

NGHIÊN CỨU TỐI ƯU KIẾN TRÚC MẠNG NƠ RON VỚI THUẬT TOÁN TOÁN DI TRUYỀN NSGA-II, ỨNG DỤNG TRONG BÀI TOÁN DỰ ĐOÁN ĐỘ NHÁM BỀ MẶT SAU KHI TIỆN CNC VỚI THÉP SUM24L

A RESEARCH ON NEURAL NETWORK ARCHITECTURE OPTIMIZATION USING THE NSGA-II GENETIC ALGORITHM FOR SURFACE ROUGHNESS PREDICTION IN CNC TURNING OF SUM24L STEEL

TS. Nguyễn Duy Anh, ThS. Nguyễn Thị Hiếu Thảo*

Trường Đại học Giao thông vận tải

*Email: hieuthaont@utc.edu.vn

TÓM TẮT

Độ nhám bề mặt sau khi gia công là yếu tố then chốt ảnh hưởng đến độ bền, ma sát và tính năng vận hành của chi tiết cơ khí. Trong bối cảnh gia công thép SUM24L bằng phương pháp tiện, việc dự đoán chính xác độ nhám bề mặt từ các thông số cắt như tốc độ (v), chiều sâu (t) và lượng chạy dao (s) là cần thiết để tối ưu hóa quy trình và nâng cao chất lượng sản phẩm. Tuy nhiên, các mô hình dự đoán hiện tại thường thiếu cân bằng giữa độ chính xác và khả năng triển khai thực tế trong môi trường sản xuất thông minh. Bài báo này đề xuất một phương pháp tối ưu hóa đa mục tiêu dựa trên NSGA-II nhằm tự động thiết kế các mô hình mạng nơ-ron nhân tạo (ANN) cho bài toán dự đoán độ nhám sau tiện thép SUM24L. Hai mục tiêu được xem xét đồng thời là: (1) sai số dự đoán (MSE) và (2) kích thước mô hình sau lượng tử hóa (bytes). Thử nghiệm cho thấy phương pháp đã tạo ra một tập nghiệm Pareto hiệu quả, cung cấp các phương án mô hình ANN linh hoạt, sẵn sàng triển khai trên thiết bị biên với độ chính xác phù hợp yêu cầu công nghiệp.

Từ khóa: Độ nhám bề mặt; Mạng Nơ-ron nhân tạo; Tối ưu hóa đa mục tiêu; Thiết bị biên.

ABSTRACT

Surface roughness after machining is a key factor affecting the durability, friction, and operational performance of mechanical parts. In the context of turning SUM24L steel, the accurate prediction of surface roughness from cutting parameters such as speed (v), depth of cut (t), and feed rate (s) is necessary to optimize the process and enhance product quality. However, current prediction models often lack a balance between accuracy and practical deployment capability in smart manufacturing environments. This paper proposes a multi-objective optimization method based on NSGA-II to automatically design Artificial Neural Network (ANN) models for the problem of predicting surface roughness after turning SUM24L steel. Two objectives are considered simultaneously: (1) prediction error (MSE) and (2) post-quantization model size (bytes). Experiments show that the method has generated an effective Pareto front, providing flexible ANN model options, ready for deployment on edge devices with an accuracy suitable for industrial requirements.

Keywords: Surface roughness; Artificial Neural Network; Multi-objective Optimization; Edge devices. 

1. ĐẶT VẤN ĐỀ

Trong lĩnh vực cơ khí chế tạo, chất lượng bề mặt là một chỉ tiêu quan trọng, phản ánh trực tiếp mức độ hoàn thiện của quá trình gia công. Đặc biệt, độ nhám bề mặt (Ra) có ảnh hưởng lớn đến các đặc tính chức năng của chi tiết như khả năng chống mài mòn, chống ăn mòn, hệ số ma sát và độ bền mỏi. Việc kiểm soát độ nhám là yêu cầu thiết yếu để đảm bảo hiệu năng và tuổi thọ của sản phẩm, đặc biệt trong các ứng dụng yêu cầu độ chính xác cao.

Tuy nhiên, độ nhám sau gia công chịu ảnh hưởng bởi nhiều thông số công nghệ như tốc độ cắt (v), chiều sâu cắt (t) và lượng chạy dao (s), với mối quan hệ phi tuyến và khó mô hình hóa tường minh. Do đó, khả năng dự đoán trước độ nhám từ các thông số cắt không chỉ giúp rút ngắn thời gian thử nghiệm mà còn tiết kiệm vật liệu và chi phí sản xuất.

Nhiều phương pháp đã được đề xuất để giải quyết bài toán này, từ mô hình thực nghiệm tuyến tính đến phương pháp bề mặt đáp ứng (Response Surface Method – RSM). Gần đây, các kỹ thuật học máy – đặc biệt là mạng nơ-ron nhân tạo (Artificial Neural Network – ANN) – đã được áp dụng và cho thấy nhiều tiềm năng nhờ khả năng mô hình hóa các quan hệ phi tuyến phức tạp. Tuy nhiên, các mô hình ANN đạt độ chính xác cao thường có kích thước lớn, đòi hỏi nhiều tài nguyên tính toán, dẫn đến khó khăn trong việc triển khai thực tế trên các thiết bị giám sát phân tán hoặc hệ thống điều khiển tại xưởng, vốn yêu cầu độ trễ thấp và dung lượng nhỏ.

Bài báo này đề xuất một phương pháp mới nhằm giải quyết đồng thời hai mục tiêu quan trọng trong thiết kế mô hình ANN cho bài toán dự đoán độ nhám bề mặt sau khi tiện

thép SUM24L: (1) giảm sai số dự đoán và (2) giảm kích thước mô hình sau khi áp dụng kỹ thuật lượng tử hóa hậu huấn luyện (Post-Training Quantization). Phương pháp được xây dựng dựa trên thuật toán tối ưu hóa đa mục tiêu NSGA-II, cho phép tự động tìm kiếm tập hợp các kiến trúc ANN Pareto tối ưu về cả hai khía cạnh nêu trên. Đây là hướng tiếp cận phù hợp với yêu cầu triển khai thực tế trong môi trường sản xuất thông minh theo định hướng Công nghiệp 4.0. Cấu trúc của bài báo như sau: Phần 2 trình bày tổng quan các công trình liên quan. Phần 3 mô tả quy trình tối ưu hóa đề xuất. Phần 4 trình bày kết quả thực nghiệm và phân tích. Cuối cùng, Phần 5 đưa ra kết luận và hướng phát triển trong tương lai.

2. TỔNG QUAN TÌNH HÌNH NGHIÊN CỨU

2.1. Các mô hình dự đoán độ nhám bề mặt

Dự đoán độ nhám bề mặt từ các thông số cắt là một chủ đề được nghiên cứu rộng rãi trong lĩnh vực gia công cơ khí. Các phương pháp thống kê truyền thống như RSM và phương pháp Taguchi đã được sử dụng để xây dựng mô hình thực nghiệm giữa các biến đầu vào và đầu ra. Mặc dù có ưu điểm về tính đơn giản và khả năng giải thích, các mô hình này thường giả định tuyến tính hoặc gần tuyến tính, dẫn đến hạn chế trong việc mô hình hóa các mối quan hệ phi tuyến đặc trưng trong quá trình cắt gọt kim loại.

Để khắc phục điểm yếu này, các kỹ thuật học máy và trí tuệ nhân tạo đã được ứng dụng nhằm nâng cao khả năng dự đoán. Trong số đó, các mô hình như mạng ANN, máy vector hỗ trợ SVM, và hệ suy luận thích nghi neuro-mờ (Adaptive Neuro-Fuzzy Inference System – ANFIS) [1] đã cho thấy hiệu quả vượt trội. Trong các nghiên cứu gần đây, ANN nổi lên

như một công cụ phổ biến do khả năng học phi tuyến mạnh mẽ, đặc biệt phù hợp với dữ liệu gia công có tính nhiễu và biến động cao [2].

2.2. Tối ưu hóa đa mục tiêu và NSGA-II

Trong nhiều bài toán kỹ thuật, bao gồm cả thiết kế mô hình dự đoán, việc tối ưu đồng thời nhiều mục tiêu – chẳng hạn như độ chính xác và độ phức tạp của mô hình – là yêu cầu phổ biến. Tối ưu hóa đa mục tiêu (Multi-Objective Optimization – MOO) là cách tiếp cận cho phép xử lý các mục tiêu mâu thuẫn bằng cách tìm ra một tập nghiệm không bị chi phối lẫn nhau, tạo nên mặt phẳng Pareto tối ưu.

Trong số các thuật toán tiến hóa phục vụ MOO, NSGA-II (Non-dominated Sorting Genetic Algorithm II) là một trong những thuật toán nổi bật nhờ khả năng hội tụ nhanh, phân tán nghiệm tốt trên biên Pareto và chi phí tính toán hợp lý [3]. NSGA-II đã được áp dụng hiệu quả trong nhiều lĩnh vực kỹ thuật như thiết kế mạch, điều khiển hệ thống và tối ưu kiến trúc mạng nơ-ron.

2.3. Nén mô hình và lượng tử hóa cho thực thi trí tuệ nhân tạo trên các thiết bị biên

Trong bối cảnh sản xuất thông minh, các mô hình học máy không chỉ cần chính xác mà còn phải đủ nhẹ để triển khai trên các thiết bị biên hoặc các hệ thống giám sát nhúng. Việc nén mô hình trở thành một yêu cầu thiết yếu nhằm giảm độ trễ, tiết kiệm năng lượng và giảm chi phí phần cứng.

Trong số các kỹ thuật nén, lượng tử hóa là một phương pháp phổ biến và hiệu quả, cho phép chuyển đổi mô hình từ định dạng số thực 32-bit sang các định dạng số nguyên 16-bit hoặc 8-bit mà không làm suy giảm đáng kể độ

chính xác. Điều này giúp giảm kích thước mô hình và tăng tốc độ suy luận [4]. Các nền tảng như TensorFlow Lite đã hỗ trợ mạnh mẽ cho lượng tử hóa sau huấn luyện, giúp triển khai dễ dàng các mô hình nén trên thiết bị thực.

2.4. Đóng góp của nghiên cứu

Mặc dù đã có các nghiên cứu riêng lẻ ứng dụng ANN trong dự đoán độ nhám bề mặt, hoặc sử dụng NSGA-II để tối ưu các tham số mô hình, hiện chưa có công trình nào kết hợp một cách hệ thống hai hướng tiếp cận này để giải quyết bài toán tối ưu đồng thời độ chính xác và kích thước mô hình sau lượng tử hóa, đặc biệt trong bối cảnh gia công thép SUM24L. Nghiên cứu đề xuất một quy trình tự động thiết kế và nén các mô hình ANN dựa trên NSGA-II, từ đó tạo ra các mô hình Pareto tối ưu có thể triển khai trực tiếp trong hệ thống sản xuất thông minh.

3. PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Thu thập và tiền xử lý dữ liệu

Bộ dữ liệu sử dụng trong nghiên cứu được thu thập từ quá trình tiện thực nghiệm thép SUM24L (chi tiết đã được trình bày trong công trình trước đây của nhóm tác giả). Tập dữ liệu bao gồm 27 mẫu, mỗi mẫu tương ứng với một tổ hợp thông số cắt và kết quả đo độ nhám bề mặt tại 5 vị trí khác nhau. Cụ thể:

- Đầu vào: Ba thông số công nghệ gồm tốc độ cắt v (m/min), chiều sâu cắt t (mm) và lượng chạy dao s (mm/rev).
- Đầu ra: Năm giá trị độ nhám bề mặt $Ra1$ đến $Ra5$ (μm).
- Dữ liệu được chia thành hai tập: 80% cho huấn luyện và 20% cho kiểm thử. Các đặc trưng đầu vào được chuẩn hóa và đưa về dạng

phân phối chuẩn, giúp tăng độ ổn định trong huấn luyện mạng nơ-ron.

3.2. Quy trình tối ưu hóa đề xuất

Quy trình tối ưu hóa là bài toán tối ưu hóa đa mục tiêu, và được giải quyết bằng thuật toán NSGA-II. Mỗi cá thể trong quần thể được biểu diễn bằng một vector số thực, mã hóa cho cá thể. Mỗi cá thể bao gồm cấu trúc mạng nơ-ron (như số lớp ẩn, số nơ-ron trên mỗi lớp, loại hàm kích hoạt) và chiến lược lượng tử hóa tương ứng (ví dụ: số bit lượng tử hóa, kỹ thuật ánh xạ). Quần thể ban đầu được khởi tạo ngẫu nhiên và sau đó tiến hóa qua các thế hệ theo cơ chế chọn lọc tự nhiên. Trong mỗi thế hệ, các cá thể được giải mã để xây dựng mô hình ANN cụ thể, mô hình này được huấn luyện trên tập dữ liệu huấn luyện và tiếp tục được lượng tử hóa bằng kỹ thuật lượng tử hóa sau huấn luyện sử dụng TensorFlow Lite.

Sau khi lượng tử hóa, hiệu năng của từng cá thể được đánh giá dựa trên hai mục tiêu cạnh tranh: độ chính xác dự đoán và độ nhỏ gọn của mô hình. Độ chính xác được định lượng thông qua Sai số Trung bình bình phương (Mean Squared Error – MSE) trên tập dữ liệu kiểm thử độc lập, trong khi độ nhỏ gọn được đo bằng kích thước của tệp mô hình sau lượng tử hóa. Các giá trị mục tiêu của toàn bộ quần thể được truyền lại cho thuật toán NSGA-II, nơi chúng được xử lý thông qua cơ chế sắp xếp không chi phối và tính toán khoảng cách chen chúc để chọn lọc cá thể cha mẹ. Các cá thể con sau đó được tạo ra bằng các toán tử lai ghép và đột biến. Quá trình lặp tiếp tục cho đến khi đạt đến điều kiện dừng định trước, chẳng hạn như số thế hệ tối đa hoặc sự hội tụ của tập nghiệm Pareto.

3.2.1. Định nghĩa bài toán tối ưu

Bài toán tối ưu hóa trong nghiên cứu này được phát biểu dưới dạng tối ưu hóa đa mục tiêu với hai hàm mục tiêu cần được đồng thời cực tiểu hóa. Cụ thể, bài toán có dạng:

$$\min_x F(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x})]$$

Trong đó, $\mathbf{x}$ là một vector quyết định, đại diện cho một cấu hình tiềm năng của mô hình mạng nơ-ron nhân tạo (ANN) và chiến lược lượng tử hóa tương ứng. Hàm mục tiêu $F(\mathbf{x})$ bao gồm hai thành phần: $f_1(\mathbf{x})$ là sai số trung bình bình phương (Mean Squared Error – MSE) của mô hình trên tập dữ liệu kiểm thử, phản ánh mức độ chính xác trong dự đoán.

- $f_2(\mathbf{x})$ là kích thước mô hình sau lượng tử hóa, được đo bằng số byte, biểu thị độ nhỏ gọn của mô hình và mức tiêu tốn bộ nhớ lưu trữ.

Không gian tìm kiếm của bài toán được xác định bởi các giá trị khả dĩ của vector quyết định ($\mathbf{x}$) bao gồm: (i) các siêu tham số kiến trúc của mạng nơ-ron, cụ thể là số lượng lớp ẩn (giới hạn từ 1 đến 3 lớp) và số lượng nơ-ron trong mỗi lớp (chọn từ các giá trị trong khoảng từ 8 đến 128, là bội số của 8); và (ii) mức lượng tử hóa, được lựa chọn từ ba mức độ độ chính xác: 8-bit, 16-bit hoặc 32-bit. Các ràng buộc này tạo thành một không gian tìm kiếm rời rạc nhưng có thể mã hóa hiệu quả trong khuôn khổ của thuật toán tiến hóa. Mục tiêu cuối cùng là xác định tập hợp các nghiệm không chi phối (Pareto-optimal solutions), thể hiện sự đánh đổi hợp lý giữa độ chính xác dự đoán và độ nhỏ gọn của mô hình.

3.2.2. Cấu hình thuật toán NSGA-II

Thuật toán NSGA-II trong nghiên cứu được triển khai bằng thư viện pymoo, một công

cụ tối ưu hóa đa mục tiêu mã nguồn mở mạnh mẽ và linh hoạt. Cấu hình thuật toán được thiết lập nhằm đảm bảo khả năng khám phá không gian tìm kiếm hiệu quả và đạt được tập nghiệm Pareto ổn định sau số thế hệ giới hạn.

Cụ thể, kích thước quần thể được đặt là 30 cá thể, nhằm duy trì đa dạng di truyền trong khi vẫn đảm bảo chi phí tính toán hợp lý. Quá trình tiến hóa được thực hiện trong 25 thế hệ, đủ để cho phép thuật toán hội tụ đến một tập nghiệm không chi phối có chất lượng.

Trong quá trình tạo sinh thế hệ mới, toán tử lai ghép được sử dụng là Simulated Binary Crossover (SBX) với hệ số phân phối $\eta = 15$ và xác suất lai ghép là 0.9, nhằm khai thác hiệu quả thông tin từ các cá thể cha mẹ. Toán tử đột biến được áp dụng dưới dạng Polynomial Mutation với hệ số $\eta = 15$, và xác suất đột biến được thiết lập tự động theo số lượng biến trong vector quyết định, đảm bảo mức độ nhiễu vừa đủ để duy trì khả năng khám phá.

Chiến lược chọn lọc cá thể dựa trên cơ chế chọn lọc giải đấu (Tournament Selection) kết hợp với thuật toán sắp xếp không chi phối và tính khoảng cách chen chúc của NSGA-II tiêu chuẩn. Cấu hình này cho phép thuật toán duy trì cân bằng giữa khả năng khai thác (exploitation) và khám phá (exploration) trong quá trình tối ưu hóa đa mục tiêu.

3.2.3. Huấn luyện và lượng tử hóa mạng nơ-ron

Quá trình huấn luyện mạng nơ-ron được thực hiện bằng thư viện Keras, sử dụng thuật toán tối ưu Adam và hàm mất mát là sai số trung bình bình phương (Mean Squared Error – MSE). Để ngăn ngừa hiện tượng quá khớp chiến lược dừng sớm được áp dụng, theo

đôi sai số trên tập kiểm thử và tự động dừng huấn luyện khi hiệu suất không còn cải thiện sau một số epoch liên tiếp.

Sau khi hoàn tất huấn luyện, mô hình được chuyển đổi sang định dạng TensorFlow Lite (TFLite) bằng công cụ TFLiteConverter. Quá trình chuyển đổi này đồng thời áp dụng kỹ thuật lượng tử hóa sau huấn luyện (post-training quantization) với ba mức độ được hỗ trợ: 8-bit integer, 16-bit float, và giữ nguyên độ chính xác 32-bit float. Các mức lượng tử hóa này tạo điều kiện thuận lợi cho việc đánh đổi giữa độ chính xác của mô hình và độ nhỏ gọn khi triển khai trên các thiết bị đầu cuối có tài nguyên hạn chế.

4. KẾT QUẢ DỰ ĐOÁN

4.1. Kết quả về các giải pháp Pareto tối ưu

Hình 1 trình bày biểu đồ Pareto Front hai chiều thể hiện mối quan hệ giữa kích thước mô hình (tính theo kilobyte, trục hoành) và sai số dự đoán MSE (trục tung). Mỗi điểm trên đồ thị biểu diễn một giải pháp nằm trên biên Pareto – nghĩa là không tồn tại giải pháp nào khác có thể đồng thời cải thiện cả độ chính xác và mức tiêu tốn tài nguyên.



Hình 1. Biểu đồ Pareto Front 2D: Sai số dự đoán (MSE) so với kích thước mô hình (KB)

Nhằm làm rõ hơn đặc điểm các chiến lược thiết kế, Bảng 1 tổng hợp ba giải pháp đại diện cho ba khuynh hướng đối lập: mô hình có độ chính xác cao nhất, mô hình nhỏ gọn nhất, và một giải pháp cân bằng hợp lý giữa hai tiêu chí.

Bảng 1. Tóm tắt các giải pháp Pareto-tối ưu tiêu biểu

Giải pháp	MSE	Kích thước (KB)	Kiến trúc (số nơ-ron mỗi lớp)	Số tham số
Chính xác nhất	0.035496	6.52	[40, 16, 40]	293
Gọn nhẹ nhất	0.958995	2.19	[8]	29
Cân bằng	0.240387	3.34	[8, 24]	101

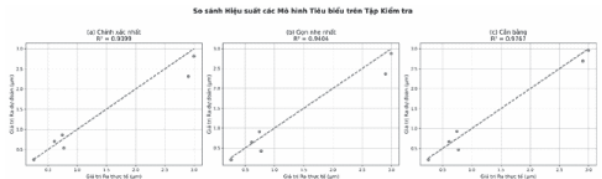
Kết quả trên Hình 1 phản ánh rõ rệt đặc trưng đánh đổi giữa độ chính xác và độ phức tạp mô hình. Cụ thể, mô hình chính xác nhất có sai số thấp nhất ($MSE = 0.035$) nhưng đòi hỏi kích thước bộ nhớ lớn hơn ba lần so với mô hình nhỏ gọn nhất. Ngược lại, mô hình gọn nhẹ nhất chỉ tiêu tốn 2.19 KB, nhưng độ chính xác suy giảm đáng kể với MSE lên đến gần 1.0. Giải pháp cân bằng có sai số vừa phải ($MSE = 0.240$) trong khi vẫn duy trì kích thước ở mức tiết kiệm (3.34 KB), cho thấy đây là một phương án thực tiễn cho các hệ thống nhúng hoặc thiết bị có hạn chế về tài nguyên phần cứng.

Từ các kết quả trên, có thể khẳng định rằng không tồn tại một mô hình “tốt nhất tuyệt đối”, mà việc lựa chọn kiến trúc phù hợp phụ thuộc vào bối cảnh ứng dụng: ưu tiên độ chính xác (ví dụ, hệ thống kiểm tra chất lượng trong công nghiệp) hay ưu tiên dung lượng bộ nhớ và tốc độ (ví dụ, thiết bị IoT biên hoặc cảm biến thông minh).

4.2. Kết quả phân tích dự đoán với các mô hình tiêu biểu

Để đánh giá sâu hơn hiệu năng dự đoán của các giải pháp Pareto-tối ưu, ba mô hình tiêu biểu được chọn gồm: (A) Chính xác nhất, (B) Nhỏ gọn nhất, và (C) Cân bằng nhất.

Mỗi mô hình được huấn luyện và áp dụng để dự đoán giá trị độ nhám bề mặt Ra trên tập kiểm tra (test set). Kết quả được trực quan hóa dưới dạng biểu đồ Dự đoán so với Thực tế. Đây là một công cụ trực quan quan trọng để đánh giá khả năng khái quát hóa và độ chính xác tuyệt đối của mô hình. Trên mỗi biểu đồ, hệ số xác định R^2 được tính và hiển thị để đo lường mức độ khớp giữa đầu ra mô hình và dữ liệu thật.



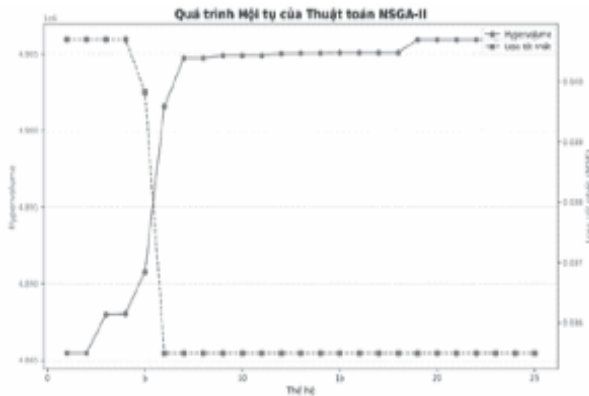
Hình 2. Biểu đồ dự đoán với thực tế trên tập kiểm tra cho ba mô hình tiêu biểu.

Ba biểu đồ trong Hình 2 cho thấy rõ sự đánh đổi giữa độ chính xác và độ phức tạp của mô hình. Mô hình (A) có hiệu năng tốt nhất nhưng chi phí tài nguyên cao, mô hình (B) cực kỳ nhỏ gọn nhưng chính xác vừa phải, trong khi mô hình (C) cung cấp giải pháp dung hòa với hiệu năng tốt và tài nguyên hợp lý. Lựa chọn giữa ba mô hình cần dựa vào ràng buộc cụ thể của từng ứng dụng – ví dụ, hệ thống biên yêu cầu nhẹ, hay hệ thống kiểm tra chất lượng yêu cầu độ chính xác cao.

4.3. Đánh giá quá trình hội tụ của thuật toán

Hình 3 trình bày tiến trình hội tụ của thuật toán NSGA-II thông qua ba chỉ số đặc

trung được ghi nhận trong suốt quá trình tiến hóa: giá trị Hypervolume, MSE tối ưu và số lượng cá thể nằm trên Pareto front tại mỗi thế hệ.



Hình 3. Quá trình hội tụ của thuật toán NSGA-II

Kết quả cho thấy Hypervolume tăng đều theo thời gian, phản ánh sự cải thiện liên tục về cả chất lượng lẫn độ đa dạng của tập nghiệm Pareto. Mức tăng nhanh trong giai đoạn đầu cho thấy thuật toán khởi đầu hiệu quả và nhanh chóng xác định được các vùng tiềm năng trong không gian tìm kiếm. Song song đó, MSE tối ưu có xu hướng giảm dần, chứng tỏ thuật toán dần tiếp cận các nghiệm có sai số dự đoán thấp hơn. Quá trình giảm diễn ra ổn định, đặc biệt rõ rệt trong các thế hệ đầu, trước khi hội tụ dần về giá trị tiệm cận.

Tổng thể, Hình 3 cung cấp bằng chứng thực nghiệm về khả năng tối ưu hóa bền vững và hiệu quả của thuật toán NSGA-II. Việc đồng thời cải thiện ba chỉ số trên trong suốt quá trình tiến hóa là dấu hiệu cho thấy sự cân bằng hợp lý giữa hai yếu tố cốt lõi trong thuật toán tiến hóa: khai phá không gian nghiệm và khai thác các vùng tối ưu cục bộ. Điều này củng cố độ tin cậy của tập nghiệm Pareto thu được và là cơ sở cho các phân tích chi tiết đã trình bày trong các mục trước.

5. KẾT LUẬN

Nghiên cứu này đã trình bày một phương pháp luận hệ thống nhằm tự động hóa quá trình thiết kế các mô hình mạng nơ-ron nhân tạo (ANN) tối ưu cho bài toán dự đoán độ nhám bề mặt. Bằng cách mô hình hóa bài toán như một bài toán tối ưu hóa đa mục tiêu và áp dụng thuật toán NSGA-II, quy trình đề xuất cho phép đồng thời tìm kiếm các cấu hình mạng hiệu quả cả về độ chính xác và độ nhỏ gọn, phù hợp với các yêu cầu triển khai trong thực tế.

Các kết quả thực nghiệm cho thấy rằng phương pháp đề xuất đã thành công trong việc tạo ra một tập hợp các nghiệm Pareto-tối ưu, thể hiện rõ ràng mối quan hệ đánh đổi giữa sai số dự đoán và kích thước mô hình sau lượng tử hóa. Ngoài ra, một số cấu hình tiêu biểu đã được xác định, đại diện cho các lựa chọn khả thi cho các mục tiêu ứng dụng khác nhau: từ độ chính xác cao đến triển khai gọn nhẹ trên thiết bị nhúng.

Về mặt thực tiễn, phương pháp này cung cấp một công cụ hữu ích và có khả năng mở rộng cho các kỹ sư và nhà nghiên cứu trong lĩnh vực sản xuất thông minh. Việc tích hợp quy trình tối ưu hóa tự động với lượng tử hóa mô hình giúp rút ngắn thời gian phát triển, đồng thời đảm bảo hiệu suất mô hình phù hợp với các ràng buộc phần cứng trong môi trường công nghiệp.

Tuy nhiên, nghiên cứu vẫn còn tồn tại một số hạn chế. Bộ dữ liệu sử dụng trong huấn luyện có quy mô tương đối nhỏ và chỉ bao gồm ba thông số đầu vào cơ bản, chưa phản ánh đầy đủ các yếu tố ảnh hưởng trong quá trình gia công. Trong các nghiên cứu tiếp theo, có thể mở rộng bộ dữ liệu với nhiều điều kiện vận

hành hơn (như độ mòn dao, mức rung động), đồng thời khám phá các kiến trúc mạng tiên tiến hơn như LSTM hoặc Transformer để cải thiện khả năng mô hình hóa. Bên cạnh đó, việc đánh giá hiệu năng triển khai thực tế của các mô hình đã lượng tử hóa trên phần cứng cụ thể (ví dụ: thời gian suy luận, mức tiêu thụ năng lượng) cũng là một hướng phát triển quan trọng nhằm đảm bảo tính khả thi khi ứng dụng trong môi trường sản xuất thực.

Lời cảm ơn:

Nghiên cứu này được tài trợ bởi Trường Đại học Giao thông vận tải (ĐH GTVT) trong đề tài mã số T2025-DT-007.

(*) Tác giả liên hệ: Nguyễn Thị Hiếu Thảo (email: hieuthaont@utc.edu.vn)❖

Ngày nhận bài: **23/6/2025**

Ngày phản biện: **07/7/2025**

Tài liệu tham khảo:

- [1]. Asiltürk, İlhan, and Mehmet Çunkaş, “*Modeling and prediction of surface roughness in turning operations using artificial neural network and multiple regression method*”. Expert systems with applications 38.5 (2011): 5826-5832.
- [2]. Zain, Azlan Mohd, Habibollah Haron, and Safian Sharif, “*Prediction of surface roughness in the end milling machining using Artificial Neural Network*”. Expert Systems with Applications 37.2 (2010): 1755-1768.
- [3]. Deb, Kalyanmoy, et al., “*A fast and elitist multiobjective genetic algorithm: NSGA-II*”. IEEE transactions on evolutionary computation 6.2 (2002): 182-197.
- [4]. Jacob, Benoit, et al., “*Quantization and training of neural networks for efficient integer-arithmetic-only inference*”. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.