УДК 004.942
DOI: 10.32626/2308-5916.2024-26.54-63

О. О. Фомін, д-р техн. наук, професор,
В. О. Канєвський, аспірант,
Д. С. Мельник, аспірант,
А. В. Бурбенко
Національний університет «Одеська політехніка», м. Одеса

ОПТИМІЗАЦІЯ АРХІТЕКТУРИ НЕЙРОННИХ МЕРЕЖ З УРАХУВАННЯМ АУГМЕНТАЦІЇ ДАНИХ

Робота присвячена вирішенню протиріччя між підвищенням стійкості моделі до завад і спотворень та ускладненням задачі навчання моделі в умовах обмежених обчислювальних ресурсів. Метою роботи є визначення архітектури моделей нелінійної динаміки в умовах обмежених навчальних даних при забезпеченні заданої точності моделювання. Ця мета досягається шляхом розвитку методу підбору архітектури нейронних мереж NAS. Наукова новизна роботи полягає у подальшому розвитку методу підбору архітектури нейронної мережі NAS при ідентифікації нелінійних динамічних об'єктів з врахуванням спотворень навчального датасету шляхом додавання аугментованих даних. На відміну від традиційного підходу до попереднього навчання, розроблений метод дозволяє будувати більш стійкі до дії завад моделі при забезпеченні необхідної точності. Практична користь роботи полягає в розвитку підходу до адаптації архітектури в залежності від методів аугментації, що використовуються, шляхом розробки алгоритму методу підбору архітектури нейронної мережі NAS з врахуванням аугментації даних, що дозволяє будувати більш надійні моделі без втрати точності моделювання. Наведено результати експериментів з моделювання тестових об'єктів з нелінійними динамічними характеристиками, проаналізовано вплив аугментації даних на якість і стійкість отриманих моделей. Цінність проведеного дослідження полягає у визначенні області ефективного використання запропонованого методу, як задач з нестачею розмічених даних при відсутності суворих вимог до швидкості процесу моделювання.

Ключові слова: *нелінійна динаміка, нейронні мережі, аугментація даних, оптимізація архітектури, автоматизація навчання.*

1. Вступ. Сучасні технології моделювання, засновані на машинному навчанні, стали невід'ємною частиною сучасних рішень у найрізноманітніших галузях – від медицини та фінансів до промисловості

та IoT. Ці технології дають змогу моделювати складні системи, аналізувати великі обсяги даних і робити прогнози з високою точністю. Особлива увага приділяється завданням нелінійної динаміки, таким як обробка сигналів, аналіз часових рядів і прогнозування [1, 2]. У цих завданнях моделі, що працюють із затримками та послідовностями даних, набувають дедалі більшого значення.

Однією з основних причин успіху сучасних інтелектуальних систем є доступ до великих і якісних наборів даних. Що більше даних є доступним для навчання, то точнішою і стійкішою стає модель. Достатній обсяг даних дає змогу моделі ефективно витягувати ключові ознаки, враховувати різні варіації та уникати перенавчання. У таких галузях, як комп'ютерний зір або оброблення природної мови, успіхи в застосуванні нейронних мереж пояснюються наявністю масштабних датасетів, як-от ImageNet або Common Crawl, а також простотою збирання даних [2]. Однак, у низці прикладних завдань, пов'язаних із часовими рядами та сигналами, отримати достатню кількість даних буває важко. Однак у низці прикладних завдань, пов'язаних із часовими рядами та сигналами, отримати достатню кількість даних буває важко. Це може бути пов'язано з високою вартістю збирання даних, складністю їх анотування або обмеженим доступом до рідкісних подій. Брак даних призводить до зниження точності моделей, їх недостатньої генералізації та схильності до перенавчання [2]. Особливо це критично в задачах прогнозування та аналізу сигналів, де дані можуть бути унікальними і неповторюваними.

Одним із найефективніших методів подолання нестачі даних є аугментація [3]. Вона дає змогу штучно збільшувати обсяг доступних даних, створюючи нові приклади на основі наявних, урізноманітнювати дані та допомагати моделям ефективно справлятися з мінливістю реальних умов. Для часових даних аугментація включає такі підходи, як тимчасові зрушення, додавання шумів, зміна частотних характеристик або комбінування декількох сигналів. Ці методи дають змогу моделям стати стійкішими до мінливості даних і поліпшити їхню здатність узагальнювати. Аугментація довела свою ефективність у завданнях оброблення мовлення, аналізу біосигналів і фінансових часових рядів в умовах обмеженого обсягу навчальної вибірки [3].

Однак, незважаючи на очевидні переваги, використання аугментації може мати значний вплив на продуктивність моделей. Наприклад, додавання шуму, масштабування, часових зсувів або інверсії сигналу вимагає, щоб модель була більш гнучкою та могла витягувати інваріантні ознаки [4]. Тому підхід, заснований на аугментації, може привести до ряду протиріч, які необхідно враховувати при проектуванні моделей. Так, проблема використання аугментації полягає

в порушенні семантичної значущості даних і балансу класів, що призводить до зниження точності моделей.

Основним протиріччям при використанні аугментації є баланс між посиленням моделі і підвищенням її складності [3]. З одного боку, аугментація урізноманітнює тренувальні дані, роблячи модель більш стійкою до шумів і спотворень реального світу. Наприклад, моделі для аналізу мовних сигналів можуть краще обробляти дані, що містять різні акценти або фонові шуми, якщо в тренувальному процесі використовувалася відповідна аугментація. З іншого боку, додавання нових варіантів даних ускладнює тренувальне завдання. Це може призвести до повільнішої конвергенції або навіть перенавчання, якщо архітектура моделі недостатньо оптимізована для роботи з доповненими даними.

Це особливо важливо в задачах, де сигнали мають складну часову структуру і залежать від контексту, як, наприклад, у завданнях прогнозування часових рядів або аналізу динамічних систем.

Таким чином, ключове завдання полягає в тому, щоб адаптувати архітектуру нейронної мережі таким чином, щоб вона могла ефективно обробляти підвищену складність даних, зберігаючи при цьому здатність виділяти ключові особливості. Для цього потрібна оптимізація параметрів архітектури моделі, таких як глибина мережі, кількість нейронів і структура зв'язків, так і продуманий вибір методів аугментації, адаптованих до характеристик часових сигналів.

2. Огляд літератури. Останніми роками проблема отримання даних і зниження їх вартості набула широкого поширення серед дослідників [1, 3]. Методи аугментації даних являють собою ключові інструменти, що дають змогу поліпшити якість навчання моделей, особливо в умовах обмеженого обсягу даних. При цьому одночасно розвивається кілька підходів аугментації [3-6]:

- *геометричні трансформації*: повороти, відображення, масштабування та обрізання даних. Часто застосовуються в обробці зображень [3];
- *додавання шуму*: білий шум, гаусівський шум і випадкові викиди для підвищення стійкості моделей [3, 5];
- *комбіновані методи*: такі як Mixup і CutMix [5], які створюють нові образи шляхом змішування вихідних даних і їхніх міток;
- *автоматизовані підходи*: AutoAugment і RandAugment [3, 4], які оптимізують процес вибору аугментацій.

Ці методи демонструють високу ефективність у навчанні моделей, підвищуючи їхню здатність до узагальнення. Однак їхнє застосування до сигналів потребує адаптації. Обробка сигналів, таких як тимчасові ряди, аудіо або біосигнали, вимагає специфічних підходів

до аугментації, що враховують їхню часову і частотну структуру. Серед таких підходів найбільш затребуваними є:

- *часові перетворення*: розтягнення або стиснення за часом, тимчасові зсуви та інверсії сигналу [4];
- *частотні перетворення*: додавання шумів, фільтрація частот або зміна амплітудно-частотних характеристик сигналу [6];
- *маскування даних*: тимчасове і спектральне маскування, що застосовується для аудіосигналів (наприклад, SpecAugment для розпізнавання мови) [5];
- *змішування сигналів*: комбінування двох або більше сигналів, що дає змогу моделям краще впоратися із зашумленими даними (аналог Mixup для часових рядів) [7].

Кожен із цих методів призначено для розв'язання конкретних завдань, таких як підвищення стійкості моделей до завад або поліпшення їхньої здатності витягувати значущі ознаки. Для усунення протиріччя між посиленням моделі та збільшенням складності навчання під час опрацювання сигналів традиційно використовують такі підходи.

Регуляризація. Dropout і L2-регуляризація знижують ризик перенавчання, що виникає через надмірну складність даних після аугментації.

Адаптивні методи навчання. Використання методів контролю навчання, таких як рання зупинка або зміна швидкості навчання залежно від складності даних, допомагає справлятися зі збільшеною варіативністю.

Оптимізація архітектури. Використання архітектур із малим числом параметрів (наприклад, згорткових мереж із каскадним застосуванням фільтрів) або гібридних підходів (згорткові та рекурентні мережі) дає змогу краще впоратися з аугментаційними сигналами [8]. Останнім часом використовують автоматизований підбір архітектури моделі NAS (Neural Architecture Search) [9]. Поки цьому напряму приділяється недостатньо уваги, хоча він видається перспективним, завдяки таким перевагам: здатність моделі до узагальнення при збереженні її точності; врахування конкретних методів аугментації, що робить модель більш стійкою до особливостей даних.

Ця робота спрямована на розвиток методу оптимізації архітектури нейронних мереж NAS, який дозволяє максимально використовувати переваги аугментації, мінімізуючи пов'язані з нею ризики.

3. Постановка проблеми. Нехай $\mathbf{D} = \{(x_i(t), y_i(t))\}$, де $i = 1,\ldots, n$, $x_i(t) \in \mathbf{X}$ ($\mathbf{X}$ – множина вхідних сигналів), $y_i(t)$ – вихідний сигнал відповідний вхідному сигналу $x_i(t)$. Нехай $\mathbf{A} = \{a_j\}$, де $j = 1,\ldots,k$, a_j: $\mathbf{X} \to \mathbf{X}$ –

функція, яка перетворює початкові дані $x_i(t)$ з метою збільшення їх варіативності. Перетворений набір даних $\mathbf{D}_A = \{(a_j(x_i(t)), y_i(t))\} \cup \mathbf{D}$.

Задача оптимізації архітектури нейронної мережі з урахуванням аугментації даних полягає у визначенні таких гіперпараметрів λ (кількість шарів, кількість нейронів в шарах, типи активації), які визначають архітектуру нейронної мережі моделі $f(\theta, \lambda)$, де θ – ваги нейронної мережі. Гіперпараметри λ доставляють мінімум помилки L на валідаційному наборі даних $\mathbf{D}_{val}$:

$$\min_{\lambda} E_{(x,y)\in\mathbf{D}_{val}}[L(x, f(\theta^*, \lambda)), y], \tag{1}$$

де $\theta^* = \arg\min_{\theta} E_{(x,y)\in\mathbf{D}_{val}}[L(x, f(\theta^*, \lambda)), y]$.

Архітектура λ повинна належати допустимого простору архітектур $\lambda \in \Lambda$, що визначається обмеженнями на типи шарів, їхню кількість і параметри. Аугментація має призводити до стійкості моделі $f(\theta, \lambda)$ до різноманітних даних:

$$E_{a_j\in A}[L(x, f(\theta, \lambda)), y] \le \varepsilon, \tag{2}$$

де ε – допустима помилка на аугментованих даних.

В якості функції втрат L_T зазвичай використовується середньоквадратична помилка (*mse)* [10].

Метою роботи є визначення архітектури моделей нелінійної динаміки в умовах обмежених навчальних даних при забезпеченні заданої точності моделювання шляхом розвитку підходу на основі підбору архітектури нейронних мереж.

4. Викладення основного матеріалу. Для автоматизації проектування архітектури нейронної мережі використовується підхід на основі оптимізації моделі. За цільову функцію приймається продуктивність на заданій задачі за умови використання мінімальних ресурсів. Такий підхід включає три ключові компоненти:

- простір пошуку архітектур: усі можливі архітектури *arc*, які можуть бути досліджені (кількість шарів, кількість нейронів у шарах, типи активацій);
- стратегія пошуку: метод *strategy*, що визначає, як шукати найкращі архітектури в заданому просторі (еволюційні алгоритми, навчання з підкріпленням, градієнтний пошук тощо);
- оцінка продуктивності: механізм оцінки продуктивності кожної архітектури *eval* (використання невеликої вибірки даних або скороченого часу навчання тощо).

4.1. Автоматизований підбір архітектури нейронних мереж. Алгоритм формування архітектури нейронної мережі на основі метода NAS у вигляді псевдокоду наведено нижче.

Алгоритм 1: *neural_architecture_search*

1: Input: arc, strategy, eval, ε
*2: Output: arc**
3: foreach arc as arc_i
4: if $eval(arc_i) < \varepsilon$
5: return arc_i
6: end if
7: end foreach

В якості стратегії пошуку зазвичай використовуються підходи на основі:

- направленого пошуку: послідовний перебір архітектур від найпростіших до складних;
- випадкового пошуку: випадкова генерація архітектур;
- градієнтний пошук: неперервна оптимізація архітектури на основі градієнтів.

Стримуючими факторами використання цих стратегій є значне обчислювальне навантаження та невизначеність в обмеженнях простору пошуку рішень.

4.2. Автоматизація побудови архітектури нейронних мереж з аугментацією даних. Використання аугментації даних у NAS змінює процес оцінювання архітектур і робить модель стійкішою до різноманітності даних. Це вимагає додаткового аналізу, як архітектура взаємодіє з даними, що пройшли аугментацію, і ускладнює стратегію пошуку.

У зв'язку з цим, при використанні аугментації даних метод NAS отримує подальший розвиток:

- врахування варіативності даних у результаті різних трансформацій (аугментації) **A**;
- оцінка стійкості до аугментації: доповнення цільової функції метриками, що вимірюють стійкість архітектури до даних після трансформацій $eval_A$;
- балансування якості та обчислювальних витрат: використання аугментації збільшує обсяг даних і складність навчання, що вимагає врахування обчислювальних обмежень.

Алгоритм формування архітектури нейронної мережі на основі методу NAS з урахуванням аугментації даних у вигляді псевдо-коду наведено нижче.

Алгоритм 2: *neural_architecture_search_aug*

1: Input: arc, strategy, $eval_A$*,* ε*,* $\boldsymbol{A}$
*2: Output: arc**
3: foreach arc as arc_i

4: if $eval_A(\mathbf{A}, arc_i) < \varepsilon$
5: return arc_i
6: end if
7: end foreach

Переваги запропонованого підходу полягають у наступному: врахування стійкості моделі до варіацій даних, можливість виявлення архітектур, які більше підходять для складних та різнорідних даних, та покращення узагальнюючих здібностей моделі.

5. Постановка експерименту. Дослідження архітектури моделей нелінійної динаміки в умовах обмежених навчальних даних при забезпеченні заданої точності моделювання проводиться на прикладі тестового об'єкта. Імітаційна модель тестового об'єкта у вигляді послідовності нелінійної ланки з насиченням та динамічної ланки першого порядку наведена на рис. 1 [10].

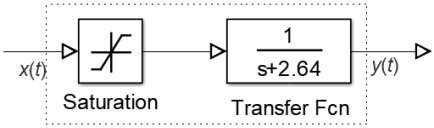

Рис. 1. *Імітаційна модель тестового об'єкта*

Для тестового об'єкта сформовано розмічений набір даних **D** на основі сигналів $x(t)$ на вході об'єкта та відгуків $y(t)$ на його виході. В якості вхідних сигналів використовуються імпульсні $x(t) = a\delta(t)$, ступінчасті $x(t) = a\Theta(t)$, лінійні $x(t) = at$ і гармонічні $x(t) = a\sin(t)$ сигнали різної амплітуди $a \in (0, 1]$. Доповнений набір даних $\mathbf{D}_A$ формується шляхом додавання завад та масштабування сигнала до існуючих сигналів.

На базі датасетів **D** та $\mathbf{D}_A$ збудовано дві моделі $f(\theta_1, \lambda_1)$ та $f_A(\theta_2, \lambda_2)$ відповідно. Моделі у вигляді нейронних мереж навчається за допомогою зворотнього розповсюдження помилки та визначенням вагових коефіцієнтів мережі методом Левенберга-Марквардта [11]. Це забезпечує високу точність моделювання при швидкій збіжності навчального процесу. Процес навчання продовжується до тих пір, поки обрана функція втрат L не досягне мінімуму або не буде виконано умову зупинки (2).

Експеримент полягає у дослідженні швидкості навчання моделей, збудованої на датасетах **D** та $\mathbf{D}_A$; дослідженні стійкості збудованих моделей при дії завад на вхідні сигнали. Моделі $f(\theta_1, \lambda_1)$ та $f_A(\theta_2, \lambda_2)$ являють собою тришарові нейронні мережі з кількістю параметрів $p_1 = 1000$ та $p_2 = 2000$ відповідно.

На рис. 2 наведено залежності функцій втрат *mse* від кількості епох навчання. На рис. 3 наведено залежності функцій втрат *mse* від рівня завад, що діють на вхідні сигнали.

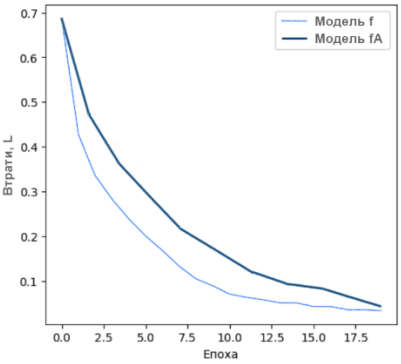

Рис. 2. *Залежності функцій втрат mse під час навчання моделей від кількості епох навчання*

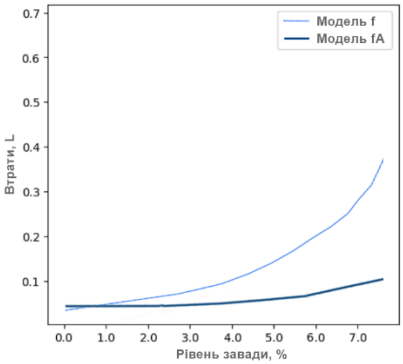

Рис. 3. *Залежності функцій втрат mse від рівня завад, що діють на вхідні сигнали*

З експерименту видно перевагу використання моделі $f_A(\theta_2, \lambda_2)$ під час ідентифікації нелінійних динамічних об'єктів на вхідних даних, які підлягають дії завад різного рівня, а саме, істотне зменшення функції втрат при рівні завад, що перевищує 3% у порівнянні з моделлю $f(\theta_1, \lambda_1)$. При цьому, час побудови моделі $f_A(\theta_2, \lambda_2)$ збільшується несуттєво у порівнянні з часом побудови моделі $f(\theta_1, \lambda_1)$.

6. Висновки. В роботі успішно розв'язано задачу визначення архітектури моделей нелінійної динаміки в умовах обмежених навчальних даних при забезпеченні заданої точності моделювання. Для вирішення протиріччя між підвищенням стійкості моделі до шумів і спотворень та ускладненням задачі навчання моделі в умовах обмежених обчислювальних ресурсів набув подальшого розвитку метод підбору архітектури нейронних мереж NAS.

Ефективність розробленого методу ідентифікації нелінійних динамічних об'єктів в умовах обмежених навчальних даних при забезпеченні заданої точності моделювання доведено під час розв'язання задачі ідентифікації тестового нелінійного динамічного об'єкта. Експеримент демонструє суттєве зменшення функції втрат на даних розширеного датасету за рівня завад, що перевищує 3%, порівняно з моделлю, що навчалася на даних без аугментації. При цьому, час побудови моделі збільшується несуттєво порівняно з часом побудови моделі, яка навчалася на даних без аугментації.

Перевагами запропонованого підходу до визначення правильної архітектури моделі є мінімізація негативних ефектів аугментації, таких як перенавчання. Автоматизація процесу налаштування архітектури під методи аугментації дозволяє поліпшити ефективність навчання моделі за браку розмічених даних.

Недоліками запропонованого підходу є підвищення складності моделі, а отже, часу її навчання. Тому, областю ефективного застосування запропонованого методу є задачі з нестачею розмічених даних за відсутності суворих вимог до швидкості процесу моделювання.

Автоматизований підбір архітектури моделі в поєднанні з аугментацією даних має значний потенціал для автоматизації та поліпшення якості моделей. Однак висока обчислювальна вартість, ризик перенавчання та залежність від якості аугментації вимагають ретельного підходу до його використання.

Список використаних джерел:

1. Kariri E., Louati H., Louati A., Masmoudi F. Exploring the Advancements and Future Research Directions of Artificial Neural Networks. *A Text Mining Approach. Appl. Sci.* 2023. Vol. 13. 3186. DOI: 10.3390/app13053186.
2. Islam M., Chen G., Jin S. An Overview of Neural Network. *American Journal of Neural Networks and Applications*. 2019. Vol. 5 (1). P. 7-11. DOI: 10.11648/j.ajnna.20190501.12.
3. Duc Haba. Data Augmentation with Python: Enhance deep learning accuracy with data augmentation methods for image, text, audio, and tabular data. *Packt Publishing*. 2023.
4. Wada S., Morimoto N., Investigating Relationship between Data Augmentation Intensity and Model Performance in Natural Language Processing. *2024 International Conference on Consumer Electronics – Taiwan (ICCE-Taiwan)*, Taichung, 2024. P. 445-446. DOI: 10.1109/ICCE-Taiwan62264.2024.10674562.
5. Guo P., Yang H., Sano A., Empirical Study of Mix-based Data Augmentation Methods in Physiological Time Series Data. *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI).* Houston, 2023. P. 206-213. DOI: 10.1109/ICHI57859.2023.00037.
6. Vaish P., Wang S., Strisciuglio N. Fourier-Basis Functions to Bridge Augmentation Gap: Rethinking Frequency Augmentation in Image Classification. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, 2024. P. 17763-17772. DOI: 10.1109/CVPR52733.2024.01682.
7. Muthumari M., Bhuvaneswari C. A., Kumar Babu J. E. N. S., Raju S. P. Data Augmentation Model for Audio Signal Extraction. *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*. Coimbatore, 2022. P. 334-340. DOI: 10.1109/ICESC54411.2022.9885539.
8. J. Chen, Y. Hong, C. Liu, Q. Xu, G. Zhou. Decoupling and Refilling: A Simple Data Augmentation Method for Aspect Term Extraction. *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Seoul, 2024. P. 12582-12586. DOI: 10.1109/ICASSP48485.2024.10446120.
9. Elsken T., Metzen J. H., Hutter F. Neural architecture search: A survey. *Journal of Machine Learning Research* / Hutter F., Kotthoff L., Vanschoren J. (eds) Cham, 2019. 20. DOI: 10.1007/978-3-030-05318-5_3.
10. Фомін О. О., Сперанський В. О., Орлов А. А. та ін. Метод опорних моделей синтезу інтелектуальних систем ідентифікації нелінійних динамічних об'єктів. *Математичне та комп'ютерне моделювання. Серія: Технічні науки.* 2024. Вип. 25, С. 129-139. DOI: 10.32626/2308-5916.2024-25.129-139.

11. Fomin O., Polozhaenko S., Bidyuk P., Tataryn O., Prokofiev A. Improving measurements accuracy in weight-in-motion systems using dynamic neural networks. *ICST-2024: Information Control Systems & Technologies, September, 23-25, 2024*. Odesa: CEUR Workshop Proceedings, 2024, 3790, p. 483–493.

OPTIMIZATION OF NEURAL NETWORK ARCHITECTURE WITH REGARD TO DATA AUGMENTATION

The paper is devoted to resolving the contradiction between increasing the model's resistance to interference and distortion and complicating the task of model training under conditions of limited computational resources. The aim of the work is to determine the architecture of nonlinear dynamics models under conditions of limited training data while ensuring a given modeling accuracy. This goal is achieved by developing a method for selecting the architecture of NAS neural networks. The scientific novelty of the work lies in the further development of the method of selecting the architecture of the NAS neural network for identifying nonlinear dynamic objects, taking into account the distortions of the training dataset by adding segmented data. In contrast to the traditional approach to pre-training, the developed method allows us to build more robust models while ensuring the required accuracy. The practical significance of the work is to develop an approach to adapting the architecture depending on the augmentation methods used by developing an algorithm for selecting the architecture of a NAS neural network taking into account data augmentation, which allows building more reliable models without losing modeling accuracy. The results of experiments on modeling test objects with nonlinear dynamic characteristics are presented, and the influence of data augmentation on the quality and stability of the obtained models is analyzed. The value of the study is to determine the area of effective use of the proposed method, as tasks with a lack of labeled data in the absence of strict requirements for the speed of the modeling process.

Keywords: *nonlinear dynamics, neural networks, data augmentation, architecture optimization, learning automation.*

Отримано: 9.12.2024