SynXAI: Synergistic Explainability for Neural Networks

Benedikt Farag
Yale University
New Haven, CT 06511
benedikt.farag@yale.du

Abstract

Trustworthy deployment of neural networks requires understanding about feature importance, but also how the network processes them. We present SynXAI, a novel framework to analyze the synergistic interpretability of Multilayer Perceptrons (MLPs). By defining a quantifiable *Synergy score*, we characterize neurons as synergistic, redundant or independent based on their pairwise interactions. We validate this approach using a real-world fish-catch dataset from the *Icelandic Marine & Freshwater Institute*. Our proposed Synergy Pruning method significantly outperforms the baseline importance pruning, retaining model accuracy (less than 10% drop) with 85% pruned neurons, compared to a $> 45\%$ drop in accuracy for the baseline method. Furthermore, we provide explainability analysis of the MLP model which is, to our knowledge, the first of its kind for fish catch prediction which sheds light on how deep neural networks couple spatial and oceanographic variables to predict Pollock presence in the North Atlantic. Our code is available at: `https://github.com/benedikt20/synXAI`

1 Problem Definition

Here we will investigate neuron interaction in a traditional MLP framework. For these purposes, we will train a simple MLP binary classification network on a real-world dataset on fish catches in the North-Atlantic. The activation layers of the network will be explored in terms of neuron synergy, that we define as follows:

- Two neurons are synergistic if $\Delta_{ij} > \Delta_i + \Delta_j$
- Two neurons are redundant if $\Delta_{ij} < \Delta_i + \Delta_j$
- But independent if $\Delta_{ij} \approx \Delta_i + \Delta_j$

where Δ_k is the absolute change in output score (softmax probability) after masking out neuron k and Δ_{ij} is the change in output score after masking out neurons i and j simultaneously.

This methodology reveals important information about how the neurons collaborate in predicting the output, which is not clear for MLP models which are black-box models. This mechanistic interpretability has not been examined to date and provides a meaningful insight into the cooperative dynamics driving the model's predictions.

In addition to the synergistic explanations, the MLP will be explained to give interpretable explanations on the model's behavior, such as Shapley values, LIME and decision boundaries. The primary goal of this project is to develop a methodology for synergistic explanations of MLPs, while also examining the behavior of the model to predict presence and absence for a fish specie from temporal, spatial and oceanographic variables, which is a challenging task.

2 Relevance to Trustworthy Aspects

This work is focused on the explainability aspects of this course. The work tries to adapt a new method for interpretability purposes by looking at synergistic effects of neurons. Some neurons might benefit from each other, while the other might not depend on other neurons. Few aspects to mention include:

- **Interpretability:** The method tries to explain how the inner structure of the model combines information from other model sources.
- **Robustness:** With knowledge about if neurons of a model are redundant, it can indicate robustness of the model since masking out a robust neuron might not influence the prediction as for a synergistic neuron.
- **Transparency:** By shedding light on how the model structure behaves, it opens up the black box to some extent, providing transparency of the model.

In this project, we will primarily be extending available methods for interpretability of neural networks, by introducing a novel technique on synergistic behavior of neurons.

3 Related work

A number of studies have studied explainability of neural networks at both neuron-level and feature-level. First to mention is SHAP framework that assigns each feature an importance value, and is hence a *feature level explanation* [9]. Another work proposed a generalization of Sapley values with presenting Shapley-Taylor index that quantifies the attribution of interaction of a subset of features of the model, instead of a single one as Shapley values do [3]. Other works have focused on layer-level explanations such as [1], which revealed that different layers in CNNs represent different structures and have different categories of meaning, which is linked to the training techniques. Explanations have also been done on neuron-level, where neuron importance was computed with *conductance* [4], measuring the flow of attributions through a hidden unit. An ablation study revealed that it was sufficient to ablate 3.7 filters on average to change the label, over 100 images from ImageNet.

For the synergy analysis outlined above, it will be used to perform pruning of the MLP. Several works have demonstrated efficient structural pruning methods (i.e. removing structural components such as layers, neurons etc), for example *skeletonization pruning* [12] and *oracle pruning* [11], which measure the sensitivity of training loss to the removal of individual neurons. Other techniques involve ranking neurons based on their total ℓ_1 sum such as [8]. Here we will be using **Importance pruning** as a baseline based on the ranking of the mean absolute difference in output score for every neuron, which is fundamentally different from magnitude based approaches focusing on the weights. While this measures the effect on the output, the majority of pruning literature relies on the loss function as the primary metric for estimating neuron importance [6, 10].

There have also been publications on deep learning architectures for oceanographic catch data. One paper implemented MLP neural network to predict catch weight using only on-board sensors on vessels for real-time predictions [2]. This aligns with the proposed technique in this project to train a MLP on tabular catch data. However other architectures such as convLSTM and CNN have also been found effective for fish prediction [14, 5].

4 Proposed Approach

4.1 Model Architecture and Training

To investigate the neuron-synergy mechanism, we train a Multilayer Perceptron (MLP) to predict the binary presence of Pollock (a demersal fish specie) from tabular oceanographic data. The network consists of two hidden layers, with 128 and 64 neurons, each with batch normalization and `ReLU` activation. The output of the network is a sigmoid probability score, which is transformed into binary labels with a threshold of 0.5. The architecture is illustrated in Figure 1.

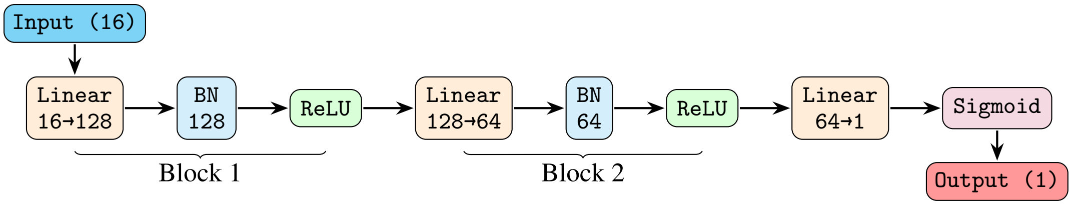


Figure 1: MLP architecture

The model is trained using binary cross entropy (BCE) loss with the Adam optimizer [7] using a learning rate of $5 \cdot 10^{-4}$. A temporal split was used for training-, validation- and test sets, with training range from 2010-2022, validation on 2023 and test on 2024-2025. The model was trained with an early stopping (5 iterations) to prevent overfitting.

4.2 Neuron Synergy Quantification

Our primary contribution is a framework to quantify the synergistic behavior of neurons within layer. This aims to shed light on the mechanistic interpretability of neural networks, that is important for understanding the behavior of the network at inference time. We define a neuron importance Δ_i as the mean absolute change in the model's output probability when that neuron is masked (where the neuron's activation is set to zero).

The synergy analysis was conducted for each activation layer. We define a synergy score for every combination of neurons in each layer as

$$S_{ij} = \Delta_{ij} - (\Delta_i + \Delta_j) \tag{1}$$

that represents the absolute difference in the output score by masking out two neurons versus the cumulative effect of masking out each neuron separately. Algorithm 1 demonstrates this process.

Algorithm 1 Neuron Synergy

```
Require: Model $f$, Dataset $X$, Layer neurons $N$
Ensure: Synergy matrix $S$
$P_{\text{base}} \leftarrow f(X^{\text{test}})$
Step 1: Single Neuron Importance
for each neuron $i \in N$ do
    $P_i \leftarrow f(X \setminus \{i\})$
    $\Delta_i \leftarrow \frac{1}{|X|} \sum |P_{\text{base}} - P_i|$
end for
Step 2: Pairwise Synergy
for each pair $(i, j)$ where $i < j$ do
    $P_{ij} \leftarrow f(X \setminus \{i, j\})$
    $\Delta_{ij} \leftarrow \frac{1}{|X|} \sum |P_{\text{base}} - P_{ij}|$
    $S_{ij} \leftarrow \Delta_{ij} - (\Delta_i + \Delta_j)$
end for
return $S$
```

The neurons are then ranked in ascending order by synergistic-importance, where the synergy scores S_{ij} were sorted by the ReLU-sum, i.e.

$$\mathcal{R}_i = \sum_j \text{ReLU}(S_{ij}) = \sum_j \max(0, S_{ij}) \tag{2}$$

since synergistic neurons have a positive synergy score $S_{ij} > 0$ that benefit from other neurons.

4.3 Model Explainability

The synergistic explainability outlined here above provides information about how the network operates. However it does not provide any information about feature importance or which features drive the predictions. For trustworthy aspects of the model, we will also perform interpretable analysis at instance level and model level, by showing decision boundaries in a two dimensional space. Shapley values and LIME [13] are very commonly used methods for explaining feature importance of black box models, e.g. [9].

5 Experiment

5.1 Dataset

The dataset that we will be using for this project is a real-world dataset from the *Icelandic Marine & Freshwater Institute* for historical catches of demersal fish species in the North-Atlantic Ocean around Iceland. The dataset contains about 12M catch records from Icelandic fisheries dating back to 1960s. The raw data contained information about species caught, latitude/longitude coordinates, weight (in kilograms), and the catch date. A subset of the data will be used herein, with records after `2010-01-01`, until `2025-08-31`.

The data was grouped by hauls (with same latitude, longitude and date) with aggregated species and associated weight. The parameter of interest was computed as the *ratio of Pollock in every haul* as:

$$r = \frac{\text{weight of Pollock}}{\text{total haul weight}}$$

and for hauls without any Pollock caught, this ratio was set to zero. This parameter r gives information about *the density of Pollock* in each haul that is a metric to be used on the presence of Pollock in each haul. The resulting Pollock ratio r is shown spatially in Figure 2, particularly dominant south and west of Iceland.

The data was augmented with oceanographic variables such as temperature and oxygen at ocean bottom from *Copernicus Marine Data Store* . This was saved as an csv file to be used for training the MLP network shown in Figure 1.

The resulting data contained 16 input features which are summarized in Table 1. The target variable is the *Pollock ratio*, r, in each haul, which was made binary with a threshold of 0.1 (i.e. 10% Pollock of total weight of each haul). The training data was split temporally into training (1,016,881), validation (65,776) and test (108,489) with a class distribution of 78.4% non-Pollock ($r < 0.1$) and 21.6% Pollock ($r \geq 0.1$).

Table 1: Input features used in the model

Variable	Description	Unit
latitude	Latitude	$^\circ$
longitude	Longitude	$^\circ$
depth	Depth	m
thetao	Temperature	$^\circ C$
uo	East velocity	m/s
vo	North velocity	m/s
so	Salinity	g/kg
thetao_grad	Temp. gradient	$^\circ C/$unit
chl	Chlorophyll	mg/m^3
no3	Nitrate	$mmol/m^3$
nppv	Primary production	$mmol/m^3/$day
o2	Oxygen	$mmol/m^3)$
po4	Phosphate	$mmol/m^3$
si	Silicate	$mmol/m^3$
day_cos	cos(day-of-year)	-
day_sin	sin(day-of-year)	-

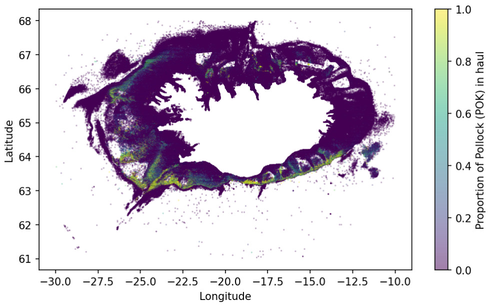


Figure 2: Pollock ratios of haul weight

5.2 Training Results

The two layer-block MLP network in Figure 1 was trained while monitoring for validation loss, where the best checkpoint occurred at epoch 14. The train-loss, val-loss and val-accuracy are shown in Appendix (Figure A.1). The model achieved a ROC-AUC score of 0.90 and an accuracy of 0.85, with other metrics reported in Appendix (Table A.1). While the model's performance would ideally be slightly better, explainability methods would still provide meaningful insight into the model.

5.3 Baselines

To justify the use of MLP for this classification problem, a few simpler methods were implemented for this same dataset. Logistic regression, KNN, random forest and XGBoost yielded ROC-AUC scores of 0.824, 0.828, 0.908 and 0.908 respectively on the test set. The random forest and XGBoost models provided comparable performance as the MLP which achieved a best ROC-AUC of 0.905. However for the mechanistic explainability purposes of the neural network the performance is competitive to other model architectures.

The performance of the model was analyzed in Table A.2 by adding features recursively demonstrated that the latitude and longitude alone achieved an ROC-AUC score of 0.886 while other variables do not seem to contribute a lot to the performance of the model. This is not surprising as the Pollock is rather stationary in specific areas around Iceland, that does vary by season.

To act as a pruning baseline, we used the average difference in the output (probability) score of the MLP by ablating every neuron in each layer. Histograms of the distribution of the mean absolute output probability score of the MLP, in each layer, is shown in Appendix (Figure A.2). The neurons were ranked in ascending order based on the neuron importance for each of the layers, and pruned from the lowest importance to highest, see Appendix Figures A.3 and A.4, to compare importance pruning to random pruning based on ROC-AUC and accuracy. The importance pruning appeared to be much more effective than random pruning, maintaining performance with a number of neurons pruned.

5.4 Synergy Results

The computed synergy scores S_{ij} by eq. (1) for each of the activation layers are shown in Figure 3, ordered by ascending synergistic-importance $\mathcal{R}$ by eq. (2). The synergy scores are much stronger in the second activation layer compared to the first activation layer. Given the complexity of the task of predicting the presence and absence of Pollock, this indicates how the network is strongly dependent on the neuron interactions of the second layer. This attribute has been well established in the literature where neural networks have more profound compositional feature representations in deeper layers [15]. The block-like regions in the second activation function suggest that there is a set of 21 redundant neurons (the blue). The blue (redundant) interactions provide similar information for the prediction while the red (synergistic) neurons focus on different patterns of the inputs that are dependent on each other.

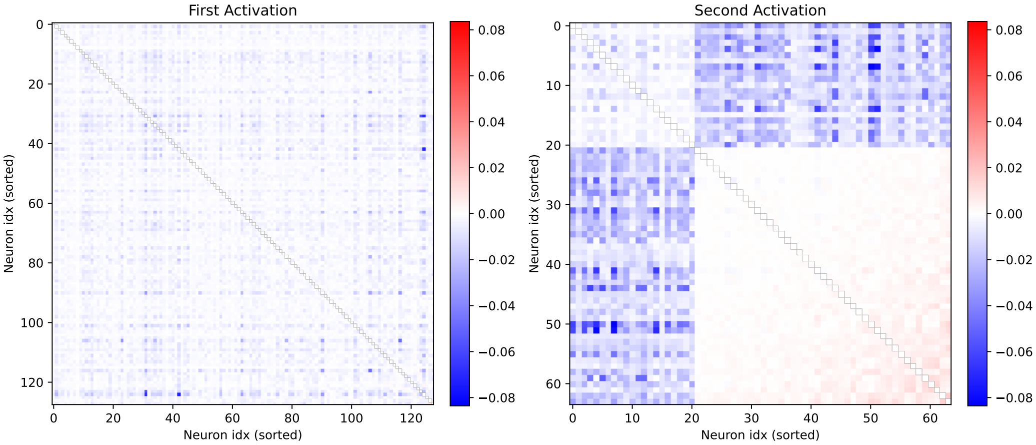


Figure 3: Synergy matrices for each activation layer

These synergy scores were subsequently used to develop a novel pruning strategy that we call **Synergy Pruning**. Unlike the baseline importance pruning, which evaluates neurons in isolation, this method assesses the total contribution for each neuron based on both its individual importance and its cooperative potential (synergy) within the layer. We define a *Synergistic Importance Score*, $\mathcal{Z}_i$, for each neuron i as

$$\mathcal{Z}_i = \Delta_i + \alpha \cdot \left(\frac{1}{N} \sum_{j \neq i} \mathrm{ReLU}(S_{ij}) \right) \tag{3}$$

where Δ_i is the individual importance (absolute difference in output probability), and the second term represents the average positive synergy that the neuron contributes to the other neurons within layer. The hyperparameter α controls the weight on the synergy contribution relative to Δ_i, where greater α emphasizes synergistic neurons over individual importance.

Synergy pruning was performed by iteratively pruning neurons, starting with the lowest $\mathcal{Z}_i$ scores for each of the two layers of the MLP. Figure 4 shows an ablation study on the hyperparameter α for synergy pruning on the second activation layer for $\alpha = 5$ and $\alpha = 20$ compared to the baseline importance pruning presented before both for AUC-ROC and accuracy (see Figure A.5 for the first activation layer).

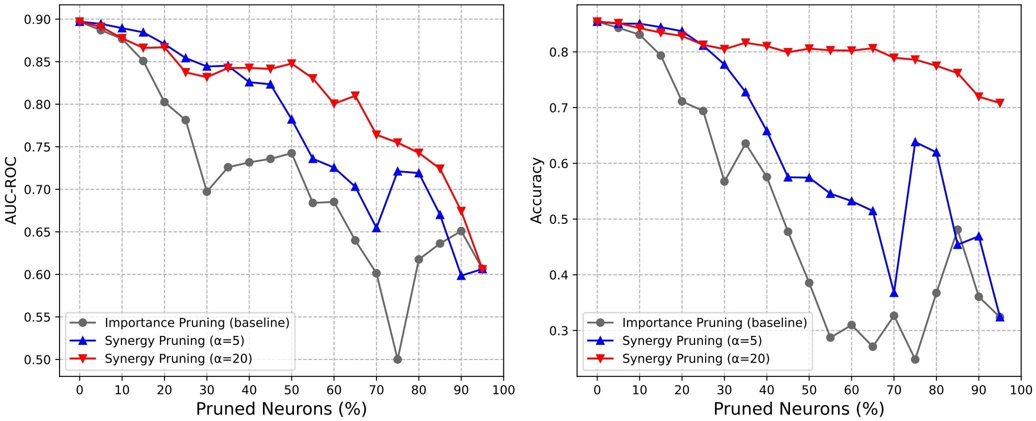


Figure 4: Synergy pruning for the second layer

The synergy pruning outperforms the baseline importance pruning for both layers. While the performance improvement is less pronounced in the first layer (Figure A.5), both AUC-ROC and accuracy is improved by some extent with greater α showing better performance than lower α, while not being consistent for all pruning levels. However, the synergy pruning significantly outperforms importance pruning in the second layer. The higher $\alpha = 20$ shows very stable accuracy with pruning up to 85% of neurons with an accuracy drop by only 10% compared to the non-pruned network. This highlights the benefit of retaining cooperative neurons rather than relying only on individually dominant neurons.

5.5 Trustworthy Aspects

5.5.1 Shapley Values

Since the Shapley calculations are computationally expensive, the reference distribution was approximated with 10,000 randomly drawn inputs, and independent 1,000 samples were randomly selected to be explained. First, local explanations were analyzed for high ($r > 0.8$) and low ($r < 0.2$) Pollock ratios as waterfall plots (see Appendix Figures A.6 and A.7). While the spatial coordinates of each location (longitude and latitude) typically drive the prediction score for each of the six examples, the oceanographic features such as NO_3, thetao (temperature) and PO_4 apparently influence the prediction as well. However these are randomly chosen local explanation for six inputs and does not quite generalize for the model. However it provides a valuable insight into the impact of the features on the model prediction.

For a more general explanation of the feature attributions, the 1,000 examples are plotted altogether with their impact on the model output in Figure 5. The color is relative to the distribution of every feature while the horizontal axis represents the impact on the output score of the model. The spatial coordinates are clearly the most impactful for the model, where low latitude values are strong drivers for higher output score (translated as higher probability for Pollock presence). The depth also shows a sharp impact on the model output, but higher depth (deeper ocean) has negative impact on the model output. Out of the oceanographic features, NO_3, PO_4 and temperature (thetao) have the strongest influence on the model's output with others be much less pronounced. Surprisingly, cyclically-encoded day of the year are not very influential for the model.

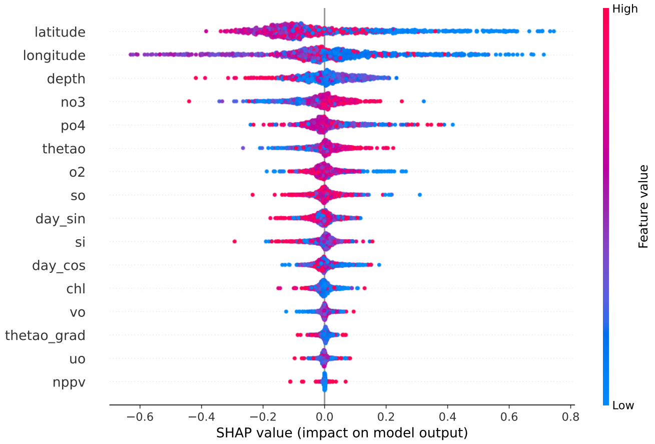


Figure 5: SHAP summary plot for 1,000 inputs

5.5.2 LIME Explanations

As an alternative to the SHAP explanations, we look at LIME explanations for the model. We select two (a FP and FN sample) incorrectly classified data points x_0 and perturb the input and approximate the decision boundary in a neighborhood of x_0, by analyzing the regression coefficients (see Appendix Figure A.8). For those two samples, the latitude appeared to be the main driver in the incorrectly predicted class.

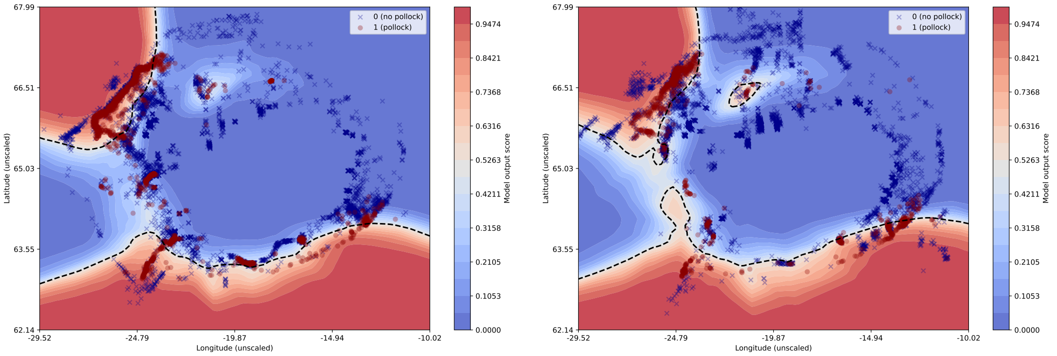
Since LIME local explanations are instance-level and do not provide an intuitive explanation for the model, we sampled 10,000 data points from the test-set and summarized the statistics in Table 2. The spatial dimensions and the depth are the primary drivers of the prediction of the model, and other features aligning with the Shapley feature importance in Figure 5. Most features have a near-zero mean, indicating that the influence-direction is highly dependent on the instances, resulting in a zero global average. It is also interesting that latitude is by far the most influential feature on the model, based on local surrogate model coefficients.

Table 2: Summary statistics from LIME coefficients (top 10 features)

Feature	Abs Mean	Mean	SD
latitude	0.115	-0.015	0.134
longitude	0.064	0.007	0.079
depth	0.052	0.008	0.060
no3	0.043	-0.015	0.050
thetao	0.041	-0.000	0.049
po4	0.024	0.004	0.030
si	0.020	-0.001	0.022
so	0.012	-0.001	0.014
thetao_grad	0.010	0.000	0.012
chl	0.010	-0.001	0.012

5.5.3 Decision Boundaries

One direct approach to intuitively understand the model's classification decision is to visualize the decision boundary in a 2D space. From the SHAP summary plot in Figure 5, the spatial coordinates are the most influential features. Figure 6 shows the decision boundary in latitude-longitude space for two different times (fixing `day_cos` and `day_sin`), with other features at their means. The decision boundaries fit reasonably well with true data points from the test set for the same `day-of-year` (with a 10-day neighborhood). While the decision boundary is comparable for the two times, it indicates flexibility and is dynamic with time.



(a) Decision boundary for `day-of-year` $= 150$

(b) Decision boundary for `day-of-year` $= 200$

Figure 6: Decision boundary for latitude-longitude space (other features at their means)

These decision boundaries demonstrate spatial and temporal dependence of the model, revealing intuitive regions capturing most of the true data points. This facilitates the interpretation of the prediction mechanism of the MLP. When combined with the global and local Shapley values and LIME explanations, these explanations make the black-box model more interpretable and more trustworthy guidance tool for Icelandic skippers targeting Pollock.

6 Conclusion

6.1 Findings

In this work, we have framed synergy analysis and *synergy pruning* for neural networks, which is a novel approach to quantify the cooperative behavior of neurons. We demonstrated that synergy pruning significantly outperforms a baseline importance pruning. Pruning 85% of neurons resulted in less than 10% accuracy drop while importance pruning resulted in an accuracy reduction by 46%. Furthermore, we found that synergistic effects are more pronounced in deeper layers of the network.

This study also highlighted trustworthy aspects of using neural networks for fish presence and absence in the North-Atlantic. Through Shapley, LIME and decision boundary visualization, we provided meaningful insight into the model's behavior. The geographical features (lat, lon, depth) were the most influential drivers, while a few oceanographic variables also played a significant, but secondary, role in predicting the presence of Pollock.

6.2 Limitations

A primary limitation of this work is the problem formulation as a binary classification task, with an arbitrary threshold of $r = 0.1$ from the observed data. In practice, predicting fish presence is better suited as a multi-class or a regression problem, which would capture the continuous nature of the data.

Additionally, a generalization of synergy pruning remains an open question. Although we observed that synergy pruning is more effective in the second layer of the two-layer MLP, further investigation is required to determine if this holds for deeper architectures and different model types. While this work framed synergistic explanations for MLPs, the same methodology can generalize for transformer models by treating attention heads as neurons. This needs to be analyzed in future work.

Finally, the dataset used herein has strong spatial dependence (latitude and longitude). While the synergy scores in Figure 3 revealed a clear pattern-like structures it is unclear whether this would emerge for a dataset without dominant features. Future work will need to validate this for broader source of datasets.

References

[1] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[2] Tianbai Chen, Li Zhong, Naweiluo Zhou, and Dennis Hoppe. Catch weight prediction for multi-species fishing using artificial neural networks. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1545–1552. IEEE, 2021.

[3] Kedar Dhamdhere, Ashish Agarwal, and Mukund Sundararajan. The shapley taylor interaction index. *arXiv preprint arXiv:1902.05622*, 2019.

[4] Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. How important is a neuron? *arXiv preprint arXiv:1805.12233*, 2018.

[5] Haibin Han, Chao Yang, Bohui Jiang, Chen Shang, Yuyan Sun, Xinye Zhao, Delong Xiang, Heng Zhang, and Yongchuang Shi. Construction of chub mackerel (scomber japonicus) fishing ground prediction model in the northwestern pacific ocean based on deep learning and marine environmental variables. *Marine Pollution Bulletin*, 193:115158, 2023.

[6] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1389–1397, 2017.

[7] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

[9] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[10] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11264–11272, 2019.

[11] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016.

[12] Michael C Mozer and Paul Smolensky. Skeletonization: A technique for trimming the fat from a network via relevance assessment. *Advances in neural information processing systems*, 1, 1988.

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[14] Mingyang Xie, Bin Liu, Xinjun Chen, Wei Yu, and Jintao Wang. Short-to medium-term forecasting of fishing ground distribution based on deep learning model. *Canadian Journal of Fisheries and Aquatic Sciences*, 82:1–17, 2024.

[15] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

A Appendix

A.1 Model Performance Metrics

Table A.1: Classification metrics for the MLP network

Class	**Precision**	**Recall**	**F1**
0 (negative)	0.91	0.91	0.91
1 (positive)	0.65	0.64	0.65
Accuracy		0.85	
Confusion Matrix			
	Pred 0	**Pred 1**	
True 0	77,753	7,831	
True 1	8,154	14,751	

Table A.2: Performance of the model with added features

Step	**Added Feature(s)**	**Accuracy**	**AUC**
1	latitude	0.801	0.751
2	longitude	0.840	0.886
3	depth	0.846	0.892
4	po4	0.845	0.895
5	no3	0.852	0.896
6	thetao	0.843	0.889
7	day_cos, day_sin	0.860	0.905
8	o2	0.854	0.898
9	so	0.857	0.898
10	si	0.853	0.897
11	chl	0.852	0.896
12	vo, uo	0.855	0.898
13	thetao_grad	0.853	0.898
14	nppv	0.853	0.897

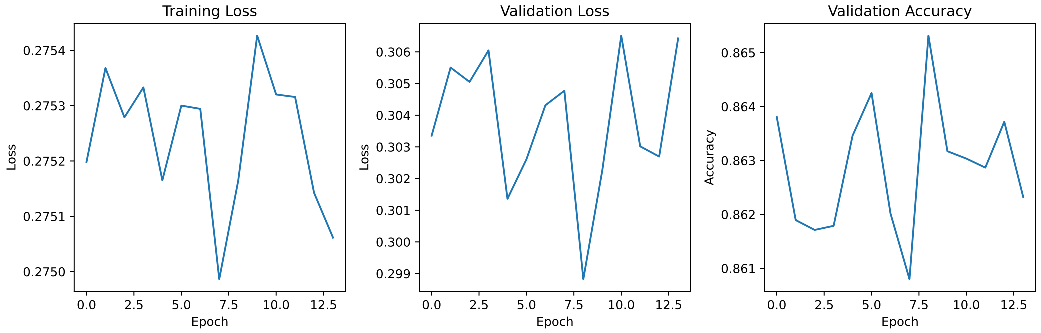


Figure A.1: Training metrics for the MLP network

A.2 Synergy and Pruning Analysis

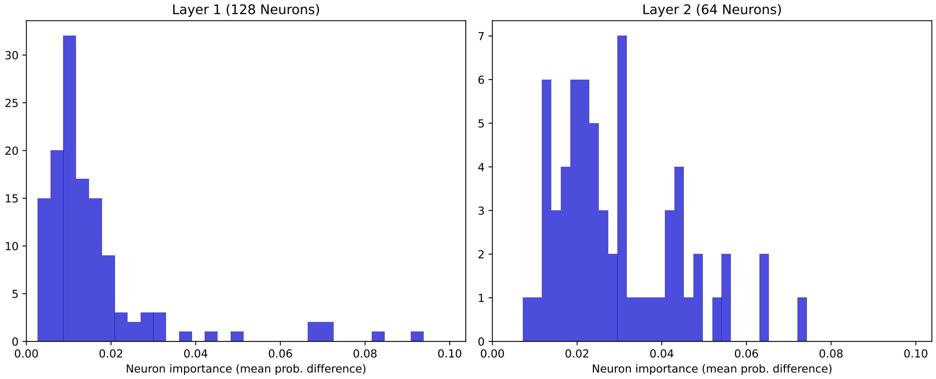


Figure A.2: Distribution of neuron importance (i.e. mean absolute output score)

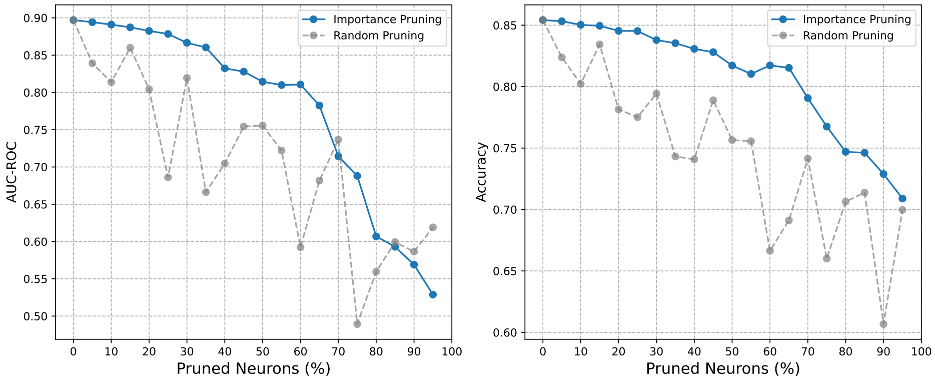


Figure A.3: Performance versus proportion of pruned neurons in layer 1

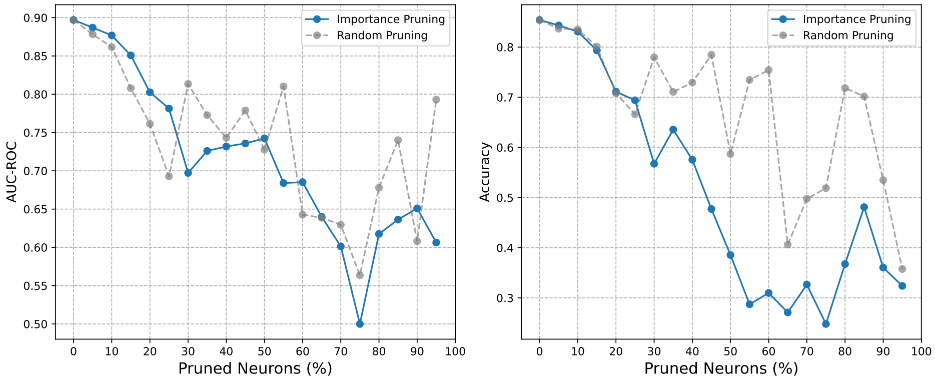


Figure A.4: Performance versus proportion of pruned neurons in layer 2

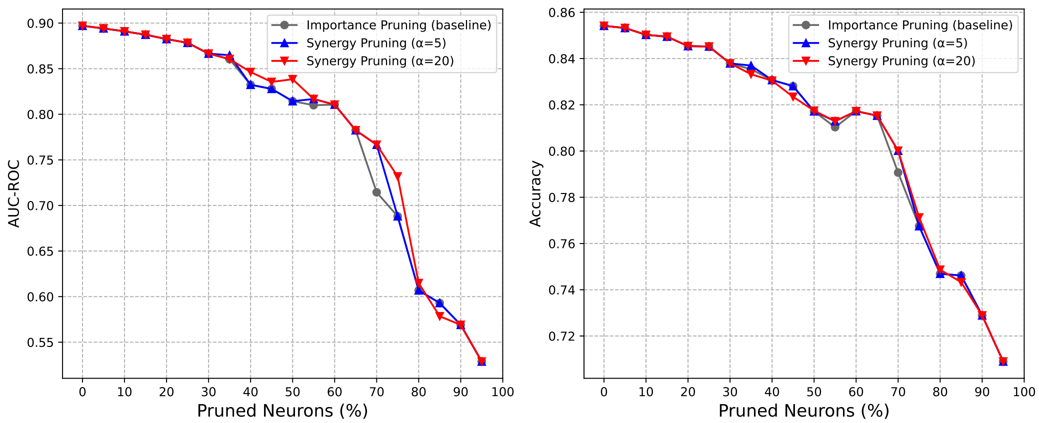
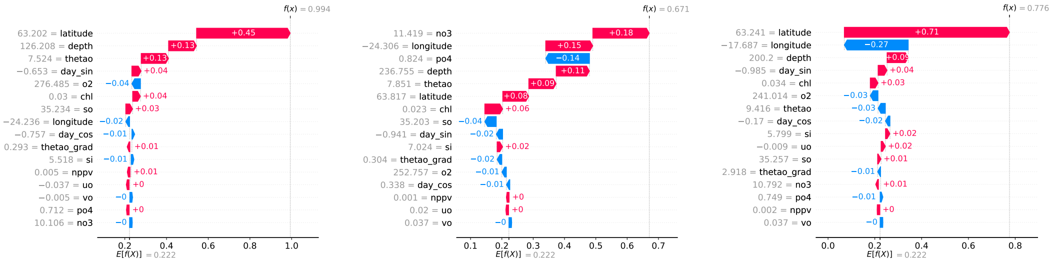


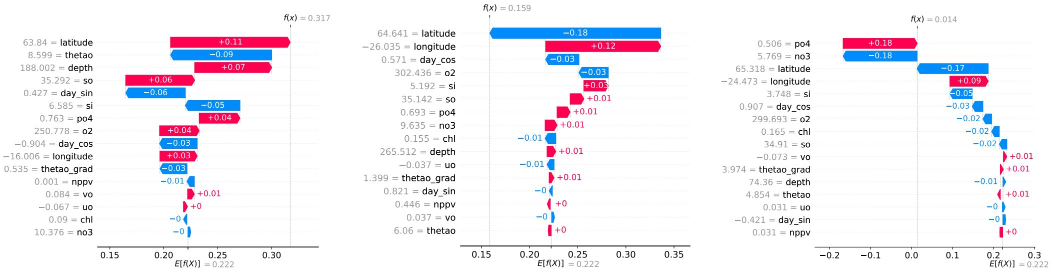
Figure A.5: Synergy pruning for the first layer

A.3 Local Shapley values



(a) True label: 1 ($r = 1.00$) **(b)** True label: 1 ($r = 1.00$) **(c)** True label: 1 ($r = 0.95$)

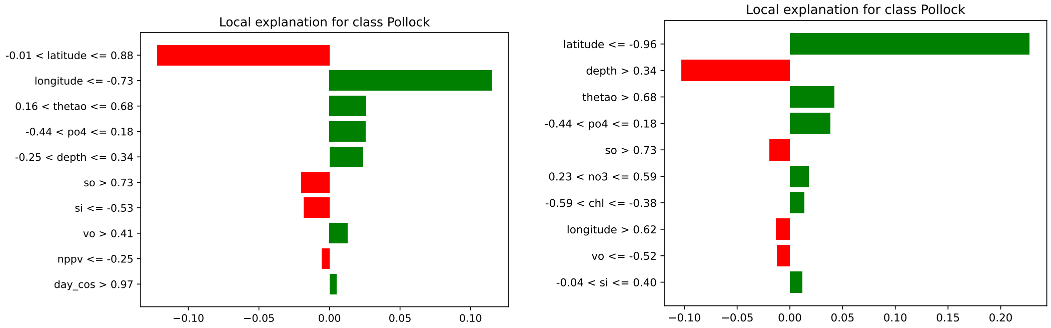
Figure A.6: SHAP waterfall plots for three high-ratio ($r > 0.8$) datapoints



(a) True label: 0 ($r = 0.00$) **(b)** True label: 0 ($r = 0.00$) **(c)** True label: 0 ($r = 0.00$)

Figure A.7: SHAP waterfall plots for three low-ratio ($r < 0.2$) datapoints

A.4 LIME Plots



(a) FN: pred: 0 (0.09), true label: 1 ($r = 0.38$)

(b) FP: pred: 1 (0.87), true label: 0 ($r = 0.05$)

Figure A.8: LIME explanations for FN and FP datapoints