BurstDetector: Using Machine Learning to Predict Starbursts in Galaxy Mergers

Sakethram Badri and Vishesh Mishra

Abstract

When two galaxies merge together, several phenomena such as starbursts can ensue. Studying such phenomena is crucial to understanding how galaxies and stars interact, with the analysis of star formation rates providing insight into the many mysteries of our universe. Current methods of direct imaging and spectral analysis in analyzing mergers are mostly manual and not automated, oftentimes being prone to error as well. With large datasets emerging online through more and more readily available simulation data, more efficient methods must be developed to study such data. Machine learning techniques can expedite such processes, with this paper aiming to evaluate 3 techniques known for successful image classification in their success with automating the analysis of images from these datasets: Convolutional Neural Networks (CNN), Random Forest Algorithms (RF), and Support Vector Machines (SVM). Trained and cross-validated on image data from the Sloan Digital Sky Survey, our CNN ”BurstDetector” yielded the most success with an accuracy of 92.7% in detecting the occurrence of starbursts, demonstrating that CNNs tend to experience the most success in this image classification task of the 3 models evaluated. BurstDetector can also be run on a multitude of computers regardless of their GPU, making it computationally efficient. A computationally efficient model like BurstDetector is essential to being able to interpret the tremendous amount of data online. The study of the resultant stars forming in merging galaxies through their images can be pivotal to making new discoveries in the field of physics and astronomy, opening the door to revelations in the structure of the universe and even progress with dark matter.

Introduction

Galaxy mergers occur when two roughly equally sized galaxies collide against one another, resulting in a violent interaction between the contents of the two cosmic bodies. These events can cause phenomena such as intense bursts of star formation, known as starbursts. Understanding such phenomena is important for advancing our knowledge of galaxy evolution and the processes that shape the universe. Predicting when and where starbursts occur in merging galaxies could potentially provide insight into the mechanisms that trigger intense periods of star formation in the universe. Research on the activity of starburst mergers can even reveal the unseen distribution of dark matter within galaxies, which is one of the biggest mysteries in astrophysics.

Rationale

Having an accurate and efficient method of knowing which galaxies to study further will save time when analyzing mergers. With most current methods on conducting such analysis being manual, doing this research is extremely time consuming. Machine learning can enable the efficient analysis of vast astronomical datasets, which in recent times have begun to develop more and more, and uses algorithms to predict phenomena with greater accuracy and speed.

Machine learning techniques can revolutionize the way scientists interpret interactions between galaxies, providing predictive power that was previously unavailable through manual techniques. The scientific impact of the project extends beyond galaxy evolution, contributing to the field of astronomy by accelerating discoveries and improving data analysis techniques. Moreover, the predictive power of machine learning can transfer over to various other fields, ultimately advancing scientific research processes and technological innovation in society.

Data Collection

Firstly, we constructed an SQL query to filter for galaxy mergers and gather images of galaxy mergers in FITS format with their corresponding star formation rates from the Sloan Digital Sky Survey (SDSS) SQL image query [1], [2]. Given the limited availability of high-quality merger images, TensorFlow-based data augmentation techniques [3] were utilized to artificially expand the dataset while preserving astrophysical accuracy. 14517 images were created, and were split into 11613 training images and 2904 cross validation images (80/20 split).

- Geometric Transformations (Rotating, Flipping, Scaling)
- Color Transformations (Brightness, Saturation Adjustment)
- Addition of Artificial Noise (Blur, Sharpness)

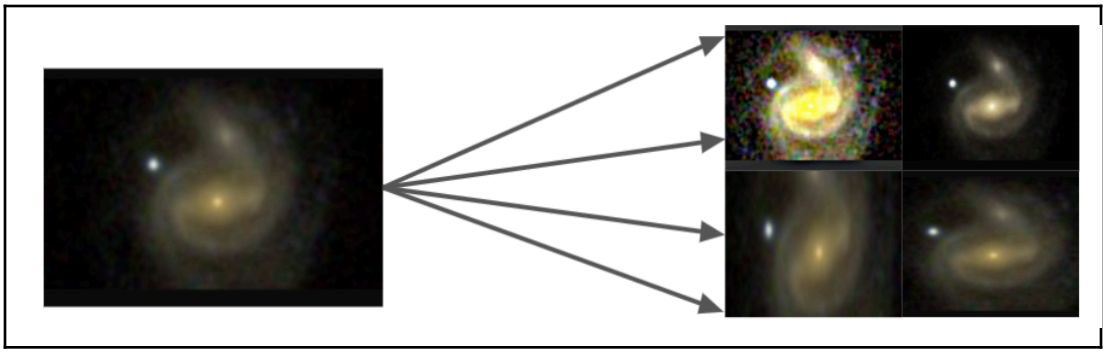

Figure 1: Example of how Tensorflow augments images to expand a dataset

Model Development

Three models with three different machine learning algorithms were trained and developed: Convolutional Neural Networks (CNN), Random Forest (RF), and Support Vector Machine (SVM) [4], [5]. Each model was programmed and executed in VSCode Jupyter Notebooks to implement and fine-tune the training process. To determine the most effective starburst classification model, the cross-validation dataset was fed through all three models, and their performances were assessed using multiple evaluation metrics, including accuracy, precision,

sensitivity, specificity, and F1-Score. After initial evaluations, we refined the most effective model, the CNN, by tuning parameters such as learning rates and improved preprocessing techniques to enhance key astrophysical features, such as gas density and tidal tails.

Results

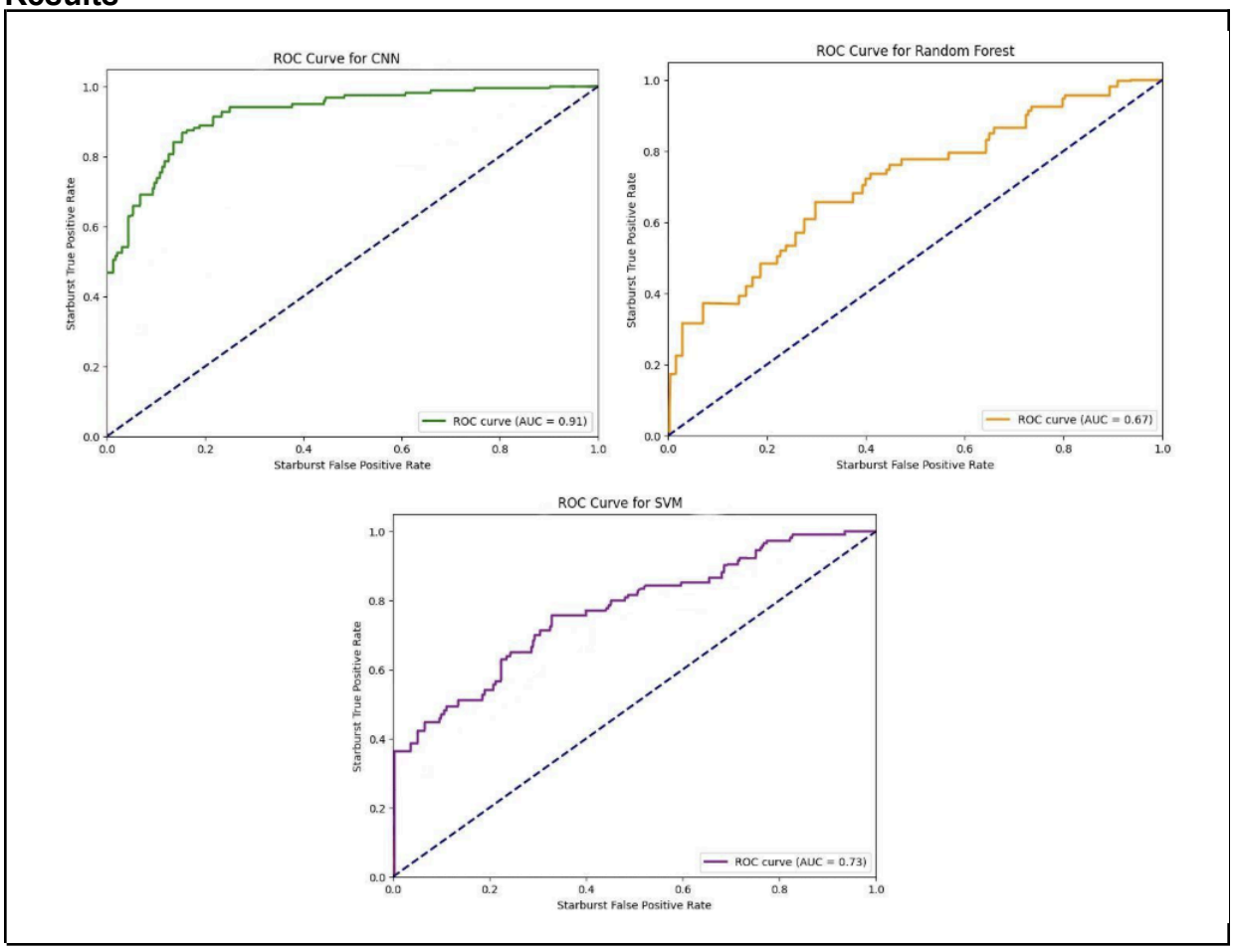

Figure 2: ROC curves comparing AUC across starburst classification models

As seen in Figure 2, the Convolutional Neural Network (CNN) yielded the most effectiveness with the highest AUC (Area Under Curve) of 0.91 on the ROC curves. A higher AUC signifies a more effective model in making accurate predictions.

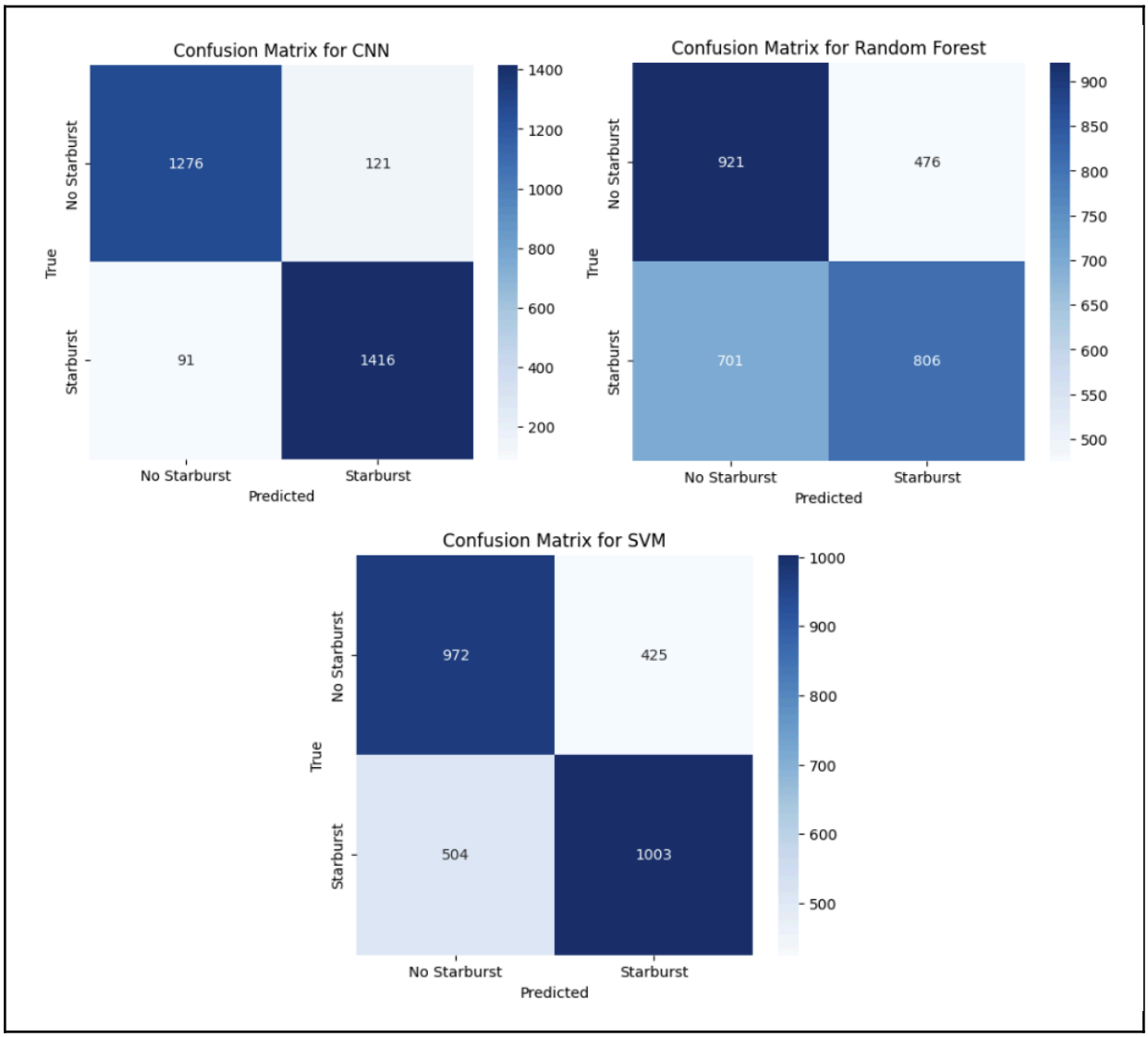

Figure 3: Confusion Matrices visually representing true and false positives and negatives

The Confusion Matrix further confirmed the CNN's success, showing the largest proportion of correct predictions, minimizing Type I and Type II errors.

CNN Metrics	Value
Accuracy	0.927
Precision	0.921
Recall (Sensitivity)	0.939
F1 Score	0.931
Specificity	0.913

RF Metrics	Value
Accuracy	0.595
Precision	0.629
Recall (Sensitivity)	0.535
F1 Score	0.578
Specificity	0.659

SVM Metrics	Value
Accuracy	0.681
Precision	0.702
Recall (Sensitivity)	0.666
F1 Score	0.684
Specificity	0.696

Figure 4: Metrics for each classification model. The CNN outperforms both of the other models in every metric.

Additionally, the CNN model achieved an overall accuracy of 92.7% in detecting starbursts, indicating its high reliability and effectiveness in identifying the phenomena of starbursts. These observations as well as the direct comparisons in precision, recall, accuracy, F1-score, sensitivity, and specificity between the 3 techniques make it clear that the Convolutional Neural Network was the most effective, and underscores the potential of machine learning techniques to improve astronomical predictions.

The Convolutional Neural Network (CNN) used multiple layers to progressively learn and extract important features from the galaxy merger images. The strength of CNNs lies in their ability to automatically learn hierarchical features from raw image data, without the need for manual feature extraction [5].

- 1. Edge Detection and Low-Level Features: The initial convolutional layers were responsible for detecting basic visual features such as edges and textures. Filters were applied to identify simple structures like gradients and pixel intensity, which allows the model to detect the boundaries between merging galaxies.
- 2. Shape Recognition and Mid-Level Features: As the image data passed through subsequent convolutional layers, the network recognized more complex patterns, such as shapes and spatial relationships. These features include the identification of spiral arms, elliptical galaxy shapes, and merging tidal tails.
- 3. Gas Density and High-Level Features: The deeper layers of the network were critical in detecting astrophysical phenomena such as gas density. By analyzing regions with high gas concentration and star formation activity, the model recognized features of starburst

events, where star formation occurs due to gravitational interactions.

- 4. Starburst Classification: After extracting the layered features (edges, shapes, gas density) and combining the respective SFR data, the model used a fully connected layer to combine the learned information into a final classification. The final classification decision was based on the recognition of specific astrophysical features like tidal tails, gas density variations, and the morphology of interacting galaxies that typically precede starburst events

Processing each image in just 23 milliseconds, the CNN algorithm proves to be significantly faster than traditional human manual analysis. Such swift classification not only makes the process more efficient, but also allows for the rapid analysis of vast datasets. The resultant combination of high accuracy, speed, and ability to handle large datasets positions the CNN as a highly effective tool for astronomical research, saving time and resources.

Figure 5: Region of high gas concentration in merger

The main sequence of star-forming galaxies (SFMS) is an observed relationship between stellar mass (Mstar) and star formation rate (SFR) for galaxies that are actively forming stars. It describes how the majority of star-forming galaxies follow a nearly linear correlation between their SFR and stellar mass, meaning more massive galaxies tend to form stars at a higher rate. The SFMS is typically expressed as a power-law relation [6]:

$$\log(SFR) = \alpha \log(M*) + \beta$$

where:

$M*$ is the stellar mass of the galaxy (in solar masses, $M\odot$)

α (slope) is describes how SFR scales with stellar mass (often ~0.7 to 1.0)

B (intercept) depends on the cosmic epoch and specific dataset

The slope was set to 0.7 to suggest that more massive galaxies have slightly lower SFRs

than lower-mass galaxies—meaning they form stars less efficiently relative to their mass. The intercept of -7 ensures that at M* = 1010 MO, the observed SFR is close to 1 MO/yr. Starburst galaxies lie significantly above the main sequence, so the threshold for starburst formation from the SDSS data was 5 times the SFMS.

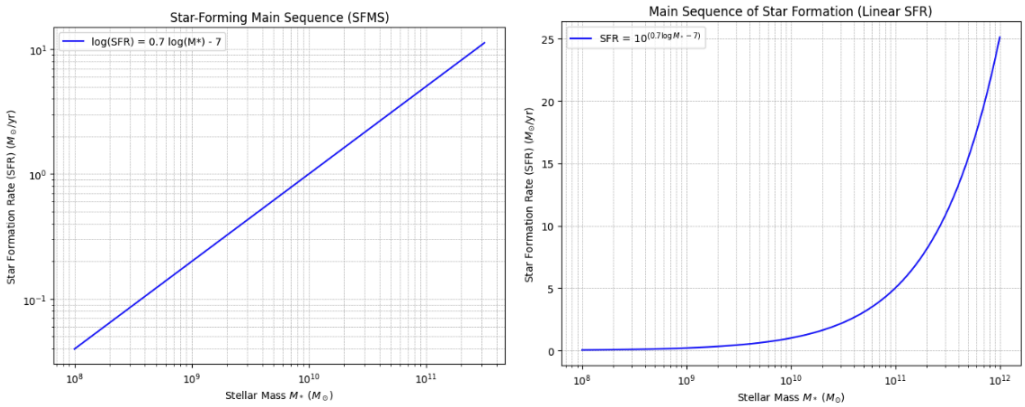

Figure 6: Correlation between SFR and Stellar Mass

Conclusion / Future Research

Since dark matter is theorized to shape the structure of the universe, understanding its role and connection to star formation is vital to the future of astronomy. Developing an efficient and accurate algorithm for classifying starbursts, such as BurstDetector, presents a unique opportunity to improve the accuracy of research in galaxy evolution and the role dark matter halos play in the universe. Starbursts, rapid periods of intense star formation, are often observed in galaxies influenced by the gravitational pull of dark matter halos which can trigger these explosive bursts of star activity. Missions such as the James Webb Space Telescope (JWST), Nancy Grace Roman Space Telescope, and upcoming NASA supported deep-sky surveys aim to study the structure and evolution of galaxies across cosmic time. Integrating our classification model, BurstDetector, with data from these missions could assist astronomers understand how dark matter-driven interactions shaped early galaxy formation. Additionally, our solution can be deployed to analyze data from NASA's Euclid mission, which is dedicated to studying dark energy and dark matter by mapping billions of galaxies.

Overall, our model BurstDetector is a necessity in today's world where classifications of star bursts are manual, slow, and inaccurate. BurstDetector not only opens up new possibilities for scientists to shift their focus and time from classifying starbursts to researching them, but also allows computers of any GPU to perform these classifications. This results in discoveries relating to the field of physics and astronomy being made at a much faster rate.

References

[1] Pearson, W. J., Wang, L., Trayford, J. W., Petrillo, C. E., & Van Der Tak, F. F. S. (2019). Identifying galaxy mergers in observations and simulations with deep learning. Astronomy & Astrophysics, 626, A49. https://doi.org/10.1051/0004-6361/201935355 Accessed 21 Dec. 2024.

[2] SDSS SkyServer DR12, skyserver.sdss.org/dr12/en/tools/search/sql.aspx. Accessed 3 Jan. 2025.

[3] Shorten, Connor, and Taghi M. Khoshgoftaar. ”A survey on image data augmentation for deep learning.” Journal of big data 6.1 (2019): 1-48. https://link.springer.com/article/10.1186/s40537-019-0197-0?code=b33ae7db-07a1-485c ac3b-4f409f373507 Accessed 5 Jan. 2025

[4] Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 6308-6325. https://doi.org/10.1109/JSTARS.2020.3026724 Accessed 26 Dec. 2024.

[5] Wu, J. (2017). Introduction to convolutional neural networks. National Key Lab for Novel Software Technology. Nanjing University. China, 5(23), 495. https://cs.nju.edu.cn/wujx/paper/CNN.pdf Accessed 26 Dec. 2024.

[6] Pearson, W. J., Wang, L., Alpaslan, M., Baldry, I., Bilicki, M., Brown, M. J. I., ... & van Der Tak, F. F. S. (2019). Effect of galaxy mergers on star-formation rates. Astronomy & Astro physics, 631, A51. https://doi.org/10.1051/0004-6361/201936337 Accessed 17 Dec. 2024.