Comparison of Machine Learning Algorithms for Hotel Booking Cancellation in Automated Method

[1st]R. Prabha
Associate Professor,
Department of Electronics and Communication Engineering,
Sri Sai Ram of Institute of Technology,
Chennai, India.
r.praba05@gmail.com

[2nd]Senthil G. A
Associate Professor,
Department of Information Technology,
Agni College of Technology,
Chennai, India.
senthilga@gmail.com

[3rd]A. Sahaya Anselin Nisha
Professor,
Department of Electronics and Communication Engineering,
Sathyabama Institute of Science and Technology, Chennai, India.
anselinnisha.ece@sathyabama.ac.in

[4th]Snega S
Student, Department of Electronics and Communication Engineering,
Sri Sai Ram of Institute of Technology,
Chennai, India.
sit19ec040@sairamtap.edu.in

[5th]Keerthana. L
Student, Department of Electronics and Communication Engineering,
Sri Sai Ram Institute of Technology,
Chennai, India.
sit19ec021@sairamtap.edu.in

[6th]Sharmitha S
Student, Department of Electronics and Communication Engineering,
Sri Sai Ram Institute of Technology,
Chennai, India
sit19ec082@sairamtap.edu.in

2022 International Conference on Computer, Power and Communications (ICCPC) | 979-8-3503-9784-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCPC55978.2022.10072135

***Abstract*— People usually book hotels online and cancel booking due to various reasons, which is a loss to the business, which is an important problem for hotel managers. These articles examined how artificial intelligence is used to determine which reservations can be cancelled and therefore avoid some losses. The study compares algorithms that forecast cancellation, such as Decision Trees, Naive Bayes, KNN, Logistic Regression, and Random Forest. The data is obtained from the publicly available dataset. Lots of insights from the data can be fetched and pre-processing, feature encoding and engineering is applied. This method is carried out to develop a model which gives minimum error and good accuracy in an automated process.**

Keywords*— *Machine Learning, Data Science, Automation, Hotel Management, K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Logistic Regression, Algorithm, Naive Bayes, Support Vector Machine (SVM), Booking, Cancellation, Dataset, Data Visualization.

I. INTRODUCTION

When it comes to tourism, hotel plays an important role. Tourists from various countries prefer hotels to stay. The main aim of hotels is to provide service and for profit. Cancellations can affect the hotel's perspective among the customers. Nowadays people can be easily influenced by others. In some cancellation policies, there is some certain date until when we can cancel the booking for free and money will be refunded. At some places, the money is non-refundable. In reality, according to a survey done by D-Edge Hospitality Services, the cancellation cost across all platforms has increased by 6% during last four years, reaching approximately 40% in 2018. This rise in cancellations makes it more difficult for hotels to estimate properly, resulting in suboptimal utilization and revenue loss.

With the emerging tendency of cancellation season to season, some hoteliers believe that widespread cancellation in hotels is the new industry standard, which is entirely incorrect. One out of every five hotel visitors cancel hotel bookings ahead of stay. This cancellation pattern has resulted in the guesthouse being unable to effectively estimate number of rooms within their revenue management, as well as a lost in economic cost.

These article tries to prevent cancellations, over policies, and rigid administration, all of which can have a detrimental impact on the hotel. Reservation characteristics may be strong indicators of the likelihood that a reservation will be cancelled. For example, with an advance time of 60 days, the average duration of stay of cancelled bookings is 65% longer than that of non-cancelled reservations. It is critical to understand why customers cancel and what sorts of appointments are being cancelled. To answer this challenge, we will develop a customer feature extraction study using a real-world hotel accommodation dataset in order to acquire insights about the consumers.

To resolve this problem, a machine learning-based system model was established. Machine Learning has recently become a popular technology. Instead of hard coding, Machine Learning automates jobs.

Researchers programme the machine to understand and deliver reliable results. The dataset should be utilized to train the model. It is used to test the algorithm's accuracy and prediction capacity. We are using numerous different machine learning technique to predict cancellation, such as Decision Trees, Naive Bayes, KNN, Logistic Regression, Random Forest, AdaBoost, Gradient Boosting, which would be especially in comparison in the manuscript by performing collection of data, data pre-processing, statistical modelling, visualization of data, feature engineering, and feature encrypting. Once we have the appropriate findings, it is simple to take the necessary actions to optimise more cancellations and give the greatest customer experience possible.

II. RELATED WORK

Intrinsic part Rhetoric Summarization of Accommodation Reviews - Customer feedback is also significant in

determining cancellations. This article describes how to do sentiment analysis on feedback and also how hotels may leverage this to enhance their service. [1,10].

Using a Machine Learning Classification Model to Predict Hotel Booking Cancellation - Since hotel booking cancellation results is loss to hotels, this paper tells how artificial intelligence is used to predict the cancellations.

A Critical Review of Application of Machine Learning in the Hotel Industry - Nowadays Machine Learning became a trending technology, this study tells how this trending technology can be used in hotel and tourism [2,11].

Predicting hotel cancellations using data science. The cancellation may be predicted with great accuracy, according to the Handbook of Research on Holistic Optimization Algorithms in the Hospitality, Tourism, and Travel Industry. In order for hotels to enhance their allocation strategy and cancellation policies [3,12].

An iterative method for estimating hotel daily utilization using competing datasets. This study describes how to enhance accuracy by utilizing a recursive technique to forecast cancellation using a competition set of inputs [14-16].

Güven,., and F. Simsir (2020). Demand Forecasting Using Artificial Neural Networks (ANN) and Support Vector Machines in the Retail Apparel Industry (SVM) Techniques. The purpose of this article is to build predictive models for textile retail establishments using Artificial Neural Networks (ANN) and Support Vector Machines (SVM).

S. Hwang, J. Kim, E. Park, and S. J. Kwon Who will be your next client: A machine learning approach to returning customers in aircraft services - One such article includes K Nearest Neighbors and Support Vector Machine to assess feedback to determine if a consumer would return [11].

Nugroho Rendi Septian, Adi Putro. Deep Neural Network and Logistic Regression forecasting of hotel booking cancellation - In this, article analyzes Deep Neural Network with Logistic Regression machine learning models.

Eleazar CSánchez, Agustn J. Sánchez-Medina Using machine learning and big data to anticipate hotel booking cancellations efficiently. It uses Artificial Neural Network (ANN) and Support Vector Machine to determine which Table 1. Dataset customers are most likely to cancel their reservation (SVM).

L. Masiero, G. Viglia, and M. Nieto-Garcia (2020). Consumer behaviour in online hotel booking that is opportunistic. Using an Analytical Model, it studies customer behaviour to discover approaches for improved strategy [13].

III. DATASET DESCRIPTION

This dataset is publicly available. There are119391 values with two classes. There are 9 categorical columns. The dataset also contains missing values. The dataset contains values like arrival date, year, month, number of adults and babies. reserved room type, repeated guest, already cancelled etc. The following columns are present in the dataset.

Table I. Dataset

Column Name	Datatype
hotel	object
is_canceled	int64
lead_time	int64
arrival_date_year	int64
arrival_date_month	object
arrival_date_week_number	int64
arrival_date_day_of_month	int64
stays_in_weekend_nights	int64
stays_in_week_nights	int64
adults	int64
children	float64
babies	int64
meal	object
country	object
market_segment	object
distribution_channel	object
is_repeated_guest	int64
previous_cancellations	int64
previous_bookings_not_canceled	int64
reserved_room_type	object
assigned_room_type	object
booking_changes	int64
deposit_type	object
agent	float64
company	float64
days_in_waiting_list	int64
customer_type	object
adr	float64
required_car_parking_spaces	int64
total_of_special_requests	int64
reservation_status	object
reservation_status_date	object

IV. METHODOLOGY

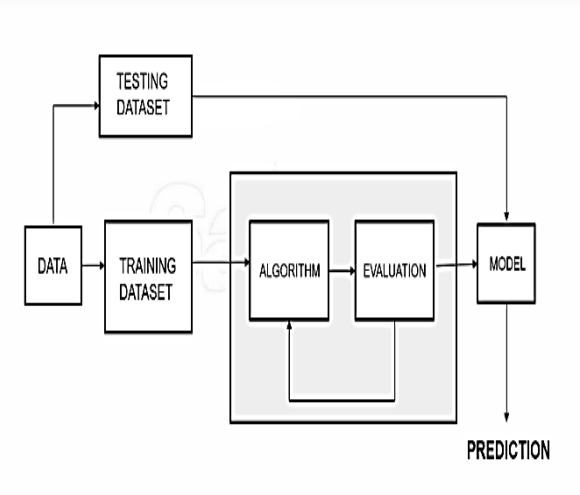

Fig. 1. Flow Diagram for Steps in Model Building

A. Data Preparation

It is the first step of Machine Learning. The real data contains various inconsistencies. It should be cleaned before using. There may be some missing values in the dataset, which is due to glitch in data collection. There will be outliers in the dataset which is. There may be negative values in the dataset. It is important to find relevant features. The dataset should be cleaned to get proper accurate results. The missing values can be replaced with mean, mode or neighbor values. The outliers are detected and removed. The incorrect and irrelevant data is removed. It is the important part in Machine Learning.

Fig. 2. Algorithm Accuracy

B. Data Visualization

We use libraries such as seaborn, matplotlib, folium, plotly for data visualization.

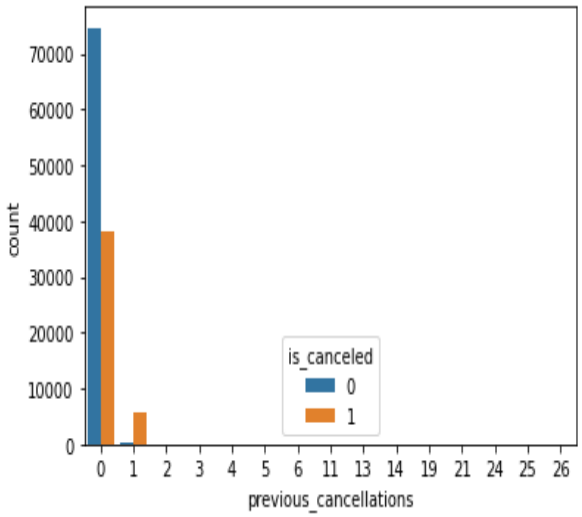

Fig. 3. Previous Cancellations

Customers with the most cancellations have none. Customers are less inclined to cancel their present reservation. Customers who have previously cancelled are more likely to cancel the current booking. This is consistent with the harmonious relationship between previous cancellations with is cancelled.

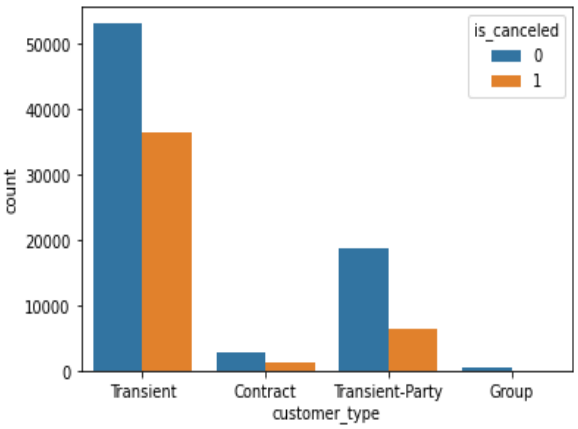

Fig. 4. Total Number of Guests per Month

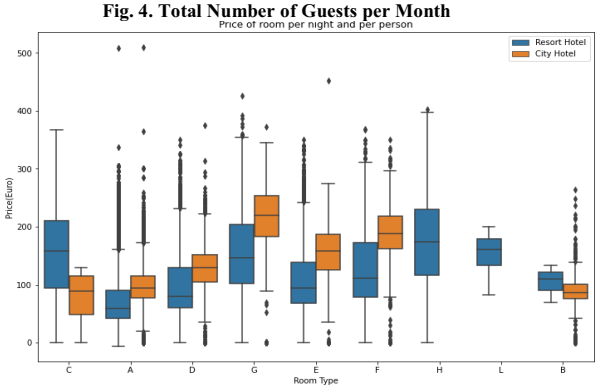

Fig. 5. Resort vs City

Transient clients account for 75% of all bookings. It also has the highest cancellation rate in any category. As can be observed, city hotels have substantially longer waited times in days than resort hotels, indicating that their demand is stronger.

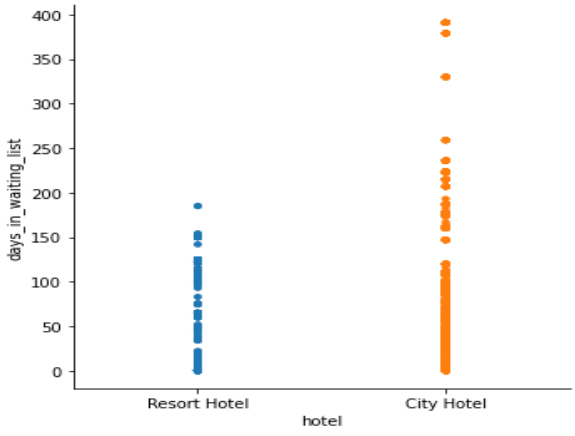

Fig. 6. Room Price Per Night Over the Month

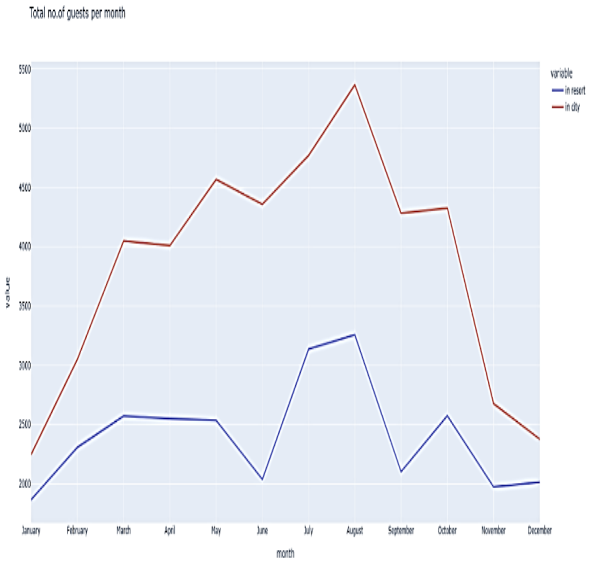

Fig. 7. Total No. of Guests Per Month

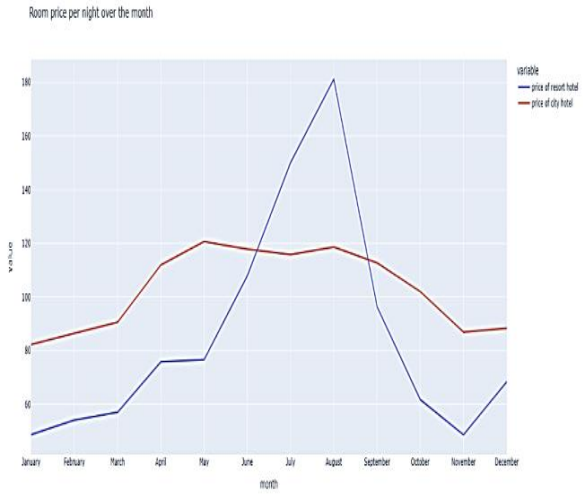

Fig. 8. Room Price Per Night

C. *Feature Selection*

After the dataset is cleaned, finding relevant features for the model. The features can be picked logically or we can use feature selection methods. By giving all features to the model, the model may learn irrelevant data and will be accurate. This can reduce size of the data. LASSO Regression is used to feature selection in this study. Least Absolute Shrinkage and Selection Operator minimizes cost function and select useful features and discards other features. The coefficients that are different from zero are selected in this method. The alpha hyperparameter is tuned.

D. *Classification Model*

The dataset is split up into two parts: test and training. StandarScaler is used to scale the numerical values. Also, MinMaxScaler. For predictions, we employ methods such as Decision Tree Classifier, Random Forest Classifier, Gradient Boosting, Ada Boosting, K-Nearest Neighbors, and Nave Bayes. We divided the dataset into two parts: both training and testing. Sklearn libraries include Decision Tree Classifier, Random Forest Classifier, GaussianNB, AdaBoost Classifier, Kneighbors Classifier, and Gradient Boosting Classification algorithm. An object named model is constructed, and all of the algorithms are applied automatically. All of the parameters' hyperparameters have been tuned. The tunable model's assessment matrix is validated. Imported metrics include the discriminant function, calibration Scores, and f1 score. Several metrics are used to assess the model.

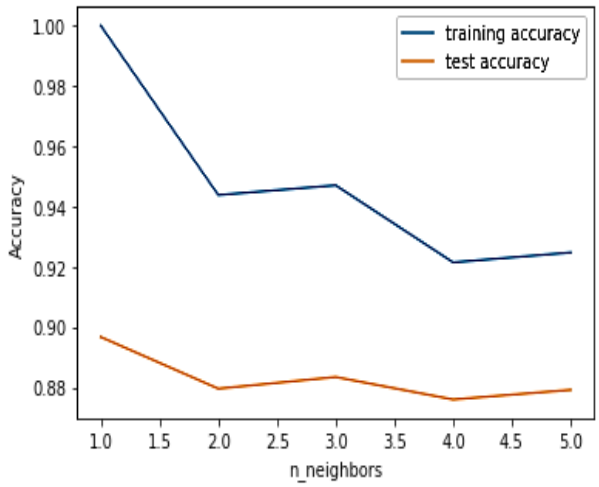

Fig. 9. KNN Accuracy

E. *Mathematical Equation*

The main idea in logistic regression is to model p(X) as;

$$p(X) = \frac{e\ \beta 0+\beta 1X1\cdots+\beta iXi}{1 + e\ \beta 0+\beta 1X1\cdots+\beta iXi}. \quad (1)$$

Gini Index of Random Forest:

$$Gini = \sum K\ k=1\ p\hat{}nk(1 - p\hat{}nk) \quad (2)$$

XGBoost Boosted Model Output

$$f(x) = \sum M\ m=1\ \lambda\ \hat{}f\ m(x) \quad (3)$$

Information Gain = Entropy(S)- [(Weighted Avg) *Entropy (each feature)]

$$Entropy(s)= -P(yes)log2\ P(yes)- P(no)\ log2\ P(no) \quad (4)$$

$$Gini\ Index= 1- \sum_j P_j^2$$

V. RESULT AND DISCUSSION

This article compares the underlying algorithms. When all of the algorithms, such as Logistic Regression, Nave Bayes, KNN, Random Forest, Decision Tree, AdaBoost, and Gradient Boosting, are compared, we can find that Random Forest and KNN provided about 93% accuracy, and this model may be employed. The Decision Tree algorithm is a classifier made up of several decision trees. Rather of relying on a single decision tree, the Random Forest collects predictions from all trees and speculates on the eventual outcome depending on the majority vote of projections. The greater the percentage of trees in the forest, the higher the accuracy and smaller the danger of premature convergence.

Table II. Comparison of Algorithms

ALGORITHMS	CONFUSION MATRIX	ACCURACY	F1 SCORE
Naive Bayes	[[24421 31898] [3208 29880]]	0.6073461	0.629941
KNN	[[56030 289] [7273 25815]]	0.915420	0.87224
Random Forest Classifier	[[55805 514] [4975 28113]]	0.938606	0.911058
Decision Tree Classifier	[[52588 3731] [3544 29544]]	0.918630	0.89037
AdaBoost	[[53132 3187] [13238 19850]]	0.81628	0.70734
Gradient Boosting	[[56038 281] [9306 23782]]	0.89277	0.83225

Confusion matrix tells the performance of the classification model.

$$\text{False positive rate} = \frac{FP}{(FP + TN)} \quad (5)$$

$$\text{False negative rate} = \frac{FN}{(FN + TP)} \quad (6)$$

.

For accuracy divide correct number of classifications with all classifications.

$$\text{Accuracy} = \frac{TP + TN}{(FN + FP + TP + TN)} \quad (7)$$

$$\text{Precision} - \frac{TP}{(TP + FP)} \quad (8)$$

$$\text{Recall} - \frac{TP}{(TP + FN)} \quad (9)$$

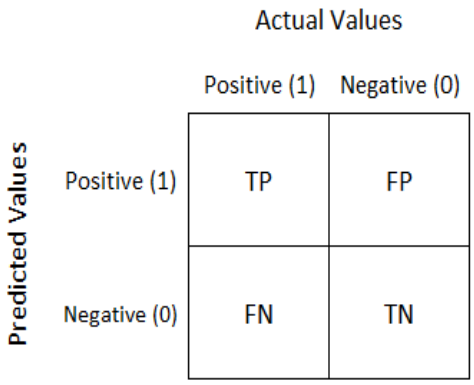

Fig. 9. Confusion Matrix

i. True Positive is predicted is true and it is actually true.

ii. False Positive is predicted true but not actually true.

iii. False Negative is predicted false and is actually false.

iv. True Negative is predicted false but no actually false.

VI. CONCLUSION

We have been able to develop a robust model that did not overfit to the data. On the test data, we received a f1 score of 0.91 and an accuracy of 0.93. Models for booking cancellations enable hospitality industry to create less stringent cancellation procedures. This improves business performance since less stringent cancellation procedures result in more reservations. The cancellation provision should be changed to make it more stringent. It should also conduct a small poll of consumers who cancel reservations to see why they're there.

From a business perspective if a booking is flagged as a cancellable booking, we could send the customer an e-mail to confirm their booking. If they are not able confirm the booking would be deemed cancelled and the customer has to reserve a new booking. This would ensure a more vigilant booking method on part of the hotel and would also reduce the number of false bookings. Since we were working here with a limited amount of data there is a pretty good chance of model decay in the future. To counter this, periodic remodelling and configuring of the model would be required.

These approach permits hotel management to take action on possibly cancelled reservations even while producing more exact market estimates. It is also a good idea to perform a satisfaction survey with clients who confirm their stay to seek out their overall opinions of the services provided. According to the data, the bed and breakfast regime is the most popular, so it is critical to ensure that its agreed to implement the clients' expectations. This knowledge might assist hotels in determining when they need to boost their advertising budget, such as gradually raising their advertising

budget as the peak season approaches. They also can improve staffing to handle the increased volume.

VII. FUTURE SCOPE

These are the initial steps. Prediction of trips which are likely to be cancelled can also be used to find out real users and the other users who reserve in advance and sell to others in increased price and cancel when no one buys. A better comprehension of the issue may need further hotel-specific knowledge (such as surrounding parking availability or deposit policies). This shows that working more closely with hotels and even developing hotel-specific models is a smart further move. Implementation of advanced chatbots that can make guess what customer is going to ask next can speed the process, which alerts and monitors. For production, the process could be automated to reduce the workload on hotel staff and management. The data produced by Travel Industry cannot be processed by human beings. Highly supervised algorithms should be used.

REFERENCES

[1] António, N. (2019). Predictive models of hotel booking cancellation: a semi-automated analysis of the literature. Tourism & Management Studies, 15(1), 7-21.

[2] Lee, Misuk. "Modeling and forecasting hotel room demand based on advance booking information." Tourism Management 66 (2018): 62-71, https://doi.org/10.1016/j.tourman.2017.11.004.

[3] Schwartz, Z.; Uysal, M.; Webb, T.; Altin, M. Hotel daily occupancy forecasting with competitive sets: A recursive algorithm. *Int. J. Contemp. Hosp. Manag.* **2016**, *28*, 267–285.

[4] Cirillo, C., Bastin, F., & Hetrakul, P. (2018). Dynamic discrete choice model for railway ticket cancellation and exchange decisions. Transportation Research Part E: Logistics and Transportation Review, 110, 137–146.

[5] Park, J. Y., & Nagy, Z. (2018). Comprehensive analysis of the relationship between thermal comfort and building control research — A data driven literature review. Renewable and Sustainable Energy Reviews, 82, 2664–2679.

[6] Antonio, N.; De Almeida, A.; Nunes, L. An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations. *Data Sci. J.* **2019**, *18*, 1–20.

[7] Antonio, N. Predictive models for hotel booking cancellation: A semi-automated analysis of the literature. *Tour. Manag. Stud.* **2019**, *15*, 7–21.

[8] Dimiduk, D.M.; Holm, E.A.; Niezgoda, S.R. Perspectives on the Impact of Machine Learning, Deep Learning, and Artificial Intelligence on Materials, Processes, and Structures Engineering. *Integrating Mater. Manuf. Innov.* **2018**, *7*, 157–172.

[9] Attaran, M.; Deb, P. Machine learning: The new 'big thing' for competitive advantage. *Int. J. Knowl. Eng. Data* Min. 2018, 5, 277–305.

[10] S. Soundararajan, R. Prabha, M. Baskar and T. J. Nagalakshmi, "Region centric gl feature approximation based secure routing for improved qos in manet," Intelligent Automation & Soft Computing, vol. 36, no.1, pp. 267–280, 2023.

[11] Senthil, G. A., R. Prabha, A. Pomalar, P. Leela Jancy, and M. Rinthya. "Convergence of Cloud and Fog Computing for Security Enhancement." In 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 1-6. IEEE, 2021. DOI: 10.1109/I-SMAC52330.2021.9640872.

[12] R. Prabha, S. G. A, N. N. Saranya, A. M, K. Somasundaram and K. C, "Design of Adaptive Priority Based IoT Communication in Wireless Network," 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA), 2022, pp. 1-5, doi: 10.1109/ICACTA54488.2022.9753550.

[13] S. G. A, R. Prabha, M. Razmah, S. Sridevi, D. Roopa and R. M. Asha, "A Big Wave of Deep Learning in Medical Imaging - Analysis of Theory and Applications," *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2022, pp. 1321-1327, doi: 10.1109/ICICCS53718.2022.9788412.

[14] K.C. Suresh, R. Prabha, N. Hemavathy, S Sivarajeswari, D Gokulakrishnan, M. Jagadeesh kumar, A Machine Learning Approach for Human Breath Diagnosis with Soft Sensors, Computers and Electrical Engineering, Volume 100, 2022.

[15] S. Sridevi, R. Prabha, K. N. Reddy, K. M. Monica, G. A. Senthil and M. Razmah, "Network Intrusion Detection System using Supervised Learning based Voting Classifier," 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), 2022, pp. 01-06, doi: 10.1109/IC3IOT53935.2022.9767903.

[16] R-Prabha M, Prabhu R, Suganthi SU, Sridevi S, Senthil GA, Babu DV. Design of Hybrid Deep Learning Approach for Covid-19 Infected Lung Image Segmentation. In Journal of Physics: Conference Series 2021 Oct 1 (Vol. 2040, No. 1, p. 012016). IOP Publishing. doi:10.1088/1742-6596/2040/1/012016.