**AIMM 2024 Workshop Group Activity: Addressing Data Imbalance in Machine Learning**

**Objective:**

To explore various techniques for handling data imbalance in machine learning and train AI/machine learning models effectively using these methods.

**Activity Overview:**

The class will be divided into 5 groups. Each group will focus on a specific method to address data imbalance and will train a machine learning model using this method. The goal is to compare the effectiveness of different techniques before and after data balancing and understand their impact on model performance.

**Group Assignments:**

1. **Group 1**: Random Oversampling

- Task: Apply random oversampling to balance the dataset.

- Model: Train ten models and compare

2. **Group 2**: SMOTE (Synthetic Minority Over-sampling Technique)

- Task: Implement SMOTE to generate synthetic samples for the minority class.

- Model: Train ten models and compare

3. **Group 3**: Random Undersampling

- Task: Apply random undersampling to reduce the size of the majority class.

- Model: Train ten models and compare

4. **Group 4**: ADASYN (Adaptive Synthetic Sampling)

- Task: Implement ADASYN to create synthetic samples focusing on hard-to-classify examples.

- Model: Train ten models and compare

5. **Group 5**: Hybrid Methods (Combination of Over and Undersampling)

- Task: Apply a hybrid approach by combining oversampling and undersampling techniques.

- Model: Train ten models and compare

**Instructions:**

1**. Data Preparation:**

- Use the provided imbalanced malaria dataset.

- Apply your group’s assigned method to address the imbalance.

- Study more about the imbalance method and the model selected

2. **Model Training:**

- Split the dataset into training and testing sets (75/25 split).

- Train the assigned machine learning model using the processed data.

- Evaluate the model using appropriate metrics (e.g., Accuracy/Balanced Accuracy, Recall, F1-Score, ROC-AUC etc).

- Compare your method or model with one or two other methods.

3. **Presentation:**

- Prepare a 5-minute presentation summarizing your approach, the challenges faced, and the results obtained.

- Highlight the impact of the imbalanced data handling technique on model performance.

- Engage the possibility of developing your ideas into a collaborative research publication.

4. **Discussion:**

- We shall engage in a class-wide discussion to compare the effectiveness of the different techniques and models.

**Expected Outcomes:**

- A deeper understanding of various data imbalance handling techniques.

- Insight into how different methods affect machine learning model performance.

- Enhanced collaboration and problem-solving skills among participants.

**Note:** Be creative and collaborative! Each group's work contributes to the collective learning experience of this workshop. We shall conduct a quiz on the third day of this workshop before issuing your certificates.