Modelo CNN para clasificación de Galaxias

María José Soto Castro A01705840

Introducción

La clasificación de galaxias (por estructura y forma) es un factor clave para la astrofísica moderna. La morfología de las galaxias puede derivarse de factores importantes para entender cómo las galaxias evolucionan y se forman. Las galaxias se clasifican por la secuencia de Edwin Hubble, la cual identifica las galaxias por sus características de espiral, suavidad y forma elíptica.

La clasificación de las galaxias ha sido un cuello de botella debido al tiempo de preparación de datos por medio de participantes y sitios de clasificación públicos como Sloan Digital Sky Survey (SDSS) o Dark Energy Survey (DES). A pesar de resolver la alta carga de clasificación manual tradicional con estos sitios, debido a la falta de profesionales en el proceso pueden ser inviables para la investigación de expertos.

Debido a esto, las técnicas de deep learning, particularmente las Redes Neuronales Convolucionales (CNN) han surgido como una solución efectiva en tiempo.

Descripción de datos utilizados

Para el entrenamiento de los modelos utilizados se utilizaron imágenes de Galaxy Zoo 2 disponible en Kaggle (Mifsud, 2016). La clasificación de dichas imágenes se hicieron por medio del Sloan Digital Sky Survey (SDSS) el cual se hace de manera pública, por lo cual estos datos pueden presentar un sesgo alto al no ser clasificadas por profesionales. Debido a esto, los datos utilizados en estos modelos fueron recolectados en base al criterio de aceptación de Dominguez Sanches, et. al., permitiendo trabajar solo con las galaxias más robustas en su clasificación.

La clasificación de las galaxias está basada en la clasificación de Hubble. En la cual, existen 3 clasificaciones predominantes, elípticas, espirales y espirales en barra. En el proyecto de Galaxy Zoo 2, estas clasificaciones se determinan en base a series de preguntas que determinan la clasificación por árboles de decisión (Willett et al., 2013).

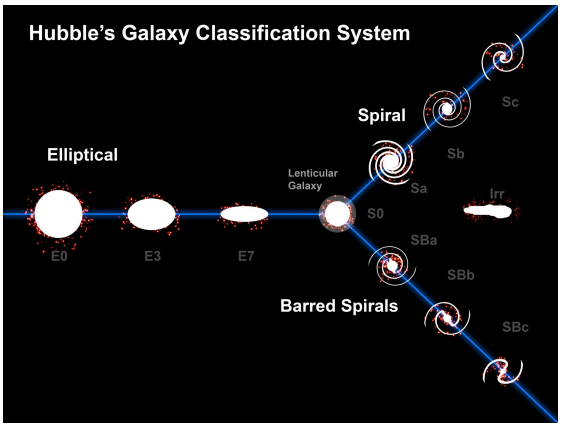


Figura 1. Clasificación de Hubble para Galaxias.
(Un ejemplo popular de una galaxia espiral es la vía láctea :))

Conjunto de datos y preprocesamiento

El preprocesamiento de datos es una etapa crucial para asegurar que el modelo de red neuronal convolucional (CNN) aprenda de manera efectiva, sea robusto a variaciones y generalice bien a nuevas imágenes de galaxias. Este proceso se gestiona principalmente a través de la aplicación de dos conjuntos de transformaciones de torchvision.transforms:

1. Transformaciones a datos

Para poder entrenar el modelo, las imágenes se transformaron para optimizar la carga técnica y mantener los tiempos de entrenamiento óptimos.

Dimensiones: Las imágenes fueron redimensionadas a el tamaño (128 x 128 píxeles)

Normalización: Las imágenes pasaron por un proceso de estandarización de píxeles entre [1, -1].

2. Aumento de datos

Para el conjunto de datos se implementó al grupo de imágenes de entrenamiento transformaciones que reducen el sobreajuste y simulan una variedad más grande del dataset.

RandomHorizontalFlip(): Aplica un espejo horizontal a la imagen en un 50% de veces. Esto permite al modelo entrenar con diferentes perspectivas y el modelo no depende de la simetría.

RandomRotation(): Se rota la imagen a 15°. Debido a que la orientación de la galaxia no está sujeta a espacio rígido, la rotación obliga al modelo a reconocer los patrones de galaxia sin depender del ángulo en el que se tomó.

3. División y Balanceo de Clases

División del dataset

Debido a la capacidad computacional, el total de imágenes fue de 15,000. Los datos se dividieron en 80% entrenamiento y 20% para pruebas. Para la validación se extrajeron 10% de las imágenes de entrenamiento.

Balanceo de Clases

Debido a que el dataset presenta más galaxias de clasificación Elíptica que de Espiral o Espiral Barrada, se utilizó un peso inverso a la frecuencia en el dataset. Esto significa que si el modelo da un falso positivo en una clase con menor frecuencia, será penalizado más que si da un falso positivo de la clase con mayor frecuencia.

Arquitectura del Modelo

GalaxyCNN version. 1.0

La arquitectura del modelo GalaxyCNN, se inspiró en el modelo de clasificación de morfología en galaxias y no en tipo de galaxias (Dieleman et al., 2015). El cual propone el uso de 4 capas convolucionales y 3 capas densas. El diseño permite la reducción de la imagen sin perder las características dado a la utilización de Max Pooling de 2.

En este modelo se utilizó la siguiente arquitectura:

- Bloque 1 (conv1):
  - Capa Convolucional de 32 canales, kernel = 6, padding = 2
  - Activación ReLU
  - Regularización: Dropout (0.2)
- Bloque 2 (conv2 y pool2):
  - Capa Convolucional de 64 canales, kernel = 5, padding 2
  - Activación ReLU
  - Pooling: MaxPooling de 2
  - Regularización: Dropout (0.2)
- Bloque 3 (conv3 y pool3):
  - Capa Convolucional de 128 canales, kernel = 3, padding 1
  - Activación ReLU
  - Pooling: MaxPooling de 2
  - Regularización: Dropout (0.2)
- Bloque 4 (conv4):
  - Capa Convolucional de 128 canales, kernel = 3, padding = 1
  - Activación ReLU
  - Regularización: Dropout (0.5)
- Capa Densa 1 (fc1):
  - Capa Lineal
  - Salida de 64 unidades
  - Activación ReLU
- Capa Densa 2 (fc2):
  - Capa Lineal
  - Salida de 3 unidades

El Maxpooling en los bloques 2 y 3 reduce las dimensiones espaciales a la mitad, crucial para disminuir la complejidad computacional. Se utiliza un Dropout (Hinton et al., 2012) alto de 0.5 en el último bloque para prevenir el

sobreajuste, mientras que el resto de capas se regulan con un valor más moderado de 0.25. La salida final (fc2) produce 3 clasificaciones, que se convierten en probabilidades.

GalaxyCNN version. 2.0 con Batch Normalization

La segunda versión de esta arquitectura, mantiene la estructura básica del diseño inspirado en Dieleman et. al. (2015), pero incorpora Batch Normalization (BN) en cada bloque convolucional para mejorar el sobreajuste en el entrenamiento.

- Bloque 1 (conv1, bn1):
  - Capa Convolucional de 32 canales, kernel = 6, padding = 2
  - Batch Normalización
  - Activación ReLU
  - Regularización: Dropout (0.5)
- Bloque 2 (conv2, bn2, pool2):
  - Capa Convolucional de 64 canales, kernel = 5, padding 2
  - Batch Normalización
  - Activación ReLU
  - Pooling: MaxPooling de 2
  - Regularización: Dropout (0.25)
- Bloque 3 (conv3, bn3, pool3):
  - Capa Convolucional de 128 canales, kernel = 3, padding 1
  - Batch Normalización
  - Activación ReLU
  - Pooling: MaxPooling de 2
  - Regularización: Dropout (0.25)
- Bloque 4 (conv4, bn4):
  - Capa Convolucional de 128 canales, kernel = 3, padding = 1
  - Batch Normalización
  - Activación ReLU
  - Regularización:Dropout (0.25)
- Capa Densa 1 (fc1):
  - Capa Lineal
  - Salida de 64 unidades
  - Activación ReLU
- Capa Densa 2 (fc2):
  - Capa Lineal
  - Salida de 3 unidades

La principal mejora de la versión 2.0 es la inclusión de batch normalization (Ioffe, 2015), permite que antes de pasar a una capa no lineal, como ReLU, se estandaricen los valores. Además, ayuda a reducir el uso de dropout, en esta versión invertimos los valores de las capas iniciales de dropout a las finales, debido a que las características generales (color, forma) son triviales en comparación con las características específicas como espiral, barra y centro.

Gráficas GalaxyCNN v. 1.0

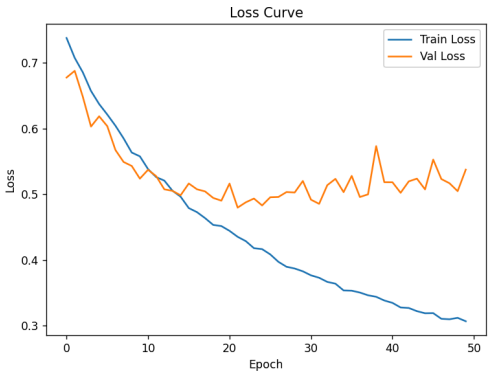


Figura 2. Curva de Pérdida GalaxyCNN v. 1.0.

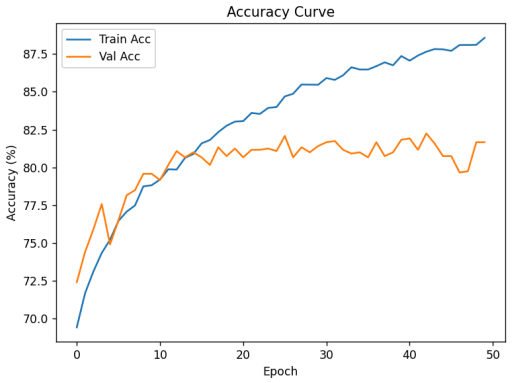


Figura 3. Curva de Accuracy GalaxyCNN v. 2.0.

Análisis de Resultados

La diferencia entre el primero modelo y el segundo es inicialmente que el primer modelo presentaba un alto grado de overfitting, resultando en un accuracy en entrenamiento mayor a 87.5% pero en validación de 81%. Esto significa que el modelo, no está aprendiendo correctamente las características importantes en las diferentes clasificaciones de las galaxias. También se puede apreciar el overfitting en la gráfica de curva de pérdida a partir de la época 15. Esto significa que el modelo empieza a memorizar los datos de entrenamiento en lugar de aprender características generalizables.

Gráficas GalaxyCNN v. 2.0.

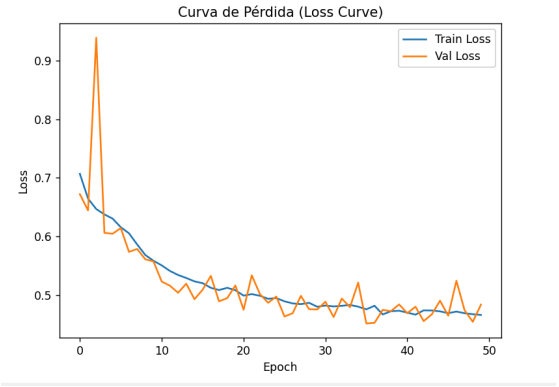


Figura 4. Curva de Perdida GalaxyCNN v. 2.0.

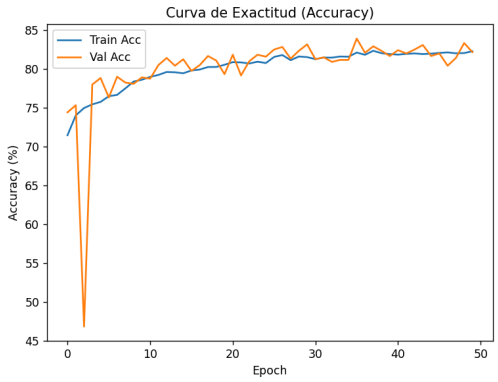


Figura 5. Curva de Accuracy GalaxyCNN v. 2.0.

Análisis de Resultados

En la segunda versión de GalaxyCNN, el aprendizaje se ha vuelto más lento, pero no muestra señales de overfitting, ni underfitting. Debido a que las curvas de accuracy y pérdida de entrenamiento y validación se encuentran cercanas una de la otra. Al finalizar el entrenamiento el accuracy quedó en alrededor de 81.5% para entrenamiento y validación.

El entrenamiento del modelo puede ser explicado por la inclusión de pesos altos en dropout, debido a que el modelo pierde peso con el dropout.

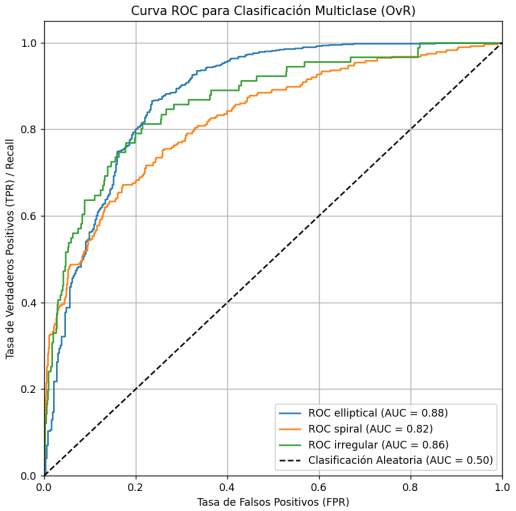


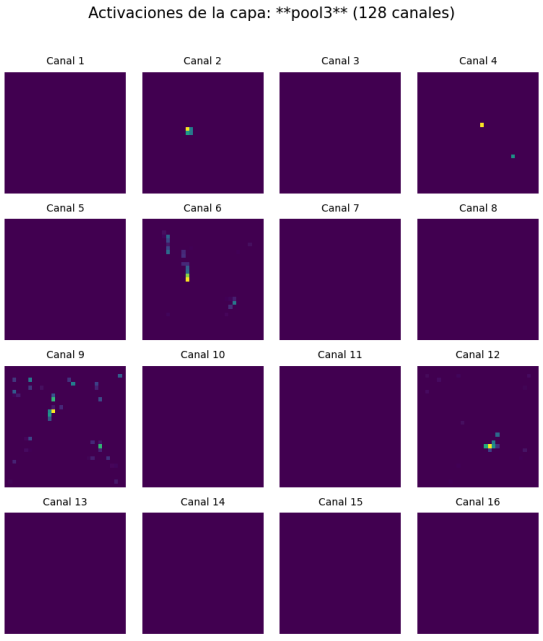
Figura 6. Curva de ROC para Clasificación Multiclase

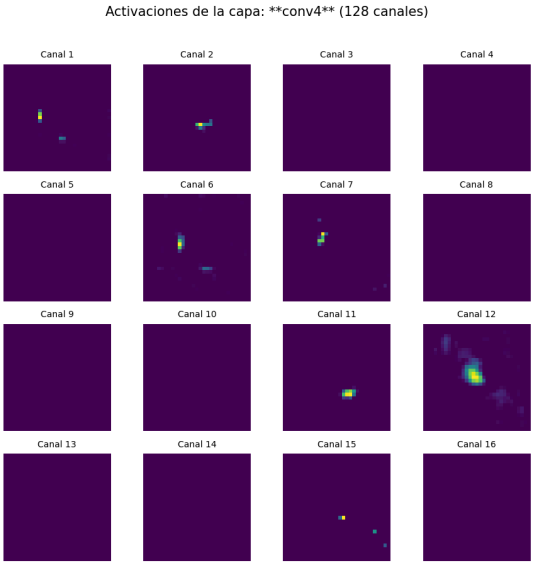
Una vez se soluciona el problema de overfitting, se analiza la capacidad del modelo de identificar correctamente los casos positivos de cada clase. En el GalaxyCNN v. 2.0. Todas las clases tienen un valor de Área Bajo la Curva (AUC) mayor a 0.8. Por lo que el modelo, puede predecir todas las clases con un 80% de probabilidad de no ser un falso positivo. La clase con mayor probabilidad es la elíptica, debido al desequilibrio en la cantidad de imágenes por clase, y la clase elíptica

siendo la clase con más imágenes. Debido al balance de clases, el modelo puede clasificar todas las clases arriba de un 80%.

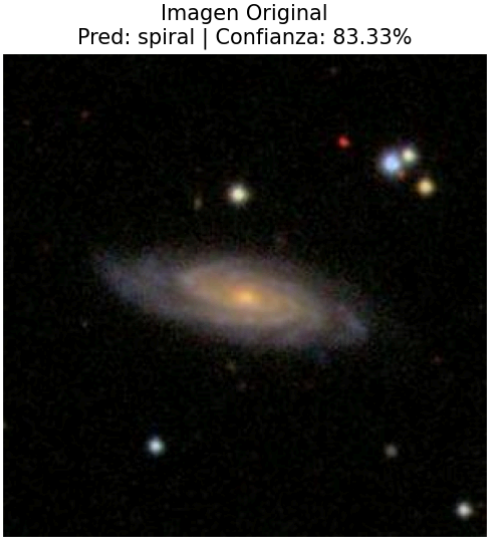
Análisis de Capas por Clase GalaxyCNN v. 2.0.

Clase 1. Elíptica

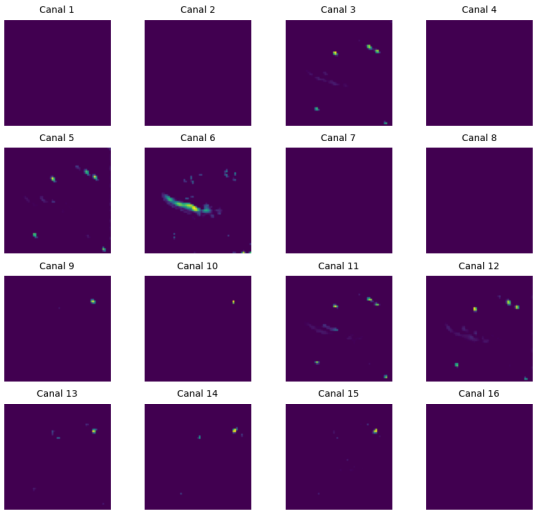




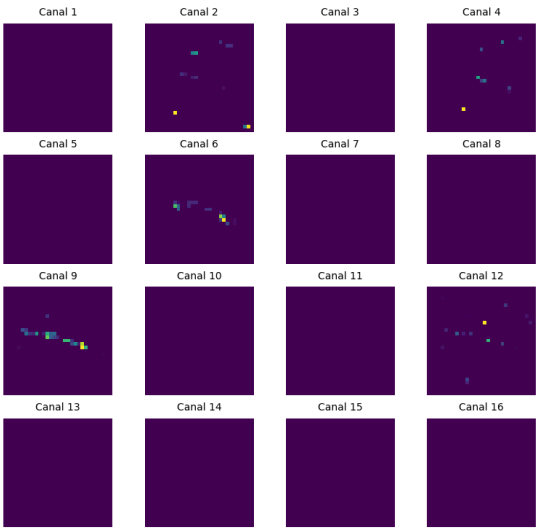
Clase 2. Espiral



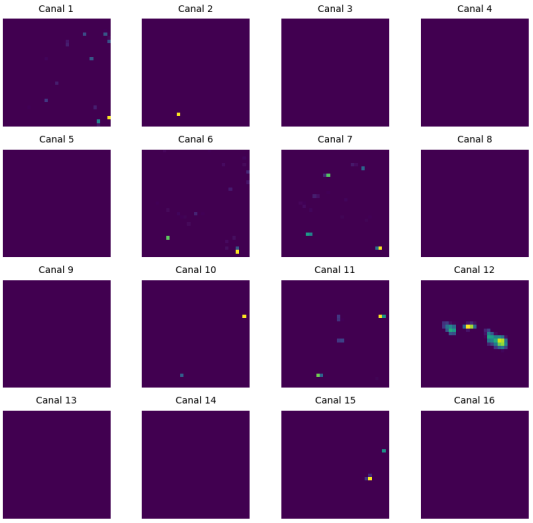
Activaciones de la capa: **pool2** (64 canales)



Activaciones de la capa: **pool3** (128 canales)

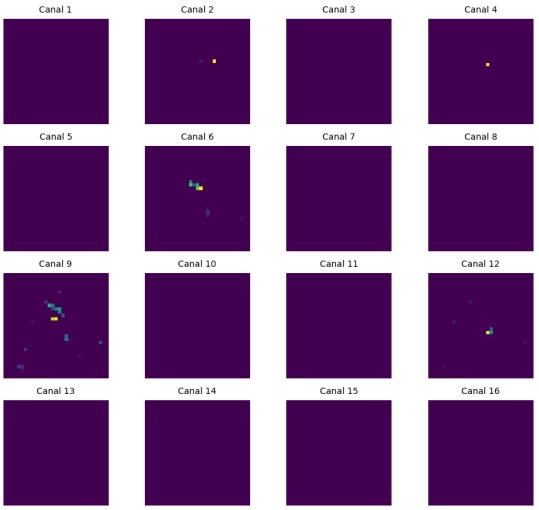


Activaciones de la capa: **conv4** (128 canales)

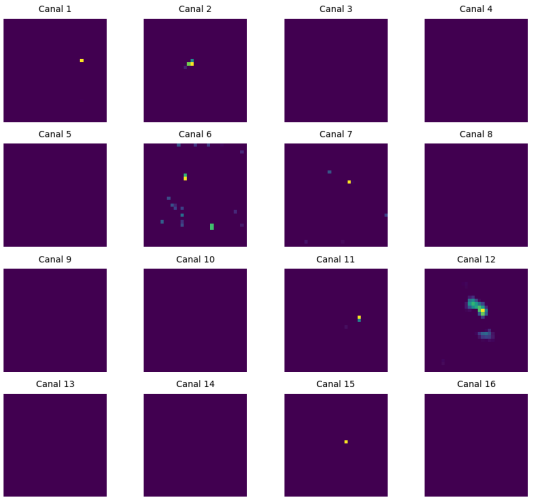


Clase 3. Espiral Barrado

Activaciones de la capa: **pool3** (128 canales)



Activaciones de la capa: **conv4** (128 canales)



Conclusión

El estudio se centró en la clasificación automatizada de galaxias utilizando un modelo CNN. La GalaxyCNN v. 1.0 inicial evidenció un sobreajuste severo a partir de la época 15, logrando una alta precisión de entrenamiento (~87.5%) que no se reflejó en la validación (~81%). Este problema fue crucialmente abordado en la GalaxyCNN v. 2.0 mediante la incorporación de Batch Normalization (BN) y un ajuste en la estrategia de Dropout. Esta optimización logró una convergencia sana y estable, eliminando el sobreajuste y manteniendo la precisión de validación y entrenamiento en un nivel competitivo de aproximadamente 81.5%.El análisis final de la Curva ROC confirmó la robustez de la v. 2.0, mostrando un Área Bajo la Curva (AUC) superior a 0.8 en todas las clases. Este resultado, apoyado por una estrategia efectiva de balanceo de clases, demuestra que el modelo es altamente fiable y tiene una excelente capacidad para distinguir entre los diferentes tipos de galaxias (Elípticas, Espirales y Espirales en Barra). En conclusión, la GalaxyCNN v. 2.0 es un clasificador sólido y generalizable, validando el uso de CNNs como una herramienta eficiente para automatizar la clasificación morfológica en grandes bases de datos astronómicas.

Referencias

D. Dobrycheva, V. Khramtsov, M. Vasylenko, & I. Vavilova. (2020). The CNN classification of galaxies by their image morphological peculiarities. Proceedings of the International Astronomical Union, 16(S362), 111–115. https://doi.org/10.1017/s1743921322001259

Dieleman, S., Willett, K. W., & Dambre, J. (2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. Monthly Notices of the Royal Astronomical Society, 450(2), 1441–1459. https://doi.org/10.1093/mnras/stv632

H Domínguez Sánchez, M Huertas-Company, Bernardi, M., Tuccillo, D., & Fischer, J. L. (2018). Improving galaxy morphologies for SDSS with Deep Learning. Monthly Notices of the Royal Astronomical Society, 476(3), 3661–3676. https://doi.org/10.1093/mnras/sty338

Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. https://arxiv.org/pdf/1207.0580

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. ArXiv.org. https://arxiv.org/abs/1502.03167

Mifsud, R. (2016). Resized and Reduced Galaxy Zoo 2 Images. Kaggle.com. https://doi.org/10.5281/zenodo.3565489

Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Kevin, Edmondson, E. M., Fortson, L. F., Sugata Kaviraj, Keel, W. C., Melvin, T., Nichol, R. C., M. Jordan Raddick, Schawinski, K., Simpson, R. J., Skibba, R. A., Smith, A. M., & Thomas, D. (2013). Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society, 435(4), 2835–2860. https://doi.org/10.1093/mnras/stt1458