Seasonal Pattern And Solar Irradiance Prediction Using Machine Learning: A GHI Based Analysis For Bangladesh

Sumaya Islam1, a), Sk. Shahed Ali1, b), M R Wasik Ahmed Apon1, c), MD Abdul Aziz1, d), Jubaer Bin Murshed1, e), Md Saef Ullah Miah1, f), Md. Jakir Hossen2, g)

1*Department of Computer Science, American International University-Bangladesh, Dhaka, Bangladesh*

2*Faculty of Engineering and Technology (FET), Multimedia University, 75450, Melaka, Malaysia*

*g) Corresponding author:* [*jakir*](mailto:g)jakir.hossen@mmu.edu.my)[*.hossen@mmu.edu.my*](mailto:.hossen@mmu.edu.my)

*a)* [*sumayaislam688@gmail.com*](mailto:sumayaislam688@gmail.com)

*b) mkbadhon500@gmail.com*

*c) aponw*[*asikahmed@gmail.com*](mailto:asikahmed@gmail.com)

*d) aata*[*wsif@gmail.com*](mailto:wsif@gmail.com)

*e)Jbmha[wlader@gmail.com](mailto:wlader@gmail.com)*

*f) md.saefullah@gmail.com*

**Abstract.** This study explores an effective and practical approach for predicting Global Horizontal Irradiance (GHI) using advanced machine learning techniques. By focusing on feature selection and model evaluation, we aimed to enhance the accuracy and efficiency of GHI predictions. Through Random Forest feature importance analysis, we identified five key predictors: Clearsky GHI, DNI, DHI, Cloud Type, and Solar Zenith Angle as the most significant factors influencing GHI. XGBoost and LightGBM achieved impressive performance, with RMSE values of 0.0046 and 0.0047, MAE values of 0.0024 and 0.0023, and *R*2 scores of 0.9997, using all 19 features. Even with the top five features, similar results were maintained. Bandarban, Jhenaidah, and Cox’s Bazar were identified as the most promising locations for solar energy in Bangladesh. Seasonal analysis indicated that spring and summer are optimal periods for energy production. This study demonstrates the effectiveness of combining feature selection with machine learning for accurate, scalable GHI prediction, supporting efficient solar energy planning.

# INTRODUCTION

Solar energy increasingly becoming an important part of the global effort to transition to renewable energy sources, and Bangladesh holds an immense potential [1]. Bangladesh is located at 24*◦*N latitude and 90*◦*E longitude, and it receives plenty of sunlight throughout the year, making it an ideal location for solar power generation [2]. However, the efficiency of solar energy systems depends on accurately forecasting solar radiation, especially Global Horizontal Irradiance (GHI), which measures the total amount of solar radiation reaching a flat surface, making it a crucial factor in solar power generation [3]. GHI prediction is challenging, as it is affected by various atmospheric factors such as weather conditions, cloud cover, and ozone levels [4]. These make accurate GHI forecasting difficult, particularly in Bangladesh because of its prominent seasonal variations. Bangladesh has set an ambitious goal to derive 20% of its total energy from renewable sources by 2030. However, progress has been slow—renewable energy contributed just 3.15% of total consumption in 2018 and slightly rising to 3.25% in 2019, according to SREDA [5]. This study enhances GHI prediction in Bangladesh by applying machine learning models with carefully selected meteorological features. Instead of relying on a large set of features, we focus on the most relevant ones to improve efficiency while maintaining high accuracy. Experimental results using two years of data (2019 - 2020) confirm that the selected five features provide nearly identical accuracy compared to using all 19 features, validating the effectiveness of feature selection in reducing model complexity. The key contributions of this study are:

* Developing a feature selection strategy that enhances prediction efficiency without compromising accuracy.
* Evaluating multiple machine learning models and demonstrating that ensemble methods, particularly XGBoost and LightGBM, achieve superior performance.
* Conducting a geographical and seasonal analysis to identify optimal locations and peak periods for solar energy generation in Bangladesh.

The significance of this study lies in its targeted feature selection approach and its practical insights into solar energy planning, which can help optimize photovoltaic deployment and resource allocation. The rest of this paper is organized as follows: the Literature Review presents relevant prior work, the Methodology section outlines the data processing and modeling approach, the Results section discusses key findings, and the Conclusion summarizes contributions and suggests future research directions.

# LITERATURE REVIEW

Hiremath et al. [6] aimed to prognosticate GHI by using colorful environmental factors and several Machine Learning algorithms. In their study, they used eight different algorithms such as Linear Regression, Decision Tree, and RNN, among others. Out of all eight of their model, the Complex-Valued Neural Network(CVNN) has the smallest MSE and RMSE, at 58.7 and 7.6, independently. For the unborn compass, to ameliorate the RMSE values, more sophisticated deep learning models and larger data sets can be enforced.

Richardson et al. [7] described details of the development of a low-cost all-sky imaging system and an intra-hour cloud motion prediction methodology that produces minutes-ahead irradiance forecasts. For their study, their software was written in Python 2.7 and employed the open-source computer vision package OpenCV. By far, the most expensive part of the SkyImager was the security camera enclosure, but this approach was used to save time and effort. Their exploration delved into retaining grayscale images of the red-blue ratio throughout the entire process. More generally, a totally data-driven deep learning approach may prove superior.

Chodakowska et al. [8] presented a study on the application of auto-regressive integrated moving average (ARIMA) models for the seasonal forecasting of solar radiation in different climatic conditions. ARIMA models of two time series (for monthly and hourly data) were erected for two locations, and a forecast was developed. All models showed a very good fit to the data, as measured by the model’s standard error of judgment and the *R*2 coefficient of determination *R*2 *>* 85%.

In this study [9], they stressed the significance of feature engineering in prognosticating solar radiation. The models show a better fit for hourly data for the summer months in the case of both Amman and Warsaw. Despite notable progress, future exploration is demanded in these directions. They compared the model’s performance to other models such as Decision Tree, Linear Regression, and eXtreme Gradient Boosting. Their results showed that the RF model performed better than the other models in terms of MSE, RMSE, MAE, and *R*2. The RF model offers excellent performance in prognosticating solar radiation, with a *R*2 value of 0.96. The findings could have been better if further meteorological variables data and deep-learning methods were included.

Hayajneh et al. [10] explored an innovative operation of tiny machine learning to give real-time, low-cost fore- casting of solar energy yield on resource-constrained edge Internet of Things devices, such as micro-controllers, for better domestic and artificial energy management. They have conducted a comprehensive evaluation of four machine learning models named LSTM, BiRNN, BiGRU, and BiLSTM. Despite the challenges posed by vanishing gradients, the BiRNN model’s performance stands out but the BiGRU model serves as a balanced alternative and can generally enhance the performance across all architectures. Their study can be improved by including other ML or DL models and innovative hybrid architectures for time series prediction to enhance the accuracy of solar energy yield prediction.

Sharma and Whig [11] aimed to produce robust prediction models for three key solar irradiance parameters. In their study, they used a dataset of ten years at a 30-minute interval, encompassing various meteorological variables. The system had resulted in a 12% improvement in load forecasting accuracy and their developed model achieved an accuracy of 92.5%. Their study can be improved by using deep learning and expanding the evaluation to various regions with diverse climatic conditions.

These studies demonstrate significant progress in solar energy forecasting but punctuate the need for larger datasets, improved model accuracy, and advanced machine learning approaches. Our research bridges these gaps by exploring multiple machine-learning models, analyzing two years of GHI data from Bangladesh, and optimizing prediction accuracy for seasonal patterns.

# METHODOLOGY

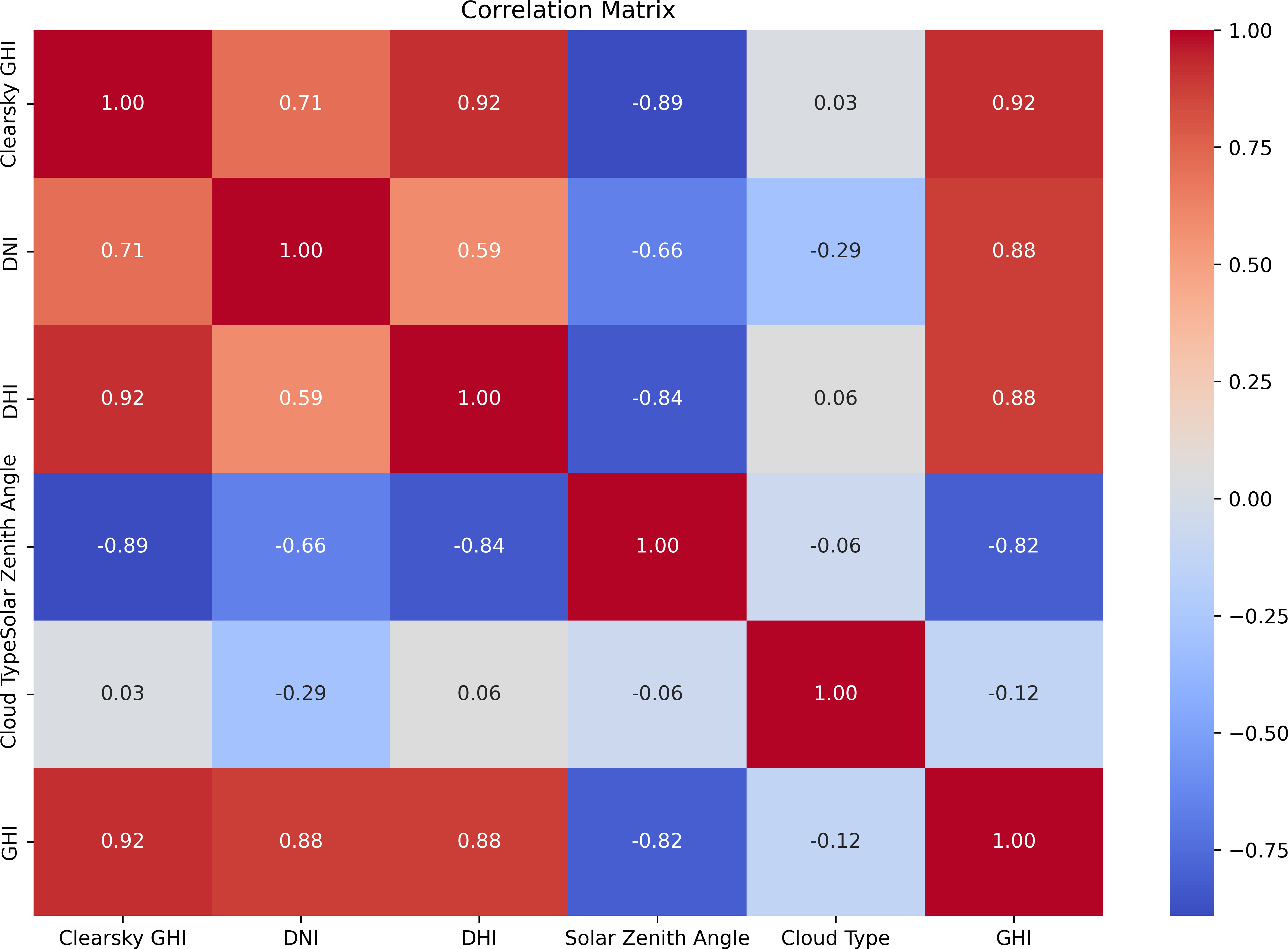
**Dataset**

This study uses a dataset collected from the National Solar Radiation Database(NSRDB) [12] and it covers 46 cities in Bangladesh for the years of 2019 and 2020. The dataset contains about 807,000 hourly entries, including detailed readings of meteorological conditions and solar radiation. A total of 19 features are included in the dataset: Clearsky DHI, Clearsky DNI, Clearsky GHI, Cloud Type, Fill Flag, Ozone, Solar Zenith Angle, Precipitable Water, Temper- ature, Dew Point, DHI, DNI, GHI, Relative Humidity, Surface Albedo, Pressure, Wind Direction, and Wind Speed. Global Horizontal Irradiance (GHI) is the main variable for this study, which is a key indicator of solar energy po- tential. As the dataset was complete with no missing values, the main preprocessing step involved applying min-max normalization to the continuous variables. This technique scaled continuous values to fall between 0 and 1, helping ensure that no single feature dominated the learning process due to differences in units or scale. After normalization, the dataset was then used for subsequent machine-learning analyses.

## Feature Selection

The Random Forest algorithm was implemented to identify the most influential predictors for Global Horizontal Irradiance (GHI). By ranking the importance of all 19 features in the dataset, the top five contributors were selected for further analysis. Clearsky GHI, DNI, DHI, Cloud Type, and Solar Zenith Angle. These features demonstrated the highest potential for accurately predicting GHI, as determined by their importance scores derived from the algorithm.

To further examine the relationship between the top five features and their association with GHI, a correlation matrix was generated. The matrix presented in Figure 1, reveals strong positive correlations among solar irradiance-related features: Clearsky GHI exhibits a correlation of 0.92 with GHI, while DNI and DHI are both highly correlated with GHI (r = 0.88).



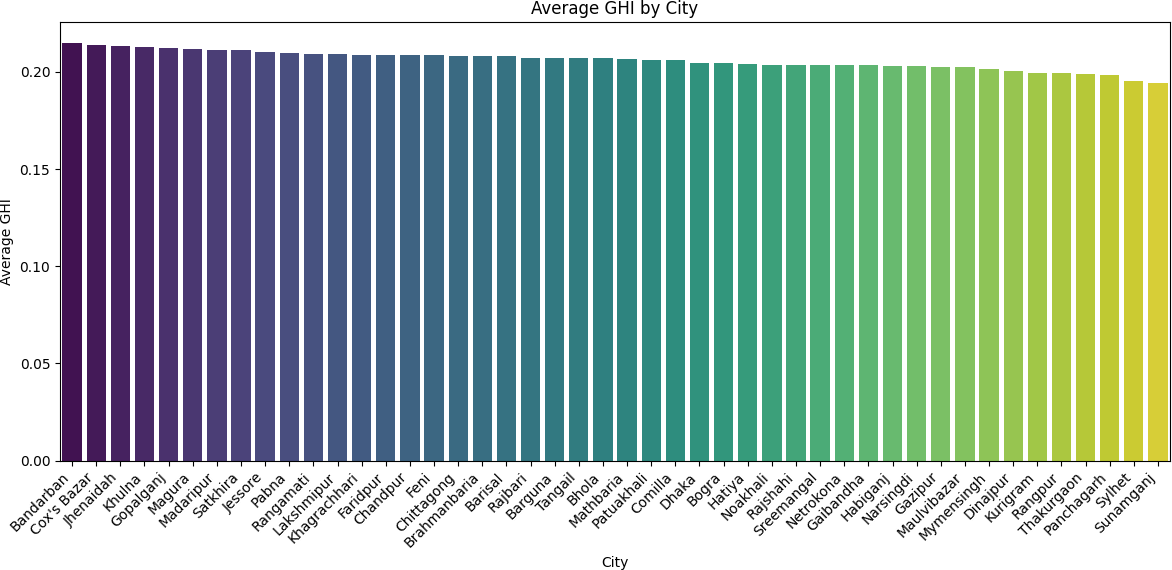
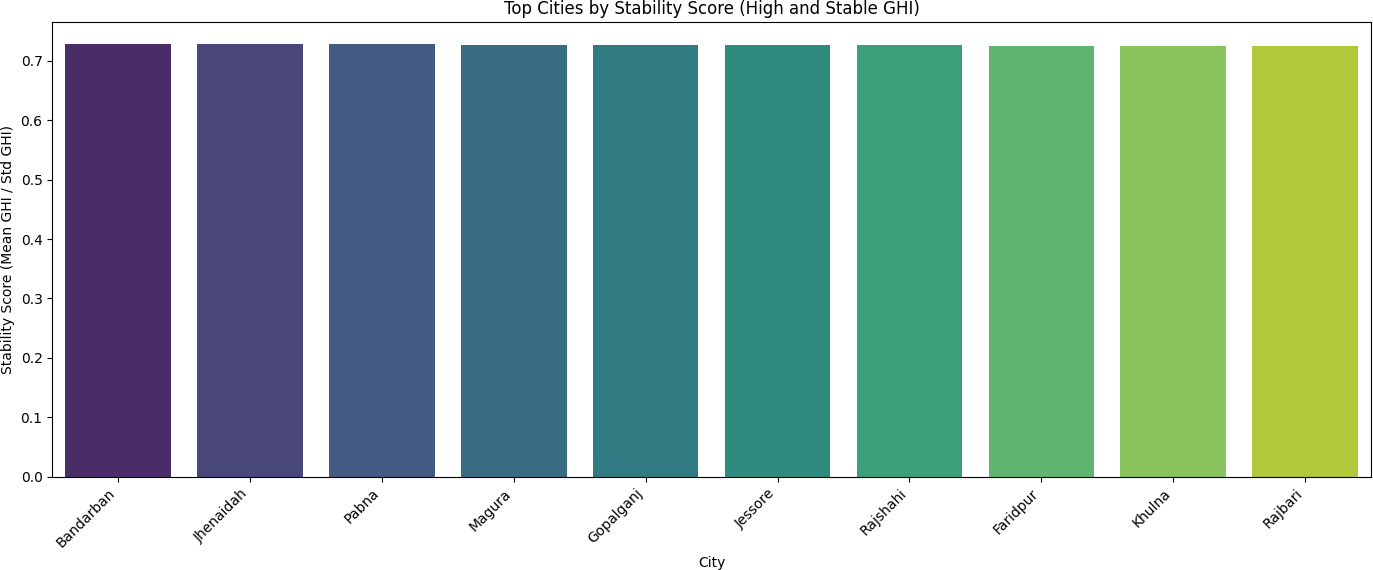
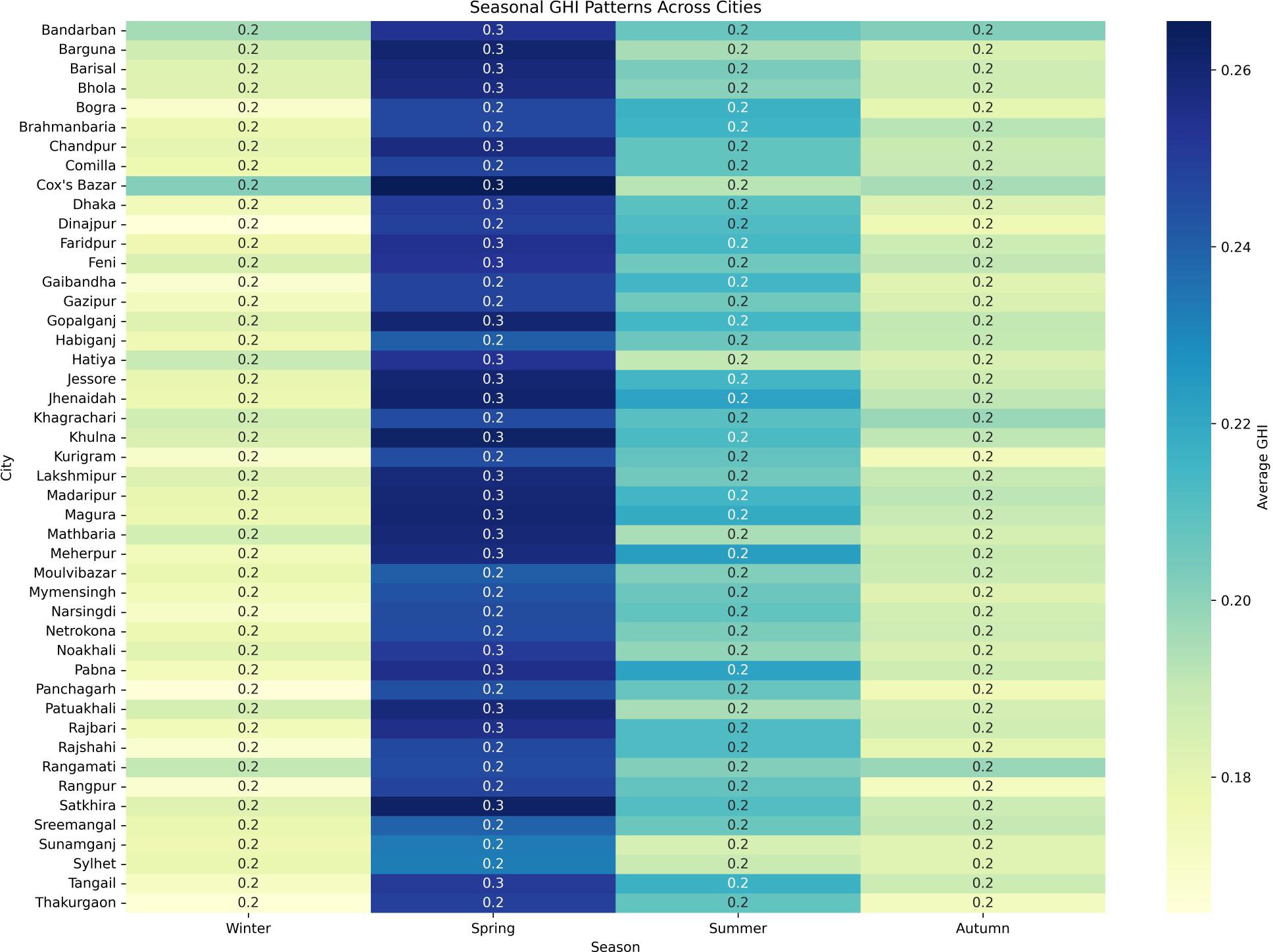
**FIGURE 1.** Correlation matrix of the top five features

# Exploratory Data Analysis

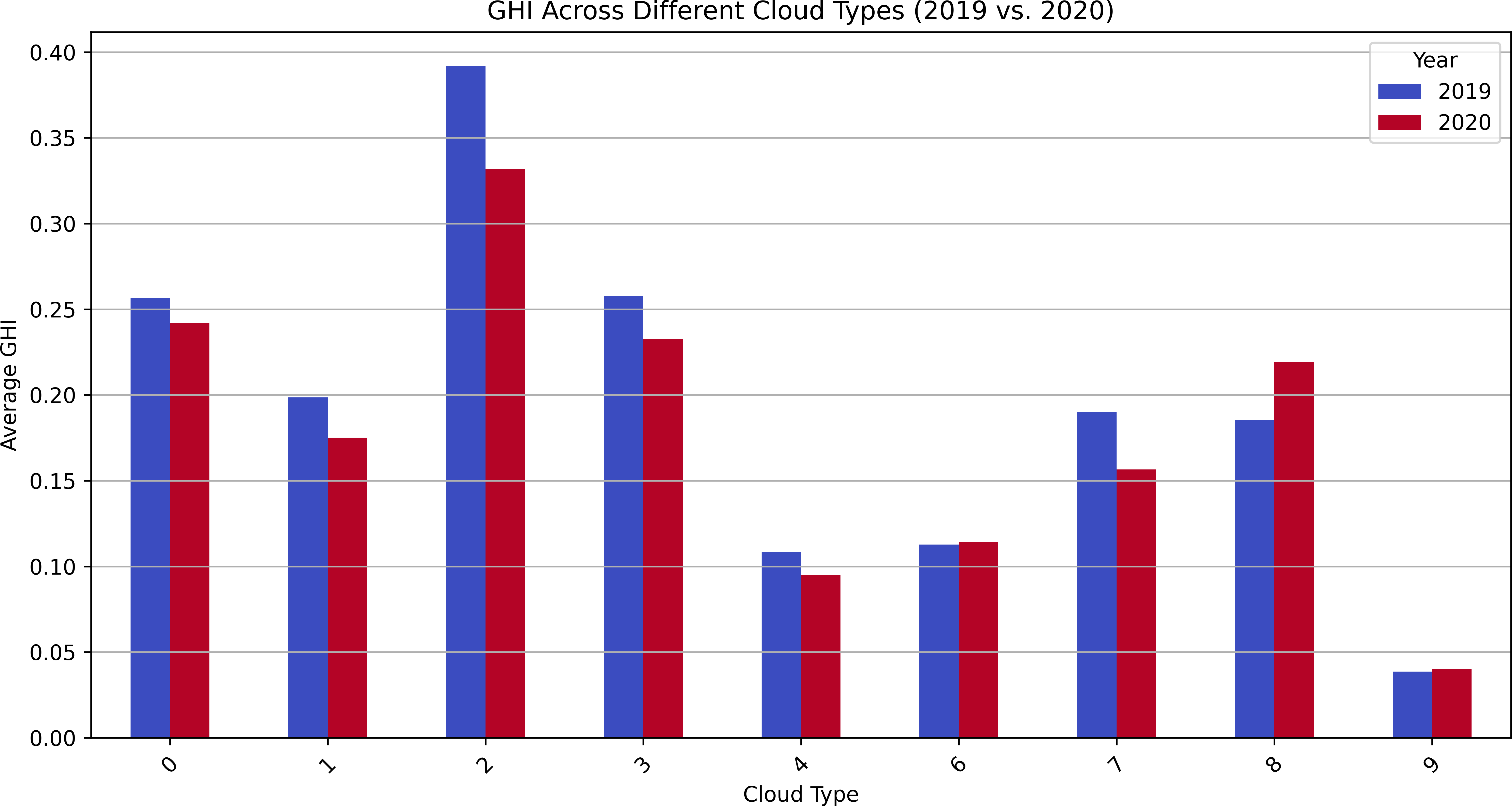
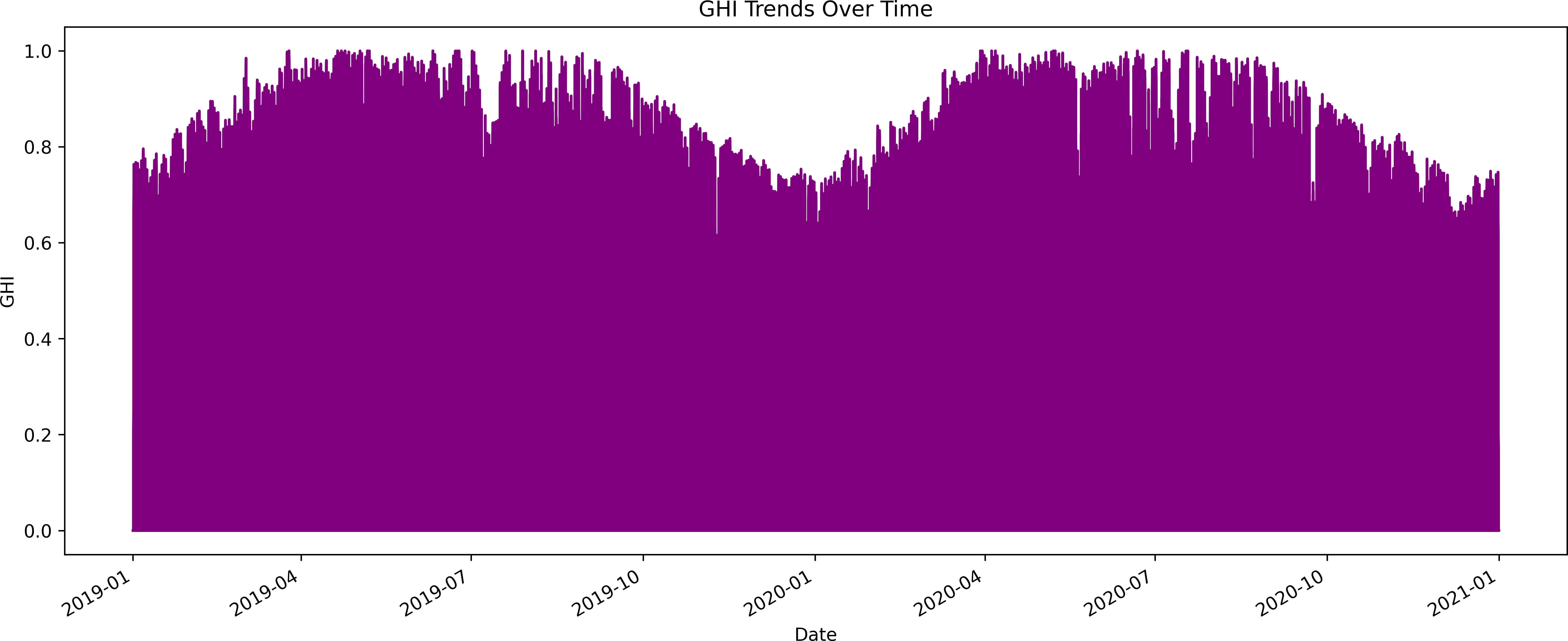
A variety of data visualization methods were utilized to explore the spatial and temporal variations of Global Hori- zontal Irradiance (GHI) across 46 cities in Bangladesh. These visualizations provided essential insights into city-wise differences, temporal stability, and seasonal patterns, aiding the identification of regions with solar energy potential.

Figure 2 presents a comprehensive visual analysis of Global Horizontal Irradiance (GHI) across cities in Bangladesh, highlighting spatial, temporal, and atmospheric factors relevant to solar energy planning. Figure 2(a) illustrates the av- erage GHI values for each city during 2019–2020, revealing that regions such as Bandarban, Cox’s Bazar, Jhenaidah, Khulna, and Gopalganj receive the highest solar irradiance, thus offering strong potential for solar energy deploy- ment. Figure 2(b) displays the temporal stability of GHI across cities, with Bandarban, Jhenaidah, Pabna, Magura,

and Gopalganj demonstrating the lowest variability. This consistency in solar irradiance makes them particularly reliable for long-term energy harvesting.

1. Average GHI by city (2019–2020) (b) GHI stability across cities (c) Seasonal GHI heatmap (Spring Peak)



(d) Daily GHI trends (2019–2020) (e) GHI distribution by cloud type

**FIGURE 2**. GHI analysis across Bangladesh: (a) Average GHI by city (2019–2020), highlighting high-potential regions like Bandarban and Cox’s Bazar; (b) GHI stability across cities, where Bandarban, Jhenaidah, and Pabna show consistent irradiance; (c) Seasonal GHI heatmap showing peak irradiance in spring (March–May); (d) Daily GHI trends over 2019–2020 confirming seasonal variation; (e) GHI distribution by cloud type, with Fog (Type 2) showing the highest irradiance and Overshooting clouds (Type 9) the lowest

Figure 2(c) shows a seasonal heatmap of GHI across 46 cities, emphasizing a distinct peak in solar irradiance during the spring months (March to May). This pattern underscores the need to align solar energy generation strategies with seasonal trends for optimal performance. On the bottom row, Figure 2(d) depicts daily GHI trends over the two- year period, reinforcing the earlier observation of seasonal peaks in spring and dips during the winter months, thus confirming the reliability of these temporal patterns for solar planning.

Lastly, Figure 2(e) compares GHI values under different cloud types. Interestingly, Fog (Cloud Type 2) exhibits the highest average GHI, likely due to the diffusion of solar radiation. Clear skies (Type 0) and water-based clouds (Type 3) also support high irradiance, while Overshooting clouds (Type 9) consistently produce the lowest GHI due to their dense and obstructive nature. Together, these insights provide a holistic view of the factors influencing solar energy potential and offer practical guidance for selecting optimal locations and timeframes for solar infrastructure deployment.

## Model Selection and Data Splitting

In this study, five tree-based regression models namely Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost and two linear regression models namely, Linear Regression and Ridge Regression were selected for evalu- ation. These models were chosen based on their effectiveness in capturing both linear and non-linear relationships in datasets.

The dataset was split into training (70%), validation (15%), and testing (15%) subsets using the train\_test\_split function from the Scikit-learn library. The training set was utilized to train the models, while the validation set was used for hyperparameter tuning and overfitting mitigation. Finally, the testing set was reserved for evaluating the model’s generalization performance. A fixed random seed was implemented during the splitting process to ensure reproducibility. The resulting dataset partitions contained 564,900 samples in the training set, 121,050 samples in the validation set, and 121,050 samples in the testing set.

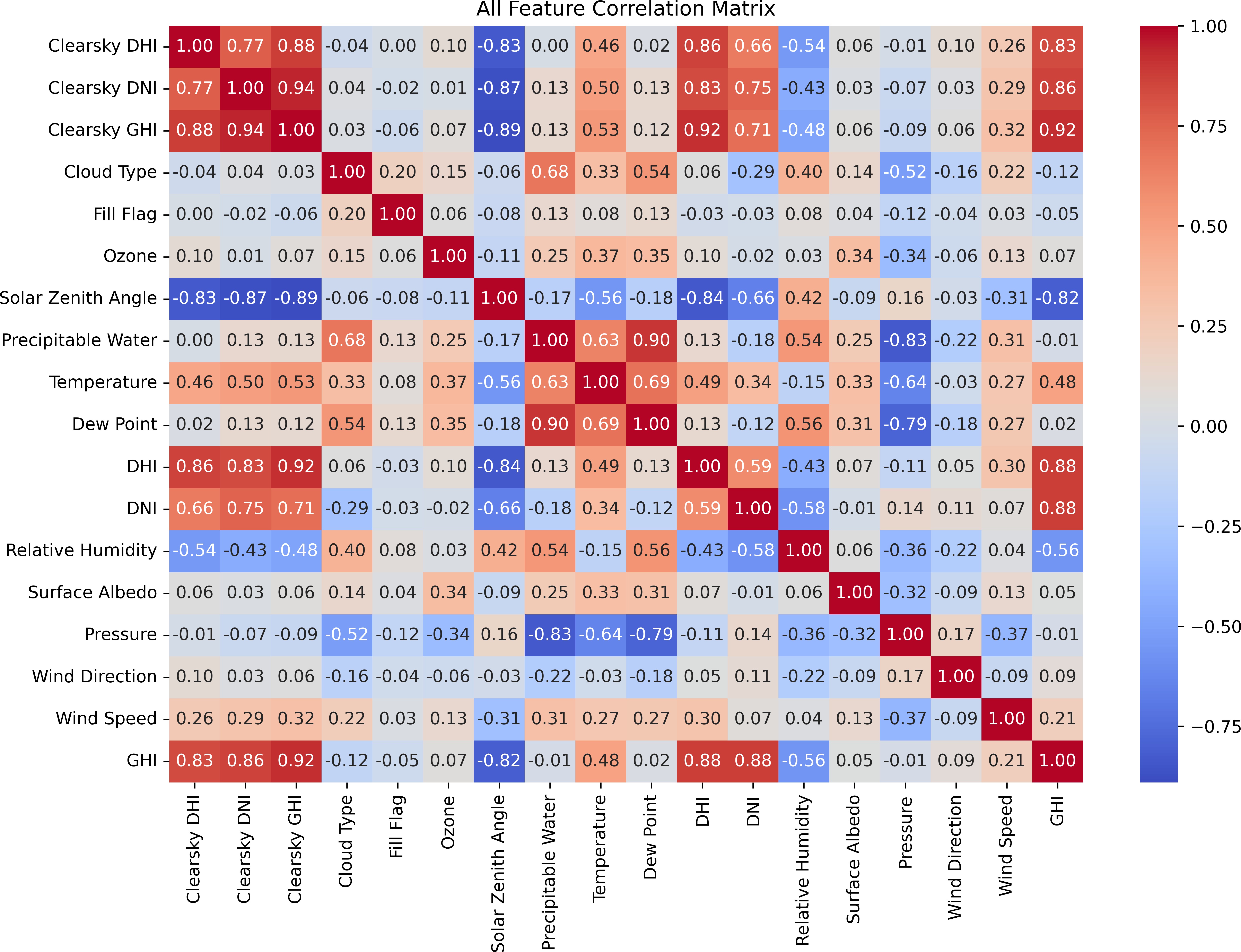
## Performance Metrics

The performance of the regression models was assessed using three key metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination *R*2. These metrics were chosen to evaluate the accuracy, consistency, and explanatory power of the models. These metrics were computed for all models to provide a comprehensive evaluation of their predictive performance. The results of these evaluations are presented in the Results section.

# RESULT AND DISCUSSION

The correlation matrix (Figure 3) revealed that Clearsky GHI, DNI, DHI, Cloud Type, and Solar Zenith Angle had the strongest influence on GHI, with high positive correlations for the first three (up to 0.94) and strong negative cor- relation for Solar Zenith Angle ( 0*.*82). Temperature and Precipitable Water showed moderate positive correlations, while Relative Humidity exhibited a moderate negative correlation. Other features like Wind Speed, Pressure, and Cloud Type showed weaker relationships with GHI.

*−*



**FIGURE 3.** Correlation matrix of all features

Among all the evaluated models, XGBoost and LightGBM consistently demonstrated superior performance. When trained on all 19 features, both models achieved the lowest RMSE (0.0046 and 0.0047) and highest *R*2 scores (0.9997), outperforming traditional methods such as Linear Regression and Ridge Regression, which had significantly higher RMSE values of 0.0307 and lower *R*2 scores of 0.9886. Even when trained with only the top five selected features, both models retained high accuracy—XGBoost and LightGBM achieved RMSEs of 0.0051 and 0.0057 and *R*2 scores of 0.9997 and 0.9996, outperforming all other models. These results are summarized in Table 1.

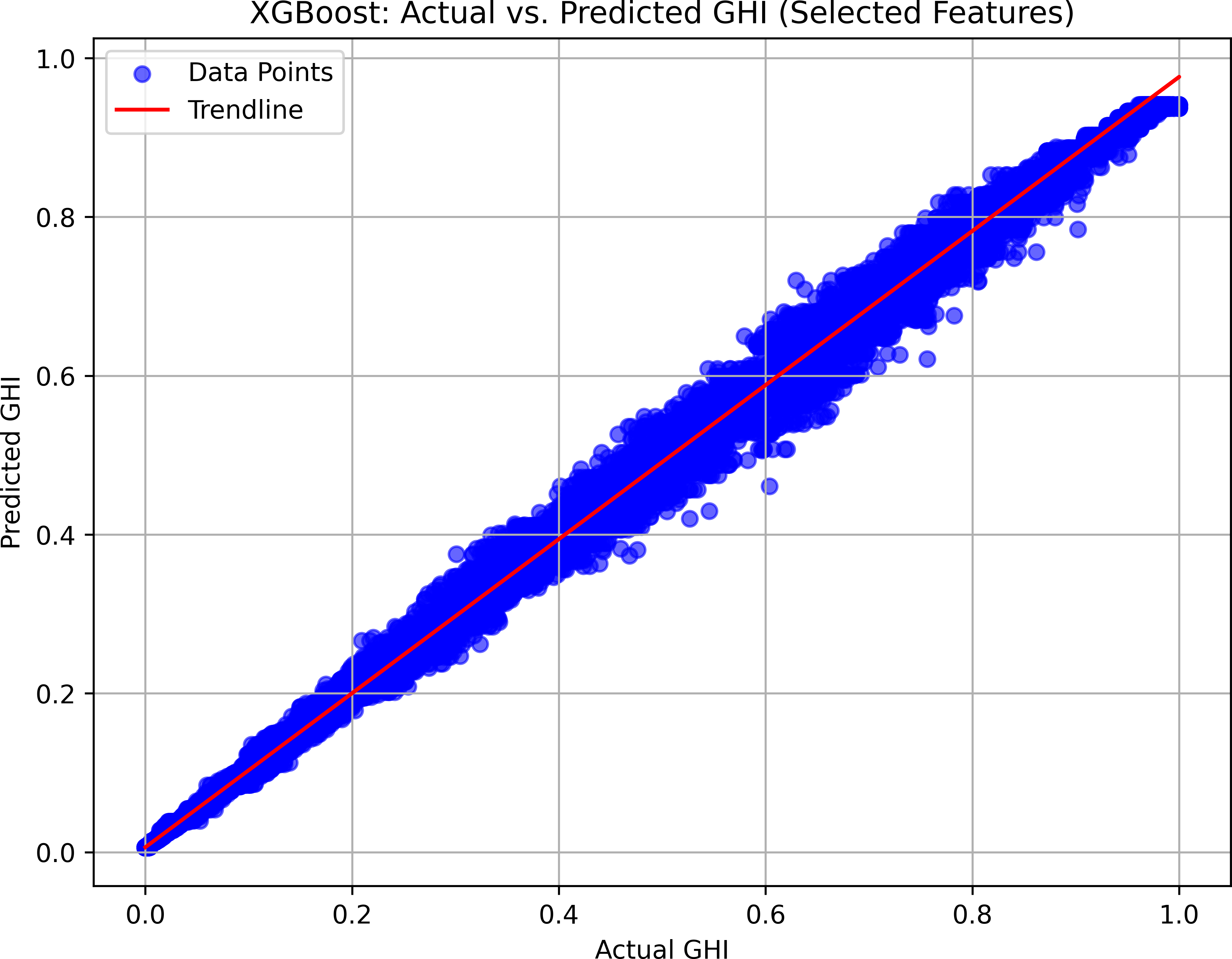
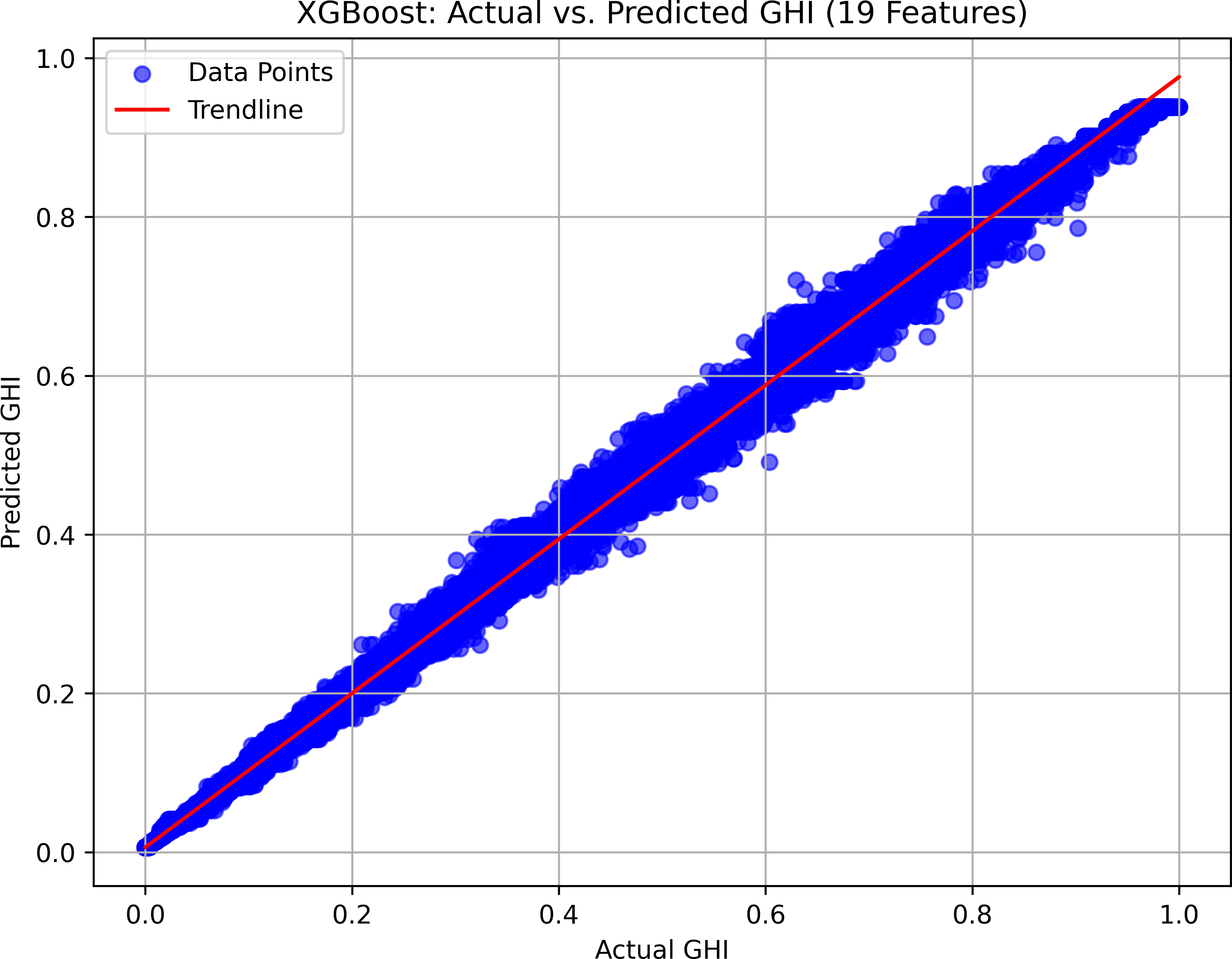
Their superior performance can be attributed to their ability to model complex non-linear relationships, incorporate regularization techniques, and iteratively minimize prediction errors through gradient boosting [13, 14]. Compared to traditional models such as Decision Tree and Linear Regression, ensemble methods like XGBoost and LightGBM demonstrate improved generalization and robustness against overfitting [15]. These findings confirm that ensemble- based gradient boosting models are not only more accurate but also resilient to dimensionality reduction, making them ideal for practical solar irradiance forecasting.

**TABLE 1.** Comparison of model performances using (a) all features and (b) selected features

1. (b)

|  |  |  |  |
| --- | --- | --- | --- |
| Models | **RMSE MAE*****R*2** | Models | **RMSE MAE*****R*2** |
| Decision Tree | 0.0331 0.0163 0.9867 | Decision Tree | 0.0331 0.0163 0.9867 |
| Random Forest | 0.0059 0.0025 0.9996 | Random Forest | 0.0061 0.0026 0.9996 |
| XGBoost | 0.0046 0.0024 0.9997 | XGBoost | 0.0051 0.0026 0.9997 |
| LightGBM | 0.0047 0.0023 0.9997 | LightGBM | 0.0057 0.0027 0.9996 |
| CatBoost | 0.0080 0.0046 0.9992 | CatBoost | 0.0085 0.0047 0.9991 |
| Linear Regression | 0.0307 0.0191 0.9886 | Linear Regression | 0.0469 0.0331 0.9732 |
| Ridge Regression | 0.0307 0.0191 0.9886 | Ridge Regression | 0.0469 0.0331 0.9732 |

Scatter plots in Figure 4 further validated the models’ accuracy, showing a strong alignment between predicted and actual GHI values for both full and reduced feature sets. The minimal drop in performance supports the scalability and computational efficiency of using fewer features, making this approach well-suited for real-time and large-scale applications.



(a) Scatter plot of all features. (b) Scatter plot of selected 5 features.

**FIGURE 4.** Scatter plots showing (a) all features and (b) selected features used for model training

# CONCLUSION

This study highlights the effectiveness of combining feature selection with machine learning for accurate GHI predic- tion. Using Random Forest, five key features—Clearsky GHI, DNI, DHI, Cloud Type, and Solar Zenith Angle—were identified. XGBoost and LightGBM achieved excellent performance, with RMSEs of 0.0046 and 0.0047, and *R*2 scores of 0.9997 using all features. Even with the reduced set, models maintained high accuracy, confirming the selected features’ efficiency. Scatter plots showed strong agreement between predicted and actual GHI values, under- scoring the method’s scalability for real-time applications. Bandarban, Jhenaidah, and Cox’s Bazar emerged as prime solar locations, with spring and summer offering peak solar potential. Future work may explore real-time integration, hybrid models, and long-term forecasting to further enhance reliability and applicability.

# REFERENCES

1. Bangladesh renewable energy sector opportunities,” https://www.trade.gov/market-intelligence/bangladesh-renewable-energy-sector-opportunities, accessed: 2025-1-31.
2. S. Paul, “A report on "solar energy and its potential for bangladesh" august 2020,” (2024).
3. A. A. Imam, A. Abusorrah, M. M. A. Seedahmed, and M. Marzband, “Accurate forecasting of global horizontal irradiance in saudi arabia: A comparative study of machine learning predictive models and feature selection techniques,” Mathematics **12** (2024), 10.3390/math12162600.
4. L. Wu, T. Chen, N. Ciren, D. Wang, H. Meng, M. Li, W. Zhao, J. Luo, X. Hu, S. Jia, L. Liao, Y. Pan, and Y. Wang, “Development of a machine learning forecast model for global horizontal irradiation adapted to tibet based on visible all-sky imaging,” Remote Sensing **15** (2023), 10.3390/rs15092340.
5. “SDG 7: Affordable and clean energy, too ambitious to achieve? — tbsnews.net,” [https://www.tbsnews.net/opini](http://www.tbsnews.net/opinion/)on/ sdg-7-affordable-and-clean-energy-too-ambitious-achieve-41193, [Accessed 05-06-2025].
6. S. V. Hiremath, R. Naik, S. Naik, S. Shettar, and K. Chachadi, “Ghi prediction based on weather data using machine learning,” in *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)* (2024) pp. 1–6.
7. W. Richardson, H. Krishnaswami, R. Vega, and M. Cervantes, “A low cost, edge computing, all-sky imager for cloud tracking and intra-hour irradiance forecasting,” Sustainability **9** (2017), 10.3390/su9040482.
8. E. Chodakowska, J. Nazarko, Nazarko, H. S. Rabayah, R. M. Abendeh, and R. Alawneh, “Arima models in solar radiation forecasting in different geographic locations,” Energies **16** (2023), 10.3390/en16135029.
9. M. A. I. Rafi and M. R. Sohan, “Enhanced solar radiation prediction with machine learning: A comprehensive analysis of meteorological data using random forest,” in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (2024) pp. 1–6.
10. A. M. Hayajneh, F. Alasali, A. Salama, and W. Holderbaum, “Intelligent solar forecasts: Modern machine learning models and tinyml role for improved solar energy yield predictions,” IEEE Access **12**, 10846–10864 (2024).
11. P. Sharma and P. Whig, “Maximizing solar energy utilization through predictive machine learning techniques,” International Journal of Sus- tainable Development Through AI, ML and IoT **2**, 1–13 (2023).
12. National Renewable Energy Laboratory, “National Solar Radiation Database (NSRDB),” (2023), accessed: 2025-01-24.
13. T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016) pp. 785–794.
14. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “Lightgbm: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems*, Vol. 30 (2017) pp. 3146–3154.
15. A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” Frontiers in Neurorobotics **7**, 21 (2013).