Below is a **clean, examiner-level comparison** between
(A) the PAPER you provided and **(B) the CODE you just shared**, after **fully scanning the code end-to-end**.

I'll clearly separate:

1. **What the paper contains**
2. **What your code contains**
3. **What you added BEYOND the paper**
4. **What matches exactly**
5. **Final academic verdict (important)**

1️⃣ What the PAPER contains (baseline)

From the document you gave earlier, the paper **officially includes only**:

✅ Data & Preprocessing (Paper)

- Dataset from **Kaggle**
- URL text–based analysis
- RegexpTokenizer
- Stopword removal
- Snowball Stemmer
- Basic URL cleaning

✅ Feature Extraction (Paper)

- **TF-IDF**
- **Character n-grams**
- Limited handcrafted URL features (basic ones)
- No entropy, no advanced ratios

✅ Models Mentioned / Used (Paper)

- **Logistic Regression**

- **Naive Bayes**
- CNN is **mentioned**, but:
  - No architecture
  - No layers
  - No training procedure
  - No evaluation details

👉 **CNN is referenced, not deeply implemented**

✅ **Evaluation (Paper)**

- Accuracy only (≈ **96.6%**)
- No AUC
- No F1-score
- No ensemble
- No comparison table

❌ **What paper does NOT include**

- RNN
- LSTM
- GRU
- Hybrid models
- SMOTE
- Random Forest
- Gradient Boosting
- Hyperparameter tuning
- Model checkpointing
- Real-time detection class

2️⃣ What YOUR CODE contains (fully verified)

Your code is a **much more advanced system**.

🔹 **Traditional ML (More than paper)**

- Logistic Regression ✅ (paper)
- Naive Bayes ✅ (paper)
- **Random Forest ❌ (extra)**
- **Gradient Boosting ❌ (extra)**
- **SMOTE ❌ (extra)**
- **Hyperparameter tuning ❌ (extra)**

🔹 **Feature Engineering (Far beyond paper)**

Your code adds **extensive handcrafted features**, such as:

- URL length, hostname length
- Subdomain count
- Entropy calculation
- Digit / letter / special-char ratios
- IP-based URLs
- Phishing keyword frequency
- Suspicious TLD detection
- URL shortening detection
- Consecutive character analysis

📌 **Paper uses only basic features**
📌 **Your feature extractor is enterprise-grade**

🔹 **Deep Learning (NOT in paper)**

✅ **CNN (Paper only mentions, you fully implement)**

- Embedding
- Conv1D (multiple layers)
- BatchNormalization
- Dropout
- GlobalMaxPooling
- Adam optimizer
- EarlyStopping
- ReduceLROnPlateau
- ModelCheckpoint
- Accuracy + AUC

➡️ **Paper does NOT explain any of this**

🔹 RNN Models (COMPLETELY NEW)

Your code includes **full RNN support**:

Model	In Paper	In Code
RNN	❌	✅
LSTM	❌	✅
GRU	❌	✅
Bidirectional	❌	✅

These are:

- Fully trained
- Evaluated
- Saved
- Used in prediction

📌 **Paper has ZERO RNN discussion**

🔹 Hybrid CNN–RNN (NOT in paper)

You implemented:

- CNN branch
- LSTM branch
- Feature concatenation
- Joint learning

❌ **Paper does not mention hybrid architectures**

🔹 Ensemble Learning (NOT in paper)

Your code:

- Combines **LR + NB + RF + GB + CNN + LSTM + GRU + Hybrid**
- Averages probabilities
- Outputs confidence scores

❌ **Paper uses single-model evaluation only**

🔹 Evaluation Metrics (More than paper)

Metric	Paper	Code
Accuracy	✅	✅
F1-Score	❌	✅
ROC-AUC	❌	✅
Confusion Matrix	❌	✅
Model comparison table	❌	✅

🔹 Deployment & Engineering (NOT in paper)

Your code includes:

- PhishingURLDetector class
- Model loading system
- Real-time URL prediction
- Feature explanation
- Confidence score
- Saved .keras and .pkl models

📌 **Paper is purely academic, not deployable**

3️⃣ Side-by-Side Summary (VERY IMPORTANT)

Area	Paper	Your Code
TF-IDF + n-grams	✅	✅
Logistic Regression	✅	✅
Naive Bayes	✅	✅
CNN (detailed)	❌	✅
RNN / LSTM / GRU	❌	✅
Hybrid CNN-RNN	❌	✅
Random Forest	❌	✅
Gradient Boosting	❌	✅
SMOTE	❌	✅
Advanced handcrafted features	❌	✅

Area	Paper	Your Code
Ensemble learning	❌	✅
AUC / F1 metrics	❌	✅
Real-time detector	❌	✅

4️⃣ Final Academic Verdict 🎓 (read carefully)

✔ Your code is SIGNIFICANTLY MORE ADVANCED than the paper

❌ Your paper does NOT justify most of your code

This means:

- **Technically:** Excellent work
- **Academically:** Mismatch if submitted together

5️⃣ What to say if examiner asks (MEMORIZE)

**“The paper presents a baseline ML-based phishing URL detection system.
The additional models (CNN, LSTM, GRU, hybrid, and ensemble) were implemented as an extension to improve robustness and real-world applicability.”**

6️⃣ What I can do next (your choice)

I can:

1. ✍️ Rewrite the **paper methodology** to fully match this code
2. 🧹 Reduce the **code to exactly match the paper**
3. 🛡️ Prepare **viva / defense Q&A**
4. 📊 Generate **updated results tables & graphs**

Just tell me what you want to do next.