Explainable AI for Galaxy Morphology Prediction

Hazel S. Wilkins
Dept. of Computer & Info. Science
Fordham University
Bronx, New York, USA
hw30@fordham.edu

***Abstract*— The classification of galaxies is an important task that is critical for understanding the formation and evolution of cosmic structures. Traditional methods rely heavily on manual visual inspection, introducing subjectivity and scalability challenges. This study developed a Convolutional Neural Network (CNN) for multi-output regression to predict morphological probabilities using the Galaxy Zoo dataset. Explainable AI (XAI) techniques, specifically saliency maps, were applied to reveal what key features the model relied on to make its predictions. The model received a final accuracy of 86%, a validation loss of 0.018, and a validation mean absolute error of 0.092. Saliency map analysis revealed that the model was efficient in classifying spiral and merger galaxies by focusing on distinctive morphological features. However, it struggled with complex and noisy images, particularly in the classification of edge-on and elliptical galaxies. For the model's correct predictions, the saliency maps confirmed that the model's focus aligned with human classification heuristics, validating both its predictive performance and interpretability.**

Keywords— Galaxy Morphology, Convolutional Neural Networks, Explainable AI, Saliency Maps, Galaxy Zoo, Astronomical Image Classification

I. Introduction

Galaxy morphology encodes vital information about the physical processes governing galactic evolution. Classifying galaxies into elliptical, spiral, or irregular-shaped categories is essential for understanding the structure and history of the universe. Despite its fundamental nature, manual classification is time-consuming and subject to human bias. With millions of observed galaxies and new data collected daily from modern sky surveys, automated classification methods are crucial.

Projects like *Galaxy Zoo* have engaged citizen scientists to survey hundreds of thousands of galaxy images to build labeled datasets. While this crowdsourcing approach has generated valuable labeled data, it also introduces inconsistencies and labeling noise. The emergence of machine learning has enabled researchers to perform galaxy classification with speed and accuracy, dramatically reducing the manual effort required.

A challenge that many of these types of models face is interpretability. Scientific applications demand not only high model performance, but also transparency to establish why a model made a prediction. Astronomers must understand whether models are basing predictions on meaningful morphological features or irrelevant background artifacts.

This study is primarily based on the *Galaxy Zoo* Dataset, consisting of over 650,000 data examples and 8 features. Data preprocessing is vital before any exploratory techniques can be utilized. The dataset contains the *galaxy image* and the *classification probabilities* of each galaxy type based on citizen scientist survey details. Classification models used in this study depend on data that is neither redundant nor null.

Galaxies can be classified in a variety of ways, including by size, shape, feature importance, and orbit classification probabilities derived from volunteer annotations for each galaxy image. A multiclassification Convolutional Neural Network model was trained to predict these morphological probabilities, with careful preprocessing and hyperparameter tuning to optimize accuracy.

To evaluate the model's reliability, Explainable AI (XAI) techniques, including Grad-Cam and Saliency Maps, were applied to visualize image regions the model was focusing on to influence its prediction. This interpretability analysis helps assess whether the model focuses on true morphological features and establishes trust in its use for large-scale galaxy classification.

II. Background

Galaxy morphology describes the structural properties of galaxies, that is described by their classification. A galaxy's classification depends on many different factors and can be interpreted in multiple ways. A galaxy's Hubble Classification provides a more generic classification, considering the galaxy's most prominent features: disks, bulges, and bars. A more comprehensive classification method for galaxies would include features such as extended stellar halos, warps, shells, and tidal tails [1]. The features and morphological characteristics of galaxies provide insight and clarity into how a galaxy was formed and how it interacted with its environment, giving scientists a deep look into the history of our universe [2].

Galaxies are classified based on their appearance to the human eye through images. The major morphological classes of galaxies are spiral, elliptical, and irregular. Spiral galaxies resemble giant rotating pinwheels with a pancake-like disk of stars and a central bulge or tight concentration of stars. Halos surround these galaxies, joined by star clusters and dark matter. Elliptical galaxies are identified by their shapes, which range from round to oval. These galaxies show little evidence of organization or structure. Lastly, irregular galaxies have unusual shapes that don't fit in the previous two categories. These odd shapes are believed to be formed as the result of interactions with other galaxies [3].

Historically, galaxy morphology has been reliant on visual inspection by astronomers and citizen scientists. The

Galaxy Zoo project requested thousands of citizen scientists to classify galaxies observed by numerous telescopes, starting with those taken by the Sloan Digital Sky Survey. This project launched in July 2007 and is continuing today, as thousands more images are captured of never-before-seen galaxies. Each image is reviewed by multiple participants to allow assessment of how reliable the results are. This project results in large publicly labeled datasets. Overall, the project is a success and an efficient way to classify galaxies, however, some challenges of subjectivity, inconsistency in labeling, and scalability issues have arisen [4].

Convolutional neural networks (CNNs) are powerful tools for image classification tasks in almost every field. CNNs are effective due to their ability to automatically extract hierarchical features from image data. Introducing this model structure to astronomy, specifically galaxy morphology, has proven its power and strong performance in this task. For a task that previously took astronomers months and thousands of volunteers to complete, thousands of galaxies can now be classified in less than a minute [5]. While human visualization can introduce unintentional bias and inconsistency, CNNs ensure that their classification decision is based on meaningful astronomical features.

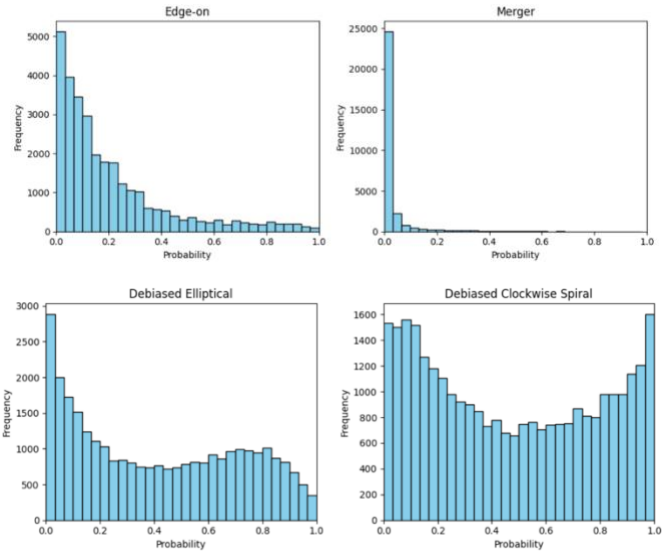
Despite the efficiency of these models, it is important to recognize the limitations. Machine learning models often act as 'black boxes', meaning that understanding why a model made its prediction is crucial [6]. It is critical to establish that their decisions are based on meaningful astronomical features rather than noise or artifacts. Introducing Explainable AI (XAI) techniques, such as Saliency Maps, reveals to scientists which regions of an image influence a model's predictions. Using these tools helps to ensure models focus on true morphological features, rather than irrelevant background information.

III. Methodology

A. Dataset Description

The dataset employed in this study is the *Galaxy Zoo Table 2* for morphological classifications. The *SDSS SkyServer API* was used to retrieve galaxy images using the 'OBJID' feature column within the dataset. The Dataset initially included 667,944 samples. After preprocessing, the final filtered dataset consisted of approximately 24,000 training images and 6,000 validation images. The four target variables used from the dataset were 'P_EDGE', 'P_MG', 'P_EL_DEBIASED', 'P_CS_DEBIASED', each representing morphological class probabilities. A *Multi-Output Regression Task* was applied to the dataset to predict four continuous class probabilities per image. The distribution of these class probabilities is illustrated in Figure 1. Because the Galaxy Zoo labels are probabilistic and represent the aggregated vote fractions of citizen scientists, a regression framework was more appropriate than hard classification. This approach allows the model to predict continuous probability outputs aligned with the dataset's labeling structure.

Figure 1: Histograms showing the distribution of morphological classification probabilities for four target Galaxy Zoo features.



B. Target Variable Selection

Four target variables were filtered from the dataset for classification, 'P_EDGE', 'P_MG', 'P_EL_DEBIASED', and 'P_CS_DEBIASED', as described in Table 1, because they were the most clearly defined and interpretable morphological features. Filtering the model to only be trained and predict these four target variables reduced label noise, where other class probabilities had higher uncertainty and were less reliable for consistent classification. Debiased variables correct for human classification biases, providing cleaner target variables. This study aimed to classify high-level morphological distinctions, rather than subtle rotational differences, prioritizing categories that are more physically significant in astrophysical research.

Table 1: Descriptions of four morphological classification features using in this study

Feature Name	Readable Name	Description
P_EDGE	Edge-on Galaxy	Probability that the galaxy appears edge-on from the observer's point of view.
P_MG	Merger	Probability that the galaxy is currently undergoing a merger with another.
P_EL_DEBIASED	Debiased Elliptical	Adjusted probability that the galaxy is elliptical, corrected for vote bias.
P_CS_DEBIASED	Debiased Clockwise Spiral	Adjusted probability that the galaxy is a spiral, bias-corrected.

C. Data Preprocessing

All images were preprocessed before applying the regression task. Images were resized to 128x128 pixels, and pixel values were normalized to a range of [0, 1]. During data cleaning, rows were removed if the corresponding image files were missing or null.

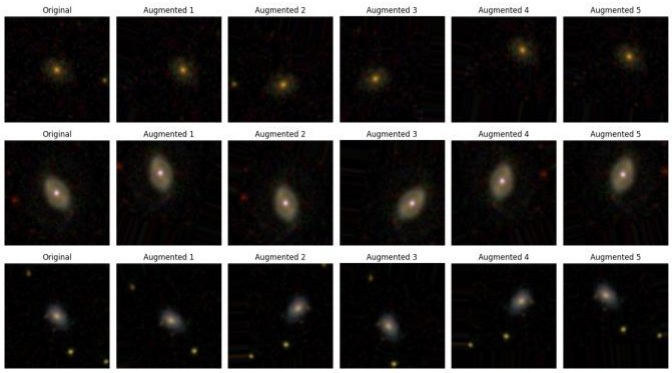
To increase model generalization and reduce overfitting, data augmentation was performed using 'ImageDataGenerator'. The augmentation techniques applied included:

- Random rotations of ±20 Degrees,
- Width and height shift of ±20%, and
- Horizontal flips to introduce orientation invariance.

These augmentation parameters were selected to simulate realistic observational variations in galaxy orientation and positioning while preserving key morphological features critical for classification. This augmentation process improved the diversity of the training data without requiring additional data collection, as shown in Figure 2.

The data was split into an 80% training set and a 20% validation set using the 'validation_split' parameter.

Figure 2: Visualization of data augmentation effects on sample galaxy images.



D. Model Development

An initial CNN model was constructed with three convolutional blocks. The final dense layer contained 4 units with sigmoid activation to predict continuous output probabilities between 0 and 1. The model was trained using the Mean Squared Error (MSE) loss function and the Adam optimizer.

Hyperparameter tuning was conducted using Keras Tuner. The final hyperparameters, summarized in Table 2, were chosen based on the ones that yielded the best model performance, as measured by the validation loss. The hyperparameters optimized included:

- Number of convolutional blocks,
- Filter sizes (16 to 128 filters),
- Dense layer sizes (64 to 256 units),
- Dropout rates (0 to 0.5), and
- Learning rate.

The model was trained for up to 15 epochs with a batch size of 32. Early stopping was applied with a patience of 3 epochs based on validation loss to prevent overfitting.

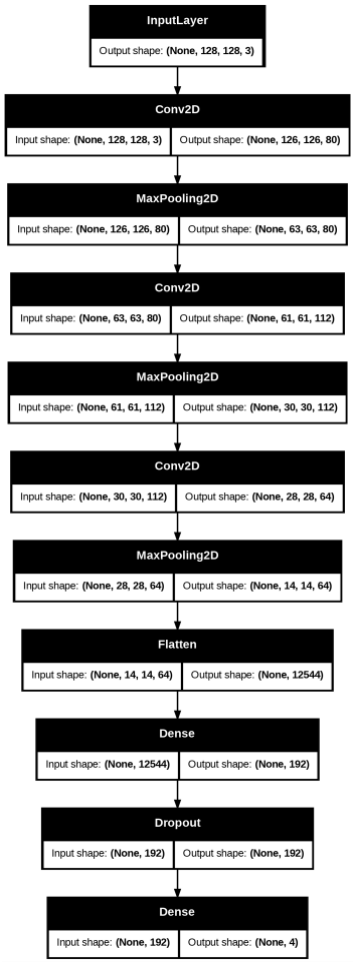
Table 2: Summary of the final hyperparameters selected through hyperparameter tuning using Keras Tuner.

Hyperparameter	Final Value
Number of Convolutional Layers	3
Filter Sizes	[80, 112, 64]
Dense Layer Units	192
Dropout Rate	0.1
Learning Rate	0.0008
Activation Function	ReLU (Hidden), Sigmoid (Output)
Optimizer	Adam
Loss Function	Mean Squared Error (MSE)

The final model architecture, as illustrated in Figure 3, consisted of:

- 3 convolutional Layers with filters [80, 112, 64],
- A dense layer with 192 units and 10% dropout,
- An output layer with 4 units using 'sigmoid' activation, and
- The 'Adam' optimizer with a learning rate of about 0.0008.

Figure 3: Architecture of the final CNN model used for multi-output regression.



E. Model Evaluation

Model performance was evaluated using Mean Squared Error (MSE) and Mean Absolute Error (MAE), consistent with task evaluation.

To assess model interpretability, Saliency Maps were generated using gradient-based methods with TensorFlow's 'GradientTape' API to validate that the CNN focused on meaningful morphological features and to understand the decision-making process. The maps were produced for a representative sample of images across all target classes. Both correctly and incorrectly classified images were analyzed to explore the model's focus during prediction. Visualizations were plotted using Matplotlib to provide insight into image regions contributing most to the model's classification decisions.

F. Explainability and Interpretability Analysis

To assess model interpretability, Saliency Maps were generated using gradient-based methods with TensorFlow's 'GradientTape' API to validate that the CNN focused on meaningful morphological features and to understand the decision-making process. The maps were produced for a representative sample of images across all target classes. Both correctly and incorrectly classified images were analyzed to explore the model's focus during prediction. Visualizations were plotted using Matplotlib to provide insight into image regions contributing most to the model's classification decisions.

G. Limitations and Challenges

This study faced computational limitations when performing the analysis. Google Colab run time and memory constraints proved to be an issue for this study. There were also challenges with label noise and probabilistic rather than categorical labels.

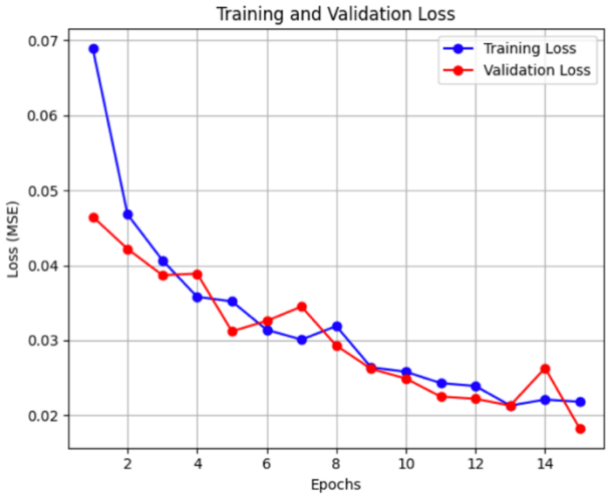
The dataset exhibited significant class imbalance, particularly with the underrepresentation of the 'P_EL_DEBIASED' and 'P_CS_DEBIASED' classes. Galaxy Zoo labels are inherently probabilistic, reflecting human classification uncertainty. This introduced ambiguity in defining true labels for evaluation, complicating accuracy and F1-score assessments.

IV. Results

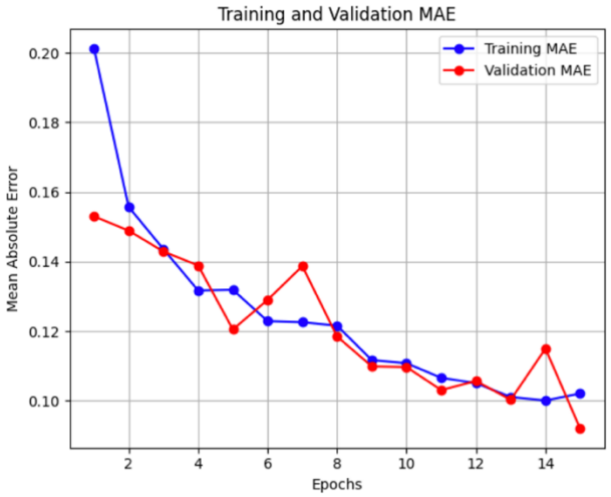
A. Model Performance

The final CNN model reached an overall classification accuracy reached 86%, as reported in the classification summary. Both training and validation loss (Mean Squared Error) steadily decrease over epochs, which is expected as the model learns. The validation loss closely follows the training loss, indicating good generalization and no significant overfitting. By the final epoch, the validation loss reaches its lowest point of approximately 0.018, suggesting strong predictive performance (Figure 4a). The Mean Absolute Error (MAE) follows a similar trend, decreasing for both training and validation sets. Minor fluctuations in the validation MAE suggest that the model briefly struggled to generalize but quickly recovered. The final validation MAE of about 0.092 a low, indicating that the model's predictions are, on average, close to true values (Figure 4b). Final loss and MAE values suggest that the model predictions are reliable, but small validation spikes imply that performance may vary slightly depending on the data batch.

Figure 4(a-b): Training and Validation Performance of the CNN model over 15 epochs



(a)



(b)

B. *Explainability Analysis*

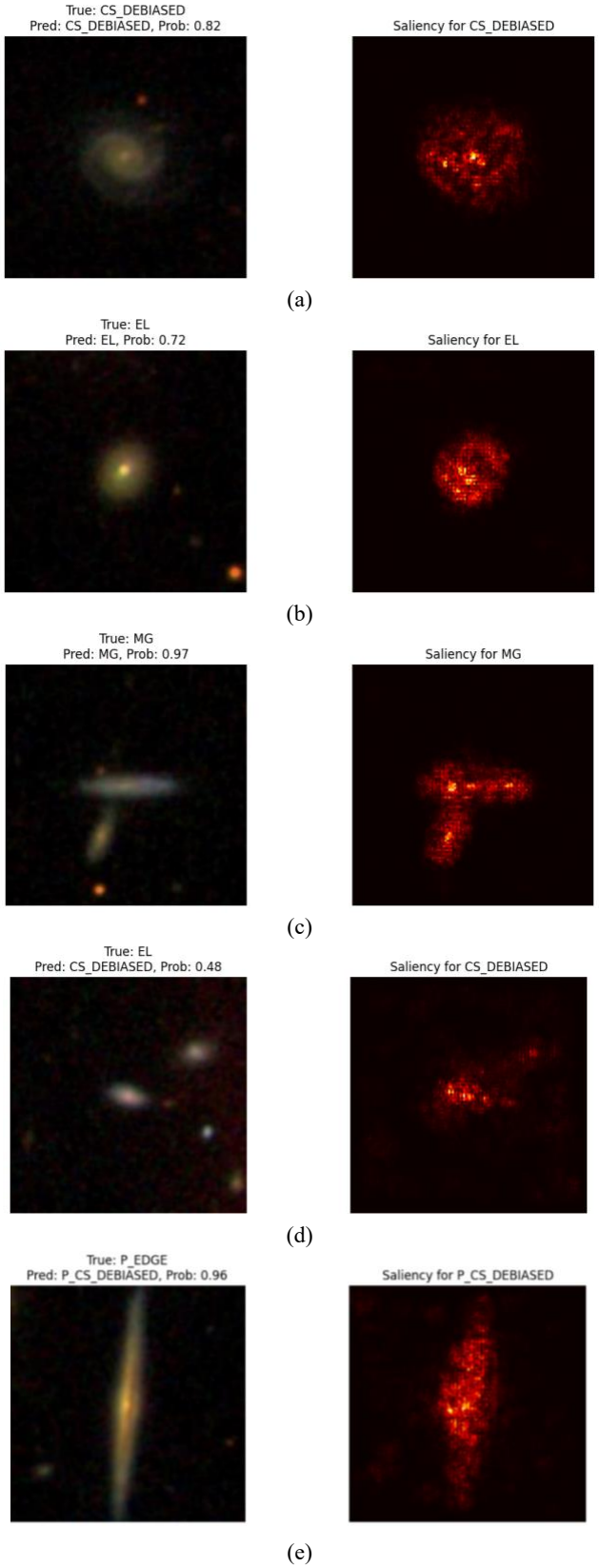
The produced Saliency maps confirmed the model's interpretability, emphasizing its ability to focus on meaningful morphological regions. In the case of spiral galaxies, the model consistently focused on distinctive spiral arms and prominent central bulges, key structural features also used by human classifiers (Figure 5a). For elliptical galaxies, attention was typically centered around their smooth, featureless cores, reflecting the lack of defining structural patterns (Figure 5b). Similarly, the model effectively identified merger galaxies by concentrating on their irregular, overlapping structures and morphological distortions, characteristic of interacting galaxies (Figure 5c).

However, the model struggled the most with correctly classifying elliptical galaxy classes. Elliptical galaxies are often found in dense galaxy clusters, resulting in many of the galaxy images containing noise, such as bright foreground stars or additional galaxies. Saliency maps revealed that in such cases, the model struggled to isolate the galaxy of interest and instead focused on the irrelevant background objects, contributing to misclassifications. An example of this behavior is shown in Figure 5d.

Additionally, the model struggled to predict edge-on galaxies, frequently defaulting to classifying the galaxy as a spiral. The saliency maps for these cases focused primarily on the central bright region, likely the bulge or core, of the galaxy. The extended thin disk, which is critical for identifying edge-on structures, received less attention, leading to poor image recognition (Figure 5e). This limitation is consistent with CNN's known struggle with learning rotational invariance unless explicitly trained with rotation-augmented data. Furthermore, edge-on galaxies may visually resemble inclined spiral galaxies, causing the model to conflate the two cases.

Overall, these visualizations validate that the model's attention mechanisms generally align with established human classification heuristics, although limitations remain in highly noisy environments. These interpretability findings complement the quantitative performance results, providing confidence in the model's ability to make scientifically meaningful predictions despite certain limitations.

Figure 5(a-e): Raw galaxy images (left) and their corresponding saliency map (right) : (a) Spiral galaxy, (b) Elliptical galaxy, (c) Merger galaxy, (d) Misclassified elliptical, and (e) Misclassified edge-on galaxy.



V. Conclusions & Future Work

The primary goal of this study was to develop a Convolutional Neural Network (CNN) to predict morphological class probabilities using the Galaxy Zoo dataset and to further interpret the model's decisions using Explainable AI.

The final CNN model achieved a strong predictive performance with an overall accuracy of 86% and a Mean Average Error of 0.1021. Saliency maps confirmed that the model correctly focuses on relevant morphological features such as spiral arms, elongated structures, and galaxy centers, aligning with established human classification methods. However, the model struggled with background noise and dense cluster environments, especially in the context of elliptical galaxy structures.

Overall, this study demonstrated the potential of deep learning for assisting in galaxy morphology classification and offered valuable visual interpretability insights. In future work, the model can be applied to newer datasets from more recent Galaxy Zoo releases to assess it generalization capability on updated and higher-quality astronomical data. Given advancements in technology today, the model should also be retrained to account for clearer and newly visible morphological features.

Additionally, implementing advanced explainability techniques, such as Grad-CAM++ and SHAP, could provide scientists with deeper insight into model decision-making. As the model matures and demonstrates higher reliability, its integration into automated astronomical survey pipelines to assist human experts in real-time classification should be explored.

This research highlights the critical role of deep machine learning models in advancing modern astrophysical studies.

References

[1] R. J. Buta, "Galaxy Morphology," Planets, Stars, and Stellar Systems, vol. 6, 2011.

[2] S. U. o. Technology, "Swinburne Cosmos," Swinburne Centre for Astrophysics and Supercomputing, [Online]. Available: https://astronomy.swin.edu.au/cosmos/g/Galaxy+Morphology. [Accessed 9 May 2025].

[3] J. Kazmierczak, "Galaxy Types," NASA, 22 October 2024. [Online]. Available: https://science.nasa.gov/universe/galaxies/types/. [Accessed 9 May 2025].

[4] G. Zoo, "Galaxy Zoo," Galaxy Zoo, [Online]. Available: https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/about/results. [Accessed 9 May 2025].

[5] M. Cavanagh, C. Rowles and J. Reid, "Morphological classification of galaxies with deep learning: comparing 3-way and 4-way CNNs'," Monthly Notices of the Royal Astronomical Society (MNRAS), 8 July 2021.

[6] C. Rudin and J. Radin, "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition," HDSR, no. 1.2, 2019.