In [20]:

```
pip install pyarrow
```

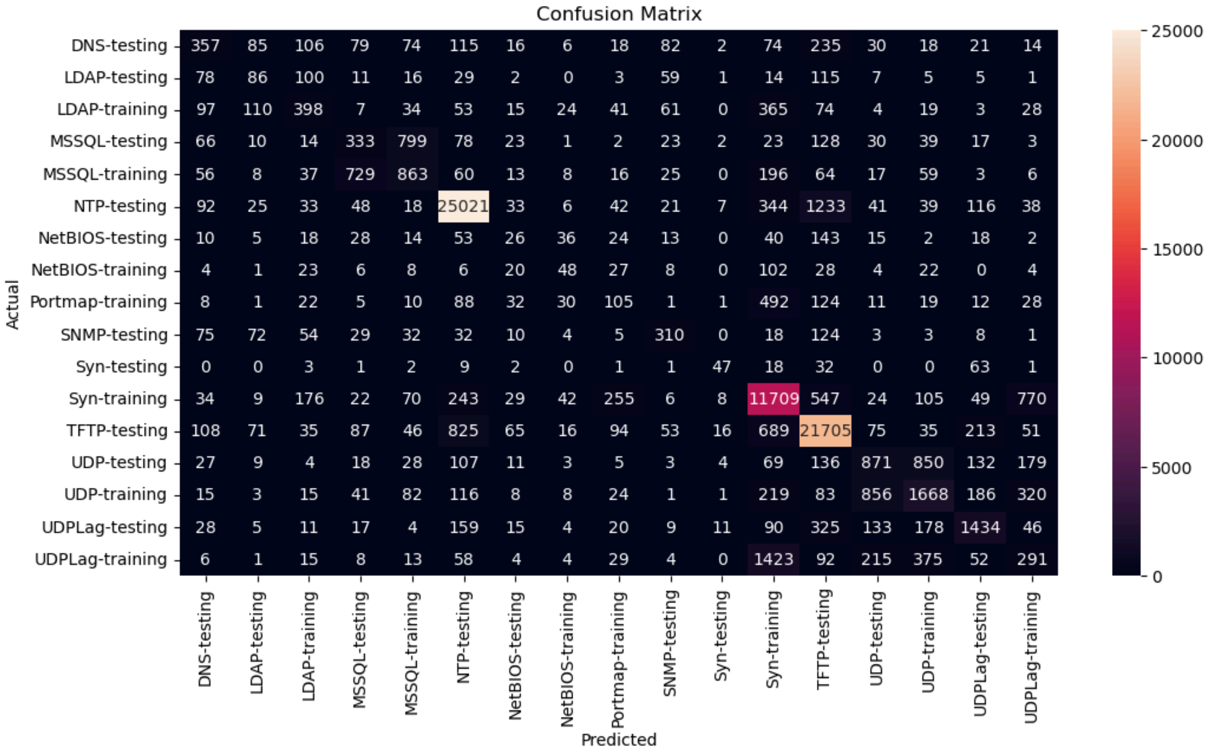
```
Requirement already satisfied: pyarrow in c:\users\sahrv\anaconda3\lib\site-packages
(16.1.0)
Requirement already satisfied: numpy>=1.16.6 in c:\users\sahrv\anaconda3\lib\site-pa
ckages (from pyarrow) (1.26.4)
Note: you may need to restart the kernel to use updated packages.
```

In [24]:

```
# 📦 1. Imports
import pandas as pd
import os
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

# 📂 2. Load all Parquet files
data_dir = 'CICDDoS2019/'  # Update path if needed

all_data = []

for file in os.listdir(data_dir):
    if file.endswith('.parquet'):
        print(f"Loading: {file}")
        df = pd.read_parquet(os.path.join(data_dir, file))
        df['attack_type'] = file.replace('.parquet', '')
        all_data.append(df)

print(f"\n✅ Total files loaded: {len(all_data)}")

# 🧹 3. Combine and clean data
df = pd.concat(all_data, ignore_index=True)
print("✅ Combined Data Shape:", df.shape)

# Drop rows with all NaNs or where labels are missing
df.dropna(how='all', inplace=True)
df.dropna(subset=['attack_type'], inplace=True)

# Keep only numeric columns + label
df = df.select_dtypes(include=['float64', 'int64']).copy()
df['attack_type'] = pd.concat([d['attack_type'] for d in all_data], ignore_index=Tr

# 🎯 4. Features & Target
X = df.drop('attack_type', axis=1)
y = df['attack_type']

# Encode labels
le = LabelEncoder()
y_encoded = le.fit_transform(y)

# ✂️ 5. Train/Test Split
X_train, X_test, y_train, y_test = train_test_split(X, y_encoded, test_size=0.2, ra
```

```python
# 🤖 6. Train Random Forest
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# 📊 7. Evaluate
y_pred = model.predict(X_test)

print("\n✅ Accuracy:", accuracy_score(y_test, y_pred))
print("\n📄 Classification Report:\n", classification_report(y_test, y_pred, targe

# 🌀 8. Confusion Matrix
plt.figure(figsize=(12, 6))
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d',
            xticklabels=le.classes_, yticklabels=le.classes_)
plt.title("Confusion Matrix")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()
```

```
Loading: DNS-testing.parquet
Loading: LDAP-testing.parquet
Loading: LDAP-training.parquet
Loading: MSSQL-testing.parquet
Loading: MSSQL-training.parquet
Loading: NetBIOS-testing.parquet
Loading: NetBIOS-training.parquet
Loading: NTP-testing.parquet
Loading: Portmap-training.parquet
Loading: SNMP-testing.parquet
Loading: Syn-testing.parquet
Loading: Syn-training.parquet
Loading: TFTP-testing.parquet
Loading: UDP-testing.parquet
Loading: UDP-training.parquet
Loading: UDPLag-testing.parquet
Loading: UDPLag-training.parquet

✅ Total files loaded: 17
✅ Combined Data Shape: (431371, 79)

✅ Accuracy: 0.756557519559548

📄 Classification Report:
                  precision    recall  f1-score   support

     DNS-testing       0.34      0.27      0.30      1332
    LDAP-testing       0.17      0.16      0.17       532
   LDAP-training       0.37      0.30      0.33      1333
   MSSQL-testing       0.23      0.21      0.22      1591
  MSSQL-training       0.41      0.40      0.40      2160
     NTP-testing       0.92      0.92      0.92     27157
 NetBIOS-testing       0.08      0.06      0.07       447
NetBIOS-training       0.20      0.15      0.17       311
Portmap-training       0.15      0.11      0.12       989
    SNMP-testing       0.46      0.40      0.42       780
     Syn-testing       0.47      0.26      0.34       180
    Syn-training       0.74      0.83      0.78     14098
    TFTP-testing       0.86      0.90      0.88     24184
     UDP-testing       0.37      0.35      0.36      2456
    UDP-training       0.49      0.46      0.47      3646
  UDPLag-testing       0.61      0.58      0.59      2489
 UDPLag-training       0.16      0.11      0.13      2590

        accuracy                           0.76     86275
       macro avg       0.41      0.38      0.39     86275
    weighted avg       0.74      0.76      0.75     86275
```



	DNS-testing	LDAP-testing	LDAP-training	MSSQL-testing	MSSQL-training	NTP-testing	NetBIOS-testing	NetBIOS-training	Portmap-training	SNMP-testing	Syn-testing	Syn-training	TFTP-testing	UDP-testing	UDP-training	UDPLag-testing	UDPLag-training
DNS-testing	357	85	106	79	74	115	16	6	18	82	2	74	235	30	18	21	14
LDAP-testing	78	86	100	11	16	29	2	0	3	59	1	14	115	7	5	5	1
LDAP-training	97	110	398	7	34	53	15	24	41	61	0	365	74	4	19	3	28
MSSQL-testing	66	10	14	333	799	78	23	1	2	23	2	23	128	30	39	17	3
MSSQL-training	56	8	37	729	863	60	13	8	16	25	0	196	64	17	59	3	6
NTP-testing	92	25	33	48	18	25021	33	6	42	21	7	344	1233	41	39	116	38
NetBIOS-testing	10	5	18	28	14	53	26	36	24	13	0	40	143	15	2	18	2
NetBIOS-training	4	1	23	6	8	6	20	48	27	8	0	102	28	4	22	0	4
Portmap-training	8	1	22	5	10	88	32	30	105	1	1	492	124	11	19	12	28
SNMP-testing	75	72	54	29	32	32	10	4	5	310	0	18	124	3	3	8	1
Syn-testing	0	0	3	1	2	9	2	0	1	1	47	18	32	0	0	63	1
Syn-training	34	9	176	22	70	243	29	42	255	6	8	11709	547	24	105	49	770
TFTP-testing	108	71	35	87	46	825	65	16	94	53	16	689	21705	75	35	213	51
UDP-testing	27	9	4	18	28	107	11	3	5	3	4	69	136	871	850	132	179
UDP-training	15	3	15	41	82	116	8	8	24	1	1	219	83	856	1668	186	320
UDPLag-testing	28	5	11	17	4	159	15	4	20	9	11	90	325	133	178	1434	46
UDPLag-training	6	1	15	8	13	58	4	4	29	4	0	1423	92	215	375	52	291