Journal of Soft Computing Paradigm (ISSN: 2582-2640)
www.irojournals.com/jscp/

Performance Comparison of different Disease Detection using Stacked Ensemble Learning Model

Arunya Paul[1], Tejaswini Kar[2], Sasmita Pahadsingh[3], Priya Chandan Satpathy[4], Biswaranjan Behera[5]

[1-3]School of Electronics Engineering, KIIT Deemed to be University, Bhubaneswar, India

[4,5]Aryan Institute of Engineering and Technology, Arya Vihar, Bhubaneswar, India.

E-mail: [1]arunyapaul@gmail.com, [2]tkarfet@kiit.ac.in, [3]spahadsinghfet@kiit.ac.in, [4]priyachandansatapathy@gmail.com, [5]biswaranjanbehera30123@gmail.com

Abstract

Malignancy risks and genetic disorders have long been challenging due to procedures that lack precision and predictability, thereby complicating the precise identification of diseases and their root causes. Machine learning classifiers have emerged as more suitable and effective tools. Various machine learning classifiers have been utilized to examine different genetic disorders, and the results from these classifiers have been further compared to determine their superiority. In this study, a variety of classifiers, including the SVM, KNN, decision tree, random forest, and logistic regression algorithms, are examined. These classifiers utilize specific training variables to analyze how input values correspond to the respective class. After successfully implementing each classifier, we proceeded to employ Stacking, an ensemble machine learning technique that aggregates predictions from individual classifiers on the same dataset. Four datasets, including the breast cancer, diabetes, Parkinson's, and genomic datasets, were successfully implemented using the aforementioned methods, and the results obtained showed how the input values correspond to the class using a few training variables. SVM classifier was shown to be the most effective of the five described classifiers, having the highest accuracy in most of the cases. It provided accuracies of 97.43%, 97.46%, 97.45%, and 97.44% for each of the genome cancer, diabetes, Parkinson's, and breast cancer datasets. The

DOI: https://doi.org/10.36548/jscp.2024.1.003
Received: 04.02.2024, received in revised form: 02.03.2024, accepted: 17.03.2024, published: 26.03.2024

KNN and Random Forest models also came out to be very effective, with accuracy around 95% and 91%, respectively, for various disease datasets. The Logistic Regression and Decision Tree models also worked well. However, the ensemble method of Stacking proved to be highly efficient above all other base models and generated accuracies above 97.5% for all the aforementioned diseases.

Keywords: Machine Learning Classifiers, Disease Detection, SVM, KNN, Decision Tree, Random Forest, Logistic Regression Algorithms, and Stacking.

1. Introduction

Machine Learning, as a subject of investigation, draws upon and integrates principles from various closely linked disciplines, within the field of artificial intelligence [1][2]. Learning, or gaining the necessary abilities or information through practical application, is the main focus. In the broadest sense, this refers to gathering relevant insights from provided historical data [3]. Linear regression, logistic regression, Decision Trees, Support Vector Machine (SVM), and K-nearest neighbor classifier (KNN) are examples of some Machine Learning Algorithms [4]. When describing the support vector machine, or SVM, algorithm, we have SVM kernel functions that assist in altering the dimensions of the data. SVM methods make use of kernels, a collection of small functions [5]. A kernel's job is to take data as input and change it into the form that is desired [6]. Linear Kernels are the most basic kind of kernels, typically being one dimensional in nature [7]. When there are many features, it works well. Comparatively speaking to other functions, linear kernel functions are quicker [8]. When separating data using a straight line is not possible, non-linear kernels are utilized. They convert a space from nonlinear to linear. Data is transformed into a different dimension so that it may be categorized. By adding, it turns the two variables x and y into three variables. One of the most popular methods for evaluating models is K-fold cross-validation. Although even though this strategy is less well-known than the validation set approach, it may help us understand our data and model better [9][10][11].

In our proposed work, we investigated a range of classifiers, including the SVM, KNN, decision tree, random forest, and logistic regression methods. These classifiers will employ certain training variables and examine the relationship between the input values and the class. After each classifier was successfully constructed, we continued to use Stacking, an ensemble

machine learning approach that combines all of the predictions from each individual classifier on the same provided data set.

2. Description of the Dataset

For this work we have considered four different types of datasets with three different diseases. The first genome cancer dataset was collected from NCBI and rest two datasets i.e. diabetic and Parkinson's disease dataset were collected from Kaggle. The first one is a genome date set containing 390 samples, 44 gene features and five variants of cancer. This cancer dataset contains 78 samples from each class. The second one is the Diabetes disease dataset containing 768 samples and 8 features having two classes. The third one is the Parkinson's disease dataset containing 195 samples and 23 features having two classes. The Parkinson's dataset is composed of a range of biomedical voice data measurements of 31 people, out of which 23 are having Parkinson's disease (PD). Each column in the data corresponds to a particular voice measure, and each row corresponds to one of 195 voice recordings from these individuals. The two classes are set to 0 and 1 where 0 represents health and 1 represents a person with PD. There are around six recordings per patient. Last one is Breast cancer dataset collected from Kaggle. All the data are in CSV format.

3. Methodology

Here, different standard machine learning based classifiers, such as SVM, KNN, Random Forest, Decision Tree, and Logistic Regression classifiers are used and Stacking is employed for an optimized result.

1) **SVM** - also known as support-vector machines—are supervised learning models with associated learning algorithms that look at data for classification and regression analysis. In order to widen the gap between the two classes, SVM maps produce recommendations to focus in space.

2) **k-NN -** is an acronym for the k-closest neighbors' algorithm, which is a non-parametric classification technique. It is used for regression and classification. The information in both situations consists of the k closest preparing models in a data collection. The item is then simply demoted to the class of that single nearest neighbor if k = 1.

3) **Logistic Regression:** This logistic model may be used to display a variety of situations akin to determining what an image includes. Each object spotted in the picture would be assigned a probability between 0 and 1, with a total of one.

4) **Random Forest:** The supervised machine learning algorithm includes random forest. Its foundation is the idea of ensemble learning. It is the practice of integrating various classifiers to address complex issues and enhance model performance.

5) **Decision Tree**- A supervised learning technique that may be used to address both classification and regression issues, decision trees are categorized under this category. It is a structured classifier that resembles a tree, with internal nodes that stand in for the dataset's characteristics. Branches are used to indicate decision-making processes. The result is represented by each leaf node.

6) **Ensemble Technique:** It combines different learning algorithms for generating one optimal predictive model. The model has better performance than the individual base learners. In this work we considered the Stacking model for the performance evaluation. Stacking model considers different heterogeneous weak learners allowing them to learn in parallel, and finally combines them by training a meta-learner to generate a prediction based on the individual weak learner’s predictions. A meta learner takes inputs as the predictions, as the features and the ground truth data as the target, and tries to combine the input predictions in the best possible way for making a better output prediction [12]. The important steps of the stacked Ensemble model are as follows.

Step 1: Apply a K-Fold cross validation by separating the data set into K-Folds.

Step 2: Out of K fold data one-fold is deployed for testing and other folds are deployed for training for different base models and the process is repeated K times.

Step 3: All the out of sample predictions are fed as features to the meta model.

Step 4: The final output is predicted using the meta model. The Complete Stacked Ensemble learning model is shown in Figure. 1.

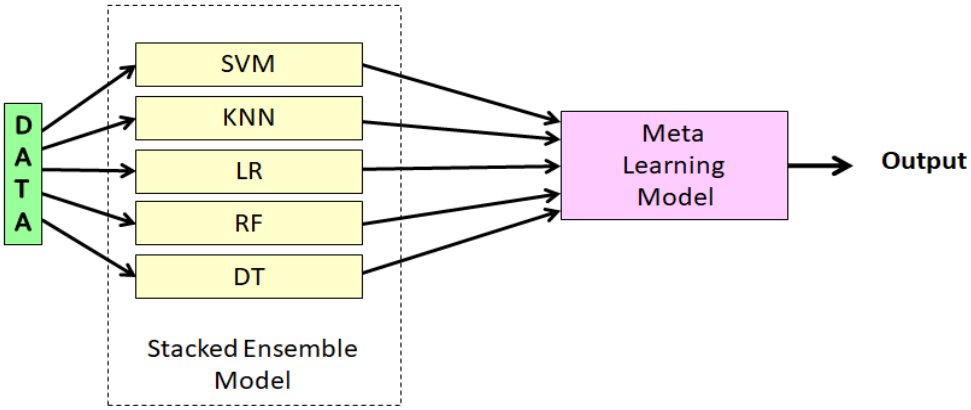

Figure 1. Stacked Ensemble Learning Model

4. Results and Discussion

For simulation and results analysis, we have considered three performance parameters such as precision, recall and F1 measure, accuracy and support. The "support" shows as the number of outcomes of the desired label. For handling imbalanced data sets we have used the SMOTE (synthetic minority oversampling technique) algorithm. SMOTE aims at balancing the class distribution by randomly increasing minority class samples through replication [13][14].

To conduct the simulation, all datasets were divided into training and testing sets, with the ratio being 80% for training and 20% for testing.[15] .

4.1 Genome Cancer Dataset

The performances of the genome cancer dataset of individual classifier and using stacking models were given in Table 1 and Table 2 respectively. The mean accuracy and the standard deviation are obtained through cross validation.

Table 1. Performance of Individual Classifiers for Genome Cancer Dataset.

Base Model	Accuracy (%)
SVM	97.43

K-NN	95.91
Random forest	90.80
Logistic Regression	83.50
Decision Tree	86.61

Table 2. Performance of Stacked Ensemble Model for Genome Cancer Dataset

Base Model	Mean Accuracy (standard deviation)
SVM	0.974 (0.005)
K-NN	0.959 (0.004)
Random forest	0.908 (0.006)
Logistic Regression	0.835 (0.011)
Decision Tree	0.866 (0.010)
Stacking	0.977(0.005)

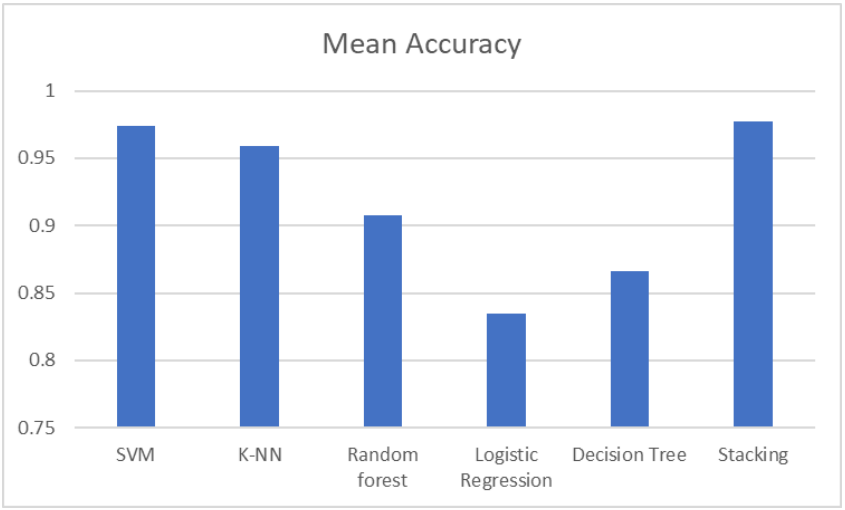

Figure 2. Mean Accuracy for Genome Cancer Dataset

The Figure.2 shows the performance of the individual classifiers and the stacked ensemble classifier for genome dataset.

Table 3. Performance of Individual Classifiers for Diabetes Dataset

Base Model	Accuracy (%)
SVM	97.46
K-NN	95.90
Random forest	91.01
Logistic Regression	83.51
Decision Tree	86.50

4.2 Diabetes Dataset

The performance of diabetic dataset for individual classifiers and using stacking model was given in Table 3 and Table 4 respectively.

Table 4. Performance of Stacked Ensemble model for Diabetes Dataset

Base Model	**Mean Accuracy (standard deviation)**
SVM	0.974(0.005)
K-NN	0.959(0.004)
Random forest	0.910(0.008)
Logistic Regression	0.835(0.011)
Decision Tree	0.865(0.010)
Stacking	0.975(0.005)

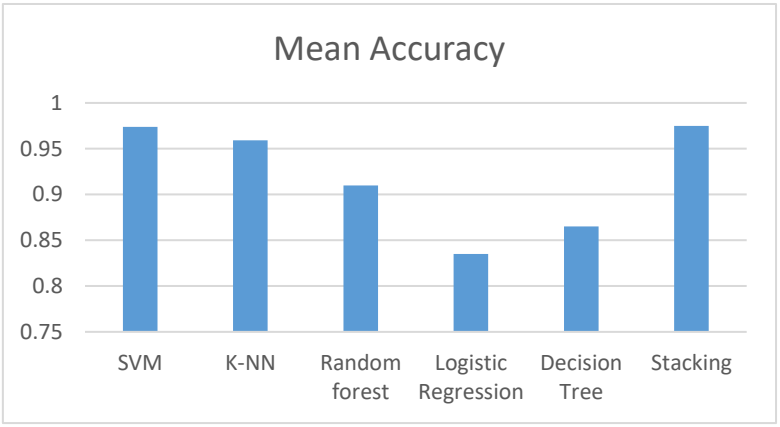

Figure 3. Mean Accuracy for Diabetes Dataset

The Figure.3 shows the performance of the individual classifiers and the stacked ensemble classifier for Diabetes dataset.

4.3 Parkinson's Dataset

The performances of each classifier and ensemble model were presented in Table 5 and 6.

Table 5. Performance of Individual Classifiers for Parkinson's Dataset.

Base Model	Accuracy (%)
SVM	97.45
K-NN	95.90
Random forest	90.90
Logistic Regression	83.52
Decision Tree	86.60

Table 6. Performance of Stacked Ensemble Model for Parkinson's Dataset.

Base Model	Mean Accuracy (standard deviation)
SVM	0.974(0,005)
K-NN	0.959(0.004)
Random forest	0.909(0.009)
Logistic Regression	0.835(0.011)
Decision Tree	0.866(0.010)
Stacking	0.976(0.005)

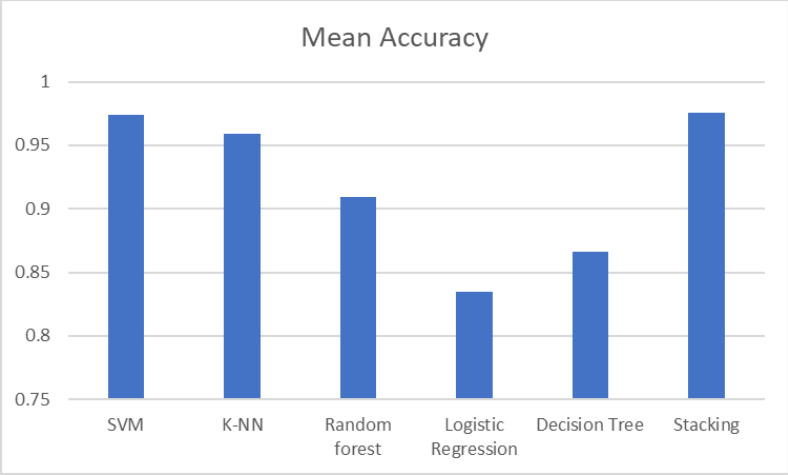

Figure 4. Mean Accuracy for Parkinson's Dataset.

The Figure.4 shows the performance of the individual classifiers and the stacked ensemble classifier for Parkinson's dataset.

4.4 Breast Cancer Dataset

The performance of individual classifiers and Stacked Ensemble models for the above data set were presented in Table 7 and Table 8 respectively.

Table 7. Performance of Individual Classifiers for Breast Cancer Dataset.

Base Model	Accuracy(%)
SVM	97.44
K-NN	95.92
Random forest	90.80
Logistic Regression	83.50
Decision Tree	86.62

Table 8. Performance of Stacked Ensemble Model for Breast Cancer Dataset.

Base Model	Mean Accuracy (standard deviation)
SVM	0.974(0,005)
K-NN	0.959(0.004)
Random forest	0.908(0.007)
Logistic Regression	0.835(0.011)
Decision Tree	0.866(0.009)
Stacking	0.975(0.009)

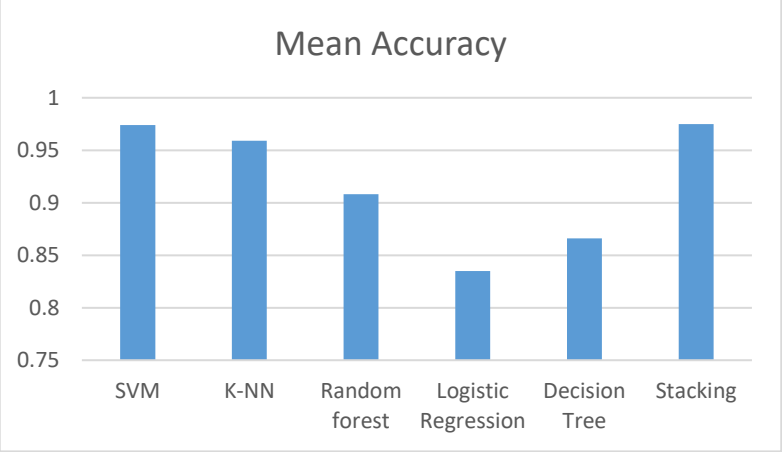

Figure 5. Mean Accuracy for Breast Cancer Dataset

The Figure.5 shows the performance of the individual classifiers and the stacked ensemble classifier for Breast Cancer dataset. To encase the overall performance of the base classifiers in comparison with the Stacked Ensemble model, a Bar Graph plot has been shown in Figure.6. From the simulation results and the Bar plot, it is observed that for all the datasets, the Stacked Ensemble model provides the best performance as compared to the individual base models.

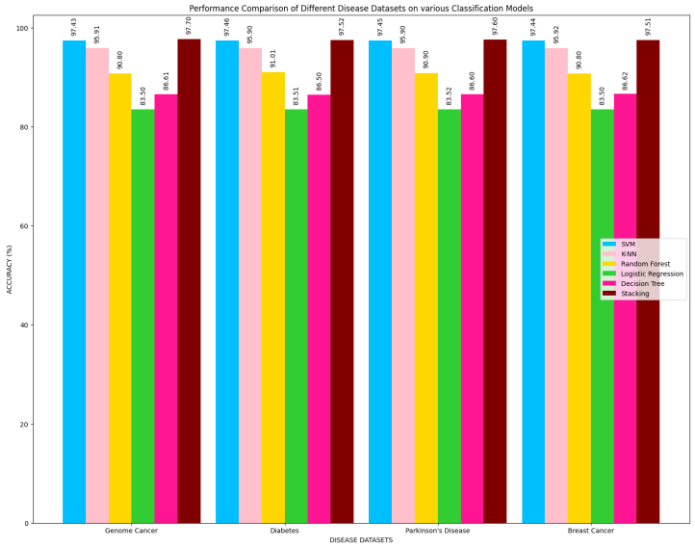

Figure 6. A Comparative Bar Graph Plot of all Classifiers and Proposed Stacking Algorithm used for Various Disease Datasets.

5. Conclusion

We tested for a total of five basic classifiers (SVM, KNN, decision tree, logistic regression, and random forest) and effectively determined the accuracy of each classifier. Moreover, the performance of stacked Ensemble model was also evaluated. An exhaustive performance study of the base models and the ensemble model for four different categories of dataset through a comparative bar graph plot was performed. We have considered the genomic cancer Data set, the Diabetes dataset, the Parkinson's Dataset, and the Breast Cancer dataset. The stacking algorithm exhibited promising results above all other base models, providing a consistent accuracy around and above 97.5% across all disease datasets. These findings emphasize the importance of selecting appropriate classifiers tailored to the dataset

characteristics to achieve optimal accuracy and predictive performance. Also, the advantage of incorporating a Stacking ensemble model for generating better accuracy above other base models, has been highlighted. The results shed light on the strengths and weaknesses of each classifier, providing valuable insights for future applications in medical diagnosis and decision-making processes. However, it is essential to consider that the choice of classifier should be context-dependent and may vary depending on the dataset and the specific problem at hand. Further research and exploration of advanced classifiers and future engineering techniques could lead to even better results in similar machine learning tasks.

References

[1] A. Mahapatra, S. Pahadsingh and T. Kar, “Transfer learning based COVID-19 detection Using Radiological Images,” 2021 IEEE 2nd International Conference on Applied Electromagnetics, Signal Processing, & Communication (AESPC), Bhubaneswar, India, 2021, pp. 1-4,

[2] S. Acharya, T. Kar, U. C. Samal, and P. K. Patra, “Performance Comparison between SVM and LS-SVM for Rice Leaf Disease detection ”, EAI Endorsed Scal Inf Syst, vol. 10, no. 6, Sep. 2023.

[3] S. Mohan, C. Thirumalai and G. Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,” in IEEE Access, vol. 7, pp. 81542-81554,2019,

[4] Mei, Jie, Christian Desrosiers, and Johannes Frasnelli. "Machine learning for the diagnosis of Parkinson's disease: a review of literature." Frontiers in aging neuroscience 13 (2021): 633752.D.

[5] Dahiwade, Dhiraj, Gajanan Patle, and Ektaa Meshram. "Designing disease prediction model using machine learning approach." In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1211-1215. IEEE, 2019.

[6] Ahmadi, Hossein, Marsa Gholamzadeh, Leila Shahmoradi, Mehrbakhsh Nilashi, and Pooria Rashvand. "Diseases diagnosis using fuzzy logic methods: A systematic and meta-analysis review." Computer Methods and Programs in Biomedicine 161 (2018): 145-172.

[7] Sajal, Md Sakibur Rahman, Md Tanvir Ehsan, Ravi Vaidyanathan, Shouyan Wang, Tipu Aziz, and Khondaker Abdullah Al Mamun. "Telemonitoring Parkinson's disease using machine learning by combining tremor and voice analysis." Brain informatics 7, no. 1 (2020): 12.

[8] Zeng, Ling-Li, Liang Xie, Hui Shen, Zhiguo Luo, Peng Fang, Yanan Hou, Beisha Tang, Tao Wu, and Dewen Hu. "Differentiating patients with Parkinson's disease from normal controls using gray matter in the cerebellum." The Cerebellum 16, no. 1 (2017): 151-157.

[9] Swapna, G., R. Vinayakumar, and K. P. Soman. "Diabetes detection using deep learning algorithms." ICT express 4, no. 4 (2018): 243-246.

[10] Shen, Li. "End-to-end training for whole image breast cancer diagnosis using an all convolutional design." arXiv preprint arXiv:1711.05775 (2017).

[11] Asuntha, A., and Andy Srinivasan. "Deep learning for lung Cancer detection and classification." Multimedia Tools and Applications 79, no. 11 (2020): 7731-7762.

[12] Atallah, Rahma, and Amjed Al-Mousa. "Heart disease detection using machine learning majority voting ensemble method." In 2019 2nd international conference on new trends in computing sciences (ictcs), pp. 1-6. IEEE, 2019.

[13] Chang, Victor, Vallabhanent Rupa Bhavani, Ariel Qianwen Xu, and M. A. Hossain. "An artificial intelligence model for heart disease detection using machine learning algorithms." Healthcare Analytics 2 (2022): 100016.

[14] Shruthi, U., V. Nagaveni, and B. K. Raghavendra. "A review on machine learning classification techniques for plant disease detection." In 2019 5th International conference on advanced computing & communication systems (ICACCS), pp. 281-284. IEEE, 2019.

[15] Umbare, R. T., Omkar Ashtekar, Aishwarya Nikhal, Bhagyashri Pagar, and Omkar Zare. "Prediction and Detection of Liver Diseases using Machine Learning." In 2023 IEEE 3rd International Conference on Technology, Engineering, Management for

Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET), pp. 1-6. IEEE, 2023.

[16] Dutta, Supratik, Sibasish Choudhury, Adrita Chakraborty, Sushruta Mishra, and Vikas Chaudhary. "Parkinson Risks Determination Using SVM Coupled Stacking." In *International Conference On Innovative Computing And Communication*, pp. 283-291. Singapore: Springer Nature Singapore, 2023.

Author's biography

Arunya Paul - Arunya Paul is a 3rd Year B.Tech Student, from the Electronics and Telecommunications department, School of Electronics Engineering, KIIT Deemed to be University Bhubaneswar, Odisha, India. His research interests include AI and ML, Tiny ML, IOT and Embedded Systems, RF and Communication technologies.

Dr. Tejaswini Kar - Dr. Tejaswini Kar is currently an Assistant Professor with the School of Electronics Engineering, KIIT deemed to be University, Bhubaneswar, Odisha, India. Her current research areas include signal processing, image processing, video processing, CBIR systems, AI and ML.

Dr. Sasmita Pahadsingh - Dr. Sasmita Pahadsingh, is an Associate Professor in School of Electronics Engineering, KIIT Deemed to be University, Bhubaneswar, Odisha, India. She has more than 15 years of teaching and research experience. Her current research areas include Microwave, MIMO antennas, RF Sensors, and Communication technologies. She is a member of IEEE, ISTE, and Indian Science Congress.

Priya Chandan Satpathy - Assistant Professor & HOD (Electronics & Communication Engineering Department), AIET, Arya Vihar, Bhubaneswar, India.

Biswaranjan Behera - Assistant Professor, AIET, Arya Vihar, Bhubaneswar, India.