Advanced Prediction of Specific Star Formation Rate (SSFR) Using Synthetic Data and Machine Learning Models

Satvik Raghav
Department of Electronics and Communication Engineering
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham
Bengaluru, India
satvikraghav007@gmail.com

Vikhyath B. M.
Department of Electronics and Communication Engineering
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham
Bengaluru, India
vikyath@gmail.com

Prasanth Ayitapu
Department of Electronics and Communication Engineering
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham
Bengaluru, India
payitapu@gmail.com

Raja Karthikeya
Department of Electronics and Communication Engineering
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham
Bengaluru, India
raja.karhikeyawork@gmail.com

Dr. T. K. Ramesh
Department of Electronics and Communication Engineering
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham
Bengaluru, India
tk_ramesh@blr.amrita.edu

Dr. Beena B. M.
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru,
Amrita Vishwa Vidyapeetham
India
bm_beena@blr.amrita.edu

***Abstract*—The accurate prediction of the Specific Star Formation Rate (SSFR) in galaxies is critical for understanding the evolution and dynamics of the universe. This study presents an advanced approach to the prediction of SSFR using real and synthetic data generated from the Sloan Digital Sky Survey Data Release 7 (SDSS-DR7). The primary goal is to improve predictive performance using machine learning models, including Linear Regression, Ridge, Lasso, Random Forest, and Gradient Boosting. Novel features, such as celestial position and derived feature transformations, were engineered to improve the accuracy of the model. A synthetic data set was generated to augment the original data, preserving the correlations among features while ensuring diversity in galaxy characteristics. This synthetic augmentation helped to overcome limitations posed by smaller datasets, ultimately improving the robustness of the model.**

The models were extensively evaluated using key performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) scores. Among the models tested, Random Forest demonstrated the best balance of predictive performance and computational efficiency. The project also involved clustering techniques like DBSCAN to identify regions of interest in the celestial map, highlighting zones with significant SSFR variations for potential satellite observations. The interactive visualisations further provided a user-friendly exploration of the SSFR predictions. The findings suggest that the incorporation of synthetic data and feature engineering strategies significantly improves the accuracy and generalisability of SSFR prediction models, providing valuable insights for astronomical research and observations.

***Index Terms*—Star Formation Rate, Synthetic Data Generation, Gaussian Copula, Regressors, Gradient Boosting, Clustering (DBSCAN)**

I. Introduction

Star Formation Rates (SFRs) are pivotal in tracing the evolutionary history of galaxies. Accurately measuring SFRs is crucial for understanding galaxy formation and evolution, as these rates reflect the amount of new stellar material being created over time. Traditionally, SFRs have been measured using spectroscopy, a technique that provides detailed information about the physical properties and stellar populations of galaxies. Spectroscopy, while powerful, is both time-consuming and expensive, requiring large telescopes and substantial observational effort. Consequently, it is challenging to apply this method to the extensive datasets produced by large-scale astronomical surveys like the Sloan Digital Sky Survey (SDSS).

The advent of large-scale photometric surveys has revolutionised the ability to estimate SFRs. Unlike spectroscopy, which measures the light spectra of individual galaxies, photometric surveys capture the total light across multiple wavelengths, offering a more accessible but less detailed view. Photometric data is more readily available and cover millions of galaxies, surpassing the spectral coverage of a few thousand. However, this approach presents its own challenges, as photometric data lack the spectral detail necessary to directly measure SFRs.

Recent advances in machine learning (ML) and deep learning (DL) have provided promising solutions to these challenges. These techniques excel in modelling complex, nonlinear relationships within data, making them well-suited for capturing the intricate dependen-

cies between photometric observations and SFRs. By training on existing spectroscopic datasets, ML and DL models can learn patterns [1]that generalise to photometric data, offering a cost-effective and scalable alternative to traditional methods.

This paper introduces a comprehensive approach to estimating SFRs using photometric data from SDSS Data Release 7 (SDSS-DR7), which includes photometry for over 27 million galaxies. A range of advanced ML algorithms, including linear regression, ridge, lasso, random forest, and gradient boosting, have been used to enhance the accuracy of SFR predictions. The approach also incorporated synthetic data [2]generation to improve model robustness and generalisation.

The findings demonstrate that these ML models can achieve high accuracy in the estimation of SSFR based solely on photometric parameters. By comparing the performance of different algorithms, the strengths and limitations in the context of photometric data have been highlighted. Furthermore, regions of interest in the sky with high variability in SSFR were identified, which could inform future observational campaigns. In addition, an interactive web interface was developed to facilitate the exploration and visualisation of the results, making the data accessible to both researchers and the public.

This work represents a significant step forward in the use of ML techniques for astrophysical applications, offering a scalable and effective method for SSFR estimation and contributing valuable insights into the distribution of star formation across the universe.

The rest of the paper is organised as follows. Section II provides a literature review of the existing approaches to estimating the SSFR (Specific Star Formation Rate) and the application of synthetic data in astronomy. Section III covers the methodology, including the description of the dataset,the data preprocessing steps, the synthetic data generation, and the modelling procedure. Section IV details the evaluation metrics and cross-validation approach used to assess the models. In Section V, results have been presented, including the performance of models on both original and synthetic data, and the key visualisations. Finally, Section VI concludes the study and outlines potential directions for future research.

II. Literature Review

Alpha MAML [3], an adaptive version of the Model-Agnostic Meta-Learning (MAML) technique, was introduced by Harkirat Singh Behl et al. Their goal was to reduce the requirement for manual intervention by automating hyperparameter tuning and improving few-shot learning performance. Alpha MAML showed better training stability and convergence speed, particularly when initial hyperparameters were not optimal. Unlike other strategies, this focused on making learning models more versatile across a range of tasks rather than using artificial data.

SynSys [4] is a system created by Jessamyn Dahmen and Diane Cook that is intended for the synthesis of synthetic data in healthcare applications, particularly for the recognition of smart home activities. The system mimics human behavior patterns by creating realistic synthetic data using Hidden Markov Models (HMMs). Their method improved the effectiveness of machine learning models by effectively combining synthetic data with real datasets, especially in situations where there was a shortage of real data. By adding synthetic data to tiny annotated datasets, they were able to significantly increase the accuracy of activity recognition.

Generative Adversarial Networks (GANs) [5] were used by Frid-Adar et al. to analyse medical data, producing artificial data to supplement real datasets for tasks like liver lesion categorization. Their research showed that adding synthetic data to original datasets improved the quantity and diversity of training datasets, which in turn improved classifier performance. This method worked especially well in medical applications where there is a dearth of real, annotated data.

An extensive analysis on synthetic data generation techniques was carried out by Alvaro Figueira et al. [6] with an emphasis on the incorporation of Generative Adversarial Networks (GANs). The distribution of the original dataset can be replicated in realistic data samples by GANs, which is a notable capability. This paper is notable for its thorough assessment of GAN architectures used in a variety of fields, including picture production, tabular data, and healthcare. The focus on techniques such as Gaussian Mixture Models (GMMs) for managing multimodal data, which allows precise synthetic data generation, is an important aspect of this review. In order to validate the effectiveness of machine learning models trained on synthetic data, the authors also examine various methods for evaluating the quality of synthetic data.

M. Delli Veneri et al. [7] employed various machine learning methods to estimate star formation rates (SFR) for photometric samples of galaxies, comparing their predictions with those from traditional methods. Their findings indicate that machine learning techniques effectively predicted SFRs, consistently outperforming conventional approaches. This study highlights the promise of machine learning for analysing large photometric datasets; however, it also points out limitations, particularly concerning the quality and completeness of the input data. The authors note that the current datasets are still insufficiently representative, which hampers the development of more robust and generalisable models.

Aufort et al. [8] used an Approximate Bayesian Computation (ABC) method to analyze the recent star formation history of galaxies. By utilizing a substantial

collection of mock galaxy spectra to approximate posterior distributions, they showed that ABC can yield dependable estimates of star formation histories, even when observational data is sparse. This research emphasizes the usefulness of ABC in astrophysics, especially for navigating the complexities and ambiguities found in galactic spectra.

At lower redshift $z \lesssim 2$, Georgios E Magdis et al. [9] presented an evolutionary sequence of the interstellar medium (ISM) in star-forming galaxies throughout their lifetimes, utilizing red and infrared spectral energy distributions. Their findings helped resolve the degeneracy in determining the dust and gas properties of galaxies from mid-infrared to millimeter data, yielding reliable estimates of dust masses—key factors in galaxy evolution—though at a relatively coarse resolution. To fully understand the ISM's complexity, additional high-resolution observations and simulations are necessary.

Using deep learning to estimate SFR from SDSS galaxy spectra [10]. By applying the convolutional neural network and using spectroscopic training data, Lovell et al. were able to achieve an amazingly high accuracy of around 92%. This demonstrates the potential of deep learning for handling spectral datasets effectively. The study also highlighted the promising prospects for automatic SFR determination while cautioning against overfitting and challenges related to training data size. The authors recommended further tuning of the models to enhance robustness and generalization capabilities.

V. Bonjean et al. [11] applied machine learning techniques to estimate SFR and stellar masses from photometric data. Their approach achieved significant accuracy improvements, with approximately 88% accuracy in SFR and stellar mass estimates. The study showcased the effectiveness of machine learning in managing multi-wavelength data but also pointed out challenges in data integration. The authors suggested employing advanced feature selection and data preprocessing methods to enhance model performance and address these challenges.

III. Methodology

This flowchart (fig.1) demonstrates the overall workflow of the research from data preprocessing phase with its related essential steps like feature selection, quality controlling, and database cross-matching, to model architecture which includes several machine learning and deep learning models, to training phase consisted of splitting train/test data in 80/20 ratio, performing k-fold cross validation and hyperparameter tuning, to inference phase consisting of MAE, MSE, RMSE, and R^2 metrics followed by result plotting. This process culminates in evaluating models on both the original and synthetic datasets, identifying regions of interest, developing interactive visualisations, and saving results for final analysis.

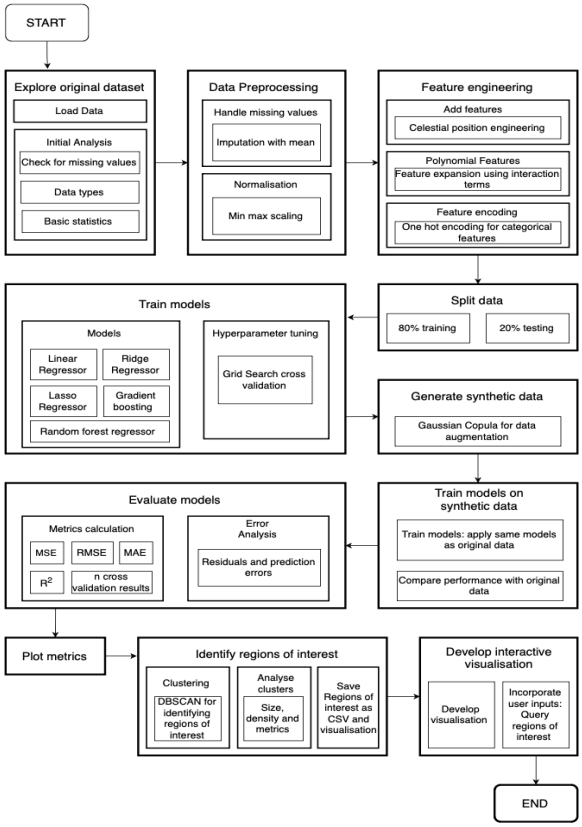

Fig. 1. Flow diagram explaining the model

A. Dataset

The primary dataset used in this research is derived from the Sloan Digital Sky Survey Data Release 7 (SDSS-DR7). It contains over 27 million galaxies, though a subset of 1,048,576 rows and 7 columns was utilised for this study. The columns in the dataset are: `dr9objID` (unique object ID), `objID` (SDSS object identifier), `RAdeg` (right ascension), `DEdeg`(declination), `photoz` (photometric redshift), `Qual` (categorical quality flag), and `SSFR` (target variable, specific star formation rate).

The dataset includes photometric data, which is more readily available and scalable than traditional spectroscopic data. However, estimating SSFR from photometric data is challenging because photometric measurements lack the precision of spectra. By using this dataset, the project aims to estimate SSFR accurately using machine learning models and synthetic data generation to augment the dataset and improve model performance.

Initial exploration of the dataset involved loading data from the `photometric_sfr.csv` file. The dataset was then analysed for missing values, data types, and basic statistics. Fortunately, no missing values were found, ensuring consistency for further preprocessing.The main columns of the dataset are described below:

TABLE I
DESCRIPTION OF DATASET COLUMNS USED FOR SSFR PREDICTION

Column	Format	Description
dr9objID	I19	Unique identifier for objects in the SDSS-DR9 catalog, used for cross-referencing.
objID	I18	Object identifier in SDSS-DR7, from which the photometric data was sourced.
RAdeg	E10.6 deg	Right Ascension (J2000) used to compute the celestial position of galaxies.
DEdeg	E10.6 deg	Declination (J2000), part of the celestial position feature engineering process.
photoz	E8.6	Photometric redshift, a key feature used to estimate distances to galaxies and calculate SSFR.
Qual	I1	Quality flag for redshift measurement accuracy: 1 = high accuracy, 2 = medium accuracy, 3 = low accuracy, 0 = no flag (used for feature encoding).
SSFR	F8.4 [yr^{-1}]	Specific Star Formation Rate (target variable), representing the rate of star formation relative to the galaxy's mass.

B. Data Pre-Processing

Preprocessing the data was a crucial step to ensure that the machine learning models could perform optimally. The following preprocessing steps were applied:

1) **Handling Missing Values**: Although the dataset was free of missing values, a robust strategy was prepared in case future data might have gaps. This strategy involved median or mean imputation for missing values. As there were no missing data, this step was not required but remains a core part of the preprocessing pipeline for reproducibility.
2) **Normalisation and Scaling**: Since machine learning models often perform better when numerical features are on a similar scale, Min-Max scaling was applied to features such as RAdeg, DEdeg, and photoz. This transformation rescaled the values to a range between 0 and 1, which is essential for gradient-based [12] models like Linear Regression [13], Ridge, and Lasso. Standardisation was also tested for some models that benefit from normally distributed data.
3) **One-Hot Encoding**: The Qual feature, which represented the quality of observations, was categorical. To handle this, one-hot encoding was applied, creating binary columns that allowed machine learning models to interpret the quality flags as features.

Data was then split into training and testing sets using an 80-20 split, where 838,860 rows were assigned to the training set, and 209,716 rows to the testing set. Stratified sampling ensured that the target variable, SSFR, maintained a balanced distribution in both sets, crucial for minimising prediction bias.

C. Feature Engineering

Feature engineering played a pivotal role in enhancing the dataset's ability to model SSFR accurately. Several key transformations and augmentations were performed:

1) **Celestial Position Engineering**: A new feature, celestial_position, was generated by combining the right ascension (RAdeg) and declination (DEdeg) values. This transformation helped capture the spatial positioning of galaxies, allowing models to infer spatial relationships linked to star formation rates.
2) **Polynomial Features and Interaction Terms**: Interaction terms between photometric redshift (photoz) and the celestial coordinates were created to capture non-linear relationships between these variables and SSFR. Polynomial expansion of certain features was also applied, improving the models' ability to capture complex dependencies.
3) **One-Hot Encoding for Categorical Variables**: The one-hot encoding of Qual not only transformed this categorical variable into binary features but also allowed us to incorporate quality flags in the models, which is important in observational astronomy where data quality can significantly affect predictions.
4) **Synthetic Data Generation**: To further enhance the dataset, synthetic data was generated using Gaussian Copula models. This method preserved the correlations between the original features while creating new samples. The synthetic data provided additional training samples, particularly for underrepresented regions in the feature space, helping the models generalise better on unseen data.

D. k-Fold Cross-Validation

To ensure robust model evaluation and mitigate potential bias or overfitting during the training process, K-Fold Cross-Validation was employed. This technique divides the dataset into *k* equally sized subsets, referred to as "folds." The model is trained on *k-1* folds while one fold is reserved for validation. This process is repeated *k* times, with each fold serving as the validation set once. The final evaluation metric is averaged over all *k* iterations, providing a more generalised estimate of model performance.

In this study, k=5 was selected to achieve a balance between computational efficiency and the need for

a sufficiently diverse set of training-validation splits. A lower value of *k* (e.g., 3) might yield less stable estimates of the model's performance, while a higher value (e.g., 10) would increase computational costs without offering substantial improvements in results. Given the relatively large size of the dataset, partitioning it into five folds allowed for adequate representation in both training and validation sets.

To enhance the model's exposure to diverse samples, Stratified K-Fold Cross-Validation was implemented. This method preserves the distribution of the target variable across the folds, which is particularly crucial in datasets with potential imbalances in specific regions of the parameter space (e.g., certain ranges of SSFR).

The cross-validation framework was integrated into the analysis using Scikit-learn's `StratifiedKFold` function. Model performance metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2, were computed across all folds. The averaged metrics provided a more reliable indication of each model's generalisation capability prior to final testing.

E. Model Selection and Saving the Best Model

Several machine learning models were trained using both the original and synthetic datasets. The models included linear regression, ridge, lasso, random forest, and gradient boosting. These models were chosen for their robustness in handling non-linear data, ability to capture complex patterns, and scalability. Each model was trained using 5-fold cross-validation to ensure reliability, and hyperparameter tuning was performed using grid search

For hyperparameter optimisation, parameters such as the number of trees, learning rate, and maximum depth were tuned to minimise the Mean Squared Error (MSE) on the validation set. Cross-validation results were averaged to prevent overfitting.

F. Synthetic Data Integration

After generating synthetic data using the Gaussian Copula model, the same machine learning models were retrained on the augmented dataset. This allowed for a direct comparison between the models trained on the original data and those trained on the expanded dataset. The performance differences were analysed, particularly in terms of the accuracy and generalisation capacity of the models. The inclusion of synthetic data improved the models' ability to predict SSFR in less-represented regions, validating the effectiveness of data augmentation in this domain.

IV. Evaluation Metrics

In this study, several evaluation metrics have been used to assess the performance of machine learning models in predicting star formation rates (SFRs) using photometric data. The primary metrics used are Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2. Each metric provides different insights into the accuracy and reliability of the predictions.

TABLE II
Comparison of Original and Synthetic Data Statistics

Feature	Data Type	Mean	Std. Dev.	Min-Max
`RAdeg`	Original	190.12	45.57	[150.12, 230.46]
	Synthetic	190.22	45.68	[150.01, 230.57]
`DEdeg`	Original	-0.88	0.46	[-1.23, -0.12]
	Synthetic	-0.88	0.46	[-1.24, -0.12]
`photoz`	Original	0.46	0.12	[0.10, 0.90]
	Synthetic	0.46	0.12	[0.10, 0.90]
`SSFR`	Original	-9.12	0.57	[-9.57, -8.12]
	Synthetic	-9.13	0.57	[-9.60, -8.10]

A. Mean Absolute Error (MAE)

MAE is a widely used metric in regression tasks, representing the average magnitude of errors between predicted and actual values. It is calculated as:

$$\boxed{\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|} \tag{1}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations. MAE provides an intuitive measure of model accuracy, with lower values indicating better performance. Unlike other metrics, MAE is less sensitive to outliers, making it particularly useful when dealing with noisy data [8].

B. Mean Squared Error (MSE)

MSE measures the average of the squares of the errors, that is, the average squared difference between the predicted values and the actual values. It is calculated as mentioned below:

$$\boxed{\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations. MSE provides a measure of the average squared deviation, with lower values indicating better performance. It is sensitive to outliers due to the squaring of errors.

C. Root Mean Squared Error (RMSE)

RMSE is the square root of MSE and provides an error measure in the same units as the target variable. It is calculated as:

$$\boxed{\text{RMSE} = \sqrt{\text{MSE}}} \tag{3}$$

where MSE is the mean squared error. RMSE is useful for understanding the typical magnitude of the error in the predictions, with lower values indicating better performance. Like MSE, RMSE is sensitive to outliers.

D. R^2 (Coefficient of Determination)

R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated as:

$$\boxed{R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (4)$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of the actual values, and n is the number of observations. R^2 provides a measure of how well the model explains the variability of the data, with higher values indicating better performance and a better fit of the model to the data.

V. Implementation and Analysis

A. Synthetic data

To address the limitation of the size of the original dataset and enhance the performance of the SSFR prediction model, synthetic data was generated. The Gaussian Copula method was used to ensure the preservation of key data correlations. Below are the distributions of four essential features: SSFR, RAdeg, DEdeg, and photoz. These distributions for both original and synthetic data demonstrate the high similarity between the two, indicating the fidelity of the synthetic data generation process.

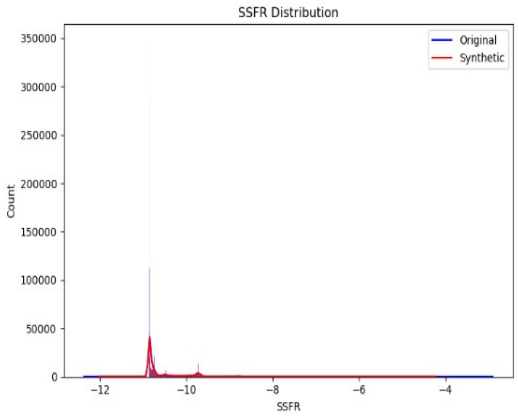

Fig. 2. SSFR Distribution

The distributions reveal how the synthetic data successfully mirrors the characteristics of the original dataset, ensuring valid model training and augmenting the available data for improved predictions.

Additionally, the celestial position feature was engineered by combining RAdeg and DEdeg into a single coordinate pair, calculated as follows:

$$\boxed{\text{celestial\_position} = \sqrt{\text{RAdeg}^2 + \text{DEdeg}^2}} \quad (5)$$

This feature enabled us to capture the spatial characteristics of galaxies within the data.

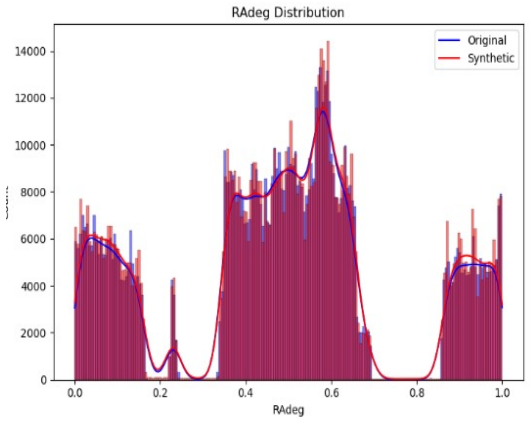

Fig. 3. RAdeg Distribution

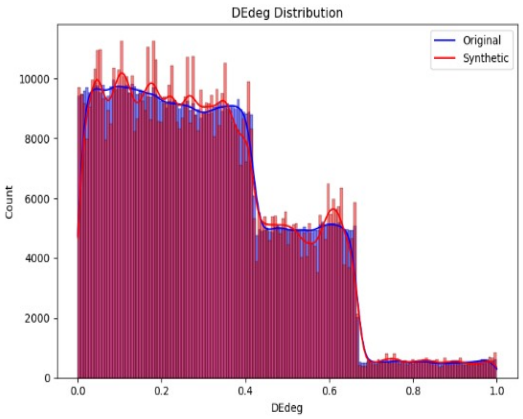

Fig. 4. DEdeg Distribution

B. SSFR

1) SSFR Prediction Metrics: The performance metrics of various regression models on the Specific Star Formation Rate (SSFR) prediction task are summarised in the table below. Each model was evaluated using several metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R^2 Score, Adjusted R^2 Score, and Training Time. The results provide insight into the effectiveness

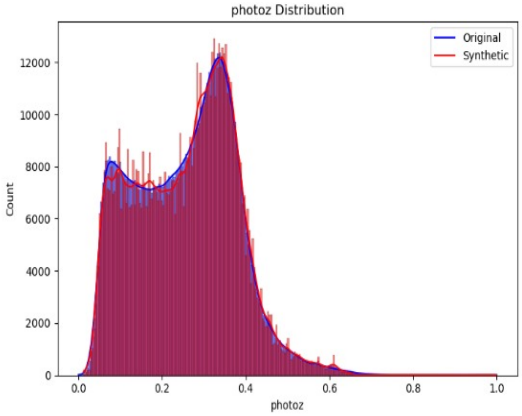

Fig. 5. Photoz Distribution

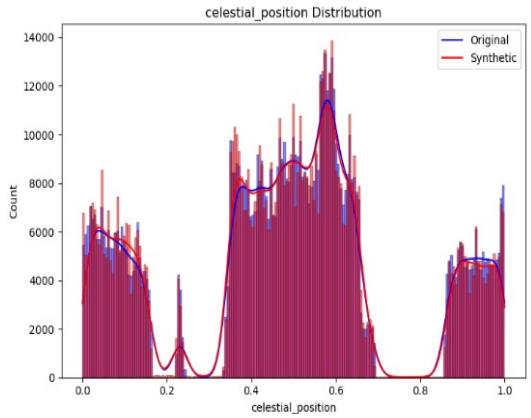

Fig. 6. Celestial Position Distribution

TABLE III
PERFORMANCE METRICS OF VARIOUS REGRESSION MODELS

Model	MSE	RMSE	MAE
Linear Regression	0.19211	0.43831	0.31453
Ridge Regression	0.19212	0.43831	0.31454
Lasso Regression	0.19714	0.44400	0.32390
Random Forest	0.18895	0.43469	0.30946
Gradient Boosting	0.18915	0.43491	0.31063

Model	R^2 Score	Adj R^2 Score
Linear Regression	0.02547	0.02636
Ridge Regression	0.02546	0.02636
Lasso Regression	-1.36E-07	-2.11E-06
Random Forest	0.04150	0.04174
Gradient Boosting	0.04052	0.04158

TABLE IV
MODEL PERFORMANCE METRICS

and efficiency of each model for SSFR prediction.

Among the models evaluated, Random Forest exhibited the lowest MSE (0.18895) and RMSE (0.43469), indicating superior predictive accuracy compared to other methods. It also achieved the lowest MAE (0.30946), reflecting its effectiveness in minimising prediction errors. The Random Forest model, however, has a notably higher training time (415.27 seconds) compared to other models, which may impact its practical utility in scenarios requiring quick model training.

Gradient Boosting showed comparable performance with a slightly higher MSE (0.18915) and RMSE (0.43491) than Random Forest. Despite these marginal differences, it had a higher training time (723.86 seconds), which might limit its applicability in resource-constrained environments.

Ridge Regression and Lasso Regression provided competitive results with MSE values of 0.19212 and 0.19714, respectively. These models have significantly lower training times compared to the ensemble methods, making them suitable for scenarios where computational efficiency is critical. Ridge Regression demonstrated marginally better performance than Lasso Regression, particularly in terms of R^2 Score and Adjusted R^2 Score.

2) SSFR Spatial Distribution: In this analysis, the Specific Star Formation Rate (SSFR) predictions were plotted spatially. A heatmap was generated to identify regions in the sky where SSFR values were concentrated, providing astronomers with insights into galaxy formation rates. This visualisation also illustrates how galaxies in specific regions tend to have varying star formation rates, which can be crucial for satellite observation planning.

C. Regions of interest

Using DBSCAN clustering, regions of interest were identified based on SSFR predictions and their gradients. These regions are of particular interest for satellite observation as they represent areas of high star formation activity. The top 10 clusters identified in the analysis are shown below:

TABLE V
SUMMARY OF CLUSTERS WITH PREDICTED SSFR AND RELATED METRICS

Cluster	Pred SSFR	SSFR Grad	RAdeg	DEdeg	Photoz
11	-10.49	44.64	0.14	0.38	0.52
24	-10.49	21.98	0.37	0.50	0.52
22	-10.49	19.92	0.98	0.35	0.52
23	-10.49	19.73	0.97	0.04	0.52
34	-10.49	18.52	0.92	0.04	0.53
43	-10.52	17.99	0.23	0.10	0.44
38	-10.49	17.91	0.99	0.30	0.53
16	-10.49	16.74	0.91	0.29	0.52
6	-10.49	16.48	0.51	0.21	0.53
30	-10.49	13.68	0.97	0.26	0.53

Where Pred SSFR is the Predicted SSFR and SSFR Grad is the SSFR Graident

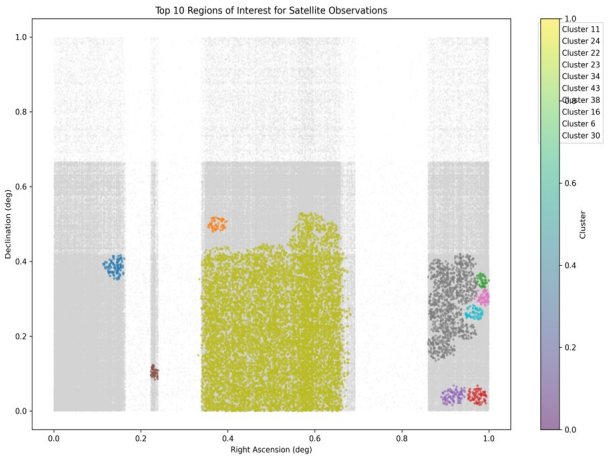

Fig. 7. Top 10 Regions of interest for Satellite Observations

These regions were selected based on their SSFR gradient, a metric that highlights areas with rapid changes in star formation rates. These regions are optimal targets for further astronomical study and satellite observation.

D. Interactive Visualisation tool

An interactive tool was developed to allow users to explore SSFR predictions and regions of interest visually. This tool offers dynamic zoom, panning capabilities, and feature overlays to assist researchers in examining areas of the sky with high star formation activity.

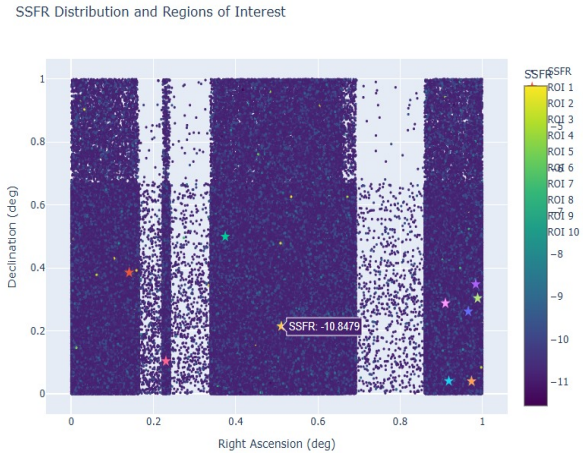

Fig. 8. Screenshot of the interactive tool developed

E. Summary

This research investigates the use of synthetic data to enhance the prediction of Specific Star Formation Rate (SSFR) in galaxies using the Sloan Digital Sky Survey Data Release 7 (SDSS-DR7). The project employs advanced machine learning models, including Linear Regression, Ridge Regression, Lasso Regression, Random Forest, and Gradient Boosting, to analyse the photometric data of galaxies. The study integrates synthetic data generated through Gaussian Copula and evaluates the performance of various models using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). Significant insights were obtained through the implementation of K-Fold Cross Validation and DBSCAN clustering, identifying key regions of interest for potential astronomical observations.

The analysis of synthetic data and model performance demonstrates that models trained with synthetic data can effectively supplement real datasets, improving the robustness and accuracy of SSFR predictions. The project further explores spatial distribution through heatmaps and identifies the top 10 regions of interest for future observations. This comprehensive approach highlights the effectiveness of synthetic data in bridging gaps in observational data and advancing astronomical research methodologies.

VI. Conclusion

The integration of synthetic data in the prediction of SSFR has proven to be a valuable strategy, enhancing the predictive capabilities of machine learning models. By augmenting real-world datasets with synthetic data, the research has successfully improved model performance and identified significant patterns and regions of interest. The use of advanced clustering techniques and spatial analysis has provided deeper insights into the spatial distribution of galaxies, which can inform future astronomical research and observational strategies.

Overall, this study underscores the potential of synthetic data in overcoming limitations of real datasets and contributing to more accurate and comprehensive analyses in astronomy. Future work could focus on refining synthetic data generation techniques and expanding the analysis to include additional variables and datasets, further advancing the understanding of galaxy formation and evolution.

VII. Future Scope

Future research could build on this study by exploring more sophisticated synthetic data generation methods, such as incorporating generative adversarial networks (GANs) or variational autoencoders (VAEs) to produce even more realistic and diverse datasets. Expanding the analysis to include additional astronomical variables and leveraging multi-wavelength observations could provide a more comprehensive view of galaxy formation and evolution. Integrating these advanced synthetic data techniques with emerging deep learning architectures and larger datasets could enhance model performance and uncover new insights into cosmic phenomena. Additionally, developing and testing these methods on other astronomical surveys and datasets could validate their effectiveness and applicability across various domains of astrophysical research.

References

[1] T. Ramesh, A. Shashikanth *et al.*, "A machine learning based ensemble approach for predictive analysis of healthcare data," in *2020 2nd PhD Colloquium on Ethically Driven Innovation and Technology for Society (PhD EDITS)*. IEEE, 2020, pp. 1–2.

[2] L. Y. NA, G. Bharathraj, S. P. AS, S. Shreyas, and C. Rajesh, "Generating synthetic dataset for cotton leaf using dcgan," in *2023 4th International Conference On Signal Processing And Communication (ICSPC)*. IEEE, 2023, pp. 249–252.

[3] H. S. Behl, A. G. Baydin, and P. H. Torr, "Alpha maml: Adaptive model-agnostic meta-learning," *arXiv preprint arXiv:1905.07435*, 2019.

[4] J. Dahmen and D. Cook, "Synsys: A synthetic data generation system for healthcare applications," *Sensors*, vol. 19, no. 5, p. 1181, 2019.

[5] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[6] A. Figueira and B. Vaz, "Survey on synthetic data generation, evaluation methods and gans," *Mathematics*, vol. 10, no. 15, p. 2733, 2022.

[7] M. Delli Veneri, S. Cavuoti, M. Brescia, G. Longo, and G. Riccio, "Star formation rates for photometric samples of galaxies using machine learning methods," *Monthly Notices of the Royal Astronomical Society*, vol. 486, no. 1, p. 1377–1391, Mar. 2019. [Online]. Available: http://dx.doi.org/10.1093/mnras/stz856

[8] G. Aufort, L. Ciesla, P. Pudlo, and V. Buat, "Constraining the recent star formation history of galaxies: an approximate bayesian computation approach," *AA*, vol. 635, p. A136, 2020. [Online]. Available: https://doi.org/10.1051/0004-6361/201936788

[9] G. E. Magdis, E. Daddi, M. Béthermin, M. Sargent, D. Elbaz, M. Pannella, M. Dickinson, H. Dannerbauer, E. da Cunha, F. Walter, D. Rigopoulou, V. Charmandaris, H. S. Hwang, and J. Kartaltepe, "The Evolving Interstellar Medium of Star-forming Galaxies since z = 2 as Probed by Their Infrared Spectral Energy Distributions," , vol. 760, no. 1, p. 6, Nov. 2012.

[10] C. C. Lovell, V. Acquaviva, P. A. Thomas, K. G. Iyer, E. Gawiser, and S. M. Wilkins, "Learning the relationship between galaxies spectra and their star formation histories using convolutional neural networks and cosmological simulations," *Monthly Notices of the Royal Astronomical Society*, vol. 490, no. 4, p. 5503–5520, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1093/mnras/stz2851

[11] V. Bonjean, N. Aghanim, P. Salomé, A. Beelen, M. Douspis, and E. Soubrié, "Star formation rates and stellar masses from machine learning," *Astronomy amp; Astrophysics*, vol. 622, p. A137, Feb. 2019. [Online]. Available: http://dx.doi.org/10.1051/0004-6361/201833972

[12] V. Daliya and T. Ramesh, "A parameter tuned ensemble model for accurate prediction of diabetic progression," in *2021 IEEE 3rd PhD Colloquium on Ethically Driven Innovation and Technology for Society (PhD EDITS)*. IEEE, 2021, pp. 1–2.

[13] D. VK and T. Ramesh, "Optimized stacking ensemble models for the prediction of diabetic progression," *Multimedia Tools and Applications*, vol. 82, no. 27, pp. 42 901–42 925, 2023.