УДК 004.8:004.93
DOI: 10.18523/2617-3808.2025.8.50-56

A. Mykytyshyn, N. Shvai

VALIDATING ARCHITECTURAL HYPOTHESES IN NEURAL DECISION TREES WITH NEURAL ARCHITECTURE SEARCH

This article introduces an automated and unbiased framework for validating architectural hypotheses for neural network models, with a particular focus on Neural Decision Trees (NDTs). The proposed methodology employs Neural Architecture Search (NAS) as an unbiased tool to explore architectural variations and empirically assess theoretical claims. To demonstrate this framework, we investigate a hypothesis found in the literature: that the complexity of decision nodes in NDTs decreases monotonically with tree depth. This assumption, initially motivated by the task of monocular depth estimation, suggests that deeper nodes in the tree require fewer parameters due to simpler split functions.

To rigorously test this hypothesis, we conduct a series of NAS campaigns over the CIFAR-10 image classification dataset. The search space parameterizes each node by the number of convolutional blocks and fully connected layers, while all other architectural components are held constant to isolate the effect of node depth. By applying Tree-structured Parzen Estimator (TPE)-based NAS and evaluating over 300 architectures, we quantify complexity metrics across tree levels and analyze their correlations using Spearman's rank coefficient.

The results provide no statistical or visual evidence supporting the hypothesized trend: node complexity does not decrease with depth. Instead, complexity remains nearly constant across levels, regardless of tree depth or search space size. These results suggest that assumptions derived from specific applications may not generalize to other domains, underscoring the importance of empirical validation and careful search-space design. The presented framework may serve as a foundation for verifying other structural assumptions across various neural network families and applications.

Keywords: Neural Architecture Search (NAS), Neural Decision Trees (NDTs), Automated Machine Learning (AutoML), Computer Vision, Node Complexity.

Introduction

Automated Machine Learning (AutoML) has significantly streamlined the development and deployment of complex machine learning models by automating various aspects of the machine learning pipeline, including data preprocessing, feature engineering, model selection, and hyperparameter optimization [4]. A major subset of AutoML is Neural Architecture Search (NAS), a technique focused specifically on automating the design of neural network architectures. NAS has demonstrated remarkable success across various domains, consistently delivering architectures that match or even outperform those designed by human experts [3, 5].

At the same time, Neural Decision Trees (NDTs) have emerged as an innovative class of hybrid machine learning models, integrating the interpretability and intuitive decision logic of classical decision trees with the expressive capabilities of neural networks [2]. Despite their practical benefits, several underlying theoretical assumptions about their structure remain untested. In particular, a hypothesis proposed by Roy and Todorovic (2016) states that nodes deeper within an NDT should be structurally simpler, since learning split functions presumably becomes easier deeper within the tree [9].

However, this assumption has not been validated beyond the original context in which it was proposed (monocular depth estimation). Unverified theoretical assumptions pose a significant risk, potentially leading researchers and practitioners towards suboptimal architectural decisions and incorrect generalizations across different applications. Therefore, rigorous empirical testing of such architectural hypotheses is necessary to ensure reliable design decisions for neural decision trees in diverse tasks.

The objective of this research is to empirically verify the hypothesis that the complexity of neural decision tree nodes decreases with increasing depth using Neural Architecture Search as an unbiased experimental tool.

The scientific novelty of the obtained results includes:

- The first systematic, NAS-based empirical validation of an architectural hypothesis regarding the complexity of neural decision tree nodes.
- Empirical evidence demonstrating that the assumption of decreasing complexity with increasing depth, proposed by Roy and Todorovic, does not generalize across tasks.

The practical value of the obtained results lies in providing a validated methodological approach to critically evaluate theoretical claims about NDT architectures. Researchers and practitioners can leverage these findings to guide more reliable and empirically grounded architectural decisions in neural decision tree design.

Problem Definition

Roy and Todorovic [9] introduce the Neural Regression Forest (NRF) for monocular depth estimation, in which each split node is implemented by a "shallow" CNN – specifically, one with only one or two convolutional layers followed by one or two fully-connected layers – and the overall network depth emerges from the stacking of these modules. They conjecture that "it becomes easier to learn the split functions as we go down the tree," suggesting that nodes at lower levels should require progressively fewer layers to achieve the best possible predictive performance.

As a result of this conjecture, they propose an architecture where the nodes in a tree get "simpler" as they get closer to the leaves. Specifically, they split the tree into three equally deep layers. "For the top one third of the tree height, we use CNNs with 2 convolution + pooling layers, and 2 fully connected perceptron layers. For the lower one third of the tree height (closer to the leaf nodes), we use CNNs with 2 convolution + pooling layers and 1 fully connected perceptron layer. Finally, for the bottom third of the tree height, we use CNNs with 1 convolution + pooling layer and 1 connected perceptron layer" [9].

Drawing directly on their assumption and generalizing it, we formulate our central hypothesis for neural decision trees (NDTs):

Hypothesis. *In neural decision trees, the complexity of each split node, quantified by the number of convolutional blocks and fully connected layers, decreases monotonically with increasing tree depth.*

Proposed Approach

To verify the hypothesis that node complexity in neural decision trees decreases with tree depth, we employ neural architecture search (NAS) solely as an empirical tool. NAS systematically explores a predefined set of discrete architectural choices by optimizing for a target metric (in our case, accuracy), thus revealing which per-node configurations best support the task under identical training conditions [1, 3]. By automating the search, we eliminate human bias in selecting convolutional and fully connected layers, thus obtaining an unbiased measurement of how complexity varies across different levels.

All candidate architectures are trained and evaluated on the CIFAR-10 dataset, which comprises 60,000 color images of size 32×32 across 10 classes [6]. We choose CIFAR-10 for its status as a standard benchmark. Another benefit is that its modest image resolution and well-established augmentation protocols enable rapid NAS iterations. It is also worth noting that the decision to select a dataset different from the one used in the paper where this idea was initially introduced ([9]) is deliberate, as this allows us to test the general applicability of the hypothesis.

We define our search space by parameterizing each split node at tree level ℓ with two integer hyperparameters:

- b_ℓ – the number of convolutional blocks at level ℓ
- f_ℓ – the number of fully connected layers at level ℓ

Each convolutional block replicates the module from [2] – namely, a 3×3 convolution with 256 channels, followed by batch normalization and ReLU activation. We consider two value ranges for (b_ℓ, f_ℓ): {1, 2} (matching the original setup from [9] at depth ten), and {1, 2, 3} (to probe a broader spectrum of complexity).

All other architectural and optimization parameters remain fixed across experiments: the overall tree depth d is set per experiment, a double-block stem precedes branching, channel widths are constant at 256, and the optimizer, learning rate schedule, and augmentation pipeline are identical. This ensures that variations in NAS outcomes reflect only the relative efficacy of different (b_ℓ, f_ℓ) choices, allowing for a clear test of the monotonic complexity hypothesis.

We employ the Tree-structured Parzen Estimator (TPE) algorithm. Each candidate model undergoes 5 epochs of training ("proxy evaluation"), after which the top 10% of architectures (by validation accuracy) are retrained for 15 epochs to refine the final rankings.

All experiments employ mixed-precision training and a gradient scaler to speed up training. Input images are augmented with random horizontal flips and random crops to 32×32 (with four-pixel padding), followed by tensor conversion and normalization using the CIFAR-10 channel means and standard deviations. Optimization uses the AdamW optimizer with a fixed initial learning rate of 1×10^{-3}.

Architectures are ranked by accuracy on a held-out CIFAR-10 validation set. Accuracy directly measures the quality of learned split functions in a classification context, providing a principled and unbiased basis for comparing node complexities [3].

Implementation Details and Reproducibility

All experiments were conducted on Google Colab using an NVIDIA A100 GPU with 32 GB of RAM, subject to Colab's 12-hour session limit. The code was written in Python 3 and relied on the PyTorch deep learning framework for model definition and training [8], as well as Microsoft's NNI library for the implementation of NAS [7]. To ensure reproducibility, a fixed random seed of 42 was used for all data splits, network initialisations, and sampling in the NAS algorithm.

Experimental Campaigns

To test the node complexity hypothesis, we ran six NAS campaigns, each fixing tree depth d and the discrete knob set for (b_ℓ, f_ℓ) using TPE with median-stop early stopping after 3 epochs. Each trial required approximately 1 minute for 5 epochs (shorter if stopped early), and all campaigns ran no longer than the 12-hour Colab limit (approximately 700 minutes). All campaigns used a 5-epoch proxy evaluation and 15-epoch retraining of the top 10 % of candidates.

Table 1. **Experimental runs that were performed**

Run ID	Depth d	$(b_\ell, f_\ell) \in$	Search space size
A	3	{1, 2}	$2^3 \times 2^3 = 64$
B	4	{1, 2}	$2^4 \times 2^4 = 256$
C	6	{1, 2}	$2^6 \times 2^6 = 4{,}096$
D	3	{1, 2, 3}	$3^3 \times 3^3 = 729$
E	4	{1, 2, 3}	$3^4 \times 3^4 = 6{,}561$
F	6	{1, 2, 3}	$3^6 \times 3^6 = 531{,}441$

- Runs A-C probe the original Roy and Todorovic complexity range {1, 2} at depths 3, 4, and 6.
- Runs D-F expand the search to {1, 2, 3} at the same depths.

Each trial sampled a full configuration of $\{b_0, ..., b_{d-1}, f_0, ..., f_{d-1}\}$ and returned validation accuracy after 5 epochs; the top 10 % of trials per run were then retrained for 15 epochs for final evaluation.

Run-wise accuracy summaries

Figure 1 (next page) displays a table listing the top-five architectures returned by each campaign after the 15-epoch retraining phase. Three regularities are evident:

- **Tight within-run spread.** All five models in a given run differ by ≤ 0.01 in top 1 accuracy, showing that the parameters we searched over most likely do not influence the accuracy that much.
- **Depth penalty.** Mean accuracy declines as depth increases—from ~0.83 for depth 3 (Runs A, D) to ~0.79 for depth 6 (Runs C, F). Deeper NDTs entail more parameters and longer inference paths, making them harder to optimize given the fixed 15-epoch budget and 256-channel bottleneck.
- **Effect of knob range.** For each depth, the runs with the broader option set {1, 2, 3} (D–F) surpass their {1, 2} counterparts (A–C) by ~0.01. Allowing an extra layer choice evidently enables NAS to fine-tune node expressiveness without overfitting, hence the modest but consistent gain.

run	accuracy	rank	fc_l0	stg_l0	fc_l1	stg_l1	fc_l2	stg_l2	fc_l3	stg_l3	fc_l4	stg_l4	fc_l5	stg_l5
A	0.83	1	1	2	2	1	2	2						
	0.828	2	2	2	2	2	2	1						
	0.826	3	1	2	1	1	1	1						
	0.824	4	1	2	1	2	2	1						
	0.822	5	1	2	2	2	2	1						
B	0.81	1	1	1	2	2	2	1	1	1				
	0.808	2	1	1	2	1	1	2	1	1				
	0.806	3	1	1	1	2	2	1	1	1				
	0.804	4	1	1	2	1	1	2	1	1				
	0.802	5	1	1	2	1	2	2	2	1				
C	0.79	1	2	1	2	1	2	2	1	2	2	1	1	2
	0.788	2	1	2	2	2	2	2	1	2	1	2	2	2
	0.786	3	2	1	2	1	1	2	2	1	2	1	2	1
	0.784	4	1	1	1	1	2	2	1	1	1	2	2	1
	0.782	5	2	1	1	2	1	2	2	2	2	2	2	1
D	0.84	1	1	1	2	3	1	1						
	0.838	2	1	1	3	2	1	1						
	0.836	3	3	1	3	2	1	2						
	0.834	4	2	1	1	3	1	3						
	0.832	5	1	1	1	3	3	2						
E	0.82	1	3	3	1	3	2	1	2	3				
	0.818	2	1	1	3	1	1	1	1	1				
	0.816	3	1	3	1	3	2	2	2	1				
	0.814	4	2	2	2	1	2	3	1	2				
	0.812	5	3	1	3	1	2	3	3	2				
F	0.8	1	1	2	2	1	3	3	3	3	2	3	3	3
	0.798	2	3	3	1	2	3	3	2	3	2	1	3	3
	0.796	3	1	3	1	1	3	1	3	3	3	1	1	1
	0.794	4	2	3	1	2	3	3	3	2	1	1	1	1
	0.792	5	3	3	1	1	1	2	3	1	1	2	1	3

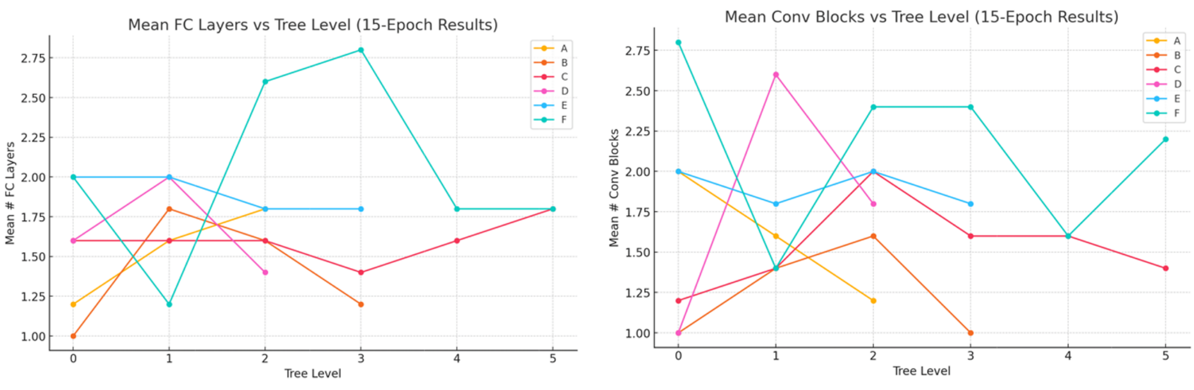
Figure 1. Table, depicting the top 5 results for each run, measured by validation accuracy after 15 epochs

Taken together, the accuracies fall in the narrow band 0.78 – 0.84, confirming that all architectures are viable classifiers and that subsequent analyses can focus on node complexity rather than gross performance differences.

Complexity profiles visualized

Figure 2 (left, 1) plots the mean number of convolutional blocks per level ($\bar{b}_\ell$) for every run, while Figure 2 (right, 2) shows the analogous curve for fully-connected layers ($\bar{f}_\ell$). Shaded markers denote tree depth. Below are the observations we draw from visually examining the plots:

- The curves are essentially flat; no run exhibits a consistent downward slope.
- Local fluctuations are noise-like, reflecting the stochastic nature of NAS sampling rather than a systematic preference for simpler nodes at deeper levels.
- Runs with the larger knob range (D–F) unsurprisingly have higher absolute means, yet their level-wise profiles are again flat.



1 *2*

Figure 2. Plots showing the mean FC layers (*1*) and convolution blocks (*2*) vs tree level

Visually, therefore, the data do not support the conjectured monotonic decrease in node complexity.

Statistical test of the monotonic-simplicity hypothesis

To formalize the visual impression, we apply Spearman's rank correlation coefficient ρ, a non-parametric measure of monotone association between two variables [10]:

$$c = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

$$d_i = \text{rank}(x_i) - \text{rank}(y_i),$$

where x_i is the tree level and y_i is the corresponding complexity metric (either b_ℓ or f_ℓ), and n is the total number of node observations pooled across runs.

We define the null hypothesis and the alternative hypothesis as follows:

- Null hypothesis H_0: c = 0 (no monotonic relationship between level and complexity).
- Alternative hypothesis H_1: c < 0 (complexity decreases with depth).

Using the pooled dataset, we obtain:

Table 2. **The numeric results of the Spearman's rank-correlation test**

Metric	ρ	p-value
Conv-blocks b_ℓ	+0.03	0.75
FC-layers f_ℓ	+0.11	0.21

Because both p-values are ≫ 0.05, we fail to reject H_0. Moreover, the positive signs of c, although small, run counter to H_1, reinforcing the qualitative conclusion: there is no statistical evidence that node complexity diminishes with tree depth under the examined search space and training regime.

Together, the near-identical accuracies, flat complexity profiles, and non-significant Spearman coefficients collectively refute the monotonic-simplicity hypothesis for neural decision trees on CIFAR-10, at least within the constraints of the present experimental design.

Hypothesis Evaluation

The central hypothesis under investigation posited that, in neural decision trees, node complexity, measured by the number of convolutional blocks b_ℓ and fully connected layers f_ℓ, would decrease monotonically with increasing tree depth. In other words, deeper split nodes should require fewer layers to achieve comparable classification accuracy.

However, the empirical evidence fails to support this conjecture. As shown in the previous section, the mean complexity profiles $\overline{b_\ell}$ and $\overline{f_\ell}$ are essentially constant across levels, without any discernible downward trend. The formal Spearman rank test further confirms that the observed correlations are small and positive (ρ ≈ +0.03 for b_ℓ, c ≈ +0.11 for $\overline{f_\ell}$) and statistically non-significant (p ≫ 0.05), leading us to retain the null hypothesis of no monotonic association.

Several factors may explain this departure from the original expectation of Roy and Todorovic. First, their depth-10 architecture was tailored to monocular depth estimation, a regression task with spatial continuity, whereas our CIFAR-10 classification problem may impose different representational requirements at all tree levels. Second, our fixed 256-channel convolutional blocks and limited training budget (15 epochs plus early stopping) may attenuate any subtle benefits of reduced complexity in lower nodes. Finally, NAS optimizes for overall accuracy, not explicitly for per-node efficiency, so it may favor uniformly expressive nodes to maximize global performance under the given constraints.

In sum, within the confines of our search space and training regime, there is no evidence that node complexity in neural decision trees decreases with depth. This negative result suggests that the "easier-to-learn" assumption articulated by Roy and Todorovic does not generalize straightforwardly from depth estimation to image classification, or that more tailored search strategies are required to expose such a trend.

Conclusion

This thesis provided an empirical evaluation of a critical architectural hypothesis within Neural Decision Trees (NDTs), specifically the claim by Roy and Todorovic (2016) that nodes deeper within the tree require progressively simpler neural structures. Employing Neural Architecture Search (NAS) as an unbiased methodological tool, we systematically explored variations in node complexity, measured by the number of convolutional blocks and fully connected layers, across multiple tree depths using the CIFAR-10 image classification dataset.

Our comprehensive experiments and subsequent statistical analyses revealed no support for the monotonic simplicity hypothesis. Contrary to expectations, node complexity remained effectively constant across all tree levels, and no statistically significant correlation was observed between node depth and complexity. This finding suggests that architectural assumptions derived from specialized tasks, such as monocular depth estimation, may not generalize across different domains and datasets, underscoring the importance of empirically validating theoretical claims in neural architecture research.

While our experiments provide strong evidence refuting the generalized hypothesis, several limitations remain. The search space, though carefully chosen, was constrained by practical computational considerations, including limited training epochs and fixed convolutional channel widths. Future research could explore expanded or more nuanced search spaces, including varying parameters per node rather than per level, different datasets and tasks, and incorporating more advanced NAS strategies or evaluation metrics that explicitly target node-level complexity.

In summary, this study underscores the necessity for rigorous empirical testing of architectural assumptions in neural network research. The negative result regarding node simplicity in NDTs provides important insights for future architecture design and highlights the critical role of NAS methodologies in facilitating unbiased, systematic explorations.

Список літератури

1. A comprehensive survey of neural architecture search: Challenges and solutions [Electronic resource] / P. Ren et al. // ACM Computing Surveys (CSUR). — 2021. — Vol. 54, no. 4. — Pp. 1–34. — Mode of access: https://doi.org/10.1145/3447582.
2. Deep neural decision forests [Electronic resource] / P. Kontschieder et al. // Proceedings of the IEEE international conference on computer vision. — 2015. — Pp. 1467–1475. — Mode of access: https://openaccess.thecvf.com/content_iccv_2015/papers/Kontschieder_Deep_Neural_Decision_ICCV_2015_paper.pdf.
3. Elsken T. Neural architecture search: A survey [Electronic resource] / T. Elsken, J. H. Metzen , F. Hutter // Journal of Machine Learning Research. — 2019. — Vol. 20, no. 55. — Pp. 1–21. — Mode of access: https://www.jmlr.org/papers/volume20/18-598/18-598.pdf.
4. He X. AutoML: A survey of the state-of-the-art [Electronic resource] / X. He, K. Zhao, X. Chu // Knowledge-based systems. — 2021. — Vol. 212. — Pp. 106622. — Mode of access:https://doi.org/10.1016/j.knosys.2020.106622.
5. Hutter F. Automated machine learning: methods, systems, challenges [Electronic resource] / F. Hutter, L. Kotthoff, J. Vanschoren. — Springer Nature, 2019. — Mode of access: https://doi.org/10.1007/978-3-030-05318-5.
6. Krizhevsky A. Learning multiple layers of features from tiny images [Electronic resource] / A. Krizhevsky, G. Hinton, et al. — 2009. — Mode of access: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.
7. NNI: Neural Network Intelligence [Electronic resource] / H. Liu et al. — 2019. — Mode of access: https://github.com/microsoft/nni.
8. Pytorch: An imperative style, high-performance deep learning library [Electronic resource] / A. Paszke et al. // Advances in neural information processing systems. — 2019. — Vol. 32. — Mode of access: https://pytorch.org.
9. Roy A. Monocular depth estimation using neural regression forest [Electronic resource] / A. Roy, S. Todorovic // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — Pp. 5506–5514. — Mode of access:https://doi.org/10.1109/cvpr.2016.594.
10. Spearman C. The proof and measurement of association between two things [Electronic resource] / C. Spearman // International journal of epidemiology. — 2010. — Vol. 39, no. 5. — Pp. 1137–1150. — Mode of access:https://doi.org/10.1093/ije/dyq191.

References

Elsken, T., Metzen, J. H., & Hutter, F. (2019). Neural architecture search: A survey. *Journal of Machine Learning Research*, *20* (55), 1–21.

He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. *Knowledge-based systems*, *212*, 106622.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated machine learning: methods, systems, challenges*. Springer Nature.

Kontschieder, P., Fiterau, M., Criminisi, A., & Bulo, S. R. (2015). Deep neural decision forests. In *Proceedings of the IEEE international conference on computer vision* (pp. 1467–1475).

Krizhevsky, A., Hinton, G., & others. (2009). *Learning multiple layers of features from tiny images*.

Liu, H., Li, Y., Shen, Y., Zhao, D., & Xie, C. (2019). *NNI: Neural Network Intelligence*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., & others. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Chen, X., & Wang, X. (2021). A comprehensive survey of neural architecture search: Challenges and solutions. *ACM Computing Surveys (CSUR)*, *54* (4), 1–34.

Roy, A., & Todorovic, S. (2016). Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5506–5514).

Spearman, C. (2010). The proof and measurement of association between two things. *International journal of epidemiology*, *39* (5), 1137–1150.

Микитишин А. П., Швай Н. О.

ВАЛІДАЦІЯ АРХІТЕКТУРНИХ ГІПОТЕЗ У НЕЙРОННИХ ДЕРЕВАХ РІШЕНЬ ЗА ДОПОМОГОЮ ПОШУКУ НЕЙРОННИХ АРХІТЕКТУР

У цій роботі запропоновано автоматизовану та об'єктивну методику перевірки архітектурних гіпотез за допомогою пошуку нейронних архітектур (Neural Architecture Search, NAS). Основна ідея полягає в застосуванні NAS як інструменту для оцінки теоретичних припущень щодо структури моделей без ручного налаштування архітектур або впливу суб'єктивних рішень дослідника. Для демонстрації підходу було перевірено гіпотезу про те, що складність вузлів у нейронних деревах рішень (Neural Decision Trees, NDTs) зменшується зі збільшенням глибини дерева. Це припущення зустрічається в науковій літературі та використовується як обґрунтування для побудови спеціалізованих архітектур, однак раніше не було перевірене на систематичній експериментальній основі.

У межах дослідження було розроблено повністю автоматизований експериментальний фреймворк для генерації, навчання та оцінювання сотень архітектур NDT з різними конфігураціями вузлів. Для пошуку ефективної структури дерев було використано метод баєсівської оптимізації (Tree-structured Parzen Estimator, TPE). Складність вузлів оцінювали за кількома метриками: кількістю параметрів, кількістю обчислювальних операцій, кількістю нейронів у шарі та глибиною шару. Для аналізу зв'язку між глибиною вузлів і їхньою складністю застосовували коефіцієнт рангової кореляції Спірмена (Spearman's rank correlation coefficient).

За результатами обчислювального експерименту, що охопив понад 300 згенерованих моделей на синтетичному класифікаційному датасеті, не було виявлено жодної стабільної або статистично значущої залежності між глибиною вузла та його складністю. Отримані результати свідчать про те, що припущення, сформовані на основі окремих прикладів або інтуїції, можуть не узагальнюватися на інші задачі або домени. Це підкреслює важливість емпіричної перевірки теоретичних архітектурних міркувань, а також необхідність уважного проєктування простору пошуку в NAS.

Запропонований підхід може бути використаний для перевірки інших архітектурних гіпотез у різноманітних типах нейронних мереж, що робить його перспективним інструментом у дослідженнях у сфері автоматизованого машинного навчання.

Ключові слова: пошук нейронних архітектур, нейронні дерева рішень, автоматизоване машинне навчання, комп'ютерний зір, складність вузлів.

Матеріал надійшов 16.06.2025