CNN-Based Measurement of Galaxy Ellipticities from Fourier-Plane Radio Interferometric Data: A New Approach to Weak Lensing Analysis

Sofie Seilnacht[1]

[1] School of Physics & Astronomy, Cardiff University, Cardiff, UK

Abstract

Dark matter, though not visible, greatly influences the Universe's structure and evolution, and can be studied through gravitational lensing, particularly by observing weak lensing distortions in light from background objects. However, detecting these subtle distortions is challenging, particularly when dealing with the complexities introduced by the Point Spread Function (PSF) in radio interferometry. This study explores the use of convolutional neural networks (CNNs) to measure galaxy ellipticity values from Fourier-transformed images, bypassing traditional deconvolution techniques required to handle PSF distortions. The aim is to leverage CNNs' ability to process data in Fourier space for improved accuracy in measuring weak lensing effects.

We simulated 10,100 galaxy images with varying ellipticity values and generated corresponding truth label catalogs. The images were preprocessed by transforming pixel values into Fourier space, separating them into real and imaginary components. We trained a linear regression CNN model on this data, experimenting with various architectures and training strategies. Model performance was evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and an astrophysical difference metric. The best-performing CNN model demonstrated an MAE of 1.48×10^2, an MSE of 4.6×10^4, and an RMSE of 2.145×10^2, indicating precise prediction of galaxy ellipticities. The model effectively generalizes to unseen data, though it is acknowledged that the simulated data's simplicity may limit its applicability to real-world scenarios. This research presents a promising approach to weak lensing analysis using CNNs in Fourier space, offering an alternative to traditional PSF deconvolution techniques. Future work will involve enhancing the model with more realistic data and exploring more complex architectures to improve performance and applicability.

1. Introduction

The study of dark matter has revealed the existence of an unseen yet dominant component of the Universe, one that significantly shapes its structure, influences its evolution, and challenges our current understanding of fundamental physics [1]. Understanding dark matter provides insights into various cosmological parameters, such as the parameter that describes the dark energy equation of state (w_p), the Hubble constant, and the substructure of dark matter halos, furthering our knowledge of the Universe's framework [2]. Only a small fraction of the Universe is composed of baryonic matter, the rest is dominated by dark matter and dark energy, components whose precise structural nature remains inconclusive, though their effects are observable, primarily through gravitational interactions [3].

1.1 Gravitational Lensing

One of the primary methods for detecting and studying dark matter is through gravitational lensing. In gravitational lensing, light from a background object is distorted by the gravitational field of a lens, with weak lensing causing only minor distortions that are difficult to observe without sophisticated statistical techniques [4]. The changes to the shape and size of background objects are so subtle that they cannot be detected by looking at individual objects.

Gravitational lensing was first predicted in 1916 by Albert Einstein, when he predicted that massive objects, such as stars, galaxies, and other Large Scale Structures, would curve the fabric of spacetime, causing light from distant objects passing near them to follow a curved path [5]. His theory was confirmed in 1920 when Sir Arthur Eddington and his team photographed a star field during the solar eclipse of May 1919, providing the first observational evidence of Einstein's theory.

This groundbreaking discovery sparked a pursuit to understand the implications of gravitational lensing. In 1979, the phenomenon of strong lensing was observed for the first time with the discovery of an Einstein ring around the double quasar Q0957+561, identified by Walsh, Carswell, and Weymann [6].

However, it wasn't until 2000 that the first evidence of weak lensing was detected, thanks to the work of Bacon et al. [7], Kaiser et al. [8], Van Waerbeke et al. [9], and Wittman et al. [10]. In the 24 years since, weak lensing has become a

cornerstone of cosmology, offering profound insights into the Universe's history and structure [11].

Building on this foundation, three major large-scale weak lensing surveys have emerged in the optical domain: the Dark Energy Survey (DES) and the Kilo-Degree Survey (KiDS), which have collaborated to analyze their results jointly [12], and Subaru Hyper Suprime-Cam (HSC) [13]. Together, these surveys have made substantial progress in the optical regime, enabling the inference of various cosmological parameters, such as w_p and the matter density parameter. They have also advanced our understanding of dark matter halo clustering, provided observational constraints on key cosmological constants like the Hubble constant, and contributed to a broader understanding of the Universe's structure and evolution.

To study weak lensing, researchers typically use statistical analysis methods such as the two-point correlation function and the power spectrum [4]. The two-point correlation function measures how the shapes of galaxy pairs are correlated based on their separation distance, providing insights into the distribution of cosmic structures in real space. The power spectrum, derived from the Fourier transform (or spherical harmonic expansion) of the correlation function, describes the distribution of "power" across different spatial frequencies, revealing how structures of varying scales contribute to the observed lensing effects [14].

Most studies in this area are conducted using observations in the optical and near-infrared bands of the electromagnetic spectrum. However, observing galaxies in these bands can be challenging due to the blending effects of light from larger galaxies and atmospheric distortions, which are reflected in the Point Spread Function (PSF) of a telescope [14].

1.2 The Point Spread Function

The PSF, which describes how a telescope's optics spread light from a point source, is a critical factor in these observations of galaxies. For optical telescopes, the PSF typically does not require significant scrutiny, as it has minimal impact on raw images [14]. However, in radio interferometry, where images are reconstructed from signals gathered by multiple antennae, the PSF becomes far more complex and difficult to manage. The distortions in galaxy shapes caused by weak gravitational lensing are typically only around 1%, much smaller than the intrinsic ellipticities of galaxies, and often smaller than the distortions introduced by the PSF itself [11]. Accurately removing PSF effects is both crucial and challenging, often necessitating computationally intensive deconvolution techniques such as CLEAN and the maximum entropy method (MEM) [15]. These methods, implemented using general radio data packages like DIFMAP and CASA [16], work by decomposing the intensity distribution into point-source components that reflect the original imaging process. Each component is then replaced with a sidelobe-free beam response, addressing the issues of incomplete imaging inherent in interferometry, where multiple radio antennas observe a single point in the sky.

Due to the complexities introduced by the PSF, traditional cleaning techniques often utilize "peeling." This process involves iteratively removing or deconvolving overlapping sources to isolate the target objects, thereby mitigating the influence of background sources and preventing them from overwhelming the target galaxies. However, these techniques are heavily dependent on the user's experience, leading to results that are often irreproducible and subjective [16].

Given the intricacies of these challenges, there has been a growing interest in utilizing machine learning (ML) techniques, particularly convolutional neural networks (CNNs), to measure the shear of galactic data in Fourier space. Since radio data is naturally collected in this domain, this approach allows us to bypass the need for deconvolution in favor of a model that can effectively learn patterns directly from the Fourier data. This method presents a promising new direction for research by circumventing the intricate process of deconvolving radio interferometry PSFs.

1.3 Machine Learning in Cosmology

Since interferometric data is inherently represented in Fourier space, we can leverage this data to train a linear regression CNN model to measure galaxy ellipticity values. CNNs are particularly well-suited for this task due to their capacity to process and interpret information across both spatial and frequency domains [17]. To evaluate the accuracy of our predictions and the performance of the model, we employ ML metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), normalized MAE (nMAE), root mean squared error (RMSE), and an astrophysical difference metric that compares true labels with predicted ones, demonstrated in Equation 1, in which y_i denotes the true ellipticity label and $\hat{y}_i$ refers to the predicted label.

$$diff = y_i - \hat{y}_i \qquad \text{(Eqn. 1)}$$

MAE gives the average absolute difference between the predicted and actual values. It describes, on average, how much the model's predictions differ from the true values [18]. A lower MAE indicates better model performance, as it reflects closer alignment between the model's predictions and the actual values, which can be calculated using Equation 2, below. The nMAE metric is the MAE metric divided by the range of values and multiplied by 100 to further contextualize the model results, showing the MAE value in the form of a percentage.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |diff| \qquad \text{(Eqn. 2)}$$

MSE gives a sense of the average magnitude of the prediction errors. Since the errors are squared, larger errors contribute disproportionately more to the MSE, making it sensitive to outliers [18]. Much like the MAE metric, a lower MSE indicates better model performance, signifying that the predictions are closer to the actual values on average. The MSE can be calculated in a similar manner, as shown in Equation 3.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (diff)^2 \qquad \text{(Eqn. 3)}$$

RMSE is simply the square root of the MSE value. It can be used to compare the performance of different models on the same dataset. A model with a lower RMSE is generally considered better in terms of prediction accuracy [18]. For all of the ML metrics mentioned, 0.0 would indicate perfect predictions.

CNNs offer notable advantages over traditional methods, including faster computation, efficient feature extraction, and greater reproducibility [19]. Unlike conventional statistical approaches, CNNs reduce variability introduced by human intervention and are generally less computationally demanding in the long term [20]. Despite their potential, a conclusive application of ML for measuring weak lensing shear remains to be found [21].

In this paper, we explore the development of a linear regression model using a CNN to predict the ellipticity values of simulated, Fourier-transformed single-galaxy images. In Section 2, we address the collection and processing of our input data, as well as explore the construction and training of various CNN architectures. We compare, examine, and document our model comparison findings in Section 3, using the aforementioned ML metric as our methods of assessment. Section 4 discusses potential improvements to our study, particularly in making the data more realistic and the model more complex. Section 5 concludes our analysis.

2. Methodology

2.1 Simulating Galaxy Images

Since observational data on gravitationally lensed galaxies in the radio domain are scarce, simulations are essential for generating sufficient data to effectively train the network [21]. Using a version of the SimuCLASS pipeline [22, 23], we generated 10,100 simulations with ellipticity values ranging from 0.0 to 0.4 in increments of 0.004 (1%). For each ellipticity value, 100 images were produced. These values were chosen because, as mentioned in Section 1.1, they align with the expected order of magnitude for weak lensing effects [24].

For each simulation, we generated data files and corresponding truth catalogs for each image. Each image had a resolution of 200×200 pixels, with each pixel representing 0.45 arc seconds on the sky. The simulations varied in several aspects: galaxy size, orientation, and noise level, while maintaining Gaussian profiles and constant shear values. Galaxies ranged in size from 0.75 to 10.0 arc seconds and had fluxes varying between 5.0×10^{-5} and 5.0×10^{-4} Janskys. The images included both random Gaussian pixel noise and correlated Gaussian noise, with signal to noise ratios that ranged between 5.15×10^{2} and 1.4×10^{6} depending on the flux and noise values of each image, calculated in Equation 4, below:

$$SNR = \frac{f_{peak}}{noise} \qquad \text{(Eqn. 4)}$$

We conducted the entire simulation process within a Conda-based virtual environment [23]. Subsequently, we transitioned to a custom virtual environment installed with TensorFlow for preprocessing, building, and training our frequency domain CNN (FCNN) [25].

2.1.1 Preprocessing.

For each pixel file generated from the simulations, the pixel values were transformed into the Fourier space. The Fourier-transformed pixel values were then separated into their real and imaginary components and normalized. This preparation was done in advance of inputting the data into the FCNN for further analysis and modeling. Although separating the data into real and imaginary components results in a loss of the inherent relationship between them, this approach was taken because TensorFlow and Keras do not natively support complex numbers. Thus, we chose to preprocess the data in this way to streamline the model building, training, and evaluation processes, considering the constraints of time and resources.

The dataset was divided as follows: 33% was reserved for testing the model, while 67% was used for training. Within the training set, 25% was further allocated for validation during the training process.

2.2 Network Architecture

Throughout this project, we experimented with various models to identify one with satisfactory performance. Each linear regression model followed a similar base structure: an input layer, a convolutional layer followed by a pooling layer, a flattening layer, and dense layers (including the output layer). The primary differences among the models lie in the number of layers and the incorporation of various optimization techniques, such as regularization (L1 and L2), advanced optimization algorithms (e.g., Adam, stochastic gradient descent (SGD), and root mean square propagation (RMSProp)), and hyperparameter optimization.

Using Keras' Functional API, we designed the input layer for each model to handle two channels of data: the real and imaginary components of the Fourier-transformed pixel values. To evaluate model performance, we employed several metrics discussed in Section 1.3 to measure the discrepancy between the label values and the predicted values for each prediction across the training, validation, and testing datasets.

In total, we tested 10 different models to determine the best-performing one. In our best performing model, shown in Figure 1, the architecture is as follows: it begins with an input layer, followed by three 2D convolutional layers. Each convolutional layer is succeeded by a 2D MaxPooling layer, totaling three MaxPooling layers. After these, the network includes a flattening layer, three fully connected (dense) layers, and concludes with an output dense layer.

2.3 Model Training

In the model-building phase of our project, we employed a variety of training strategies to enhance model performance.
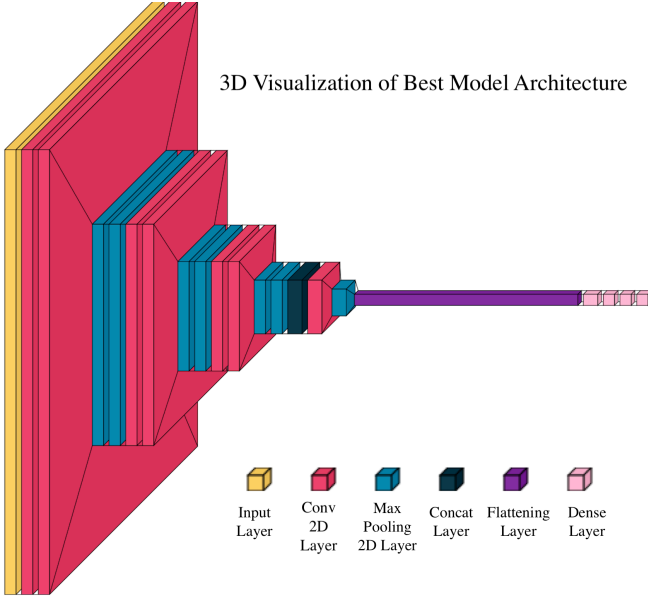
Figure 1: A 3D layering architecture visualization of the model. The diagram illustrates how the layers are structured, with distinct doubling indicating the separate processing paths for the real and imaginary components of the Fourier input data.

We experimented with different numbers of epochs, batch sizes, learning rates, and optimization algorithms, adjusting each factor to identify the optimal configuration for performance and convergence within our timeframe.

The configuration of our optimal model was as follows: the model was trained for 24 out of 30 epochs using early stopping with a patience of 5 epochs. It utilized the MAE loss function with a batch size of 64, the Adam optimizer with a learning rate of 0.001, and no regularization.

The model architecture and parameters described above provided a solid foundation for further development and optimization, leading to the results discussed below.

3. Results and Discussion

Plotting true labels against predicted labels in an optimal regression model should ideally produce a straight line with a slope of 1, passing through the origin, indicating a perfect match between predictions and actual values. Figure 2 visualizes this relationship for our top-performing model by plotting the true and predicted labels for each training and evaluation epoch. Given that our dataset contains 100 points per ellipticity value, we averaged the predicted labels at each step, resulting in one average predicted label for each true label.

We excluded the first 6 epochs from the visualizations of our training and validation sets due to a 'burn-in' effect during model training, where predictions are often unstable and do not accurately reflect the model's true capability [26]. From the 7th epoch onward, however, the model shows significant improvement, with the validation MAE dropping
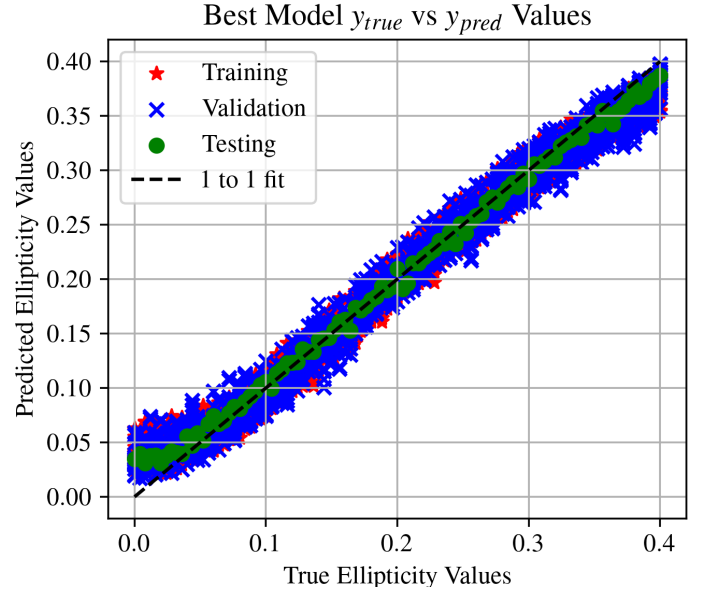

Figure 2: A scatterplot illustrating the performance of the model across training, validation, and testing datasets for the 7th epoch of training and evaluation and beyond. The plot compares the true labels and the predicted labels of the ellipticity values, showing how the model's performance evolves over time. Different markers or colors represent training, validation, and testing data, allowing for an easy comparison of how each dataset's performance metrics change throughout the training process.

to 0.03 or lower and MSE decreasing to 0.002 or below. This is in contrast to the first 6 epochs, where the validation MAE remains around 0.067 or higher and the MSE above 0.007. This improvement in metrics, seen in Figure 3, particularly in the validation data, led us to focus our analysis on epochs 7 and beyond, where the model's predictions more reliably reflect its learned patterns and overall effectiveness.

For comparison, Figure A.1 in the Appendix shows a visualization of Figure 2 with the first 6 epochs. Figures A.2 and A.3 illustrate the training and validation data "burn-in" imbalance by displaying the frequency of predictions. Additional performance metrics for these initial epochs are detailed in Table A.1.

Examining the histogram of testing data for our best-performing model, shown in Figure 4, we observe a slight tendency for predictions to cluster closer to 0.0 ellipticity. This could potentially be due to the loss function we chose to use, since MAE does not weigh larger errors more heavily. This can cause the model to predict values that reduce the average magnitude of errors, potentially resulting in a concentration of predictions near lower values like 0.0. However, the overall distribution of predictions aligns decently with the true labels, with a significant overlap between the two. This, along with the results in Figure 2, suggests that our model maintains a good balance between variance and bias.
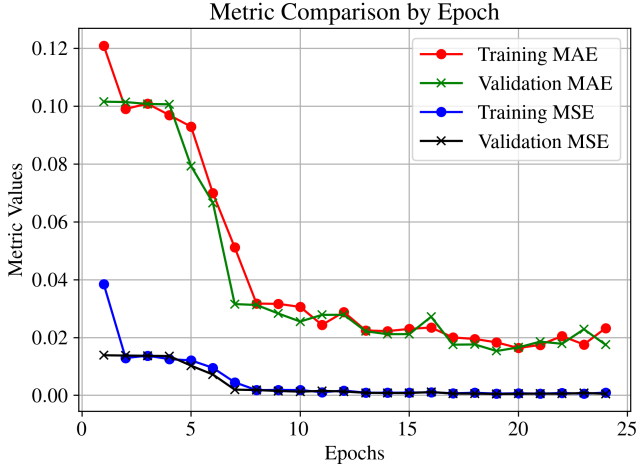
Figure 3: A plot illustrating the "burn-in" effect observed during the model's training process, which stabilizes after approximately 7 epochs.

3.1 True vs. Predicted Labels

Focusing on the green testing data points in Figure 2, and supported by the training and validation data, we find that the model generalizes well to unseen data. After 7 epochs, the model shows consistent performance with the testing data, indicating that it is no longer noticeably underfitting the data.

While our best-performing model demonstrates good generalization to unseen data, it is important to note that, as shown in Figure 2, we average the true and predicted labels for every epoch across 100 samples per ellipticity value. This averaging may not fully capture the frequency of close predictions or the distribution of predictions for each value. Although the testing data indicates promising performance, noticeable errors remain, especially when compared to the standards typically accepted in the astrophysical community, which is often around 1% [11].

3.2 Errors and Metrics

The performance of our top 3 models is detailed in Table 1, focusing on the testing data metrics gathered from our model evaluation, rounded to 5 decimals places. For brevity, training and validation metrics are provided in the Appendix. As outlined in Section 1.3, lower MAE, MSE, and RMSE scores indicate better alignment between predictions and actual values.

	MAE	nMAE	MSE	RMSE	Diff
1	0.01480	3.70%	0.00046	0.02145	-0.00093
2	0.01596	3.99%	0.00052	0.02280	-0.00137
3	0.02882	7.21%	0.00138	0.03715	0.01733

Table 1: The testing metrics for the top 3 performing models.

Our best-performing model demonstrates an MAE of 0.0148 units, meaning that, on average, the predictions deviate from actual values by this amount. Since MAE is less sensitive to outliers, it provides a more accurate representation of typical prediction errors. To contextualize this result, we calculate the normalized MAE (nMAE) using
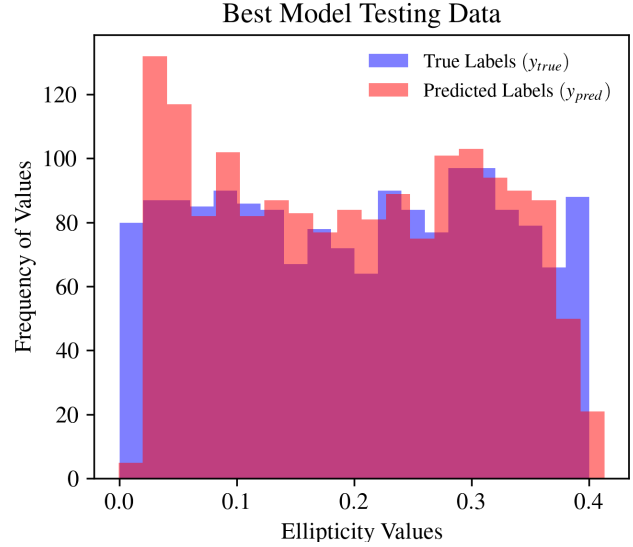

Figure 4: A histogram showing the distribution of overlap between true and predicted labels for the testing data across each epoch of training and evaluation. This histogram details how the model's predictions align with the actual true labels over time, providing insight into the accuracy and consistency of the model's performance during the training process.

Equation 5, where the denominator represents the range of values [27]. Converting the nMAE to a percentage highlights that the average prediction error is relatively small compared to the total variation in the target values:

$$nMAE = (\frac{MAE}{0.4}) * 100 \qquad \text{(Eqn. 5)}$$

The MSE of 4.6×10^{-4} indicates that the average squared difference between predictions and actual values is very low, emphasizing that larger errors are minimal. The corresponding RMSE of 0.02145 units reinforces this, indicating that typical deviations between predictions and actual values are small, supporting the model's strong generalization to unseen data. On average, the predictions deviate only slightly from the true values, indicating precise performance. Across all metrics, our best model consistently outperforms the others, showcasing its overall accuracy and reliability.

The errors mentioned above could stem from several factors, some of which have already been discussed. For instance, the separation of our input data into real and imaginary components might impair the model's ability to fully and accurately align predictions with the true labels. Additionally, the hyperparameters, including the choice of optimizer, early stopping patience value to prevent overfitting, and various metrics, may not have been optimized as thoroughly as possible. These elements could contribute to discrepancies between the predicted and actual values.

While the model performs well with the provided data, it is important to acknowledge that this data is not fully representative of real-world scenarios. Consequently, the

model is likely to underperform when faced with actual observational data. The simplicity of the model reflects the straightforward task it was given—analyzing single galaxy images simulated with Gaussian noise profiles and constant shear values. To excel with real-world data, a significantly deeper model would be required to capture the complexities of varying noise patterns, shear values, and signal-to-noise ratios (SNR), as well as to accurately differentiate between overlapping galaxy fluxes.

4. Future Work

The results presented in this paper represent an initial effort to predict shear values from galaxy simulations with varying noise levels in radio interferometer data using CNNs. Several avenues for further optimization were not explored due to time constraints. For instance, the model architecture could be enhanced, specifically when implementing more realistic data to train the model, by leveraging deeper architectures that have demonstrated strong performance, such as pre-trained networks like AlexNet or VGGNet, each comprising over 100 layers. Additionally, employing Bayesian optimization could refine the hyperparameters and further improve the model's performance.

Beyond that, another network could be constructed in which each layer is custom built to handle real and complex data at the same time. This can be done using SubClassing Models in Keras to build custom layers, or branching out to other tools such as PyTorch or MatLab to explore the ways in which they naturally handle complex valued inputs.

In terms of data enhancement, our simulations could be made more realistic. The noise models used in our input data are simplified and do not fully capture the complexities of weak lensing. Specifically, our current simulations assume a constant gravitational shear applied to single galaxy images, which does not reflect the variability found in real observations [20]. In addition, adding relative calibration errors between different Fourier space pixels would aid in the realism of our data, as we currently do not take that into account in our model or our simulations. Once an effective model for predicting constant gravitational shear is developed, the next step would be to introduce more complexity. This involves incorporating multiple galaxies with cosmologically varying shear fields. This approach would account for variations in the number, size, shape, and orientation of galaxies within an image, as well as the shear field distortions. Such an enhancement would require extensive simulations to cover a broad range of parameter distributions and more accurately represent real-world conditions.

Finally, to enhance the realism of the simulations, we could mask specific regions of the Fourier plane to mimic the limited UV coverage (spatial frequency coordinates) typical of actual interferometers. In radio interferometry, the UV coverage is determined by the arrangement and spacing of antennas in the interferometer. Due to practical limitations, such as the finite number of antennas and their specific configuration, the interferometer can't sample all spatial frequencies equally [28]. This results in incomplete coverage of the "UV plane," meaning some regions of the Fourier domain are not observed.

5. Conclusions

Analyzing radio interferometry data is a complex task, particularly due to the intricate PSF that complicates deconvolution. ML offers a promising alternative by potentially bypassing deconvolution and enabling the direct measurement of shear values caused by weak lensing using Fourier-domain radio interferometry data.

Our work demonstrates a successful model for measuring ellipticity values in single-galaxy simulations, as reflected in our performance metrics. However, significant progress is still required to enhance the accuracy of ML approaches in this domain. The suggestions for future work indicate the need for addressing more realistic data, complex noise patterns, and exploring deeper model architectures. Nevertheless, this research provides a solid foundation for future advancements, offering a stepping stone towards more effective and reliable methods in the field of cosmological analysis using machine learning.

References

1. Pires, S. et al. "FAst STatistics for Weak Lensing (FASTLens): Fast Method for Weak Lensing Statistics and Map Making." Monthly Notices of the Royal Astronomical Society 395, no. 3 (May 21, 2009): 1265–79.
2. Wilde, Joshua et al. "Detecting Gravitational Lenses Using Machine Learning: Exploring Interpretability and Sensitivity to Rare Lensing Configurations." Monthly Notices of the Royal Astronomical Society 512, no. 3 (May 21, 2022): 3464–79.
3. Xiao, Huangyu, et al. "Detecting Dark Matter Substructures on Small Scales with Fast Radio Bursts." Physical Review D 110, no. 2 (July 12, 2024): 023516.
4. Schneider, P. "Weak Gravitational Lensing." In Gravitational Lensing: Strong, Weak and Micro, edited by Peter Schneider, Christopher S. Kochanek, and Joachim Wambsganss, 269–451. Berlin, Heidelberg: Springer, 2006.
5. Refsdal, Sjur, and Jean Surdej. Highlights of Astronomy - Chapter: Gravitational Lensing. Vol. Vol. 9, 3, 1992.
6. Wambsganss, Joachim. "Gravitational Lensing in Astronomy." Living Reviews in Relativity 1, no. 1 (December 1, 1998): 12.

7. Bacon, David, Alexandre Refregier, and Richard Ellis. "Detection of Weak Gravitational Lensing by Large-Scale Structure." Monthly Notices of the Royal Astronomical Society 318, no. 2 (October 2000): 625–40.

8. Nick Kaiser. "A New Shear Estimator for Weak-Lensing Observations." The Astrophysical Journal, 537:555-577. July 10, 2000.

9. Van Waerbeke, L., Y. Mellier, T. Erben, J. C. Cuillandre, F. Bernardeau, R. Maoli, E. Bertin, et al. "Detection of Correlated Galaxy Ellipticities from CFHT Data: First Evidence for Gravitational Lensing by Large-Scale Structures." Astronomy and Astrophysics 358 (June 1, 2000): 30–44.

10. Wittman, David M., J. Anthony Tyson, David Kirkman, Ian Dell'Antonio, and Gary Bernstein. "Detection of Weak Gravitational Lensing Distortions of Distant Galaxies by Cosmic Dark Matter at Large Scales." Nature 405, no. 6783 (May 2000): 143–48.

11. Mandelbaum, Rachel, Barnaby Rowe, James Bosch, Chihway Chang, Frederic Courbin, Mandeep Gill, Mike Jarvis, et al. "The Third Gravitational Lensing Accuracy Testing (GREAT3) Challenge Handbook." The Astrophysical Journal Supplement Series 212, no. 1 (April 17, 2014): 5.

12. Survey, Dark Energy, Kilo-Degree Survey Collaboration, T. M. C. Abbott, M. Aguena, A. Alarcon, O. Alves, A. Amon, et al. "DES Y3 + KiDS-1000: Consistent Cosmology Combining Cosmic Shear Surveys." The Open Journal of Astrophysics 6 (October 20, 2023): 10.21105/astro.2305.17173.

13. Miyatake, Hironao, Sunao Sugiyama, Masahiro Takada, Takahiro Nishimichi, Xiangchong Li, Masato Shirasaki, Surhud More, et al. "Hyper Suprime-Cam Year 3 Results: Cosmology from Galaxy Clustering and Weak Lensing with HSC and SDSS Using the Emulator Based Halo Model." PHYSICAL REVIEW D 108, April 6, 2023.

14. Woody, David. "Theory and Statistics of Radio Interferometer Array Point Spread Functions," 2001.

15. A. Richard Thompson et al. Interferometry and Synthesis in Radio Astronomy. 3rd ed. Springer International Publishing, 2017.

16. Hunt, Lucas R., Megan C. Johnson, Phillip J. Cigan, David Gordon, and John Spitzak. "Imaging Sources in the Third Realization of the International Celestial Reference Frame." The Astronomical Journal 162, no. 3 (August 2021): 121.

17. Maulud, Dastan, and Adnan M. Abdulazeez. "A Review on Linear Regression Comprehensive in Machine Learning." Journal of Applied Science and Technology Trends 1, no. 2 (December 31, 2020): 140–47.

18. Chicco, Davide, Matthijs J. Warrens, and Giuseppe Jurman. "The Coefficient of Determination R-Squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation." PeerJ Computer Science 7 (July 5, 2021): e623.

19. Schmidt, K., F. Geyer, S. Fröse, P.-S. Blomenkamp, M. Brüggen, F. De Gasperin, D. Elsässer, and W. Rhode. "Deep Learning-Based Imaging in Radio Interferometry." Astronomy & Astrophysics 664 (August 2022): A134.

20. Zitha, Simphiwe, Arun Aniyan, Oleg Smirnov, and Risuna Nkolele. "ZCal: Machine Learning Methods for Calibrating Radio Interferometric Data," October 2, 2020.

21. Rezaei, S, J P McKean, M Biehl, W de Roo, and A Lafontaine. "A Machine Learning Based Approach to Gravitational Lens Identification with the International LOFAR Telescope." Monthly Notices of the Royal Astronomical Society 517, no. 1 (November 21, 2022): 1156–70.

22. Harrison, Ian, Michael L. Brown, Ben Tunbridge, Daniel B. Thomas, Tom Hillier, A. P. Thomson, Lee Whittaker, et al. "SuperCLASS -- III. Weak Lensing from Radio and Optical Observations in Data Release 1." Monthly Notices of the Royal Astronomical Society 495, no. 2 (June 21, 2020): 1737–59.

23. Harrison, Ian. "Itrharrison/Rwl_sims." Python, August 13, 2024. https://github.com/itrharrison/rwl_sims.

24. Croft, Rupert A. C., and Christopher A. Metzler. "Weak-Lensing Surveys and the Intrinsic Correlation of Galaxy Ellipticities." The Astrophysical Journal 545, no. 2 (December 20, 2000): 561–71.

25. Sofie Seilnacht. "fourierCNN/Simuclass.Yml at Main · Sofieseilnacht/fourierCNN." GitHub. Accessed August 28, 2024. https://github.com/sofieseilnacht/fourierCNN/blob/main/simuclass.yml.

26. Jian-Fang Hu et al. "Early Action Prediction by Soft Regression." IEEE Transactions on Pattern Analysis and Machine Intelligence 41, no. 11 (n.d.).

27. Piotrowski, Paweł, Inajara Rutyna, Dariusz Baczyński, and Marcin Kopyt. "Evaluation Metrics for Wind Power Forecasts: A Comprehensive Review and Statistical Analysis of Errors." Energies 15, no. 24 (January 2022): 9657.

28. Ségransan, Damien. "Observability and UV Coverage." New Astronomy Reviews, Observation and Data Reduction with the VLT Interferometer, 51, no. 8 (October 1, 2007): 597–603.

Appendix:
Table A.1: Best Model Training and Validation Metrics per Epoch

	Train MAE	Train nMAE	Train MSE	Train RMSE	Train Diff	Val MAE	Val nMAE	Val MSE	Val RMSE	Val Diff
1	0.12085	30.21%	0.03842	0.19601	-0.01103	0.10155	25.39%	0.01388	0.11781	-0.01166
2	0.09904	24.76%	0.01293	0.11371	0.00682	0.10142	25.34%	0.01378	0.11739	0.00615
3	0.10086	25.22%	0.01374	0.11722	-0.00957	0.10076	25.19%	0.01370	0.11705	-0.00997
4	0.09692	24.23%	0.01253	0.11194	0.00405	0.10066	25.17%	0.01361	0.11666	0.00359
5	0.09290	23.23%	0.01211	0.11005	0.05728	0.07927	19.82%	0.01028	0.10139	0.05946
6	0.06990	17.48%	0.00953	0.09762	0.01712	0.06660	16.65%	0.00722	0.08497	0.01875
7	0.05115	12.79%	0.00453	0.06731	-0.00422	0.03157	7.89%	0.00200	0.04472	-0.00435
8	0.03175	7.94%	0.00183	0.04278	-0.00570	0.03127	7.81%	0.00185	0.04301	-0.00594
9	0.03165	7.91%	0.00189	0.04347	0.01717	0.02831	7.08%	0.00153	0.03912	0.01615
10	0.03060	7.65%	0.00185	0.04301	0.00967	0.02554	6.39%	0.00130	0.03606	0.00874
11	0.02437	6.09%	0.00111	0.03332	0.01552	0.02785	6.96%	0.00157	0.03962	0.015156
12	0.02883	7.21%	0.00168	0.04099	-0.00644	0.02787	6.97%	0.00136	0.03688	-0.00686
13	0.02245	5.61%	0.00094	0.03066	-0.00521	0.02219	5.55%	0.00090	0.03000	-0.005832
14	0.02223	5.56%	0.00093	0.03050	-0.00433	0.02117	5.29%	0.00083	0.02881	-0.00482
15	0.02299	5.75%	0.00096	0.03098	-0.00323	0.02120	5.30%	0.00082	0.02864	-0.00350
16	0.02348	5.87%	0.00109	0.03302	-0.01903	0.02720	6.80%	0.00117	0.03421	-0.01919
17	0.02003	5.01%	0.00073	0.02702	-0.00363	0.01754	4.39%	0.00064	0.02530	-0.00357
18	0.01952	4.88%	0.00092	0.03033	-0.00493	0.01762	4.41%	0.00063	0.02510	-0.00493
19	0.01833	4.58%	0.00062	0.02490	-0.00078	0.01535	3.84%	0.00050	0.02236	-0.00056
20	0.01638	4.10%	0.00075	0.02739	-0.00052	0.01660	4.15%	0.00057	0.02387	-0.00026
21	0.01747	4.37%	0.00059	0.02429	0.00476	0.01859	4.65%	0.00063	0.02510	0.00512
22	0.02052	5.13%	0.00085	0.02915	-0.00379	0.01785	4.46%	0.00057	0.02387	-0.00332
23	0.01760	4.40%	0.00059	0.02429	0.01869	0.02287	5.72%	0.00086	0.02933	0.01830
24	0.02329	5.82%	0.00098	0.03130	0.00113	0.01751	4.38%	0.00056	0.02366	0.00082

Table A.1: Includes the training and validation metrics per epoch for the best performing model.
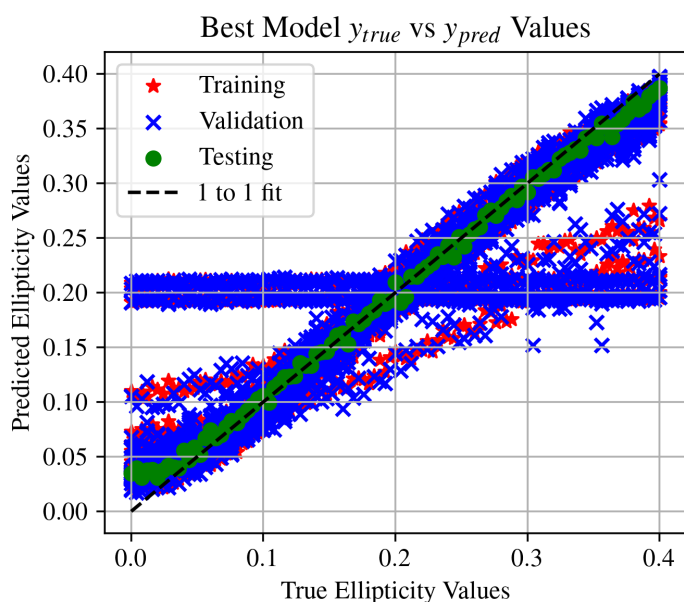
Figure A.1:

Figure A.1: A scatterplot illustrating the performance of the model across all epochs of training, validation, and testing datasets for each epoch of training and evaluation. The plot compares the true labels and the predicted labels of the ellipticity values, showing how the model's performance evolves over time. Different markers or colors represent training, validation, and testing data, allowing for an easy comparison of how each dataset's performance metrics change throughout the training process.
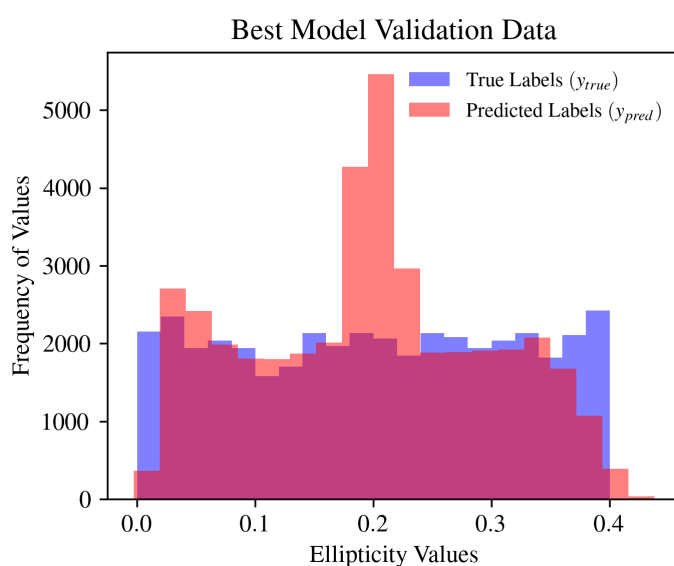
Figure A.3:

Figure A.3: A histogram showing the distribution of overlap between true and predicted labels for the validation data across each epoch of training. This illustrates the "burn-in" effect with the frequency of predicted values in the first 6 epochs.
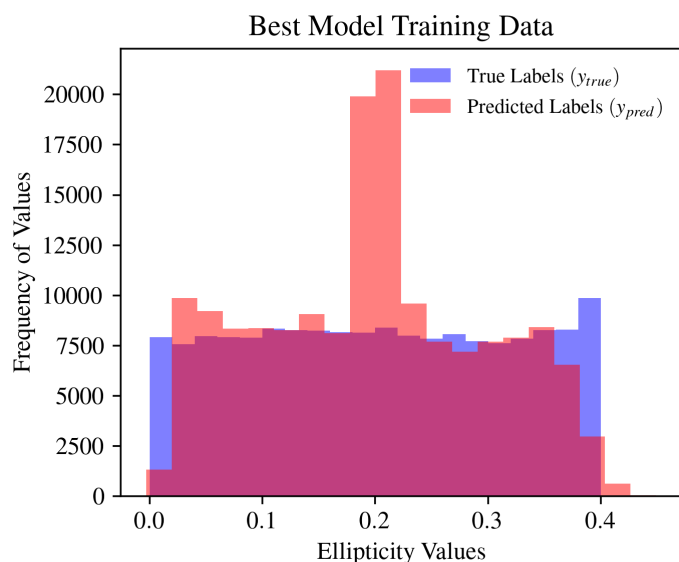
Figure A.2:

Figure A.2: A histogram showing the distribution of overlap between true and predicted labels for the training data across each epoch of training. This illustrates the "burn-in" effect with the frequency of predicted values in the first 6 epochs.