Artificial Intelligence Approach to Detecting and Estimating Wind and Rainfall Over the Oceans from Deep-Sea Hydro-Acoustic Data

1 Abstract

This paper proposes a comprehensive artificial intelligence-based methodology for detecting and estimating rainfall and wind patterns over oceanic regions using deep-sea hydro-acoustic data. The system addresses key challenges related to data handling, feature extraction, and model deployment while ensuring continual improvement via a feedback learning loop.

2 1. Introduction

Climate modeling and marine weather prediction are vital for navigation, disaster prevention, and environmental monitoring. Deep-sea hydro-acoustic sensors offer untapped potential for detecting meteorological events like rain and wind, which can be analyzed using modern AI techniques. This study integrates signal processing, machine learning, and real-time system deployment for effective environmental monitoring.

3 2. Timeline and Project Phases

Phase 1 (April): Background Research and Setup

- Study background theory and review related work.
- Define research objectives and system scope.
- Set up the technical environment (install libraries, tools).

Phase 2 (Early–Mid May): Data Preprocessing and Feature Extraction

- Load raw hydro-acoustic and satellite/weather datasets.
- Perform data cleaning, temporal alignment, and normalization.
- Extract initial feature sets (e.g., spectrograms, MFCCs).
- Analyze correlation between acoustic features and real-world events.

Phase 3 (Mid May–End of June): Model Development and Testing

- Develop baseline and deep learning models for event classification.
- Train using labeled acoustic segments for wind/rain.
- Perform preliminary testing and tune hyperparameters.

Phase 4 (July): Model Refinement and Pipeline Integration

- Improve model robustness with augmentation and ensemble strategies.
- Integrate feature extraction, prediction, and post-processing into a single pipeline.
- Validate pipeline with hold-out test datasets.

Phase 5 (August): Paper Writing and Documentation

- Draft the manuscript (introduction, methods, results).
- Finalize experimental results and visualizations.
- Write 70–80% of the paper.

Phase 6 (Early–Mid September): Review and Presentation

- Review and revise manuscript based on feedback.
- Prepare slide deck and presentation materials.
- Submit paper or thesis and deliver oral presentation.

4 3. Data Sources

- **MBARI S3 Bucket**: Provides long-term acoustic data from underwater hydrophones deployed in deep-sea environments.
- **AMADEUS**: A data archive containing synchronized oceanic and meteorological datasets. Useful for ground-truth comparisons and supervised learning.

Data includes audio recordings with different sampling rates, geospatial locations, and environmental metadata (e.g., time, depth, temperature).

5 4. Identified Problems

- **Huge Data Volume**: The large scale of acoustic data demands significant storage and processing resources.
- **Imbalanced Data (Rare Event Detection)**: Rain and wind events are infrequent, leading to class imbalance in classification tasks.
- **Disjoint Temporal Distribution**: Rain/wind events occur in bursts separated by long silent periods, complicating time-series modeling.
- **Hardware and Computational Limitations**: Deep learning on high-resolution audio data requires powerful GPUs and memory.
- **Heterogeneity in Formats and Quality**: Data comes from multiple sources with inconsistent formatting and resolution.
- **Noise**: Background noise from ocean currents, marine life, and ships can interfere with signal clarity.

6 5. Solutions and Data Processing

- **Cloud Storage + GPU Processing**: Audio data is stored on cloud platforms and processed locally using high-performance computing.
- **Python + SQL Pipelines**: For data cleaning, querying, and efficient retrieval based on timestamp and location.
- **Temporal Grouping**: Data is segmented into time-based blocks (hour/day) to standardize analysis and reduce variability.
- **Noise Filtering**: Techniques such as spectral subtraction, band-pass filtering, and wavelet transforms are used to reduce noise.
- **Data Labeling and Selection**: Only 10–12% of data is labeled with rain events; these are carefully selected for training.

7 6. Feature Engineering

- **Formulas for Meteorological Inference**: Rain, drizzle, and wind speeds are estimated using sound intensity and frequency shifts.
- **Spectrograms**: 2D representations of frequency content over time. Used as visual input to CNNs.
- **MFCCs**: Extracted to represent acoustic features that match human auditory perception. Useful for time-series classification.
- **Labeling**: Ground truth is generated by correlating acoustic signals with meteorological records or manually verified patterns.

8 7. Model Design and Training

- **Deep Learning Models**: CNNs for spectrograms, RNNs/LSTMs for sequential MFCCs, and Transformer-based architectures for context-aware learning.
- **Rule-Based Systems**: For baseline comparison using simple thresholds in frequency and amplitude.
- **Training Strategy**: Supervised learning using balanced datasets. Augmentation is applied to simulate rare events.
- **Output Classes**: Event type (rain, wind, no event), intensity level (low, medium, high).

9 8. Evaluation Metrics

- **Testing Environments**: Simulated and real-world testing with diverse sea states and locations.
- **Metrics Used**:
  - Accuracy: Overall correctness
  - Precision and Recall: Critical for rare event detection
  - ROC-AUC: Model discrimination power
  - MAE: For continuous outputs like wind speed estimation
- **Cross-validation**: Ensures generalizability across different data subsets.

10 9. Deployment and Feedback Loop

- **Real-Time Inference Pipeline**:
  - Input: Raw hydro-acoustic stream
  - Processing: Spectrogram/MFCC extraction
  - Inference: Trained model outputs predictions
  - Output: Label, confidence, timestamp
- **Deployment Tools**: Docker, REST API with Flask/FastAPI, and cloud orchestration (e.g., Kubernetes).
- **Continuous Learning**:
  - Data and predictions are logged
  - Human-in-the-loop for validation
  - Periodic model fine-tuning with new labeled data
- **Versioning**: Using DVC or MLflow for experiment tracking and model management.

11 10. Conclusion

This framework demonstrates a scalable and adaptive approach to leveraging deep-sea acoustic data for meteorological analysis. It integrates data engineering, feature extraction, modern deep learning, and operational deployment for real-world use.

References

1 Relevant research papers and datasets (to be included)

2 MBARI Data Repository, AMADEUS documentation

3 Deep Learning for Audio Processing - IEEE Tutorials