Understanding Neural ODE prediction decision using SHAP

Phuong Dinh[1], Deddy Jobson[2], Takashi Sano[1], Hirotada Honda[1], and Shugo Nakamura[1]

[1]Toyo University, Tokyo, Japan
[2]Mercari, Inc., Tokyo, Japan

Abstract

Neural ordinary differential equations (NODEs) have emerged as a powerful approach for modelling complex dynamic systems using continuous-time transformations. Although NODEs offer superior modelling capabilities, little research has been conducted on understanding the factors that contribute to their predictions on image datasets. In this paper, we propose the leveraging of SHapley Additive exPlanations (SHAP), which is an influential explainable artificial intelligence method, to gain insights into the NODEs prediction process. We enable the interpretable analysis of important pixels that contribute to the prediction decisions of NODEs by adapting SHAP to the continuous-time nature thereof. Experiments on synthetic datasets demonstrate the efficacy of our proposed approach in revealing the dynamics and important features that drive NODEs predictions. Our empirical findings provide insights into how NODEs determine important features and the distributions of the Shapley values of each class. The proposed integration of SHAP with NODEs contributes to the broader goal of enhancing transparency and trustworthiness in the application of continuous-time models to complex real-world systems.

1 Introduction

Neural ordinary differential equations (NODEs) [1], which have been proposed as a variant of ResNet and RNNs, have infinite hidden layers and a continuous transition from input to output. Although NODEs have mainly been used for modelling time-series data, they have also been applied to image data. They have exhibited competitive performance with equivalent neural networks despite using fewer parameters.

However, NODEs still require a large number of parameters, which leads to opaque predictions. This limits the trustworthiness of the predictions made by NODEs and hinders their practical adoption.

The lack of interpretability of advanced machine learning models is not a new problem. Gradient-boosted decision trees and neural networks also face interpretability issues.

We propose the use of Shapley values to explain the relevance of the input features to NODEs.

Our contributions are as follows:

1. We propose the use of SHapley Additive exPlanations (SHAP) values to understand how NODEs make predictions.
2. We perform multiple experiments on a publicly available image dataset to verify our proposal.

2 Related Work

The importance of explainable artificial intelligence (XAI) [2] has been researched extensively in recent years.

Although no studies have been conducted to explain the interpretability of NODEs, several authors have used XAI methods to understand the predictions of NODEs. Laurie and Lu [3] combined NODEs and XGBoost for dynamic tumor modelling and used SHAP values to explain the inferences of their model.

Jha et al. [4] used SHAP and other XAI methods to highlight and mitigate the problems in NSDEs using their proposed neural stochastic differential equation (NSDE) architecture.

The authors who proposed NSDEs used model interpretability methods in addition to improvements to NODEs, such as SHAP values, to compare the robustness of their proposal and that of NODEs. However, they did not use SHAP values to explore the interpretability of the NODEs itself.

Jha et al. [5] used XAI to evaluate the robustness of NODEs and treated the robustness of attributions as a qualitative metric for the target. However, they did not analyze the higher-order moments of SHAP values.

3 Methodology

Our goal was to understand the importance of features selected by NODEs for prediction and to verify the correlation between the important features identified by SHAP and decisions made by NODEs. We considered several mainstream XAI methods that are used to explain deep neural network models, including Grad-CAM [6], LIME [7], and SHAP [8]. We decided to use SHAP, specifically Deep SHAP [8], for two main reasons. First, SHAP provides delicacy in explaining important features in both

Proceedings of the 5th Northern Lights Deep Learning Conference (NLDL), PMLR 233, 2024.

the selected classes and other classes in the classification task, which is the main focus of our research. Second, it is an equitable metric for data evaluation [9]. The SHAP values correspond to the importance of each pixel in the final model prediction. A better understanding of how the model determines the correct class can be achieved by identifying the regions that correspond to the most positive SHAP values (and, therefore, predictions). As the purpose was to enhance the understanding of the NODEs model, we employed SHAP with DeepLIFT [10] to strengthen the ability to explain the NODEs predictions in terms of the accuracy, missingness, and consistency [8].

SHAP was proposed by Lundberg and Lee [8] and is an XAI method that is derived from Shapley values, a game theory concept. Shapley values are used to assign credit to each participant in a cooperative game. In the context of machine learning, each feature is a participant and the goal of the "cooperative game" is to generate a prediction for each instance. As such, Shapley values can be used as a type of feature attribution XAI. The Shapley value for each feature at each data point is the expected value of the increase in the predicted value from a model that is trained with the feature compared to that when the model is trained without the feature. If N features exist in a dataset, 2^N different models will be trained. However, this is not feasible with hundreds of features. SHAP values are a modification of Shapley values that use sampling and integration for the feasible computation of approximates. SHAP values maintain the desirable properties of local accuracy, missingness, and consistency, whereas other additive feature attribution-based methods violate at least one of these properties.

DeepSHAP, the main XAI method we employed in this study, is a unified version of SHAP and DeepLIFT [8]. DeepLIFT is an additive feature attribution method that uses backpropagation to determine the effect of an input by comparing the change to its reference value selected by the user [10, 11]. DeepLIFT applies the linear composition rule using the gradient and, thus, can linearize non-linear components in a neural network [8]. Therefore, DeepSHAP can handle both non-linear components and dependent features. DeepSHAP can provide better accuracy in interpretability for neural network models such as NODEs by cooperating with DeepLIFT and SHAP.

We further investigated the distribution of the SHAP values using the mean and variance of the Shapley values of the potential classes. Given k possible classes, with each image I_i represented by a 3D array of height h_i, width w_i, and 3 colour channels, NODEs will consider the Shapley value of I_i to be predicted as $j^{th}, (j = 1, 2, ..., k)$ to be p_{ij}, which means a size of $h_i \times w_i \times 3$: $p_{ij} = \left[p^{(j)}_{abc}\right]_{a=1,...,h_i,b=1,...,w_i,c=0,1,2}$. For potential p_{ij}, the Shapley values have the same shape as image I_i, which means a size of $h_i \times w_i \times 3$.
The mean of the Shapley value if image I_i belongs to class j^{th} is defined as follows:

$$m_{ij} = \frac{\sum_{a=1}^{h_i} \sum_{b=1}^{w_i} \sum_{c=0}^{2} p^{(j)}_{abc}}{3h_i w_i}$$

The variance of the Shapley value if image I_i belongs to class j^{th} is defined as follows:

$$Var(p_{abc}) = E_{a,b}\left[\left(p^{(j)}_{abc}\right)^2\right] - \left(E_{a,b}\left[p^{(j)}_{abc}\right]\right)^2,$$

where $E_{a,b}[\cdot]$ denotes the average w.r.t. a and b and $Var(X)$ is the variance of vector X.

4 Experiments

4.1 Experimental setup

Overview. We conducted three experiments to ascertain the performance of NODEs. We aimed to demonstrate (1) the interpretability of SHAP of NODEs predictions, (2) the speed performance of DeepExplainer and GradientExplainer, and (3) the prediction difference between NODEs and its predecessor ResNet.
Dataset. The CIFAR-10 dataset [12] was used for training, validating, and diagnosing NODEs performance using the SHAP technique in all experiments. The training and test/validation datasets contained 50,000 and 10,000 images, respectively. The data loader removed 128 test images from the training set and 1000 images from the test set for evaluation in each iteration.
Model architecture and training. The NODEs architecture was that proposed in the original paper of Chen [1]. We trained both the NODEs and ResNet models once, and then employed the model state dict for all experiments. Convolution was used as the downsampling method, and an adjoint solver was employed as the ODE solver. The NODEs model was trained for 10 epochs, with 128 batches per epoch and 0.01 learning rate, which maintained an accuracy of 0.78 in the validation set. Such configuration was determined based on trial-and-error to select the setting that can compromise both learning speed and accuracy. Similarly, ResNet was trained to maintain an accuracy of 0.80. The model state and implementation will be made available on GitHub.
Explainer. DeepExplainer is the representative explainer for SHAP owing to its explainability in deep learning models. An explainer was constructed on the training dataset, including correct and incorrect predictions from the neural network models. Following calculations from the 1000-image dataset,

which differed from the training dataset, the SHAP values were separated into two groups: correct and incorrect predictions.

4.2 Explaining Local Performance of NODEs with DeepExplainer

A DeepExplainer (DeepLift combined with the Shapley value) was created using the trained NODEs and a dataset. We constructed a SHAP explainer for 10 classes with 10 images each. The detailed visualization of the images is similar to that of the example in Figure 1. We randomly selected 10 correctly predicted images and 10 incorrectly predicted images per class in each iteration and considered their SHAP values. Thus, for each type of prediction (correct or incorrect), we obtained an array of $32 \times 32 \times 3 \times 10 \times 10$ dimensions. Subsequently, we calculated the mean of the SHAP values for the selected prediction per image. Hence, for each class, we obtained 2 mean SHAP values: those for the correct and incorrect predictions. By iterating the process, we obtained a 2×10 array of average mean SHAP values (Table 1). We performed a one-sided paired t-test with dof=9 and found that the difference between the mean of the SHAP values of the correct and incorrect predictions was statistically significant ($p = 0.01$). In the heatmap of the mean SHAP values (Figure 1), the diagonal cells (selected class) are highlighted, which means that they had higher values than other cells in the same row. This indicates an intuitive analogy to the NODEs prediction decision, which tends to select the class with the highest mean SHAP value as the best prediction.

Correct	Incorrect	Difference
0.007	-0.002	0.008
0.006	0.002	0.005
0.007	0.009	-0.003
0.015	0.017	-0.002
0.013	0.003	0.011
0.017	0.016	0.002
0.028	0.019	0.008
0.039	0.007	0.032
0.016	0.003	0.012
0.021	0.004	0.018

Table 1. Difference between mean SHAP values of correct and incorrect predictions, which empirically demonstrates that the mean SHAP value of the correctly predicted class is larger than the mean SHAP value of the incorrectly predicted class.

4.3 Comparison of NODEs Explainability of DeepExplainer and GradientExplainer

Similarly to DeepExplainer, an explainer of type GradientExplainer is created on the same train dataset and the same test dataset. Then, we obtained the run time comparison (Table 2) and mean SHAP values between correct and incorrect predictions of GradientExplainer (Figure 3).

Task	Deep Explainer	Gradient Explainer
Generating explainer	178 ms	140 ms
SHAP values calculation (correct)	1m51s	1m31s
SHAP values calculation (incorrect)	1m43s	1m27s

Table 2. Runtime comparison between DeepExplainer and GradientExplainer. GradientExplainer was approximately 20 s faster in the SHAP values calculation.

4.4 Performance of NODEs and ResNet

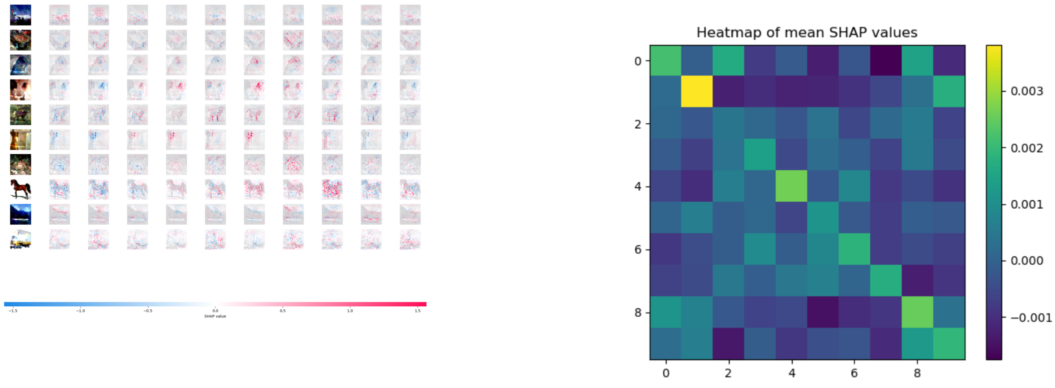
We created two explainers for NODEs and ResNet using the same DeepExplainer and dataset. The mean SHAP values of ResNet are depicted in Figure 4, and a comparison between the variance of the SHAP values of NODEs and ResNet is shown in Figure 5.

5 Discussion

The experimental results of the NODEs prediction with DeepExplainer distinguish between the NODEs decision and mean SHAP values. The diagonal in the heatmap of 1 shows the mean SHAP values of the correct predictions. The cells on the diagonal are highlighted, which indicates that the correctly predicted classes had higher mean SHAP values than the others. The experiment comparing GradientExplainer and DeepExplainer clarifies the performance-time tradeoff. Although GradientExplainer is faster than DeepExplainer, the correlation between the NODEs decision and mean SHAP values computed by GradientExplainer is less clear than that computed by DeepExplainer.

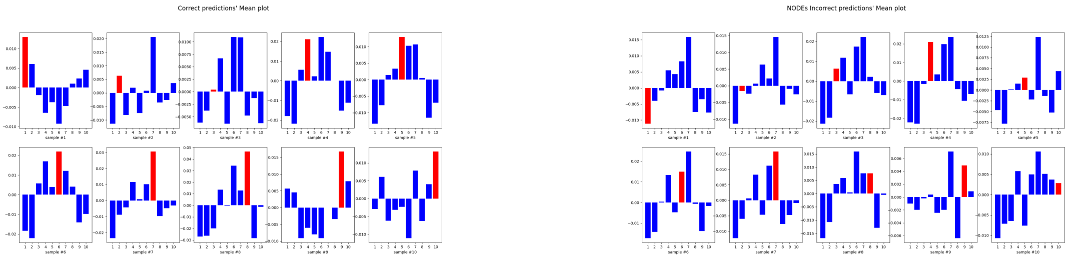
6 Conclusion

We have explored the understanding of NODEs prediction decisions using SHAP. We aimed to demonstrate the intricate relationship between the mean



(a) SHAP values of 10 images on 10 classes **(b)** Heatmap of mean SHAP values

Figure 1. SHAP values of 10 images in 10 classes. Each row represents how NODEs recognize the feature of each image that correlates to each of the 10 classes. Positive SHAP values are denoted as red in the sub-images, indicating the features that are shared by the current image and other images in that class.



(a) Correct predictions' mean plot **(b)** Incorrect predictions' mean plot

Figure 2. Comparison of mean of SHAP values for correct and incorrect predictions.Among the mean SHAP values in 10 classes of an image, all the images that gets correct prediction (a) illustrate a better correlation between mean SHAP values and the fact that one class is chosen

SHAP values and prediction decisions within the context of NODEs experimentally.

The core findings of our study reveal a significant pattern: the mean SHAP value of the correctly predicted class consistently surpasses that of the incorrectly predicted class. This empirical evidence underscores the power of SHAP values in explaining the decision-making process of NODEs by providing insights into the internal dynamics of the model that contribute to correct and incorrect predictions.

Our results demonstrate a correlation between the mean SHAP values and prediction decisions, which suggests that SHAP values can serve as a viable indicator of the prediction tendencies of NODEs.

The results also offer insight into the choice between DeepExplainer and GradientExplainer. The DeepExplainer method is better at drawing correlations between the mean SHAP values and prediction decisions. In contrast, GradientExplainer offers the advantage of runtime efficiency, which indicates a trade-off between the methodological depth and computation time.

Furthermore, we found a similarity between the correlation of the mean SHAP values and decision outcomes for NODEs and ResNet, thereby suggesting the validity of the NODEs decision. The difference in the variance between ResNet and NODEs also indicates the possibility of researching the robustness of the features of ResNet and NODEs.

In conclusion, this study presents empirical evidence of the link between the mean SHAP values and prediction decisions in NODEs, and highlights the efficacy of SHAP as a tool for deciphering the model behavior. Our results demonstrate the potential for both interpretability and performance optimization in NODEs. We encountered drawbacks in runtime during the research, and hence, we would like to provide NODEs pretrained models on the CIFAR-10 dataset alongside its implementation to support the reproduction of this work.

On top of the benefit of increased transparency and accountability of NODEs, the insights learned in this study can be used to improve the performance of NODEs on image-processing tasks. The formulation of mean SHAP values being employed in this paper suggests that not only pixels with high SHAP values

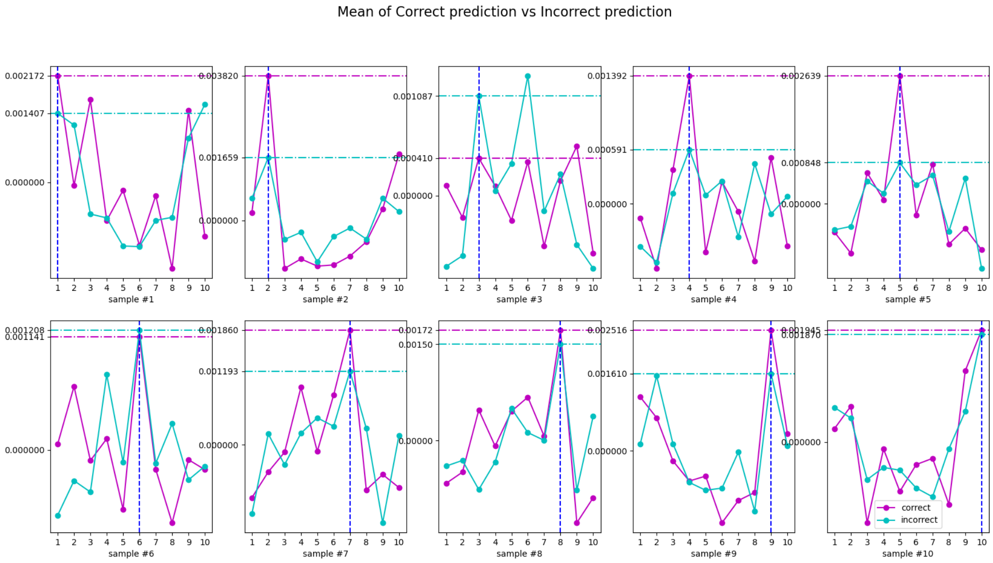


Figure 3. Mean SHAP values of correct and incorrect predictions generated by GradientExplainer. GradientExplainer had approximate correlations between the prediction correctness and mean SHAP values, with lower confidence, compared to those generated by DeepExplainer.

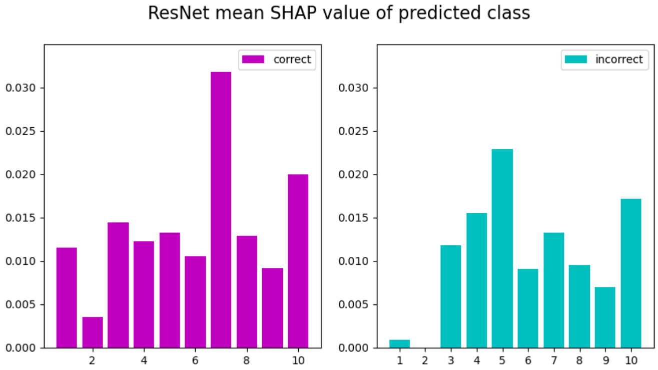
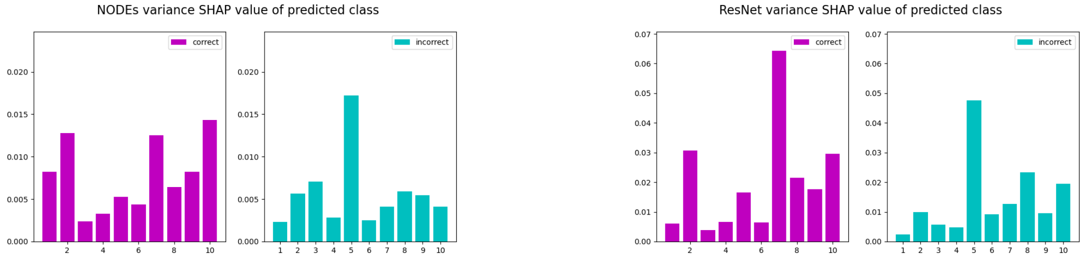


Figure 4. Mean of SHAP values of correct and incorrect predictions for ResNet. The graph indicates that, similar to the case of NODEs, the ResNet classification decision was correlated with the mean of the SHAP values, and the mean of the SHAP values of the correct predictions was higher than that of the incorrect predictions.



(a) NODEs Variance SHAP value of predicted class

(b) ResNet Variance SHAP value of predicted class

Figure 5. Comparison of variance of SHAP values between NODEs and ResNet. The variance of ResNet was higher than that of NODEs, indicating stronger fluctuation in the SHAP values within the features in an image.

but also all pixels will be taken into consideration. For example, the correspondence between the mean of the SHAP values and that of the prediction can be used to improve NODEs performance by cropping misguiding borders off of images, similar to other work [13].

7 Acknowledgement

This research is funded by Toyo University Top Priority Research Program.

References

[1] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. *Neural Ordinary Differential Equations*. 2019. arXiv: 1806.07366 [cs.LG].

[2] D. Gunning, E. Vorm, J. Y. Wang, and M. Turek. "DARPA's Explainable AI (XAI) Program: A Retrospective". In: *Applied AI Letters* 2.4 (Dec. 2021). DOI: 10.1002/ail2.61. URL: https://doi.org/10.1002/ail2.61.

[3] M. Laurie and J. Lu. *Explainable Deep Learning for Tumor Dynamic Modeling and Overall Survival Prediction using Neural-ODE*. 2023. arXiv: 2308.01362 [q-bio.QM].

[4] S. Jha, R. Ewetz, A. Velasquez, and S. Jha. "On Smoother Attributions using Neural Stochastic Differential Equations". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Z.-H. Zhou. Main Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 522–528. DOI: 10.24963/ijcai.2021/73. URL: https://doi.org/10.24963/ijcai.2021/73.

[5] S. K. Jha, R. Ewetz, A. Velasquez, A. Ramanathan, and S. Jha. "Shaping Noise for Robust Attributions in Neural Stochastic Differential Equations". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.9 (June 2022), pp. 9567–9574. DOI: 10.1609/aaai.v36i9.21190. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21190.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. URL: http://dx.doi.org/10.1007/s11263-019-01228-7.

[7] D. Garreau and U. von Luxburg. "Explaining the Explainer: A First Theoretical Analysis of LIME". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 26–28 Aug 2020, pp. 1287–1296. URL: https://proceedings.mlr.press/v108/garreau20a.html.

[8] S. Lundberg and S.-I. Lee. *A Unified Approach to Interpreting Model Predictions*. 2017. arXiv: 1705.07874 [cs.AI].

[9] S. Shobeiri and M. Aajami. "Shapley Value is an Equitable Metric for Data Valuation". In: *Iraqi Journal for Electrical And Electronic Engineering* 18 (May 2022). DOI: 10.37917/ijeee.18.2.2.

[10] A. Shrikumar, P. Greenside, and A. Kundaje. *Learning Important Features Through Propagating Activation Differences*. 2019. arXiv: 1704.02685 [cs.CV].

[11] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. *Not Just a Black Box: Learning Important Features Through Propagating Activation Differences*. 2017. arXiv: 1605.01713 [cs.LG].

[12] A. Krizhevsky. "Learning Multiple Layers of Features from Tiny Images". In: (2009), pp. 32–33. URL: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[13] V. Bento, M. Kohler, P. Diaz, L. Mendoza, and M. A. Pacheco. "Improving deep learning performance by using Explainable Artificial Intelligence (XAI) approaches". en. In: *Discover Artificial Intelligence* 1.1 (Oct. 2021), p. 9. ISSN: 2731-0809. DOI: 10.1007/s44163-021-00008-y. URL: https://doi.org/10.1007/s44163-021-00008-y (visited on 12/10/2023).