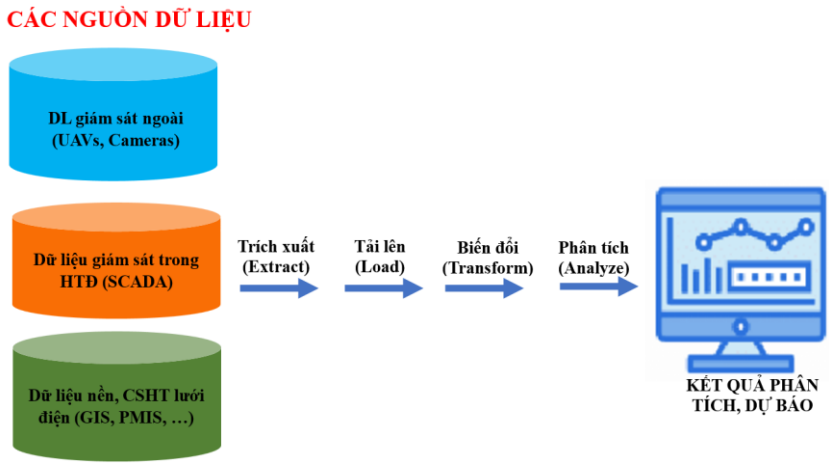
thiết bị thường gặp liên quan dây dẫn và cách điện, dây chống sét, dây cáp quang, các điều kiện cấu trúc của cột. Với mục đích giám sát và cảnh báo sớm các hư hỏng trên hệ thống truyền tải, nghiên cứu này tập trung vào hệ thống đường dây truyền tải 110 kV, cụ thể là lưới điện 110 kV thuộc Tổng công ty Điện lực Hà Nội quản lý.

2.2. Phương pháp thực hiện

Để xây dựng hệ thống giám sát tình trạng vận hành và cảnh báo sớm sự cố của các thiết bị và đường dây tải điện 110 kV cần phải kết hợp dữ liệu từ nhiều nguồn khác nhau như dữ liệu bên trong hệ thống từ hệ thống giám sát SCADA, dữ liệu cơ sở hạ tầng lưới điện từ GIS, PMIS, dữ liệu hình ảnh giám sát bên ngoài thu nhận được từ UAV. Do vậy, để đảm bảo độ chính xác cho các thuật toán AI cần phải tích hợp và đồng bộ hóa và gán nhãn dữ liệu cho hệ thống.

Tích hợp và đồng bộ hóa dữ liệu được sử dụng với tần suất ngày càng nhiều khi mà khối lượng và nhu cầu chia sẻ dữ liệu hiện nay rất lớn. Có nhiều hướng tiếp cận tích hợp và đồng bộ hóa dữ liệu như: tích hợp và đồng bộ hóa dữ liệu thủ công, tích hợp và đồng bộ hóa dữ liệu phần mềm trung gian, tích hợp và đồng bộ hóa dữ liệu dựa trên ứng dụng, tích hợp và đồng bộ hóa dựa trên truy cập thống nhất, tích hợp và đồng bộ hóa dữ liệu lưu trữ chung (kho dữ liệu, hồ dữ liệu), trong đó hướng tiếp cận tích hợp và đồng bộ hóa dữ liệu lưu trữ chung hiện được đánh giá là hướng tiếp cận phức tạp nhất. Tuy nhiên đây lại là hướng tiếp cận hiện đại và hiệu quả nhất, rất phù hợp để giải quyết bài toán dữ liệu lớn (big data) [12]. Trên cơ sở đó, nhóm đã lựa chọn hướng tiếp cận này để đề xuất phương pháp tích hợp và đồng bộ hóa dữ liệu đa nguồn của hệ thống. Cụ thể, để tích hợp và đồng bộ hóa dữ liệu, nhóm đề xuất xây dựng các đường ống dữ liệu (data pipline) dựa trên các mô hình ELT (Extract – Load – Transform) [13], [14].

ELT là một công nghệ tương đối mới, được tạo ra nhờ các công nghệ máy chủ hiện đại, dựa trên đám mây. Kho dữ liệu dựa trên đám mây cung cấp khả năng lưu trữ gần như vô tận và khả năng xử lý có thể mở rộng. Ví dụ: các nền tảng như Amazon Redshift và Google BigQuery làm cho các đường ống ELT trở nên khả thi chính nhờ vào khả năng xử lý đáng kinh ngạc của chúng. ELT được ghép nối với một hồ dữ liệu, cho phép ngay lập tức nhập một nhóm dữ liệu thô có quy mô ngày càng mở rộng. ELT không yêu cầu chuyển đổi dữ liệu thành một định dạng đặc biệt trước khi lưu nó vào hồ dữ liệu. Ưu điểm chính của ELT là liên quan đến tính linh hoạt và dễ dàng lưu trữ dữ liệu mới, không có cấu trúc. Với ELT, hệ thống có thể lưu bất kỳ loại thông tin nào, ngay cả khi không có thời gian hoặc khả năng để chuyển đổi và cấu trúc thông tin đó trước. Hơn nữa, không nhất thiết phải phát triển các quy trình ETL phức tạp trước khi nhập dữ liệu và tiết kiệm thời gian cho các nhà phát triển và nhà phân tích AI khi xử lý thông tin mới.



Hình 1. Mô hình tích hợp, đồng bộ hóa dữ liệu

3. THIẾT KẾ, XÂY DỰNG MODULE ĐỒNG BỘ HÓA DỮ LIỆU CHO HỆ THỐNG

3.1. Mô hình tích hợp và đồng bộ hóa dữ liệu

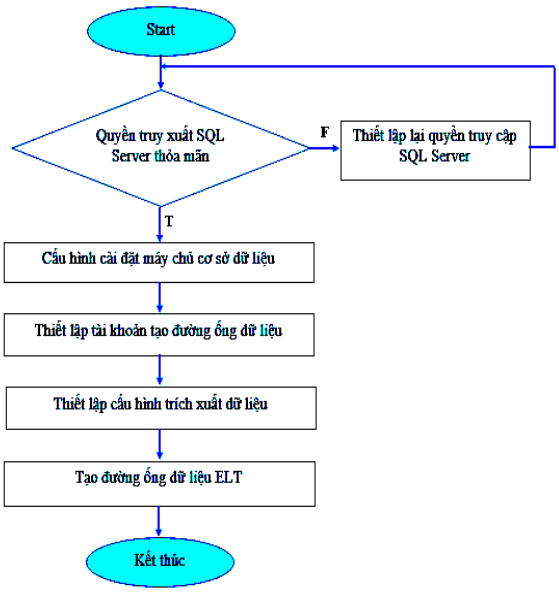
Từ các dữ liệu đa nguồn khác nhau, dựa trên việc tạo đường ống dữ liệu theo mô hình ELT, mô hình tích hợp và đồng bộ hóa dữ liệu được thể hiện trên Hình 1.

Dữ liệu giám sát, bao gồm: Dữ liệu video/hình ảnh thu nhận được từ các thiết bị camera (quang, nhiệt) gắn trên UAV và một số camera cố định gắn trên cột (tại một số vị trí trọng yếu), dữ liệu định vị GPS của UAV; thông số môi trường (nhiệt độ, độ ẩm/sương mù, cấp gió,…) và các tham số về độ cao dây dẫn, khoảng cách giữa các pha và các cột điện. Toàn bộ dữ liệu này thuộc dạng dữ liệu phi cấu trúc và được quản trị bằng nhóm lưu trữ dưới dạng cơ sở dữ liệu và quản trị bằng hệ quản trị cơ sở dữ liệu MongoDB. Dữ liệu giám sát bên trong hệ thống điện (dữ liệu từ SCADA) thường được lưu trữ dạng file ".csv" hoặc ".xlsx", dữ liệu cơ sở hạ tầng lưới điện được kế thừa từ các hệ thống PMIS hoặc GIS của ngành điện, thường được lưu trữ dạng cơ sở dữ liệu SQL hoặc NoSQL. Theo nguyên tắc tạo đường ống dữ liệu ELT, mô hình tích hợp và đồng bộ hóa dữ liệu bắt đầu từ công đoạn trích xuất dữ liệu từ các cơ sở dữ liệu MongoDB cho dữ liệu giám sát ngoài UAV, từ các file dữ liệu flat (.csv, .xlsx) cho dữ liệu SCADA, từ cơ sở dữ liệu SQL cho dữ liệu cơ sở hạ tầng lưới điện (PMIS, GIS), sau khi trích xuất, dữ liệu được tải lên kho dữ liệu/hồ dữ liệu để sử dụng cho quá trình học máy.

3.2. Module tích hợp đồng bộ hóa dữ liệu

Mô hình tích hợp dữ liệu trên Hình 1 được cụ thể hóa trong môi trường ngôn ngữ lập trình python, đã tích hợp các hệ quản trị SQL (MySQL), SQL Server, MongoDB, Singer.io và Pipelinewise. Để thực hiện bước trích xuất dữ liệu từ các hệ quản trị cơ sở dữ liệu đa nguồn, nghiên cứu đã xây dựng 03 module tương ứng với 03 cơ sở dữ liệu. Cụ thể, module

SQL_Extract(.) được xây dựng để trích xuất dữ liệu từ hệ quản trị cơ sở dữ liệu SQL, CSV_Extract(.) để trích xuất dữ liệu dạng file ".csv", MongoDB_Extract(.) để trích xuất dữ liệu từ hệ quản trị cơ sở dữ liệu MongoDB. Quy trình thực hiện của Module SQL_Extract(.) được tiến hành như trên Hình 2.



Hình 2. Các bước thực hiện của module SQL_Extract(.)

Bước đầu tiên cần kiểm tra các thông tin đăng nhập cần thiết để sao chép dữ liệu từ SQL Server đã đầy đủ và thỏa mãn hay chưa? Nếu chưa, cần thiết lập lại bằng cách xác định đầy đủ giá trị cho các thuộc tính xác định quyền truy xuất:

- CREATE USER hoặc INSERT đặc quyền (đối với cơ sở dữ liệu MySQL), đặc quyền được yêu cầu để tạo người dùng cơ sở dữ liệu cho PipelineWise.CREATE USER.
- GRANT OPTION đặc quyền trong SQL (MySQL), đặc quyền được yêu cầu để cấp các đặc quyền cần thiết cho người dùng cơ sở dữ liệu PipelineWise.GRANT OPTION.
- SUPER đặc quyền trong SQL, nếu sử dụng dựa trên nhật ký, SUPER đặc quyền được yêu cầu để xác định cài đặt máy chủ thích hợp. Thông tin xác thực kết nối cơ sở dữ liệu (sẽ cung cấp cho chi tiết kết nối trong lần nhấn) cũng cần có quyền truy cập vào INFORMATION_SCHEMA của cơ sở dữ liệu SQL để lấy siêu dữ liệu của các bảng (như chi tiết khóa chính, chi tiết kiểu dữ liệu cột,…).

Việc thiết lập tài khoản tạo đường ống dữ liệu bao gồm thiết lập quyền lựa chọn (SELECT) trên cơ sở dữ liệu và mọi bảng cần sao chép, thiết lập các quyền thay thế (REPLICATION CLIENT, REPLICATION SLAVE) nếu sử dụng Log_Based. Các tham số cấu hình cho việc trích xuất dữ liệu từ SQL được thiết lập cụ thể như sau:

```
id: "SQL_Extract"
name: "Project MySQL Database"
type: "tap-mysql"
owner: "*********"
send_alert: False
slack_alert_channel: "#tap-channel"
db_conn:
  host: "<HOST>"
  port: 3306
  user: "<USER>"
  password: "<PASSWORD>"
  dbname: "<DB_NAME>"
  use_gtid: <boolean>
```

```
  engine: "mariadb/mysql"
  filter_dbs: "power_line_schema
fastsync_parallelism: <int>
# Optional: size of multiprocessing pool
target: "snowflake"
# ID of the target connector where the data will be loaded
batch_size_rows: 20000
# Batch size for the stream to optimise load performance
stream_buffer_size: 0
# In-memory buffer size (MB) between taps
schemas:
  - source_schema: "powerline_db"
# Source schema (aka. database) in MySQL/ MariaDB with tables
    target_schema: "repl_ powerline_db "
# Target schema in the destination Data Warehouse
    target_schema_select_permissions:
# Optional: Grant SELECT on schema and tables that created
      - grp_stats
    tables:
      - table_name: "powerline_db "
        replication_method:
"INCREMENTAL"
 # One of INCREMENTAL, LOG_BASED and FULL_TABLE
        replication_key: "last_update"
      - table_name: "tranmission_two"
            replication_method:
"LOG_BASED"
```

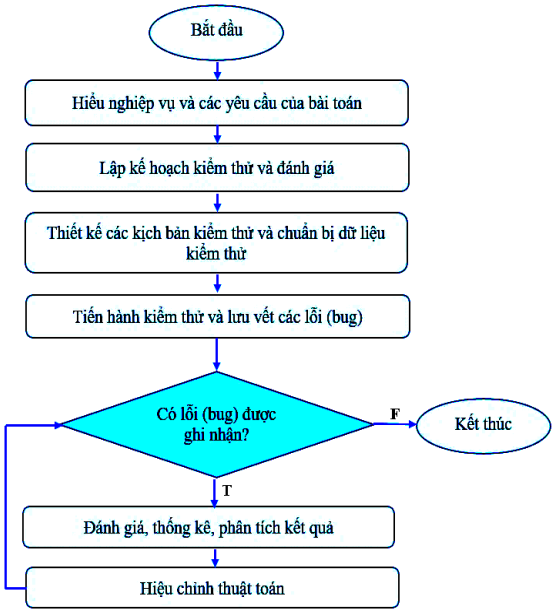
Tương tự như việc kết nối và trích xuất dữ liệu từ SQL, để trích xuất dữ liệu phi cấu trúc từ MongoDB, trước tiên cũng cần kiểm tra xem quyền đăng nhập để kết nối và sao chép dữ liệu từ MongoDB đã được thiết lập đúng chưa? Nếu chưa thiết lập hoặc thiết lập chưa đúng thì cần thực hiện thiết lập lại các quyền này. Có thể sử dụng các giá trị thuộc tính sau để thiết lập: *read* (chỉ đọc), *readWrite* (đọc và ghi), *readAnyDatabase* (đọc bất kỳ cơ sở dữ liệu nào), *readWriteAnyDatabase* (đọc và ghi bất kỳ cơ sở dữ liệu nào), *dbOwner* (là người tạo ra CSDL), *backup* (sao lưu), *root* (dữ liệu gốc), *find* & *changeStream* (tìm và thay đổi trên các luồng dữ liệu được tải). Sau khi thiết lập các quyền truy xuất dữ liệu, bước tiếp theo cần cài đặt máy chủ cơ dữ liệu bắt buộc. Sau đó tiến hành thiết lập quyền tạo đường ống dữ liệu. Cuối cùng, tiến hành và hoàn thiện các công việc trích xuất (sao chép), tải dữ liệu và các biến đổi trong quy trình ELT.

4. KIỂM THỬ THUẬT TOÁN ĐỒNG BỘ HÓA DỮ LIỆU

4.1. Quy trình kiểm thử

Để tích hợp và đồng bộ hóa dữ liệu đa nguồn trong môi trường dữ liệu lớn đạt được độ tin cậy và chính xác, nghiên cứu đã tập trung kiểm thử tính đúng đắn của module phần mềm tạo đường ống dữ liệu (kiểm thử đơn vị), các cơ sở dữ liệu dữ liệu và đường ống dữ liệu (data pipline). phương pháp tích hợp và đồng bộ hóa dữ liệu theo hướng tiếp cận tạo đường ống dữ liệu ETL và ELT hiện đang là hướng tiếp cận hiện đại, đảm bảo độ tin cậy và hiệu quả cao cho các bài toán phân tích dữ liệu [15]. Việc kiểm thử đường ống dữ

liệu không chỉ kiểm thử được độ chính xác của các thuật toán thực thi trên dữ liệu như tích hợp, đồng bộ và biến đổi mà còn kiểm thử được các điều kiện ràng buộc dữ liệu, lược đồ dữ liệu. Việc kiểm thử đường ống dữ liệu ELT và các thuật toán biến đổi cũng như đồng bộ hóa dữ liệu trong đó thường được tiến hành theo một quy trình các bước thực hiện một cách rõ ràng, cụ thể (Hình 3).



Hình 3. Quy trình kiểm thử đường ống dữ liệu

Toàn bộ các lỗi (bug) phát sinh trong quá trình kiểm thử sẽ được ghi nhận và lưu vết. Trên cơ sở đó, nghiên cứu sẽ tiến hành thống kê, phân tích và đánh giá nguyên nhân và đưa ra phương án cải tiến, hiệu chỉnh thuật toán. Quá trình kiểm thử được thực hiện lặp đi lặp lại trong suốt quá trình phát triển hệ thống **Error! Reference source not found.**.

4.2. Kịch bản kiểm thử

Để kiểm thử quá trình tích hợp và đồng bộ dữ liệu, nghiên cứu tập trung vào các kịch bản kiểm thử sau đây:

- Kịch bản xác nhận tài liệu ánh xạ: Nhằm xác minh tài liệu ánh xạ xem liệu thông tin ETL tương ứng có được cung cấp hay không. Thay đổi Log nên duy trì trong mọi tài liệu ánh xạ.
- Kịch bản xác nhận dữ liệu: nhằm xác nhận cấu trúc bảng nguồn và đích dựa trên tài liệu ánh xạ có tương ứng không; loại dữ liệu nguồn và dữ liệu đích có giống nhau không; độ dài của các loại dữ liệu nguồn và đích bằng nhau; các loại và định dạng trường dữ liệu được chỉ định; độ dài loại dữ liệu nguồn không nhỏ hơn độ dài loại dữ liệu đích; tên của các cột trong bảng dựa trên tài liệu ánh xạ đã được xác thực không.
- Kịch bản xác nhận ràng buộc: đảm bảo các ràng buộc được xác định cho bảng cụ thể có đúng với mục đích và kết quả có như mong đợi hay không.
- Kịch bản về nhất quán dữ liệu: nhằm kiểm tra các kiểu dữ liệu và độ dài cho một thuộc tính cụ thể có khác nhau hay không trong các tệp dữ liệu hoặc bảng dữ liệu nếu ngữ nghĩa là giống nhau; kiểm tra xem các ràng buộc về tính toàn vẹn dữ liệu có bị xâm phạm hay không.
- Kịch bản kiểm thử về tính đầy đủ: nhằm đảm bảo rằng tất cả các dữ liệu mong đợi được tải vào bảng đích và so sánh số lượng bản ghi giữa nguồn và đích.
- Kịch bản kiểm thử về tính đúng đắn: nhằm kiểm tra xem dữ liệu có bị đọc sai cú pháp hoặc ghi không chính xác; kiểm

tả xem các giá trị dữ liệu có bị rỗng (null)? không duy nhất? hoặc ngoài phạm vi?

- Kịch bản kiểm thử về sự biến đổi dữ liệu: nhằm kiểm tra xem dữ liệu được biến đổi như thế nào? có chính xác và theo đúng mục đích của bài toán?

- Kịch bản kiểm thử chất lượng dữ liệu: nhằm kiểm tra các con số cần kiểm tra và xác nhận; kiểm tra ngày tháng; kiểm tra độ chính xác; kiểm tra giá trị rỗng (null).

- Kịch bản kiểm thử sự trùng lặp.

- Kịch bản kiểm thử sự đồng bộ hóa dữ liệu theo nhãn thời gian.

- Kịch bản kiểm thử về tính hoàn chỉnh của dữ liệu.

- Kịch bản kiểm thử về độ sạch dữ liệu: nhằm xác thực xem dữ liệu còn nhiễu hay không: Đối với dữ liệu giám sát bên trong (SCADA) hoặc dữ liệu hạ tầng lưới điện thì nhiễu thường tồn tại ở phần giá trị của các thuộc tính dữ liệu. Đối với dữ liệu hình ảnh thì các loại nhiễu thường rất đa dạng, nhiễu có thể phát sinh do thiết bị thu nhận (chẳng hạn ống kính camera bị bẩn), nhiễu có thể sinh ra do điều kiện thời tiết không tốt (nắng chói, mưa, mây mù,…) trong khi thu nhận dữ liệu.

4.3. Dữ liệu kiểm thử

Dữ liệu kiểm thử được chuẩn bị theo các nguyên tắc sau:

- Tạo các tập dữ liệu kiểm thử tốt nhất: Tạo dữ liệu tốt nhất trong khả năng có thể, không quá dài dòng nhưng đảm bảo nhận biết được lỗi cho nhiều loại ứng dụng khác nhau; với điều kiện không tốn nhiều chi phí và thời gian trong việc chuẩn bị dữ liệu kiểm thử và thực hiện kiểm thử.

- Thiết lập những dữ liệu không hợp lệ: Để kiểm tra sự đúng đắn của dữ liệu, cần phải tạo các dữ liệu có format sai. Những dữ liệu kiểu như vậy sẽ không được chấp nhận bởi hệ thống và sẽ xuất hiện các thông báo lỗi. Các ca kiểm thử trên tập dữ liệu này nhằm đảm bảo hệ thống sẽ sinh ra các thông báo báo lỗi.

- Tạo bộ dữ liệu đúng: Tạo bộ dữ liệu này nhằm đảm bảo rằng hệ thống hoặc ứng dụng sẽ phản ứng giống như yêu cầu kỹ thuật, hoặc nhận biết dữ liệu đúng hay sai đã được lưu lại vào cơ sở dữ liệu hay tệp.

- Tạo bộ dữ liệu sai: Tạo bộ dữ liệu này nhằm xác nhận phản ứng của hệ thống đối với các giá trị phủ định, các chuỗi đầu vào chứa ký tự hoặc số. Tạo bộ dữ liệu cho việc kiểm thử hiệu quả, tải dữ liệu, các thao tác truy vấn trên dữ liệu và kiểm thử hồi qui.

- Tạo ra dữ liệu trống hoặc dữ liệu mặc định: Thực hiện các test case với dữ liệu trống hoặc dữ liệu mặc định nhằm mục đích kiểm tra xem các thông báo lỗi có được hiển thị đúng hay không.

Để kiểm thử đường ống dữ liệu cũng như các thuật toán tích hợp và đồng bộ hóa dữ liệu, nhóm đã thu thập và xây dựng được 03 tập cơ sở dữ liệu tương ứng từ 03 nguồn dữ liệu chính: Dữ liệu giám sát ngoài (UAV), dữ liệu giám sát trong HT (SCADA), dữ liệu cơ sở hạ tầng lưới điện 110 kV do Công ty Lưới điện cao thế

thành phố Hà Nội quản lý.

Dữ liệu giám sát ngoài bao gồm: Tập dữ liệu video/hình ảnh, tọa độ GPS của các đối tượng trên lưới như cột điện, móng cột, đường dây truyền tải, dây chống sét, dây nối đất, cách điện (cách điện thủy tinh, cách điện silicon), phụ kiện cách điện, tạ chống rung, thanh xà, hành lang tuyến... Tập dữ liệu chứa các thông số môi trường (nhiệt độ, độ ẩm, mức độ gió, mức độ sương mù,…) tại từng thời điểm bay chụp UAV trên lưới (Hình 4a).

(a)

PC Time	F(Hz)	Pt(kW)	Qt(kVAR)	Cos phi	Ua(V)	Ub(V)	Uc(V)	Ia(A)	Ib(A)	Ic(A)
26-11-21 16:12	50.03	24736.38	-13625.87	0.85	66554.67	66668.27	66732.75	150.55	139.75	136.56
26-11-21 15:41	49.89	23285.35	-10069.13	0.9	65972.36	66087.78	66160.17	137.37	126.9	123.98
26-11-21 15:12	50	21141.35	-15995.76	0.77	66240.81	66346.68	66423.03	140.68	129.54	126.78
26-11-21 14:42	50.07	20096.15	-15402.33	0.76	66136.77	66248.94	66314.91	135.25	124.87	121.6
26-11-21 14:12	50.01	19376.38	-18365.64	0.71	66095.52	66206.62	66287.53	139.81	128.62	125.26
26-11-21 13:42	50.15	18560.9	-21918.55	0.63	66440.15	66558.56	66642.12	151.33	140.41	136.4
26-11-21 13:12	50.04	17760.73	-24288.44	0.56	66767.7	66898.59	66956.07	158.96	147.42	143.07
26-11-21 12:42	50.13	17565.47	-25475.29	0.54	67997.66	68167.58	68220.85	162.05	150.71	146.17
26-11-21 11:41	49.93	19908.55	-23105.41	0.63	68024.81	68184.3	68248.4	160.86	149.89	146.33
26-11-21 11:11	50.06	24008.95	-26658.32	0.65	68104.8	68251.24	68310.96	181.91	170.48	167.29
26-11-21 10:41	50.08	25670.55	-19548.67	0.78	67516.76	67674.85	67732.42	163.44	152.41	149.37
26-11-21 10:12	50.08	24724.89	-18365.64	0.77	67649.95	67799.08	67836.06	160.29	149.64	146.93
26-11-21 9:41	49.94	24326.72	-19548.67	0.78	67862.03	67985.44	68038.98	159.8	148.74	146.17
26-11-21 9:11	50.07	22351.18	-20735.52	0.71	67975.47	68091.71	68153.04	156.48	145.61	142.92
26-11-21 8:41	50.13	23595.46	-15995.76	0.81	67941.73	68042.29	68123.38	146.98	136.69	132.44
26-11-21 8:11	50.05	22071.69	-15402.33	0.82	66071.95	66206.27	66262.65	144.93	134.48	131.8

(b)

	A	B	C	D
1	Mã thiết bị	Đường dây/TBA	Thiết bị/Công trình cha	Thiết bị/Công trình
2	PD-{PREFIX}.779090	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 1	Thép hình - N122-29C - Néo góc
3	PD-{PREFIX}.779120	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 2	Thép hình - N122-24C - Néo góc
4	PD-{PREFIX}.779127	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 3	Thép hình - N122-29C - Néo góc
5	PD-{PREFIX}.779141	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 5	Thép hình - N122-29C - Néo góc
6	PD-{PREFIX}.779147	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 7	Thép hình - N122-29B - Néo góc
7	PD-{PREFIX}.779153	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 8	Thép hình - N122-29B - Néo góc
8	PD-{PREFIX}.779159	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 21	Thép hình - N122-29C - Néo góc
9	PD-{PREFIX}.779165	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 33	Thép hình - N122-29B - Néo góc
10	PD-{PREFIX}.779173	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 39	Thép hình - N122-29B - Néo góc
11	PD-{PREFIX}.779179	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 41	Thép hình - N122-29B - Néo góc
12	PD-{PREFIX}.779185	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 42	Thép hình - N122-29C - Néo góc
13	PD-{PREFIX}.779191	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 49	Thép hình - N122-29C - Néo góc
14	PD-{PREFIX}.779197	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 51	Thép hình - N122-29C - Néo góc
15	PD-{PREFIX}.779203	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 6	Thép hình - N122-33C - Néo góc
16	PD-{PREFIX}.779212	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 15	Thép hình - N122-33C - Néo góc
17	PD-{PREFIX}.779220	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 16	Thép hình - N122-33C - Néo góc
18	PD-{PREFIX}.779226	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 23	Thép hình - N122-33C - Néo góc
19	PD-{PREFIX}.779232	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 25	Thép hình - N122-33C - Néo góc
20	PD-{PREFIX}.779238	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 27	Thép hình - N122-33C - Néo góc
21	PD-{PREFIX}.779244	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 30	Thép hình - N122-33C - Néo góc
22	PD-{PREFIX}.779250	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 31	Thép hình - N122-33C - Néo góc
23	PD-{PREFIX}.779269	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 43	Thép hình - N122-33C - Néo góc
24	PD-{PREFIX}.779277	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 53	Thép hình - N122-33C - Néo góc
25	PD-{PREFIX}.779289	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 4	Thép hình - Đ122-27C - Đỡ thẳng
26	PD-{PREFIX}.779304	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 9	Thép hình - Đ122-27C - Đỡ thẳng
27	PD-{PREFIX}.779310	ĐZ 171 E10.5 Xuân Mai - 171E1.51 Phú Nghĩa	VT 10	Thép hình - Đ122-27C - Đỡ thẳng

Cột điện 110kV | Nút 110kV | Sứ 110kV | Đường dây 110kV | Dây chống sét 110kV | Móng cột 110kV | Dây dẫn 1 ...

(c)

Hình 4. Các nguồn dữ liệu: a. Dữ liệu giám sát ngoài, b. Dữ liệu giám sát bên trong hệ thống điện, c. Dữ liệu cơ sở hạ tầng lưới điện

Dữ liệu giám sát bên trong hệ thống điện hiện tại được thu nhận bằng cơ chế import dữ liệu từ hệ thống giám sát SCADA của EVN. Dữ liệu sau khi import được lưu dưới dạng flat file (.xlsx hoặc .csv) (Hình 4b).

Dữ liệu cơ sở hạ tầng lưới điện hiện tại được kế thừa (trích xuất) từ hệ thống PMIS và GIS của EVN. Dữ liệu sau khi trích xuất được lưu trữ dưới dạng flat file (.csv, .xlsx) (Hình 4c).

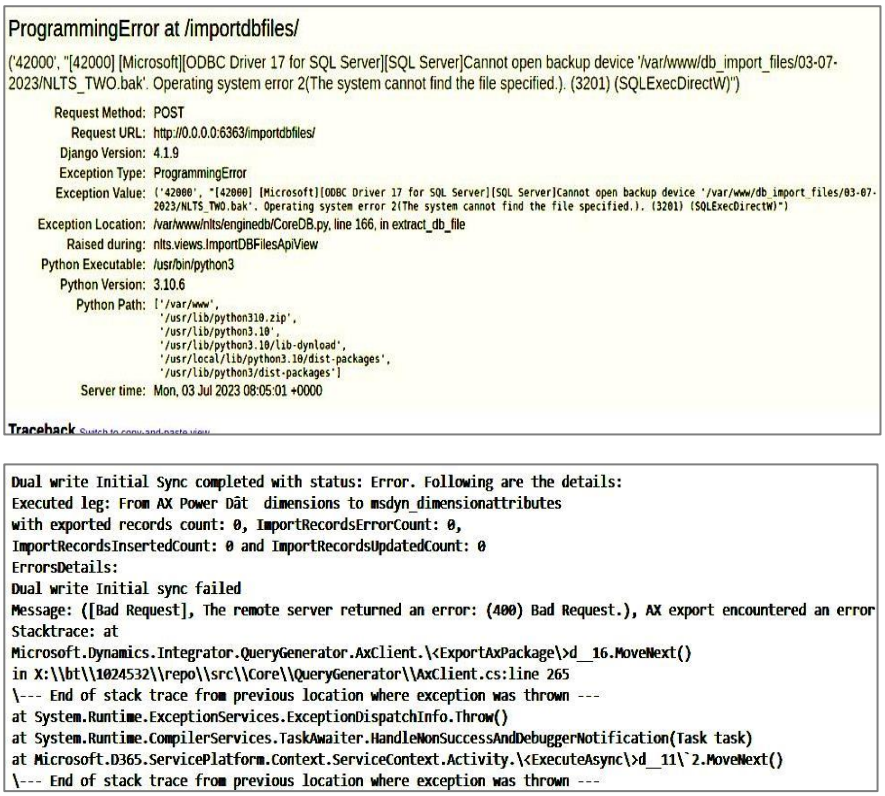
4.4. Tiến hành kiểm thử và kết quả

Trong quá trình kiểm thử module tích hợp và đồng bộ hóa dữ liệu, nghiên cứu đã ghi nhận được một số loại lỗi (Hình 5).

Các lỗi điển hình gồm: Dữ liệu bị thiếu, dữ liệu nhiễu, dữ liệu không cân bằng, dữ liệu không chính xác hoặc không đúng, dữ liệu dư thừa. Ngoài ra, còn một số lỗi phát sinh do các nguyên nhân như: Định dạng kiểu dữ liệu ngày tháng khác nhau giữa các bảng dữ liệu, định danh đối tượng từ các CSDL nguồn chưa rõ ràng, điều kiện ràng buộc và các quyền tải và trích xuất dữ liệu chưa chặt chẽ, hoặc do không tương thích giữa các phiên bản của các thư viện trong môi trường python.

Giải pháp để khắc phục lỗi dữ liệu bị thiếu bao gồm thu thập thêm dữ liệu và sử dụng các kỹ thuật imputation như mean imputation, median imputation, hoặc phương pháp mô phỏng để điền vào dữ liệu thiếu.

Giải pháp để khắc phục lỗi dữ liệu nhiễu: Dữ liệu nhiễu được định nghĩa là dữ liệu có giá trị khác biệt hoặc sai lệch so với những dữ liệu xung quanh. Để khắc phục lỗi này, nhóm sử dụng các phương pháp lọc nhiễu như smoothing, hoặc outlier detection để xác định và xử lý các dữ liệu nhiễu.



Hình 5. Một số hình ảnh trong quá trình kiểm thử module tích hợp và đồng bộ dữ liệu hệ thống

Giải pháp để khắc phục vấn đề dữ liệu không cân bằng: Dữ liệu không cân bằng được coi là dữ liệu không được phân bổ đồng đều giữa các lớp hoặc các kết quả khác nhau. Hướng tiếp cận để khắc phục vấn đề này gồm thu thập thêm dữ liệu và sử dụng các kỹ làm giàu dữ liệu (data augmentation), bao gồm các kỹ thuật sinh dữ liệu và biến đổi dữ liệu hoặc kỹ thuật resampling như SMOTE (Synthetic Minority Over-sampling Technique) để tạo ra dữ liệu mô phỏng và cân bằng dữ liệu.

Đối với vấn đề dữ liệu không chính xác hoặc không đúng cần kiểm định dữ liệu hoặc làm sạch dữ liệu để loại bỏ, chỉnh sửa hoặc thay thế dữ liệu không chính xác.

Giải pháp để khắc phục vấn đề dữ liệu dư thừa: Dữ liệu dự thừa là dữ liệu bị trùng lặp hoặc không cần thiết cho việc phân tích hoặc mô hình hóa. Để khắc phục vấn đề này, chúng tôi sử dụng các kỹ thuật lựa chọn đặc trưng (feature selection) giảm kích thước dữ liệu (dimensionality reduction) để loại bỏ dữ liệu dư thừa.

Đối với các vấn đề về định dạng kiểu dữ liệu ngày tháng khác nhau giữa các bảng

dữ liệu, chúng tôi đề xuất tích hợp sử dụng thư viện hỗ trợ xử lý ngày tháng như Pandas trong Python để chuyển đổi định (convert) hoặc hợp nhất (merge) các định dạng ngày tháng.

Đối với các vấn đề định danh đối tượng từ các CSDL nguồn chưa rõ ràng cần bổ sung hoặc tạo thêm trường định danh duy nhất, ví dụ như khóa chính (primary key) hoặc tạo một trường mới để lưu trữ định danh duy nhất.

Đối với các vấn đề về điều kiện ràng buộc và các quyền tải và trích xuất dữ liệu chưa chặt chẽ cần bổ sung thêm các test case để đảm bảo tính nhất quán.

Vấn đề không tương thích giữa các phiên bản của các thư viện trong môi trường python được khắc phục bằng cách kiểm tra và lựa chọn, tích hợp các phiên bản tương thích của các thư viện trong quá trình cài đặt.

5. KẾT LUẬN

Nghiên cứu đã sử dụng giải pháp tích hợp và đồng bộ hóa dữ liệu bằng phương pháp tạo đường ống dữ liệu ELT, đồng thời thực hiện thiết kế và lập trình xây dựng mô hình này sử dụng ngôn ngữ lập trình python, tích hợp các hệ quản trị SQL (MySQL), SQL Server, MongoDB, Singer.io và Pipelinewise để tích hợp và đồng bộ các nguồn dữ liệu của hệ thống. Quá trình thử nghiệm và đánh giá thuật toán cũng được thực hiện theo các kịch bản thử nghiệm một cách chi tiết sử dụng 03 nguồn dữ liệu chính: Dữ liệu giám sát ngoài (UAV), dữ liệu giám sát trong hệ thống (SCADA), dữ liệu cơ sở hạ tầng lưới điện 110 kV do Công ty Lưới điện cao thế thành phố Hà Nội quản lý để phát hiện các bug và đưa ra phương án cải thiện thuật toán và dữ liệu đảm bảo độ chính xác cho mô hình tích hợp và đồng bộ hóa dữ liệu, hướng tới đảm bảo độ chính xác cho các thuật toán phân tích và dự báo sự cố.

LỜI CẢM ƠN

Nhóm tác giả trân trọng cảm ơn Chương trình hỗ trợ nghiên cứu, phát triển và ứng dụng công nghệ của công nghiệp 4.0, mã số: KC-4.0.31/19-25, đã hỗ trợ trong quá trình nghiên cứu.

TÀI LIỆU THAM KHẢO

[1] Bộ Công Thương, "Thông tư số 31/2019/TT-BCT của Bộ Công Thương: Sửa đổi, bổ sung một số điều của Thông tư số 28/2014/TT-BCT ngày 15 tháng 9 năm 2014 của Bộ trưởng Bộ Công Thương quy định quy trình xử lý sự cố trong hệ thống điện quốc gia, Thông tư số 40/2014/TT-BCT ngày 05 tháng 11 năm 2014 của Bộ trưởng Bộ Công Thương quy định quy trình điều độ hệ thống điện quốc gia và Thông tư số 44/2014/TT-BCT ngày 28 tháng 11 năm 2014 của Bộ trưởng Bộ Công Thương quy định quy trình thao tác trong hệ thống điện quốc gia", 2019.

[2] "14/08/2003: Mất điện diện rộng ở Đông Bắc Hoa Kỳ", *Nghiên cứu quốc tế,* 2003.

[3] "Khắc phục sự cố lưới điện tại khu công nghiệp", *Báo điện tử Bắc Giang*, 2015.

[4] "Điện không ổn định, doanh nghiệp thiệt hại lớn", *Báo điện tử Hải Dương*, 2019.

[5] "Progress Report for OE ARRA Smart Grid Demonstration Program Aggregation of RDSI, SGDP, and SGIG Results", *U.S. Departement of Energy*, 2015.

[6] "Oncor's Pioneering Transmission Dynamic Line Rating (DLR) Report", *U.S. Department of Energy*, 2014.

[7] C. Deng, S. Wang, Z. Huang, Z. Tan, J. Liu, "Unmanned aerial vehicles for power line inspection: A cooperative way in platforms and communications", *Journal of Communications*, vol. 9, no. 9, pp. 687–692, 2014.

[8] Ei Phyo Thwe, Min Min Oo, "Fault Detection and Classification for Transmission Line Protection System Using Artificial Neural Network", *Journal of Electrical and Electronic Engineering*, vol. 4, no. 5, pp. 89–96, 2016.

[9] Bas Geerdink, Tobias Macey, 97 Things Every Data Engineer Should Know: Collective Wisdom from the Experts, *O'Reilly Media*, 2021.

[10] Martin Kleppmann, Designing Data-Intensive Applications, *O'Reilly*, 2017.

[11] Laura Madsen, Disrupting Data Governance: A Call to Action, *Technics Publications*, 2019.

[12] Nathan Marz, Big Data: Principles and best practices of scalable realtime data systems, *Manning, 1st edition,* 2015.

[13] Paul Crickard, Data Engineering with Python: Work with massive datasets to design data models and automate data pipelines using Python, *Packt Publishing*, 2020.

[14] Michael Walker, Python Data Cleaning Cookbook: Modern techniques and Python tools to detect and remove dirty data and extract key insights, *Packt Publishing*, 2020.

[15] James Densmore, Data Pipelines Pocket Reference: Moving and Processing Data for Analytics, O'Reilly Media, 2021.

[16] Paul Jorgensen, Software Testing: A Craftsman's Approach, Fourth Edition, *Auerbach Publications*; 4th edition, 2013.

Giới thiệu tác giả:

Tác giả Vũ Thị Thu Nga tốt nghiệp đại học ngành hệ thống điện năm 2004, nhận bằng Thạc sĩ ngành kỹ thuật điện năm 2007 tại Đại học Bách khoa Hà Nội; nhận bằng Tiến sĩ ngành kỹ thuật điện năm 2014 tại Đại học Toulouse - Pháp. Hiện nay tác giả là giảng viên Trường Đại học Điện lực.

Lĩnh vực nghiên cứu: Tích điện không gian, HVDC, vật liệu cách điện, kỹ thuật điện cao áp, rơle và tự động hóa trong hệ thống điện.

Tác giả Nguyễn Thị Thanh Tân tốt nghiệp đại học ngành khoa học máy tính năm 1999 và nhận bằng Thạc sĩ ngành công nghệ thông tin năm 2001 tại Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội; nhận bằng Tiến sĩ ngành khoa học máy tính tại Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Hiện nay tác giả công tác tại Trường Đại học Điện lực.

Lĩnh vực nghiên cứu: Xử lý ảnh, học máy, công nghệ tri thức, khai phá dữ liệu, trí tuệ nhân tạo (AI).