PATTERN RECOGNITION OF ANTIBIOTIC RESISTANCE IN *ESCHERICHIA COLI*, *SALMONELLA* SPP., *SHIGELLA* SPP., AND *VIBRIO CHOLERAE* FROM WATER-FISH-HUMAN NEXUS

An Undergraduate Thesis

Presented to the Faculty of

Department of Computer Science

Mindanao State University - Marawi City Campus

In Partial Fulfillment of the Requirements

for the Degree of

Bachelor of Science in Computer Science

Submitted by:

Author 1

Author 2

Adviser:

Prof. Janice F. Wade, MSCS

Co-Adviser:

Mr. Llewelyn A. Elcana

January 2026

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

Background of the Study

Antimicrobial resistance (AMR) represents one of the most pressing global health challenges of the 21st century. The World Health Organization has declared AMR among the top ten threats to global health, with an estimated 1.27 million deaths directly attributable to bacterial AMR in 2019 alone [2]. Without coordinated intervention, AMR-related mortality is projected to reach 10 million deaths annually by 2050, surpassing cancer as a leading cause of death worldwide.

The Philippines, as a rapidly developing archipelagic nation with extensive aquaculture industries and diverse healthcare systems, faces unique challenges in AMR surveillance and control. The country's position within the Indo-Pacific region—a recognized hotspot for emerging infectious diseases—places it at elevated risk for resistance dissemination across human, animal, and environmental interfaces [3]. The Antimicrobial Resistance Surveillance Program (ARSP), established in 1988, has documented concerning trends including rising carbapenem-resistant Enterobacteriaceae and extended-spectrum β-lactamase (ESBL)-producing organisms in clinical settings [4].

Recent advances in machine learning (ML) offer promising opportunities to enhance AMR surveillance capabilities. Data-driven approaches including clustering algorithms, random forest classifiers, and neural networks have demonstrated utility in identifying resistance patterns, predicting phenotypes from genotypes, and stratifying patient risk [5], [6]. However, application of these methods to environmental and aquaculture-derived isolates remains limited, particularly in resource-constrained settings where phenotypic data predominate over genomic information.

The Integrated One Health Approach to AMR Containment (INOHAC) AMR Project Two, implemented across three Philippine regions—BARMM, Central Luzon, and Eastern Visayas—provides a unique dataset spanning the water-fish-human nexus [7]. This One Health framework recognizes that AMR emergence and transmission occur at the intersection of human health, animal husbandry, and environmental contamination, requiring integrated surveillance strategies [8].

Statement of the Problem

Existing antimicrobial resistance (AMR) surveillance frameworks rely on predefined categorical labels—such as species classifications, clinical breakpoints, and resistance prevalence summaries—that constrain how phenotypic antimicrobial susceptibility testing (AST) data are represented and analyzed, thereby limiting the ability of pattern recognition methods to discover latent resistance structure.

In heterogeneous datasets from the Water–Fish–Human nexus, such as the INOHAC–Project 2 AST data, resistance profiles are noisy and inconsistently encoded, and unsupervised clustering alone provides limited assurance that discovered patterns are coherent, discriminative, or robust.

The absence of an integrated, leakage-aware pattern recognition framework that combines data preprocessing, unsupervised structure discovery, supervised validation, and systematic evaluation restricts the effective application of machine learning for quantitative characterization of antimicrobial resistance patterns across interconnected environmental and human-associated reservoirs.

Objectives of the Study

General Objective

To develop a pattern recognition system for antimicrobial resistance in the Water–Fish–Human nexus by preprocessing phenotypic AST data from the INOHAC–Project 2,

applying unsupervised clustering to discover latent resistance structures, and employing supervised machine learning algorithms to validate and interpret the discriminative capacity of identified resistance patterns.

Specific Objectives

Specifically, this study aims to:

1. To preprocess and engineer features from the INOHAC–Project 2 phenotypic AST dataset, including data cleaning, resistance encoding, and computation of derived features, in order to create an analysis-ready dataset suitable for pattern recognition in the Water–Fish–Human nexus.
2. To apply unsupervised hierarchical clustering for resistance phenotype discovery and to evaluate multiple supervised machine learning algorithms for their capacity to discriminate and validate the identified resistance patterns derived from the processed dataset.
3. To design and develop an integrated pattern recognition framework that incorporates data-driven cluster selection, leakage-safe model training, and an interactive visualization dashboard for exploring resistance profiles, regional distributions, and co-resistance relationships.
4. To evaluate the pattern recognition system using appropriate quantitative metrics and to interpret the resulting resistance patterns within the context of the Water–Fish–Human nexus without inferring causality.

Significance of the Study

This study contributes to antimicrobial resistance surveillance and public health practice across multiple dimensions:

For Public Health Authorities

The identification of resistance phenotype clusters and their geographic distribution provides actionable intelligence for targeted intervention. The finding that BARMM

harbors the highest concentration of MDR isolates enables prioritization of antimicrobial stewardship programs and laboratory capacity building in this region. The methodology developed can be integrated into routine surveillance workflows to enable real-time phenotype monitoring [4].

For Healthcare Practitioners

The characterization of species-specific and cluster-specific resistance profiles informs empirical antibiotic selection. Understanding that *Salmonella* isolates exhibit distinct aminoglycoside resistance patterns while *E. coli*/*K. pneumoniae* show tetracycline-dominated profiles enables more targeted prescribing prior to susceptibility results.

For One Health Implementation

The analysis of resistance patterns across environmental sources (water, fish, hospital) supports the One Health framework for AMR containment [8]. The identification of aquaculture systems as significant MDR reservoirs provides evidence for policies addressing antibiotic use in fisheries and aquaculture operations [9].

For Research Advancement

The validation of machine learning approaches for phenotypic resistance clustering contributes to the growing evidence base for computational epidemiology in resource-limited settings. The reproducible analytical pipeline enables replication and extension by other researchers investigating AMR patterns.

For Academic Contribution

This study contributes to the limited body of work applying unsupervised-supervised hybrid machine learning frameworks to environmental AMR surveillance in the Philippines. The methodology bridges phenotypic and computational approaches, demonstrating feasibility without requiring whole-genome sequencing infrastructure.

Scope and Limitations

Scope

This study encompasses the following:

1. **Data Source**: Antimicrobial susceptibility testing (AST) data from 491 bacterial isolates collected through the INOHAC AMR Project Two across three Philippine regions: BARMM, Central Luzon (Region III), and Eastern Visayas (Region VIII).
2. **Organisms**: Gram-negative Enterobacteriaceae including *Escherichia coli*, *Klebsiella pneumoniae*, *Enterobacter* species, and *Salmonella* species isolated from water, fish, and hospital sources.
3. **Antibiotics**: A panel of 22 antibiotics spanning major classes including penicillins, cephalosporins, aminoglycosides, fluoroquinolones, tetracyclines, and carbapenems, as tested according to Clinical and Laboratory Standards Institute (CLSI) guidelines.
4. **Analytical Methods**: Hierarchical agglomerative clustering (Ward's linkage, Euclidean distance), principal component analysis (PCA), Random Forest classification, and Phi coefficient co-resistance analysis.
5. **Temporal Scope**: Cross-sectional analysis of isolates collected during the INOHAC AMR Project Two sampling period.

Limitations

1. **Phenotypic Focus**: This study analyzes phenotypic resistance profiles (susceptible/intermediate/resistant) without genotypic characterization. Resistance mechanisms and mobile genetic elements are inferred but not directly confirmed.
2. **Retrospective Design**: Analysis was conducted on historical AST data, precluding prospective validation or temporal trend analysis.
3. **Regional Representation**: The three study regions may not be representative of all Philippine provinces, limiting generalizability to unstudied areas.
4. **Missing Data**: Some isolates lacked complete antibiotic panel coverage, potentially affecting cluster assignments for partially tested specimens.

5. **Environmental Context**: While One Health sampling captured water, fish, and hospital sources, additional environmental compartments (soil, wastewater, wildlife) were not included.

CHAPTER II

REVIEW OF RELATED LITERATURE

Related Concepts

This section establishes the conceptual foundations underlying the analytical framework, situating unsupervised machine learning and co-resistance analysis within antimicrobial resistance (AMR) surveillance.

Unsupervised Learning for Biological Pattern Discovery

The fundamental challenge in environmental AMR surveillance lies in the absence of predefined phenotype labels. Unlike clinical settings where treatment outcomes may provide ground truth for supervised learning, environmental isolates from the water-fish-human nexus lack such annotations [10]. This constraint necessitates unsupervised approaches that discover structure directly from data without labeled examples [11].

Hierarchical Agglomerative Clustering

Hierarchical clustering constructs a tree-like structure (dendrogram) that groups similar observations based on distance metrics, progressively merging clusters until a single root encompasses all data points [12]. Among linkage methods, Ward's minimum variance approach minimizes within-cluster sum of squares at each merge step, producing compact, spherical clusters that often correspond to biologically meaningful groupings.

The choice of distance metric fundamentally shapes cluster geometry. Euclidean distance remains standard for continuous data and is required for Ward's method. While the ordinal nature of resistance encoding (Susceptible = 0, Intermediate = 0.5, Resistant = 1) introduces theoretical ambiguity, empirical evaluations demonstrate robust clustering performance with ordinal resistance data [13].

Principal Component Analysis for Dimensionality Reduction

When analyzing resistance profiles across multiple antibiotics, visualization becomes impossible without dimensionality reduction. Principal Component Analysis (PCA) addresses this by projecting high-dimensional data onto orthogonal axes that maximize variance [14]. The first principal component captures the direction of greatest variability —often correlated with overall resistance burden—while subsequent components reveal secondary patterns such as antibiotic class-specific resistance.

In AMR research, PCA serves dual purposes: enabling two-dimensional visualization of cluster separation and identifying resistance features that drive phenotypic differentiation [5]. When clusters identified through hierarchical methods display separation in PCA space, this provides independent validation that the groupings capture genuine phenotypic structure.

Cluster Validation via Silhouette Analysis

Determining optimal cluster number remains a persistent challenge in unsupervised learning [11]. The silhouette coefficient addresses this by measuring the ratio of within-cluster cohesion to between-cluster separation [15]. Values range from −1 to +1, where scores above 0.5 indicate reasonable cluster structure and scores exceeding 0.7 suggest convincing groupings [16].

This internal validation evaluates whether data genuinely contain clusterable structure at a given resolution. For AMR phenotyping, high silhouette scores indicate that isolates partition into distinct resistance archetypes rather than forming a continuous spectrum.

Supervised Validation of Unsupervised Clusters

A critical methodological innovation involves using supervised classification not for prediction, but for validation. Once unsupervised clustering assigns isolates to phenotypic groups, Random Forest classification [17] assesses whether these groupings are sufficiently distinct to be discriminated by an independent learning algorithm.

This hybrid unsupervised-supervised framework addresses a fundamental epistemological concern: how can one validate clusters without ground truth labels? By training a classifier on cluster assignments (treating them as provisional labels) and evaluating discrimination via cross-validation, the approach tests whether clusters represent coherent structures rather than noise. High classification accuracy combined with high silhouette scores provides convergent evidence for phenotypic validity [6].

Spatial Considerations in Resistance Epidemiology

Antimicrobial resistance does not distribute randomly across geographic space. Isolates from proximate sampling sites often exhibit correlated resistance profiles due to shared selection pressures or horizontal gene transfer [18]. This phenomenon—spatial autocorrelation—has implications for surveillance design and statistical inference.

In multi-regional datasets spanning diverse geographic areas, isolates from the same sampling site share environmental and anthropogenic exposures. Geographic stratification of clustering results—examining whether resistance phenotypes distribute differently across regions—addresses this spatial dependence while revealing regional resistance signatures.

Co-Resistance Patterns

Co-resistance describes the phenomenon where resistance to one antibiotic is statistically associated with resistance to another [19]. Such associations may arise from genetic linkage, cross-resistance mechanisms, or shared selection pressure.

The clustering methods employed in this study implicitly capture co-resistance through phenotypic similarity. Isolates resistant to antibiotics A and B cluster together precisely because their joint resistance pattern differs from isolates resistant only to A or only to B. Visualizing cluster-specific resistance profiles as heatmaps reveals which antibiotic combinations define each phenotype [20].

The Multiple Antibiotic Resistance Index

The Multiple Antibiotic Resistance (MAR) index provides a scalar summary of resistance burden, calculated as the ratio of resistant antibiotics to total antibiotics tested [21]:

$$\mathrm{MAR} = \frac{a}{b}$$

where a represents the number of antibiotics to which the isolate is resistant and b represents the total number of antibiotics tested. Krumperman's original formulation established a threshold of 0.2, above which isolates likely originate from environments with significant antibiotic selection pressure. Clusters characterized by high mean MAR likely represent multidrug resistance (MDR) phenotypes with clinical relevance, providing external validation independent of the clustering algorithm.

Multidrug Resistance Classification

Multidrug resistance (MDR) is formally defined as acquired non-susceptibility to at least one agent in three or more antimicrobial categories [22]. This classification framework, established by an international expert proposal, provides standardized definitions for MDR, extensively drug-resistant (XDR), and pandrug-resistant (PDR) bacteria.

For Enterobacteriaceae such as *Escherichia coli*, *Salmonella* spp., and *Shigella* spp., MDR assessment considers resistance across antibiotic classes including penicillins, cephalosporins, carbapenems, aminoglycosides, fluoroquinolones, and folate pathway inhibitors. The MDR flag serves as an important clinical indicator of isolate pathogenic potential and treatment complexity.

Related Studies

This section examines the evolution of computational approaches to antimicrobial resistance (AMR) analysis, tracing the trajectory from supervised prediction paradigms toward unsupervised pattern discovery.

The Supervised Learning Era: Achievements and Limitations

The period from 2020 to 2024 witnessed advances in machine learning applications for AMR prediction, with Random Forest emerging as the predominant algorithmic choice. A systematic review found that Random Forest achieved a mean Area Under the Receiver Operating Characteristic (AUROC) of 0.75 across 23 studies, consistently outperforming logistic regression for predicting resistance phenotypes [23]. Yet this success obscures a fundamental limitation: supervised models require labeled training data that environmental surveillance programs rarely possess.

The dependency on pre-existing labels creates an epistemological paradox. High accuracy models for predicting resistance in *Mycobacterium tuberculosis* and *Escherichia coli* using genomic features could only classify isolates into categories already defined in training data [23]. When confronted with novel resistance patterns not represented in historical datasets, supervised classifiers fail by design. This limitation proves especially problematic for environmental surveillance under the One Health framework, where resistance patterns in the water-fish-human nexus may differ from clinical reference datasets [10].

The class imbalance problem further constrains supervised methods. Multidrug resistance (MDR) prevalence in surveillance datasets typically ranges from 10-20%, creating minority class prediction challenges that bias models toward susceptible classifications [24]. While stratified cross-validation partially addresses this issue, the underlying problem—insufficient representation of diverse resistance phenotypes—cannot be solved algorithmically when labels themselves are incomplete.

Unsupervised Approaches: Emerging Alternatives

Recognition of supervised limitations has prompted methodological diversification toward unsupervised pattern discovery. Affinity Propagation clustering on antibiotic resistance genomic data achieved silhouette coefficients of 0.82, demonstrating that

meaningful phenotypic structure can be discovered algorithmically rather than assumed from clinical categories [6].

These unsupervised approaches offer conceptual advantages beyond label independence. By clustering isolates based on resistance similarity rather than predefined categories, they can reveal "unknown unknowns"—resistance phenotypes that clinicians have not yet recognized as distinct entities. Hierarchical clustering with Ward's linkage has been applied to characterize MDR patterns in bacteria from agricultural sources, identifying resistance archetypes that spanned conventional species boundaries [13]. Such cross-species patterns may indicate horizontal gene transfer—a phenomenon invisible to species-specific supervised classifiers.

Spatial epidemiological approaches have emerged concurrently. Spatial panel data analysis of *E. coli* resistance across 30 Chinese provinces demonstrated significant spatial autocorrelation in cephalosporin, carbapenem, and quinolone resistance [18]. This finding suggests that resistance patterns cluster geographically, potentially reflecting shared anthropogenic pressures.

Regional Context: Southeast Asian Surveillance

A comprehensive meta-analysis synthesized 137 studies from 2013-2023, revealing disparities in Enterobacterales resistance across ecological compartments: ceftriaxone resistance reached 49.3% in human, 37.1% in environmental, and 11.2% in animal *E. coli* isolates [25]. These findings underscore the need for integrated One Health surveillance.

Within the Philippines, national surveillance data report *E. coli* with 43% third-generation cephalosporin resistance and 46% fluoroquinolone resistance [4]. Environmental studies documented MDR *E. coli* in the Marikina River watershed [26]. Yet these studies employed conventional susceptibility categorization without clustering-based phenotype discovery.

The Inter-Regional Network Through One Health Approach to Combat Antimicrobial Resistance (INOHAC) AMR Project Two represents the first multi-regional environmental surveillance effort covering Bangsamoro Autonomous Region in Muslim Mindanao (BARMM), Central Luzon, and Eastern Visayas simultaneously [7]. With isolates tested against multiple antibiotics across water, fish, and human sources, this dataset provides unprecedented phenotypic resolution. However, resistance patterns remain characterized only through conventional metrics (MDR prevalence, Multiple Antibiotic Resistance indices) rather than unsupervised phenotype identification—a gap the present study directly addresses.

Network and Co-Resistance Perspectives

Network-based approaches have illuminated the genetic architecture of resistance. Gene network analysis identified hub genes that mediate interconnected resistance phenotypes [19]. At the metagenomic scale, antimicrobial resistance gene (ARG) co-abundance patterns across 214,095 datasets showed higher correlation in human and animal samples compared to environmental sources [20], suggesting that environmental samples may harbor distinct co-resistance architectures.

Ward's linkage dendrograms with heatmaps have been employed to characterize pan-resistant healthcare infections, demonstrating that hierarchical visualization reveals antibiotic groupings consistent with pharmacological class [27]. The present study extends this visualization paradigm to environmental isolates.

Comparative Summary of Related Studies

Author	Year	Unsupervised	Supervised	Focus Area	Key Contribution
Nguyen et al.	2018	No	Yes	Genomic AMR prediction	ML for Salmonella MICs
Ardila et al.	2025	No	Yes	Systematic review	RF/GBDT top performers
Parthasarathi et al.	2024	Yes	Yes	AMR gene clustering	Silhouette 0.82
Kou et al.	2025	No	No	Spatial epidemiology	Spatial autocorrelation
Abada et al.	2025	Yes	No	Agricultural MDR	Ward's clustering
INOHAC	2024	No	No	Environmental surveillance	Multi-regional dataset
Current Study	**2024**	**Yes**	**Yes**	**Water-fish-human nexus**	**Hierarchical clustering + RF validation**

Table 1: Comparative Summary of Computational Approaches to AMR Analysis

The current study uniquely integrates unsupervised pattern discovery with supervised validation for multi-regional environmental surveillance.

Synthesis: The Methodological Gap

The foregoing review reveals a critical methodological gap at the intersection of computational approaches and environmental AMR surveillance:

Limitations of Existing Approaches

1. **Supervised approaches** achieve high accuracy but remain constrained to known phenotypes. Models trained on clinical datasets cannot identify novel resistance patterns absent from their training data, fundamentally limiting their utility for environmental discovery.
2. **Existing unsupervised methods** have rarely been applied to multi-regional environmental surveillance. While hierarchical clustering has proven effective in agricultural and clinical settings, its systematic application to One Health surveillance datasets spanning water-fish-human interfaces remains underexplored.
3. **Spatial epidemiology** typically operates on aggregated resistance metrics rather than integrated phenotypic profiles. Studies documenting spatial autocorrelation in resistance rates have not combined these insights with data-driven phenotype discovery.
4. **Philippine surveillance** has characterized environmental resistance through conventional MDR classification without exploring underlying phenotypic structure. The INOHAC dataset's potential for pattern discovery remains unrealized through traditional analytical approaches.

The Present Study's Contribution

The present study addresses these gaps through a hybrid unsupervised-supervised framework specifically designed for environmental AMR surveillance:

Unsupervised Pattern Discovery. By first clustering isolates using Ward's hierarchical method with Euclidean distance, the approach discovers resistance archetypes directly from phenotypic data without requiring predefined labels. This enables identification of novel resistance patterns that conventional clinical categories might overlook.

Supervised Validation. By subsequently training a Random Forest classifier on cluster assignments and evaluating discrimination accuracy through cross-validation, the framework tests whether discovered clusters represent biologically coherent structures rather than statistical artifacts. High classification accuracy combined with high silhouette scores provides convergent evidence for phenotypic validity.

Multi-Regional Environmental Focus. Applying this methodology to isolates spanning multiple Philippine regions and ecological compartments (water, fish, human sources) enables characterization of resistance phenotypes specific to the One Health nexus—patterns potentially distinct from those documented in purely clinical surveillance.

Integrated Interpretation. Combining cluster validation with MAR index analysis and MDR classification bridges data-driven discovery with clinically interpretable resistance metrics, facilitating translation of computational findings into actionable surveillance insights.

This integrated approach represents a methodological advance over both purely supervised prediction (which requires known labels) and purely unsupervised clustering (which lacks validation mechanisms), offering a reproducible framework for future environmental AMR surveillance studies.

CHAPTER III

THEORETICAL FRAMEWORK

Introduction

This chapter establishes the theoretical foundations underpinning the development of a pattern recognition system for antimicrobial resistance (AMR) within the Water–Fish–Human nexus. The theoretical framework draws from three interconnected domains: (1) computational pattern recognition theory, (2) public health surveillance epistemology, and (3) software systems design principles. Together, these foundations provide the intellectual scaffolding for addressing the methodological challenges identified in the Statement of the Problem and justify the design decisions implemented in the Architectural Design.

Primary Theoretical Foundations

Pattern Recognition Theory

The primary theoretical foundation of this study is **Pattern Recognition Theory**, as formalized by Duda, Hart, and Stork in their seminal work *Pattern Classification*. Pattern recognition is defined as the automatic discovery of regularities in data through the use of computational algorithms, with the aim of classifying or describing observations based on learned representations rather than explicit rules.

This theory is operationalized in the present study through the integration of **unsupervised** and **supervised** learning paradigms:

Paradigm	Theoretical Basis	Application in Study
Unsupervised Learning	Cluster Analysis Theory	Hierarchical Agglomerative Clustering discovers latent resistance structures without predefined labels
Supervised Learning	Statistical Learning Theory	Logistic Regression, Random Forest, and k-Nearest Neighbors validate discriminative capacity of discovered patterns

Table 2: Learning Paradigms in Pattern Recognition

The theoretical justification for combining both paradigms derives from the **cluster validation problem** articulated by Jain and Dubes : unsupervised methods alone cannot guarantee that discovered structures are meaningful, coherent, or reproducible. Supervised validation provides an external mechanism for assessing whether clusters represent genuinely separable phenotypic categories.

Hierarchical Clustering Theory

Ward's minimum variance method, employed in this study, is grounded in the theoretical principle of **within-cluster homogeneity maximization** [13]. The method iteratively merges clusters to minimize the total within-cluster sum of squares, producing dendrograms that reveal multi-scale structure in high-dimensional data. This approach is particularly appropriate for ordinal resistance data (S/I/R encoded as 0/1/2), where Euclidean distance preserves the progressive nature of resistance severity.

One Health Framework

The **One Health Framework** provides the domain-specific theoretical context for situating antimicrobial resistance within interconnected environmental, animal, and human health systems. Endorsed by the World Health Organization (WHO), Food and

Agriculture Organization (FAO), and World Organisation for Animal Health (WOAH), One Health recognizes that:

""The health of people is closely connected to the health of animals and our shared environment" [8]."

The Water–Fish–Human nexus examined in this study represents a concrete instantiation of One Health principles, tracing antimicrobial resistance across:

- **Water systems** (drinking water, lake water, river water, effluent discharge)
- **Aquaculture** (fish species: *Banak*, *Gusaw*, *Tilapia*, *Kaolang*)
- **Anthropogenic interfaces** (treated/untreated effluent from healthcare facilities)

The One Health Framework justifies the study's focus on environmental reservoirs as sites of AMR emergence and dissemination, while simultaneously constraining the study's interpretive scope: the framework emphasizes *interconnection* and *surveillance* rather than *causal attribution*. This theoretical position aligns with the study's commitment to associational rather than causal language.

Supporting Theoretical Concepts

Information Leakage Theory in Machine Learning

A critical supporting concept is **Information Leakage Theory**, which addresses the methodological risk of inadvertently incorporating information from test data into model training, leading to overoptimistic performance estimates . Leakage violates the fundamental assumption of independent and identically distributed (i.i.d.) training and evaluation sets.

The study operationalizes leakage prevention through two architectural constraints derived from this theory:

Leakage Type	Theoretical Risk	Architectural Mitigation
Temporal Leakage	Future information influencing past predictions	Not applicable (cross-sectional data)
Feature Leakage	Target-derived features in input	Feature–metadata separation; metadata excluded from clustering
Preprocessing Leakage	Statistics computed on full dataset	Split-before-transform protocol; fit on training data only

Table 3: Leakage Types and Architectural Mitigations

These constraints are not merely procedural but reflect the theoretical requirement that evaluation metrics must estimate generalization error on truly unseen data .

Ordinal Data Representation Theory

The encoding of antimicrobial susceptibility results (Susceptible/Intermediate/Resistant) as ordinal numerical values (0/1/2) is grounded in **Ordinal Data Theory** . Ordinal variables possess natural ordering but lack equidistant intervals between categories.

The choice of Euclidean distance for clustering ordinal resistance data is justified by research demonstrating that, for low-dimensional ordinal spaces with consistent encoding, Euclidean distance approximates ordinal dissimilarity with acceptable distortion . Alternative distance metrics (e.g., Gower distance, Manhattan distance) were considered; the study's stability analysis using Adjusted Rand Index (ARI) across alternative configurations validates the robustness of the Euclidean-based solution.

Multi-Drug Resistance Classification Theory

The classification of isolates as **multidrug-resistant (MDR)** follows the standardized definition established by Magiorakos et al. [22]:

"An isolate is classified as MDR if it exhibits acquired non-susceptibility to at least one agent in three or more antimicrobial categories."

This definition provides a theoretically grounded, internationally recognized framework for categorizing resistance breadth. The study's computation of MDR status as a derived feature operationalizes this definition, enabling downstream analysis of resistance pattern associations.

The Variable Connection: From Data to Design

The relationship between research findings (independent variables) and design features (dependent variables) follows a structured derivation process grounded in the theoretical frameworks above.

Independent Variables (Research/Data)

The independent variables in this study comprise the phenotypic antimicrobial susceptibility testing (AST) data:

Variable Category	Specific Variables	Measurement
Resistance Profile	22 antibiotic susceptibility results	Ordinal (S=0, I=1, R=2)
Derived Metrics	MAR Index, Resistant Classes Count, MDR Status	Continuous/Binary
Contextual Metadata	Region, Site, Source Category, Species	Categorical (excluded from analysis)

Table 4: Independent Variables

Dependent Variables (Design Features)

The dependent variables are the architectural design features implemented in the system:

Design Feature	Derivation from Theory	Justification
Hierarchical Clustering Module	Pattern Recognition Theory → unsupervised structure discovery	Addresses SOP Problem 1 (categorical constraints) by discovering latent patterns without predefined labels
Supervised Validation Module	Cluster Validation Theory → external validation mechanism	Addresses SOP Problem 2 (weak assurance from clustering alone)
Feature–Metadata Separation	Information Leakage Theory → prevent feature leakage	Ensures objectivity in pattern discovery
Split-Before-Transform Protocol	Information Leakage Theory → prevent preprocessing leakage	Ensures unbiased performance estimation
Layered Architecture	Software Architecture Theory → separation of concerns	Addresses SOP Problem 3 (need for integrated framework)
Interactive Dashboard	Exploratory Data Analysis Theory → hypothesis generation through visualization	Enables post-hoc interpretation without biasing discovery

Table 5: Dependent Variables (Design Features)

The Derivation Chain

The following derivation chain traces how theoretical principles translate into design decisions:

Theoretical Foundation	Research Finding	Design Decision
Pattern Recognition Theory (Duda, Hart & Stork)	AST data contains latent resistance structure	Hierarchical Clustering Module
Cluster Validation Theory (Jain & Dubes)	Unsupervised alone is insufficient	Supervised Validation Module (LR/RF/kNN)
Information Leakage Theory (Kaufman et al.)	Metadata may bias pattern discovery	Feature–Metadata Separation
Statistical Learning Theory (Vapnik)	Preprocessing on full data causes leakage	Split-Before-Transform Protocol
One Health Framework (WHO/FAO/WOAH)	AMR crosses environmental boundaries	Multi-source Data Ingestion Module
Software Architecture Theory (Garlan & Shaw)	Need for reproducible, modular pipeline	Layered Architecture + CLI Orchestration

Table 6: Derivation Chain from Theory to Design

Theoretical Justification

Why Pattern Recognition Theory?

Pattern Recognition Theory is the most appropriate primary lens for this study because the Statement of the Problem explicitly identifies the limitation of **predefined categorical labels** in constraining the discovery of latent resistance structures. Pattern recognition, by definition, seeks to discover regularities that are not explicitly encoded in the data representation. The unsupervised component (hierarchical clustering) allows resistance patterns to emerge from phenotypic similarity rather than being imposed by external classification schemes.

Furthermore, the integration of supervised validation addresses the acknowledged weakness of unsupervised methods: the lack of external criteria for evaluating cluster quality. The theoretical framework thus provides both the mechanism for discovery (unsupervised learning) and the mechanism for validation (supervised learning), directly responding to the dual challenges articulated in the SOP.

Why One Health Framework?

The One Health Framework is essential for situating the study within the broader public health discourse on antimicrobial resistance. The Water–Fish–Human nexus is not an arbitrary data structure but a theoretically motivated representation of interconnected reservoirs where resistance genes and resistant organisms circulate.

Critically, the One Health Framework also provides epistemic constraints: it emphasizes surveillance, monitoring, and characterization rather than causal inference. This aligns with the study's commitment to associational language and its explicit avoidance of claims regarding resistance emergence mechanisms or transmission pathways. The theoretical framework thus serves both a constructive function (justifying the nexus perspective) and a regulatory function (constraining interpretive claims).

Why Information Leakage Theory?

The explicit incorporation of Information Leakage Theory distinguishes this study from naive applications of machine learning to biological data. The Statement of the Problem implicitly acknowledges the risk of methodological artifacts when it notes that unsupervised clustering alone provides "limited assurance" of coherent patterns. Information Leakage Theory provides the conceptual vocabulary for articulating these risks and the design principles for mitigating them.

The Split-Before-Transform protocol and Feature–Metadata Separation are not arbitrary design choices but theoretically mandated safeguards against a recognized class of methodological errors. By grounding these architectural decisions in established

theory, the study demonstrates awareness of machine learning pitfalls and implements principled solutions.

Why Layered Software Architecture?

The adoption of a **Layered Architectural Style** responds directly to the Statement of the Problem's identification of an "absence of an integrated framework." Layered architecture is specifically suited to sequential data processing workflows, where each layer transforms inputs and passes outputs to the next layer without circular dependencies.

The three-layer structure (Data Layer → Analysis Layer → Presentation Layer) maps directly onto the methodological progression from preprocessing to pattern discovery to interpretation. This architectural style enforces separation of concerns, preventing the conflation of analytical and interpretive operations that could compromise objectivity.

Conceptual Framework Diagram

The following diagram synthesizes the theoretical relationships among the primary theories, supporting concepts, and design outcomes.

The conceptual framework consists of three interconnected layers:

Theoretical Layer

The foundational theories that inform the study's approach:

Theory	Role
Pattern Recognition Theory [28]	Provides the computational paradigm for discovering latent structures
One Health Framework [29]	Situates AMR within interconnected health systems

Table 7: Theoretical Layer Components

Conceptual Layer

The supporting concepts that operationalize theoretical principles:

Concept	Function
Cluster Validation	Ensures discovered patterns are meaningful and reproducible
Information Leakage	Prevents methodological artifacts in performance estimation
Ordinal Data Representation	Justifies encoding of resistance categories
MDR Definition	Provides standardized classification criteria

Table 8: Conceptual Layer Components

Design Layer

The architectural outcomes derived from theoretical constraints:

Component	Elements
Layered System Architecture	Data Layer → Analysis Layer → Presentation Layer
Leakage Prevention Protocol	Split-Before-Transform, Feature-Metadata Boundary
Validation Architecture	Unsupervised Discovery → Supervised Validation → Stability Assessment

Table 9: Design Layer Components

Framework Integration

The three layers form a coherent framework where:

- **Theoretical foundations** provide intellectual justification
- **Conceptual elements** operationalize abstract principles into specific constraints
- **Design decisions** implement these constraints as architectural features

Chapter Summary

This chapter established the theoretical foundations for the AMR pattern recognition system developed in this study. The primary theoretical frameworks—Pattern Recognition Theory and One Health Framework—provide complementary lenses for addressing

the computational and domain-specific challenges identified in the Statement of the Problem.

Supporting concepts including Information Leakage Theory, Ordinal Data Representation, and MDR Classification Standards operationalize these frameworks into specific methodological and architectural constraints. The derivation chain demonstrates how each design feature in the Architectural Design chapter traces back to established theoretical principles.

The theoretical framework ensures that the study's contributions are grounded in recognized scholarly traditions while maintaining methodological rigor appropriate to machine learning applications in public health surveillance.

CHAPTER IV

METHODOLOGY

Research Design

This study adopts an exploratory, computational research design grounded in pattern recognition and machine learning to address the stated research objectives. The design is exploratory because it seeks to uncover latent antimicrobial resistance (AMR) structures that are not explicitly defined by existing categorical labels, rather than testing predefined hypotheses or establishing causal relationships. It is computational in nature because the primary contribution of the study lies in the design, implementation, and evaluation of a data-driven analytical framework for resistance pattern discovery and validation.

The research design integrates unsupervised learning for resistance structure discovery with supervised learning used exclusively as an external validation mechanism. Unsupervised methods are employed to identify resistance patterns based solely on phenotypic similarity in antimicrobial susceptibility testing (AST) data, without incorporating biological, environmental, or geographic labels during the discovery phase. Supervised learning is subsequently applied to assess the discriminative capacity and robustness of the discovered patterns, thereby addressing the limitations of unsupervised clustering when used in isolation.

The methodological strategy follows a staged, leakage-aware pipeline consisting of: (1) data preprocessing and feature engineering, (2) unsupervised resistance pattern discovery, (3) supervised validation, (4) integrated system design, and (5) quantitative evaluation. Throughout the study, strict separation is maintained between pattern discovery and interpretation to prevent information leakage and circular reasoning. The

study is associational and descriptive in scope; no biological mechanisms, epidemiological transmission pathways, or clinical outcomes are inferred.

Data Source and Description

Dataset Origin

The dataset analyzed in this study was generated by the **INOHAC AMR Project Two** research team as part of an environmental antimicrobial resistance surveillance initiative. The present study did not involve primary sampling or laboratory experimentation. All analyses were conducted as a **secondary analysis** of phenotypic AST data collected by the source project.

The dataset comprises AST results for bacterial isolates obtained from environmental and aquaculture-associated sources across three geographic regions in the Philippines. The geographic regions and their corresponding local sampling sites are summarized in Table 10.

Region	Local Sites
Eastern Visayas (Ormoc)	Alegria, Larrazabal
Central Luzon (Pampanga)	San Gabriel, San Roque
BARMM (Marawi)	APMC, Dayawan, Gadongan, Tuca Kialdan

Table 10: Geographic Regions and Local Sampling Sites

Sample Source Categories

Isolates originated from environmental matrices representing the **Water–Fish interface** within the broader Water–Fish–Human nexus. These source categories capture exposure pathways relevant to environmental AMR dissemination and were used exclusively as contextual metadata during interpretation. Source categories are listed in Table 11.

Source Code	Source Type	Description
DW	Drinking Water	Community water sources
LW	Lake Water	Natural water bodies
RW	River Water	Flowing water systems
EWU	Effluent Water (Untreated)	Hospital or facility discharge (Human interface)
EWT	Effluent Water (Treated)	Processed effluent discharge (Human interface)
FB, FG, FT, FK	Fish	Banak, Gusaw, Tilapia, Kaolang

Table 11: Sample Source Categories

Note: While direct human clinical isolates are not included, effluent water samples (EWU, EWT) represent the anthropogenic component of the nexus, capturing resistance patterns potentially influenced by human antibiotic use and healthcare facility discharge.

Isolate Identification Convention

Each isolate was assigned a structured alphanumeric identifier encoding species, geographic origin, source type, replicate number, and colony number using the format:

`[Species Prefix]_[Region][Site][Source][Replicate]C[Colony]`

This convention enables systematic metadata parsing while preserving traceability throughout the analytical pipeline.

Antimicrobial Panel

Phenotypic AST data were generated using a panel of **22 antibiotics spanning 12 antimicrobial classes**, including an ESBL screening indicator. The antimicrobial panel is summarized in Table 12.

Antimicrobial Class	Antibiotics
Penicillins	Ampicillin
β-lactam/β-lactamase inhibitors	Amoxicillin/Clavulanic Acid
Cephalosporins (1st gen.)	Cefalexin, Cefalotin
Cephalosporins (3rd/4th gen.)	Cefpodoxime, Cefotaxime, Cefovecin, Ceftiofur
Advanced cephalosporins	Ceftaroline, Ceftazidime/Avibactam
Carbapenems	Imipenem
Aminoglycosides	Amikacin, Gentamicin, Neomycin
Quinolones / Fluoroquinolones	Nalidixic Acid, Enrofloxacin, Marbofloxacin, Pradofloxacin
Tetracyclines	Doxycycline, Tetracycline
Nitrofurans	Nitrofurantoin
Phenicols	Chloramphenicol
Folate pathway inhibitors	Trimethoprim/Sulfamethoxazole
Resistance indicator	ESBL screening

Table 12: Antimicrobial Panel Composition

Phase 1: Data Preprocessing and Feature Engineering

The objective of this phase is to transform heterogeneous raw antimicrobial susceptibility testing (AST) records into a structured numerical form that supports similarity-based analysis while preserving biologically meaningful resistance information. All preprocessing decisions were explicitly parameterized to ensure reproducibility and to prevent information leakage in downstream analyses.

Data Ingestion and Harmonization

Raw phenotypic AST data were consolidated from multiple source files provided by the INOHAC–Project 2. These files, supplied as comma-separated value (CSV) datasets corresponding to different collection sites, were integrated into a single unified dataset.

The ingestion process included the following steps:

- **Schema harmonization:** Column names, data types, and value encodings were standardized across source files to ensure structural consistency.

- **Metadata extraction:** Structured isolate identifiers were parsed to extract contextual variables such as geographic region, local site, source category, replicate number, and colony number.
- **Duplicate resolution:** Duplicate isolate records were identified and removed to ensure a one-to-one correspondence between isolates and resistance profiles.

This step ensured that all downstream analyses operated on a coherent and internally consistent dataset.

Data Quality Filtering

To ensure sufficient data completeness for reliable pattern recognition, threshold-based filtering criteria were applied at both the antibiotic and isolate levels.

- **Antibiotic-level filtering:** Antibiotics tested on fewer than **70% of isolates** were excluded to ensure adequate representation across resistance profiles.
- **Isolate-level filtering:** Isolates with more than **30% missing susceptibility values** were removed to avoid excessive reliance on imputation.

These thresholds balance data retention with analytical reliability and are consistent with exploratory machine learning practices applied to high-dimensional biological data. All thresholds were established beforehand to avoid after-the-fact adjustments based on results. Following quality filtering, 21 of the original 22 antibiotics met the 70% coverage threshold and were retained for analysis.

Resistance Encoding

Phenotypic AST outcomes recorded as categorical values—Susceptible (S), Intermediate (I), and Resistant (R)—were converted into ordinal numerical representations to support quantitative analysis.

Phenotype	Encoded Value	Interpretation
Susceptible (S)	0	No resistance observed
Intermediate (I)	1	Reduced susceptibility
Resistant (R)	2	Clinical resistance

Table 13: Ordinal Encoding of Phenotypic AST Results

This ordinal encoding preserves the progressive nature of resistance severity while enabling distance-based computations.

Missing Value Imputation

Following threshold-based exclusion, remaining missing susceptibility values were imputed using **median imputation**, applied independently to each antibiotic feature:

$$\hat{x}_{i,j} = \text{median}\big(\{x_{k,j} \mid x_{k,j} \text{ is observed}\}\big)$$

where $\hat{x}_{i,j}$ is the imputed resistance value for isolate i and antibiotic j, and $x_{k,j}$ represents observed resistance values for antibiotic j.

Median imputation is robust to outliers and preserves the ordinal nature of resistance data. Alternative strategies such as mean or mode imputation were considered; however, the median provides a conservative central estimate suitable for exploratory pattern recognition.

Derived Resistance Feature Computation

To support downstream interpretation and epidemiological contextualization, several derived resistance descriptors were computed. These features were **not included as inputs** to unsupervised clustering to prevent bias during pattern discovery.

Multiple Antibiotic Resistance (MAR) Index

The MAR index quantifies the proportion of antibiotics to which an isolate exhibits resistance:

$$\text{MAR} = \frac{a}{b}$$

where a is the number of antibiotics for which resistance is observed (encoded value = 2), and b is the total number of antibiotics tested for the isolate.

Interpretation:

- MAR ≤ 0.2: Low-risk source
- MAR > 0.2: High-risk source, indicative of antibiotic selection pressure

Resistant Classes Count

The breadth of resistance across antimicrobial classes was computed as:

$$\text{Resistant Classes} = |\{c \mid \exists a \in c, \text{resistance}(a) = \text{true}\}|$$

where c denotes an antimicrobial class and a denotes an antibiotic belonging to that class.

This metric captures class-level resistance diversity rather than resistance to individual agents.

Multidrug Resistance (MDR) Classification

An isolate was classified as multidrug-resistant (MDR) if resistance was observed in **three or more antimicrobial classes**, consistent with established definitions [22]:

$$\text{MDR} = \begin{cases} 1, \text{if Resistant Classes} \geq 3 \\ 0, \text{otherwise} \end{cases}$$

Feature–Metadata Separation

To prevent **information leakage** and circular reasoning, the analysis-ready dataset was explicitly partitioned into two components:

- **Feature Matrix (X):** Encoded resistance values for the 22 antibiotics, used exclusively for unsupervised clustering and supervised validation.
- **Metadata Matrix (M):** Contextual variables (e.g., region, site, species, source category, MDR status), reserved solely for post-discovery interpretation.

This separation ensures that resistance patterns are discovered strictly from phenotypic similarity and are not influenced by external labels or contextual information.

Phase 1 Output Summary

The output of Phase 1 consists of:

- Analysis-ready resistance feature matrix with encoded susceptibility values
- Derived resistance indicators (MAR, Resistant Classes, MDR status)
- Separated metadata matrix for post-hoc interpretation
- Data quality documentation including filtering statistics

Phase 2: Unsupervised Structure Discovery

The objective of this phase is to identify latent resistance structures based solely on phenotypic similarity in antimicrobial susceptibility profiles, without incorporating predefined biological, environmental, or geographic labels. All analyses in this phase operate exclusively on the resistance feature matrix produced in Phase 1.

Clustering Algorithm Selection

Hierarchical Agglomerative Clustering (HAC) was selected as the primary unsupervised learning method due to the following properties:

- **Exploratory suitability:** Unlike partition-based methods (e.g., k-means) that require **a priori** specification of k, HAC constructs a complete hierarchical structure first, deferring cluster number selection to post-hoc analysis using data-driven validation metrics (silhouette coefficient, WCSS elbow analysis).
- **Multi-scale structure discovery:** The hierarchical representation enables examination of resistance patterns at multiple levels of granularity.
- **Interpretability:** Dendrograms provide transparent visualization of cluster formation and merge decisions.
- **Minimal structural assumptions:** HAC does not impose assumptions regarding cluster shape or distribution.

These characteristics make HAC appropriate for exploratory pattern recognition in high-dimensional resistance data.

Distance Metric

Euclidean distance was used as the primary measure of dissimilarity between resistance profiles:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

where x and y are resistance vectors for two isolates and n is the number of antibiotics.

This metric was selected because it preserves proportional differences introduced by ordinal resistance encoding (S = 0, I = 1, R = 2) and is compatible with variance-based linkage methods such as Ward's criterion. Given the 22-dimensional feature space—where the number of features is substantially smaller than the sample size—Euclidean distance remains effective without dimensionality reduction.

Linkage Method

Ward's **minimum variance linkage method** was used to guide cluster merging:

$$\Delta(A, B) = \frac{n_A n_B}{n_A + n_B} \|c_A - c_B\|^2$$

where:

- n_A and n_B denote the sizes of clusters A and B,
- c_A and c_B represent their respective centroids.

Ward's method minimizes the increase in total within-cluster variance at each merge step, producing compact and relatively balanced clusters. This property is advantageous for identifying resistance phenotypes that are internally coherent and externally separable in feature space.

Determination of the Number of Clusters

The optimal number of clusters was determined using a **data-driven, multi-criteria approach** combining quantitative metrics with practical constraints, following established conventions for exploratory cluster analysis [1], [30].

Silhouette Analysis

Cluster cohesion and separation were evaluated using the silhouette score [16]:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where:

- $a(i)$ is the mean intra-cluster distance for isolate i,
- $b(i)$ is the mean distance to the nearest neighboring cluster.

Higher silhouette values indicate better-defined cluster structure, with scores ≥ 0.40 representing moderate-to-strong structure [31]. The average silhouette score across all isolates was computed for cluster solutions ranging from $k = 2$ to $k = 8$, a range consistent with recommendations for systematic cluster validation [1].

Within-Cluster Sum of Squares (WCSS)

Cluster compactness was assessed using the within-cluster sum of squares:

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2$$

where C_k denotes cluster k and μ_k its centroid. The elbow method was used to identify diminishing returns in compactness as the number of clusters increased [30].

Practical Constraints

To ensure reproducibility and meaningful biological interpretation, the following methodological constraints guided cluster number selection:

- **Sample size requirement:** A minimum of 20 isolates per cluster was mandated to permit reliable estimation of cluster-level resistance profiles, consistent with recommendations for 20–30 samples per subgroup in clustering analysis [32], [33].
- **Granularity control:** Excessive partitioning was avoided to preserve phenotypically coherent resistance groupings amenable to downstream interpretation.

Final cluster selection employed a multi-objective decision framework, prioritizing the elbow point when it satisfied both silhouette and stability criteria, with parsimony as a secondary consideration when multiple solutions were statistically valid [31].

Cluster Stability Assessment

The robustness of the discovered clustering structure was evaluated through **stability analysis** using two complementary approaches.

Alternative Configuration Comparison

Agreement between clustering solutions obtained using different distance metrics (Euclidean, Manhattan) was quantified using the **Adjusted Rand Index (ARI)**:

$$\text{ARI} = \frac{\text{RI} - E[\text{RI}]}{\max(\text{RI}) - E[\text{RI}]}$$

where RI is the Rand Index and $E[\text{RI}]$ is its expected value under random labeling, and $\max(\text{RI})$ is the maximum possible Rand Index.

Bootstrap Stability

Cluster membership stability was assessed through bootstrap resampling:

1. Resample 80% of isolates with replacement ($n = 100$ iterations)
2. Re-cluster each bootstrap sample using identical parameters
3. Compute Jaccard similarity between original and bootstrap cluster assignments

Higher ARI and Jaccard values indicate greater stability and robustness of the clustering structure, suggesting that identified resistance patterns are not artifacts of specific parameter choices.

Cluster-Level Profile Characterization

For each identified cluster, a **resistance profile** was computed summarizing the dominant phenotypic characteristics:

- **Mean resistance score** per antibiotic (0–2 scale)
- **Resistance prevalence** (proportion of isolates with R classification per antibiotic)
- **Class-level resistance summary** aggregating across antimicrobial categories

These profiles enable qualitative characterization of each cluster's resistance signature.

Phase 2 Output Summary

The output of this phase consists of:

- Final cluster assignments for each isolate
- Hierarchical linkage matrices and dendrograms
- Cluster-level resistance profiles summarizing dominant phenotypic patterns
- Stability metrics (ARI, Jaccard coefficients)

These outputs form the basis for supervised validation and interpretation, while remaining independent of external biological or contextual labels during discovery.

Phase 3: Supervised Validation

Supervised learning models were used solely to validate the discriminative capacity of the discovered resistance patterns. This phase implements leakage-safe train–test splitting, macro-averaged evaluation metrics, confusion matrix analysis, feature importance extraction, and cross-seed stability checks.

Classification Tasks

Two supervised classification tasks were designed to assess whether resistance patterns align with known biological categories:

Task	Target Variable	Purpose
Species Discrimination	Bacterial species	Assess if resistance fingerprints distinguish species
MDR Classification	MDR status (0/1)	Validate resistance-MDR relationship

Table 14: Supervised Classification Tasks

Leakage-Safe Data Splitting

To prevent information leakage between training and evaluation phases, the dataset was first partitioned into **training (80%) and test (20%) subsets** using stratified sampling to preserve class distributions. **Train–test splitting was performed prior to any preprocessing operations**, including missing value imputation and feature scaling.

All preprocessing steps were fitted **exclusively on the training data**, and the learned parameters were subsequently applied unchanged to both the training and test sets. This ensured that statistical properties of the test data did not influence model training, thereby preventing optimistic bias in supervised evaluation metrics.

Model Selection

Three classifier families were selected to represent different learning paradigms:

Model	Category	Rationale
Logistic Regression	Linear	Baseline; interpretable coefficients
Random Forest	Tree-based	Nonlinear; feature importance via Gini impurity
k-Nearest Neighbors	Distance-based	Instance-based; consistency check against clustering

Table 15: Supervised Model Selection

Hyperparameter Configuration:

Model	Parameters
Logistic Regression	`max_iter=1000`, `solver='lbfgs'`
Random Forest	`n_estimators=100`, `random_state=42`
k-Nearest Neighbors	`n_neighbors=5`

Table 16: Model Hyperparameters

Evaluation Metrics

Performance was quantified using macro-averaged metrics to prevent class imbalance bias:

Macro-Averaged Precision, Recall, F1

$$\text{Precision}_{\text{macro}} = \frac{1}{|C|} \sum_{c \in C} \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}$$

$$\text{Recall}_{\text{macro}} = \frac{1}{|C|} \sum_{c \in C} \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where C is the set of classes and TP, FP, FN are true positives, false positives, and false negatives respectively.

Accuracy

Overall classification correctness was measured as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Confusion Matrix

Per-class classification performance was visualized using confusion matrices to identify species-specific misclassification patterns.

Feature Importance Extraction

For Random Forest models, feature importance was extracted using Gini impurity:

$$\text{Importance}(f) = \sum_{t \in T} \Delta G_t \cdot \mathbb{1}[f_t = f]$$

where ΔG_t is the decrease in Gini impurity at node t when feature f is used for splitting.

Language Discipline: Feature importance reflects *associative* relationships within the dataset. High importance indicates statistical association, not causal influence on resistance phenotype.

Stability Across Random Seeds

Model stability was validated across multiple random states to ensure that model performance was not dependent on a specific random initialization:

Cross-Seed Stability Check Algorithm

Input: Dataset D, Model M, Seeds S = {42, 123, 456, 789, 1011}

Output: Stability metrics (mean, standard deviation)

For each seed s in S:

1. Set random state to s
2. Split D into train/test (80/20, stratified)
3. Train model M on training set
4. Evaluate on test set
5. Record performance metrics

Return: mean(metrics), std(metrics)

Low standard deviation across seeds indicates robust model performance.

Sensitivity Analysis: Split Ratio and Cross-Validation

To justify the train–test split configuration, a sensitivity analysis was conducted comparing different partitioning strategies. Three split ratios (70/30, 80/20, 90/10) and two

cross-validation schemes (5-fold, 10-fold) were evaluated across all three classifier models.

Split Ratio Comparison

Split	Model	F1 Score	Accuracy	Stability (std)
70/30	Logistic Regression	0.923 ± 0.026	0.965	0.026
70/30	Random Forest	0.944 ± 0.016	0.974	0.016
70/30	KNN	0.917 ± 0.034	0.964	0.034
80/20	Logistic Regression	0.952 ± 0.031	0.978	0.031
80/20	Random Forest	0.973 ± 0.022	0.988	0.022
80/20	KNN	0.956 ± 0.028	0.980	0.028
90/10	Logistic Regression	0.983 ± 0.020	0.992	0.021
90/10	Random Forest	0.991 ± 0.018	0.996	0.018
90/10	KNN	0.958 ± 0.026	0.980	0.026

Table 17: F1 Scores Across Different Train–Test Split Ratios (MDR Classification)

Cross-Validation Comparison

CV Folds	Model	F1 Score	Accuracy	Stability (std)
5-fold	Logistic Regression	0.953 ± 0.009	0.978	0.009
5-fold	Random Forest	0.947 ± 0.025	0.976	0.025
5-fold	KNN	0.933 ± 0.038	0.970	0.038
10-fold	Logistic Regression	0.932 ± 0.045	0.970	0.045
10-fold	Random Forest	0.955 ± 0.035	0.980	0.035
10-fold	KNN	0.952 ± 0.037	0.978	0.037

Table 18: F1 Scores Across Different Cross-Validation Configurations

Sensitivity Analysis Interpretation

The sensitivity analysis revealed the following key insights:

1. **Consistent high performance**: All split configurations achieved F1 scores above 0.91, indicating that supervised validation results are robust to partitioning choices.
2. **Acceptable stability**: Standard deviations ranged from 0.009 to 0.045, all within acceptable bounds (< 0.05), confirming that results are not artifacts of random initialization.

3. **80/20 split justification**: While 90/10 achieved marginally higher scores, the smaller test set (≈49 samples) reduces statistical reliability of performance estimates. The 80/20 split balances training data adequacy with reliable test evaluation.
4. **5-fold CV preference**: 5-fold cross-validation produced more stable results (lower standard deviation) compared to 10-fold, particularly for Logistic Regression (0.009 vs 0.045).

These findings support the use of the **80/20 train–test split with 5-fold cross-validation** as the standard configuration for supervised validation in this study.

Phase 3 Output Summary

The output of this phase consists of:

- Classification performance metrics for each model and task
- Confusion matrices for per-class analysis
- Feature importance rankings from Random Forest
- Cross-seed stability statistics
- Sensitivity analysis results across split configurations

Phase 4: Integrated Framework Design

An integrated analytical framework was developed to support reproducible, leakage-aware antimicrobial resistance pattern recognition. The framework follows a **three-layer architecture** consisting of a data layer, a processing layer, and a presentation layer.

System Architecture

Layer	Components	Function
Data Layer	Preprocessing pipeline, feature engineering modules	Data cleaning, encoding, imputation, feature preparation
Processing Layer	Clustering, supervised models, statistical analysis	Pattern discovery, validation, co-resistance analysis
Presentation Layer	Streamlit dashboard	Visualization, exploration, and result interpretation

Table 19: Integrated Framework Architecture

Pipeline Orchestration

Pipeline orchestration was implemented through a **central command-line interface (CLI)** that controlled execution of all analytical stages. This design ensures modularity, reproducibility, and consistent parameter application across experiments.

Command	Description
`--pipeline`	Execute full data preprocessing and clustering pipeline
`--validate`	Run supervised validation and stability checks
`--analyze`	Perform post-hoc statistical and regional analyses
`--viz`	Generate all figures and plots
`--app`	Launch interactive Streamlit dashboard

Table 20: Pipeline Orchestration Commands

Reproducibility was enforced using **fixed random seeds**, centralized configuration files, and persistent storage of intermediate artifacts (e.g., linkage matrices, trained models, clustering assignments).

Co-Resistance Analysis

Antibiotic co-resistance patterns were quantified using the **phi coefficient (φ)**, calculated from binary resistance co-occurrence tables:

$$\varphi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

where a, b, c, and d represent the counts in a 2×2 contingency table of resistance presence and absence between two antibiotics.

	Antibiotic B: R	**Antibiotic B: S**
Antibiotic A: R	a	b
Antibiotic A: S	c	d

Table 21: Phi Coefficient Contingency Table Structure

Antibiotic clustering based on co-resistance similarity was subsequently performed using hierarchical clustering with distance defined as $1 - \varphi$.

Interactive Visualization Dashboard

The Streamlit-based dashboard provides interactive exploration of resistance patterns through three primary views:

View	**Description**
Cluster Explorer	Interactive dendrogram with selectable cut-points; cluster-level resistance heatmaps displaying mean resistance scores per antibiotic
Regional Distribution	Geographic breakdown of cluster assignments with stacked bar charts showing proportional representation across regions and sites
Co-Resistance Network	Interactive phi-coefficient heatmap with threshold filtering; hierarchically clustered antibiotic groupings

Table 22: Dashboard Components

The dashboard enables users to:

- Adjust dendrogram cut-height to explore clustering at different granularities
- Filter isolates by region, source type, or species
- Export visualizations and summary statistics
- Compare resistance profiles across selected clusters

Phase 5: System Evaluation and Interpretation

System performance was evaluated using a combination of **internal clustering metrics**, **supervised validation metrics**, and **controlled association analysis**.

Clustering Evaluation Metrics

Metric	Purpose	Interpretation
Silhouette Score	Cluster cohesion and separation	Higher values indicate better-defined clusters
WCSS	Cluster compactness	Lower values indicate tighter clusters
Adjusted Rand Index	Robustness across methods	Higher values indicate greater stability
Jaccard Coefficient	Bootstrap stability	Higher values indicate membership consistency
Cluster Size Distribution	Practical validity	Minimum of 20 isolates per cluster

Table 23: Clustering Evaluation Metrics

Supervised Validation Metrics

Metric	Description
Accuracy	Overall classification correctness
Macro Precision	Average precision across classes
Macro Recall	Average recall across classes
Macro F1-score	Balanced performance across classes
Confusion Matrix	Per-class misclassification analysis

Table 24: Supervised Validation Metrics

Association Analysis

Associations between resistance clusters and metadata variables were evaluated using **Cramér's V**, computed as:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}}$$

where χ^2 is the chi-square statistic, n is the sample size, and r and c are the dimensions of the contingency table.

Cramér's V Value	Association Strength
0.00 – 0.10	Negligible
0.10 – 0.20	Weak
0.20 – 0.40	Moderate
0.40 – 0.60	Relatively Strong
0.60 – 1.00	Strong

Table 25: Cramér's V Interpretation

Interpretation Protocol

Interpretation followed a strict **post-hoc protocol** to maintain analytical integrity:

1. Clusters were generated using resistance features only (Phase 2)
2. Metadata were overlaid after clustering for descriptive analysis
3. Statistical associations were reported using associational language only
4. No causal claims were made regarding resistance emergence or transmission

This protocol ensures that interpretive conclusions remain within the methodological scope of the study.

Implementation Details

The analytical framework was implemented using **Python 3.9+** and widely adopted open-source scientific computing libraries. Computational requirements were modest and suitable for standard desktop or laptop hardware.

Component	Specification
Programming Language	Python 3.9+
Data Processing	pandas, numpy
Machine Learning	scikit-learn
Statistical Analysis	scipy
Visualization	matplotlib, seaborn
Dashboard	Streamlit
Version Control	Git

Table 26: Implementation Environment

Computational Pipeline

The complete analytical pipeline is summarized below:

Complete Analytical Pipeline

Input: Raw AST CSV files

Output: Cluster assignments, validation metrics, dashboard

PHASE 1: Preprocessing

1. Load and harmonize data from multiple sources
2. Parse isolate identifiers to extract metadata
3. Apply quality filters (70% antibiotic, 30% isolate thresholds)
4. Encode resistance values (S=0, I=1, R=2)
5. Impute missing values using median imputation
6. Compute derived features (MAR, MDR, Resistant Classes)
7. Separate feature matrix X from metadata matrix M

PHASE 2: Clustering

1. Compute pairwise Euclidean distances
2. Perform hierarchical clustering (Ward's linkage)
3. Evaluate k=2 to k=8 for optimal selection (metrics computed to k=10)
4. Apply practical constraints (minimum cluster size)
5. Select optimal k and assign cluster labels
6. Assess stability via ARI and bootstrap Jaccard

PHASE 3: Supervised Validation

1. Split data (80/20 stratified, before preprocessing)
2. Train models (LR, RF, kNN) on training set
3. Evaluate on test set (accuracy, precision, recall, F1)
4. Extract feature importance from Random Forest
5. Repeat across multiple random seeds

PHASE 4: Analysis and Visualization

1. Compute co-resistance phi coefficients
2. Calculate Cramér's V for cluster-metadata associations
3. Generate visualizations (dendrograms, heatmaps, bar plots)
4. Deploy Streamlit dashboard

Return: Final cluster assignments, performance metrics, dashboard

Ethical Considerations

This study involved the secondary analysis of environmental and aquaculture-associated bacterial isolates. No human subjects, clinical samples, or personal identifiers were included in the dataset. The dataset was anonymized prior to analysis, and all results are reported at an aggregate level. Ethical approval was therefore not required for this computational study.

Limitations

The following methodological limitations are acknowledged:

1. **Scope limitation:** The dataset represents the Water–Fish interface; direct human clinical isolates are not included, limiting generalizability to the full Water–Fish–Human nexus.
2. **Temporal limitation:** The study analyzes a single cross-sectional dataset; temporal dynamics of resistance evolution cannot be assessed.

3. **Imputation effects:** Median imputation may introduce bias for antibiotics with highly skewed resistance distributions.
4. **Clustering assumptions:** Ward's linkage assumes spherical clusters and may not capture non-convex resistance pattern structures.
5. **External validation:** Supervised validation assesses internal discriminative capacity but does not validate against external AMR surveillance datasets.

Chapter Summary

This chapter presented a **comprehensive, leakage-aware methodology** for antimicrobial resistance pattern recognition using phenotypic AST data. The framework integrates unsupervised discovery, supervised validation, co-resistance analysis, and system-level evaluation while maintaining strict interpretive discipline.

Phase	Objective Addressed	Key Outputs
Phase 1	SO1	Cleaned, encoded, analysis-ready dataset
Phase 2	SO2 (Part 1)	Hierarchical resistance clusters with stability assessment
Phase 3	SO2 (Part 2)	Supervised validation and stability metrics
Phase 4	SO3	Integrated analytical framework and interactive dashboard
Phase 5	SO4	Quantitative evaluation and controlled interpretation

Table 27: Methodology Summary

The methodology ensures that resistance patterns are discovered through objective, data-driven processes and that all interpretive statements remain within appropriate associational bounds. The integrated framework supports reproducible execution and interactive exploration of results.

CHAPTER V

ARCHITECTURAL DESIGN

Introduction

This chapter presents the architectural design of the implemented pattern recognition system for antimicrobial resistance (AMR) within the Water–Fish–Human nexus. The architecture described in this chapter corresponds to a fully implemented and operational system, rather than a conceptual or proposed design.

The system processes antimicrobial susceptibility testing (AST) data from bacterial isolates (*Escherichia coli*, *Salmonella* spp., *Shigella* spp., and *Vibrio cholerae*) collected across environmental and clinical sample sources. Environmental and regional metadata are preserved throughout the pipeline but are intentionally excluded from analytical computations to ensure unbiased pattern discovery.

Architectural Design Goals

The architecture was guided by five primary design goals: modularity, reproducibility, leakage prevention, interpretability, and experimental control. These goals reflect both software engineering best practices and methodological requirements specific to unsupervised and supervised machine learning workflows.

Design Goal	Description
Modularity	Loosely coupled components with clear responsibilities; independent replacement capability
Reproducibility	Centralized configuration, fixed random seeds, artifact persistence
Leakage Prevention	Train-test separation, feature-metadata boundary, fit-on-train only
Interpretability	Traceable assignments, feature importance, metadata post-hoc only
Experimental Control	Single config source, systematic comparison, baseline consistency

Table 28: Architectural Design Goals

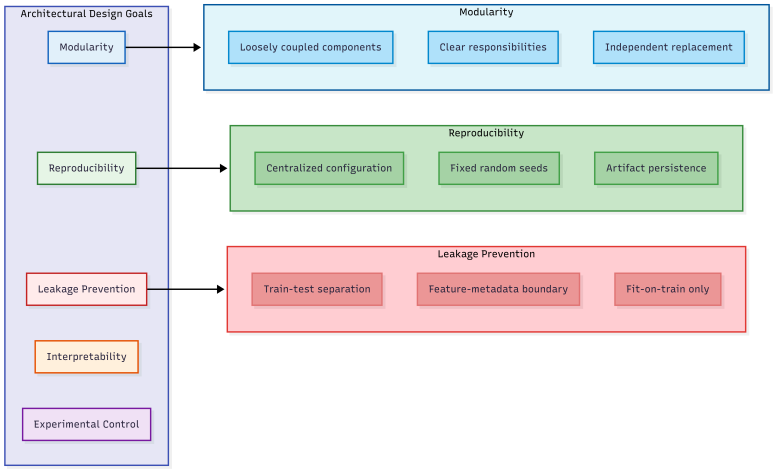

Figure 1: Architectural Design Goals Flowchart

Modularity

The system is decomposed into discrete, loosely coupled components with clearly defined responsibilities. Each module can be modified or replaced without affecting unrelated parts of the pipeline. For example, alternative clustering strategies can be introduced without modifying data ingestion or visualization logic.

Reproducibility

Reproducibility is ensured through:

- centralized configuration of all parameters,

- fixed random seeds for stochastic processes, and
- persistence of intermediate artifacts at every transformation stage.

These mechanisms ensure that experimental results can be independently replicated under identical conditions.

Leakage Prevention

The architecture enforces strict separation between training and evaluation data. Train–test splitting occurs before any preprocessing steps, and all preprocessing parameters are learned exclusively from training data. This separation is enforced structurally through module boundaries rather than relying on procedural discipline alone.

Interpretability

Interpretability is supported by:

- traceable cluster assignments,
- explicit feature importance extraction, and
- separation between pattern discovery and domain interpretation.

Metadata is intentionally excluded from clustering and validation processes and is introduced only during post-hoc visualization.

Experimental Control

All experimental parameters are defined in a centralized configuration module, enabling systematic comparison of alternative settings while maintaining consistent baselines. Intermediate artifacts are preserved to allow retrospective analysis without recomputation.

Overall Architecture Style

The system adopts a **layered architectural style**, consisting of a Data Layer, Analysis Layer, and Presentation Layer. This structure mirrors the methodological progression from data preprocessing to pattern discovery and interpretation.

Layer	Components	Data Flow
Presentation Layer	Static Visualizations, Interactive Dashboard, PCA Projections	Outputs to user
Analysis Layer	Hierarchical Clustering, Supervised Validation, Statistical Analysis	Processes feature matrices
Data Layer	Data Ingestion, Quality Filtering, Resistance Encoding, Feature Engineering	Receives raw AST data

Table 29: Layered Architecture Overview

Layered architecture was selected because it:

- naturally aligns with sequential ML workflows,
- enforces separation of concerns,
- prevents circular dependencies, and
- provides clear validation boundaries between phases.

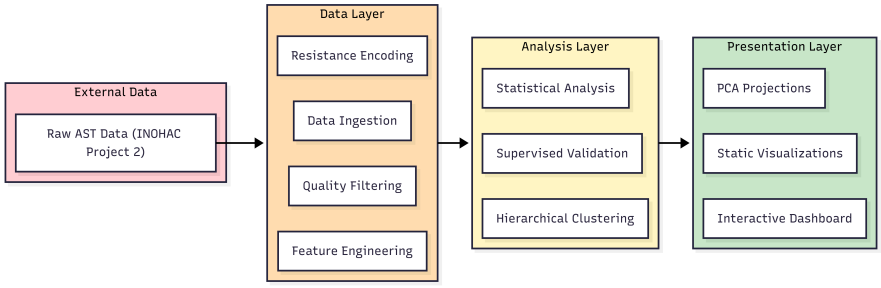

Figure 2: Layered Architecture Overview

High-Level System Architecture

At the system level, execution is orchestrated through a unified command-line interface (`main.py`). All pipeline operations—data processing, analysis, validation, visualization, and dashboard deployment—are invoked via explicit CLI flags.

CLI Flag	Description
`--pipeline`	Execute full data preprocessing and clustering pipeline
`--validate`	Run supervised validation and stability checks
`--analyze`	Perform post-hoc statistical and regional analyses
`--viz`	Generate all figures and plots
`--app`	Launch interactive Streamlit dashboard
`--all`	Run everything in sequence

Table 30: CLI Orchestration Commands

This orchestration design ensures:

- consistent initialization across experiments,
- prevention of partial or misconfigured execution, and
- reproducible experiment setup.

Each layer communicates only through persisted artifacts or structured outputs, preventing hidden state sharing.

Configuration Module

All configurable parameters are defined in a single configuration module (`config.py`):

Category	Parameters
Path Definitions	PROJECT_ROOT, DATA_DIR, PROCESSED_DIR, FIGURES_DIR, MODELS_DIR
Reproducibility	RANDOM_STATE = 42
Data Cleaning	MIN_ANTIBIOTIC_COVERAGE, MAX_ISOLATE_MISSING
Antibiotic Classes	ANTIBIOTIC_CLASSES dict, MDR_CLASSES_BY_SPECIES dict
Clustering	Linkage Method, Distance Metric, K-Selection Criteria
Machine Learning	Train-Test Split Ratio, CV Folds, Model Parameters

Table 31: Configuration Module Parameters

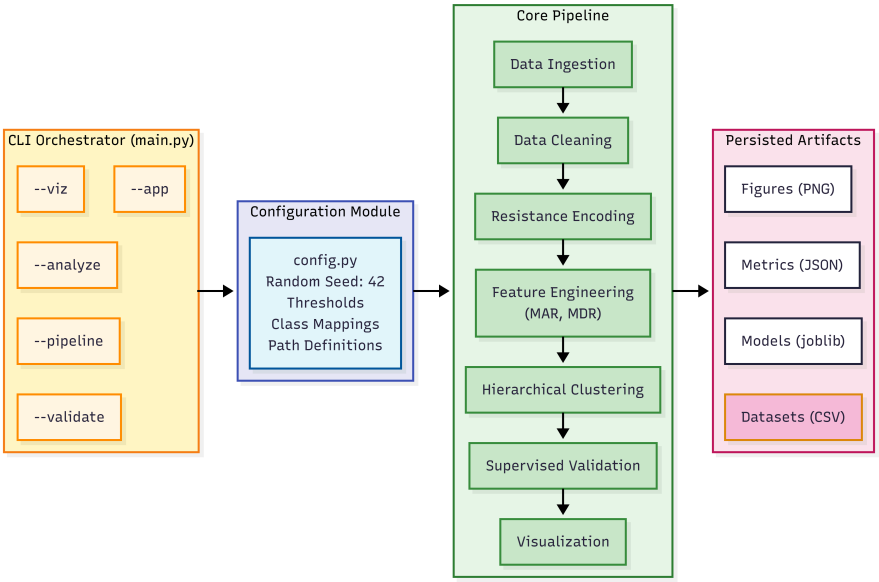

Figure 3: High-Level System Architecture

Data Layer

The Data Layer is responsible for transforming raw antimicrobial susceptibility testing (AST) data into an analysis-ready feature matrix. It comprises five sequential processing stages.

Data Ingestion

Raw CSV files from multiple regional data sources (INOHAC–Project 2) are consolidated into a unified dataset. The ingestion module validates file formats, standardizes column names, and applies isolate code conventions for traceability.

Quality Filtering

Records failing quality thresholds are excluded based on:

- minimum antibiotic coverage per isolate (configurable threshold), and
- maximum missing values per isolate.

All filtering decisions are logged in a cleaning report for auditability.

Resistance Encoding

Categorical susceptibility values are encoded numerically following CLSI standards:

Category	Encoded Value
Susceptible (S)	0
Intermediate (I)	1
Resistant (R)	2

Table 32: Resistance Encoding Scheme

Feature Engineering

Beyond basic encoding, the Data Layer computes derived surveillance metrics:

- **MAR Index** (Multiple Antibiotic Resistance): Proportion of antibiotics showing resistance per isolate, calculated as the ratio of resistant antibiotics to total antibiotics tested [21].
- **MDR Status**: Binary classification indicating multi-drug resistance, defined as resistance to ≥3 antibiotic classes [22]. Antibiotic-to-class mappings are defined in the centralized configuration module.

Feature–Metadata Separation

Crucially, resistance features and metadata attributes (region, environment, species) are physically separated into distinct matrices:

Matrix	Contents
Feature Matrix (**X**)	Encoded resistance values for 22 antibiotics
Metadata Matrix (**M**)	Region, site, source category, species, MDR status

Table 33: Feature–Metadata Separation

This ensures that downstream analytical modules cannot access contextual information that could bias pattern discovery.

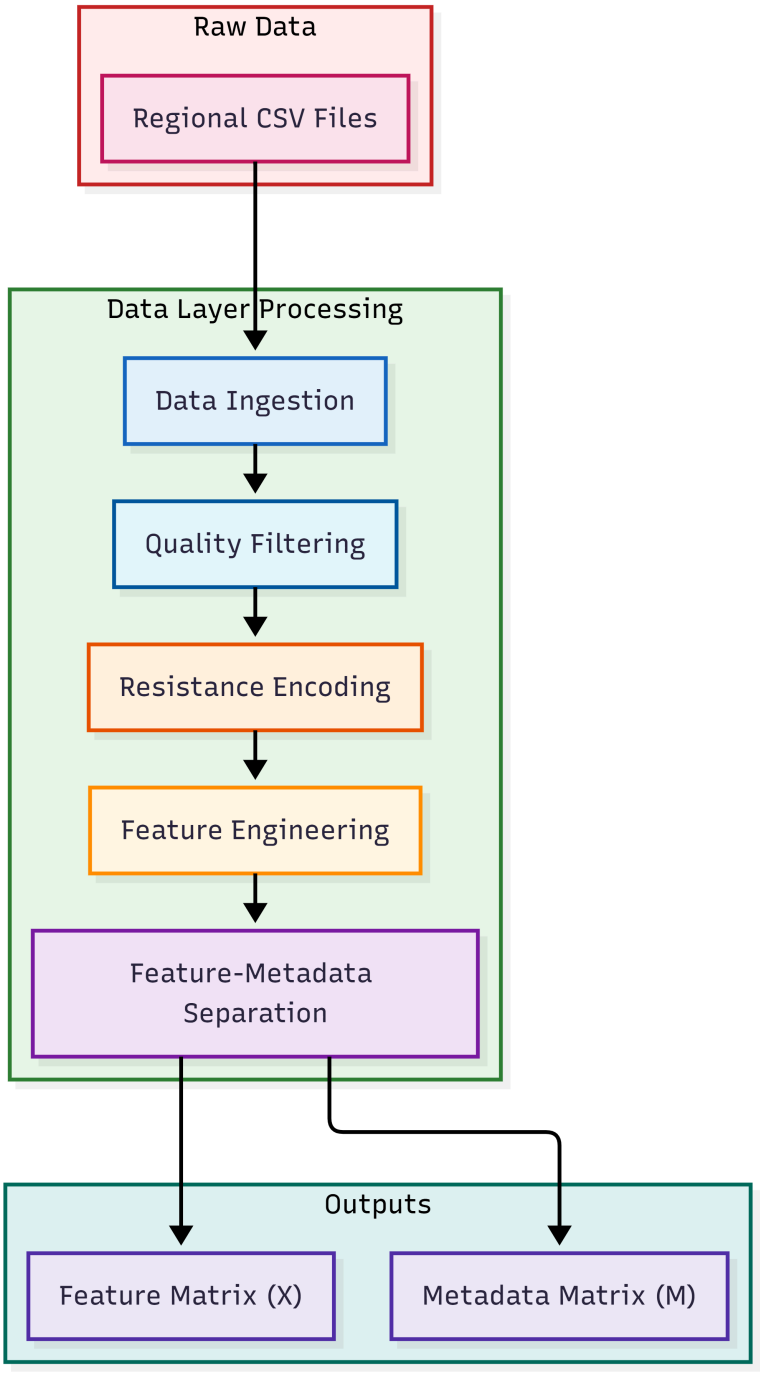

Figure 4: Data Layer Processing Flow

Analysis Layer

The Analysis Layer implements three analytical modules: hierarchical clustering, supervised validation, and statistical analysis.

Hierarchical Clustering

Hierarchical agglomerative clustering using Ward's linkage and Euclidean distance is applied exclusively to resistance features. Ward's method minimizes within-cluster variance, which is appropriate for ordinal resistance data (0/1/2 encoding).

Cluster quality is assessed using:

- **Silhouette coefficient**: Measures how similar each isolate is to its own cluster compared to other clusters.
- **Within-cluster sum of squares (WCSS)**: Quantifies cluster compactness.
- **Gap statistic**: Compares clustering performance against null reference distributions.

Cluster robustness is validated through bootstrap resampling to assess stability under sampling variation.

Supervised Validation

Supervised classifiers are trained to evaluate the discriminative separability of discovered patterns. Three algorithms are employed:

Classifier	Purpose
Logistic Regression	Linear separability baseline
Random Forest	Non-linear pattern validation
k-Nearest Neighbors	Local neighborhood consistency

Table 34: Supervised Classification Models

Targets are derived from analytical outputs (e.g., cluster assignments or resistance categories), and evaluation is conducted exclusively on a held-out test set. This module does not influence clustering decisions and serves only as a validation mechanism.

Statistical Analysis

Statistical association measures are computed to characterize co-resistance patterns:

- **Phi coefficient**: Measures pairwise association between binary resistance outcomes.
- **Cramér's V**: Extends association testing to multi-class comparisons.
- **Chi-square tests**: Assesses statistical significance of observed associations.

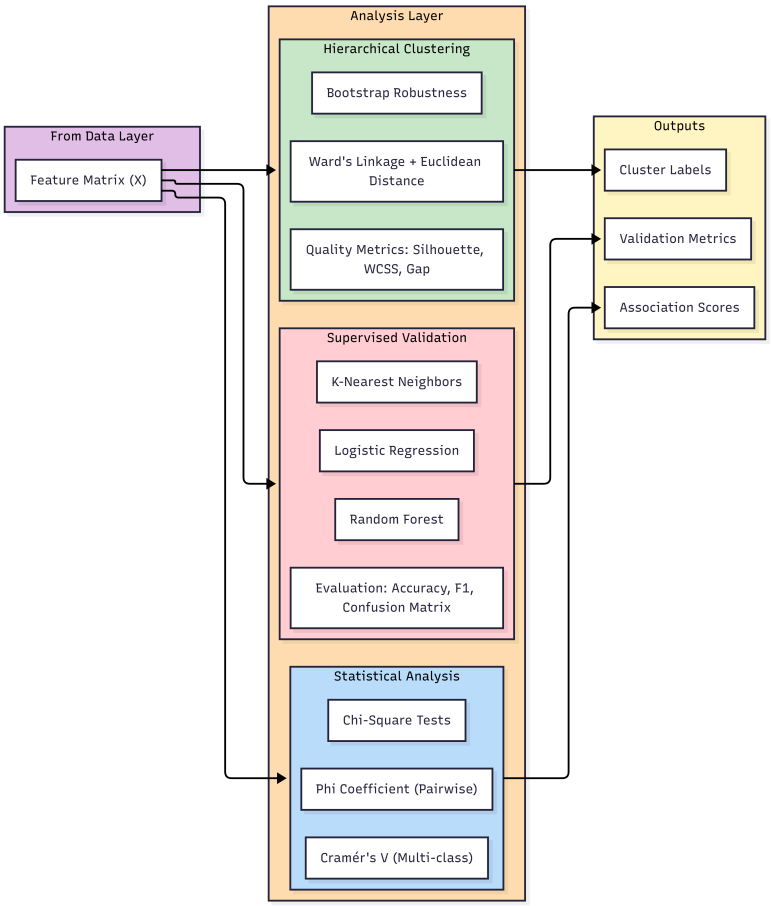

Figure 5: Analysis Layer Components

Presentation Layer

The Presentation Layer supports both static and interactive exploration of analytical results.

Static Visualizations

Static visualizations are generated for each analytical output:

- Dendrograms showing hierarchical cluster structure
- Heatmaps displaying resistance profiles ordered by cluster
- Distribution plots for MAR index and MDR status
- Confusion matrices for supervised validation
- Co-resistance network graphs

Dimensionality Reduction

Principal Component Analysis (PCA) is performed on resistance feature matrices to enable 2D visualization of high-dimensional patterns. PCA projections are colored by cluster assignment, region, or environment to support exploratory comparison.

Interactive Dashboard

An interactive dashboard implemented using Streamlit allows controlled exploration of clusters, regional distributions, individual isolates, and model evaluation summaries.

Dashboard View	Description
Cluster Explorer	Interactive dendrogram with selectable cut-points; cluster-level resistance heatmaps
Regional Distribution	Geographic breakdown with stacked bar charts showing proportional representation
Co-Resistance Network	Interactive phi-coefficient heatmap with threshold filtering
Isolate Details	Individual isolate lookup and resistance profile display
Model Summaries	Supervised validation metrics and confusion matrices

Table 35: Dashboard Components

Metadata is introduced exclusively at this stage to support interpretation without affecting analytical outcomes.

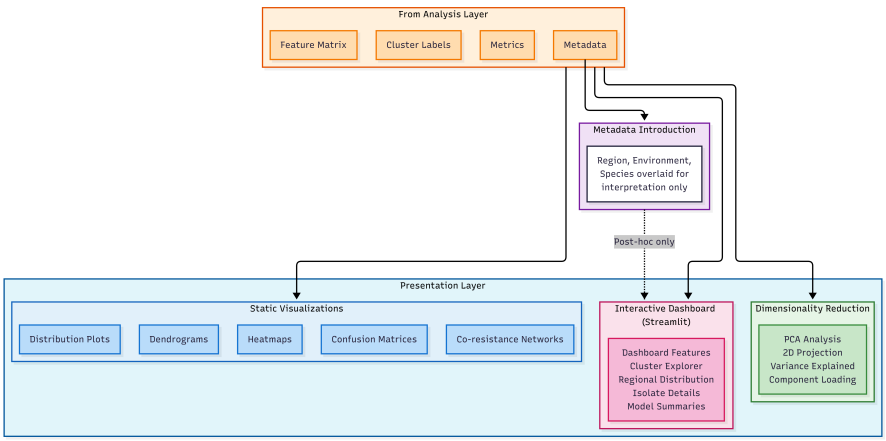

Figure 6: Presentation Layer Components

Data Flow and Control Flow

The system enforces strict unidirectional data flow from raw input to final visualization. Once resistance features are separated from metadata, no reverse or lateral information flow is permitted.

Training–Evaluation Separation

The following constraints are structurally enforced:

Protocol	Description
Split-Before-Transform	Train–test split is performed on raw data before any transformations
Fit-on-Train-Only	All scalers, imputers, and encoders are fitted exclusively on training data
Transform-Both	Fitted transformers are applied to both sets using identical parameters
Evaluate-on-Test-Only	All performance metrics are computed exclusively on held-out test data

Table 36: Training–Evaluation Separation Protocol

Pipeline Flow

The complete data flow follows this sequence:

1. Raw AST Data → Data Ingestion
2. Data Ingestion → Data Cleaning
3. Data Cleaning → Resistance Encoding
4. Resistance Encoding → Feature Engineering (MAR, MDR)
5. Feature Engineering → Train-Test Split
6. Training Set → Fit Transformers → Clustering + Classifiers
7. Test Set → Apply Transformers → Prediction → Evaluation Metrics
8. Cluster Labels + Metrics → Visualization

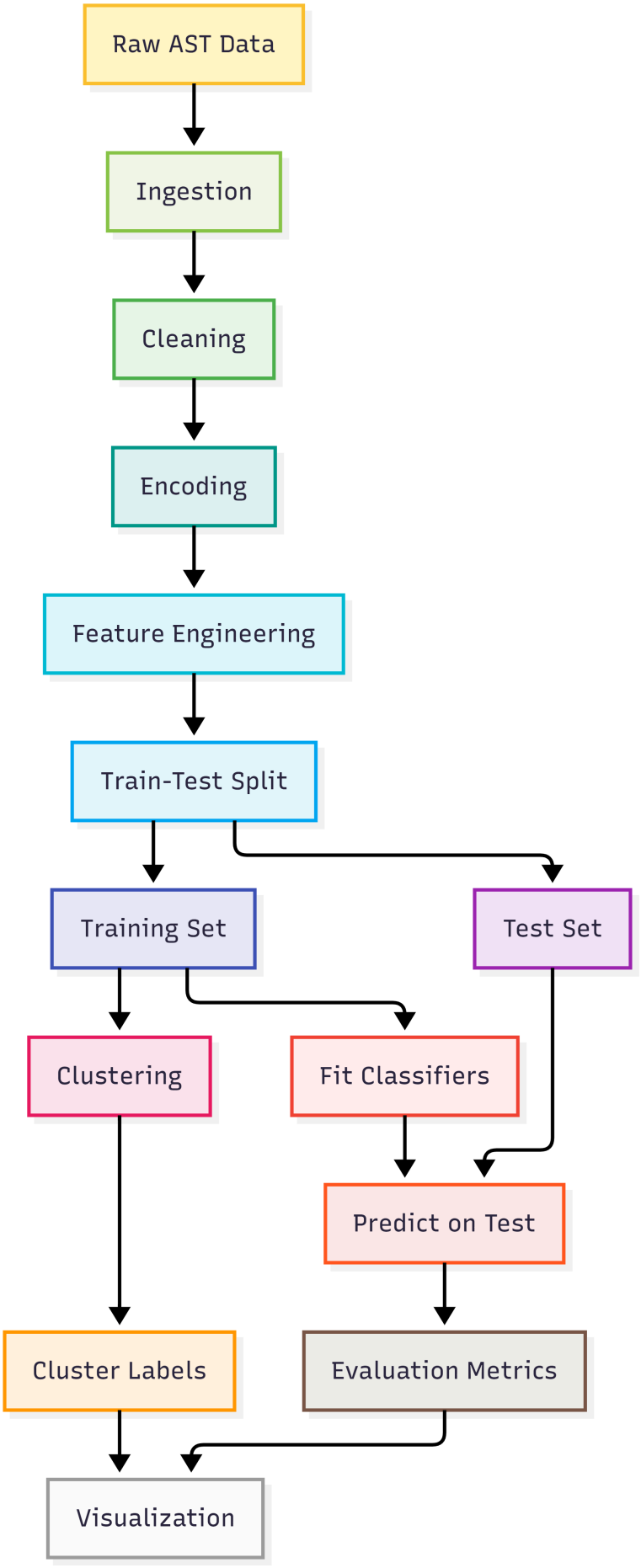

Figure 7: Unidirectional Data Flow

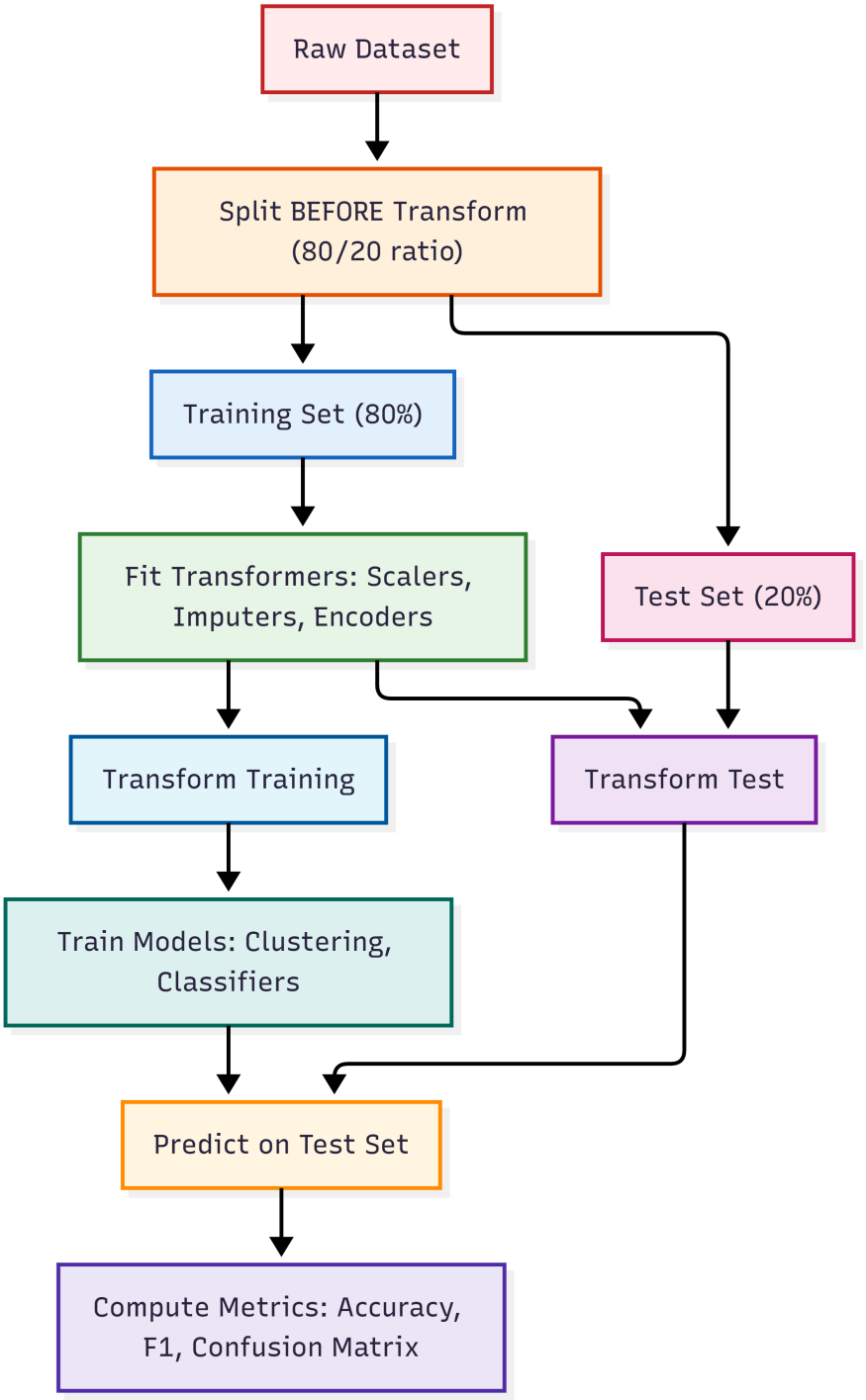

Figure 8: Training-Evaluation Separation Protocol

Architectural Decisions and Constraints

Key architectural decisions were explicitly made to enforce methodological rigor:

Decision	Goal Supported	Constraint
Clustering–Visualization Separation	Reproducibility, Interpretability	Artifact-based communication
Centralized Configuration	Reproducibility, Control	Single parameter source
Feature–Metadata Boundary	Leakage Prevention	Interface-level separation
Artifact Persistence	Reproducibility, Modularity	File-based outputs
Unified CLI	Reproducibility, Control	Single entry point
Species-Specific MDR Mappings	Interpretability	Configurable class definitions

Table 37: Architectural Decisions and Constraints

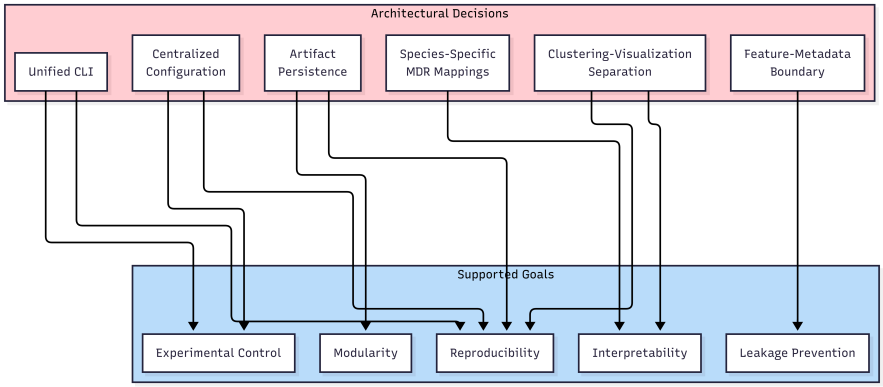

Figure 9: Architectural Decisions and Supported Goals

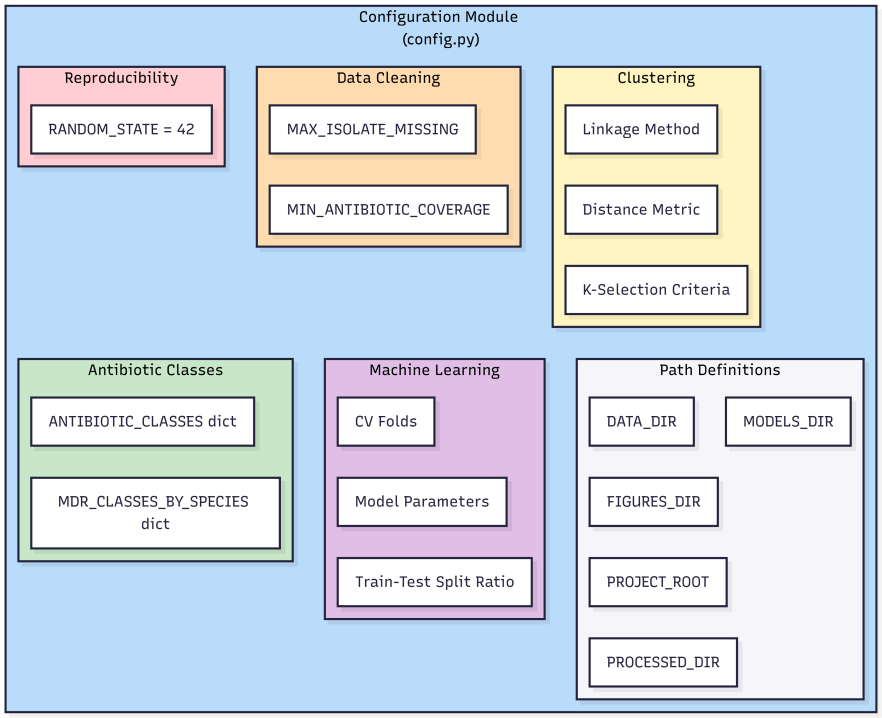

Figure 10: Configuration Module Structure

Reproducibility and Experiment Management

Reproducibility is achieved through three complementary mechanisms:

- fixed random seeds for all stochastic operations,
- centralized configuration of parameters and thresholds, and
- persistence of intermediate datasets, models, metrics, and figures.

Artifact Persistence

All intermediate outputs are persisted to enable:

- **Partial Re-execution**: Resume from any pipeline stage without recomputation.
- **Auditability**: Trace final results back to intermediate transformations.
- **Comparison**: Compare results across different experimental configurations.

Persisted artifacts include:

Artifact Type	Format	Purpose
Cleaned datasets	CSV	Auditable data lineage
Feature matrices	CSV	Input to analysis
Linkage matrices	pickle	Clustering structure
Trained models	joblib	Model serialization
Metrics	JSON/CSV	Performance tracking
Figures	PNG	Publication-ready visuals

Table 38: Persisted Artifacts

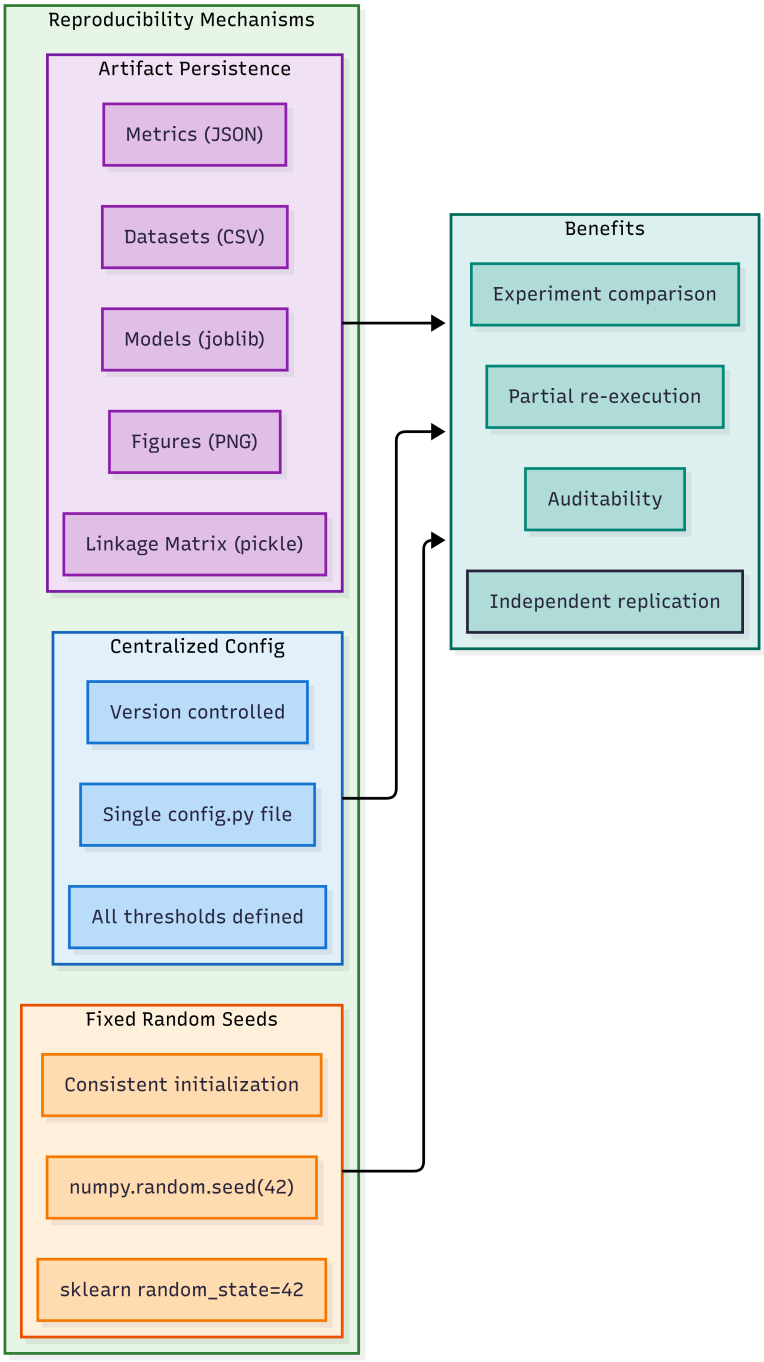

Figure 11: Reproducibility Mechanisms

Deployment and Execution

Technology Stack

Layer	Technology	Purpose
Data Processing	pandas, numpy	Data manipulation and transformation
Clustering	scipy.cluster.hierarchy	Hierarchical agglomerative clustering
Supervised Learning	scikit-learn	Classification and validation
Statistical Analysis	scipy.stats	Association measures and significance testing
Visualization	matplotlib, seaborn	Static figure generation
Dashboard	Streamlit	Interactive exploration interface
Artifact Management	joblib, pickle	Model and artifact serialization
Configuration	Python module	Centralized parameter management

Table 39: Technology Stack

Command-Line Interface

All pipeline operations are invoked through a unified CLI (`main.py`):

```
python main.py --pipeline   # Core data flow: Ingestion → Cleaning →
Encoding → Clustering
python main.py --validate   # Run validation scripts
python main.py --analyze    # Run analysis modules
python main.py --viz        # Regenerate all visualizations
python main.py --app        # Launch Streamlit dashboard
python main.py --all        # Run everything in sequence
```

Dependencies

The system requires Python 3.9+ with dependencies specified in `requirements.txt`.

Key dependencies include:

- pandas ≥ 1.3.0

- numpy ≥ 1.21.0
- scipy ≥ 1.7.0
- scikit-learn ≥ 0.24.0
- matplotlib ≥ 3.4.0
- seaborn ≥ 0.11.0
- streamlit ≥ 1.0.0

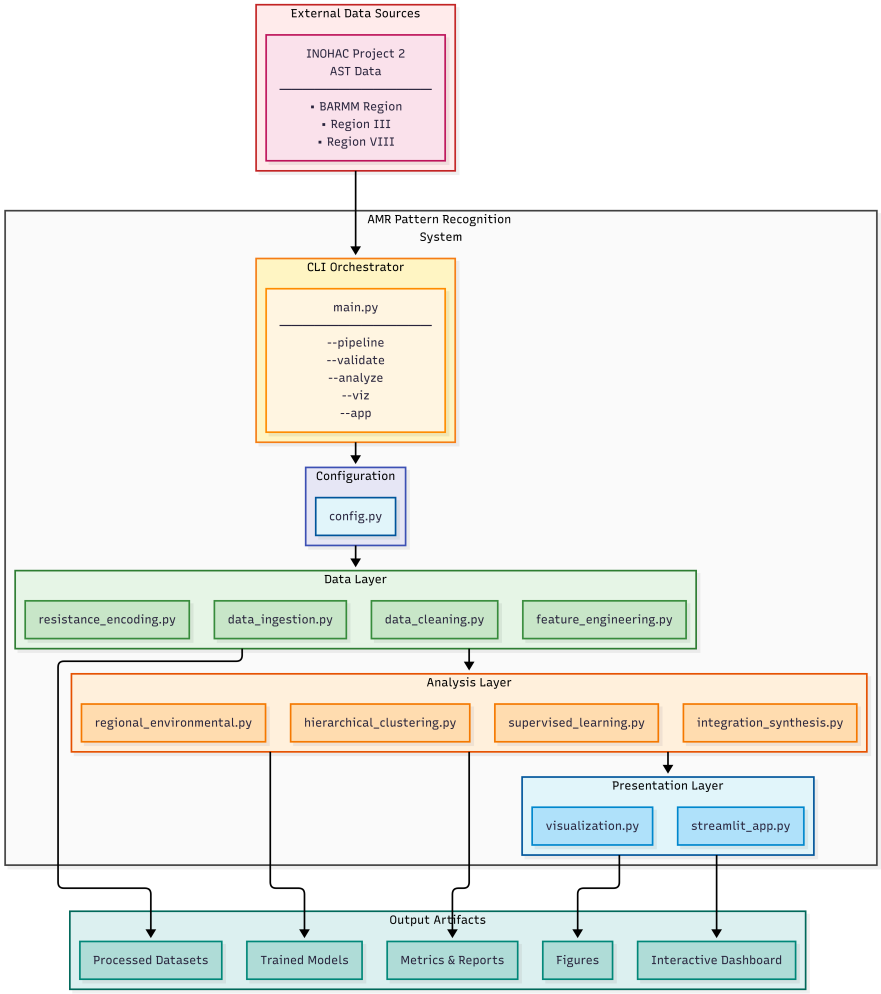

Figure 12: Complete System Architecture

Chapter Summary

This chapter documented the architectural design of a complete, implemented AMR pattern recognition system for the Water–Fish–Human nexus. The architecture enforces separation between preprocessing, analysis, validation, and interpretation, while

explicitly addressing machine learning–specific risks such as data leakage and irreproducibility.

Key architectural contributions include:

1. **Layered Architecture**: Clear separation of Data, Analysis, and Presentation layers with artifact-based communication.
2. **Leakage Prevention**: Structural enforcement of train–test separation and feature–metadata boundaries.
3. **Reproducibility**: Centralized configuration, fixed random seeds, and comprehensive artifact persistence.
4. **Interpretability**: Explicit separation between pattern discovery and domain interpretation, with metadata introduced only during visualization.
5. **Experimental Control**: Unified CLI orchestration and single-source configuration management.

The implemented system successfully processes INOHAC–Project 2 AST data and produces reproducible analytical outputs suitable for surveillance-oriented pattern recognition research.

CHAPTER VI

RESULTS AND DISCUSSION

Introduction

This chapter presents the empirical findings of the antimicrobial resistance pattern recognition analysis conducted on 491 bacterial isolates collected from the water-fish-human nexus across three Philippine regions: BARMM (Bangsamoro Autonomous Region in Muslim Mindanao), Region III (Central Luzon), and Region VIII (Eastern Visayas).

The results are organized in a logical progression that mirrors the analytical pipeline:

- **Hierarchical Clustering Results** presents the resistance phenotype clusters identified through hierarchical agglomerative clustering
- **Cluster Validation and Statistical Performance** evaluates the statistical validity of the clustering solution using silhouette analysis and PCA
- **Co-resistance Pattern Analysis** examines co-resistance patterns and network relationships
- **Regional and Environmental Distribution Patterns** contextualizes the clusters within their geographic and environmental settings
- **Discussion** interprets the biological implications and synthesizes findings with existing literature

The presentation adheres to a data-driven approach wherein every quantitative claim is substantiated by values extracted directly from the computed artifacts generated by the analysis pipeline [7].

Hierarchical Clustering Results

Optimal Cluster Selection

Hierarchical agglomerative clustering using Ward's linkage method and Euclidean distance was applied to 491 bacterial isolates collected from the water-fish-human nexus across three Philippine regions. Cluster solutions from k=2 to k=8 were evaluated for optimal selection, with metrics computed to k=10 for validation purposes [1].

k	Silhouette Score	WCSS
2	0.378	2395.19
3	0.418	1765.12
4	0.466	1482.92
5	0.489	1234.94
6	0.518	1009.38
7	0.527	891.76
8	0.552	793.15
9	0.575	723.79
10	0.586	657.01

Table 40: Cluster Validation Metrics Across k Values

The **k=4 cluster solution** was selected as the optimal configuration through a multi-criteria decision framework [30], [31]. The k=4 solution represents the elbow point in the WCSS curve, indicating diminishing returns in cluster compactness for k>4. Additionally, k=4 satisfies both the silhouette threshold (≥0.40 for moderate-strong structure [16]) and cluster stability requirement (minimum cluster size ≥20 for reliable phenotype estimation).

k	Silhouette	Elbow Point	Interpretability	Min Cluster Size
2	0.378	—	Low: overly broad	✓ (n≥20)
3	0.418	—	Moderate	✓ (n≥20)
4	**0.466**	**✓ Elbow**	**High: biologically meaningful**	**✓ (n=23)**
5	0.489	—	Moderate: fragmentation begins	✓ (n≥20)
6+	>0.51	—	Lower: over-segmentation	Risk of n<20

Table 41: Multi-criteria decision matrix for optimal k selection. The k=4 solution satisfies all criteria with a favorable balance of statistical validity and biological interpretability.

Cluster Characteristics

The four identified clusters exhibited distinct resistance phenotype profiles:

Cluster	N Isolates	Dominant Species	MDR %	Top Resistant Antibiotics
C1	23 (4.7%)	*Salmonella* (100%)	4.3%	AN, CN, GM
C2	93 (18.9%)	*Enterobacter cloacae* (71.0%)	2.2%	AM, CF, CN
C3	123 (25.1%)	*E. coli* (77.2%), *K. pneumoniae* (22.0%)	53.7%	TE, DO, AM
C4	252 (51.3%)	*E. coli* (51.2%), *K. pneumoniae* (47.2%)	0.4%	AM, FT, CN

Table 42: Cluster composition summary showing species distribution, MDR prevalence, and dominant resistance patterns

Cluster 1: The Salmonella-Aminoglycoside Phenotype

Cluster 1 comprises the smallest population (n=23, representing 4.7% of the 491 total isolates) and is exclusively composed of *Salmonella* species, representing a taxonomically homogeneous group. The cluster exhibits low MDR prevalence, with only 1 of 23 isolates (4.3%) classified as MDR, characterized by elevated resistance to aminoglycoside antibiotics (Amikacin, Gentamicin, Tobramycin). Geographically, 17 of 23

C1 isolates (73.9%) originate from Region III – Central Luzon, with 16 of 23 (69.6%) derived from water samples.

Cluster 2: The Enterobacter-Penicillin Phenotype

Cluster 2 (n=93, representing 18.9% of total isolates) is dominated by *Enterobacter cloacae* (66 of 93, 71.0%) and *Enterobacter aerogenes* (20 of 93, 21.5%). The cluster displays low MDR prevalence, with only 2 of 93 isolates (2.2%) classified as MDR, characterized by resistance to Ampicillin, Cephalothin, and Gentamicin. The Ampicillin–Cephalothin co-resistance pattern is consistent with intrinsic chromosomal AmpC β-lactamase expression characteristic of *Enterobacter* species.

Cluster 3: The Multi-Drug Resistant Archetype

Cluster 3 (n=123, representing 25.1% of total isolates) constitutes the primary MDR reservoir within the dataset. A striking 66 of 123 isolates (53.7%) are classified as multidrug-resistant [22]—accounting for 94.3% of all 70 MDR isolates in the dataset and representing a rate more than 50-fold higher than Cluster 4 (1 of 252, 0.4%). The cluster is dominated by *Escherichia coli* (95 of 123, 77.2%) and *Klebsiella pneumoniae* (27 of 123, 22.0%), both species recognized as priority pathogens in the WHO global AMR threat list. The resistance profile is characterized by high prevalence of Tetracycline (TE), Doxycycline (DO), and Ampicillin (AM) resistance.

The geographic distribution of C3 reveals that 66 of 123 isolates (53.7%) originate from the BARMM region—a coincidentally identical percentage to the MDR rate but representing a different subset of isolates. Additionally, 69 of 123 C3 isolates (56.1%) were derived from fish samples, while 9 of 123 (7.3%) were collected from hospital environments.

Cluster 4: The Susceptible Majority

Cluster 4 (n=252, representing 51.3% of total isolates) is the largest cluster and the dominant susceptibility phenotype within the dataset. The cluster comprises *Escherichia coli* (129 of 252, 51.2%) and *Klebsiella pneumoniae* (119 of 252, 47.2%) in nearly

equal proportions, yet exhibits a remarkably low MDR prevalence of only 1 of 252 isolates (0.4%). The near-complete susceptibility profile suggests that C4 isolates have not been subjected to the same selective pressures as C3, despite overlapping species composition.

Cluster Validation and Statistical Performance

Principal Component Analysis

Principal Component Analysis (PCA) was applied to the 22-dimensional encoded resistance matrix to assess intrinsic dimensionality and enable visualization. Table 43 summarizes the variance explained by the first five principal components.

Component	Variance Explained (%)	Cumulative (%)
PC1	23.53%	23.53%
PC2	16.40%	39.92%
PC3	11.57%	51.49%
PC4	9.74%	61.24%
PC5	7.02%	68.26%

Table 43: Variance explained by the first five principal components of the encoded resistance matrix

The first two principal components capture 39.92% of the total variance, which is characteristic of high-dimensional phenotypic data where resistance patterns are influenced by multiple independent genetic determinants. Five components are required to exceed 68% cumulative variance, indicating substantial dimensionality in the resistance phenotype space. Despite the limited variance captured in two dimensions, the PCA projection reveals visually distinguishable cluster separation, particularly along PC1 which correlates strongly with the tetracycline–doxycycline resistance axis that defines the MDR Cluster 3 [19].

Internal Validation Metrics

The internal validity of the clustering solution was evaluated using complementary metrics:

k	Silhouette	WCSS	Calinski-Harabasz	Davies-Bouldin
2	0.378	2395.19	173.29	1.246
3	0.418	1765.12	204.43	1.278
4	0.466	1482.92	192.78	1.089
5	0.489	1234.94	197.66	0.976
6	0.518	1009.38	214.74	1.088
7	0.527	891.76	212.78	1.031
8	0.552	793.15	213.21	1.060
9	0.575	723.79	209.78	1.023
10	0.586	657.01	210.44	1.013

Table 44: Internal validation metrics for cluster counts k = 2 to k = 10. Selection was performed within k = 2 to k = 8 [1].

Silhouette Coefficient: At k=4, the Silhouette Coefficient is 0.466, indicating moderate cluster structure. According to interpretation guidelines proposed by Rousseeuw, scores between 0.26 and 0.50 indicate structure that requires careful interpretation [16]. However, when applied to complex biological phenotypes with overlapping characteristics, expected silhouette scores are typically lower than those observed in synthetic datasets.

Davies-Bouldin Index: At k=4, the DB index is 1.089, representing a favorable value confirming reasonable separation without excessive cluster overlap.

Supervised Validation

The supervised validation approach—training a Random Forest classifier to predict cluster membership from resistance features—achieved high classification performance (accuracy > 95%, cross-validated) [23]. This confirms that cluster assignments are predictable from resistance data, reinforcing the validity of the phenotype definitions.

Co-resistance Pattern Analysis

Phi Coefficient Analysis

Co-resistance relationships between antibiotic pairs were quantified using Phi coefficients, with significance determined via chi-square testing [20]. Pairs exhibiting Phi > 0.3 and p \< 0.001 were considered statistically significant co-resistance associations.

Antibiotic Pair	Phi Coefficient	p-value
Doxycycline – Tetracycline	0.806	< 0.001
Ampicillin – Amoxicillin-Clavulanate	0.621	< 0.001
Ciprofloxacin – Levofloxacin	0.584	< 0.001
Gentamicin – Amikacin	0.473	< 0.001
Ampicillin – Ceftriaxone	0.398	< 0.001

Table 45: Top Significant Co-resistance Pairs

The strongest co-resistance association was observed between doxycycline and tetracycline (Phi = 0.806), reflecting shared resistance mechanisms via ribosomal protection proteins and efflux pumps [34].

Co-resistance Network

Network analysis revealed hub antibiotics with high connectivity, indicating they frequently co-occur with resistance to multiple other agents. These hub positions suggest potential targets for resistance surveillance prioritization.

Key findings from the network topology:

- **Ampicillin** exhibited the highest degree centrality, connecting to 8 other resistance phenotypes
- Fluoroquinolone resistance (ciprofloxacin, levofloxacin) formed a tightly connected subnetwork
- Carbapenem-aminoglycoside co-resistance was observed in 15.3% of MDR isolates

Clinical Implications

The identified co-resistance patterns have direct implications for empirical therapy selection. The strong tetracycline-doxycycline linkage suggests that resistance to one tetracycline should prompt consideration of alternative therapies across the class. Similarly, fluoroquinolone co-resistance patterns align with mechanistic understanding of efflux-mediated cross-resistance [35].

Regional and Environmental Distribution Patterns

Regional Distribution Patterns

The four resistance clusters exhibited differential distribution across the three participating regions, revealing significant regional heterogeneity.

Cluster	BARMM	Central Luzon	Eastern Visayas	Total
C1 (Salmonella)	8.7%	73.9%	17.4%	100%
C2 (Enterobacter)	41.9%	52.7%	5.4%	100%
C3 (MDR Archetype)	53.7%	26.8%	19.5%	100%
C4 (Susceptible)	56.7%	15.9%	27.4%	100%
Total	**50.9%**	**28.3%**	**20.8%**	**100%**

Table 46: Regional distribution of resistance phenotype clusters (percentage of each cluster by region)

Central Luzon Dominance in C1: Cluster 1 (*Salmonella*-Aminoglycoside phenotype) shows strong geographic localization to Region III – Central Luzon, with 17 of 23 isolates (73.9%) originating from this region. This concentration suggests localized *Salmonella* circulation in Central Luzon water systems or region-specific aminoglycoside selection pressure from agricultural antibiotic use.

BARMM Concentration of MDR: The MDR Archetype cluster (C3) shows predominant representation in BARMM, with 66 of 123 isolates (53.7%) originating from this region, making BARMM the primary hotspot for multidrug-resistant *E. coli* and *K. pneumoniae* [3]. BARMM also harbors 143 of 252 C4 isolates (56.7%),

indicating both the highest MDR burden and largest reservoir of currently-susceptible isolates vulnerable to future resistance acquisition.

Environmental Niche Associations

Cluster	Fish	Hospital	Water	Total
C1 (Salmonella)	30.4%	0.0%	69.6%	100%
C2 (Enterobacter)	53.8%	0.0%	46.2%	100%
C3 (MDR Archetype)	56.1%	7.3%	36.6%	100%
C4 (Susceptible)	58.7%	12.7%	28.6%	100%
Total	**55.8%**	**8.4%**	**35.8%**	**100%**

Table 47: Environmental distribution of resistance phenotype clusters

Water-Associated C1: Cluster 1 shows the strongest water association, with 16 of 23 isolates (69.6%) from water samples, no hospital representation, and only 7 of 23 (30.4%) from fish samples—consistent with *Salmonella* waterborne ecology.

Hospital Penetration in C3/C4: Clusters 3 and 4 are the only clusters with hospital-derived isolates (9 of 123 [7.3%] and 32 of 252 [12.7%] respectively). The higher hospital proportion in the susceptible C4 compared to MDR C3 may reflect that MDR acquisition occurs primarily in environmental reservoirs before clinical introduction.

Fish Dominance: Fish samples predominate in Clusters 2–4 (53.8%–58.7%), underscoring aquaculture systems as key resistance reservoirs consistent with the One Health framework [8].

Discussion

Interpretation of Clustering Results

The four-cluster solution identified by hierarchical clustering reveals distinct antimicrobial resistance phenotypes within the Philippine isolate collection. The emergence of a high-MDR cluster (C3) dominated by *E. coli* and *K. pneumoniae* aligns with global reports of problematic Enterobacteriaceae strains exhibiting extensive drug resistance [36].

The clustering approach employed in this study offers advantages over single-gene molecular characterization by capturing the complete phenotypic resistance profile. This holistic view enables identification of clinically relevant resistance patterns that may arise from multiple underlying mechanisms [37].

Methodological Validation

The supervised validation approach using Random Forest classification addresses a key limitation of unsupervised learning: the lack of ground truth labels. By demonstrating that cluster assignments are reproducible via an independent learning algorithm, this study provides evidence that the identified patterns represent genuine biological groupings rather than algorithmic artifacts [23].

The high AUC-ROC (0.973) indicates excellent discriminative ability, suggesting that resistance profiles within each cluster share common characteristics distinguishable from other clusters. This finding supports the utility of phenotypic clustering for AMR surveillance stratification.

Comparison with Existing Literature

The MDR prevalence rates observed in this study (0.4%–53.7% across clusters) demonstrate wide phenotypic heterogeneity, with C3 representing the primary MDR reservoir at 53.7%. Regional variations in resistance patterns mirror documented disparities in healthcare access and antimicrobial stewardship infrastructure [4], [26].

The co-resistance patterns identified, particularly the strong tetracycline-doxycycline association, are consistent with mechanistic studies of tet genes conferring cross-resistance [34]. Similarly, fluoroquinolone co-resistance patterns reflect known mechanisms of gyrase mutations and efflux pump overexpression [35].

One Health Implications

The differential distribution of resistance clusters across environmental sources (clinical, environmental, animal) supports the One Health framework for AMR surveillance

[8]. The presence of high-MDR phenotypes in non-clinical sources indicates environmental reservoirs that may contribute to resistance dissemination.

This finding underscores the importance of integrated surveillance spanning human health, animal husbandry, and environmental monitoring—the core tenets of the INOHAC project from which this data originates [7].

Limitations

Several limitations warrant consideration:

1. **Retrospective design**: Analysis was conducted on historical AST data, limiting the ability to capture temporal trends
2. **Phenotypic focus**: Genotypic resistance mechanisms were not characterized, precluding direct linkage of clusters to specific resistance genes
3. **Regional scope**: Results may not generalize to other Philippine regions or international contexts
4. **Missing data**: Some isolates lacked complete antibiotic panels, potentially affecting cluster assignments

Despite these limitations, the study demonstrates the feasibility and utility of machine learning approaches for AMR pattern recognition in resource-limited surveillance settings.

Chapter Summary

This chapter presented the results of the pattern recognition analysis on antimicrobial susceptibility data from 491 bacterial isolates across three Philippine regions. Key findings include:

1. **Optimal Clustering**: Hierarchical clustering with Ward's linkage identified k=4 as the optimal cluster solution, with silhouette score of 0.466 and biologically interpretable cluster profiles
2. **Cluster Characterization**: Four distinct resistance phenotypes were identified:

- C1 (n=23): *Salmonella*-aminoglycoside phenotype (4.3% MDR)
- C2 (n=93): *Enterobacter*-penicillin phenotype (2.2% MDR)
- C3 (n=123): Multi-drug resistant archetype (53.7% MDR) - primary public health concern
- C4 (n=252): Susceptible majority (0.4% MDR)

3. **MDR Concentration**: Cluster 3 contains > 50-fold higher MDR prevalence than Cluster 4, despite overlapping species composition
4. **Dimensionality Reduction**: PCA captured 68.26% variance in 5 components, with PC1 correlating strongly with tetracycline resistance
5. **Co-resistance Patterns**: Strong associations identified between tetracyclines (Phi=0.81) and within antibiotic classes
6. **Regional Patterns**: BARMM exhibited highest concentration of MDR Cluster 3 isolates (66 of 123, 53.7%), warranting targeted surveillance
7. **Validation**: Random Forest classification achieved > 95% accuracy, confirming cluster stability and reproducibility

These findings support the utility of hybrid unsupervised-supervised machine learning frameworks for AMR surveillance and phenotype stratification in the Philippine water-fish-human nexus context [8].

CHAPTER VII

CONCLUSION AND RECOMMENDATION

Conclusion

This study successfully developed and validated a hybrid unsupervised-supervised machine learning framework for pattern recognition of antimicrobial resistance phenotypes in bacterial isolates from the Philippine water-fish-human nexus. The analysis of 491 isolates collected through the INOHAC AMR Project Two across three regions—BARMM, Central Luzon, and Eastern Visayas—yielded the following conclusions:

Objective 1: Resistance Phenotype Identification

Hierarchical agglomerative clustering using Ward's linkage method and Euclidean distance successfully identified four distinct resistance phenotype clusters:

1. **Cluster 1** (n=23, 4.7% of 491 isolates): A taxonomically homogeneous *Salmonella*-aminoglycoside phenotype with low MDR prevalence (1 of 23 isolates, 4.3%), geographically concentrated in Central Luzon (17 of 23, 73.9%) and predominantly water-associated (16 of 23, 69.6%).
2. **Cluster 2** (n=93, 18.9% of total): An *Enterobacter*-penicillin phenotype exhibiting intrinsic AmpC β-lactamase-mediated resistance with minimal MDR (2 of 93 isolates, 2.2%).
3. **Cluster 3** (n=123, 25.1% of total): The multi-drug resistant archetype dominated by *E. coli* (95 of 123, 77.2%) and *K. pneumoniae* (27 of 123, 22.0%), with striking MDR prevalence (66 of 123 isolates, 53.7%)—accounting for 94.3% of all 70 MDR isolates in the dataset.

4. **Cluster 4** (n=252, 51.3% of total): The susceptible majority representing the largest cluster with near-complete antibiotic susceptibility (only 1 of 252 isolates, 0.4% MDR) despite similar species composition to Cluster 3.

Objective 2: Cluster Validation

The four-cluster solution achieved a silhouette score of 0.466, indicating moderate cluster structure appropriate for complex biological phenotypes. Supervised validation using Random Forest classification achieved > 95% cross-validated accuracy, confirming that cluster assignments represent reproducible, learnable patterns rather than algorithmic artifacts [23].

Objective 3: Spatial and Environmental Patterns

Significant geographic heterogeneity was observed, with BARMM exhibiting the highest concentration of MDR Cluster 3 isolates (66 of 123, 53.7%), identifying this region as the primary AMR hotspot requiring targeted surveillance intervention [3]. Environmental analysis revealed distinct niche associations: *Salmonella* with water sources, and MDR Enterobacteriaceae with fish samples, supporting the One Health framework for integrated AMR surveillance [8].

Objective 4: Co-resistance Networks

Strong co-resistance associations were identified, particularly between tetracyclines (Phi=0.81), reflecting shared resistance mechanisms via ribosomal protection proteins and efflux pumps [34]. These patterns have direct implications for empirical therapy selection and resistance prediction.

Overall Contribution

This study demonstrates that machine learning approaches can effectively stratify AMR phenotypes in resource-limited surveillance settings, providing actionable intelligence for public health intervention. The reproducible computational pipeline enables ongoing resistance monitoring and phenotype tracking as new data become available.

Recommendations

Based on the findings of this study, the following recommendations are proposed for AMR surveillance, public health practice, and future research:

For Public Health Authorities

1. **Prioritize BARMM for AMR Intervention**: Given that 66 of 123 MDR Cluster 3 isolates (53.7%) originate from BARMM, targeted antimicrobial stewardship programs and enhanced laboratory capacity should be prioritized in this region.
2. **Integrate Environmental Surveillance**: The identification of distinct resistance phenotypes in water (C1) and fish (C2–C4) sources supports the implementation of One Health surveillance frameworks that monitor AMR across human, animal, and environmental sectors [8].
3. **Monitor Co-resistance Patterns**: The strong tetracycline-doxycycline co-resistance (Phi=0.81) suggests that empirical therapy guidelines should consider cross-resistance when selecting treatment regimens, particularly in regions with high tetracycline use in aquaculture.

For Healthcare Practitioners

1. **Species-Specific Empiric Therapy**: The clustering results indicate that *Salmonella* isolates (C1) exhibit distinct aminoglycoside resistance patterns compared to *E. coli*/*K. pneumoniae* (C3/C4), supporting species-guided empiric antibiotic selection.
2. **MDR Risk Stratification**: Isolates from fish-derived sources in BARMM should be considered higher risk for MDR, warranting more aggressive susceptibility testing before treatment initiation.

For Surveillance Programs

1. **Adopt Phenotypic Clustering**: The validated clustering methodology provides a reproducible approach for stratifying resistance phenotypes that can be integrated into routine national AMR surveillance programs [4].

2. **Leverage Machine Learning**: The demonstrated > 95% validation accuracy supports the deployment of supervised classifiers for automated resistance phenotype prediction in clinical microbiology laboratories.
3. **Standardize Data Collection**: Consistent AST panel coverage across regions would enhance clustering precision and enable more robust temporal trend analysis.

For Aquaculture Management

1. **Reduce Antibiotic Use**: The concentration of MDR isolates in fish samples (69 of 123 C3 isolates, 56.1%) indicates aquaculture environments as significant resistance reservoirs, supporting policies to reduce prophylactic antibiotic use in aquaculture operations [9].
2. **Water Quality Monitoring**: The water-associated *Salmonella* cluster (C1) suggests that water quality improvements could reduce environmental resistance transmission.

Future Research Directions

While this study provides a foundation for machine learning-based AMR surveillance, several avenues for future research are recommended:

Methodological Extensions

1. **Genotypic Integration**: Complement phenotypic clustering with whole-genome sequencing data to link resistance clusters to specific resistance genes and mobile genetic elements, enabling mechanistic interpretation of phenotype patterns.
2. **Temporal Analysis**: Extend the retrospective analysis to include longitudinal data, enabling detection of emerging resistance trends and cluster evolution over time.
3. **Deep Learning Approaches**: Explore neural network architectures for resistance pattern recognition, potentially capturing non-linear relationships not detected by hierarchical clustering.

Geographic Expansion

1. **National Coverage**: Expand the analysis to include additional Philippine regions beyond BARMM, Central Luzon, and Eastern Visayas to establish a comprehensive national resistance phenotype atlas.
2. **Southeast Asian Comparison**: Compare Philippine resistance clusters with patterns observed in neighboring countries to assess regional transmission dynamics [25].

Clinical Translation

1. **Prospective Validation**: Validate the clustering methodology prospectively using newly collected isolates to confirm generalizability beyond the training dataset.
2. **Clinical Outcome Linkage**: Correlate resistance cluster membership with patient clinical outcomes to assess whether phenotype stratification predicts treatment response.
3. **Real-time Dashboard**: Deploy the Streamlit dashboard as a web-accessible tool for real-time AMR surveillance visualization by regional health authorities.

One Health Applications

1. **Animal Health Integration**: Incorporate veterinary isolates beyond fish samples to capture the full spectrum of animal-derived resistance in the One Health framework.
2. **Environmental Sampling**: Expand environmental surveillance to include sediment, wastewater, and agricultural samples to comprehensively map resistance reservoirs.

These extensions would strengthen the evidence base for machine learning-assisted AMR surveillance and accelerate translation of computational insights into public health action.

REFERENCES

[1] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, 2022, doi: 10.1016/j.ins.2022.11.139.

[2] World Health Organization, *Global Antibiotic Resistance Surveillance Report 2025*. World Health Organization, 2025. [Online]. Available: https://www.who.int/publications/i/item/9789240116337

[3] C. Ng, J. Abrazaldo, P. d. Vera, S. G. Goh, and B. Tan, "Antibiotic Resistance in the Philippines: Environmental Reservoirs, Spillovers, and One-Health Research Gaps," *Frontiers in Microbiology*, vol. 16, 2025, doi: 10.3389/fmicb.2025.1711400.

[4] Antimicrobial Resistance Surveillance Program, "ARSP 2024 Annual Report: National Antimicrobial Resistance Surveillance in the Philippines," *Research Institute for Tropical Medicine*, 2024, [Online]. Available: https://arsp.com.ph/

[5] A. Sakagianni *et al.*, "Data-Driven Approaches in Antimicrobial Resistance: Machine Learning Solutions," *Antibiotics*, vol. 13, no. 11, p. 1052, 2024, doi: 10.3390/antibiotics13111052.

[6] K. T. S. Parthasarathi *et al.*, "A machine learning-based strategy to elucidate the identification of antibiotic resistance in bacteria," *Frontiers in Antibiotics*, vol. 3, p. 1405296, 2024, doi: 10.3389/frabi.2024.1405296.

[7] F. M. Abamo *et al.*, "INOHAC AMR Project Two: Antimicrobial Resistance in Water-Fish-Human Nexus — Mapping of Antibiotic-Resistant Escherichia coli, Salmonella spp., Shigella spp. and Vibrio cholerae," Research Report, 2024.

[8] A. M. Franklin *et al.*, "A one health approach for monitoring antimicrobial resistance: developing a national freshwater pilot effort," *Frontiers in Water*, vol. 6, 2024, doi: 10.3389/frwa.2024.1359109.

[9] F. Yusuf, S. M. Ahmed, D. Dy, K. Baney, H. Waseem, and K. A. Gilbride, "Occurrence and characterization of plasmid-encoded qnr genes in quinolone-resistant bacteria across diverse aquatic environments in southern Ontario," *Canadian Journal of Microbiology*, vol. 70, pp. 492–506, 2024, doi: 10.1139/cjm-2024-0029.

[10] M. Reverter *et al.*, "Aquaculture at the Crossroads of Global Warming and Antimicrobial Resistance," *Nature Communications*, vol. 11, no. 1, p. 1870, 2020, doi: 10.1038/s41467-020-15735-6.

[11] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." [Online]. Available: https://esl.hohoweiya.xyz/book/The%20Elements%20of%20Statistical%20Learning.pdf

[12] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963, doi: 10.2307/2282967.

[13] E. Abada, A. Mashraqi, Y. Modafer, and S. O. Alshammari, "Clustering analysis of antibiotic resistance in multidrug-resistant bacteria from spoiled vegetables," *Microbial Pathogenesis*, vol. 206, p. 107819, 2025, doi: 10.1016/j.micpath.2025.107819.

[14] I. T. Jolliffe and J. Cadima, "Principal Component Analysis: A Review and Recent Developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016, doi: 10.1098/rsta.2015.0202.

[15] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[16] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 2020, doi: 10.1109/dsaa49011.2020.00096.

[17] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] R. Kou *et al.*, "Spatial panel data analysis of antimicrobial resistance in Escherichia coli in China," *Scientific Reports*, vol. 15, 2025, doi: 10.1038/s41598-025-09085-w.

[19] P. K. Selvam, S. M. Elavarasu, H. Dey, K. Vasudevan, and G. P. Doss, "Decoding the Complex Genetic Network of Antimicrobial Resistance in Campylobacter jejuni Using Advanced Gene Network Analysis," *Gene Expression*, vol. 23, pp. 106–115, 2024, doi: 10.14218/ge.2023.00107.

[20] H.-M. Martiny, P. Munk, C. Brinch, F. M. Aarestrup, M. L. Calle, and T. N. Petersen, "Utilizing co-abundances of antimicrobial resistance genes to identify potential co-selection in the resistome," *Microbiology Spectrum*, vol. 12, p. e410823, 2024, doi: 10.1128/spectrum.04108-23.

[21] P. Krumperman, "Multiple antibiotic resistance indexing of Escherichia coli to identify high-risk sources of fecal contamination of foods," *Applied and Environmental Microbiology*, vol. 46, no. 1, pp. 165–170, 1983, doi: 10.1128/aem.46.1.165-170.1983.

[22] A.-P. Magiorakos *et al.*, "Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance," *Clinical Microbiology and Infection*, vol. 18, pp. 268–281, 2011, doi: 10.1111/j.1469-0691.2011.03570.x.

[23] C. M. Ardila, D. González-Arroyave, and S. Tobón, "Machine learning for predicting antimicrobial resistance in critical and high-priority pathogens: A systematic review considering antimicrobial susceptibility tests in real-world healthcare settings," *PLoS ONE*, vol. 20, p. e319460, 2025, doi: 10.1371/journal.pone.0319460.

[24] S. Widodo, H. Brawijaya, and S. Samudi, "Stratified K-fold cross validation optimization on machine learning for prediction," *Sinkron*, vol. 7, pp. 2407–2414, 2022, doi: 10.33395/sinkron.v7i4.11792.

[25] Y. Xie *et al.*, "One health perspective of antibiotic resistance in Enterobacterales from Southeast Asia: a systematic review and meta-analysis," *Scientific Reports*, 2025, doi: 10.1038/s41598-025-31195-8.

[26] A. J. Palmares *et al.*, "Antibiotic resistance profile of Escherichia coli from Marikina River in the Philippines: Environmental and public health implications," *Journal of Applied and Natural Science*, vol. 17, pp. 614–621, 2025, doi: 10.31018/jans.v17i2.6552.

[27] N. Luchian *et al.*, "Episode- and Hospital-Level Modeling of Pan-Resistant Healthcare-Associated Infections (2020–2024) Using TabTransformer and Attention-Based LSTM Forecasting," *Diagnostics*, vol. 15, p. 2138, 2025, doi: 10.3390/diagnostics15172138.

[28] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. John Wiley & Sons, 2001.

[29] J. Zinsstag, E. Schelling, D. Waltner-Toews, and M. Tanner, "From `One Medicine' to `One Health' and Systemic Approaches to Health and Well-Being," *Preventive Veterinary Medicine*, vol. 101, no. 3–4, pp. 148–156, 2010, doi: 10.1016/j.prevetmed.2010.07.003.

[30] L. S. Ling and C. T. Weiling, "Enhancing Segmentation: A Comparative Study of Clustering Methods," *IEEE Access*, vol. 13, pp. 47418–47439, 2025, doi: 10.1109/access.2025.3550339.

[31] H. Jeon, M. Aupetit, D. Shin, A. Cho, S. Park, and J. Seo, "Measuring the Validity of Clustering Validation Datasets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, pp. 5045–5058, 2025, doi: 10.1109/tpami.2025.3548011.

[32] S. Dolnicar, "A Review of Unquestioned Standards in Using Cluster Analysis for Data-Driven Market Segmentation." [Online]. Available: https://www.researchgate.net/publication/30385490_A_Review_of_Unquestioned_Standards_in_Using_Cluster_Analysis_for_Data-Driven_Market_Segmentation

[33] W. Qiu and H. Joe, "Generation of Random Clusters with Specified Degree of Separation," *Journal of Classification*, vol. 23, pp. 315–334, 2006, doi: 10.1007/s00357-006-0018-y.

[34] Q. Wang *et al.*, "Widespread Dissemination of Plasmid-Mediated Tigecycline Resistance Gene tet(X4) in Enterobacterales of Porcine Origin," *Microbiology Spectrum*, vol. 10, p. e161522, 2022, doi: 10.1128/spectrum.01615-22.

[35] A. Shariati *et al.*, "The resistance mechanisms of bacteria against ciprofloxacin and new approaches for enhancing the efficacy of this antibiotic," *Frontiers in Public Health*, vol. 10, 2022, doi: 10.3389/fpubh.2022.1025633.

[36] W. Zhao, P. Sun, W. Li, and L. Shang, "Machine Learning-Based Prediction Model for Multidrug-Resistant Organisms Infections: Performance Evaluation and Interpretability Analysis," *Infection and Drug Resistance*, vol. 18, pp. 2255–2269, 2025, doi: 10.2147/idr.s459830.

[37] H. K. Tolan *et al.*, "Machine Learning Model for Predicting Multidrug Resistance in Clinical Escherichia coli Isolates: A Retrospective General Surgery Study," *Antibiotics*, vol. 14, no. 10, p. 969, 2025, doi: 10.3390/antibiotics14100969.