UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Technology
Robotics and Computer Engineering

Bernard Michał Szeliga

Statistical Analysis for Mechanistic Explainability of Artificial Neural Networks

Master's Thesis (30 EAP)

Supervisors:
Kallol Roy
Nesma Talaat Abbas Mahmoud

Tartu 2025

Statistical Analysis for Mechanistic Explainability of Artificial Neural Networks

Abstract:

With the growth of the prevalence of artificial intelligence (AI), significant resources have been dedicated to advancing capabilities and expanding opportunities for the implementation of artificial neural networks (ANN). At the same time, efforts to provide explanations for the processes and to ”open the black box” have been notably more limited, despite growing legislative pressures and the need for more transparency before AI can be implemented in high-risk applications. Even when advances are made, explanations often answer the ”Why?” question, and not ”How?”. The arguments for the mechanistic explainability of the neural networks have been raised to address this point, among others. This paper presents a framework for analysis of the ANNs focused on broader population trends and patterns within networks, assesses the effectiveness of suggested methods, and suggests potential paths for future development in the area of mechanistic interpretability of neural networks

Keywords: Artificial Intelligence, Statistical analysis, XAI

CERCS: P160 Statistics, operation research, programming, actuarial mathematics

CERCS: P176 Artificial intelligence

Kunstlike närvivõrkude mehhanistliku seletatavuse statistiline analüüs

Lühikokkuvõte:

Tehisintellekti (TI) leviku kasvuga on märkimisväärsed ressursid eraldatud tehisnärvivõrkude (TIN) võimekuse arendamiseks ja rakendamise võimaluste laiendamiseks. Samal ajal on protsesside selgitamise ja ”musta kasti avamise”jõupingutused olnud märkimisväärselt piiratumad, hoolimata kasvavast seadusandlikust survest ja vajadusest suurema läbipaistvuse järele enne tehisintellekti rakendamist kõrge riskiga rakendustes. Isegi kui edusamme tehakse, vastavad selgitused sageli küsimusele ”Miks?”, mitte ”Kuidas?”. Selle punkti käsitlemiseks on muuhulgas esitatud argumente närvivõrkude mehhanistliku selgitatavuse kohta. Käesolev artikkel esitab raamistiku TIN-ide analüüsimiseks, mis keskendub laiematele populatsioonitrendidele ja mustritele võrkudes, hindab pakutud meetodite tõhusust ja pakub välja potentsiaalseid edasisi arengusuundi närvivõrkude mehhanistliku tõlgendatavuse valdkonnas.

Võtmesõnad: Tehisintellekt, statistiline analüüs, XAI

CERCS: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

CERCS: P176 Tehisintellekt

Contents

1. Introduction 5
2. Related Works 6
3. Method 8
    3.0.1 Justification 9
    3.0.2 Class activation band analysis 9
  3.1 Sensitivity and correlation analysis 10
    3.1.1 Neuron sensitivity 10
    3.1.2 Neuron correlation 10
    3.1.3 Justification 10
    3.1.4 Limitations 11
4. Testing and Visualization 12
  4.1 Testing system and parameters 12
  4.2 Visualization 12
5. Results and Analysis 14
  5.1 Neuron's active classes analysis 14
  5.2 Sensitivity and Correlation Analysis 17
  5.3 Active Band Analysis 21
  5.4 Additional Parameters 23
6. Critical Appreciation 29
7. Discussion and Conclusion 30
References 31
Appendices 33
  7.1 Test network structure 33
License 34

1. Introduction

With deep neural networks becoming increasingly prevalent in everyday life ([1]), a substantial portion of the research on the topic focuses on improving performance and implementing the technology ([2]). At the same time, the efficiency and explainability of neural networks (NN) is a notoriously difficult problem. Recently, the European Union (EU) introduced mandatory regulations that require inclusion of reasoning behind AI decisions ([3], [4]). A large part of the research appears to focus on the explainability of a single decision for a single data point ([5]). This enables a detailed understanding of the decision-making process, but it is challenging to generalize and requires significant computational resources. Layer-wise Relevance Propagation (LRP) method uses explainability optimization (XAI) during training ([6]) to enhance the network's performance. Kästner et al. ([7]) argued that the majority of XAI approaches are task-specific and aim to provide the best explanation for a given situation. Although important, this approach provides insight into only a narrow set of conditions; at the same time, some requirements require a more holistic approach. Understanding a neural network's operation can be facilitated by partitioning the network into functional blocks, which improves the efficiency and accuracy of the model.

This thesis presents a statistical analysis of neural network activations during its training on larger datasets. Statistical patterns emerge from the activations of convolutional neural networks (CNNs) during its classification task. This method can be generalized to other neural network architectures. The thesis is organized as follows: Section 3 discusses current XAI methods with their benefits and limitations. In Section 3, an alternative method of analysis is proposed, and Section 4 explains the experimental setup and its visualization of the results. The experimental results and analysis are discussed in Section 5. Section 6 discusses the benefits and limitations of the proposed method. Section 7 concludes with future work.

2. Related Works

This section discusses a general overview of the interpretability of deep neural networks. Vander Haar et al. ([5]) provide an overview of such methods that focus on CNNs. In their analysis they explained the contemporary techniques, such as:

- Deconvolutional neural network
- Layer-wise relevance propagation
- Deep Taylor decomposition
- Logic Rules
- Inversion
- Local Interpretable Model-agnostic Explanations (LIME)
- Occlusion tests
- Shapley Additive Explanations (SHAP)

The majority of these methods aim to create a saliency map, or an analogous visual representation of the importance and/or meaning of specific sections of the input image, and how that information combines to result in a particular decision made by the network. The logical rules method, based on fuzzy logic, differs in this regard from other frameworks as the output is in the form of a series of logical statements.

The methods presented in the overview are case-specific and thus can provide detailed information for each case provided, but explanations cannot be easily extrapolated to create general rules. Additionally, due to the complexity of providing specific information for a single case, it is not feasible to obtain sufficient results to create a representative sample for the entire population. The LRP method proposed by Lee et at. ([6]) builds on the known methods by utilizing XAI concepts into optimization algorithms. Proposed by them, the method introduces an additional parameter β to the stochastic gradient descent, where β is calculated in correlation with the LRP scores. While optimization of the neural network algorithms as not been the focus of this work, the generated results can foreseeably be used for this purpose, although reduction of the computation complexity of the networks appears to be more feasible to implement. An additional difference between the approaches is the stage at witch the methods are applied. The

LRP algorithm, as presented by the authors, is applied during the training process, wile the framework presented in this paper is used on the fully trained model.

Lee et al. provided proof of the utility of the XAI frameworks beyond a simple understanding of the networks and, as an extension, of an additional argument for the development of novel methods. Kästner et al. ([7]) propose that as the complexity of deep neural networks increases, case-specific approaches employed by contemporary XAI models are insufficient. They proposed a more holistic approach of MI that is appropriate in some cases, and in a broader sense, that it could be a valuable extension to existing methods. Secondly, it has been shown that contemporary methods can provide a detailed answer to the question ”Why this decision was made?”, but they are unable to answer the question ”How is the decision made?”. Olah et al. proposed methods for enhancing the interpretability of deep neural networks ([8], [9]), which gained some insights from that understanding ([10]).

Feature visualization allows the development of highly activating features through an iterative process. This method can provide a high-quality representation, but it is also susceptible to the creation of synthetic images that contain no valuable information, yet result in artificially high neuron activation. This method requires implementation of constraints, or a form of supervision over the image optimization process what in turn limits the possible range of scope of implementation. Semantic dictionaries offer a promising approach for translating neural network operations into a form that is human-understandable. This approach faces an opposite problem to the feature visualization, where images that can be mapped are limited by the ability of natural language do provide clear and precise information.

The thesis adopts a ”middle of the road” approach that balances the legibility of the information with its accuracy to the operation of the model. For example, using the feature visualization, an image that is highly stimulating to the artificial neurons could be developed, and later filtered and simplified to create semantically mappable representations. The methods presented in this work could be used to facilitate this goal by providing a comparative analysis between the developed visualized feature and a simplified image. In this way, the relationship between the simplified image and the original can be proven and then mapped to semantics.

3. Method

We propose a novel method that estimates the significance of a neuron within the network in relation to either any input or a specific input and output. It achieves this through two modes of analysis. The first estimates active classes of a neuron, and the second calculates the neuron's relation to the input and output of the network, denoted as sensitivity and correlation of the neuron, respectively.

The active class of the neuron is defined as a set of inputs for which the neuron's activation level deviates significantly from an arbitrary input. Here, the neuron is denoted as a function n. If a neuron is activated using the ReLU function, then it maps its inputs to real positive number:

$$n : X \to \mathbb{R}^+ \tag{1}$$

Let C be the set of all classes and X_c denote the set of all inputs belonging to class c. For any input x from X, it belongs to exactly one class and thus:

$$\forall c \in C, X_c \subset X \tag{2}$$

$$\forall c, k \in C, X_c \cap X_k = \emptyset \tag{3}$$

$$\forall x \in X, \exists c \in C, x \in X_c \tag{4}$$

Let the network be denoted as a function f that maps inputs from X to C:

$$f : X \to C \tag{5}$$

If a neuron detects a feature typical of a given class, then:

$$\forall x, y \in X, f(x) = f(y) \Rightarrow n(x) \approx n(y) \tag{6}$$

And thus, with a sufficiently large sample size, a neuron can be interpreted to map inputs from c to a natural distribution:

$$n : X_c \to \mathcal{N}(\mu_c, \sigma_c^2) \tag{7}$$

where:

- $\mathcal{N}(\mu, \sigma^2)$ is normal distribution with average μ and standard deviation σ
- μ_c is expected value of $n(x_c)$
- σ_c is standard deviation of $n(x_c)$

Let C_A denote the set of active classes for a given neuron. The class is active for a neuron when the neuron's activation level deviates significantly from all other inputs:

$$c \in C_A \iff \mathbb{P}(\overline{n}(X_c) \neq \overline{n}(X \setminus X_c)) < \alpha \tag{8}$$

Where α is a significance level.

3.0.1 Justification

The method considered the neuron as a measurement and a random variable, and the feature as a property of the input data that is measured by neurons. Under this assessment, if the feature is typical for a given class of inputs, the measurements, given sufficient sample size, should approximate a normal distribution (central limit theorem, [11]). Alternatively, if the feature is not representative of the class, the neuron activation level distribution for that class should be analogous to the distribution for the whole population of inputs.

This description assumes that the feature is not representative of the majority of the classes, but it can also be reversed, such that distributions of classes not displaying the given feature are disanalogous to the population. The computations are indifferent to the direction of the relation.

By performing an analysis of variance (ANOVA, [12]), it can be determined whether the activation level distribution can be considered to have an equal mean to the remainder of the population within the confidence interval. The samples belonging to the analyzed class must be excluded, as ANOVA assumes the independence of the observations. This requirement is also the basis for the requirement that no sample belongs to more than one class (Equation 3).

In cases where no "passive" level can be defined, and thus none of the class distributions is analogous to the population distribution, all classes can be considered active. In this case, it can be considered that the neuron performs a rudimentary classification of samples.

3.0.2 Class activation band analysis

The class activation band analysis is an extension of the active class analysis. Using (4) and (7), a neuron can be represented as:

$$n : X \to Y = \sum_{c \in C} \mathcal{N}(\mu_c \sigma_c^2) \tag{9}$$

Interpreting the output of a neuron as a multimodal distribution, where each class represents one of the modes, allows the computation of distribution properties that can later be used for more granular analysis.

3.1 Sensitivity and correlation analysis

Using (1) and (5), let $g(x)$ be a function that maps input to its class:

$$g : X \to C \tag{10}$$

In contrast to $f(x)$ that estimates the class of the input x, $g(x)$ is assumed to be correct. In the context of machine learning, $g(x)$ is the label of input data.

3.1.1 Neuron sensitivity

Sensitivity p_1 is defined as the probability of a significant change in neuron activation level, given the change of input class:

$$p_1 = \mathbb{P}(\forall x, y \in X, |n(x) - n(y)| \geq \varepsilon \mid g(x) \neq g(y)) \tag{11}$$

Where ε is the minimal required change in neuron activation level that can be considered significant.

Estimator $\hat{p_1}$ of p_1 was calculated as:

$$\hat{p_1} = \frac{\sum\limits_{x,y \in X} 1_{\{|n(x)-n(y)|>\varepsilon \wedge g(x) \neq g(y)\}}}{\sum\limits_{x,y \in X} 1_{\{g(x) \neq g(y)\}}} \tag{12}$$

3.1.2 Neuron correlation

Correlation p_2 is defined, as probability of network output $f(x)$ changing, given significant change in neuron activation level $n(x)$:

$$p_2 = \mathbb{P}(\forall x, y \in X, f(x) \neq f(y) | \, |n(x) - n(y)| > \varepsilon) \tag{13}$$

Estimator $\hat{p_2}$ of p_2 was calculated as:

$$\hat{p_1} = \frac{\sum\limits_{x,y \in X} 1_{\{f(x) \neq f(y) \wedge |n(x)-n(y)|>\varepsilon\}}}{\sum\limits_{x,y \in X} 1_{\{|n(x)-n(y)|>\varepsilon\}}} \tag{14}$$

3.1.3 Justification

The p_1 parameter is referred to as the sensitivity of the neuron, as it assesses the relation between the network input and the neuron activation level (analogous to sensitivity analysis). The p_2 parameter, in turn, assesses the relation between the neuron activation level and the output of the network and hence is referred to as correlation.

For the calculation of sensitivity, due to a mismatch in data dimension (3-dimensional matrix and scalar), only the class of the sample is considered. This simplification, in turn, results in a difference in cardinality, preventing the use of derivative-based methods.

The use of a statistical approach and introduction of a minimal neuron activation level change threshold ε are attempts to address these limitations.

3.1.4 Limitations

The sensitivity and correlation parameters cannot be calculated, in a practical sense, for the entire population, as this would require conducting analysis for each combination of values, and access to an accurate function $g(x)$. As a result, only estimators $\hat{p_1}$ and $\hat{p_2}$ could be calculated.

An additional factor limiting the value of the gained information is the high interdependence of the neurons. As mostly linear operations, memory-less sub-systems of the network, one neuron's outputs propagate through the network, affecting subsequent neurons, making independent analysis of one neuron impossible.

4. Testing and Visualization

The analysis was tested using a convolutional neural network performing a classification task. Due to the high quantity of numerical results, their representation required a custom visualization method.

4.1 Testing system and parameters

For the testing, a deep CNN was created (Appendix 7.1). It was trained for 50 epochs using the ”cifar10” dataset. After each epoch, a validation subset was used to collect data on activation levels for each of the convolution layers and the fully connected layer. The validation set consisted of 500 total images sampled from the testing set and maintained constant between epochs. This number resulted in an average of 50 images per class, allowing for the assumption that the distributions approach normality (thanks to the central limit theorem).

The convolution neuron output was reduced to a single number by taking the average value from the output matrices.

For the active class analysis, the equality of means was tested using Welch’s t-test, as equality of variances was not guaranteed. Confidence interval α was set to 5%.

For the neurons’ sensitivity and correlation analysis, ε was selected on a neuron-by-neuron basis, equal to the standard deviation for all classes (combined).

4.2 Visualization

Results of neuron-by-neuron calculations are presented in the form of a structured heatmap (Figure 1). The color scale of the heatmap is normalized to the range of presented values, so dark blue represents the lowest value in the displayed range, and dark red represents the highest value in the displayed range. Each band of the heatmap represents one layer of the network, and each cell within the band represents a single neuron. Range, and aggregated values are included in tables following the heatmap (Table 1).

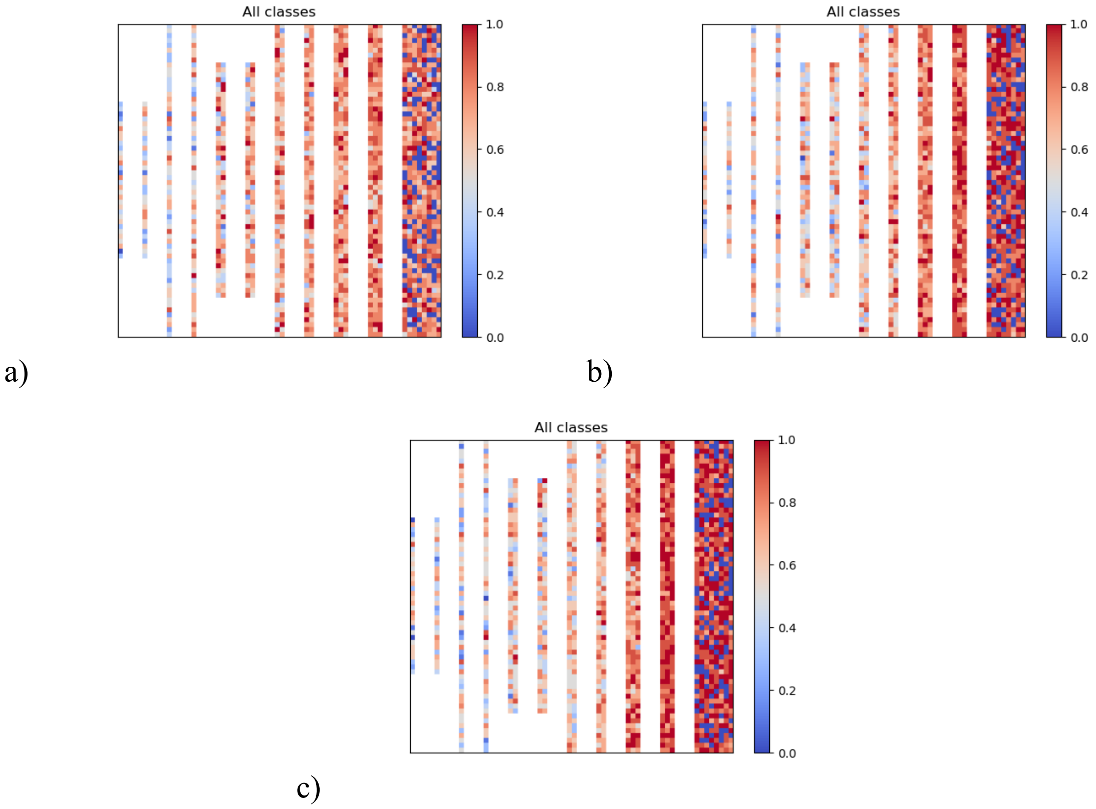


Figure 1. Example of the structured heatmap used for visualization

Table 1. Example table with the contextual information for the presented heatmap

Epoch	1	25	50								
Min	0	0	0								
Max	10	10	10								
Average number of active classes by layer											
Epoch	Layer										
	1	2	3	4	5	6	7	8	9	10	11
1	4.7	5.0	4.9	5.9	6.2	6.5	6.8	7.1	7.5	7.9	5.9
25	4.9	5.3	5.6	5.4	5.8	5.8	6.2	7.0	8.0	8.8	6.6
50	5.0	5.2	5.1	5.3	5.8	5.8	6.0	6.7	8.0	9.1	6.8

5. Results and Analysis

A significant section of the results was affected by ”dead” neurons, introducing bias, especially in the last layer of the network. To accommodate this, whenever applicable, results with and without the dead neurons are included, where filtered results are denoted with the ”*” in the layer number.

5.1 Neuron’s active classes analysis

The results of the active class analysis of the neurons can be divided into two groups: class-specific binary representation of neurons activated by a given class (Figure 2), and generalized, where only the number of classes activating the given neuron is shown, without differentiation by which of the classes (Figure 3 and Table 2).

Generalized results clearly show the increase in the average number of active classes of a neuron in later layers of the network. This result follows the expectation that early neurons of a CNN are activated by generic features that might not be attributable to any specific class, while late neurons are more specialized and thus respond to an input belonging to a different class, in a measurably different way.

Comparing the results between the epochs, an increase in specialization can also be observed as a steeper gradient between the layers.

On an individual neuron level, two phenomena can be observed: ”dead” neurons activated by no class, and ”hyperactive” neurons activated by all classes. In both cases, a more detailed analysis is required to determine the cause of their behavior. For dead neurons, they can be either activated to the same level by all classes or not activated at all. Hyperactive neurons, on the other hand, can display either a distinct activation level for each of the classes, or a low number of activation levels shared between the classes in such a way that neither can be considered a ”passive” level.

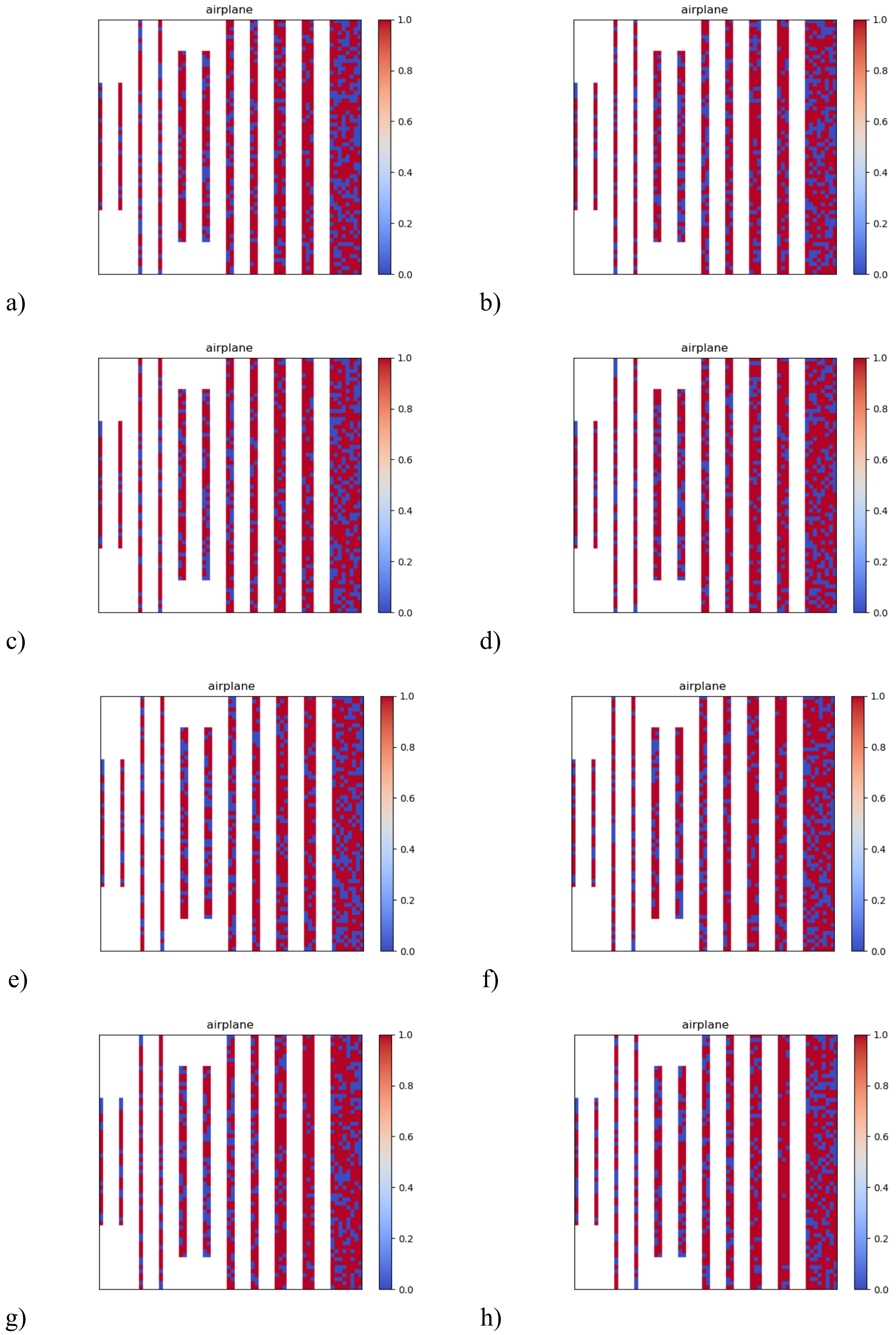


Figure 2. Neuron activation map for class ”airplane” for epochs a)1, b)2, c)3, d)5, e)7, f)10, g)25, h)50

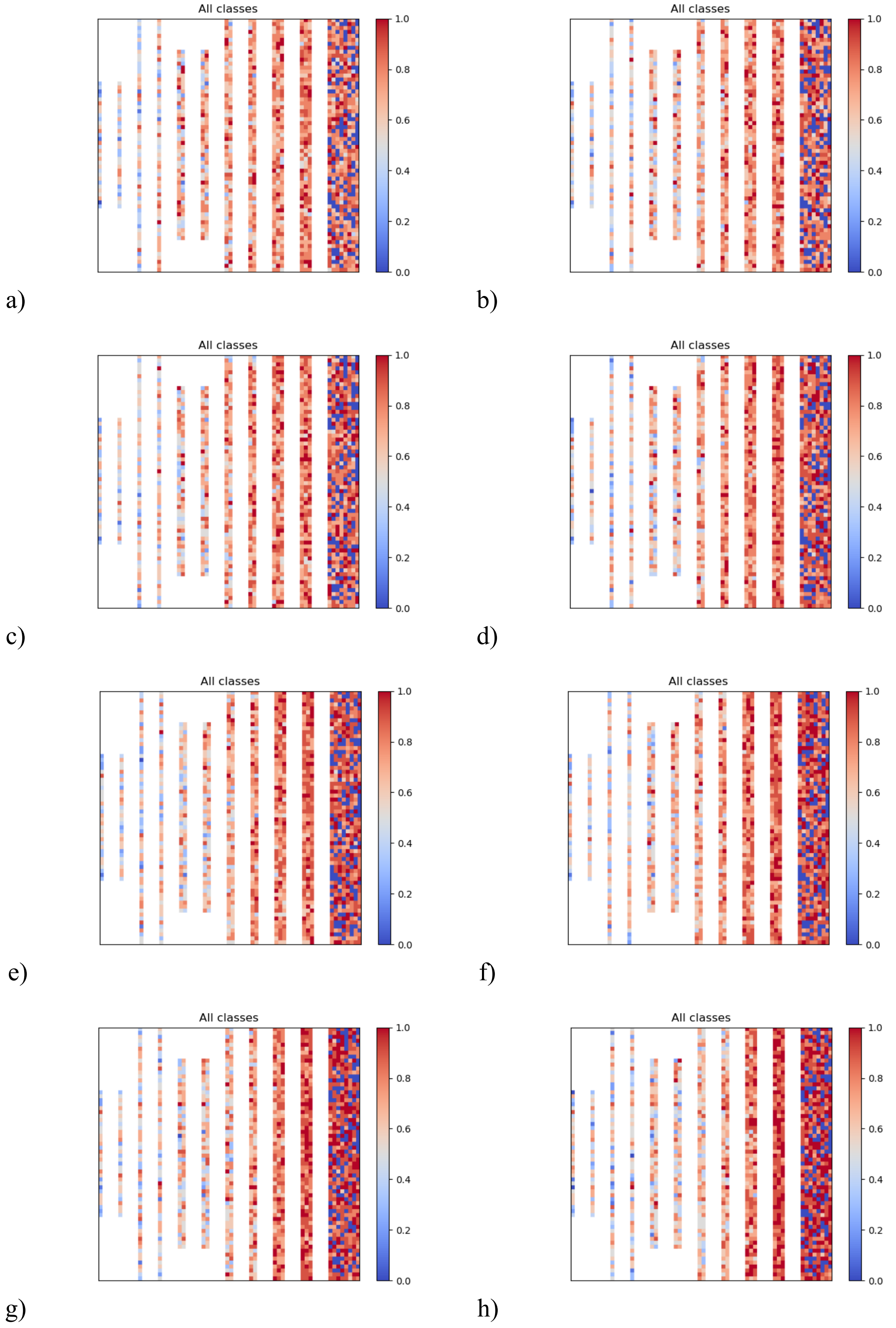


Figure 3. Heatmaps of the neuron sensitivity for epochs a)1, b)2, c)3, d)5, e)7, f)10, g)25, h)50

Table 2. Heatmap range, and averages by layer for generalized active class analysis

Epoch	1	2	3	5	7	10	25	50				
Min	0	0	0	0	0	0	0	0				
Max	10	10	10	10	10	10	10	10				
Average number of active classes by layer												
Epoch	Layer											
	1	2	3	4	5	6	7	8	9	10	11	11*
01	4.7	5.0	4.9	5.9	6.2	6.5	6.8	7.1	7.5	7.9	5.9	7.8
02	4.7	5.2	5.2	5.8	6.1	6.1	6.5	7.4	7.6	7.9	5.9	8.1
03	4.7	5.3	5.2	5.7	6.0	6.3	6.7	7.5	7.8	8.0	6.0	8.2
05	4.8	5.4	5.1	5.5	5.9	6.0	6.7	7.5	7.9	8.3	6.3	8.4
07	4.7	5.2	5.1	5.6	5.6	6.1	6.6	7.5	7.9	8.4	6.3	8.5
10	5.0	5.4	5.1	5.6	5.8	6.3	6.7	7.3	8.1	8.5	6.4	8.7
25	4.9	5.3	5.5	5.4	5.8	5.8	6.2	7.0	8.0	8.8	6.6	8.9
50	5.0	5.2	5.1	5.3	5.8	5.8	6.0	6.7	8.0	9.1	6.8	9.0

5.2 Sensitivity and Correlation Analysis

Neuron sensitivity analysis (Figure 4 and Table 3) shows that neuron sensitivity does not change significantly between most of the layers and all training epochs, and remains centered within the $0.4 - 0.5$ range. Lower sensitivity of the last two layers constitutes a notable exception to this rule.

This trend contradicts the expectation that the later layers should be more sensitive to the input class since they represent the higher-level information. Low sensitivity of the early neurons is to be expected due to the simplicity of the detected feature, which is likely shared between classes, but with the increase in the specificity, the relation between the feature and the input class was expected to grow.

Lower-than-expected values of the sensitivity of the neurons suggest that the network operation and the decision-making process are based on features that are not specific to a class. This assessment would be in agreement with the results of the active class analysis.

Additionally, it indicates that the assumption of $\varepsilon = \sigma$ was incorrect, and that ε should be lower to detect smaller changes in the neuron activation level.

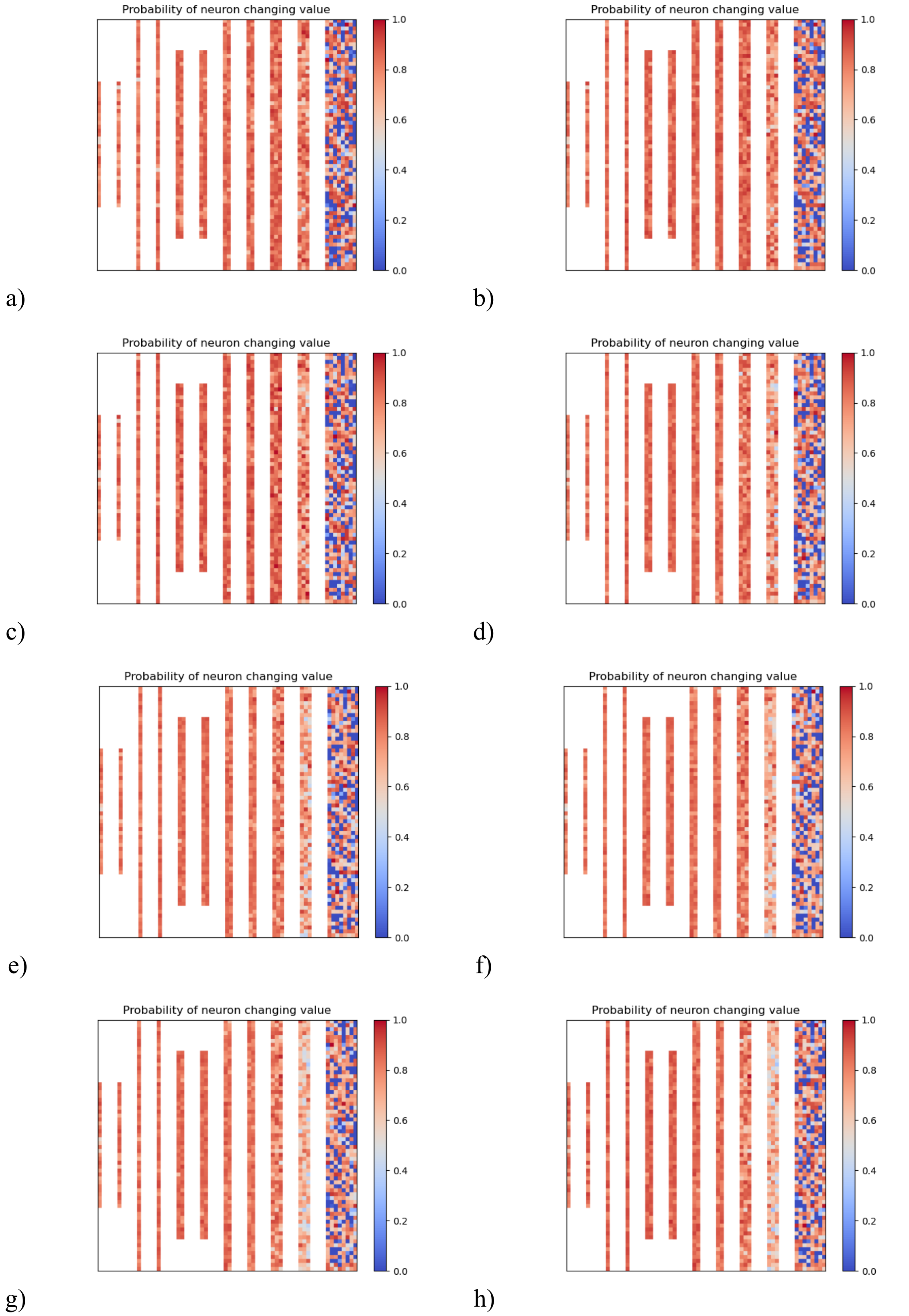


Figure 4. Heatmaps of the neuron sensitivity for epochs a)1, b)2, c)3, d)5, e)7, f)10, g)25, h)50

Table 3. Heatmap ranges, and average neuron sensitivity by layer

Epoch	1	2	3	5	7	10	25	50			
Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0			
Max	0.55	0.54	0.54	0.55	0.55	0.55	0.55	0.53			
Average neuron sensitivity by layer											
Epoch	Layer										
	1	2	3	4	5	6	7	8	9	10	11	11*
01	0.43	0.44	0.45	0.46	0.46	0.47	0.46	0.46	0.47	0.43	0.29	0.39
02	0.44	0.45	0.46	0.47	0.47	0.47	0.46	0.47	0.46	0.42	0.28	0.39
03	0.44	0.45	0.46	0.47	0.46	0.47	0.46	0.47	0.46	0.41	0.28	0.38
05	0.44	0.45	0.46	0.47	0.47	0.47	0.46	0.46	0.45	0.4	0.28	0.37
07	0.44	0.45	0.46	0.47	0.47	0.47	0.46	0.46	0.44	0.38	0.28	0.38
10	0.44	0.45	0.46	0.47	0.47	0.47	0.46	0.46	0.45	0.38	0.28	0.38
25	0.44	0.46	0.46	0.47	0.47	0.47	0.46	0.45	0.43	0.35	0.28	0.38
50	0.44	0.45	0.46	0.47	0.47	0.46	0.45	0.45	0.42	0.34	0.28	0.38

In contrast, the relation between the feature, represented by neuron activation level, and the output class, was consistently high with the exception of the dead neurons, as shown in the neuron correlation analysis results (Figure 5 and Table 4).

High values of p_2 across epochs were expected and are explained by the causal relation between neurons and the output. Low correlation would suggest a seemingly random operation of the network.

Analyzing the results of both sensitivity and correlation for a given epoch at the same time shows an interesting case, where a neuron appears to be fully insensitive ($p_1 \approx 0$), but remains highly correlated ($p_2 \approx 1$). Specific analysis of activation levels of those neurons indicates that this edge case is caused by neurons activated by a very low number of inputs (≤ 3), suggesting inherent sensitivity of the method to outliers.

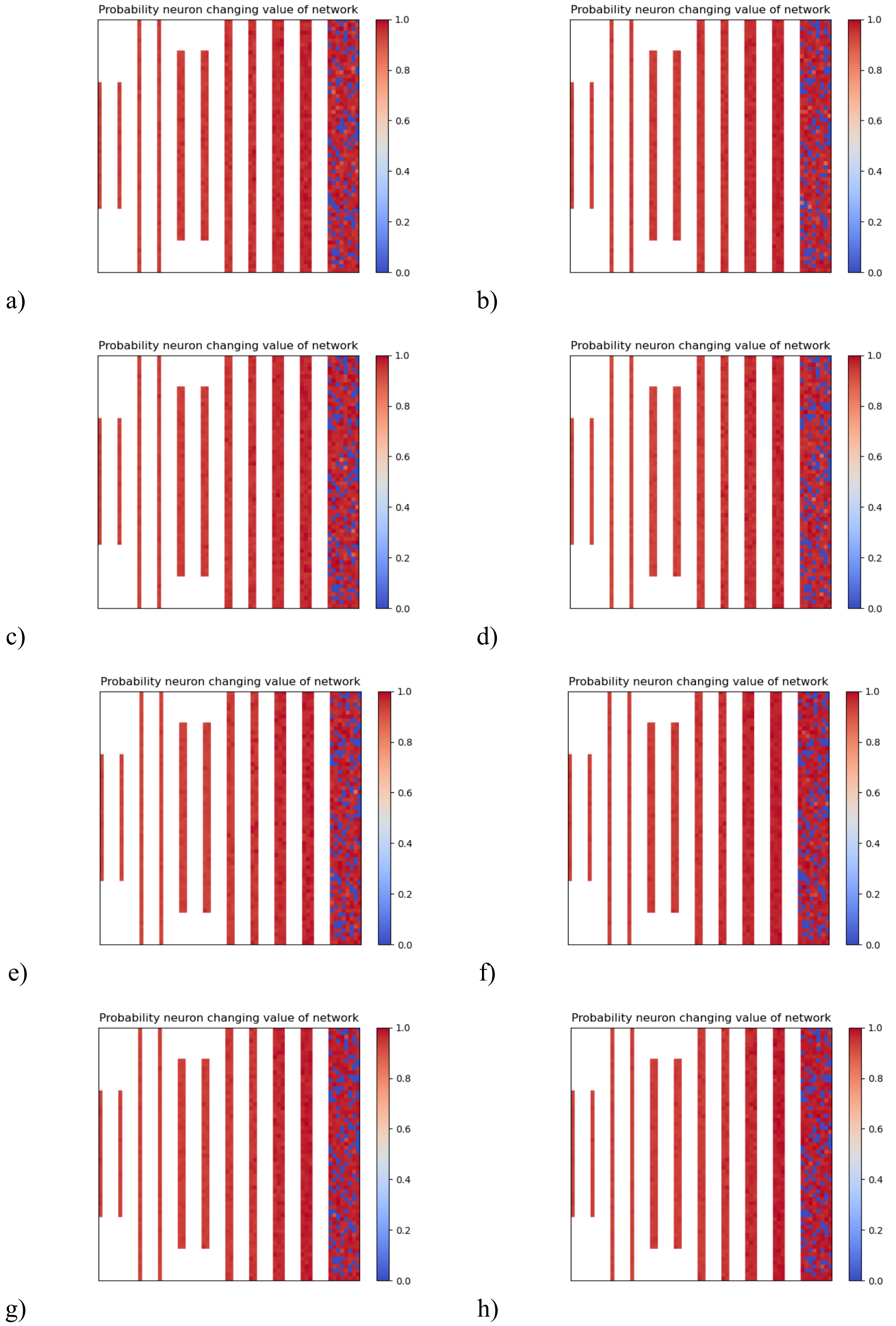


Figure 5. Heatmaps of the neuron correlation for epochs a)1, b)2, c)3, d)5, e)7, f)10, g)25, h)50

Table 4. Heatmap range, and average neuron correlation by layer

Epoch	1	2	3	5	7	10	25	50				
Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0				
Max	0.55	0.54	0.54	0.55	0.55	0.55	0.55	0.53				
Average neuron correlation by layer												
Epoch	Layer											
	1	2	3	4	5	6	7	8	9	10	11	11*
01	0.9	0.9	0.9	0.9	0.91	0.91	0.91	0.92	0.93	0.94	0.72	0.93
02	0.9	0.9	0.9	0.9	0.91	0.91	0.91	0.92	0.93	0.93	0.72	0.93
03	0.9	0.9	0.9	0.9	0.9	0.91	0.91	0.92	0.93	0.93	0.71	0.93
05	0.9	0.9	0.9	0.9	0.9	0.9	0.91	0.92	0.93	0.94	0.71	0.94
07	0.9	0.9	0.9	0.9	0.9	0.9	0.91	0.92	0.93	0.94	0.7	0.94
10	0.91	0.91	0.91	0.91	0.91	0.91	0.92	0.93	0.94	0.95	0.71	0.95
25	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.92	0.94	0.95	0.7	0.95
50	0.91	0.91	0.91	0.91	0.91	0.91	0.92	0.92	0.94	0.95	0.72	0.95

5.3 Active Band Analysis

The analysis of the activation levels as their own distinct distributions provided supplementary information to the information gathered in the neuron active class analysis.

Among others, it allowed for a more granular division of the neurons based on their activation level distributions. The majority of the neurons can be divided into one of five groups:

- Dead neurons, with activation level for all of the inputs equal to 0 (Figure 6a).
- Inactive neurons, with activation level within the common band (different from 0) for all provided inputs (Figure 6b).
- Binary neurons, with two activation levels, one of which is frequently equal to 0 for the majority of inputs (Figure 6c).
- Multi-stable neurons, with more than 2 bands without overlap in the first to third quartile range (Figure 6d).
- Continuous neurons, with overlapping bands that are still distinct (Figure 6e).

For visualization, each of the neurons is presented in the form of a box and whiskers plot.

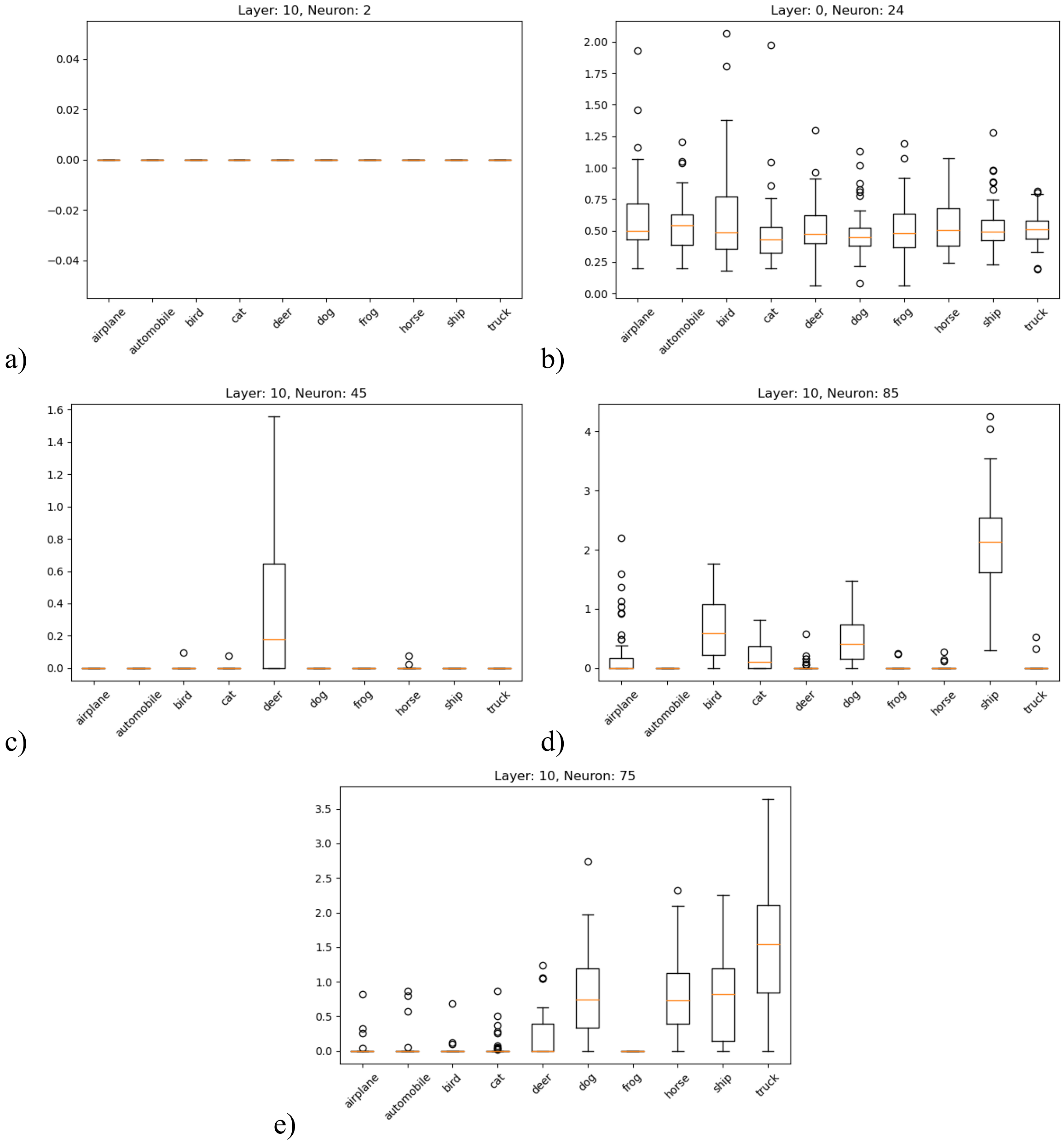


Figure 6. Types of observed neurons based on their activation distributions: a)”dead”, b)inactive, c)binary, d)multi-stable, e)continuous

Table 5. Average parameters of all distributions within the network

Epoch	Normality W	Kurtosis	Skewness
01	0.87	2.42	1.05
02	0.87	2.56	1.07
03	0.86	2.82	1.1
05	0.86	3.28	1.17
07	0.85	3.44	1.21
10	0.85	3.58	1.22
25	0.84	4.2	1.34
50	0.83	4.5	1.41

In the early layers, mainly the continuous neurons were observed, whereas in the later layers all listed types of neurons were present.

5.4 Additional Parameters

Using the neuron activation levels divided into the class-specific distributions, additional distribution parameters were calculated (Figures 7, 8, 9 and Tables 6, 7, 8), and agregated for entire network (Table 5). While they do not provide further information on the network operation, they were used to validate the method assumptions.

The normality of the distribution is the result W of the Shapiro-Wilk test, and the kurtosis was calculated using Fisher's definition (Fourth central moment of the distribution reduced by 3, 0 for a normal distribution)

High normality indicates that the assumption in (7) was correct. Positive skewness was expected due to the truncation of negative outputs to 0 by the ReLU activation function (as described in (1)). The leptokurtic nature of the distributions, as indicated by high kurtosis, suggests a higher probability of the activation level value being close to the average. While this deviates from the expectation of a normal distribution, it indicates that distributions, on average, form narrow bands. This, in turn, provides additional credence to the active band analysis method.

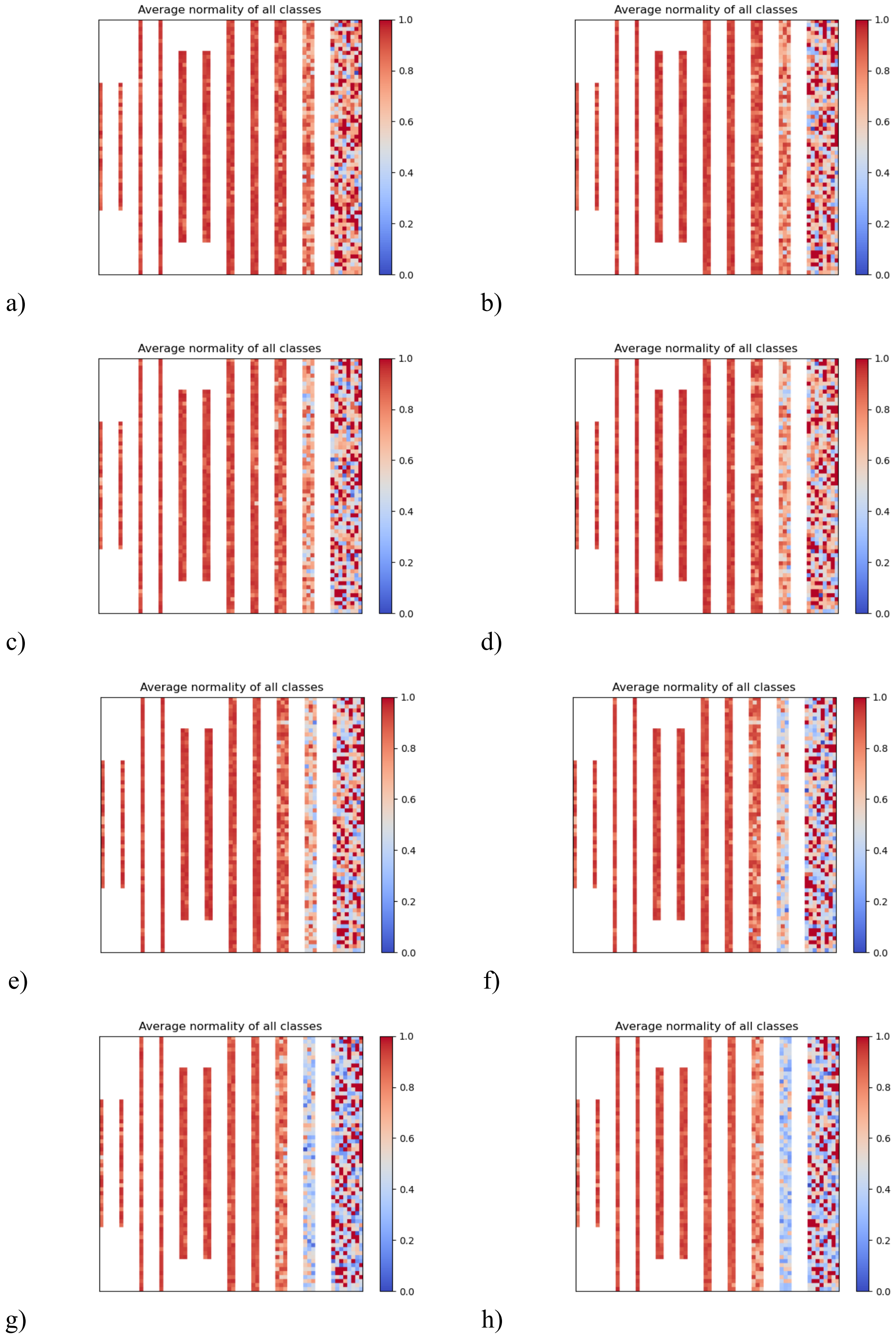
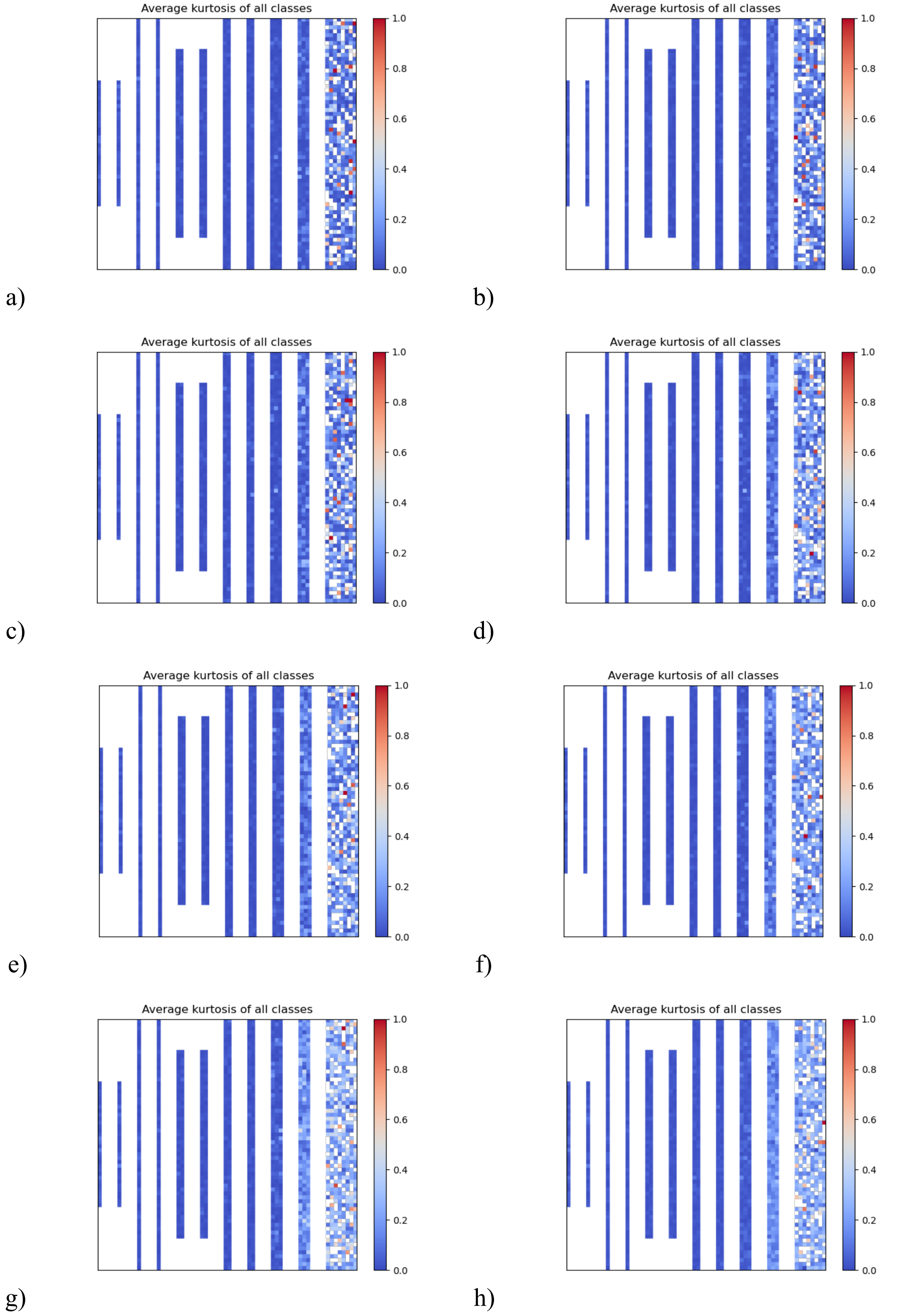


Figure 7. Heatmaps of the neuron average normality of class distribution for epochs a)1, b)2, c)3, d)5, e)7, f)10, g)25, h)50



a)

b)

c)

d)

e)

f)

g)

h)

Figure 8. Heatmaps of the neuron average kurtosis of class distribution for epochs a)1, b)2, c)3, d)5, e)7, f)10, g)25, h)50

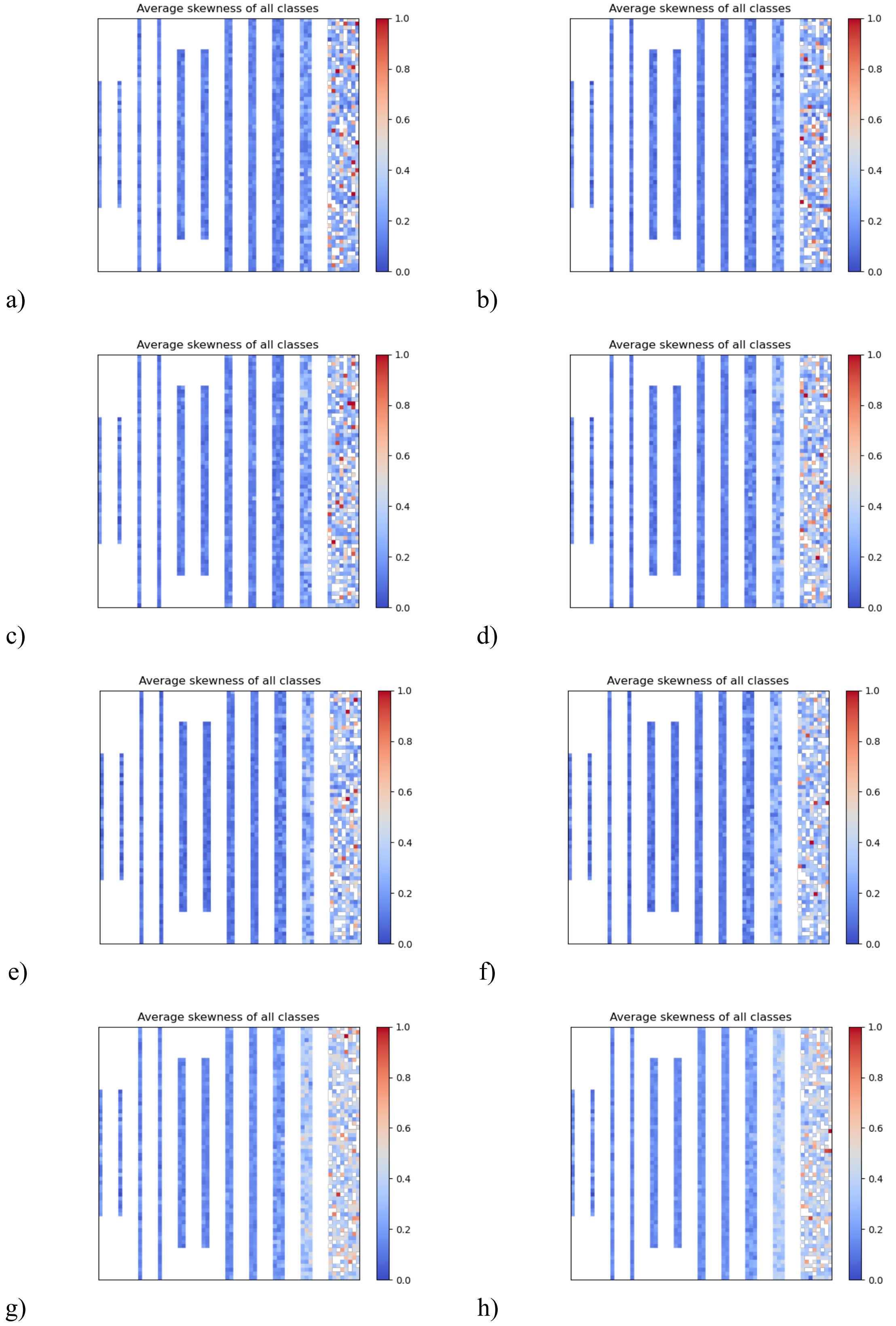


Figure 9. Heatmaps of the neuron average skewness of class distribution for epochs a)1, b)2, c)3, d)5, e)7, f)10, g)25, h)50

Table 6. Heatmap range, and average normality of class distribution by layer

Epoch	1	2	3	5	7	10	25	50				
Min	0.27	0.31	0.37	0.29	0.32	0.4	0.43	0.42				
Max	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0				
Average neuron normality by layer												
Epoch	Layer											
	1	2	3	4	5	6	7	8	9	10	11	11*
01	0.91	0.92	0.94	0.95	0.94	0.95	0.94	0.94	0.92	0.84	0.77	0.7
02	0.91	0.93	0.94	0.95	0.95	0.95	0.95	0.95	0.92	0.82	0.77	0.69
03	0.91	0.94	0.95	0.95	0.94	0.95	0.94	0.94	0.91	0.79	0.77	0.69
05	0.92	0.94	0.94	0.95	0.95	0.96	0.95	0.94	0.9	0.76	0.76	0.68
07	0.92	0.94	0.95	0.96	0.95	0.96	0.94	0.94	0.9	0.74	0.76	0.67
10	0.92	0.94	0.95	0.96	0.95	0.96	0.94	0.94	0.89	0.73	0.76	0.68
25	0.92	0.94	0.95	0.95	0.95	0.95	0.94	0.94	0.88	0.69	0.75	0.66
50	0.92	0.94	0.95	0.95	0.95	0.95	0.93	0.93	0.86	0.66	0.75	0.67

Table 7. Heatmap range, and average kurtosis of class distribution by layer

Epoch	1	2	3	5	7	10	25	50				
Min	-0.64	-0.61	-0.48	-0.59	-0.55	-0.6	-0.51	-0.86				
Max	52.02	52.02	49.02	49.81	52.02	52.02	42.94	51.03				
Average neuron kurtosis by layer												
Epoch	Layer											
	1	2	3	4	5	6	7	8	9	10	11	11*
01	1.25	1.06	0.74	0.47	0.4	0.35	0.38	0.36	0.53	2.33	6.85	6.62
02	1.16	0.76	0.59	0.4	0.4	0.34	0.39	0.38	0.66	2.74	7.24	6.92
03	1.2	0.76	0.56	0.39	0.45	0.34	0.59	0.53	0.76	3.82	7.63	7.38
05	1.16	0.63	0.64	0.35	0.43	0.29	0.43	0.47	1.02	4.75	8.88	8.8
07	1.1	0.68	0.52	0.22	0.44	0.27	0.46	0.4	1.12	5.26	9.32	9.19
10	1.18	0.62	0.5	0.25	0.44	0.3	0.44	0.49	1.2	5.67	9.7	9.61
25	1.19	0.74	0.52	0.28	0.39	0.39	0.5	0.45	1.45	7.06	11.31	11.29
50	1.17	0.78	0.48	0.37	0.47	0.45	0.6	0.57	1.53	7.78	12.09	12.03

Table 8. Heatmap range, and average skewness of class distribution by layer

Epoch	1	2	3	5	7	10	25	50				
Min	-0.37	-0.27	-0.33	-0.35	-0.17	-0.18	-0.42	-0.53				
Max	7.35	7.35	7.14	7.16	7.35	7.35	6.7	7.25				
Average neuron skewness by layer												
Epoch	Layer											
	1	2	3	4	5	6	7	8	9	10	11	11*
01	0.97	0.78	0.71	0.59	0.61	0.6	0.61	0.63	0.63	1.17	1.94	1.91
02	0.92	0.64	0.62	0.51	0.59	0.56	0.61	0.61	0.66	1.25	2.0	1.96
03	0.92	0.6	0.61	0.51	0.61	0.54	0.65	0.64	0.68	1.43	2.05	2.01
05	0.87	0.54	0.65	0.47	0.57	0.48	0.6	0.63	0.78	1.63	2.19	2.18
07	0.84	0.56	0.6	0.46	0.58	0.51	0.64	0.65	0.83	1.76	2.26	2.24
10	0.85	0.54	0.62	0.47	0.59	0.52	0.65	0.66	0.84	1.82	2.28	2.27
25	0.86	0.52	0.6	0.52	0.58	0.6	0.69	0.69	0.97	2.1	2.47	2.47
50	0.84	0.55	0.6	0.57	0.64	0.65	0.76	0.77	1.04	2.25	2.55	2.54

6. Critical Appreciation

The neuron's active classes analysis allows for visualization of the data propagation within the artificial neural network. While in this work, it was tested using the output classes as the parameter; it is not a requirement, and an arbitrary set of classes can be used. Furthermore, if some of the assumptions were to be relaxed, it is possible to use this method in conjunction with feature visualization or semantic dictionaries ([8],[9]) to provide a more comprehensive explanation of neural network inner workings. An additional benefit of the method is the independence of the calculation, facilitating its parallelization.

The fact that this method of analysis provides information on the population rather than on individual cases can also be considered its comparative advantage, as several methods can provide justifications on a case-by-case basis ([5]).

Some of the main disadvantages of all the presented methods are the requirement for access to the internal network parameters and the low legibility of the generated results, demanding a higher level of understanding of the process for utilizing it effectively.

Specific to the sensitivity and correlation analysis, was its susceptibility to outliers, and the difficulty of proper selection of the ε parameter. A more robust method without the need for careful consideration of the parameters could be able to provide the intended information more reliably.

7. Discussion and Conclusion

The neuron's active classes analysis and the active band analysis provided useful information that followed the expectation, suggesting the validity of the methods. Utilizing the gathered information, the logical next step is to implement the method in conditions where the results are not known, so the practical utility can be tested. The previously mentioned concept of the semantic dictionaries is of particular interest.

The neuron sensitivity and correlation method failed to deliver the expected information and requires deeper reconsideration. A potential avenue for consideration could be a shift in perspective, instead of calculating the probability of an event, calculate the average response to the event in case of the sensitivity, or the average required change to cause the event in the case of correlation. If normalized for the range, this approach could provide the demanded information on the strength of the relation between the network input and the neuron, and the neuron and the network output.

The active band analysis provides an additional path of future development, in the form of the use of XAI for improvements in network efficiency. If the activation bands were to be used similarly to their analog signal counterparts, like the signal source can be identified by the generated frequency, the output class could be identified by its neuron activation level. If this identification can be done uniquely for all inputs, using activation levels of early to mid network neurons, it could be possible to drastically reduce the required network size and thus its computation complexity.

On the whole, the presented framework and analysis provide a crucial first step in the development of the mechanistic interpretability of artificial neural networks using statistical analysis.

References

[1] Singla A., Sukharevsky A., Yee L. A., and Chui M. The state of AI in 2025: Agents, Innovation, and transformation. Nov. 2025. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.

[2] Ofosu-Ampong K. Artificial intelligence research: A review on dominant themes, methods, frameworks and future research directions. *Telematics and Informatics Reports* 14 (2024), p. 100127. DOI: https://doi.org/10.1016/j.teler.2024.100127. https://www.sciencedirect.com/science/article/pii/S2772503024000136.

[3] Supervisor E. D. P. TechDispatch – Explainable artificial intelligence. 2/2023. Ed. by Attoresi M., Lareo X., and Velasco L. Publications Office of the European Union, 2023. DOI: doi/10.2804/802043.

[4] Parliament T. E. and Council of the European Union the. Regulation (EU) 2024/1689 of the European Parliament and of the Council. June 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

[5] Haar L. V., Elvira T., and Ochoa O. An analysis of explainability methods for convolutional neural networks. *Engineering Applications of Artificial Intelligence* 117 (2023), p. 105606. DOI: https://doi.org/10.1016/j.engappai.2022.105606. https://www.sciencedirect.com/science/article/pii/S0952197622005966.

[6] Lee J. h., Shin I. h., Jeong S. g., Lee S.-I., Zaheer M. Z., and Seo B.-S. Improvement in Deep Networks for Optimization Using eXplainable Artificial Intelligence. *2019 International Conference on Information and Communication Technology Convergence (ICTC)*. 2019, pp. 525–530. DOI: 10.1109/ICTC46691.2019.8939943.

[7] Kästner L. and Crook B. Explaining AI through mechanistic interpretability. *European Journal for Philosophy of Science* 14.4 (Oct. 2024). DOI: 10.1007/s13194-024-00614-4. http://dx.doi.org/10.1007/s13194-024-00614-4.

[8] Olah C., Mordvintsev A., and Schubert L. Feature Visualization. *Distill* 2.11 (Nov. 2017). DOI: 10.23915/distill.00007. http://dx.doi.org/10.23915/distill.00007.

[9] Olah C., Satyanarayan A., Johnson I., Carter S., Schubert L., Ye K., and Mordvintsev A. The Building Blocks of Interpretability. *Distill* 3.3 (Mar. 2018). DOI: 10.23915/distill.00010. http://dx.doi.org/10.23915/distill.00010.

[10] Olah C., Cammarata N., Voss C., Schubert L., and Goh G. Naturally Occurring Equivariance in Neural Networks. *Distill* 5.12 (Dec. 2020). DOI: 10.23915/distill.00024.004. http://dx.doi.org/10.23915/distill.00024.004.

[11] Feller W. An Introduction to Probability Theory and Its Applications, Volume 2. An Introduction to Probability Theory and Its Applications. Wiley, 1971. https://books.google.ee/books?id=n-kmAQAAIAAJ.

[12] Girden E. R. ANOVA: Repeated measures. 84. Sage, 1992.

Appendices

7.1 Test network structure

Table 9. Structure of the testing network

Input	Image 32x32x3
Convolution 2D	32-32
Convolution 2D	32-32
Max Pooling 2D	
Convolution 2D	32-64
Convolution 2D	64-64
Max Pooling 2D	
Convolution 2D	64-96
Convolution 2D	96-96
Max Pooling 2D	
Convolution 2D	96-128
Convolution 2D	128-128
Max Pooling 2D	
Convolution 2D	128-192
Convolution 2D	192-192
Global Average Pooling 2D	
Fully connected	512
Output	10

License

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Bernard Michał Szeliga,

1. grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis

”Statistical Analysis for Mechanistic Explainability of Artificial Neural Networks”,

supervised by Kallol Roy and Nesma Talaat Abbas Mahmoud;

2. grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright;

3. am aware of the fact that the author retains the rights specified in points 1 and 2;

4. confirm that granting the non-exclusive licence does not infringe other persons’ intellectual property rights or rights arising from the personal data protection legislation.

Bernard Szeliga 31/12/2025

Bernard Szeliga