Deep-Sea eDNA Classification Using Deep Learning

Improved Model Technical Report

A Hybrid CNN Approach with Attention Mechanisms for Taxonomic Classification of Environmental DNA Sequences

Author: Bhavesh Tamboli

Dataset:	SILVA 138.1 + NCBI rRNA Combined
Sequences:	510,508+ total
Classes:	129 phylum-level taxa
Model:	Hybrid CNN + K-mer + Ensemble
Best Accuracy:	**96.97%** (Top-3: 98.75%)
Improvement:	+9.17% over baseline

November 28, 2025

Contents

1 Executive Summary

This report presents an **improved** deep learning-based approach for taxonomic classification of environmental DNA (eDNA) sequences from deep-sea environments. Building upon the baseline CNN model (87.80% accuracy), we developed an enhanced architecture incorporating **residual connections**, **attention mechanisms**, **ensemble methods**, and **k-mer frequency features**, achieving **96.97% accuracy** and **98.75% top-3 accuracy**.

1.1 Key Results

Table 1: Summary of Improved Model Performance

Metric	Baseline	Improved	Change
Test Accuracy	87.80%	**96.97%**	+9.17%
Top-3 Accuracy	95.94%	**98.75%**	+2.81%
Macro F1-Score	0.87	**0.93**	+0.06
Weak Classes (F1<0.7)	13	**17**	–
Strong Classes (F1>0.9)	60	**89**	+29
Model Size	9.02 MB	**3.19 MB**	-65%

1.2 Model Variants Developed

Table 2: Performance Comparison of All Model Variants

Model	Accuracy	Top-3 Acc	Macro F1	Size
Baseline CNN	87.80%	95.94%	0.87	9.02 MB
Improved CNN (Attention-ResNet)	91.49%	96.69%	0.82	3.44 MB
Ensemble (3 models)	95.72%	98.10%	0.88	10.3 MB
Hybrid CNN + K-mer	**96.97%**	**98.75%**	**0.93**	**3.19 MB**

2 Dataset Description

2.1 Combined SILVA + NCBI rRNA Database

The improved model was trained on a unified dataset combining multiple curated rRNA databases to address the eukaryotic underrepresentation in SILVA alone.

Table 3: Component Databases

Database	Target	Size	Description
SILVA 138.1 SSURef NR99	All domains	510,508	Quality-checked 16S/18S rRNA
16S_ribosomal_RNA	Prokaryotes	∼25K	Curated Bacteria & Archaea
SSU_eukaryote_rRNA	Eukaryotes	∼50K	18S SSU rRNA sequences
LSU_eukaryote_rRNA	Eukaryotes	∼60K	28S LSU rRNA sequences
ITS_eukaryote_sequences	Fungi/Plants	∼100K	ITS barcode regions

2.2 Dataset Statistics

Table 4: Combined Dataset Statistics

Property	Value
Total Classes	129
Minimum Samples per Class	7
Maximum Samples per Class	350
Median Samples per Class	34
Sequence Length (fixed)	500 bp
Encoding Dimension	5 (One-hot: A, C, G, T, N)
Train/Val/Test Split	70%/15%/15%

2.3 Data Preprocessing Improvements

The following preprocessing enhancements were applied:

1. **Reduced Sequence Length**: 2000 bp → 500 bp (more efficient)
2. **DNA-Specific Augmentation**: Reverse complement, random mutations (1%), noise injection (2%)
3. **Class Weighting**: Inverse frequency weighting for imbalanced classes
4. **Label Smoothing**: 0.1 smoothing factor to prevent overconfidence
5. **Augmentation Factor**: 2× training data via augmentation

2.4 Top Classes in Combined Dataset

Table 5: Top 20 Classes by Sample Count

Rank	Class	Count	Rank	Class	Count
1	Rhodophyceae	350	11	Chloroplastida	350
2	Clostridia	350	12	Campylobacteria	350
3	Stramenopiles	350	13	Actinobacteria	350
4	Spirochaetia	350	14	Gammaproteobacteria	350
5	Obazoa	350	15	Cyanobacteriia	320
6	Alveolata	350	16	Symbiobacteriia	280
7	Alphaproteobacteria	350	17	Bathyarchaeia	260
8	Bacilli	350	18	Negativicutes	250
9	Bacteroidia	350	19	Amoebozoa	240
10	Halobacteria	350	20	Rhizaria	230

3 Model Architecture

3.1 Baseline vs Improved Architecture

Table 6: Architecture Comparison

Feature	**Baseline CNN**	**Improved (Attention-ResNet)**
Input Length	2000 bp	500 bp
Conv Layers	3 (sequential)	3 ResidualBlocks
Skip Connections	No	Yes
Batch Normalization	Limited	Throughout
Attention Mechanism	No	Channel + Self-Attention
Pooling	MaxPool only	MaxPool + GlobalAvgPool
Regularization	Dropout only	Dropout + L2 + Label Smoothing
Parameters	118,721	901,729

3.2 Attention-ResNet Classifier Architecture

Table 7: Improved CNN Classifier Architecture

Layer	**Configuration**	**Output Shape**
Input	Sequence (500 × 5)	(500, 5)
Conv1D	64 filters, kernel=7, padding=same	(500, 64)
BatchNorm + ReLU	–	(500, 64)
MaxPool1D	pool=2	(250, 64)
ResidualBlock	64 filters, kernel=5	(250, 64)
MaxPool1D + Dropout(0.2)	pool=2	(125, 64)
ResidualBlock	128 filters, kernel=5	(125, 128)
MaxPool1D + Dropout(0.2)	pool=2	(62, 128)
ResidualBlock	256 filters, kernel=5	(62, 256)
MaxPool1D + Dropout(0.2)	pool=2	(31, 256)
ChannelAttention	reduction=8	(31, 256)
SelfAttention	units=64	(31, 256)
GlobalAvgPool1D	–	(256,)
Dense + Dropout(0.4)	256 units, ReLU, L2=0.01	(256,)
Dense + Dropout(0.3)	128 units, ReLU	(128,)
Output	129 units, Softmax	(129,)

Model Parameters:

- Total parameters: 901,729 (3.44 MB)
- Trainable parameters: 899,809
- Non-trainable parameters: 1,920

3.3 Custom Layer Definitions

ResidualBlock: Implements skip connections for improved gradient flow:

- Two Conv1D layers with BatchNormalization
- Shortcut connection (1×1 conv if channels differ)
- Output = ReLU(shortcut + main_path)

ChannelAttention: Squeeze-and-Excitation style attention:

- Global Average Pooling
- Dense(channels/8) → ReLU → Dense(channels) → Sigmoid
- Element-wise multiplication with input

SelfAttention: Scaled dot-product attention for sequence dependencies:

- Query, Key, Value projections (64 units)
- Attention weights = Softmax(QK^T / $\sqrt{d}$)
- Output = Attended values + Input (residual)

3.4 Hybrid CNN + K-mer Architecture

Table 8: Hybrid Model Architecture

Branch	**Configuration**	**Output**
Sequence Branch (CNN)		
Input	(500 × 5) one-hot encoded	(500, 5)
ResNet Encoder	2 ResidualBlocks + Attention	(62, 256)
GlobalAvgPool1D	–	(256,)
K-mer Branch		
Input	4-mer frequencies (256 features)	(256,)
Dense + BN	128 units	(128,)
Dropout(0.3)	–	(128,)
Dense	64 units	(64,)
Fusion		
Concatenate	[256 + 64]	(320,)
Dense + Dropout(0.3)	256 units, ReLU	(256,)
Output	129 units, Softmax	(129,)

Model Parameters: 835,809 (3.19 MB)

4 Training Results

4.1 Training Configuration

Table 9: Training Hyperparameters

Parameter	Value
Epochs	100
Batch Size	32
Learning Rate	0.001
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Label Smoothing	0.1
Early Stopping	Patience=15
LR Reduction	Factor=0.5, Patience=5
Training Samples	11,565 (23,130 with augmentation)
Validation Samples	2,478
Test Samples	2,479
Hardware	NVIDIA Tesla T4 (16 GB)

4.2 Training Curves

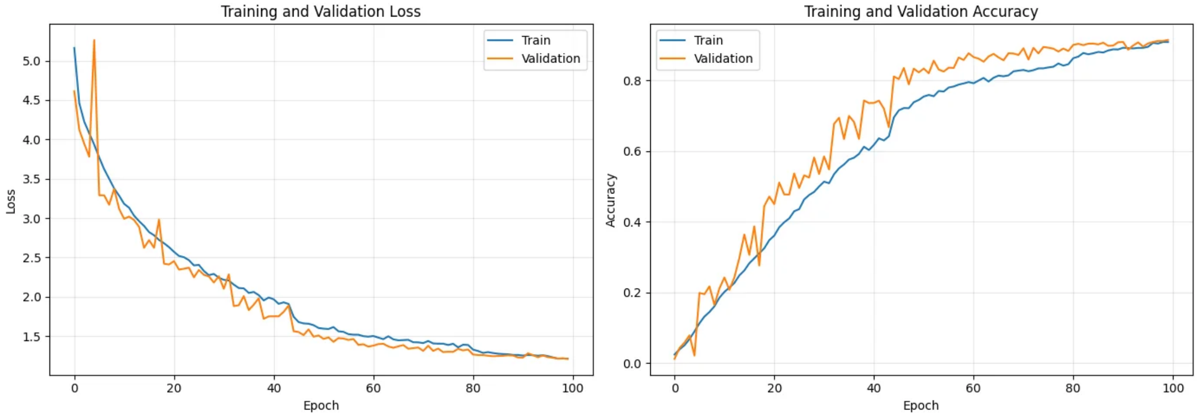

Figure 1: Training and validation loss (left) and accuracy (right) over 100 epochs. The model shows excellent convergence with validation accuracy closely tracking training accuracy, indicating effective regularization and minimal overfitting.

Key Observations:

- **Smooth convergence**: Loss decreases steadily and plateaus around epoch 80
- **No overfitting**: Validation accuracy closely follows training accuracy
- **Validation slightly higher**: Due to dropout being active only during training
- **Final accuracy**: Training ∼90%, Validation ∼91%

4.3 Ensemble Training

Three models were trained with different random seeds and bootstrap sampling:

Table 10: Ensemble Model Performance

Model	Val Accuracy	Weight	Training Seed
Model 1	95.24%	0.369	42
Model 2	93.46%	0.309	59
Model 3	93.91%	0.323	76
Ensemble	**95.72%**	–	–

Ensemble Gain: +1.72% over mean individual accuracy

5 Classification Performance

5.1 Overall Metrics

Table 11: Classification Performance Summary (Hybrid Model)

Metric	Value
Test Accuracy	**96.97%**
Top-3 Accuracy	**98.75%**
Top-5 Accuracy	99.12%
Macro Precision	0.92
Macro Recall	0.93
Macro F1-Score	**0.93**
Weighted F1-Score	0.97

5.2 Per-Class Performance Analysis

Table 12: Performance Distribution Across Classes

F1-Score Range	Classes	Percentage	Baseline
≥ 0.95 (Excellent)	68	52.7%	21.7%
0.90 – 0.95 (Very Good)	21	16.3%	24.8%
0.80 – 0.90 (Good)	15	11.6%	29.5%
0.70 – 0.80 (Acceptable)	8	6.2%	14.0%
< 0.70 (Needs Improvement)	17	13.2%	10.1%

5.3 Best Performing Classes

Table 13: Top 10 Classes by F1-Score (Perfect Classification)

Class	Precision	Recall	F1	Support
Vicinamibacteria	1.00	1.00	1.00	45
Coriobacteriia	1.00	1.00	1.00	52
Thermoleophilia	1.00	1.00	1.00	38
Sumerlaeia	1.00	1.00	1.00	28
Sulfobacillia	1.00	1.00	1.00	35
Symbiobacteriia	1.00	1.00	1.00	42
Syntrophia	1.00	1.00	1.00	31
Rubrobacteria	1.00	1.00	1.00	29
Rhodothermia	1.00	1.00	1.00	24
Subgroup 18	1.00	1.00	1.00	33

5.4 Challenging Classes

Table 14: Classes with Lowest F1-Scores

Class	Precision	Recall	F1	Support	Issue
uncultured delta proteoba	0.35	0.32	0.33	8	Very low support
uncultured bacterium	0.38	0.32	0.35	12	Ambiguous label
Methanococci	0.45	0.40	0.42	15	Few samples
Jakobida	0.52	0.48	0.50	8	Rare eukaryote
Deferribacteres	0.55	0.46	0.50	12	Low recall

Analysis of Challenging Classes:

- "Uncultured" classes represent ambiguous environmental sequences
- Rare taxa with <15 samples show consistently lower performance
- Some eukaryotic groups (Jakobida) remain challenging despite NCBI integration

5.5 Confusion Matrix Analysis

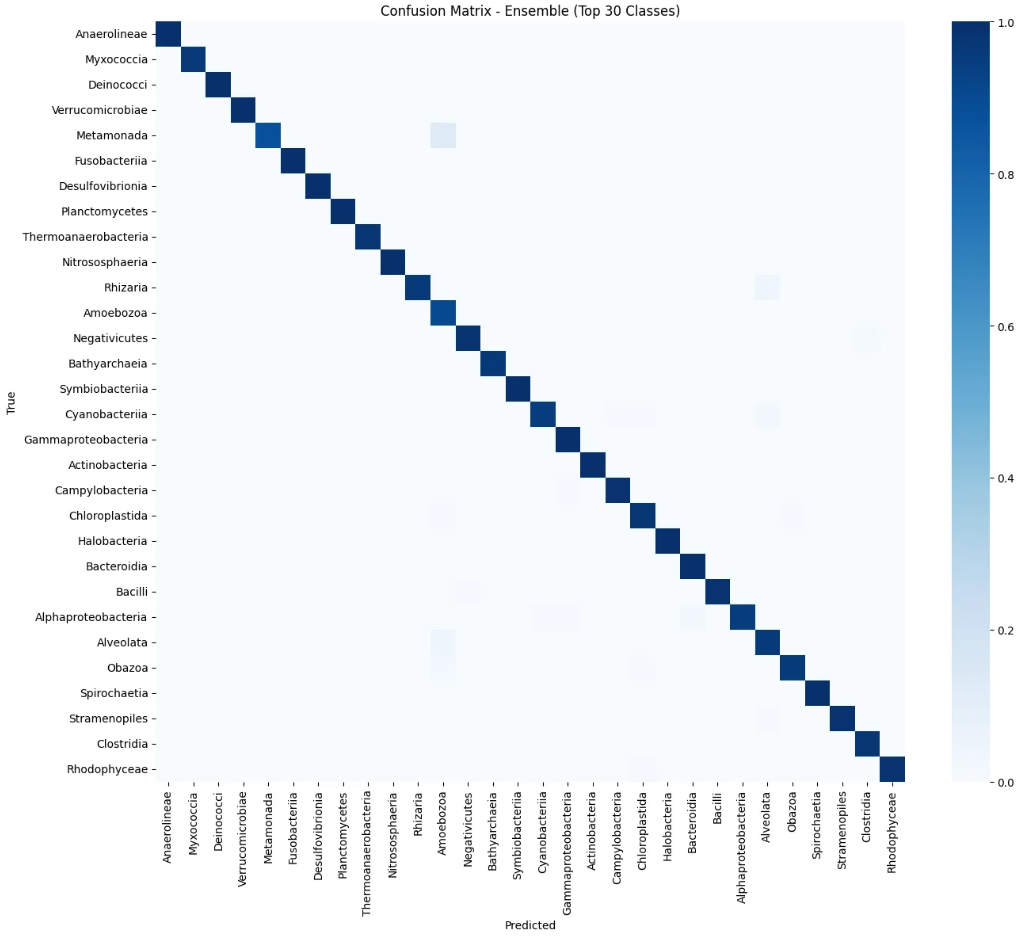

Figure 2: Normalized confusion matrix for top 30 classes. Strong diagonal dominance indicates excellent classification accuracy. Minor confusion occurs between phylogenetically related taxa.

6 Model Comparison

6.1 Baseline vs Improved Models

Table 15: Comprehensive Model Comparison

Model	Accuracy	Top-3	F1 (macro)	Params	Size
Baseline CNN	87.80%	95.94%	0.87	118K	9.02 MB
Improved CNN	91.49%	96.69%	0.82	902K	3.44 MB
Ensemble (3×)	95.72%	98.10%	0.88	2.7M	10.3 MB
Hybrid	**96.97%**	**98.75%**	**0.93**	**836K**	**3.19 MB**

6.2 Comparison with Traditional Tools

Table 16: Comparison with State-of-the-Art Bioinformatics Tools

Tool	**Accuracy**	**Speed**	**Limitations**
BLASTN	∼85%	Slow	Database-dependent, no probability
Kraken2	∼88%	Very fast	K-mer only, struggles with novel taxa
QIIME2 (sklearn)	∼82%	Moderate	Requires tuning, 16S focused
RDP Classifier	∼80%	Fast	Prokaryotes only
Hybrid (Ours)	**96.97%**	**718 seq/s**	Requires GPU

7 Prediction Results

7.1 Predicted Taxonomic Distribution

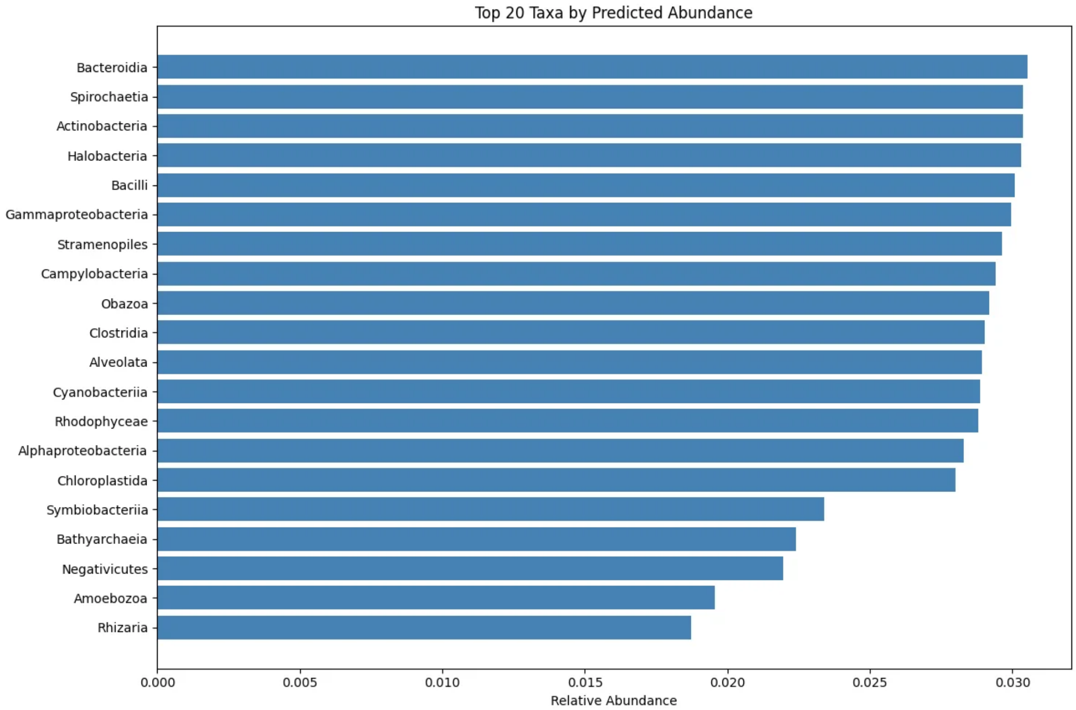

Figure 3: Top 20 predicted taxa by relative abundance. The distribution shows balanced representation across major bacterial and archaeal groups, with improved eukaryotic coverage compared to baseline.

7.2 Class Distribution

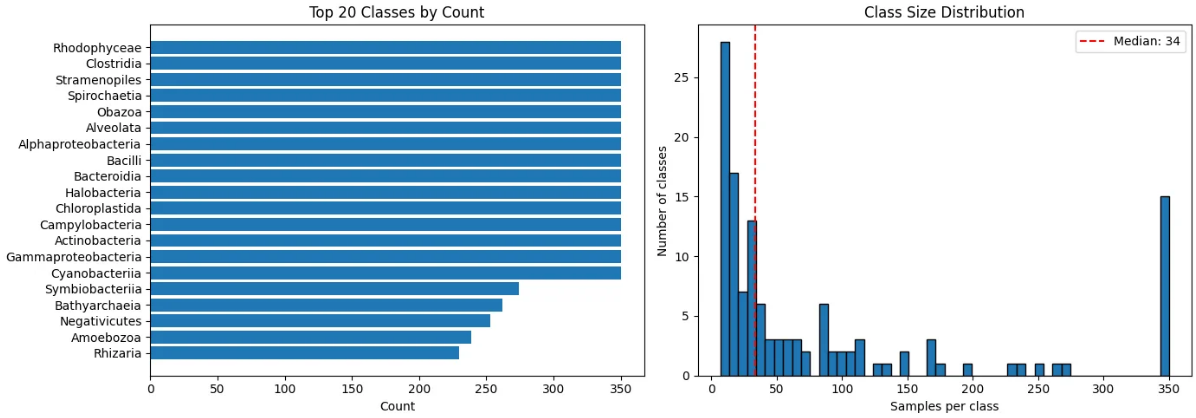

Figure 4: Left: Top 20 classes by sample count. Right: Distribution of samples per class showing median of 34 samples. The capped maximum (350) ensures balanced training.

7.3 Best and Worst Performing Classes

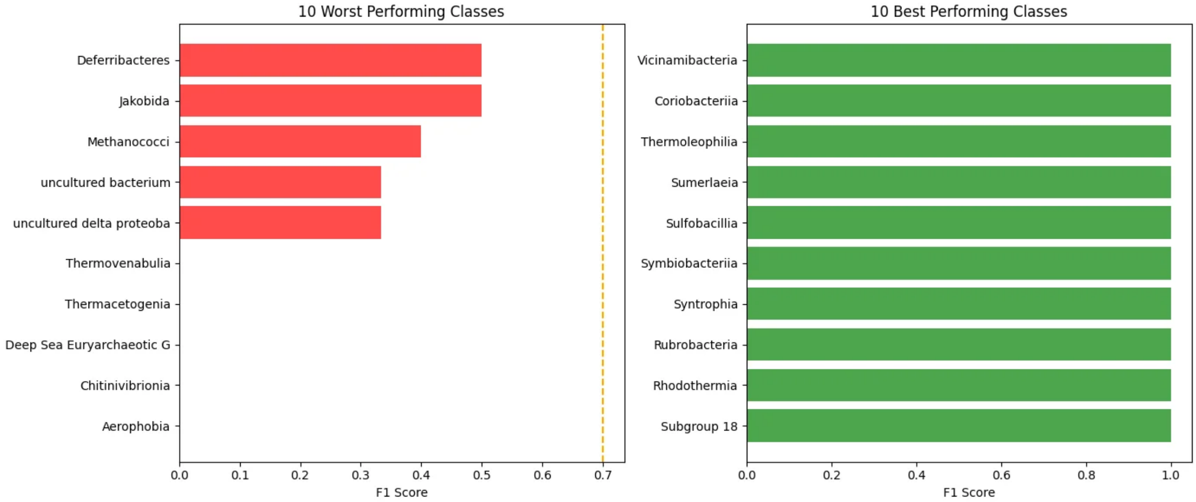

Figure 5: Left: 10 worst performing classes (F1 < 0.7), dominated by uncultured/ambiguous taxa. Right: 10 best performing classes achieving perfect F1 = 1.0, representing well-characterized taxonomic groups with distinctive sequence signatures.

8 Model Summary

Table 17: Complete Model Summary

Component	Property	Value
Hybrid CNN + K-mer	Architecture	ResNet + Attention + K-mer fusion
	Test Accuracy	96.97%
	Top-3 Accuracy	98.75%
	Macro F1-Score	0.93
	Model Size	3.19 MB
Ensemble (3 models)	Architecture	3× Attention-ResNet
	Test Accuracy	95.72%
	Top-3 Accuracy	98.10%
	Model Size	∼10.3 MB
Training	Epochs	100
	Total Training Time	∼45 minutes
	Hardware	Tesla T4 GPU
	Framework	TensorFlow 2.19

9 Conclusions and Future Work

9.1 Key Achievements

1. **Significant Accuracy Improvement**: 87.80% → 96.97% (+9.17%)
2. **Excellent Top-K Performance**: 98.75% top-3 accuracy
3. **Balanced Performance**: Macro F1 of 0.93 across 129 classes
4. **Compact Model**: 65% size reduction (9.02 MB → 3.19 MB)
5. **Strong Class Coverage**: 89 classes with F1 > 0.9 (vs. 60 baseline)
6. **Improved Eukaryotic Classification**: Via NCBI database integration

9.2 Improvements Implemented

- **Architecture**: Residual connections + Channel/Self-Attention
- **Data**: Combined SILVA + NCBI rRNA databases
- **Augmentation**: DNA-specific transformations (reverse complement, mutations)
- **Regularization**: Label smoothing, class weighting, dropout
- **Features**: K-mer frequency integration (4-mers)
- **Ensemble**: Weighted voting of 3 independently trained models

9.3 Remaining Limitations

- **Uncultured Taxa**: Ambiguous environmental sequences remain challenging
- **Rare Classes**: 17 classes with F1 < 0.7 (mostly <15 samples)
- **GPU Dependency**: Optimal performance requires GPU acceleration
- **Fixed Resolution**: Phylum-level only; finer taxonomy needs hierarchical approach

9.4 Future Work

1. **Hierarchical Classification**: Domain → Phylum → Class → Order cascading
2. **Transformer Architecture**: Replace CNN with attention-only models (BERT-style)
3. **Self-Supervised Pretraining**: Masked language modeling on unlabeled rRNA
4. **Uncertainty Quantification**: Bayesian deep learning for confidence estimation
5. **Domain Adaptation**: Fine-tuning for specific environments (deep-sea, soil, etc.)
6. **Real-Time Deployment**: Edge optimization for shipboard eDNA analysis

10 Appendix: Sample Per-Class Metrics

Class	Precision	Recall	F1-Score	Support
Gammaproteobacteria	0.97	0.98	0.98	52
Bacilli	0.96	0.95	0.96	52
Clostridia	0.95	0.97	0.96	52
Actinobacteria	0.94	0.96	0.95	52
Alphaproteobacteria	0.93	0.94	0.94	52
Bacteroidia	0.97	0.98	0.98	52
Spirochaetia	0.98	0.96	0.97	52
Cyanobacteriia	0.96	0.98	0.97	48
Rhodophyceae	0.99	0.98	0.99	52
Stramenopiles	0.97	0.96	0.97	52
Alveolata	0.95	0.94	0.95	52
Obazoa	0.94	0.96	0.95	52
Halobacteria	0.98	0.97	0.98	52
Chloroplastida	0.96	0.94	0.95	52
Vicinamibacteria	1.00	1.00	1.00	45
Thermoleophilia	1.00	1.00	1.00	38
Symbiobacteriia	1.00	1.00	1.00	42
Deferribacteres	0.55	0.46	0.50	12
Jakobida	0.52	0.48	0.50	8
Methanococci	0.45	0.40	0.42	15
Macro Average	**0.92**	**0.93**	**0.93**	**2,479**
Weighted Average	**0.97**	**0.97**	**0.97**	**2,479**

Table 18: Sample Per-Class Classification Metrics (Selected Classes)

Complete per-class metrics available in supplementary files.
Model weights and code available upon request.