Pseudo code

1. **Data Collection**

a. Gather a dataset of URLs labeled as 'phishing' or 'legitimate'.

b. Ensure the dataset is balanced to prevent bias.

2. **Feature Extraction**

a. For each URL in the dataset:

i. Extract lexical features (e.g., URL length, presence of special characters).

ii. Extract host-based features (e.g., domain age, WHOIS information).

iii. Extract content-based features (e.g., HTML content, presence of forms).

3. **Data Preprocessing**

a. Handle missing or inconsistent data.

b. Normalize or standardize features as required.

c. Split the dataset into training and testing subsets (e.g., 80% training, 20% testing).

4. **Model Selection**

a. Choose appropriate machine learning algorithms (e.g., Random Forest, SVM, XGBoost).

b. Consider ensemble methods for improved performance.

5. **Model Training**

a. Train the selected models using the training dataset.

b. Perform hyperparameter tuning to optimize model performance.

6. **Model Evaluation**

a. Evaluate the trained models on the testing dataset using metrics such as accuracy, precision, recall, and F1-score.

b. Select the model with the best performance metrics.

7. **Model Deployment**

a. Integrate the chosen model into the Scamify system.

b. Develop an interface for users to input URLs for evaluation.

8. **Prediction**

a. For a given input URL:

i. Extract the same features as in the training phase.

ii. Apply the trained model to predict whether the URL is 'phishing' or 'legitimate'.

iii. Return the prediction result to the user.

9. **Continuous Improvement**

a. Regularly update the dataset with new URLs to capture emerging phishing techniques.

b. Retrain and validate the model periodically to maintain detection accuracy.