

MÔ HÌNH LAI SARIMA-SVR TỐI ƯU BẰNG THUẬT TOÁN FIREFLY VỚI KỸ THUẬT NỘI SUY NÂNG CAO CHO DỰ BÁO TIÊU THỤ NĂNG LƯỢNG TÒA NHÀ

Nguyễn Văn Minh Đức^{1*}, Trần Văn Thịnh¹, Hoàng Anh²

*Tác giả liên hệ, email: nvmduc@hou.edu.vn. ORCID: 0009-0001-4356-3837

Ngày tòa soạn nhận được bài báo: 15/01/2026

Ngày phản biện đánh giá: 17/03/2026

Ngày bài báo được duyệt đăng: 14/04/2026

DOI: 10.59266/houjs.2026.1162

Tóm tắt: Nghiên cứu đề xuất một khung phương pháp toàn diện nhằm nâng cao độ chính xác dự báo tiêu thụ năng lượng từ dữ liệu thực tế chứa nhiều lỗi. Trọng tâm là mô hình lai SARIMA-SVR (Seasonal AutoRegressive Integrated Moving Average - Support Vector Regression) được tối ưu hóa siêu tham số bằng thuật toán Firefly (FA). Phương pháp có ba đóng góp chính: (1) Kỹ thuật nội suy dựa trên mẫu hình thời vụ (theo giờ, ngày, mùa) để khôi phục dữ liệu lỗi, đảm bảo chất lượng đầu vào; (2) Kiến trúc lai chặt chẽ, sử dụng đầu ra dự báo xu hướng và thời vụ từ mô hình SARIMA làm một đặc trưng ưu tiên cho mô hình SVR, cho phép SVR tập trung học các thành phần phi tuyến phức tạp; (3) Ứng dụng FA để tự động tìm bộ tham số tối ưu cho SVR, hiệu quả hơn các phương pháp truyền thống. Kết quả thực nghiệm trên tập dữ liệu thực tế chứng minh mô hình lai đề xuất đạt độ chính xác vượt trội một cách có ý nghĩa so với các mô hình cơ sở như SARIMA, SVR thuần túy, LSTM (Long Short-Term Memory). Nghiên cứu không chỉ cung cấp một công cụ dự báo khả thi cho các hệ thống năng lượng mà còn đóng góp một phương pháp luận tích hợp giữa tiền xử lý dữ liệu, mô hình hóa lai và tối ưu hóa. Hướng phát triển tương lai bao gồm tích hợp biến ngoại cảnh và thử nghiệm trên các bộ dữ liệu đa lĩnh vực.

Từ khóa: dự báo tiêu thụ năng lượng, mô hình lai SARIMA-SVR, thuật toán Firefly, nội suy dữ liệu nâng cao, xử lý dữ liệu lỗi

I. Đặt vấn đề

Trong bối cảnh chuyển đổi số, dự báo chính xác nhu cầu năng lượng là nòng cốt cho các hệ thống quản lý thông minh. Tuy nhiên, dữ liệu thực tế thường chứa nhiều nhiễu và giá trị thiếu, làm sai lệch nghiêm

trọng kết quả dự báo nếu không được xử lý đúng cách (Little và Rubin, 2019).

Lịch sử phát triển phương pháp dự báo cho thấy sự tiến hóa từ các mô hình thống kê cổ điển như ARIMA/SARIMA - hiệu quả với các mẫu hình tuyến tính

¹ Khoa Điện - Điện tử, Trường Đại học Mở Hà Nội, Hà Nội, Việt Nam

² Trường Điện - Điện tử, Đại học Bách Khoa Hà Nội, Hà Nội, Việt Nam

và chu kỳ nhưng kém linh hoạt với phi tuyến - đến các mô hình học máy như SVR, LSTM có khả năng mô hình hóa phi tuyến mạnh nhưng đòi hỏi dữ liệu lớn và dễ học quá mức (Smola và Schölkopf, 2004; Hochreiter và Schmidhuber, 1997). Xu hướng phát triển mô hình lai nhằm kết hợp ưu điểm của cả hai trường phái (Zhang, 2003). Tuy nhiên, phần lớn các nghiên cứu lai hiện tại vẫn tồn tại ba hạn chế chính: (1) xử lý dữ liệu đầu vào sơ sài; (2) kết hợp mô hình thiếu cơ chế tích hợp tối ưu và tối ưu siêu tham số hiệu quả; (3) chưa cân nhắc đầy đủ tính khả thi khi triển khai thực tế (Khashei và Bijari, 2011). Với mô hình lai SARIMA-SVR, các nghiên cứu phổ biến thường sử dụng SARIMA dự báo sơ bộ rồi dùng SVR dự báo phần dư, hoặc kết hợp tuyến tính hai đầu ra. Các cách tiếp cận này chưa tận dụng triệt để khả năng học phi tuyến của SVR. Phương pháp đề xuất khác biệt ở chỗ đầu ra dự báo của SARIMA được sử dụng trực tiếp như một đặc trưng ưu tiên đầu vào cho SVR, cho phép SVR đồng thời học bổ sung các

quan hệ phi tuyến phức tạp. Xuất phát từ đó, nghiên cứu đề xuất khung phương pháp lai toàn diện với ba đóng góp chính:

Đóng góp 1: Kỹ thuật nội suy dựa trên mẫu hình thời vụ thông minh, xử lý khác biệt lỗi ngắn/dài hạn.

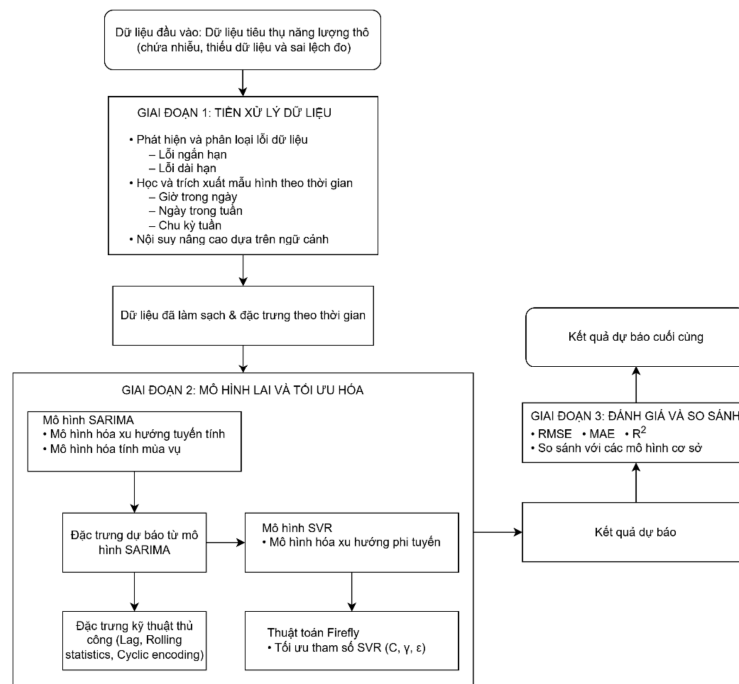
Đóng góp 2: Kiến trúc lai SARIMA-SVR tích hợp chặt chẽ, sử dụng đầu ra dự báo SARIMA làm đặc trưng đầu vào cho SVR.

Đóng góp 3: Ứng dụng thuật toán Firefly để tự động lựa chọn bộ tham số tối ưu cho SVR (Yang, 2010).

II. Phương pháp, vật liệu nghiên cứu

2.1. Tổng quan về quy trình nghiên cứu

Quy trình nghiên cứu gồm ba giai đoạn chính: (1) Tiền xử lý dữ liệu với kỹ thuật nội suy nâng cao; (2) Xây dựng và huấn luyện mô hình lai tối ưu; (3) Đánh giá, phân tích kết quả và so sánh với các mô hình tham chiếu (Hình 1).



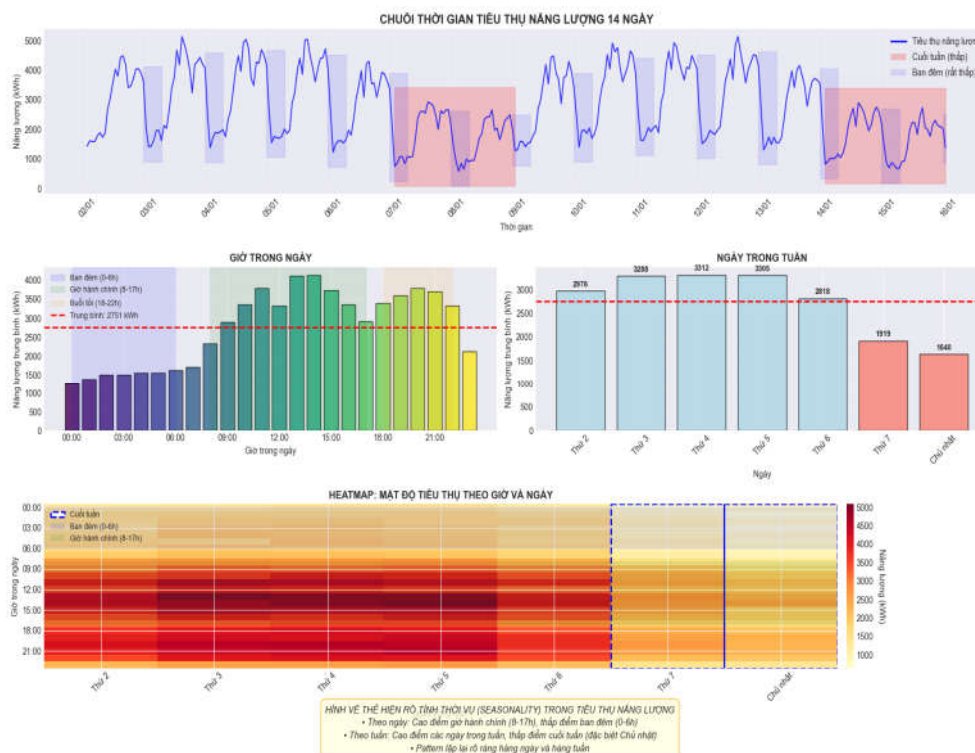
Hình 1. Sơ đồ khối quy trình mô hình lai đề xuất

2.2. Thu thập và mô tả dữ liệu nghiên cứu

Dữ liệu nghiên cứu là chuỗi tiêu thụ năng lượng theo giờ (kWh) từ hệ thống đo đếm thông minh của tòa nhà FPT Complex tại Đà Nẵng, thu thập từ tháng 01 đến tháng 06/2022. Đây là dữ liệu thực tế, phản ánh đầy đủ các đặc tính vận hành và thách thức thường gặp.

Bảng 1. Thống kê mô tả dữ liệu tiêu thụ năng lượng gốc

Chỉ số thống kê	Giá trị
Tổng số mẫu	4,344
Khoảng thời gian	01/01/2022 - 30/06/2022
Giá trị trung bình (Mean)	6,458 kWh
Độ lệch chuẩn (Std)	8,037 kWh
Giá trị nhỏ nhất (Min)	0 kWh
Giá trị lớn nhất (Max)	59,812 kWh
Số điểm dữ liệu lỗi	418 (9.6%)



Hình 2. Phân bổ dữ liệu

Dữ liệu thể hiện tính thời vụ rõ rệt theo ngày và tuần: tiêu thụ thấp vào cuối tuần và ban đêm, cao vào giờ hành chính các ngày làm việc. Tỷ lệ dữ liệu lỗi (~9.6%) là đáng kể và cần được xử lý kỹ lưỡng trước khi đưa vào mô hình dự báo.

2.3. Phương pháp tiền xử lý dữ liệu và nội suy nâng cao

Để giải quyết vấn đề dữ liệu lỗi, chúng tôi đề xuất một quy trình tiền xử lý thông minh gồm ba bước chính: Phát hiện lỗi, học mẫu hình và nội suy theo ngữ cảnh.

Phương pháp này kế thừa và phát triển từ các nghiên cứu về xử lý dữ liệu thiếu trong chuỗi thời gian (Little & Rubin, 2019), (Junninen và cộng sự, 2004).

Bước 1: Phát hiện và phân loại lỗi: Các điểm dữ liệu lỗi (giá trị âm, giá trị 0 trong giờ hành chính, hoặc vượt ngưỡng) được gộp thành cụm. Cụm ≤ 6 giờ là lỗi ngắn hạn, > 6 giờ là lỗi dài hạn.

Bước 2: Học mẫu hình tiêu thụ theo mùa: Từ phần dữ liệu hợp lệ, hệ thống tự động học các mẫu hình tiêu thụ ở nhiều cấp

độ thời gian, dựa trên nguyên lý phân rã chuỗi thời gian (Cleveland và cộng sự, 1990):

Mẫu hình theo giờ trong ngày

Mẫu hình theo ngày trong tuần

Mẫu hình theo tháng: Xu hướng tiêu thụ thay đổi theo mùa trong năm.

Việc học được thực hiện bằng cách tính toán giá trị trung bình và độ lệch chuẩn cho từng nhóm thời gian tương ứng.

$$V_{interp} = \alpha \cdot V_{pattern} + (1 - \alpha) \cdot \frac{\sum_{i=1}^n (w_i \cdot V_{neighbor,i})}{\sum_{i=1}^n w_i} \quad (1)$$

với $w_i = \frac{1}{1+d_i}$, d_i là khoảng cách thời gian từ điểm lỗi đến điểm lân cận thứ i , $V_{pattern}$ là giá trị mẫu hình theo giờ tương ứng, và α là hệ số cân bằng.

Hệ số α được xác định bằng thực nghiệm trên 20% dữ liệu hợp lệ (không chứa lỗi) bằng cách mô phỏng các cụm lỗi ngắn hạn với độ dài từ 1 đến 6 giờ, sau đó thử nghiệm với các giá trị từ 0 đến 1.

Kết quả đánh giá dựa trên RMSE và MAE cho thấy $\alpha=0,3$ đạt sai số khôi phục nhỏ nhất và bảo toàn tốt nhất tính thời vụ của chuỗi, do đó được chọn làm tham số mặc định cho toàn bộ thực nghiệm. Phương

$$f(t) = V_{before} + (V_{after} - V_{before}) \cdot \frac{1}{1 + \exp(-k(t-t_0))} \quad (2)$$

với k là hệ số điều chỉnh độ dốc (đặt $k=12$), t là vị trí thời gian trong cụm lỗi (chuẩn hóa về $[0,1]$), t_0 là điểm giữa ($t_0=0.5$).

Điều chỉnh theo mẫu hình: Điều chỉnh từng điểm trên đường cơ sở bằng hệ số tỷ lệ lấy từ mẫu hình theo giờ để đảm bảo bảo toàn tính thời vụ:

$$V_{final}(t) = f(t) \cdot \frac{P_h}{P^-} \quad (3)$$

Trong đó P_h là giá trị mẫu hình theo giờ tại thời điểm t , P^- là giá trị trung bình của tất cả các giờ.

Bước 3: Kỹ thuật nội suy nâng cao theo ngữ cảnh

Đối với lỗi ngắn hạn, giá trị được khôi phục dựa trên sự kết hợp có trọng số giữa mẫu hình và giá trị lân cận, được phát triển từ phương pháp nội suy theo khoảng cách nghịch đảo (Inverse Distance Weighting - IDW) (Shepard, 1968) trong không gian thời gian:

pháp này kết hợp thông tin dài hạn (mẫu hình) và ngắn hạn (giá trị lân cận), tối ưu hơn so với nội suy tuyến tính đơn thuần.

Đối với lỗi dài hạn, một phương pháp phức tạp hơn được áp dụng, kết hợp nội suy phi tuyến với điều chỉnh theo mẫu hình:

Xác định ngữ cảnh: Tìm giá trị hợp lệ cuối cùng trước cụm lỗi (V_{before}) và giá trị hợp lệ đầu tiên sau cụm lỗi (V_{after}).

Tạo đường cơ sở: Tạo một đường cong nội suy phi tuyến sử dụng hàm sigmoid điều chỉnh, cho phép mô phỏng chuyển tiếp mượt mà giữa hai điểm biên (Han & Moraga, 1995):

Phương pháp đề xuất tạo ra dữ liệu khôi phục một cách “tự nhiên”, liền mạch và bảo toàn được cấu trúc thống kê vốn có của chuỗi thời gian, qua đó khắc phục hiệu quả các hạn chế của phương pháp nội suy tuyến tính truyền thống.

2.4. Kiến trúc mô hình lai đề xuất

Kiến trúc lai sử dụng đầu ra dự báo SARIMA như một đặc trưng đầu vào cho SVR, cho phép SVR điều chỉnh trực tiếp đóng góp của thành phần tuyến tính - thời vụ thông qua hệ số học trong không gian

kernel, đồng thời giảm nguy cơ mất thông tin do tách mô hình.

2.4.1. Thành phần SARIMA

Mô hình SARIMA (Seasonal AutoRegressive Integrated Moving Average) được sử dụng làm nền tảng để nắm bắt cấu trúc tuyến tính, xu hướng và đặc biệt là tính thời vụ trong dữ liệu tiêu thụ năng lượng. Một mô hình SARIMA cho chuỗi thời gian được biểu diễn bởi phương trình (Hyndman & Athanasopoulos, 2018):

$$\phi_p(B)\Phi_p(B^s)(1-B)^d(1-B^s)^D y_t = \theta_q(B)\Theta_q(B^s)\epsilon_t \quad (4)$$

Trong đó:

B là toán tử dịch chuyển lùi (Backshift operator): $By_t = y_{(t-1)}$.

$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$ là đa thức tự hồi quy (AR) bậc p .

$\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q$ là đa thức trung bình trượt (MA) bậc q .

$\Phi_p(B^s)$ và $\Theta_q(B^s)$ là các đa thức AR và MA cho thành phần mùa, với bậc tương ứng là P và Q .

$(B)^d$ và $(B^s)^D$ biểu thị phép lấy sai phân bậc d và sai phân mùa bậc D để đạt được tính dừng (stationarity).

s là chu kỳ mùa (seasonal period), với dữ liệu theo giờ có chu kỳ ngày rõ rệt, chúng tôi đặt $s=24$.

ϵ_t là nhiễu trắng (white noise) tại thời điểm t .

Đầu ra dự báo từ mô hình SARIMA, ký hiệu là $\hat{y}_t^{SARIMA}$, được xem như một đặc trưng đầu vào cho mô hình SVR, vì nó mã hóa thông tin về xu hướng, tính thời vụ và các phụ thuộc tuyến tính của dữ liệu.

2.4.2. Kỹ thuật xây dựng đặc trưng cho SVR

Để cho phép mô hình SVR nắm bắt các mối quan hệ phi tuyến phức tạp vượt ra ngoài khả năng của SARIMA, một tập hợp các đặc trưng được thiết lập một cách có hệ thống. Quá trình này dựa trên nguyên lý cho rằng việc biểu diễn tốt dữ liệu là chìa khóa cho hiệu suất của các mô hình học máy.

Đặc trưng thời gian tuần hoàn: Các biến thời gian rời rạc (giờ, ngày trong tuần, tháng) được chuyển đổi thành các đặc trưng tuần hoàn liên tục sử dụng phép biến đổi sin-cos (Díaz và cộng sự, 2019).

Đặc trưng độ trễ: Giá trị quan sát trong quá khứ trực tiếp ảnh hưởng đến hiện tại. Chúng tôi chọn các độ trễ có ý nghĩa vật lý: 1, 2, 3 giờ (phụ thuộc ngắn hạn), 24 giờ (phụ thuộc cùng kỳ ngày hôm trước), và 168 giờ (phụ thuộc cùng kỳ tuần trước) (Hamilton, 2020).

$$L_t^k = y_{t-k} \text{ với } k \in \{1, 2, 3, 24, 168\} \quad (5)$$

Đặc trưng thống kê động: Các đặc trưng này nắm bắt xu hướng và độ biến động cục bộ. Chúng tôi tính trung bình trượt và độ lệch chuẩn trượt cho các cửa sổ 3, 6, 12, và 24 giờ (Shumway & Stoffer, 2017).

$$M_t^w = \frac{1}{w} \sum_{i=1}^w y_{t-i} \quad S_t^w = \sqrt{\frac{1}{w-1} \sum_{i=1}^w (y_{t-i} - M_t^w)^2} \quad (6)$$

với w là kích thước cửa sổ.

Đặc trưng dự báo từ SARIMA: Đây là đặc trưng then chốt cho sự tích hợp. Giá trị dự báo $\hat{y}_t^{SARIMA}$ từ mô hình SARIMA được đưa vào như một đặc trưng, cung cấp cho SVR một xu hướng tuyến tính đã

được làm sáng tỏ, giúp SVR tập trung vào việc điều chỉnh các sai số phi tuyến.

2.4.3. Thành phần SVR và Tối ưu hóa bằng Thuật toán Firefly

Support Vector Regression (SVR) là một phiên bản của Support Vector

Machine (SVM) cho bài toán hồi quy (Smola & Schölkopf, 2004).

Cho một tập huấn luyện $\{(x_p, y_p), \dots, (x_n, y_n)\}$, với $x_i \in \mathbb{R}^m$ là vector đặc trưng và y_i là giá trị mục tiêu, bài toán

$$\min_{w, b, \xi_i, \xi_i^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \text{ subject to } y_i - (w^T \phi(x_i) + b) \leq \epsilon + \xi_i, \quad (7)$$

$$(w^T \phi(x_i) + b) - y_i \leq \epsilon + \xi_i^*, \quad \xi_i, \xi_i^* \geq 0, i = 1, \dots, n.$$

Trong đó:

$\phi(x)$ là ánh xạ (mapping) đưa dữ liệu vào một không gian đặc trưng chiều cao hơn.

$C > 0$ là tham số phạt (penalty parameter), cân bằng giữa độ phẳng của hàm và mức độ cho phép các điểm nằm ngoài ống.

ξ_i, ξ_i^* là các biến slack cho phép vi phạm ràng buộc.

ϵ là tham số xác định độ rộng của “ống” (tube) trong đó không tính phạt sai số.

Sử dụng Kernel RBF (Radial Basis Function) $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$ cho phép SVR học các quan hệ phi tuyến phức tạp trong không gian ban đầu mà không cần tính toán tường minh phép ánh xạ $\phi(x)$ (Smola & Schölkopf, 2004). Tham số γ quyết định ảnh hưởng của từng điểm dữ liệu huấn luyện.

Hiệu năng của SVR phụ thuộc nhạy cảm vào bộ ba siêu tham số $\Theta = (C, \gamma, \epsilon)$. Thay vì sử dụng Grid Search, chúng tôi áp dụng thuật toán Firefly (Firefly Algorithm - FA) (Yang, 2009), một thuật toán siêu heuristic (metaheuristic) lấy cảm hứng từ

$$\theta_i^{t+1} = \theta_i^t + \beta_{ij}(r_{ij})(\theta_j^t - \theta_i^t) + \alpha \cdot (\text{rand} - 0.5) \quad (10)$$

với α là hệ số nhiễu ngẫu nhiên, và rand là một số ngẫu nhiên trong $[0, 1]$. Quá trình này cho phép khám phá hiệu quả không gian tìm kiếm và khai thác các vùng tối ưu tiềm năng, dẫn đến việc

tối ưu hóa của ϵ -SVR được xây dựng để tìm một hàm $f(x) = w^T \phi(x) + b$ sao cho độ lệch lớn nhất so với y_i là ϵ .

Bài toán tối ưu ban đầu được trình bày như sau (Smola & Schölkopf, 2004):

tập tính của loài đom đóm, để tìm kiếm tối ưu trong không gian siêu tham số này.

FA mô phỏng hai hành vi chính: cường độ ánh sáng và sự hấp dẫn. Trong ngữ cảnh tối ưu hóa:

Mỗi con đom đóm i đại diện cho một giải pháp (một bộ tham số $\Theta_i = [C, \gamma, \epsilon]$).

Cường độ ánh sáng I_i của một con đom đóm tỷ lệ nghịch với giá trị hàm mục tiêu $f(\Theta_i)$ (ở đây là RMSE trên tập kiểm định chéo):

$$I_i = \frac{1}{1 + f(\Theta_i)} \quad (8)$$

Độ hấp dẫn β_{ij} giữa hai con đom đóm i và j tỷ lệ nghịch với khoảng cách r_{ij} giữa chúng trong không gian tham số và tỷ lệ thuận với chênh lệch cường độ ánh sáng (Yang, 2009):

$$\beta_{ij}(r_{ij}) = \beta_0 \cdot e^{-\gamma_{ffa} r_{ij}^2} \cdot (I_j - I_i) \quad (9)$$

với β_0 là độ hấp dẫn tại khoảng cách 0, γ_{ffa} là hệ số hấp thụ ánh sáng (khác với γ của SVR), và $r_{ij} = \|\Theta_i - \Theta_j\|$. Một con đom đóm i sẽ di chuyển về phía một con đom đóm j sáng hơn nó ($I_j > I_i$) theo quy tắc cập nhật:

tìm ra bộ tham số Θ^* gần tối ưu toàn cục một cách hiệu quả hơn so với các phương pháp tìm kiếm ngẫu nhiên hoặc lưới thông thường.

2.5. Thiết lập thực nghiệm và đánh giá

2.5.1. Phân chia dữ liệu và tiền xử lý

Áp dụng phương pháp phân chia dữ liệu tuần tự theo thời gian trên chuỗi dữ liệu đã xử lý ($N = 4,344$ giờ). Dữ liệu được chia thành ba tập:

Tập huấn luyện (70%): dùng để ước lượng tham số SARIMA và huấn luyện SVR.

Tập xác thực (15%): dùng để tối ưu siêu tham số SVR bằng thuật toán Firefly và điều chỉnh mô hình, tránh quá khớp.

Tập kiểm tra (15%): chỉ dùng để đánh giá hiệu năng cuối cùng.

2.5.2. Thiết lập thực nghiệm

Các thiết lập chi tiết như sau:

Không gian tìm kiếm: $C \in [2^{(-5)}, 2^{15}]$ (tỷ lệ logarit), $\gamma \in [2^{(-15)}, 2^5]$, $\varepsilon \in [0, 0.5]$. Mỗi cá thể đom đóm biểu diễn một bộ tham số.

Khởi tạo quần thể: Kích thước quần thể $N_{pop} = 30$. Các cá thể được khởi tạo ngẫu nhiên đều trong không gian tìm kiếm. Ánh sáng ban đầu $I_i = 1/(1 + RMSE_i)$ với tính trên tập kiểm tra sau khi huấn luyện SVR với bộ tham số tương ứng.

Tham số thuật toán Firefly: Độ hấp dẫn tại khoảng cách 0: $\beta_0 = 1$. Hệ số hấp thụ ánh sáng: $\gamma_{ffa} = 0.5$. Hệ số nhiễu ngẫu nhiên: $\alpha_{ffa} = 0.2$ (giảm dần theo số thế hệ theo tỷ lệ 0.97). Số thế hệ tối đa: $T_{max} = 100$.

Quy trình di chuyển: Trong mỗi thế hệ, so sánh cường độ sáng của từng cặp đom đóm i và j . Nếu $I_j > I_i$, đom đóm i di chuyển về phía j theo công thức (10)

Sau mỗi thế hệ, áp dụng giảm nhiễu theo $\alpha_{ffa} = 0.97 \cdot \alpha_{ffa}$. Cá thể tốt nhất (có RMSE nhỏ nhất trên tập kiểm tra) được ghi nhận và duy trì qua các thế hệ.

Hàm mục tiêu và lựa chọn: Hàm mục tiêu là RMSE trên tập kiểm tra, được tính sau khi huấn luyện SVR với kernel RBF. Quá trình tối ưu lặp lại 100 thế hệ, kết quả cuối cùng là bộ (C, γ, ε) của cá thể tốt nhất. Các tham số này sau đó được sử dụng để huấn luyện lại SVR trên toàn bộ tập huấn luyện và xác thực trước khi đánh giá trên tập kiểm tra.

Bảng 2 dưới đây tóm tắt các tham số chính được sử dụng trong quy trình tiền xử lý và tối ưu hóa.

Bảng 2. Các tham số trong thuật toán

Thành phần	Tham số	Giá trị
Nội suy lỗi ngắn hạn	α	0,3
	Số điểm lân cận n	4 (mỗi phía)
Nội suy lỗi dài hạn	k	12
	t_0	0,5
Phân loại lỗi	Ngưỡng ngắn/dài hạn	6 giờ
Thuật toán Firefly tìm tham số tối ưu	C	474.4683
	γ	0.0001
		0.0144

2.5.3. Các mô hình cơ sở so sánh

Để xác định hiệu quả của mô hình đề xuất, chúng tôi so sánh với các mô hình cơ sở đại diện cho các hướng tiếp cận khác nhau:

SARIMA thuần túy: Mô hình thống kê tiêu chuẩn được mô tả trong mục 2.4.1. Đây là đại diện cho phương pháp thống kê cổ điển.

SVR với Grid Search (SVR-Grid): Mô hình SVR với kernel RBF, trong đó bộ

siêu tham số (C, γ, ϵ) được tìm kiếm bằng thuật toán tìm lưới (Grid Search) trên không gian tham số được xác định trước. Phương pháp này là cách tiếp cận tối ưu truyền thống phổ biến (Hsu và cộng sự, 2003).

Mạng LSTM (Long Short-Term Memory): Một kiến trúc mạng nơ-ron hồi tiếp (RNN) mạnh mẽ được thiết kế để học các phụ thuộc dài hạn trong chuỗi thời gian (Hochreiter & Schmidhuber, 1997).

2.5.4. Các chỉ số đánh giá

Hiệu suất dự báo được đánh giá định lượng thông qua bốn chỉ số tiêu chuẩn. Cho là giá trị thực tế, là giá trị dự báo, và là số lượng mẫu trong tập đánh giá, các chỉ số được tính như sau:

Sai số Bình phương Trung bình Căn bậc hai (Root Mean Square Error - RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (11)$$

RMSE đo lường độ lệch trung bình giữa giá trị dự báo và giá trị quan sát, có tính chất đo các sai số lớn do sử dụng bình phương (Hyndman & Koehler, 2006).

Sai số Tuyệt đối Trung bình (Mean Absolute Error - MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (12)$$

MAE cung cấp thước đo sai số trung bình đơn giản và dễ diễn giải hơn RMSE (Hyndman & Koehler, 2006).

Hệ số Xác định (Coefficient of Determination - R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (13)$$

với $\bar{y}$ là giá trị trung bình của các giá trị thực tế. R^2 đo lường tỷ lệ phương sai trong biến mục tiêu được giải thích bởi mô hình. Giá trị càng gần 1 biểu thị mức độ phù hợp càng cao (Chicco và cộng sự, 2021).

Sai số Phần trăm Tuyệt đối Trung bình (Mean Absolute Percentage Error - MAPE):

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (14)$$

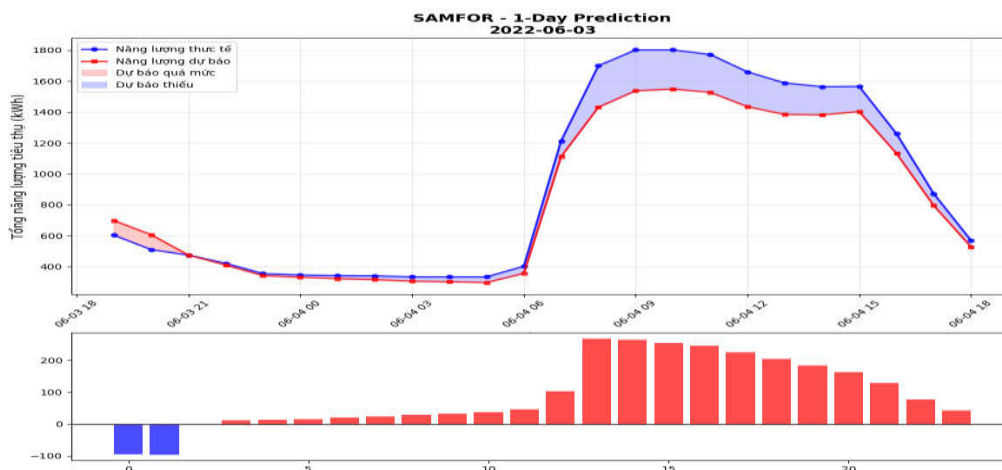
MAPE biểu thị sai số dưới dạng phần trăm, rất hữu ích để so sánh hiệu suất trên các bộ dữ liệu có quy mô khác nhau (Chicco và cộng sự, 2021).

III. Kết quả và thảo luận

3.1. Kết quả

Phân tích chi tiết dự báo trong một ngày làm việc điển hình (Hình 3) cho thấy mô hình đề xuất bám sát đường cong tải thực tế. Sai số tập trung ở một số khung giờ: 18h-21h dự báo cao hơn thực tế 100-200 kWh; 00h-09h dự báo thấp hơn 10-40 kWh, riêng 09h chênh lệch 200 kWh; từ 12h-18h sai số ổn định 20-50 kWh. Nhìn chung, mô hình thể hiện khả năng dự báo khá ổn định trong các giờ có tải trung bình và thấp, nhưng vẫn còn độ lệch đáng kể ở một số giờ cao điểm hoặc khi tải thay đổi đột ngột.

Bảng 3 dưới đây so sánh các chỉ số đánh giá giữa mô hình đề xuất và các mô hình cơ sở. Các chỉ số được tính toán dựa trên công thức tính bốn chỉ số tiêu chuẩn đã trình bày ở trên. Kết quả trong bảng 2 khẳng định một lần nữa sự vượt trội của mô hình đề xuất ngay cả trong bài toán dự báo ngắn hạn với độ phân giải cao (theo giờ). Với $R^2 = 0.948$, mô hình giải thích được gần 95% phương sai của dữ liệu thực trong ngày, một mức độ phù hợp rất cao. $MAPE = 5.12\%$ cho thấy sai số tương đối trung bình được kiểm soát rất tốt, đặc biệt có ý nghĩa khi dự báo cho các khung giờ có mức tiêu thụ thấp, nơi sai số tương đối thường bị thổi phồng.



Hình 3. Kết quả dự báo của mô hình đề xuất trong 1 ngày

Bảng 3. Tổng hợp chỉ số đánh giá cho dự báo 1 ngày

Mô hình	RMSE (kWh)	MAE (kWh)	R ²	MAPE (%)
SARIMA	412.7	325.4	0.862	9.85
SVR-Grid Search	356.2	278.9	0.888	8.12
LSTM	332.5	260.1	0.902	7.45
Đề xuất	134.1	107.3	0.948	5.12

3.2. Thảo luận

Kết quả cho thấy việc tích hợp đầu ra SARIMA làm đặc trưng cho SVR giúp SVR có đường cơ sở tin cậy, tập trung điều chỉnh phi tuyến. Tối ưu hóa bằng Firefly cho phép chọn bộ tham số kernel phù hợp, phản ứng linh hoạt ở các giờ cao điểm. So với các nghiên cứu lai SARIMA-SVR trước đây, phương pháp đề xuất cải thiện đáng kể RMSE và MAPE nhờ hai yếu tố: (i) chiến lược nội suy thông minh bảo toàn cấu trúc chuỗi khi có dữ liệu lỗi; (ii) kiến trúc lai tích hợp chặt chẽ với đầu ra SARIMA làm đặc trưng ưu tiên, thay vì chỉ dự báo phần dư. Việc sử dụng FA tự động hóa hoàn toàn quá trình tìm kiếm siêu tham số, đạt hiệu quả cao hơn Grid Search.

IV. Kết luận

Nghiên cứu đề xuất khung phương pháp tích hợp giải quyết thách thức dự báo tiêu thụ năng lượng từ dữ liệu thực tế chứa nhiều lỗi. Mô hình lai SARIMA-

SVR tối ưu bằng thuật toán Firefly đạt hiệu suất vượt trội nhờ ba đóng góp: (1) Kỹ thuật nội suy dựa trên mẫu hình thời vụ; (2) Kiến trúc lai tích hợp chặt chẽ sử dụng đầu ra SARIMA làm đặc trưng ưu tiên cho SVR; (3) Ứng dụng tối ưu hóa siêu heuristic bằng FA. Mô hình cung cấp công cụ dự báo chính xác, có thể ứng dụng trong quản lý năng lượng tòa nhà. Hướng phát triển tương lai bao gồm tích hợp biến ngoại sinh và thử nghiệm trên các bộ dữ liệu đa lĩnh vực.

Về ý nghĩa thực tiễn, mô hình cung cấp một công cụ dự báo chính xác, phù hợp với dữ liệu thực tế, có thể ứng dụng trong quản lý năng lượng tòa nhà hoặc điều độ lưới điện. Nghiên cứu cũng chỉ ra một số hạn chế, như chưa tích hợp biến ngoại sinh (thời tiết, ngày lễ), và đề xuất các hướng phát triển trong tương lai như mở rộng khung mô hình, thử nghiệm thuật toán tối ưu hóa khác, và kiểm định trên các bộ dữ liệu đa lĩnh vực.

Tài liệu tham khảo

- Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7, e623.
- Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1), 3-33.
- Díaz, S., de la Fuente, D., & Prada, M. A. (2019). A new method for time series forecasting based on intelligent hybridization of neural networks and ARIMA models. *Applied Intelligence*, 49(8), 3007-3024.
- Hamilton, J. D. (2020). *Time series analysis*. Princeton University Press.
- Han, J., & Moraga, C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International Workshop on Artificial Neural Networks* (pp. 195-201). Springer.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.
- Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). *A practical guide to support vector classification*. National Taiwan University.
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts.
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679-688.
- Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, 38(18), 2895-2907.
- Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing*, 11(2), 2664-2675.
- Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons.
- Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 ACM National Conference* (pp. 517-524).
- Shumway, R. H., & Stoffer, D. S. (2017). *Time series analysis and its applications: With R examples* (4th ed.). Springer.
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199-222.
- Yang, X. S. (2009). Firefly algorithms for multimodal optimization. In *International Symposium on Stochastic Algorithms* (pp. 169-178). Springer.
- Yang, X. S. (2010). Firefly algorithm. *Nature-Inspired Metaheuristic Algorithms*.
- Zhang, G. P. (2003). Time series forecasting using hybrid ARIMA and neural networks. *Neurocomputing*.

AN OPTIMIZED HYBRID SARIMA-SVR MODEL USING FIREFLY ALGORITHM WITH ENHANCED INTERPOLATION FOR BUILDING ENERGY CONSUMPTION FORECASTING

Nguyen Van Minh Duc¹, Tran Van Thinh¹, Hoang Anh²

Abstract: *This study proposes a comprehensive methodological framework to enhance the forecasting accuracy of energy consumption from real-world data containing numerous errors. The core is a hybrid SARIMA-SVR (Seasonal AutoRegressive Integrated Moving Average - Support Vector Regression) model with its hyperparameters optimized using the Firefly Algorithm (FA). The method offers three main contributions: (1) An interpolation technique based on seasonal patterns (hourly, daily, seasonal) to recover missing or erroneous data, ensuring input quality; (2) A tight hybrid architecture, utilizing the trend and seasonal forecasts from the SARIMA model as priority input features for the SVR model, allowing SVR to focus on learning complex nonlinear components; (3) Application of the FA to automatically find the optimal parameter set for SVR, proving more efficient than traditional methods. Experimental results on real-world datasets demonstrate that the proposed hybrid model achieves significantly higher accuracy than baseline models, including standalone SARIMA, standalone SVR, and LSTM (Long Short-Term Memory). The research not only provides a feasible forecasting tool for energy management systems but also contributes an integrated methodology combining data preprocessing, hybrid modeling, and optimization. Future research directions include integrating exogenous variables and testing on datasets from various domains.*

Keywords: *energy consumption forecasting, hybrid SARIMA-SVR model, firefly algorithm, enhanced data interpolation, erroneous data handling*

¹ Faculty of Electrical and Electronic Engineering, Hanoi Open University, Hanoi, Vietnam

² School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam