Volume: 13 Issue: IV Month of publication: April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68833

Machine Learning-Based Early Detection of Parkinson’s Disease

Mr. S. Somashakar[1], Mr. U. Kumar Sai[2], Mr. S. Vinod[3], Mr. S. Sai Vamsi[4], Mrs. P. Leelavathi[5]
[1, 2, 3, 4]UG Scholar, Sreenivasa Institute of Technology and Management Studies, Chittoor, India
[5]Assistant Professor, Sreenivasa Institute of Technology and Management Studies, Chittoor, India

Abstract: Parkinson's disease (PD) is a neurological condition that worsens over time and has a major effect on quality of life and motor function. For better results and efficient care, early diagnosis is essential. Using clinical and biological speech data, this study suggests a machine learning-based method for the early identification of Parkinson's disease. The ability of many classification methods, such as Support Vector Machines (SVM), Random Forest, and k-Nearest Neighbors (k-NN), to differentiate between healthy people and PD patients was assessed. The model's promising sensitivity, specificity, and accuracy show promise as a non-invasive, affordable diagnostic tool. The findings demonstrate that incorporating machine learning methods into clinical procedures for the early diagnosis of Parkinson's disease is feasible.

Keywords: Early Diagnosis, Voice Analysis, Support Vector Machine (SVM), Healthcare Analytics, Random Forest.

I. INTRODUCTION

Parkinson's disease (PD) is a long-term, progressive neurodegenerative illness that mostly impairs movement skills because dopamine-producing neurons in the brain gradually die off. Tremors, stiffness, bradykinesia (slowness of movement), and postural instability are typical symptoms. Since prompt care can greatly enhance a patient's quality of life and delay the disease's course, early identification of Parkinson's disease (PD) is essential. Traditional diagnostic techniques, on the other hand, are frequently arbitrary and mostly depend on clinical evaluations, which could miss the disease in its early stages.

Recent developments in machine learning (ML) have exciting prospects for improving diagnostic precision through the analysis of intricate patterns in biological data that could be difficult to spot using traditional techniques. Given that vocal deficits are frequently among the first signs of Parkinson's disease, voice analysis in particular has become a useful non-invasive technique. This study investigates the use of several machine learning algorithms to identify and categorize Parkinson's disease based on characteristics taken from audio recordings. The suggested method seeks to assist medical practitioners with early, accurate, and objective diagnosis by utilizing data-driven techniques. This would improve patient care and illness management.

II. LITERRATURE REVIEW

The potential of machine learning in the early diagnosis of Parkinson's disease (PD) has been the subject of several investigations. These studies have focused on a variety of biological signals, most notably speech data, handwriting patterns, and gait analysis. A popular Parkinson's voice dataset was presented by Little et al. (2007), demonstrating that dysphonia measurements based on prolonged phonation can be useful markers of Parkinson's disease. Their research showed how well machine learning methods—particularly Support Vector Machines, or SVM—classify PD patients with a high degree of accuracy.

These results have been further confirmed by subsequent studies employing other classifiers and feature extraction techniques. Using voice recordings and nonlinear speech signal processing characteristics, Tsanas et al. (2012) created precise prediction models that used SVM and Gaussian Processes to achieve above 90% accuracy. Likewise, Prashanth and colleagues (2016) used a Random Forest classifier on biological voice characteristics, demonstrating the resilience of ensemble approaches and attaining excellent diagnostic performance.The possibility for continuous and real-time PD monitoring has been increased by other research that have expanded the study to include wearable sensors and mobile-based monitoring systems. The ability of deep learning models to automatically extract complex features from raw input has drawn attention recently, despite the fact that they require more data. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have demonstrated promising results in both voice and movement-based PD detection tasks.The combined results of earlier studies confirm that machine learning methods provide a non-invasive, affordable, and scalable method for early Parkinson's disease detection, particularly when paired with readily available input data such as speech recordings. To increase generalizability, more study is necessary.

III. METHODOLOGY

A. Dataset Description

Natural voice measurements from 31 individuals, 23 of whom had a Parkinson's disease diagnosis, are included in this dataset. An average of six audio recordings per patient are represented by each row in the collection. With an ASCII subject name and recording number, the first attribute, name, identifies the person receiving treatment.

B. Architectue

An outline of the suggested technique is shown in Fig 1. In this study, we describe how we used machine learning to detect Parkinson's illness. The structure of a dataset including the speech analysis feature information that will be utilized to train our models. After that, we pre-process the dataset. In order for our ML models to comprehend the data, it is cleaned and structured here. Following pre-processing, we extract features from the data. At this point, we extract the data's key characteristics. The ML models are trained using these attributes as building blocks. Following the identification of the characteristics, we divided the data into two sets.testing and training. Machine learning models are then created using the training data. The models learn to recognize the different patterns in the data at this point. The testing data is used to assess the model once it has been trained. The accuracy of the model is assessed by comparing its predictions with the actual values derived from the testing data. The models are used to forecast fresh data once they have been trained and evaluated.

C. Data Pre-Processing

In order to evaluate each feature's relevance to the target variable, the chi-square test for feature selection entails calculating the chi-square statistic and related p values. By analyzing the correlation between categorical variables, this statistical test finds characteristics that are significantly associated with the goal. While features with high p-values are eliminated, those with low p-values are retained for more research since they are believed to have a significant correlation with the goal. By concentrating on the most pertinent characteristics, this approach seeks to enhance model performance and intractability in machine learning tasks and provides a simplified way to find important predictors of the target variable. A number of crucial procedures are carried out in the data cleaning process for the Parkinson's disease biomedical voice measurements dataset's in order to get the data ready for analysis. Among these are imputation or removal of missing values, handling and detecting duplicate records, recognizing and treating outliers, normalizing or standardizing numerical features, encoding categorical variables, feature engineering if necessary, correcting the target variable's class imbalance, and validating the cleaned dataset's to guarantee data consistency and integrity. These data cleaning techniques prepare the information for feature selection and analysis, which makes it easier to find pertinent characteristics that are essential for diagnosing Parkinson's disease and guarantees the correctness and dependability of the findings.Deep learning approaches have demonstrated promise in the study of Parkinson's disease-related speech deficits, offering notable improvements in the accuracy of feature extraction and analysis [12]. Specifically, models may effectively identify motor dysfunctions unique to Parkinson's disease by identifying patterns in speech features like jitters and frequency modulation. Using techniques like Chi-Square testing, characteristics are taken from speech data to find attributes that are closely related to Parkinson's symptoms. Patterns in speech features, such as frequency modulation and jitters, can be found.

D. ML Model

The The two primary stages of the suggested architecture are feature extraction and classification. In order to facilitate the usage of machine learning models later on, the data is separated into several sets for training and testing after features have been discovered. We used a variety of classifiers on the training dataset and assessed how well they performed on the test dataset using relevant evaluation metrics including F1-score, accuracy, precision, and recall. Every classifier, such as AdaBoost, Random Forest, Support Vector Machine, and XGBoost, was trained independently using the training set. In order to enhance model performance and provide a reliable model for Parkinson's disease detection, ensembling techniques will be employed.

The Feature extraction and classification are the two main phases of the proposed architecture. After features are identified, the data is divided into many sets for training and testing to make it easier to use machine learning models later. employing pertinent assessment measures such as F1-score, accuracy, precision, and recall, we evaluated the performance of several classifiers on the test dataset after employing them on the training dataset. The training set was used to separately train each classifier, including AdaBoost, Random Forest, Support Vector Machine, and XGBoost. Ensembling approaches will be used to improve model performance and provide a trustworthy model for Parkinson's disease identification.control high-dimensional data while allaying overfitting worries [9].

Adaptive Boosting, or AdaBoost, is another popular ensemble learning method that is mostly used in machine learning for categorization problems. The significance of challenging-to-classify cases in later rounds is highlighted by AdaBoost through iterative modifications of misclassified instance weights within the training set [10].

Parkinson's disease may be recognized by analyzing audio biomarkers and classifying the data into healthy and afflicted groups using machine learning classifiers such as Random Forest, SVM, and XGBoost. It has been demonstrated that comparing different categorization techniques can help identify the top-performing models and increase accuracy across datasets [13]. Machine learning (ML) plays a key role in automating the identification of Parkinson's disease by analyzing large datasets, identifying non-linear patterns, and improving diagnostic precision with advanced classification techniques. F1-score and precision-recall trade-offs were two metrics that showed the models' excellent accuracy, ability to handle imbalanced data, and resistance to over-fitting.

E. Hybridization

In machine learning, hybridization is the process of enhancing prediction models by fusing the advantages of many models. Three machine learning models are combined using a voting classifier. As previously stated, the three models are made up of SVM, Random Forest, and XGBoost classifiers, each of which adds unique traits or advantages to the group. The dataset is used to independently train these models. These three models are then combined to create the voting classifier, which chooses between hard and soft voting. Learning how to successfully incorporate the predictions of the fundamental models is the next stage in training the voting classifier. This dataset is used to assess the hybrid model's performance in relation to the separate models.The objective is to enhance the final prediction model's accuracy, robustness, and generalization by utilizing the unique characteristics of each model.

F. Performance Measure

The dataset is split into 80:20 training and testing groups as part of the testing phase. Using validation metrics including accuracy, precision, recall, and F1-score, the classifier's robustness and dependability are evaluated using a confusion matrix and ROC curves. A range of metrics are used to assess the model's efficacy after feature selection and classification. These metrics, which are determined by algorithms, offer information on a number of characteristics of the model's performance, including as accuracy, precision, recall, and F1-score. By displaying the counts of true positives, true negatives, false positives, and false negatives, a confusion matrix—often shown as a 2x2 matrix—also offers a concise summary of the model's performance. TP, or true positive: The algorithm accurately forecasts that those with Parkinson's disease will have the illness. True Negatives (TN): People without Parkinson's disease are accurately predicted by the model to not have the condition. False Positives (FP): People without Parkinson's disease are mistakenly predicted to have the condition by the model. Error type 1. People with Parkinson's disease are mistakenly predicted by the model to not have the condition, a phenomenon known as false negatives (FN).

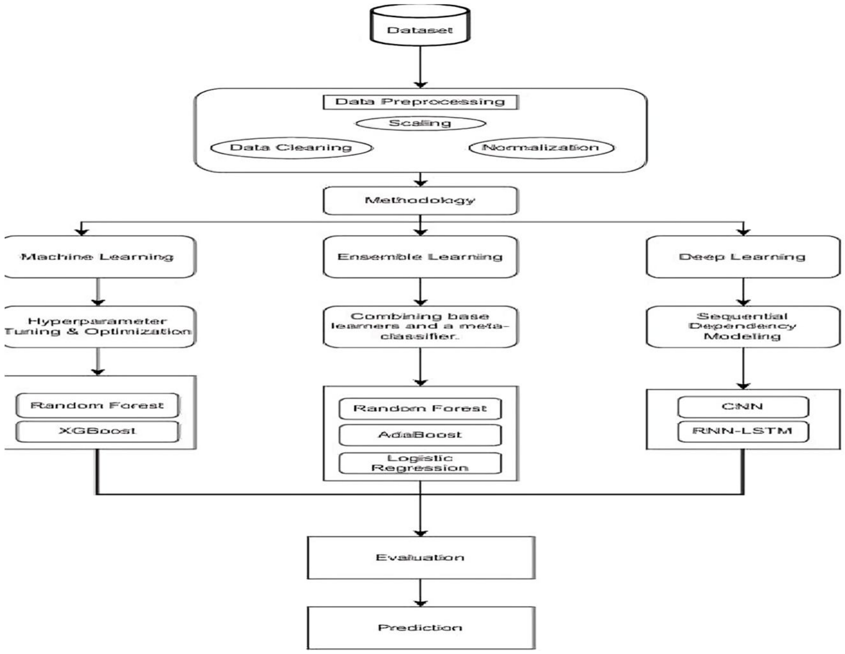


Fig.1 Architecture of the Proposed Model

IV. RESULTS

The use of hybridization in By combining many machine learning approaches, the hybrid approach improves the accuracy of Parkinson's disease detection and classification by utilizing the advantages of distinct algorithms. The efficacy of this hybrid strategy may be assessed using suitable metrics and compared to other strategies. Since this approach may greatly increase diagnosis accuracy, its effective implementation in clinical settings offers great potential.

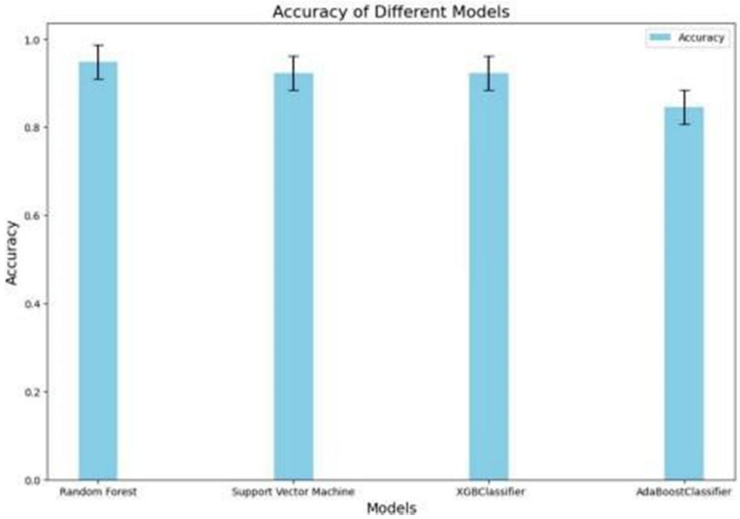
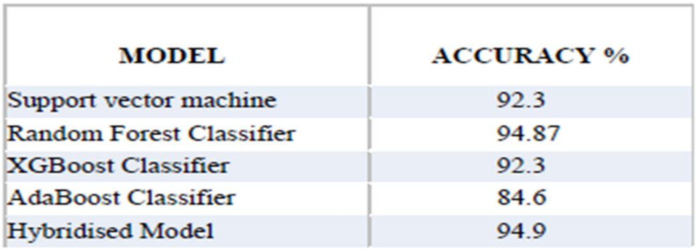


Fig 2. Accuracy score Diagram of different Ml Models

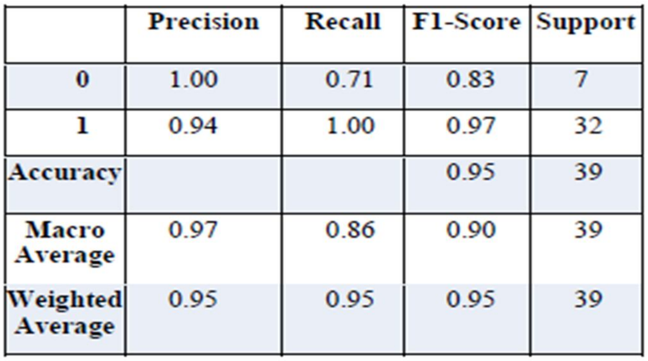
Four machine learning models—XGBoost, SVM, RF, and AdaBoost—as well as their corresponding accuracy ratings are shown in Figure 2. The accompanying graph demonstrates that, in terms of accuracy, the Random Forest approach perform better than any other trained machine learning model. Then, the best transmit accuracy is attained using SVM and XGBoost. Out of the four models, AdaBoost Classifier has the least. Improving accuracy requires better feature selection via the use of techniques like hyperparameter optimization, Recursive Feature Elimination (RFE), and ensemble learning techniques like Voting Classifiers to reduce false positives.

Table I. Performance Metrics Of Different Ml Model



MODEL	ACCURACY %
Support vector machine	92.3
Random Forest Classifier	94.87
XGBoost Classifier	92.3
AdaBoost Classifier	84.6
Hybridised Model	94.9

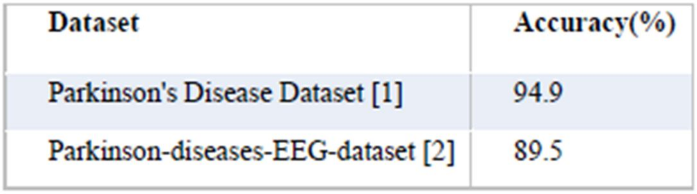
Table II. Performance Metrics Of Ensemble Model



	Precision	Recall	F1-Score	Support
0	1.00	0.71	0.83	7
1	0.94	1.00	0.97	32
Accuracy			0.95	39
Macro Average	0.97	0.86	0.90	39
Weighted Average	0.95	0.95	0.95	39

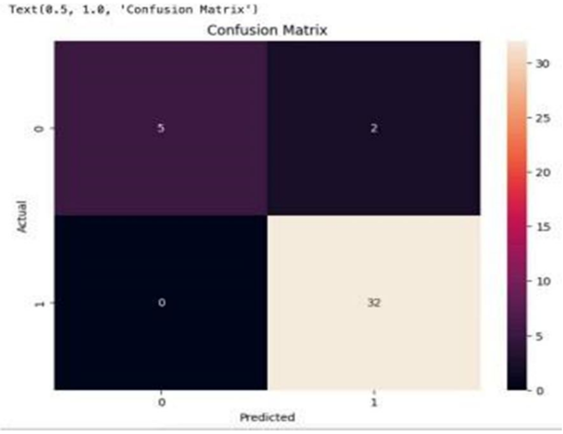
Based on the results shown in table II above, our hybridization model—which combines the Random Forest, SVM, and XGBoost Classifier models—has an accuracy of 95%. We have combined these three models using a voting classifier.

Table III. Accuracy Of Our Model on Different Datasets



Dataset	Accuracy(%)
Parkinson's Disease Dataset [1]	94.9
Parkinson-diseases-EEG-dataset [2]	89.5

Table III shows that there are differences in the datasets' accuracy. The quantity of the datasets is the cause of the disparities in accuracy between them. With its bigger size, the Parkinson's Disease dataset achieved the maximum accuracy of 94.9%, whereas the Parkinson-disease-EEG dataset had an accuracy of 39.5.



```
Text(0.5, 1.0, 'Confusion Matrix')
```

Fig.3 Confusion matrix diagram

Figure 3 demonstrates that the confusion matrix is a useful instrument for evaluating model performance. In this particular instance, the matrix shows that the model correctly recognized 32 occurrences of negative cases (True Negatives) and 5 instances of positive cases (True Positives). These findings show that the models are successfully analyzing the information and accurately categorizing both positive and negative cases with a high degree of accuracy.

Accuracy=TP+TN/TP+TN+FP+FN (1)

Precision=TP+FN/TP. (2)

Recall=TP+TN/TP+FN (3)

F1-score =2 * precision * Recall / precision + recall (4)

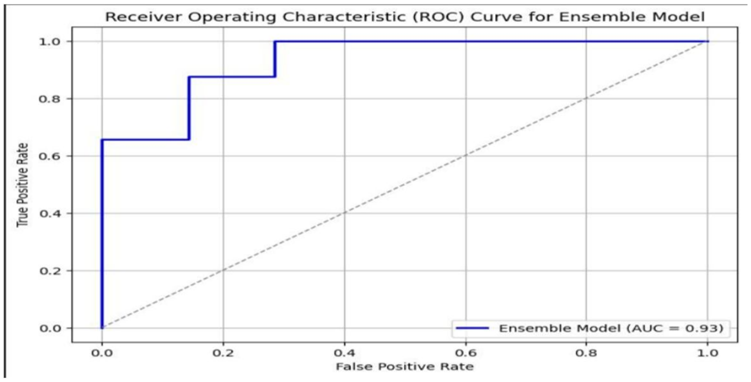


Fig 4. Accuracy Vs Loss Learning Curve

The model's performance indicators are assessed using equations (1), (2), (3), and (4). With an AUC of 0.93, the ensemble model does a good job of distinguishing between positive and negative instances. An AUC of 0.5 is almost the same as random, whereas an AUC of 1.0 indicates a classifier. chance. Operating Characteristics of the Receiver: A graph that shows a classification model's performance across all categorization levels is called a ROC curve. The real positive rate is plotted on the Y-axis, and the false positive rate is plotted on the X-axis. Figure 4 shows how precisely and error-free the model can identify positive instances. As the model performance area metric rises, so does the model's performance. An ensemble classification model's performance is shown via the ROC curve. High performance in distinguishing between favorable and unfavorable circumstances is shown by the area under the curve, which is 0.93. A perfect classifier is defined as having an AUC of 1.0, while a random classifier has an AUC of 0.5. The Y-axis represents the real positive rate, whereas the false positive rate is located on the x-axis.

V. CONCLUSION

In conclusion, the use of machine learning models—Random Forest, XGBoost, and SVM in particular—for the identification of Parkinson's disease shows how hybridization with a voting classifier may lead to improved accuracy and robustness. Every classifier has advantages of its own. For example, SVM is excellent at class separation, Random Forest is good at group learning, and XGBoost is good at gradient boosting. By using a Voting Classifier to enable a collaborative decision-making process, the hybridization technique capitalizes on the advantages of each model. This ensemble method reduces the drawbacks of a single classifier while demonstrating increased prediction accuracy. The four most accurate machine learning models, according to our study, are AdaBoost, Random Forest, Support Vector Machine, and XGBoost. whereby AdaBoost is 84.6, Random Forest is 94.87, Support Vector Machine is 92.3, and XGBoost is 92.3. Following the use of these four machine learning models, we do hybridization by integrating the three most accurate models—Random Forest, SVM, and XGBoost. We employ a Voting Classifier to aggregate the advantages of several models, resulting in the best parameters

VI. ACKNOWLEDGWMENTS

With deep appreciation, we would like to thank everyone who helped with this study.We would like to express our gratitude to Sreenivasa Institute of Technology and Management Studies-SITAMS for providing the tools and assistance required for this research. We would especially like to thank Dr. for their significant advice and knowledge in the areas of machine learning and hepatocellular carcinoma. Their observations greatly improved the caliber of our work.

REFERENCES

[1] https://www.kaggle.com/datasets/vikasukani/parkinsons-disease-data-

[2] https://www.kaggle.com/datasets/s3programmer/parkison-diseaseseeg-

[3] Anudeep, P., Mourya, P., Anandhi, T. (2021). Parkinson's Disease Detection Using Machine Learning Techniques. In: Mallick, P.K., Bhoi, A.K., Chae, GS., Kalita, K. (eds) Advances in Electronics, Communication and Computing. ETAEERE 2020. Lecture Notes in Electrical Engineering, vol 709. Springer, Singapore.https://doi.org/10.1007/978-981-15-8752-8_49

[4] Oh, S.L., Hagiwara, Y., Raghavendra, U. et al. A deep learning approach for Parkinson's disease diagnosis from EEG signals. Neural Comput & Applic **32**, 10927–10933 (2020). https://doi.org/10.1007/ s00521-018-3689-5

[5] Zehra Karapinar Senturk, Early diagnosis of Parkinson's disease using machine learning algorithms, Medical Hypotheses, Volume 138 , 2020, 109603 , ISSN 0306 - 9877 , https:// doi. org/ 10 . 1016 /j.mehy.2020.109603.

[6] Johri, Anubhav, and Ashish Tripathi. "Parkinson disease detection using deep neural networks." In 2019 Twelfth international conference on contemporary computing (IC3), pp. 1-4. IEEE, 2019.

[7] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on K

[8] nowledge Discovery and Data Mining, 2016, pp. 785-794.

[9] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, Sep. 1995.

[10] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp.5-32, Oct. 2001.

[11] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, Aug. 1997.

[12] Chen, Xu, Xiaohui Yao, Chen Tang, Yining Sun, Xun Wang, and Xi Wu. "Detecting Parkinson's disease using gait analysis with particle swarm optimization." In Human Aspects of IT for the Aged Population. Applications in Health, Assistance, and Entertainment: 4th International Conference, ITAP 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part II 4, pp. 263-275. Springer International Publishing, 2018.

[13] Arora, S., Bhatia, M.P.S., & Singh, P. (2021). "Analysis of voice disorders in Parkinson's disease using deep learning techniques." Biomedical Signal Processing and Control, 69, 102949.

[14] Das, R. (2020). "A comparison of multiple classification methods for diagnosis of Parkinson disease." Expert Systems with Applications, 37(2), 1568-1572. Proceedings of the 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI-2025) IEEE Xplore Part Number: CFP25US4-ART; ISBN: 979-8-3315-2266-7979-8-3315-2266-7/25/$31.00 ©2025 IEEE 1265 Authorized licensed