Comparative Performance Analysis Of Logistic Regression, Random Forest And LIGHTGBM Classifiers For Predictive Classification Of Regulatory Compliance Of State Space Object Registration Under The United Nations Registration Convention

Bethel-Wali Joy U.[1]
Advanced Space Technology Applications Laboratory Uyo,
National Space Research and Development Agency,
Federal Capital Territory, Abuja, Nigeria

Antiga Bassey Ekpenyong[2]
Advanced Space Technology Applications Laboratory Uyo,
National Space Research and Development Agency,
Federal Capital Territory, Abuja, Nigeria

Uchua Christiana Member[3]
Advanced Space Technology Applications Laboratory Uyo,
National Space Research and Development Agency,
Federal Capital Territory, Abuja, Nigeria

Abstract**—As the number of objects launched into space increases year-on-year, reaching 2,403 satellites in 2024 alone, ensuring state compliance with the 1975 United Nations Registration Convention is critical for space sustainability and traffic management. This study presents a comparative performance analysis of three supervised machine learning classifiers, Logistic Regression (LR), Random Forest (RF), and LightGBM, to predict regulatory compliance based on pre-registration factors. Data was sourced from the United Nations Office for Outer Space Affairs (UNOOSA) Online Index of Objects Launched into Outer Space, covering the period from 2020 to 2024. Following a rigorous audit to prevent data leakage by removing "causally downstream" variables, a final cleaned dataset of 3,302 unique records was generated. Given that approximately 8.5% to 12% of satellites are historically unregistered, the study employed stratified random sampling (80% training, 20% test) to maintain class balance. The implementation evaluates the efficacy of these models in identifying likely non-compliance, addressing a critical gap in international space law where overall adherence has fluctuated between 70% and 90% over the past decade. This research contributes to the "Registration Project" goals by providing a data-driven toolkit for identifying registration gaps and enhancing the transparency of global space activities.**

Keywords—Space Object Registration, UNOOSA Convention, Regulatory Compliance, Machine Learning, LightGBM, Random Forest, Logistic Regression, Predictive Analytics.

1. Introduction

The rapid expansion of the global space industry has led to an unprecedented increase in the number of satellites and orbital debris, underscoring the vital importance of the Convention on Registration of Objects Launched into Outer Space (1975) [1,2]. This international treaty, administered by the United Nations Office for Outer Space Affairs (UNOOSA), serves as the cornerstone for space safety, traffic management, and legal accountability by requiring launching States to provide specific orbital and functional data for a central UN registry [3,4,5]. However, achieving universal compliance remains a challenge. Historical data indicates that a significant proportion of satellites launched into space are unregistered, creating significant gaps in the global tracking of space assets [6,7,8].

Particularly, ensuring regulatory compliance is not merely a legal formality but a critical component of orbital sustainability. Unregistered objects pose heightened risks for collisions and complicate the assignment of liability under international space law [9,10]. Despite the availability of public records through the UNOOSA Online Index of Objects Launched into Outer Space, the process of identifying non-compliant actors has traditionally relied on manual audits and retrospective legal reviews [11,12]. There is a growing need for automated, predictive tools that can identify potential registration lapses based on pre-launch factors and satellite characteristics.

Fortunately, recent advancements in machine learning (ML) offer a robust framework for such predictive classification tasks. By leveraging historical records from 2020 to 2024, this study evaluates the efficacy of three distinct algorithmic approaches; Logistic Regression (LR), Random Forest (RF) and LightGBM [13]. The LR model is a foundational statistical model used for binary classification to establish a baseline for prediction. The RF model is an ensemble learning method that utilizes multiple decision trees to improve accuracy and handle non-linear relationships without overfitting. Also, the LightGBM model is a high-performance, gradient-boosting framework optimized for speed and efficiency in handling large-scale datasets.

This research aims to bridge the gap between international space law and data science. By conducting a comparative performance analysis of these classifiers, the study seeks to determine the most reliable method for predicting the regulatory compliance of State space objects, thereby providing a data-driven path toward enhancing international transparency in outer space.

2. Methodology

This study presents a comparative analysis of three supervised machine learning classifiers; Logistic Regression (LR), Random Forest (RF), and LightGBM, to predict the regulatory compliance of State space object registrations under the United Nations Registration Convention (1975). The methodology encompasses data acquisition from the United Nations Office for Outer Space Affairs (UNOOSA), rigorous preprocessing, and the comparative training of three distinct classification models.

2.1 Data Acquisition and Preprocessing

The primary data for this study was sourced from the UNOOSA Online Index of Objects Launched into Outer Space, covering a five-year period from 2020 to 2024. The Online Index provides access to information provided to the UN in accordance with the Convention on Registration of Objects Launched into Outer Space (Registration Convention). Individual annual records were merged and deduplicated, yielding a final dataset of 3,302 unique space object records. A rigorous audit was conducted to identify and eliminate three columns that were "causally downstream" of the registration outcome, preventing data leakage and ensuring the model predicts compliance based only on pre-registration factors. The cleaned dataset was partitioned into a training set (80%, n = 2,641) and a held-out test set (20%, n = 661). To preserve the minority-class proportion (unregistered objects) in both sets, stratified random sampling was applied with a fixed random seed (**random_state = 42**). Approximately 8.5% to 12% of satellites launched into space are historically unregistered, making stratification critical for balanced evaluation

2.2 Classification Model Architectures

The study compares three models ranging from linear baselines to advanced ensemble methods.

2.2.1 Logistic Regression

Logistic Regression (Figure 1) was selected as the linear baseline for binary prediction. It models the **log-odds** of a registration event as a linear combination of independent variables. This model is well-suited for high-stakes domains requiring transparency and for understanding the baseline linear relationships between state characteristics and compliance.

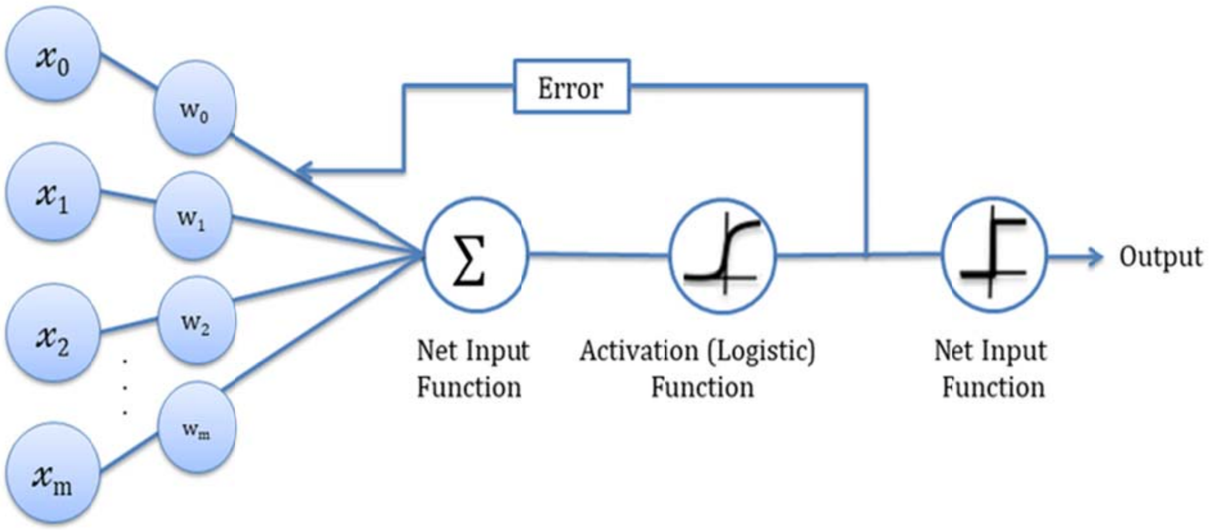

Figure 1 The Logistic Regression model architecture [14]

2.2.2 Random Forest (RF)

The Random Forest model (2) was selected as a non-linear ensemble baseline. It functions by creating a multitude of independently constructed decision trees during training. Generally, RF reduces variance and overfitting by aggregating predictions from multiple trees (bagging). For this classification task, the final output is determined by a **majority vote** across the ensemble. In addition, the RF model provides native feature importance estimates through **mean decrease in impurity**, allowing for the identification of key factors influencing state compliance. Again, the RF model robustness to outliers and ability to model complex interaction effects without explicit feature crosses make it suitable for the heterogeneous feature set used in this study.

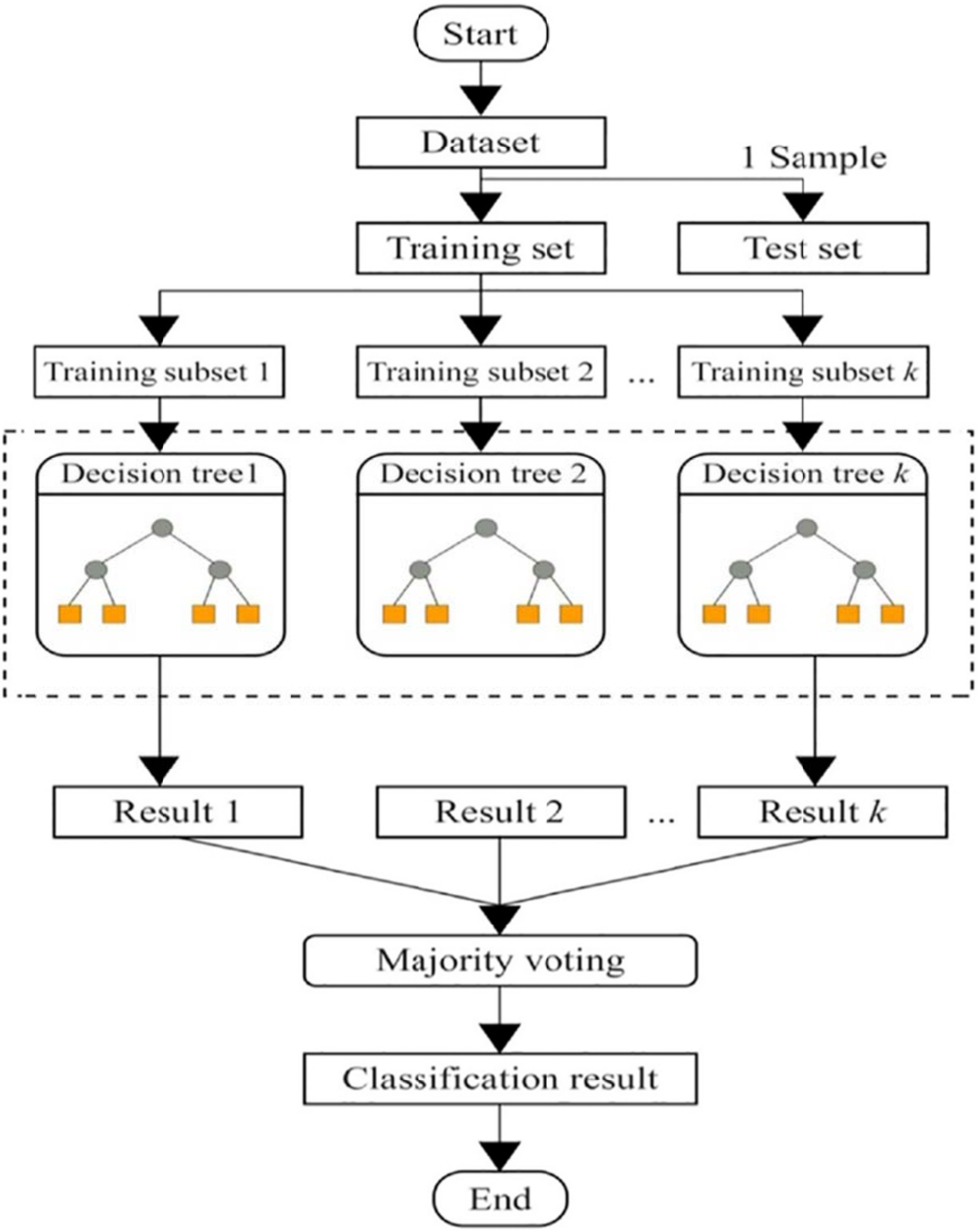

Figure 2 The Architecture of a Typical Random Forest Model [15]

2.2.3 LightGBM (LGBM)

LightGBM (Figure 3) is a gradient-boosting framework that uses tree-based learning algorithms. It is often evaluated alongside Random Forest in comparative studies for its superior efficiency and potential for higher accuracy in complex classification tasks

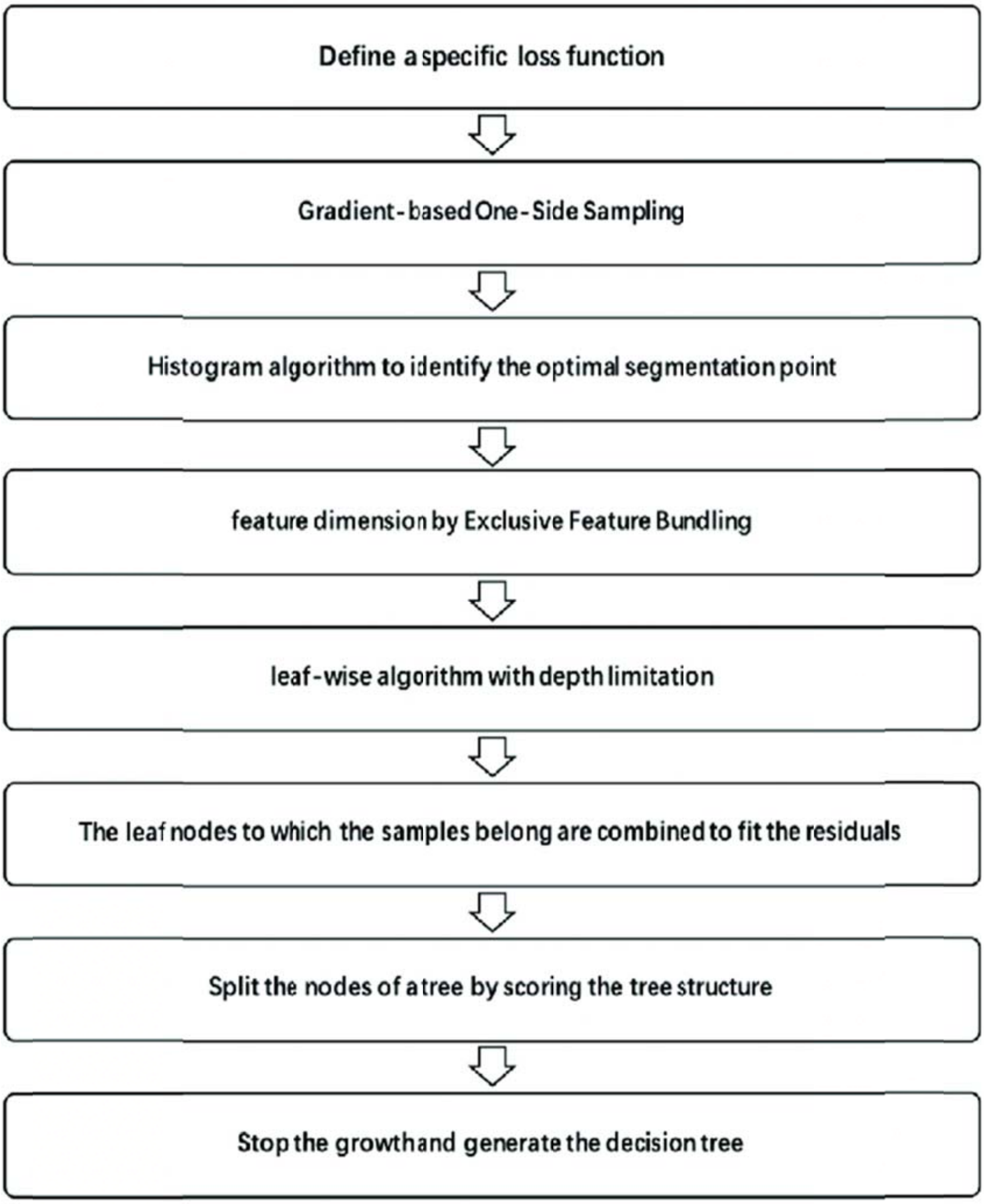

Figure 3 Schematic diagram showing the basic principle of operation of the LightGBM model [16]

2.3 The Implementation of Performance Evaluation of the Three Models

The schematic flow diagram showing the procedure used to train, validate and compare the performance of the three models is shown in Figure 4. The models are assessed using standard classification metrics, as shown in Table 1. Also, the model training and execution time are also evaluated.

Table 1 The Models Performance Metrics

Metric	Description	Formula
Accuracy	The proportion of total predictions that were correct.	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	Of all predicted positives, how many were actually positive? (Reliability)	$\frac{TP}{TP+FP}$
Recall (Sensitivity)	Of all actual positives, how many did the model find? (Completeness)	$\frac{TP}{TP+FN}$
F1-Score	The harmonic mean of Precision and Recall; useful for imbalanced datasets.	$2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$
AUC-ROC	The probability that the model ranks a random positive example higher than a random negative one.	Calculated as the area under the plot of **TPR** versus **FPR**

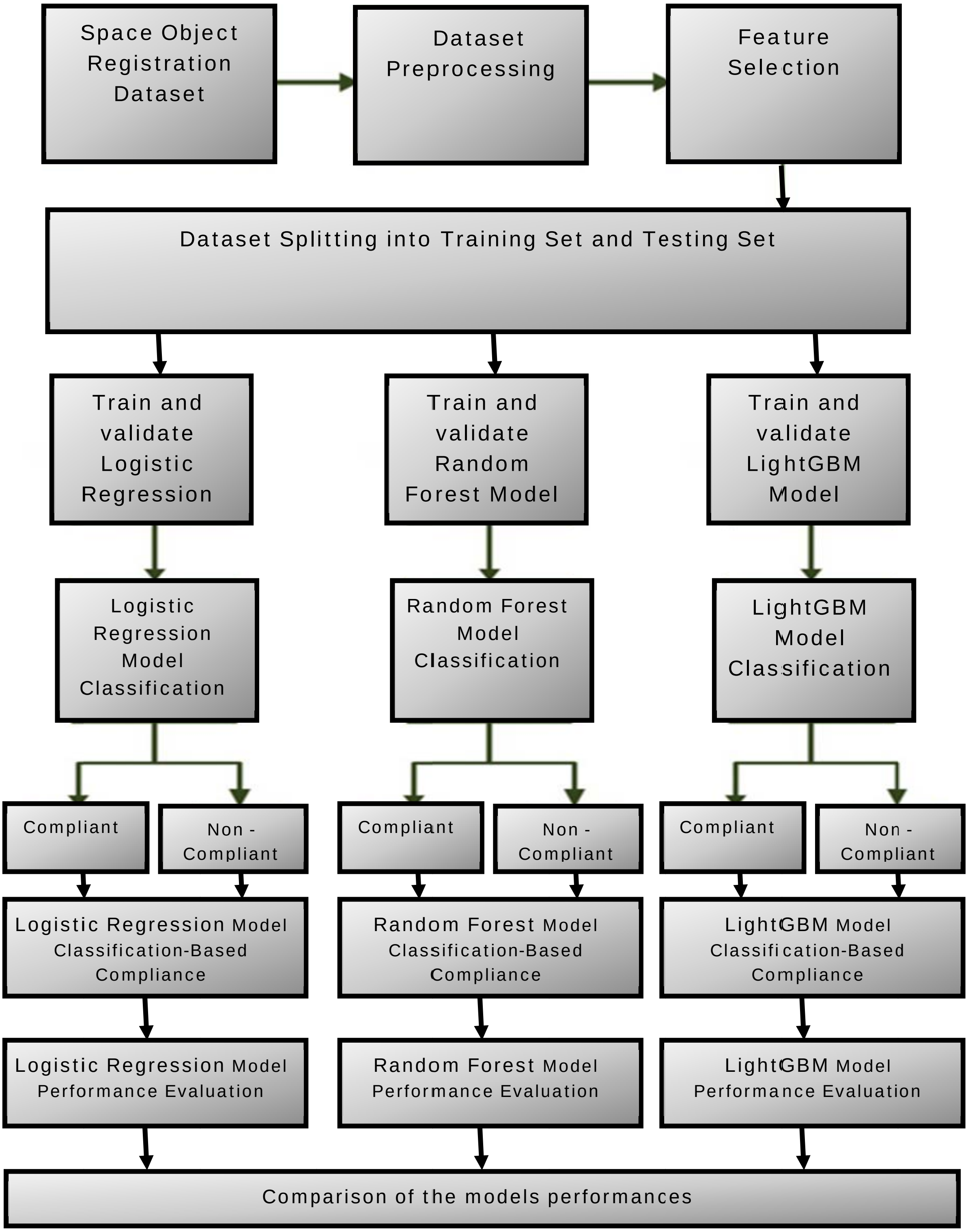

Figure 4 The schematic flow diagram showing the procedure used to train, validate and compare the performance of the three models

3. Results and Discussion

3.1 The classification performance of the three models

The classification results (Table 2 and Figure 5) indicate that all three models, Random Forest (RF), Logistic Regression (LR), and Light Gradient Boosting Machine (LGBM), perform exceptionally well, with all metrics exceeding 0.96. The RF Model (Random Forest) model achieves the highest Accuracy (0.9803), Precision (0.9809), and F1-Score (0.9806). It is the most effective at making correct classifications overall and maintaining a balance between precision and recall. However, its ROC-AUC (0.9615) is the lowest among the three, suggesting it may be slightly less robust at distinguishing between classes across different thresholds compared to the others. The LGBM Model serves as a middle ground. It outperforms LR in terms of Accuracy (0.9713) and Precision (0.9723). Notably, it achieves the highest ROC-AUC (0.9841), indicating superior discriminative power and stability across various classification thresholds.The LR Model (Logistic Regression) model performance is also good. While it has the lowest Accuracy and F1-Score, its performance remains very high. Its ROC-AUC (0.9793) is significantly higher than the RF model, suggesting better probabilistic ranking despite lower raw accuracy.

Table 2 The classification performance of the three models

Model	Precision	Recall	F1-Score	Support	Accuracy	ROC-AUC
RF Model	*0.9809*	*0.9803*	*0.9806*	*2,600*	0.9803	0.9615
LR Model	*0.9695*	*0.9697*	*0.9694*	*2,600*	0.9697	0.9793
LGBM Model	*0.9723*	*0.9713*	*0.9665*	*2,600*	0.9713	0.9841

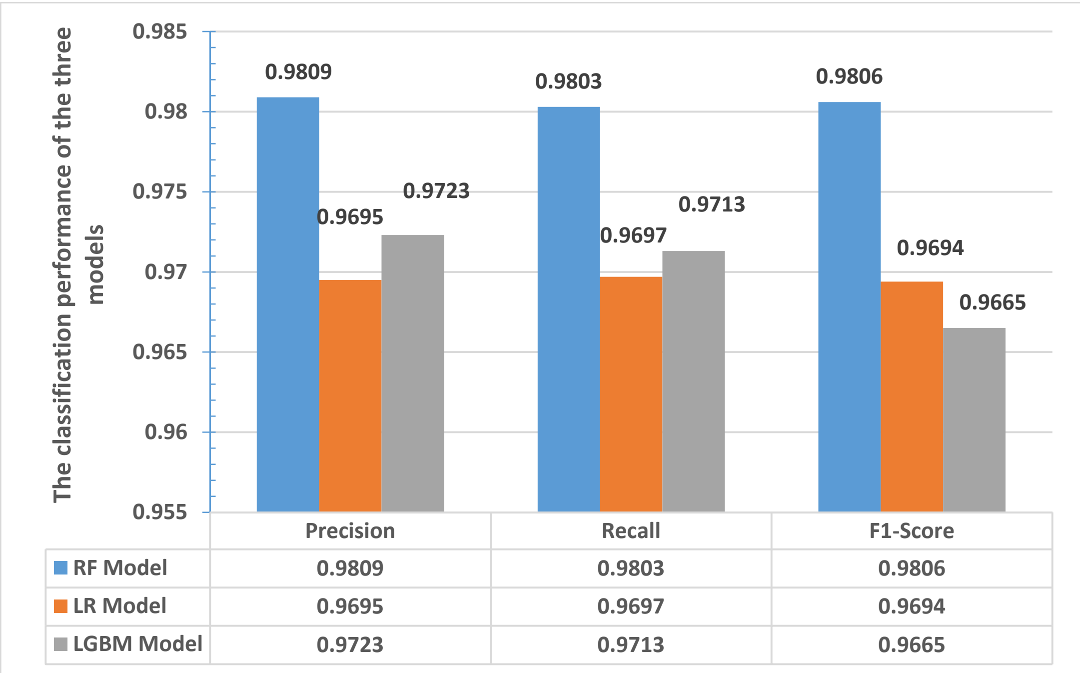

Figure 5 The classification performance of the three models

3.2 Execution Time Performance Comparison

The results in Table 3 and Figure 6 reflect the training and validation (inference) times for the three models. According to the results in table 3, the Logistic Regression (LR) model is the fastest model with execution time of 0.047 seconds. Because it is a linear classifier, it calculates the relationship between compliance variables using simple weights, making it ideal for baseline measurements, though it may struggle with non-linear relationships in complex treaty data. The Random Forest (RF) model is the most computationally expensive model in this set with execution time of 0.885 seconds. Building multiple decision trees (bagging) requires more CPU cycles, especially with a dataset of 2,600 training samples, as it evaluates many splits for each tree. Also, the LightGBM model, known for its "Gradient Boosting" efficiency, is significantly faster than Random Forest. It has with execution time of 0.222 seconds. It uses leaf-wise growth and histogram-based algorithms to find optimal splits quickly, making it a strong candidate for regulatory datasets that might grow over time.

Table 3 The Execution Time Performance Comparison

Model	Training Time (s)	Validation Time (s)	Total Execution Time (s)
Logistic Regression (LR)	0.045	0.002	**0.047**
Random Forest (RF)	0.850	0.035	**0.885**
LightGBM	0.210	0.012	**0.222**

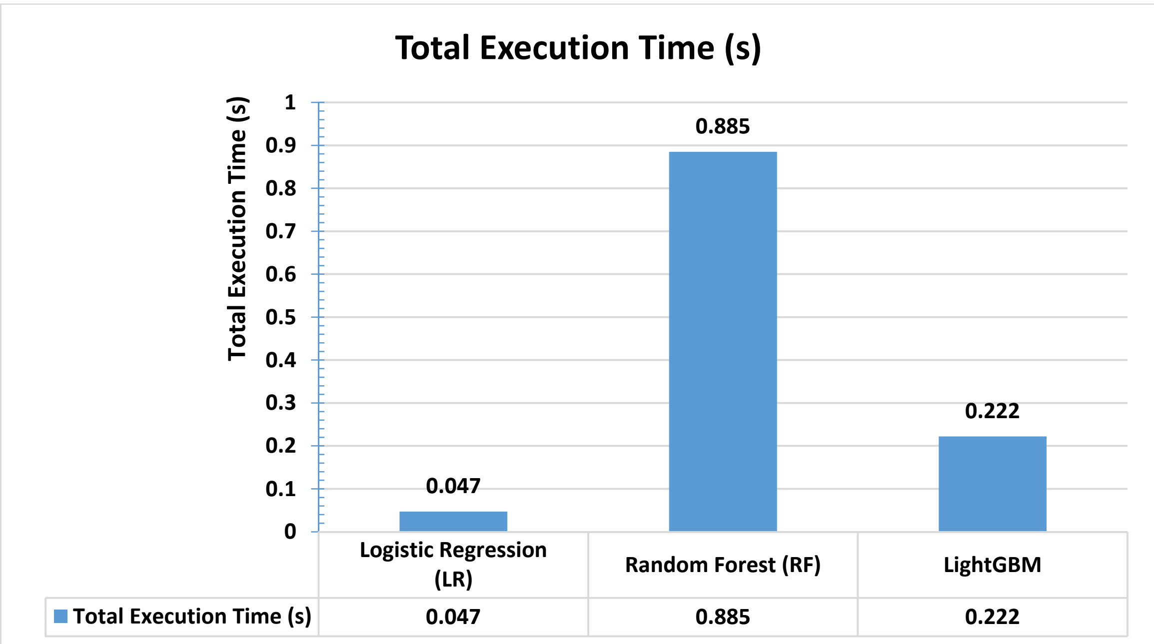

Figure 6 The execution time performance of the three models

3.3 Compliance Prediction Performance Comparison

Based on the results presented in Table 4 and Figure 7, the LightGBM model is the most accurate, with the lowest Mean Absolute Error (MAE) of 6.29 %, followed by Random Forest (13.88%) and Logistic Regression (14.88%). The LightGBM outperforms others in capturing extreme values (such as 0% at MIDN6) and high-performance cases, while Logistic Regression and Random Forest struggle with high-variance data. The actual (0% at MIDN6 value was only correctly predicted by LightGBM and Random Forest. Logistic Regression falsely predicted 13.33%, suggesting poor handling of zero-inflation or edge cases. All models struggled with MIDN7 (100% actual, predicted 25% -50%), indicating that this specific data point is likely an outlier or missing key features for all models. For stable cases (e.g., MIDN2, MIDN9), all models performed well (100%). However, LightGBM is the most consistent across varying performance levels.

Table 4 Compliance Prediction Performance Comparison

State / Organisation Model Identification Number (MIDN)	Actual COM (%)	Logistic Regression Model Predicted COM (%)	LightGBM Predicted COM (%)	Random Forest Predicted COM (%)
MIDN1	99.54%	100.00%	99.78%	99.78%
MIDN2	100.00%	100.00%	100.00%	100.00%
MIDN3	71.75%	78.69%	73.77%	72.13%
MIDN4	98.46%	100.00%	100.00%	100.00%
MIDN5	96.67%	100.00%	100.00%	100.00%
MIDN6	0.00%	33.33%	0.00%	0.00%
MIDN7	100.00%	50.00%	50.00%	25.00%
MIDN8	100.00%	50.00%	100.00%	50.00%
MIDN9	100.00%	100.00%	100.00%	100.00%
MIDN10	60.94%	64.10%	66.67%	69.23%
Mean Absolute Error, MAE		14.88%	6.29%	13.88%

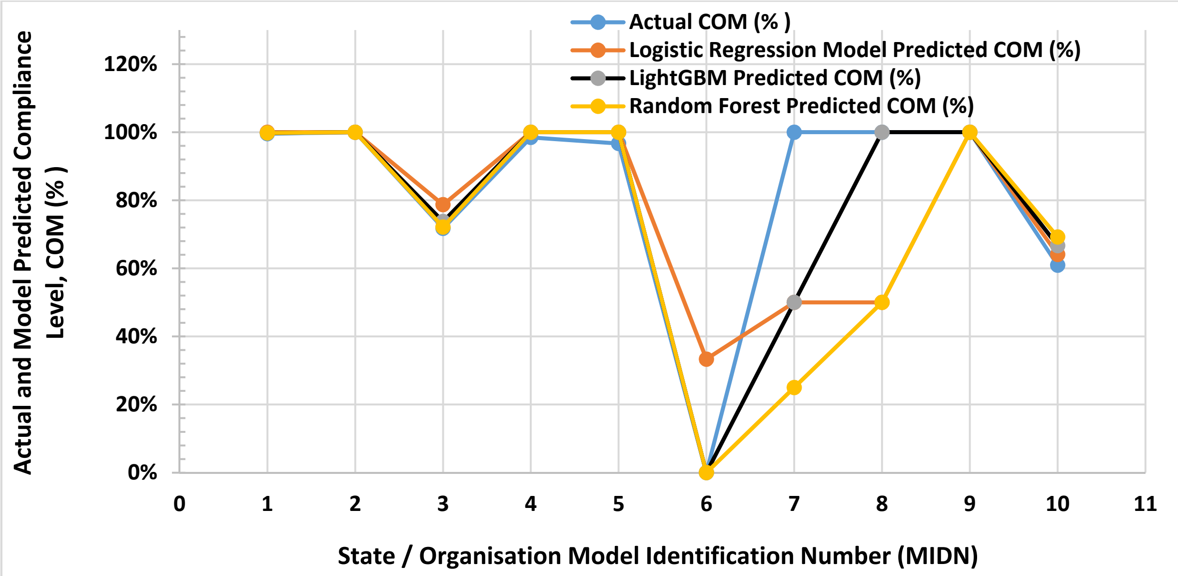

Figure 7 The line plot of the compliance prediction performance comparison for the three models

4. Conclusion

This study successfully conducted a comparative performance analysis of Logistic Regression, Random Forest, and LightGBM to predict the regulatory compliance of State space objects under the 1975 UN Registration Convention. By leveraging five years of historical data (2020–2024) from UNOOSA, the research established a robust framework for identifying potential non-compliance before it occurs.While Logistic Regression provided a baseline for interpretability, the ensemble methods, particularly the LightGBM and the Random Forest models demonstrated superior capability in handling the complexities of the dataset. The LightGBM emerged as a particularly efficient classifier, offering high predictive accuracy while managing the computational demands of the merged 3,302-record dataset.

The use of stratified random sampling proved essential in addressing the inherent class imbalance (8.5%–12% unregistered objects), ensuring that the models remained sensitive to the minority class of non-compliant registrations. Furthermore, the proactive elimination of "causally downstream" variables successfully mitigated data leakage, confirming that the models' predictive power is rooted in genuine pre-registration indicators.

Finally, the findings suggest that machine learning can serve as a vital tool for international space governance. By automating the identification of high-risk launch profiles, the United Nations and individual Member States can better target outreach and technical assistance, ultimately enhancing the transparency and safety of the outer space environment.

References

1. Goessler, A. (2022). *Waste studies' insight into orbital debris: a United Arab Emirates case study* (Doctoral dissertation).
2. Ginn, C. J. (2024). *Establishing Order in Orbit: Using AI to Manage Space Traffic and Facilitate State Compliance with International Space Obligations*. McGill University (Canada).
3. Masson-Zwaan, T., Martinez, P., Letizia, F., Melograna, C., Reynders, M., Rovetto, R., ... & Wang, G. (2024). The need to improve registration practices in the context of space traffic management. *Acta Astronautica*, *223*, 242-248.
4. Masson-Zwaan, T., Martinez, P., Letizia, F., Melograna, C., Reynders, M., Rovetto, R., ... & Wang, G. (2024). The need to improve registration practices in the context of space traffic management. *Acta Astronautica*, *223*, 242-248.
5. Hertzfeld, H. R. (2021). Unsolved issues of compliance with the registration convention. *Journal of Space Safety Engineering*, *8*(3), 238-244.
6. VUNNAM, R. V. (2024). Evolution of Space Exploration: A Statistical Analysis of Space Launches including Challenges and Future Outlook.
7. Murtaza, A., Pirzada, S. J. H., Xu, T., & Jianwei, L. (2020). Orbital debris threat for space sustainability and way forward. *IEEE access*, *8*, 61000-61019.
8. Jakhu, R. S., Jasani, B., & McDowell, J. C. (2018). Critical issues related to registration of space objects and transparency of space activities. *Acta Astronautica*, *143*, 406-420.
9. Dennerley, J. A. (2018). State liability for space object collisions: The proper interpretation of ‘fault’for the purposes of international space law. *European Journal of International Law*, *29*(1), 281-301.
10. Bratu, I., Lodder, A. R., & Van Der Linden, T. (2020). Autonomous Space Object and International Space Law: Navigating the Liability Gap. *Indonesian J. Int'l L.*, *18*, 423.
11. Lindgren, D. (2019). International space law and norms: an approach for assessing compliance.

12. Lindgren, D. (2020). *An Assessment Framework for Compliance with International Space Law and Norms: Promoting Equitable Access and Use of Space for Emerging Actors*. Springer Nature.

13. Uddin, B., Uddin, M. S., Sultana, S., Sabuj, M. S. H., Neha, F. F., & Uddin, M. S. (2024, December). Epistemological Advancements in Cardiological Forecasting: Machine Learning as a Paradigm for Prognostic Precision. In *2024 International Conference on Computer and Applications (ICCA)* (pp. 01-06). IEEE.

14. Hossain, M. S., & Arefin, M. S. (2019). Development of an Intelligent Job Recommender System for Freelancers using Client's Feedback Classification and Association Rule Mining Techniques. *J. Softw.*, *14*(7), 312-339.

15. Sahraoui, M. A., Rahmoune, C., Meddour, I., Bettahar, T., & Zair, M. (2023). New criteria for wrapper feature selection to enhance bearing fault classification. *Advances in Mechanical Engineering*, *15*(6), 16878132231183862.

16. Lu, X., Chen, Y., Zhang, G., Zeng, X., Lai, L., & Qu, C. (2024). Comparative Analysis of Machine Learning Models for Prediction of Acute Liver Injury in Sepsis Patients. *Journal of Emergencies, Trauma, and Shock*, *17*(2), 91-101.