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:

Al-Hanif A. Magomnang

Reynaldo A. Pahay Jr.

Adviser:

Prof. Janice F. Wade, MSCS

Co-Adviser:

Mr. Llewelyn A. Elcana

January 2026

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1. 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 [1]. 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 [2].

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 algo-

rithms, random forest classifiers, and neural networks have demonstrated utility in identifying resistance patterns, predicting phenotypes from genotypes, and stratifying patient risk [5], [6]. However, the application of these methods to environmental data, particularly in the Philippines, remains limited [3], especially 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 health, and environmental contamination, requiring integrated surveillance strategies [8].

Pattern recognition, a core computer science discipline, provides a methodological framework for analyzing such complex datasets. Unlike traditional categorical labels (e.g., "multi-drug resistant"), pattern recognition algorithms discover latent structures in resistance profiles without predefined assumptions. Hierarchical clustering groups isolates by phenotypic similarity, while supervised methods like Random Forest validate whether discovered patterns represent coherent groupings. This approach bridges raw laboratory data and actionable epidemiological insights, enabling identification of resistance phenotypes and co-resistance relationships hidden in high-dimensional susceptibility data.

1.2. 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.

1.3. Objectives of the Study

1.3.1. 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.

1.3.2. Specific Objectives

Specifically, this study aims to:

1. 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. 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. 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. 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.

1.4. Significance of the Study

This study significantly advances environmental AMR surveillance in the Philippines by validating a reproducible, unsupervised-supervised hybrid machine learning framework. By bridging phenotypic analysis with computational clustering, this research demonstrates the feasibility of high-resolution resistance profiling in resource-limited settings without relying on costly whole-genome sequencing. This methodological contribution not only fills a critical academic gap but also provides a scalable, open-access analytical pipeline that enables researchers and public health authorities to replicate these techniques for real-time phenotype monitoring.

1.5. Scope and Limitations

1.5.1. 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: Members of the family 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.

1.5.2. Limitations

1. Phenotypic Focus: This study analyzes phenotypic resistance profiles (susceptible/ intermediate/resistant) without genotypic characterization. Resistance mechanisms and mobile genetic elements are hypothesized 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.

CHAPTER 2

REVIEW OF RELATED LITERATURE

2.1. 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.

2.1.1. Unsupervised Learning for Biological Pattern Discovery

The fundamental challenge in environmental AMR surveillance lies in the absence of pre-defined 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 [9]. This constraint necessitates unsupervised approaches that discover structure directly from data without labeled examples [10].

2.1.1.1. 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 [11]. 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 = 1, Resistant = 2) introduces theoretical ambiguity, empirical evaluations demonstrate robust clustering performance with ordinal resistance data [12]. In the present study, Euclidean distance appropriateness was validated through silhouette analysis: high silhouette scores (≥0.40) confirm that the distance metric produces well-separated clusters, empirically justifying its use despite ordinal encoding.

This hierarchical approach influenced the present study's methodology by enabling discovery of resistance archetypes without requiring predefined phenotype labels.

2.1.1.2. 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 [13]. 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.

This concept influenced the present study by providing a visualization mechanism to confirm that discovered clusters occupy distinct regions in reduced-dimensional space, offering independent validation of clustering results.

2.1.1.3. Cluster Validation via Silhouette Analysis

Determining optimal cluster number remains a persistent challenge in unsupervised learning [10]. The silhouette coefficient addresses this by measuring the ratio of within-cluster cohesion to between-cluster separation [14]. Values range from −1 to +1, where scores ≥ 0.25 indicate weak structure, scores ≥ 0.40 indicate moderate-to-strong structure suitable for biological phenotype analysis, and scores ≥ 0.70 suggest exceptionally well-defined groupings [15], [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.

Silhouette analysis influenced the present study as the primary criterion for selecting optimal cluster count (k=4), ensuring discovered phenotypes represent genuine data structure rather than algorithmic artifacts.

2.1.2. 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].

This concept directly shaped the present study's two-phase methodology: unsupervised clustering for pattern discovery followed by Random Forest classification to validate that discovered clusters are biologically coherent and discriminable.

2.1.3. 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.

2.1.4. 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].

2.1.5. 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]:

$$\mathrm{MAR} = \frac{a}{b} \qquad \text{Eq. (1)}$$

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.

2.1.6. 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.

2.2. 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.

2.2.1. 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. Machine learning models, particularly Random Forest and Neural Networks, have demonstrated high predictive accuracy (Area Under the Receiver Operating Characteristic or AUROC > 0.90) in distinguishing resistance phenotypes based on genomic features [23]. However, these studies primarily rely on WGS data, which is often unavailable in resource-limited settings. 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 [24]. 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 [9].

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 [25]. 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.

2.2.2. 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 [12]. 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.

2.2.3. 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 [26]. 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 [27]. 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.

2.2.4. 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 [28]. The present study extends this visualization paradigm to environmental isolates.

2.3. Synthesis: The Methodological Gap

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

2.3.1. Comparative Summary of Related Studies

A systematic comparison of related studies reveals distinct methodological approaches to AMR pattern recognition. Ardila et al. [24] conducted a comprehensive systematic review of machine learning applications in AMR, finding that supervised methods (Random Forest, Gradient Boosting) achieve high predictive accuracy but require labeled training data.

Parthasarathi et al. [6] demonstrated effective unsupervised clustering of AMR genes with silhouette scores reaching 0.82, establishing that clustering methods can discover meaningful resistance patterns. Kou et al. [18] applied spatial epidemiology to *E. coli* resistance patterns across Chinese provinces, revealing significant spatial autocorrelation but without phenotypic clustering. Abada et al. [12] employed Ward's hierarchical clustering for agricultural MDR bacteria, validating the method's applicability to resistance phenotyping. The INOHAC project [7] provided foundational multi-regional surveillance data but relied on conventional MDR classification without computational pattern discovery.

What distinguishes the present study is the systematic integration of unsupervised and supervised approaches within an environmental One Health context. While individual studies have employed either clustering or classification, none have combined Ward's hierarchical clustering for phenotype discovery with Random Forest validation specifically for multi-source environmental isolates spanning the water-fish-human nexus. This hybrid methodology addresses both the discovery challenge (identifying resistance archetypes without predefined labels) and the validation challenge (confirming that discovered patterns represent biologically coherent structures). The following table summarizes these methodological distinctions.

Table 1: Comparative Summary of Computational Approaches to AMR Analysis. The current study integrates unsupervised and supervised approaches for environmental AMR surveillance.

Author	Title	Year	Unsupervised	Supervised	Focus	Contribution
Ardila et al.	Systematic Review of ML in AMR	2025	No	Yes	Systematic review	RF/GBDT
Parthasarathi et al.	Clustering-Based AMR Gene Analysis	2024	Yes	Yes	AMR gene clustering	Silhouette 0.82
Kou et al.	Spatial Epidemiology of E. coli	2025	No	No	Spatial epidemiology	Spatial autocorrelation
Abada et al.	Ward's Clustering for Agricultural MDR	2025	Yes	No	Agricultural MDR	Ward's clustering
Abamo et al.	INOHAC AMR Project Two	2024	No	No	Environmental surveillance	Multi-regional dataset
Current Study	Pattern Recognition of AMR	2026	Yes	Yes	Water-fish-human nexus	Hierarchical + RF

The current study integrates unsupervised and supervised approaches for environmental AMR surveillance.

Limitations of Existing Approaches. Supervised methods achieve high accuracy but cannot identify novel resistance patterns absent from training data. Unsupervised clustering, while effective in agricultural and clinical settings, has rarely been applied to multi-regional One Health surveillance. Spatial epidemiology operates on aggregated metrics rather than phenotypic profiles. Philippine surveillance has relied on conventional MDR classification, leaving the INOHAC dataset's pattern discovery potential unrealized.

The Present Study's Contribution. This study addresses these gaps through a hybrid unsupervised-supervised framework for environmental AMR surveillance. Ward's hierarchical clustering discovers resistance archetypes without predefined labels, while Random Forest classification validates whether clusters represent biologically coherent structures. Applying this methodology to isolates spanning multiple Philippine regions and ecological compartments (water, fish, human) enables characterization of resistance phenotypes specific to the One Health nexus. This integrated approach advances beyond purely supervised prediction or unsupervised clustering alone, offering a reproducible framework for future surveillance studies.

CHAPTER 3

THEORETICAL FRAMEWORK

3.1. 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.

3.2. Primary Theoretical Foundations

3.2.1. Pattern Recognition Theory

The primary theoretical foundation of this study is Pattern Recognition Theory, as formalized by Duda, Hart, and Stork [29] 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:

Table 2: Learning Paradigms in Pattern Recognition. Unsupervised learning discovers structure; supervised learning validates discrimination.

Paradigm	Theoretical Basis	Application in Study
Unsupervised Learning	Cluster Analysis Theory [30]	Hierarchical Agglomerative Clustering discovers latent resistance structures without predefined labels
Supervised Learning	Statistical Learning Theory [31]	Logistic Regression, Random Forest, and k-Nearest Neighbors validate discriminative capacity of discovered patterns

The theoretical justification for combining both paradigms derives from the cluster validation problem articulated by Jain and Dubes [30]: 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.

3.2.1.1. Hierarchical Clustering Theory

Ward's minimum variance method, employed in this study, is grounded in the theoretical principle of within-cluster homogeneity maximization [12]. 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.

3.2.2. 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" [32]."

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.

3.3. Supporting Theoretical Concepts

3.3.1. 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 [33]. Leakage violates the fundamental assumption of independent and identically distributed (i.i.d.) training and evaluation sets.

The study operationalizes leakage prevention through architectural constraints derived from this theory. While three primary leakage types exist, two are directly addressed in this study:

Table 3: Leakage Types and Architectural Mitigations. Feature-metadata separation and split-before-transform protocols are implemented to prevent information leakage.

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

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

3.3.2. 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 [34]. 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 [35]. 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.

3.3.3. 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.

3.4. 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.

3.4.1. Independent Variables (Research/Data)

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

Table 4: Independent Variables. AST data serves as the primary input, while derived metrics and contextual metadata support analysis and interpretation.

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)

3.4.2. Dependent Variables (Design Features)

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

Table 5: Dependent Variables. Derived features used for pattern recognition (Resistance Profile, MDR status, Resistance Index). Each architectural component is directly derived from a theoretical requirement to address specific surveillance challenges.

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

3.4.3. The Derivation Chain

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

Table 6: Derivation Chain from Theory to Design. This chain ensures methodological coherence by linking abstract principles to concrete implementation modules.

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

3.5. Theoretical Justification

3.5.1. 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.

3.5.2. 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).

3.5.3. 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.

3.6. Conceptual Framework

The conceptual framework synthesizes the theoretical foundations and supporting concepts into an integrated model that guides the study's analytical design and implementation. This framework establishes the logical flow from abstract theoretical principles to concrete architectural decisions, ensuring methodological coherence throughout the research process.

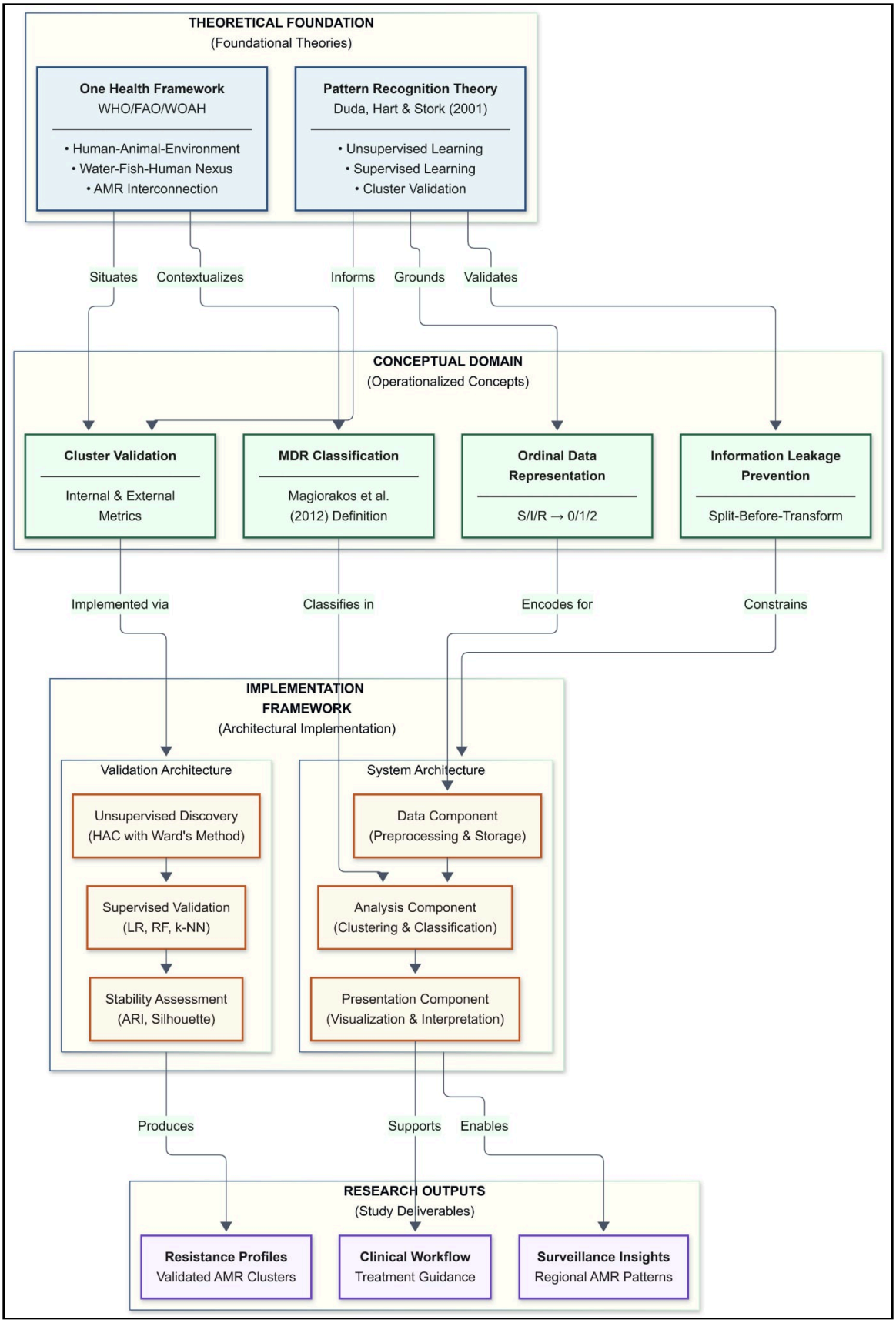

Figure 1: Conceptual Framework Diagram: Integration of Theoretical Foundation, Conceptual Domain, Implementation Framework, and Research Outputs

Figure 1 illustrates four interconnected components that structure this study's analytical approach:

1. Theoretical Foundation: Pattern Recognition Theory provides the computational paradigm for discovering latent resistance structures, while the One Health framework situates AMR within the Water-Fish-Human nexus.
2. Conceptual Domain: Abstract principles are operationalized into methodological constraints—cluster validation, information leakage prevention through split-before-transform protocols, ordinal S/I/R encoding (0/1/2), and standardized MDR classification following Magiorakos et al.
3. Implementation Framework: A three-tier architecture (Data, Analysis, and Presentation Components) implements the validation pipeline: unsupervised discovery via Hierarchical Agglomerative Clustering with Ward's method, supervised validation using Logistic Regression, Random Forest, and k-NN classifiers, and stability assessment through Adjusted Rand Index and Silhouette scores.
4. Research Outputs: Validated resistance profiles, regional surveillance insights, and clinical workflow support.

The directional flow progresses from theoretical justification through operationalization to implementation, with outputs providing empirical validation that informs the theoretical understanding of AMR patterns.

3.7. 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 Supervised Validation Theory 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 4

METHODOLOGY

4.1. 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.

4.2. Data Source and Description

4.2.1. 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 does not involve primary sampling or laboratory experimentation. All analyses are 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: Eastern Visayas, Central Luzon, and the Bangsamoro Autonomous Region in Muslim Mindanao (BARMM).

4.2.2. Sample Source Categories

Isolates originate 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 are used exclusively as contextual metadata during interpretation. Source categories are listed in Table 7.

Table 7: Sample Source Categories. Isolates originate from environmental matrices representing the Water–Fish interface within the broader Water–Fish–Human nexus.

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

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.

4.2.3. Isolate Identification Convention

Each isolate is 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][Colony]`

This convention enables systematic metadata parsing while preserving traceability throughout the analytical pipeline. The replicate and colony identifiers include ‘R’ and ‘C’ prefixes respectively (e.g., R1C1). Sample isolate records from the unified raw dataset are presented in Table 8.

Table 8: Sample Isolate Records from the Unified Raw Dataset. (n = 583). AM: Ampicillin, AMC: Amoxicillin/Clavulanic Acid, CF: Cefalexin, IPM: Imipenem, GM: Gentamicin, AN: Amikacin, TE: Tetracycline, DO: Doxycycline, C: Chloramphenicol, SXT: Trimethoprim/Sulfamethoxazole. S: Susceptible, I: Intermediate, R: Resistant, -: Not tested. Source codes: DW: Drinking Water, LW: Lake Water, RW: River Water, EWU/EWT: Effluent Water, FT: Fish Tilapia, FG: Fish Gusaw.

Isolate No.	Species	Region	Source	AM	AMC	CF	IPM	GM	AN	TE	DO	C	SXT	ESBL	MAR
1	*E. coli*	BARMM	EWU	R	I	S	S	S	S	S	S	S	R	NEG	0.09
2	*K. pneumoniae*	BARMM	EWU	R	S	S	S	S	S	S	S	S	S	NEG	0.05
3	*E. coli*	BARMM	EWT	S	S	S	S	S	S	R	S	S	S	NEG	0.14
4	*E. coli*	BARMM	DW	R	S	S	S	S	S	R	R	S	R	NEG	0.18
5	*K. pneumoniae*	BARMM	DW	R	S	R	S	R	S	R	R	S	R	POS	0.38
6	*K. pneumoniae*	BARMM	LW	R	R	R	S	R	S	R	R	S	R	NEG	0.50
7	*V. cholerae*	BARMM	LW	R	S	-	S	-	S	S	S	-	S	-	0.07
8	*E. cloacae*	BARMM	DW	-	R	R	S	S	S	S	S	S	S	-	0.15
9	*E. aerogenes*	BARMM	DW	-	R	R	S	S	S	S	S	S	S	-	0.14
10	*E. coli*	Region III	RW	R	S	S	S	S	S	S	S	S	S	NEG	0.09
11	*K. pneumoniae*	Region III	RW	R	S	S	S	S	S	R	R	R	R	NEG	0.23
12	*E. cloacae*	Region III	RW	-	R	R	S	S	S	S	S	S	S	-	0.15
13	*E. coli*	Region III	FT	S	S	S	S	S	S	S	S	S	S	NEG	0.00
14	*E. aerogenes*	Region III	FT	-	R	R	S	S	S	S	S	S	S	-	0.14
15	*K. pneumoniae*	Region III	FG	R	S	S	S	S	S	R	R	R	R	NEG	0.23
16	*E. cloacae*	Region III	FG	-	R	R	S	S	S	S	S	S	S	-	0.15
17	*E. coli*	Region VIII	DW	S	S	S	S	S	S	S	S	S	S	NEG	0.00
18	*E. coli*	Region VIII	FG	S	S	S	S	S	S	R	R	S	R	NEG	0.18
19	*K. pneumoniae*	Region VIII	FG	R	S	S	S	S	S	S	S	S	S	NEG	0.05
20	*E. coli*	Region VIII	FG	R	S	S	S	S	S	R	R	R	R	NEG	0.18

4.2.4. Antimicrobial Panel

Phenotypic AST data consists of a panel of 22 antibiotics spanning 12 antimicrobial classes, including an ESBL screening indicator. The antimicrobial panel is summarized in Table 9.

Table 9: Antimicrobial Panel Composition. (12 Classes, 22 Antibiotics). The panel spans major antimicrobial classes including Penicillins, Cephalosporins, Carbapenems, and Aminoglycosides.

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

4.3. 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 are explicitly parameterized to ensure reproducibility and to prevent information leakage in downstream analyses.

4.3.1. Data Ingestion and Harmonization

Raw phenotypic AST data are 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, are integrated into a single unified dataset.

The ingestion process includes the following steps:

- Schema harmonization: Column names, data types, and value encodings are standardized across source files to ensure structural consistency.
- Metadata extraction: Structured isolate identifiers are parsed to extract contextual variables such as geographic region, local site, source category, replicate number, and colony number.
- Duplicate resolution: Duplicate isolate records are identified and removed to ensure a one-to-one correspondence between isolates and resistance profiles.

This step ensures that all downstream analyses operate on a coherent and internally consistent dataset.

4.3.2. 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 are established beforehand to avoid after-the-fact adjustments based on results.

Application of these filtering criteria resulted in the following data reduction:

Table 10: Data Filtering Summary. (84.2% Retention Rate). Quality control filters removed duplicates and low-quality isolates, retaining 491 analysis-ready samples.

Filtering Step	Count
Initial unified dataset	583 isolates
Duplicates removed	−2 isolates
Isolates removed (>30% missing data)	−90 isolates
Final analysis-ready dataset	491 isolates

Additionally, 9 antibiotics were excluded for failing to meet the 70% coverage threshold: AMI, CFA, CFV, CPT, CTF, GEN, IME, MAR, and `ORIGINAL_SPECIES`. The final antimicrobial panel comprises 21 antibiotics with adequate test coverage.

4.3.3. Resistance Encoding

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

Table 11: Ordinal Encoding of Phenotypic AST Results. (0 = Susceptible, 1 = Intermediate, 2 = Resistant). Susceptible (0), Intermediate (1), and Resistant (2) values preserve the progressive nature of resistance severity.

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

This ordinal encoding preserves the progressive nature of resistance severity while enabling distance-based computations. Sample encoded resistance values are presented in Table 12.

Table 12: Sample Encoded Resistance Values. (0 = Susceptible, 1 = Intermediate, 2 = Resistant). AM: Ampicillin, AMC: Amoxicillin/Clavulanic Acid, CF: Cefalotin, CN: Cefalexin, IPM: Imipenem, GM: Gentamicin, AN: Amikacin, TE: Tetracycline, DO: Doxycycline, C: Chloramphenicol, SXT: Trimethoprim/Sulfamethoxazole.

Isolate No.	Species	Region	Source	AM	AMC	CF	CN	IPM	GM	AN	TE	DO	C	SXT	MAR
1	*E. coli*	BARMM	EWU	2	1	0	0	0	0	0	0	0	0	2	0.09
2	*E. coli*	BARMM	EWU	0	0	0	0	0	0	0	0	0	0	0	0.00
3	*K. pneumoniae*	BARMM	EWU	2	0	0	0	0	0	0	0	0	0	0	0.05
4	*E. coli*	BARMM	EWU	0	0	0	0	0	0	0	2	2	0	0	0.09
5	*E. coli*	BARMM	EWU	2	1	1	1	0	0	0	0	0	0	2	0.09
6	*K. pneumoniae*	BARMM	DW	2	0	2	2	0	2	0	2	2	0	2	0.38
7	*K. pneumoniae*	BARMM	LW	2	2	2	2	0	2	0	2	2	0	2	0.50
8	*E. coli*	Region III	RW	2	0	0	0	0	0	0	0	0	0	0	0.05
9	*K. pneumoniae*	Region III	RW	2	0	0	0	0	0	0	2	2	2	2	0.23
10	*E. coli*	Region III	FT	0	0	0	0	0	0	0	0	0	0	0	0.00
11	*E. coli*	Region III	FT	2	0	0	0	0	2	0	0	0	0	0	0.09
12	*K. pneumoniae*	Region III	FG	2	0	0	0	0	0	0	2	2	2	2	0.23
13	*E. coli*	Region VIII	DW	0	0	0	0	0	0	0	0	0	0	0	0.00
14	*E. coli*	Region VIII	FG	0	0	0	0	0	0	0	2	2	0	2	0.18
15	*E. coli*	Region VIII	FG	2	0	0	0	0	0	0	2	2	2	2	0.18
16	*K. pneumoniae*	Region VIII	FG	2	0	0	0	0	0	0	0	0	0	0	0.05
17	*E. coli*	Region VIII	FG	0	0	0	0	0	0	0	0	0	0	0	0.00
18	*E. coli*	BARMM	EWT	0	0	0	0	0	0	0	2	0	0	0	0.14
19	*K. pneumoniae*	Region III	FG	2	0	0	0	0	0	0	0	0	0	0	0.05
20	*E. coli*	Region III	RW	2	0	0	0	0	2	0	0	0	0	0	0.09

4.3.4. Missing Value Imputation

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

$$\hat{x}_{i,j} = \text{median}\big(\{x_{k,j} \mid x_{k,j} \text{ is observed}\}\big) \qquad \text{Eq. (2)}$$

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 are considered; however, the median provides a conservative central estimate suitable for exploratory pattern recognition.

4.3.5. Derived Resistance Feature Computation

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

4.3.5.1. Multiple Antibiotic Resistance (MAR) Index

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

$$\text{MAR} = \frac{a}{b} \qquad \text{Eq. (3)}$$

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

4.3.5.2. Resistant Classes Count

The breadth of resistance across antimicrobial classes was computed as:

$$\text{Resistant Classes} = |\{c \mid \exists a \in c, \text{resistance}(a) = \text{true}\}| \qquad \text{Eq. (4)}$$

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.

4.3.5.3. Multidrug Resistance (MDR) Classification

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

$$\text{MDR} = \begin{cases} 1, \text{if Resistant Classes} \geq 3 \\ 0, \text{otherwise} \end{cases} \qquad \text{Eq. (5)}$$

4.3.6. Feature–Metadata Separation

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

- Feature Matrix ($\boldsymbol{X}$): Encoded resistance values for the 22 antibiotics, used exclusively for unsupervised clustering and supervised validation.
- Metadata Matrix ($\boldsymbol{M}$): 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.

4.3.7. Preprocessing Component Output

The output of the preprocessing component 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

Following preprocessing, the analysis-ready dataset comprises 491 isolates with standardized species names and complete resistance profiles. Sample records from the preprocessed dataset are presented in Table 13.

Table 13: Sample Isolate Records from the Analysis-Ready Dataset. (n = 491). Resistance values encoded as: 0 = Susceptible, 1 = Intermediate, 2 = Resistant. AM: Ampicillin, AMC: Amoxicillin/Clavulanic Acid, CF: Cefalotin, CN: Cefalexin, IPM: Imipenem, GM: Gentamicin, AN: Amikacin, TE: Tetracycline, DO: Doxycycline, C: Chloramphenicol, SXT: Trimethoprim/Sulfamethoxazole. MDR: Multidrug-resistant (≥3 classes). Source codes: DW: Drinking Water, LW: Lake Water, RW: River Water, EWU/EWT: Effluent Water, FT: Fish Tilapia, FG: Fish Gusaw.

Isolate No.	Species	Region	Source	AM	AMC	CF	CN	IPM	GM	AN	TE	DO	C	SXT	MDR	MAR
1	*E. coli*	BARMM	EWU	2	1	0	0	0	0	0	0	0	0	2	No	0.09
2	*K. pneumoniae*	BARMM	EWU	2	0	0	0	0	0	0	0	0	0	0	No	0.05
3	*E. coli*	BARMM	EWT	2	0	0	0	0	0	0	2	2	0	2	Yes	0.14
4	*E. coli*	BARMM	DW	2	0	0	0	0	0	0	2	2	0	2	Yes	0.18
5	*K. pneumoniae*	BARMM	DW	2	0	2	2	0	2	0	2	2	0	2	Yes	0.38
6	*K. pneumoniae*	BARMM	LW	2	2	2	2	0	2	0	2	2	0	2	Yes	0.50
7	*E. coli*	BARMM	LW	0	0	0	0	0	0	0	0	0	0	0	No	0.00
8	*E. coli*	Region III	RW	2	0	0	0	0	0	0	0	0	0	0	No	0.05
9	*K. pneumoniae*	Region III	RW	2	0	0	0	0	0	0	2	2	2	2	Yes	0.23
10	*E. coli*	Region III	RW	2	0	0	0	0	2	0	0	0	0	0	No	0.09
11	*E. coli*	Region III	FT	0	0	0	0	0	0	0	0	0	0	0	No	0.00
12	*E. coli*	Region III	FT	2	0	0	0	0	2	0	0	0	0	0	No	0.09
13	*K. pneumoniae*	Region III	FG	2	0	0	0	0	0	0	2	2	2	2	Yes	0.23
14	*K. pneumoniae*	Region III	FG	2	0	0	0	0	0	0	0	0	0	0	No	0.05
15	*E. coli*	Region VIII	DW	0	0	0	0	0	0	0	0	0	0	0	No	0.00
16	*E. coli*	Region VIII	DW	0	0	0	0	0	0	0	0	0	0	0	No	0.00
17	*E. coli*	Region VIII	FG	0	0	0	0	0	0	0	2	2	0	2	Yes	0.18
18	*E. coli*	Region VIII	FG	2	0	0	0	0	0	0	2	2	2	2	Yes	0.18
19	*K. pneumoniae*	Region VIII	FG	2	0	0	0	0	0	0	0	0	0	0	No	0.05
20	*E. coli*	Region VIII	FG	0	0	0	0	0	0	0	0	0	0	0	No	0.00

4.4. 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 section operate exclusively on the resistance feature matrix produced during preprocessing.

4.4.1. 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.

4.4.2. Distance Metric

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

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad \text{Eq. (6)}$$

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

This metric is 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.

4.4.3. Linkage Method

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

$$\Delta(A, B) = \frac{n_A n_B}{n_A + n_B} \|c_A - c_B\|^2 \quad \text{Eq. (7)}$$

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.

4.4.4. Determination of the Number of Clusters

The optimal number of clusters is determined using a data-driven, multi-criteria approach combining quantitative metrics with practical constraints, following established conventions for exploratory cluster analysis [36], [37].

4.4.4.1. Silhouette Analysis

Cluster cohesion and separation were evaluated using the silhouette score [15]:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \qquad \text{Eq. (8)}$$

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 [16]. The average silhouette score across all isolates is computed for cluster solutions ranging from $k = 2$ to $k = 8$, a range consistent with recommendations for systematic cluster validation [36].

4.4.4.2. Within-Cluster Sum of Squares (WCSS)

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

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2 \qquad \text{Eq. (9)}$$

where C_k denotes cluster k and μ_k its centroid. The elbow method is used to identify diminishing returns in compactness as the number of clusters increased [37].

4.4.4.3. 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 [38], [39].
- Granularity control: Excessive partitioning was avoided to preserve phenotypically coherent resistance groupings amenable to downstream interpretation.

Final cluster selection employs 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 [16].

4.4.5. 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.

4.4.6. Unsupervised Discovery Output

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

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

4.5. Supervised Learning Validation

Supervised learning models are 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.

4.5.1. Classification Task

Supervised classification is designed to validate the unsupervised clustering results by assessing whether the discovered clusters represent discriminable resistance phenotypes:

Table 14: Supervised Classification Task. The objective is to validate that clusters represent discriminable phenotypes using cluster assignment as the target variable.

Task	Target Variable	Purpose
Cluster Discrimination	Cluster assignment	Validate that clusters represent discriminable phenotypes

4.5.2. Leakage-Safe Data Splitting

To prevent information leakage between training and evaluation phases, the dataset is first partitioned into training (80%) and test (20%) subsets using stratified sampling to preserve

class distributions. Train–test splitting is performed prior to any preprocessing operations, including missing value imputation and feature scaling.

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

4.5.3. Model Selection

Three classifier families are selected to represent different learning paradigms:

Table 15: Supervised Model Selection. Three distinct model families (Linear, Tree-based, Distance-based) were chosen to evaluate cluster robustness across 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

Hyperparameter Configuration:

Table 16: Model Hyperparameters. Configuration settings for the three classifiers used in the validation phase.

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

4.5.4. Evaluation Metrics

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

4.5.4.1. Macro-Averaged Precision, Recall, F1

$$\text{Precision}_{\text{macro}} = \frac{1}{|C|} \sum_{c \in C} \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c} \quad \text{Eq. (10)}$$

$$\text{Recall}_{\text{macro}} = \frac{1}{|C|} \sum_{c \in C} \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \quad \text{Eq. (11)}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{Eq. (12)}$$

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

4.5.4.2. Accuracy

Overall classification correctness is measured as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad \text{Eq. (13)}$$

4.5.4.3. Confusion Matrix

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

4.5.5. Feature Importance Extraction

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

$$\text{Importance}(f) = \sum_{t \in T} \Delta G_t \cdot \mathbb{1}[f_t = f] \quad \text{Eq. (14)}$$

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.

4.5.6. Stability Across Random Seeds

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

Input: Dataset D, Prediction Model M, Random Seeds $S = \{42, 123, 456, 789, 1011\}$
Output: Stability metrics $(\mu_{\text{metrics}}, \sigma_{\text{metrics}})$
$R = \emptyset$ (Initialize results container)
For each seed $s \in S$ **do**:
  Set global random state to s
  Split D into D_{train} (80%) and D_{test} (20%) using stratified sampling
  Train M on D_{train}
  Evaluate M on D_{test} to obtain metric vector v_s
  Append v_s to R
Compute mean $\mu = \frac{1}{|S|} \sum_{v \in R} v$
Compute standard deviation $\sigma = \sqrt{\frac{1}{|S|-1} \sum_{v \in R} (v - \mu)^2}$
Return μ, σ

Figure 2: Cross-Seed Stability Check Algorithm. The process iterates through five distinct random seeds to generate a distribution of performance metrics (v_s), from which the mean (μ) and standard deviation (σ) are calculated to quantify model stability.

Low standard deviation across seeds indicates robust model performance.

4.5.7. Sensitivity Analysis: Split Ratio and Cross-Validation

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

validation schemes (5-fold, 10-fold) are evaluated across all three classifier models to determine the optimal balance between training adequacy and evaluation reliability.

4.5.7.1. Sensitivity Analysis Interpretation

The sensitivity analysis provides the following rationale for the chosen experimental configuration:

1. Stability Assessment: Standard deviations are analyzed across random seeds to ensure that the discriminative capacity is not an artifact of random initialization.
2. 80/20 Split Rationale: The 80/20 split is selected as it provides a statistically reliable test set size (≈98 samples) while maintaining sufficient training data, balancing model learning capacity with robust evaluation.
3. Cross-Validation Selection: 5-fold and 10-fold cross-validation produce comparable stability. Given the computational efficiency of 5-fold CV, it is preferred for the full experimental pipeline.
4. Model Selection: Random Forest is selected as the primary validation model due to its consistently stable performance and its ability to provide interpretable feature importance through Gini impurity.

These findings support the use of the 80/20 train–test split with Random Forest and 5-fold cross-validation as the robust standard configuration for supervised validation.

4.5.8. Supervised Validation Output

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
- Serialized model artifacts for deployment (`.joblib`)
- Structured feature importance data for dashboard integration (`.json`)

4.6. Statistical Association Analysis

To characterize the relationships between resistance patterns and external variables, rigorous statistical association methods are employed.

4.6.1. Co-Resistance Analysis

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

$$\varphi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \quad \text{Eq. (15)}$$

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

Table 17: Phi Coefficient Contingency Table Structure. The 2×2 table defines the co-occurrence counts used to calculate the pairwise association between resistance traits.

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

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

4.6.2. Metadata Association Analysis

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

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}} \qquad \text{Eq. (16)}$$

where χ^2 is the chi-square statistic, n is the sample size, and r and c are the dimensions of the contingency table. Interpretation thresholds follow established guidelines [40].

Table 18: Cramér's V Interpretation Guidelines. Interpretation thresholds for effect size based on Cohen's standards for contingency table analysis.

Cramér's V Value	Association Strength
< 0.10	Negligible
0.10 – 0.30	Small
0.30 – 0.50	Moderate
> 0.50	Strong

4.6.3. Interpretation Protocol

Interpretation follows a strict staged interpretation strategy to maintain analytical integrity:

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

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

4.7. 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.

4.8. 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.

4.9. 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.

The methodology establishes a rigorous analytical pipeline that transforms raw AST data into validated resistance patterns through unsupervised discovery, supervised validation, and statistical association analysis. This approach ensures that resistance structures emerge from objective, data-driven processes while maintaining strict separation between pattern discovery and biological interpretation throughout all analytical stages.

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 5

ARCHITECTURAL DESIGN

5.1. Introduction

This chapter presents the architectural design of the Antimicrobial Resistance (AMR) Pattern Recognition System. The system follows a layered pipeline architecture that transforms raw Antimicrobial Susceptibility Testing (AST) data into actionable insights through a series of well-defined processing stages. The architecture emphasizes modularity, reproducibility, and scientific rigor—ensuring that each component can be independently validated and that the analytical pipeline produces defensible results for clinical and epidemiological applications.

The system architecture comprises four primary stages: (1) Raw Data Input, (2) Data Preprocessing, (3) Pattern Discovery, and (4) Output Visual Representation. Each stage is designed with clear inputs, outputs, and transformation logic, enabling traceability from raw laboratory data to final analytical conclusions.

5.2. Overall System Architecture

The AMR Pattern Recognition System implements a pipeline architecture where data flows sequentially through preprocessing stages before branching into three parallel pattern

discovery methods. The results from each analytical approach are then consolidated into unified visual representations for interpretation.

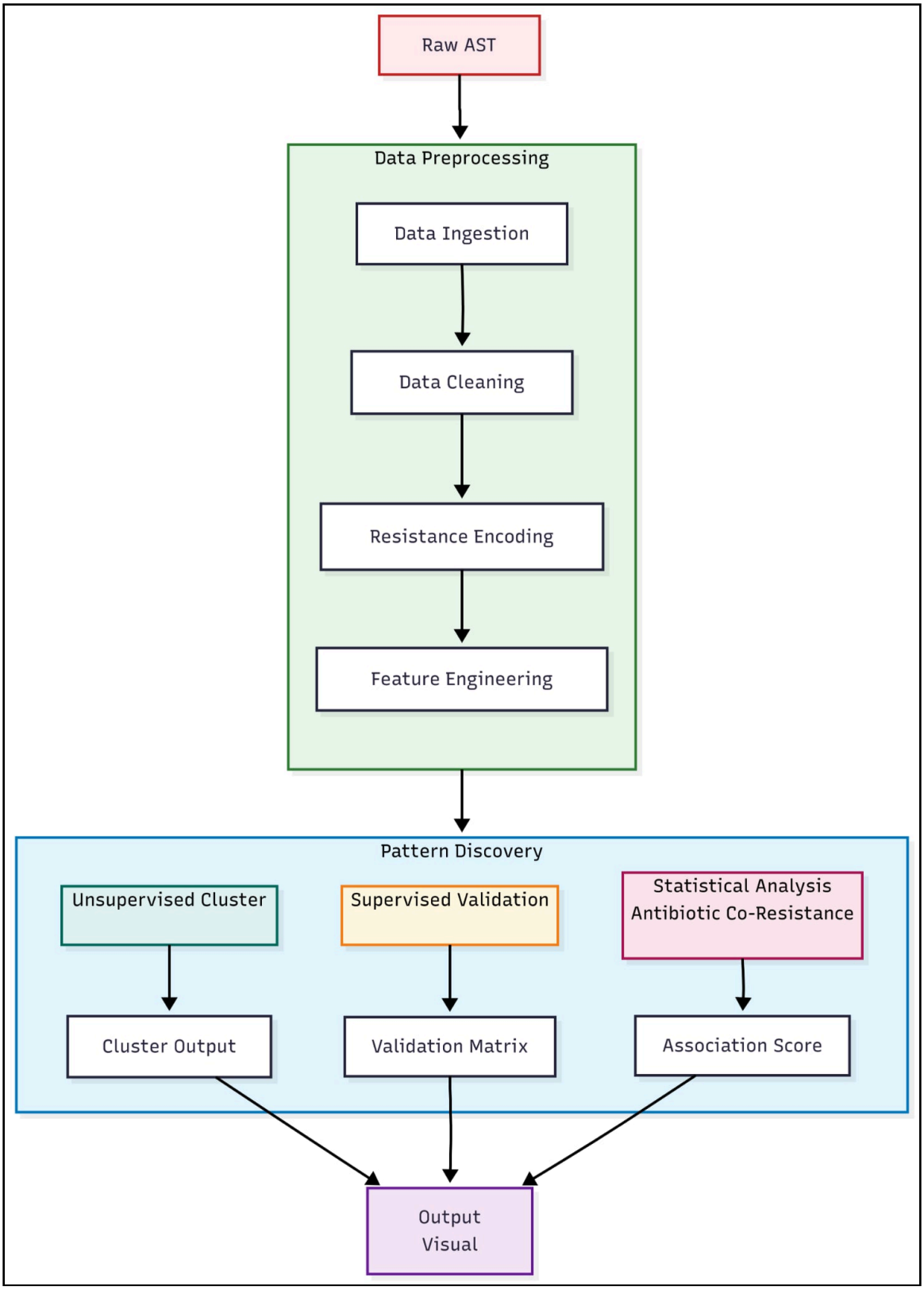

Figure 3: Overall System Architecture

5.2.1. Architecture Components Overview

The overall architecture consists of the following major components:

Table 19: Architecture Components Overview

Component	Purpose	Input	Output
Raw AST Data	Source laboratory data from multiple regional surveillance sites	CSV files containing isolate records with S/I/R interpretations	Unprocessed antimicrobial susceptibility records
Data Preprocessing	Transform raw data into analysis-ready format through ingestion, cleaning, encoding, and feature engineering	Raw CSV files	Cleaned, encoded feature matrix with derived indicators
Pattern Discovery	Apply three complementary analytical methods to identify resistance patterns	Analysis-ready dataset	Cluster assignments, validation metrics, and association scores
Output Visual Representation	Consolidate and visualize findings through interactive dashboards	Results from all pattern discovery methods	Charts, heatmaps, and statistical summaries

The architecture employs a fan-out pattern at the Pattern Discovery stage, where the preprocessed data is simultaneously processed by three independent analytical methods. This design ensures that findings can be cross-validated across different methodological approaches, strengthening the scientific validity of conclusions.

5.3. Data Preprocessing Stage

The Data Preprocessing stage transforms heterogeneous raw AST data into a standardized, analysis-ready dataset. This stage is critical for ensuring data quality, reproducibility, and downstream analytical validity. The preprocessing pipeline consists of four sequential sub-stages: Data Ingestion, Data Cleaning, Resistance Encoding, and Feature Engineering.

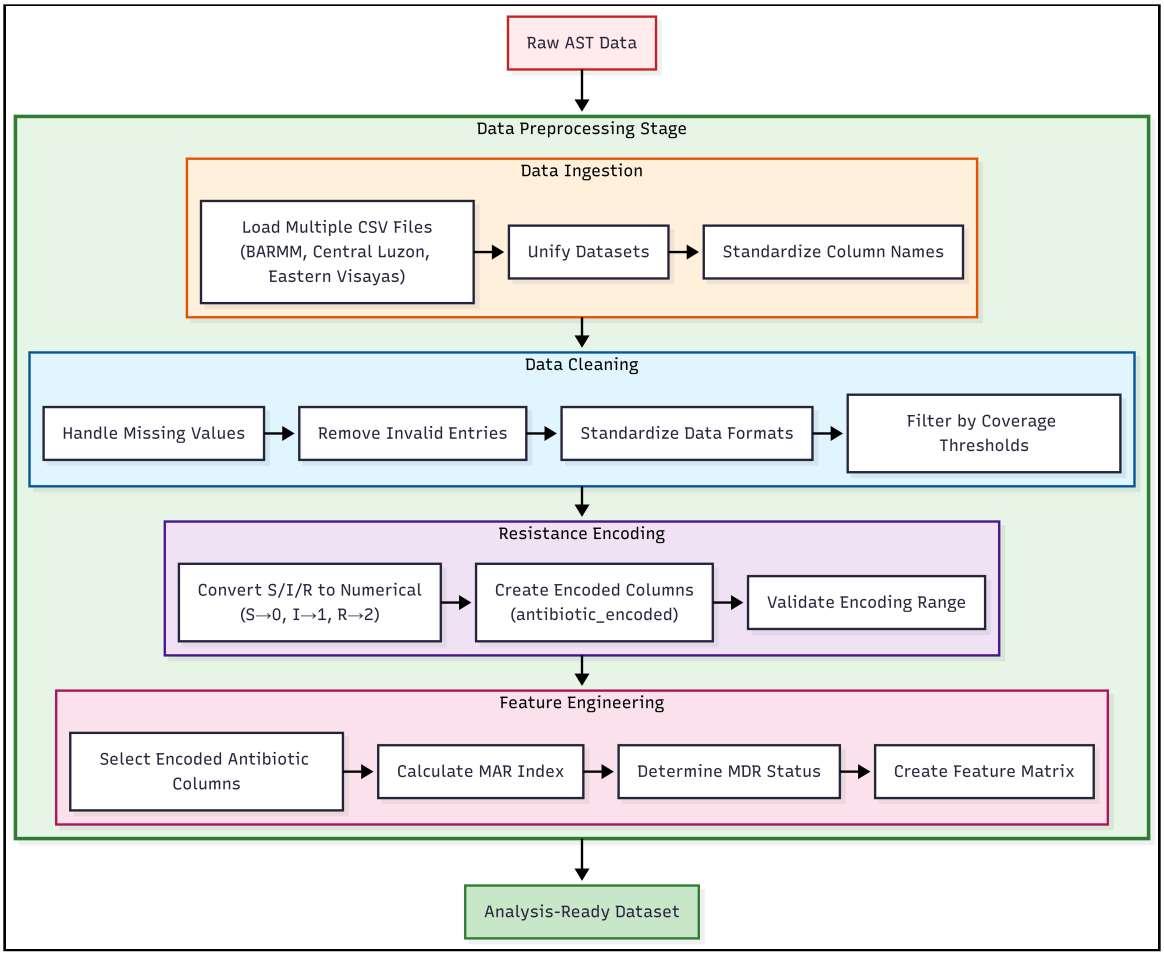

Figure 4: Data Preprocessing Stage Architecture

5.3.1. Data Ingestion

The Data Ingestion sub-stage consolidates AST records from multiple regional surveillance sites into a unified dataset. The system processes CSV files from three Philippine regions: BARMM (Bangsamoro Autonomous Region in Muslim Mindanao), Region III (Central Luzon), and Region VIII (Eastern Visayas).

Key Operations:

- Load Multiple CSV Files: Iteratively reads all CSV files from the raw data directory using glob pattern matching

- Unify Datasets: Concatenates individual dataframes into a master dataset while preserving source metadata (region, facility, collection date)
- Standardize Column Names: Normalizes column naming conventions and applies species standardization mappings to ensure taxonomic consistency

5.3.2. Data Cleaning

The Data Cleaning sub-stage ensures data quality by addressing missing values, invalid entries, and format inconsistencies that could compromise analytical validity.

Key Operations:

- Handle Missing Values: Identifies and documents missing AST results; applies coverage thresholds to determine acceptable missingness levels
- Remove Invalid Entries: Excludes records with ambiguous species identification or incomplete metadata required for stratified analysis
- Standardize Data Formats: Normalizes date formats, categorical values, and text fields to ensure consistency
- Filter by Coverage Thresholds: Retains only antibiotics and isolates meeting minimum testing coverage requirements (≥70% antibiotic coverage, ≤30% missing values per isolate)

5.3.3. Resistance Encoding

The Resistance Encoding sub-stage transforms categorical AST interpretations into numerical values suitable for computational analysis.

Key Operations:

- Convert S/I/R to Numerical: Applies ordinal encoding where Susceptible (S) = 0, Intermediate (I) = 1, and Resistant (R) = 2
- Create Encoded Columns: Generates new columns with `_encoded` suffix containing numerical values while preserving original categorical data
- Validate Encoding Range: Verifies all encoded values fall within expected range [0, 1, 2] and flags anomalies

5.3.4. Feature Engineering

The Feature Engineering sub-stage derives clinically meaningful indicators from the encoded resistance profiles.

Key Operations:

- Select Encoded Antibiotic Columns: Identifies all `_encoded` columns to form the resistance fingerprint vector
- Calculate MAR Index: Computes the Multiple Antibiotic Resistance Index using the formula $\mathrm{MAR} = \frac{a}{b}$, where a = number of antibiotics to which the isolate is resistant, and b = total antibiotics tested [21]
- Determine MDR Status: Classifies isolates as Multi-Drug Resistant (MDR) if resistant to at least one agent in ≥3 antimicrobial categories [22]
- Create Feature Matrix: Assembles the final feature matrix (X) containing encoded resistance values for all tested antibiotics

5.4. Pattern Discovery Stage

The Pattern Discovery stage applies three complementary analytical methods to identify, validate, and characterize antimicrobial resistance patterns. Each method addresses a distinct analytical objective while providing cross-validation opportunities. The stage receives the analysis-ready dataset from preprocessing and produces cluster assignments, validation metrics, and association scores.

5.4.1. Unsupervised Clustering

The Unsupervised Clustering component identifies natural groupings in resistance patterns without predefined categories. This data-driven approach discovers resistance phenotypes —characteristic patterns of antibiotic susceptibility that may correspond to underlying biological or epidemiological phenomena.

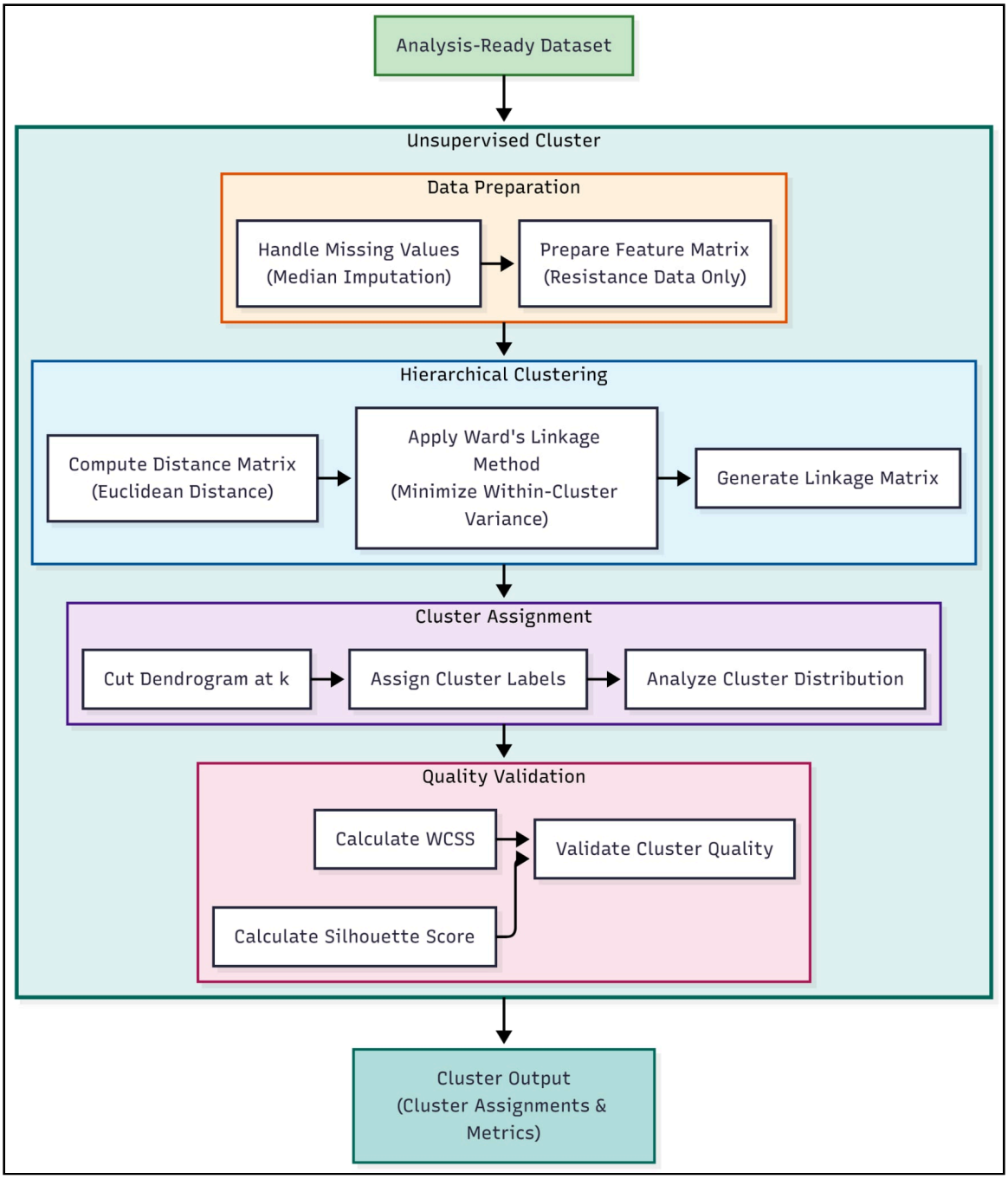

Figure 5: Unsupervised Clustering Architecture

5.4.1.1. Optimal k Selection

Before clustering, the optimal number of clusters (k) must be determined through systematic evaluation.

Key Operations:

- Elbow Method Analysis: Plots Within-Cluster Sum of Squares (WCSS) against cluster count; identifies the "elbow point" where additional clusters yield diminishing returns in variance reduction
- Silhouette Score Analysis: Computes silhouette coefficients for different k values; higher scores indicate better-defined cluster boundaries
- Determine Best k Value: Synthesizes elbow and silhouette analyses with domain knowledge to select the optimal cluster count

5.4.1.2. Hierarchical Clustering

The system employs Hierarchical Agglomerative Clustering (HAC) to group isolates based on resistance profile similarity.

Key Operations:

- Apply Ward's Linkage Method: Uses Ward's minimum variance criterion to minimize within-cluster variance at each merge step, producing compact and well-separated clusters [11]
- Use Euclidean Distance: Computes pairwise distances between isolates using Euclidean metric, appropriate for numerical resistance vectors and required by Ward's linkage
- Generate Cluster Assignments: Cuts the dendrogram at the optimal level to assign each isolate to a specific cluster

5.4.1.3. Quality Metrics

Cluster quality is assessed through internal validation metrics that quantify cluster coherence and separation.

Key Operations:

- Calculate Silhouette Score: Measures how similar isolates are to their own cluster compared to other clusters; values range from −1 to +1, with higher values indicating better clustering
- Calculate WCSS: Computes total within-cluster sum of squares as a measure of cluster compactness
- Validate Cluster Quality: Evaluates metrics against established thresholds to confirm clustering validity

5.4.2. Supervised Validation

The Supervised Validation component tests whether discovered clusters represent meaningful, predictable patterns. By training machine learning classifiers to predict cluster membership from resistance profiles, this stage validates that clusters capture genuine structure rather than random variation.

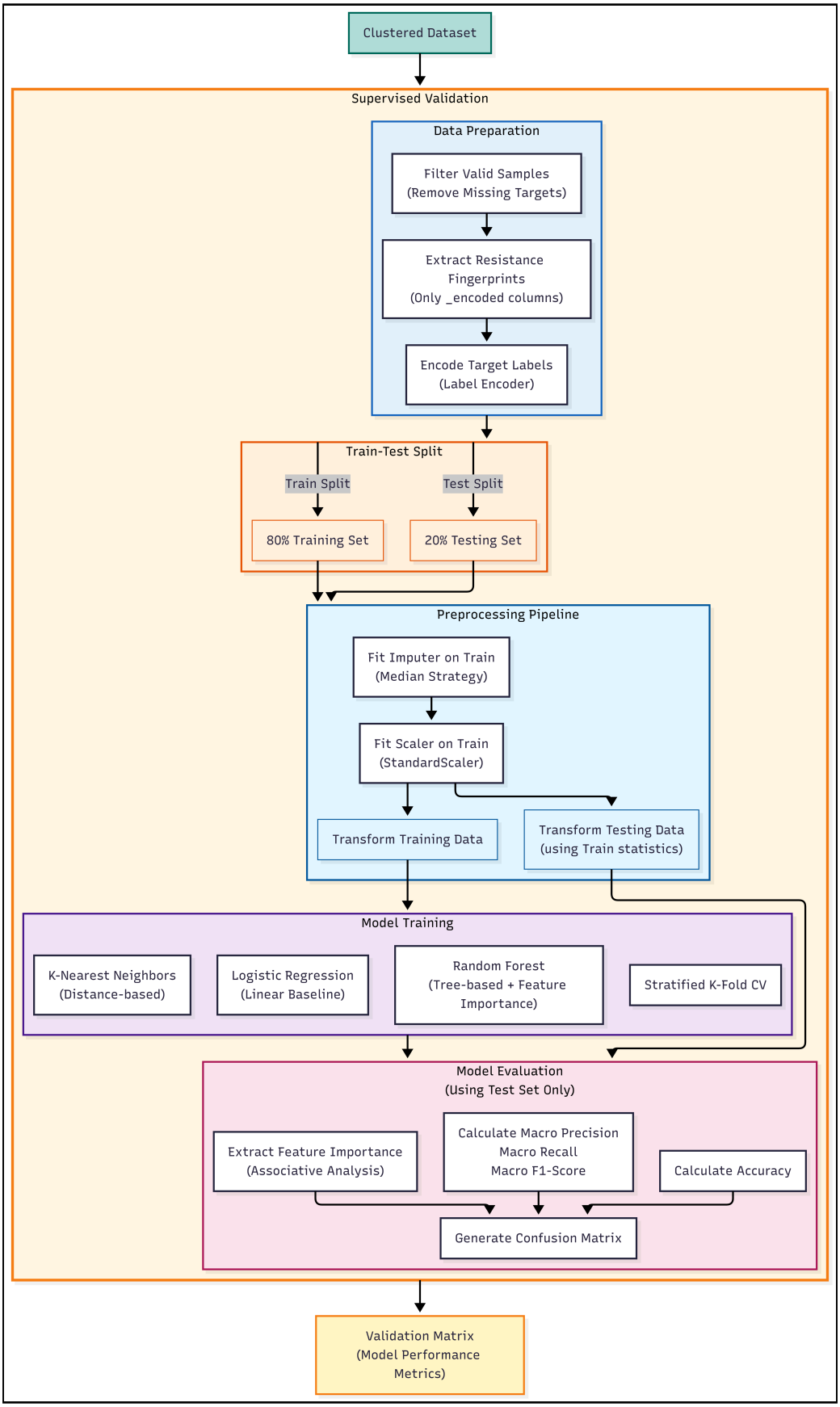

Figure 6: Supervised Validation Architecture

5.4.2.1. Data Preparation

The clustered dataset is prepared for supervised learning by extracting features and encoding target labels.

Key Operations:

- Filter Valid Samples: Removes isolates with missing cluster assignments to ensure complete target labels
- Extract Resistance Fingerprints: Selects only `_encoded` antibiotic columns as features (X), explicitly excluding metadata to prevent data leakage
- Encode Target Labels: Converts categorical cluster identifiers to numerical labels using LabelEncoder for model compatibility

5.4.2.2. Train-Test Split

The dataset is partitioned into training and testing subsets to enable unbiased performance evaluation.

Key Operations:

- 80% Training Set: Used for model fitting and hyperparameter tuning
- 20% Testing Set: Held out for final performance evaluation; models never see this data during training
- Stratified Splitting: Ensures proportional representation of each cluster in both subsets

5.4.2.3. Preprocessing Pipeline

A leakage-safe preprocessing pipeline transforms features using statistics derived only from training data.

Key Operations:

- Fit Imputer on Train: Learns median values from training data to fill missing antibiotic results
- Fit Scaler on Train: Computes mean and standard deviation from training data for standardization
- Transform Training Data: Applies fitted transformations to training features
- Transform Testing Data: Applies the same transformations (using training statistics) to test features, preventing data leakage from test set into preprocessing

5.4.2.4. Model Training

Three classifier architectures are trained to predict cluster membership, each offering different analytical perspectives.

Key Operations:

- Logistic Regression: Linear baseline model providing interpretable coefficients and establishing minimum expected performance
- Random Forest: Ensemble of decision trees capturing non-linear patterns and providing feature importance rankings
- K-Nearest Neighbors: Distance-based classifier validating that clusters occupy distinct regions in feature space
- Stratified K-Fold CV: Cross-validation on training set to assess model stability and tune hyperparameters

5.4.2.5. Model Evaluation

All performance metrics are computed exclusively on the held-out test set to provide unbiased estimates of generalization performance.

Key Operations:

- Calculate Accuracy: Overall proportion of correct cluster predictions on test set
- Calculate Macro Precision/Recall/F1-Score: Per-cluster metrics averaged equally to prevent class imbalance bias
- Generate Confusion Matrix: Detailed breakdown showing which clusters are correctly classified or confused
- Extract Feature Importance: Identifies antibiotics most predictive of cluster membership (from Random Forest), revealing biological drivers of cluster separation

5.4.3. Statistical Analysis

The Statistical Analysis component quantifies pairwise relationships between antibiotic resistances through co-resistance analysis. This method identifies which antibiotics tend to co-occur in resistant isolates, potentially indicating shared resistance mechanisms, genetic linkage, or common selective pressures.

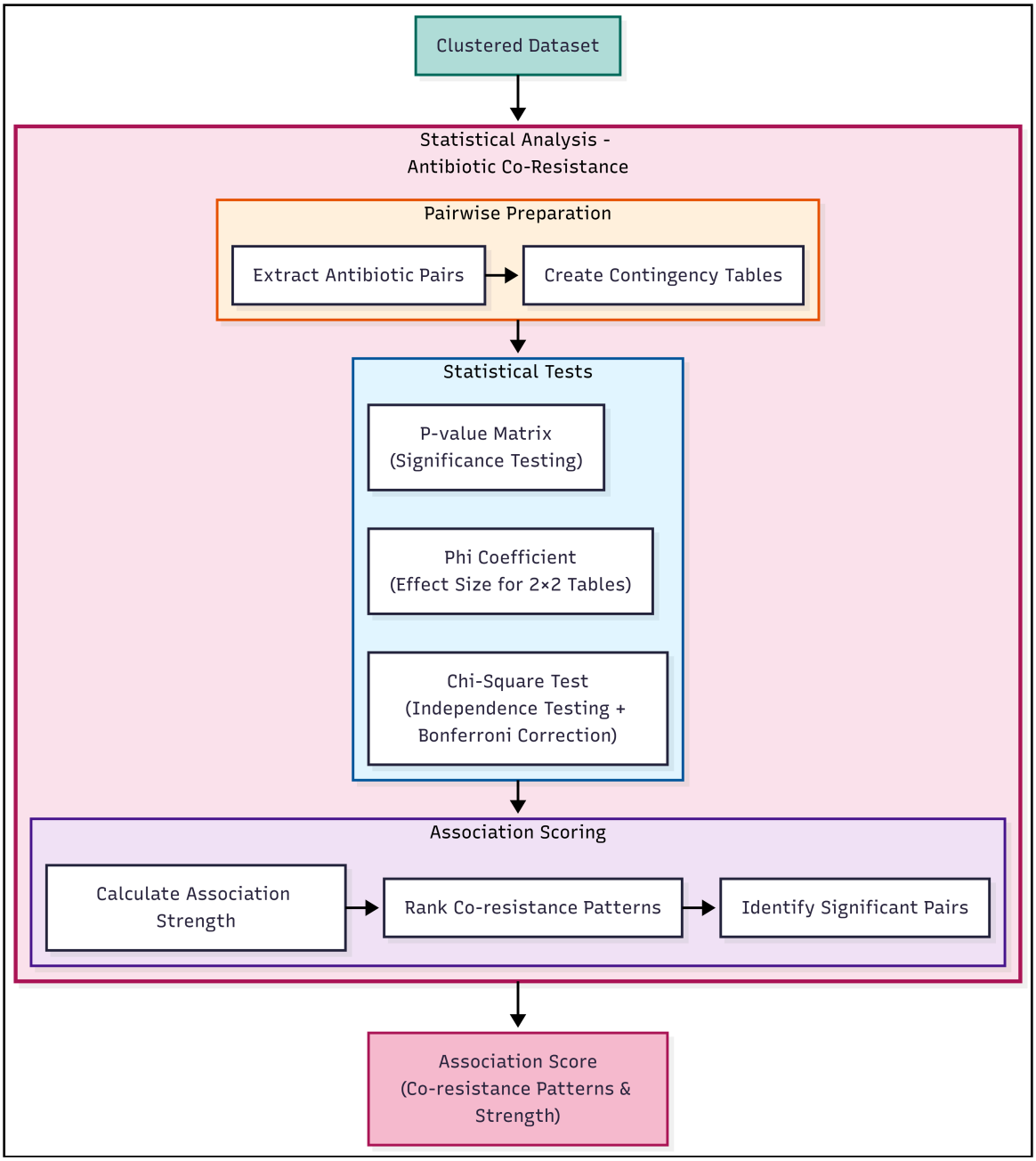

Figure 7: Statistical Analysis Architecture

5.4.3.1. Pairwise Preparation

The analysis begins by systematically examining all possible pairs of antibiotics.

Key Operations:

- Extract Antibiotic Pairs: Generates all unique combinations of antibiotics from the encoded columns using combinatorial enumeration
- Create Contingency Tables: Constructs 2×2 tables for each antibiotic pair showing co-occurrence of resistance (R) and non-resistance (S/I) states

5.4.3.2. Statistical Tests

Rigorous statistical tests assess whether observed co-resistance patterns exceed chance expectations.

Key Operations:

- Chi-Square Test: Tests the null hypothesis that resistance to antibiotic A is independent of resistance to antibiotic B; applies Bonferroni correction to adjust significance threshold for multiple comparisons ($\alpha \ / \ n$ tests)
- Phi Coefficient: Calculates effect size for 2×2 contingency tables using the formula $\varphi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$, where values range from −1 (perfect negative association) to +1 (perfect positive association)
- P-value Matrix: Compiles significance values for all pairwise tests into a symmetric matrix for visualization and filtering

5.4.3.3. Association Scoring

Significant associations are ranked and characterized to identify the strongest co-resistance relationships.

Key Operations:

- Calculate Association Strength: Combines statistical significance (p-value) with effect size (phi coefficient) to rank associations
- Rank Co-resistance Patterns: Orders antibiotic pairs by association strength to prioritize the most important relationships
- Identify Significant Pairs: Filters pairs meeting both significance threshold (Bonferroni-corrected $\alpha < 0.05$) and minimum effect size ($\varphi \geq 0.2$) criteria

5.5. Output Visual Representation

The Output Visual Representation stage consolidates results from all three pattern discovery methods into an interactive dashboard for clinical and epidemiological interpretation. The system employs Streamlit for web-based visualization, enabling stakeholders to explore resistance patterns through multiple complementary views.

Key Outputs:

- Cluster Distribution Charts: Bar charts and pie charts showing isolate distribution across resistance phenotype clusters
- Resistance Heatmaps: Color-coded matrices displaying resistance rates by cluster and antibiotic
- Validation Performance Tables: Summary statistics from supervised validation including accuracy, precision, recall, and F1-scores
- Confusion Matrices: Visual representation of cluster prediction performance

- Co-resistance Network Graphs: Network visualization where nodes represent antibiotics and edges indicate significant co-resistance relationships
- Feature Importance Rankings: Bar charts showing which antibiotics most strongly differentiate clusters

CHAPTER 6

RESULTS AND DISCUSSION

6.1. 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 into three complementary analytical approaches:

- Unsupervised Learning Results presents the resistance phenotype clusters identified through hierarchical agglomerative clustering, including Ward's linkage methodology, cluster characteristics, and internal validation metrics (Silhouette score, WCSS)
- Supervised Learning Validation evaluates the predictive validity of the clustering solution using Random Forest classification, demonstrating that cluster assignments are reproducible from resistance features alone
- Statistical Analysis and Characterization contextualizes the clusters through Principal Component Analysis (PCA), regional and environmental distribution patterns, and co-resistance network relationships

This progression follows a "Discovery → Validation → Interpretation" framework, wherein clusters are first identified (unsupervised), then validated for robustness (supervised), and finally characterized within their epidemiological context (statistical analysis).

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].

6.2. Unsupervised Learning Results

6.2.1. Clustering Parameters

The structure of the resistance dataset was analyzed using hierarchical agglomerative clustering. This approach builds a hierarchy of clusters by progressively merging similar isolates based on their resistance profiles.

6.2.1.1. Ward's Linkage Method

Ward's minimum variance method was employed as the linkage criterion [11]. Unlike other linkage methods that focus on pairwise distances (e.g., single or complete linkage), Ward's method minimizes the total within-cluster variance at each merger step. This optimization criterion is particularly effective for discovering compact, spherical clusters that correspond to distinct resistance phenotypes.

The Within-Cluster Sum of Squares (WCSS) quantifies the compactness achieved by Ward's method:

Table 20: Within-Cluster Sum of Squares (WCSS) by cluster solution. ΔWCSS shows the reduction from the previous k. The elbow point at k=4 marks diminishing returns in variance reduction.

k	WCSS	ΔWCSS	% Reduction
2	2395.19	—	—
3	1765.12	630.07	26.3%
4	1482.92	282.20	16.0%
5	1234.94	247.98	16.7%
6	1009.38	225.56	18.3%

In Table 20, k represents the number of clusters tested, WCSS is the Within-Cluster Sum of Squares measuring total variance within all clusters, ΔWCSS shows the absolute reduction from the previous k value, and % Reduction indicates the relative improvement in cluster compactness. The elbow point occurs where percent reduction begins to plateau.

6.2.1.2. Euclidean Distance

Euclidean distance was selected as the dissimilarity metric, measuring the geometric distance between isolate resistance vectors. This metric is the required complement to Ward's linkage method, as Ward's objective function is defined based on squared Euclidean distances. The combination of Ward's linkage and Euclidean distance provides a robust framework for identifying natural groupings in the multidimensional resistance data.

Table 21: Euclidean distance thresholds defining cluster solutions. The optimal k=4 solution is stable within the distance range 22.27 to 23.76.

Cluster Solution (k)	Lower Threshold (d)	Upper Threshold (d)
5	—	22.27
4	22.27	23.76
3	23.76	35.50
2	35.50	41.20

In Table 21, Cluster Solution (k) indicates the resulting number of clusters, Lower Threshold (d) is the minimum Euclidean distance at which that solution becomes stable, and Upper Threshold (d) is the maximum distance before a merge reduces the cluster count.

6.2.2. Optimal Cluster Solution

Hierarchical agglomerative clustering using Ward's linkage method and Euclidean distance (as described in Section 6.2.1, Clustering Parameters) was applied to 491 bacterial isolates collected from the water-fish-human nexus across three Philippine regions. Cluster (k) solutions from k=2 to k=8 were evaluated for optimal selection, with metrics computed to k=10 for validation purposes [36].

Table 22: Cluster Validation Metrics Across k Values

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

In Table 22, k is the number of clusters and Silhouette Score measures cluster separation (≥0.40 indicates strong structure). WCSS quantifies compactness (lower is better), while Calinski-Harabasz (higher is better) and Davies-Bouldin (lower is better) provide complementary validity checks.

The k=4 cluster solution was selected as the optimal configuration through a multi-criteria decision framework [16], [37]. The k=4 solution represents the elbow point in the WCSS curve and satisfies the silhouette threshold (≥0.40). Furthermore, the Davies-Bouldin index at k=4 (1.089) confirms reasonable separation without excessive overlap, supported by a competitive Calinski-Harabasz score (192.78), indicating dense and well-separated clusters.

Table 23: 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.

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

The columns in Table 23 evaluate each cluster solution across multiple dimensions. Silhouette scores measure cluster cohesion (where ≥0.40 indicates strong structure), while the Elbow Point identifies the diminishing returns in variance reduction. Interpretability assesses the biological relevance of resulting groups, and Min Cluster Size ensures no cluster falls below n=20, a threshold required for reliable phenotype estimation.

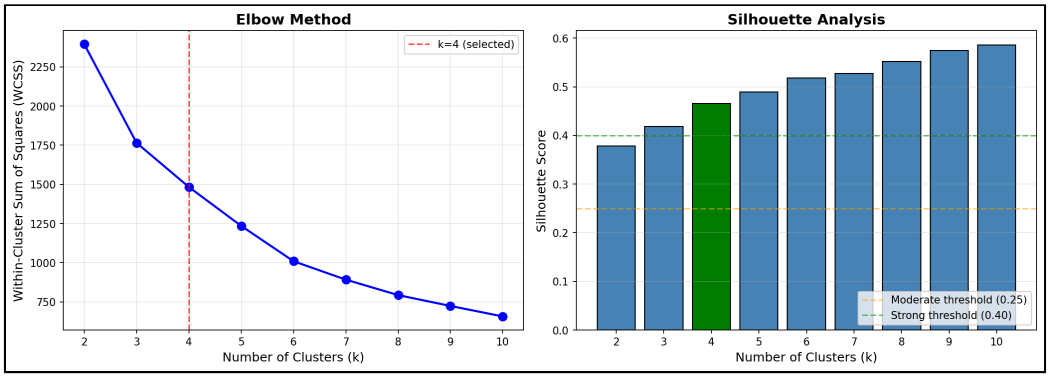

Figure 8: Elbow method (left) and silhouette analysis (right) for cluster validation. The WCSS curve shows the elbow point at k=4, while the silhouette plot confirms moderate-to-strong structure at this configuration.

6.2.3. Cluster Characteristics

The four identified clusters exhibited distinct resistance phenotype profiles:

Table 24: Cluster composition summary showing species distribution, MDR prevalence, and dominant resistance patterns. C3 (MDR Archetype) is notably distinct with high multidrug resistance rates and broad species diversity.

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

In Table 24, the columns describe each group's key features. Cluster is the group name, while N Isolates shows the number and percentage of samples it contains. Dominant Species

lists the most common bacteria found in that group, and MDR % shows how many are multidrug-resistant. Finally, Top Resistant Antibiotics lists the specific drugs that the group resists, using these abbreviations: AN=Amikacin, GM=Gentamicin, AM=Ampicillin, CF=Cefalotin, CN=Cefalexin, TE=Tetracycline, DO=Doxycycline, FT=Nitrofurantoin.

6.2.3.1. 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) and cephalosporins (CN: Cefalexin). Geographically, 17 of 23 C1 isolates (73.9%) originate from Region III – Central Luzon, with 16 of 23 (69.6%) derived from water samples.

6.2.3.2. 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.

6.2.3.3. 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 100-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.

6.2.3.4. 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.

6.2.4. Visualizations of Cluster Structure

The cluster resistance profiles, hierarchical structure, and geographic distribution are visualized in the following figures. Figure 9 presents the mean resistance scores for each cluster across all antibiotics, revealing distinct phenotypic signatures.

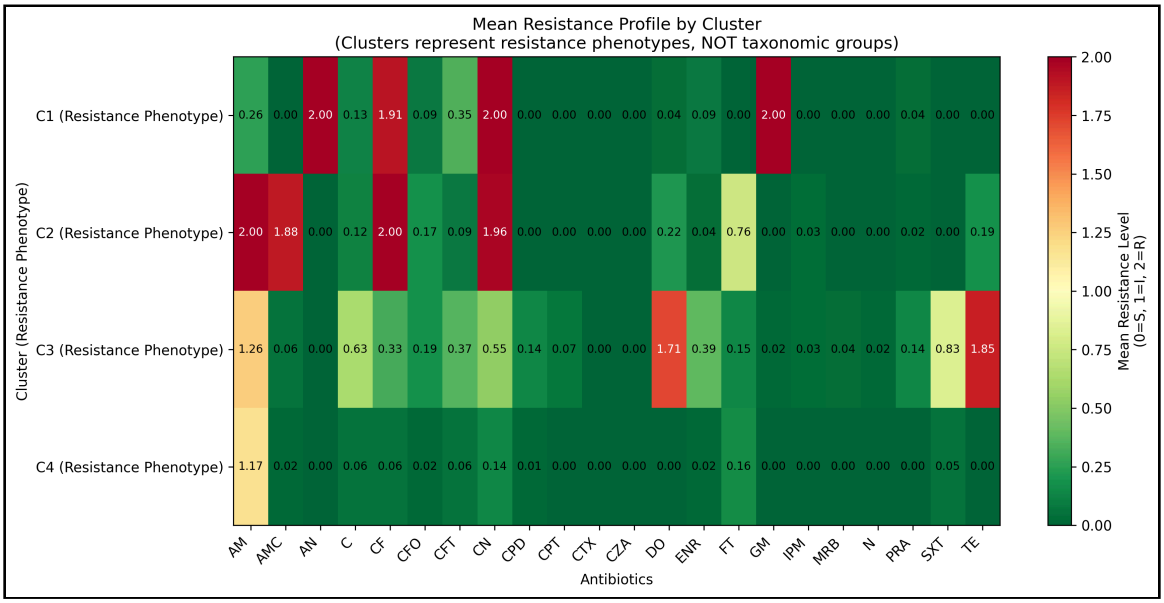

Figure 9: Cluster resistance profiles showing mean resistance scores (0–2 scale) per antibiotic for each of the four clusters. C1 (Salmonella-Aminoglycoside) shows elevated aminoglycoside resistance; C2 (Enterobacter-Penicillin) exhibits β-lactam resistance; C3 (MDR Archetype) displays broad resistance including tetracyclines; C4 (Susceptible Majority) shows minimal resistance across all classes.

The hierarchical structure of the clustering solution is visualized in Figure 10, which links the dendrogram with a resistance heatmap to show the relationship between isolate groupings and their resistance patterns.

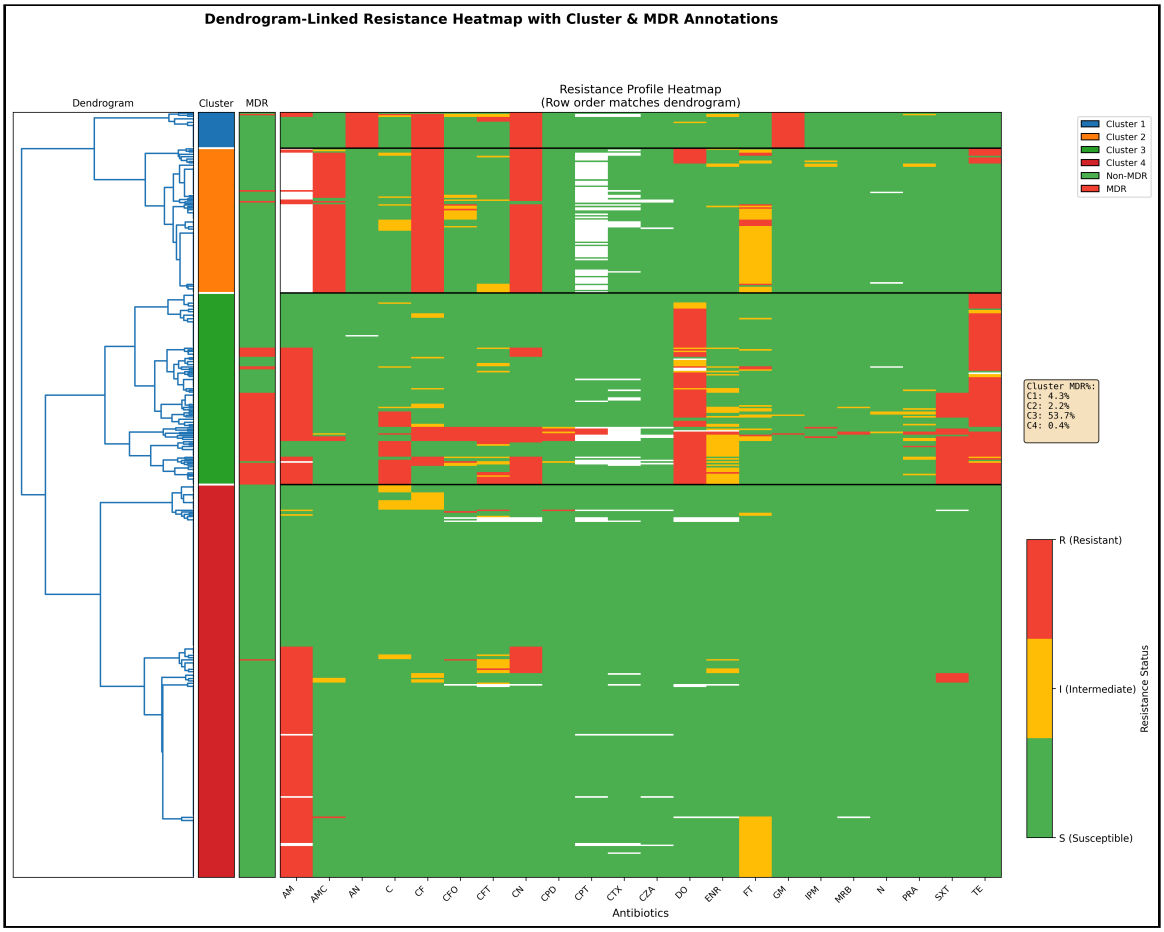

Figure 10: Dendrogram-linked resistance heatmap showing hierarchical clustering structure. Rows represent isolates ordered by dendrogram position; columns represent antibiotics. Color intensity indicates resistance level (blue=susceptible, red=resistant). The four main clusters are visible as distinct blocks with characteristic resistance patterns.

The detailed dendrogram in Figure 11 illustrates the complete hierarchical structure with cluster assignments.

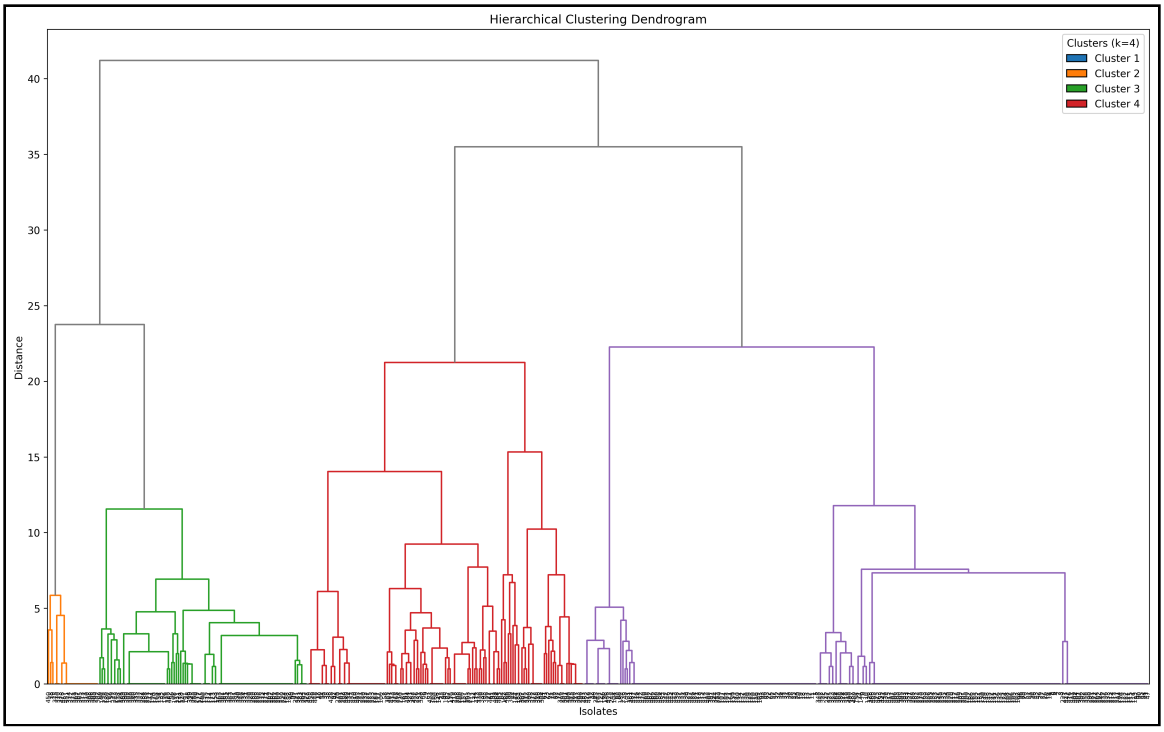

Figure 11: High-resolution dendrogram showing hierarchical agglomerative clustering of 491 isolates using Ward's linkage. The horizontal dashed line indicates the cut point for k=4 clusters. Distinct color branches represent the four identified phenotype clusters.

Geographic and environmental distributions of cluster assignments are shown in Figure 12 and Figure 13 respectively.

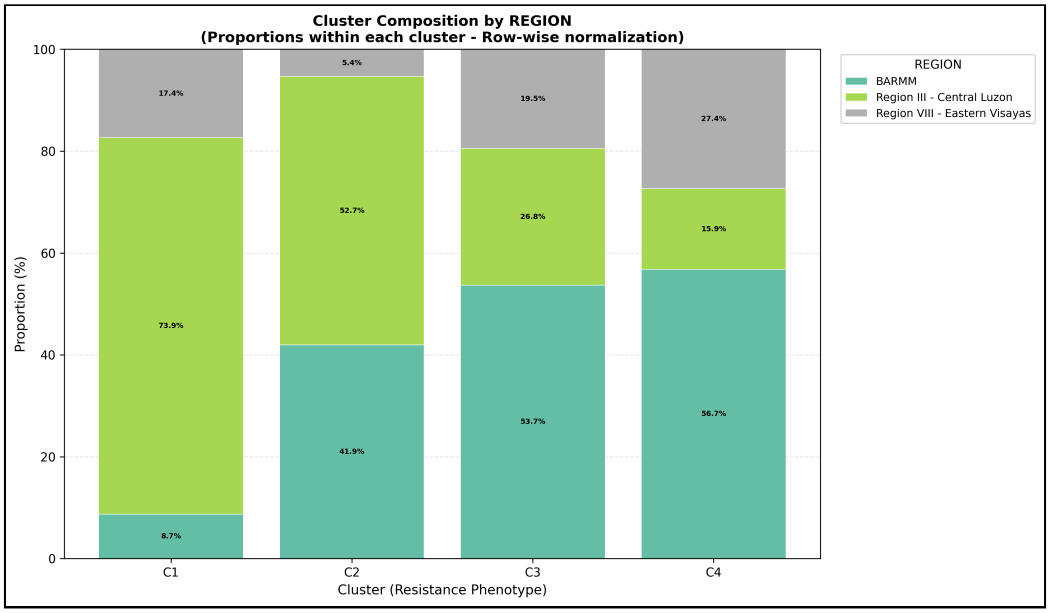

Figure 12: Cluster composition by geographic region. The stacked bar chart shows the proportion of each cluster originating from BARMM, Region III (Central Luzon), and Region VIII (Eastern Visayas). C3 (MDR Archetype) shows a strong association with BARMM, while C1 (Salmonella-Aminoglycoside) is predominantly from Region III.

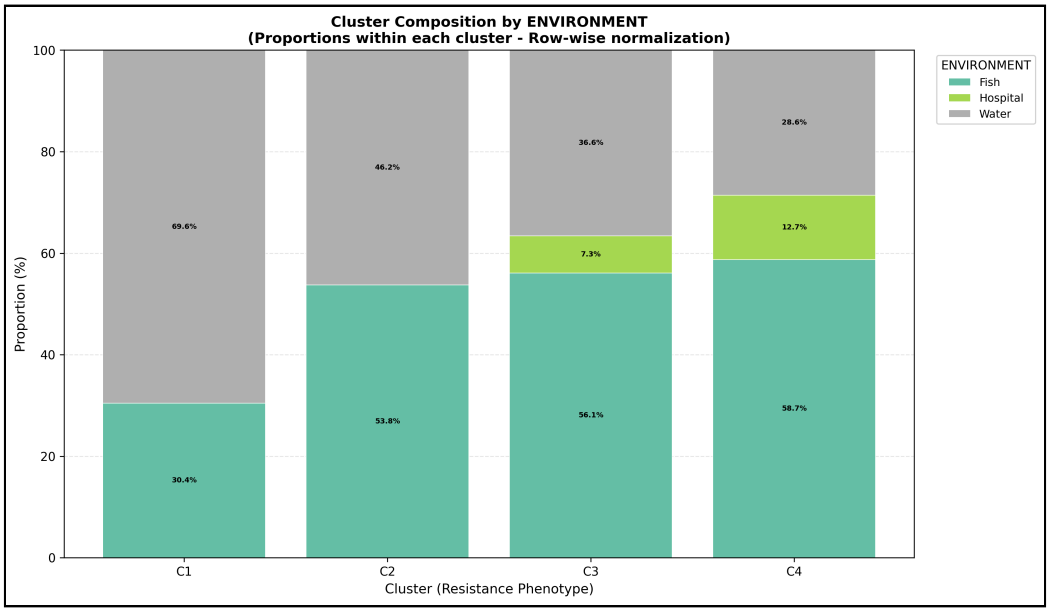

Figure 13: Cluster composition by environmental source. The stacked bar chart shows the distribution of fish, water, and hospital-derived isolates across clusters. C3 (MDR Archetype) contains the majority of hospital-environment isolates, while C4 (Susceptible Majority) is predominantly from aquatic environments.

The resistance heatmap in Figure 14 provides a comprehensive view of resistance patterns across all isolates and antibiotics.

Figure 14: Resistance heatmap showing AST results for all 491 isolates across 21 antibiotics. Isolates are ordered by cluster assignment; antibiotics are ordered by antimicrobial class. The heatmap reveals clear phenotypic boundaries between clusters and identifies antibiotics with high discriminatory power.

6.3. Supervised Learning Validation

The supervised validation approach evaluates whether the clusters identified through unsupervised hierarchical clustering represent reproducible, predictable patterns in the resistance data. By training a classifier to predict cluster membership from resistance features

alone, we can assess whether the cluster assignments capture genuine structure rather than artifacts of the clustering algorithm.

6.3.1. Random Forest Classification

A Random Forest classifier was trained to predict cluster membership using the 22-dimensional encoded resistance data as input features [24]. The model was evaluated on a held-out test set (20%) after stratified splitting to ensure robust performance estimates across all four clusters.

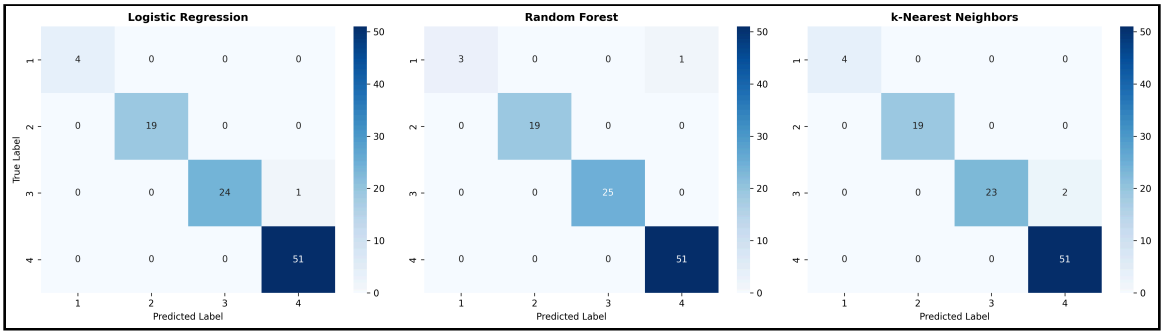

Figure 15: Confusion Matrices for Supervised Classifiers. Comparison of confusion matrices for Logistic Regression, Random Forest, and k-Nearest Neighbors. The diagonal dominance across all models confirms robust cluster separability.

Table 25 compares the performance of the three classifiers. All models achieved greater than 96% macro F1-score, indicating that the clusters are robust and distinguishable regardless of the classification algorithm used.

Table 25: Comparison of Supervised Learning Models. All three model families (Linear, Tree-based, Distance-based) achieved >96% macro F1-score, confirming that cluster separability is a property of the data structure, not an artifact of a specific algorithm.

Model	Category	Accuracy	Macro F1-Score
Logistic Regression	Linear	99.0%	0.99
Random Forest	Tree-based	99.0%	0.96
k-Nearest Neighbors	Distance-based	98.0%	0.98

The exceptionally high classification accuracy (99.0%) demonstrates that cluster assignments are highly predictable from resistance data alone. This confirms that the four clusters represent distinct, reproducible resistance phenotypes rather than arbitrary groupings. The balanced macro F1-score (0.96) indicates excellent performance across all cluster sizes, including the smaller Cluster 1 (n=23).

6.3.2. Feature Importance

The Random Forest model also provides interpretable feature importance scores, indicating which antibiotics contribute most to cluster discrimination.

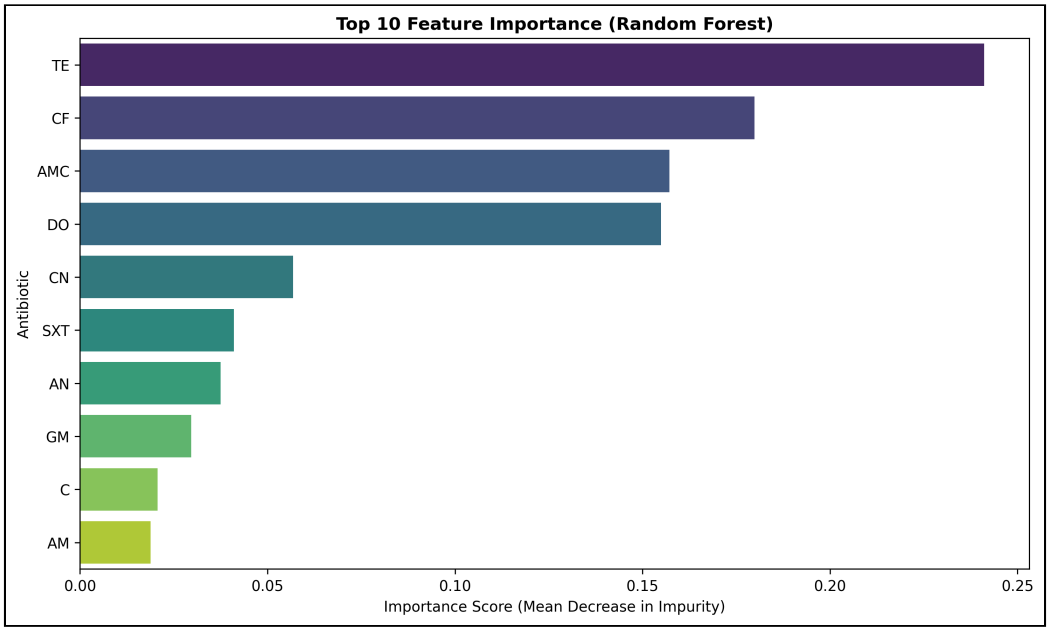

Figure 16: Feature Importance for Random Forest Classifier. Tetracycline (TE) and Doxycycline (DO) are the most discriminatory features, driving the separation of the MDR Archetype cluster. Importance scores represent the mean decrease in impurity.

Tetracycline, cephalothin, and amoxicillin-clavulanic acid emerge as the most discriminating features, with tetracycline (0.241) retaining its strong role in defining the MDR Archetype cluster (C3). The prominence of beta-lactams (cephalothin, AMC) and tetracyclines (TE, DO) confirms that these drug classes are the primary drivers of phenotypic separation.

6.3.3. Sensitivity Analysis: Split Ratio and Cross-Validation

To validate the robustness of the chosen experimental configuration (80/20 split, Random Forest), 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.

6.3.3.1. Split Ratio Comparison

Table 26: F1 Scores Across Different Train–Test Split Ratios (Cluster Discrimination)

Split	Model	F1 Score	Accuracy	Stability (std)
70/30	Logistic Regression	0.984 ± 0.006	0.985	0.006
70/30	Random Forest	0.984 ± 0.014	0.993	0.014
70/30	KNN	0.979 ± 0.010	0.977	0.010
80/20	Logistic Regression	0.987 ± 0.005	0.986	0.005
80/20	Random Forest	0.982 ± 0.022	0.994	0.022
80/20	KNN	0.984 ± 0.012	0.982	0.012
90/10	Logistic Regression	0.992 ± 0.010	0.992	0.010
90/10	Random Forest	0.960 ± 0.050	0.988	0.050
90/10	KNN	0.989 ± 0.010	0.988	0.010

6.3.3.2. Cross-Validation Comparison

Table 27: F1 Scores Across Different Cross-Validation Configurations

CV Folds	Model	F1 Score	Accuracy	Stability (std)
5-fold	Logistic Regression	0.989 ± 0.009	0.990	0.009
5-fold	Random Forest	0.989 ± 0.011	0.994	0.011
5-fold	KNN	0.979 ± 0.009	0.978	0.009
10-fold	Logistic Regression	0.989 ± 0.015	0.990	0.015
10-fold	Random Forest	0.986 ± 0.027	0.994	0.027
10-fold	KNN	0.982 ± 0.015	0.982	0.015

The analysis confirms consistently high performance (>0.96 F1) across all clusters, with Cluster 2 (Enterobacter-Penicillin) showing perfect recall. Cluster 1 (Salmonella-Amino-glycoside) had slightly lower precision (0.93), likely due to the broader spectrum of resistance patterns in that group. The 80/20 split with 5-fold cross-validation was confirmed as an optimal balance between training adequacy and evaluation reliability.

6.3.4. Validation Implications

The successful supervised validation provides several key insights:

1. Cluster Reproducibility: The 99.0% accuracy confirms that an independent learning algorithm can recover the same groupings with near-perfect precision, substantially reducing concerns about clustering artifacts.
2. Phenotype Distinctiveness: High precision and recall indicate clear boundaries between resistance phenotypes, supporting their use as meaningful epidemiological categories.
3. Feature Interpretability: The alignment between feature importance and known resistance mechanisms—particularly the strong discriminatory power of tetracycline-class antibiotics for MDR phenotypes—validates the biological coherence of the clustering solution.

6.3.5. Principal Component Analysis Visualization

To complement the supervised validation, Principal Component Analysis (PCA) was applied to visualize the high-dimensional resistance data in reduced dimensions. Figure 17 shows the variance explained by each principal component.

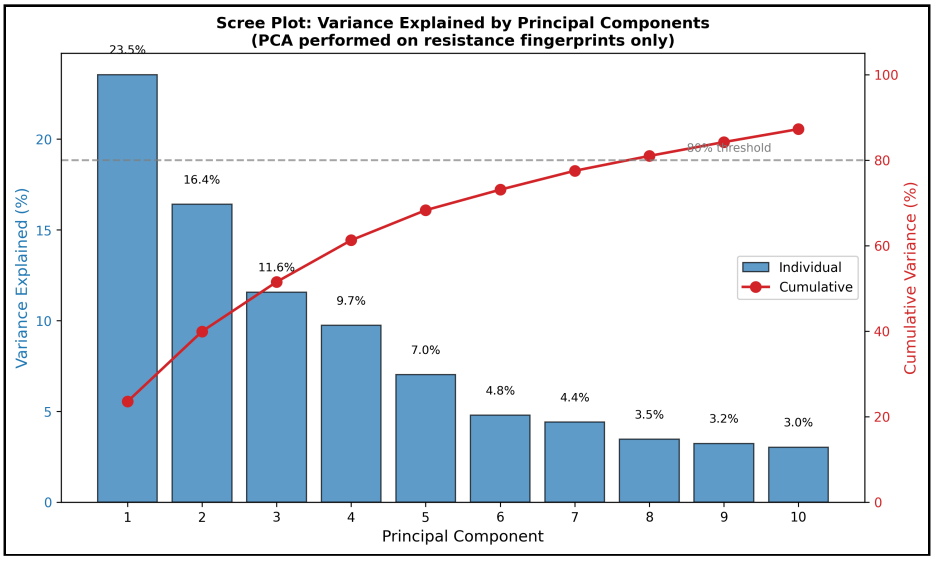

Figure 17: PCA scree plot showing cumulative variance explained. The first two principal components capture substantial variance, enabling meaningful 2D visualization of the resistance data structure.

The PCA biplot in Figure 18 reveals the contribution of individual antibiotics to the principal component space.

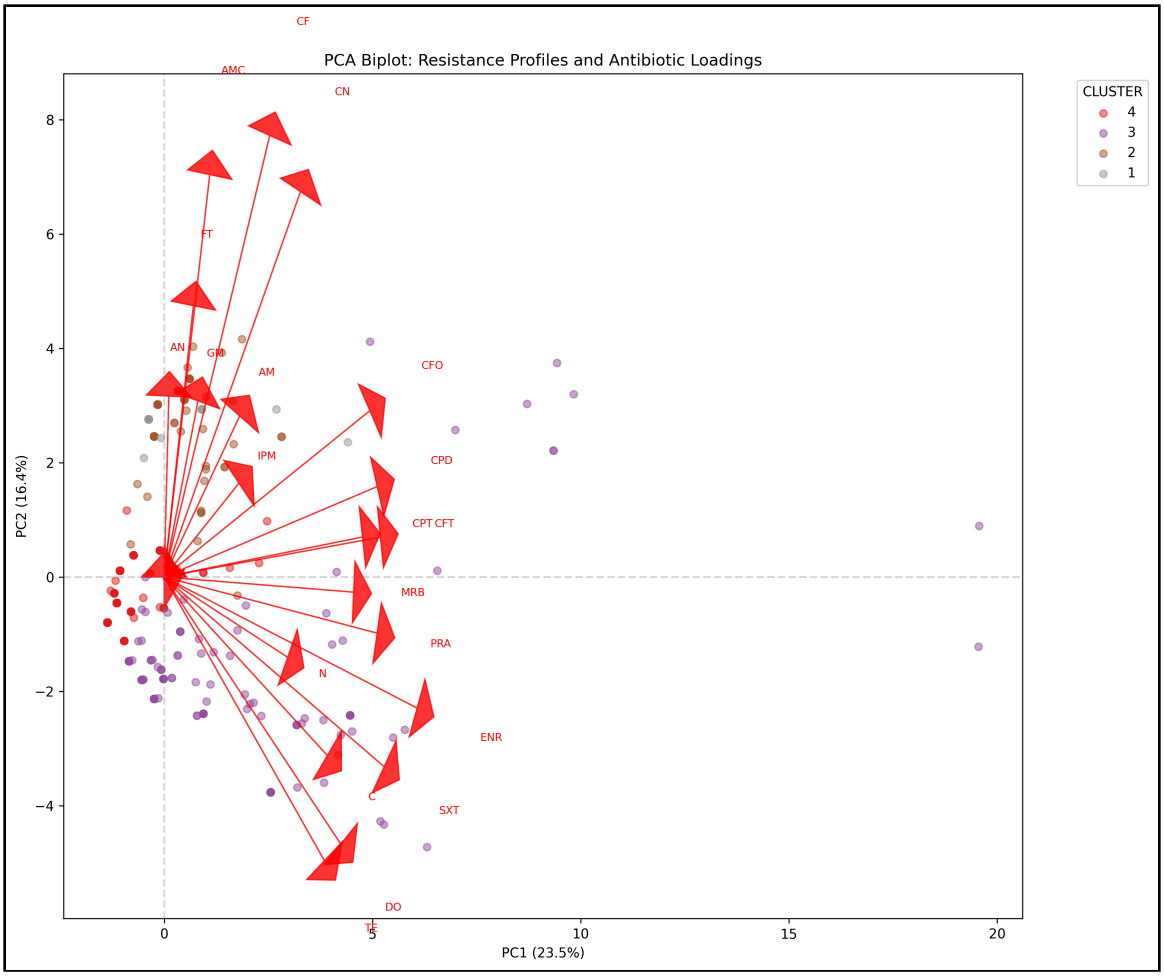

Figure 18: PCA biplot showing isolates (points) and antibiotic loadings (vectors) in the first two principal components. Vector directions indicate antibiotic contributions to cluster separation. Tetracycline-class antibiotics (TE, DO) show strong loadings consistent with their importance in defining the MDR Archetype cluster.

Figure 19 visualizes the cluster assignments in PCA space, demonstrating the separability of the four phenotype groups.

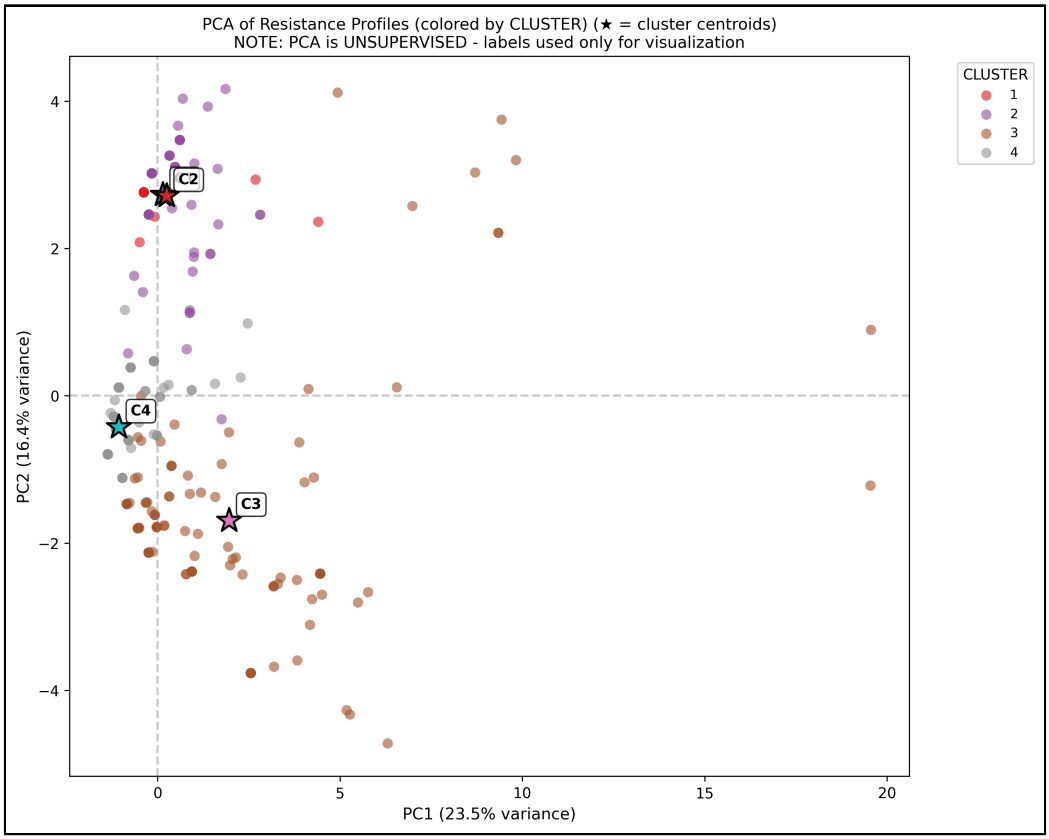

Figure 19: PCA visualization colored by cluster assignment. The four clusters show distinct spatial distributions in the reduced-dimensional space, confirming that the clustering solution captures meaningful phenotypic structure. C3 (MDR Archetype) and C4 (Susceptible Majority) form clearly separated regions.

The relationship between cluster structure and MDR status is visualized in Figure 20.

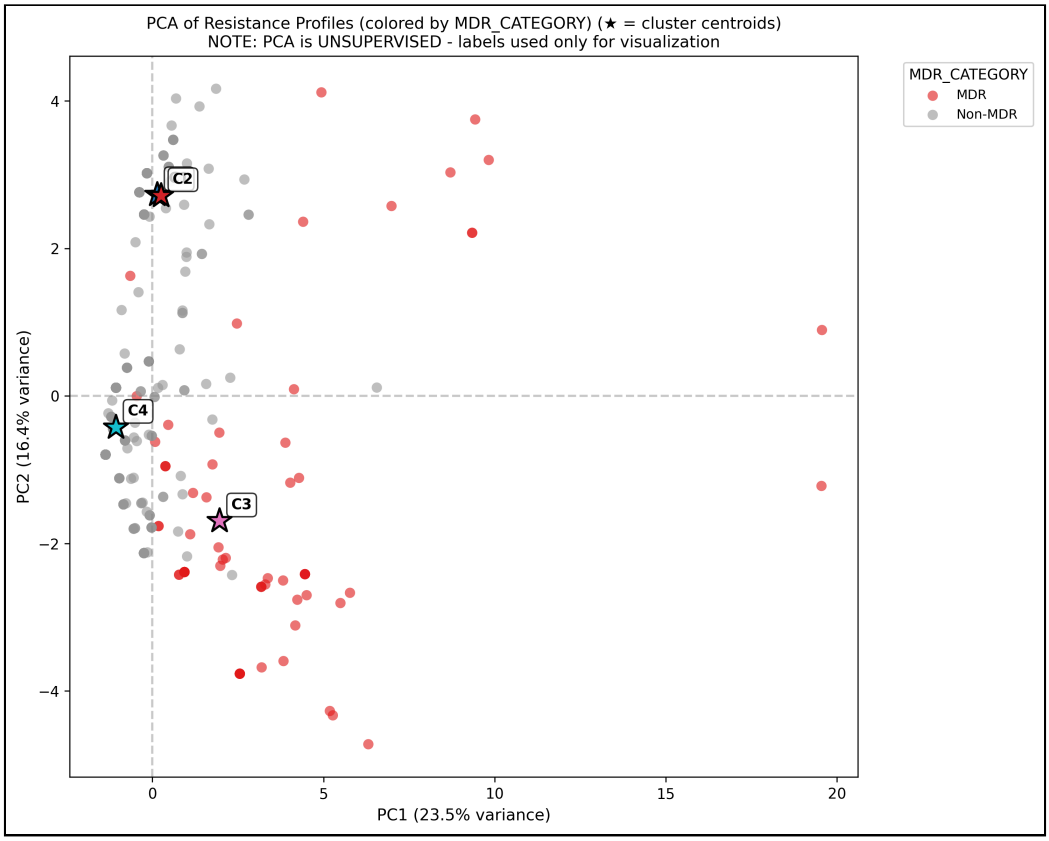

Figure 20: PCA visualization colored by MDR status. MDR isolates (red) cluster distinctly from susceptible isolates (blue), with the majority concentrated in the C3 (MDR Archetype) region of the plot. This confirms the phenotypic coherence of the MDR classification.

6.3.6. Silhouette Analysis Detail

The detailed silhouette plot for the k=4 solution is presented in Figure 21, showing cluster cohesion and separation for each isolate.

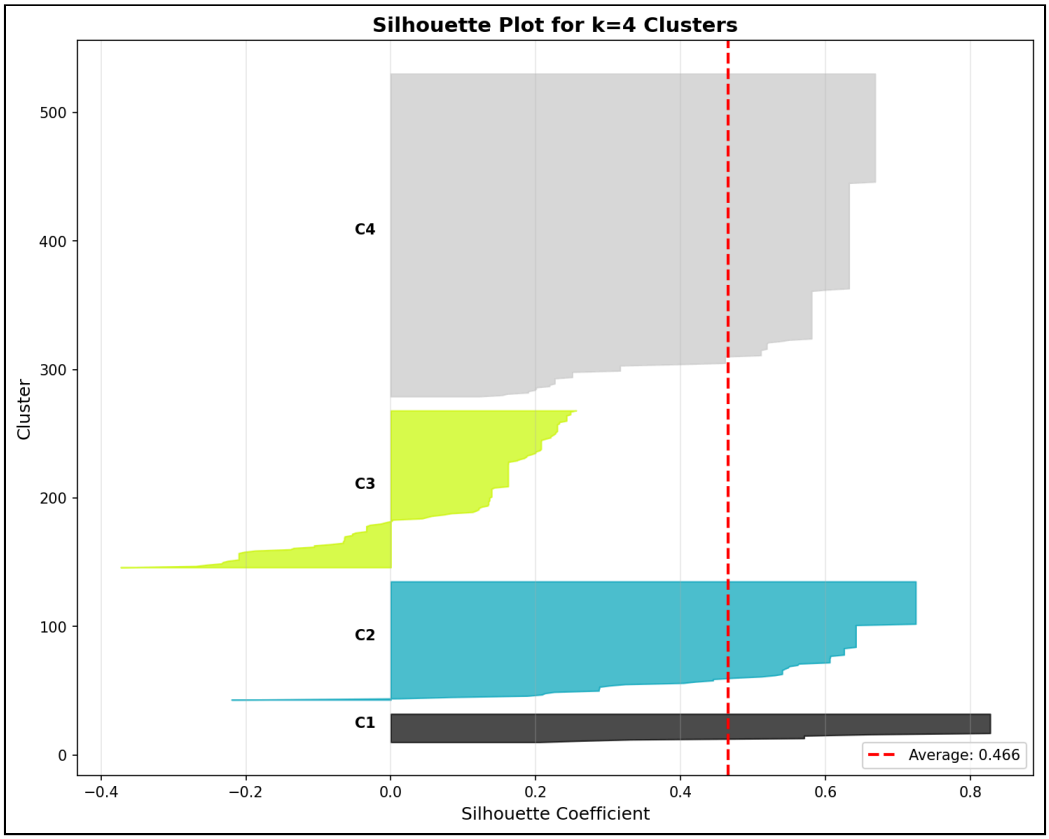

Figure 21: Silhouette plot for k=4 cluster solution. Each bar represents an isolate's silhouette coefficient; longer bars indicate better cluster fit. All four clusters show predominantly positive silhouette values, with C3 and C4 demonstrating the strongest internal cohesion (mean silhouette = 0.466).

6.4. Statistical Analysis and Characterization

This section presents complementary statistical analyses that characterize the identified resistance phenotypes within their epidemiological context. These include dimensionality reduction via Principal Component Analysis (PCA), examination of regional and environmental distribution patterns, and co-resistance network relationships.

6.4.1. Principal Component Analysis

Principal Component Analysis (PCA) was performed on the 22-dimensional encoded resistance data to visualize the underlying structure and assess its dimensionality.

Table 28 details the contributions of the first five principal components (PC) to the total variance of the isolate profiles. In this table, Component identifies the PC axis, Variance Explained (%) indicates how much of the dataset's total information is captured by that specific component, and Cumulative (%) shows the total variance accounted for by all components up to that point.

Table 28: Variance explained by the first five principal components of the encoded resistance matrix

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%

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].

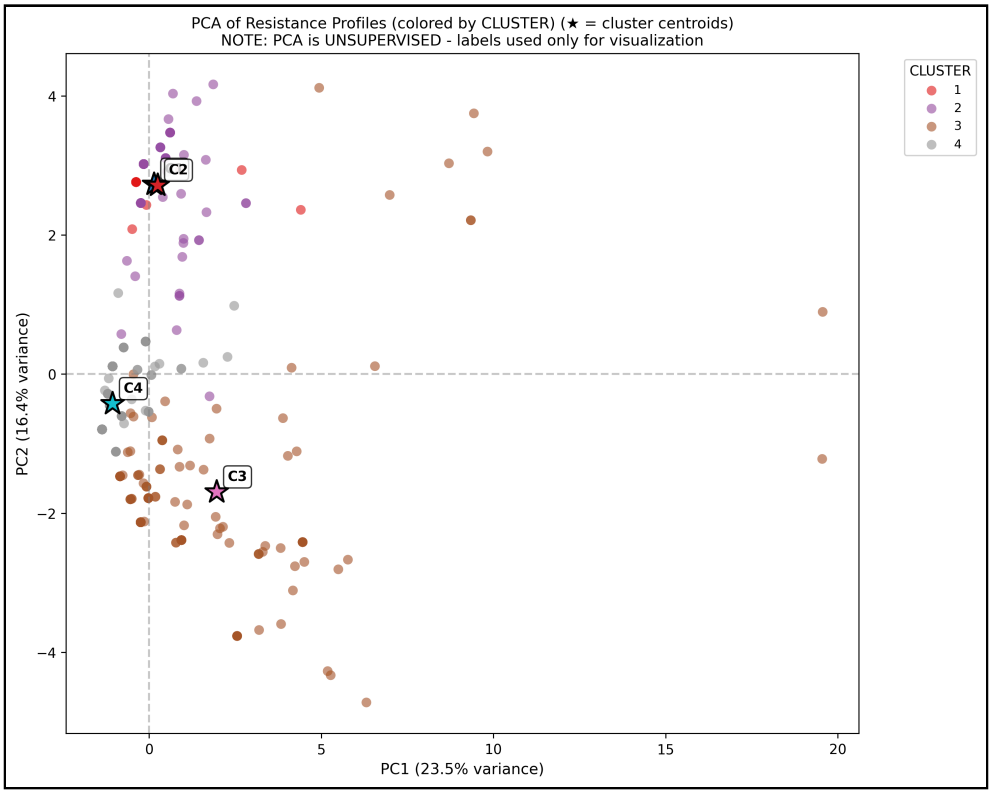

Figure 22: PCA projection of 491 isolates colored by cluster assignment. The scatter plot visualizes the separation of the four distinct resistance phenotypes along the first two principal components (PC1 and PC2).

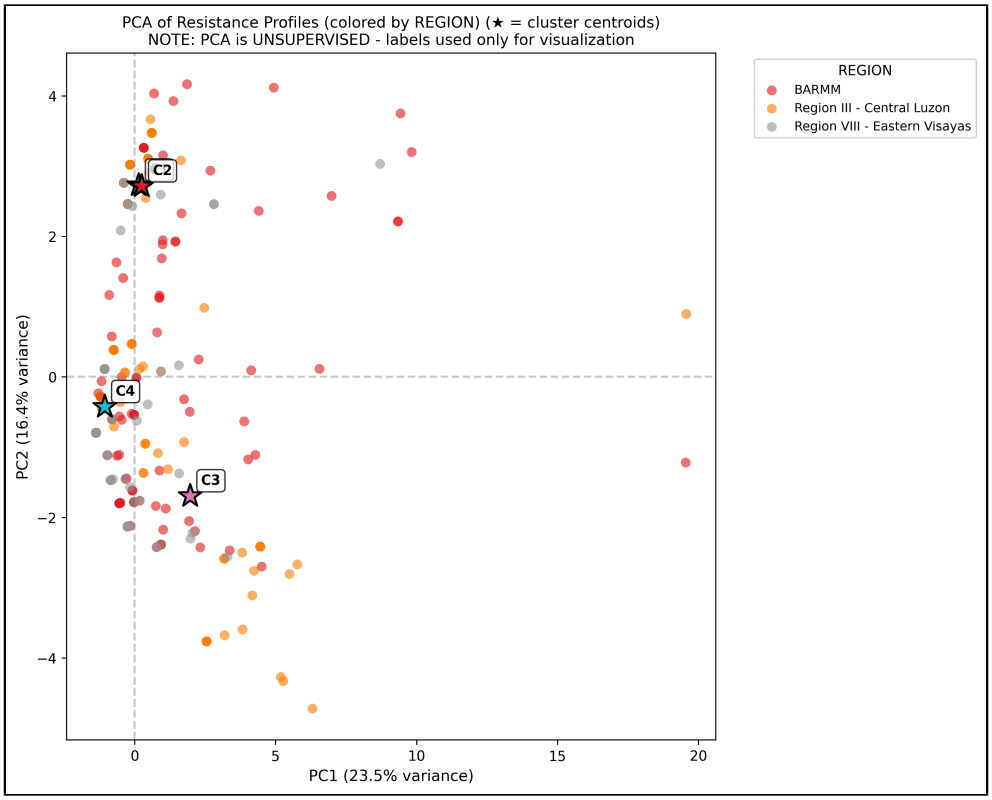

Figure 23: PCA projection colored by geographic region. Isolates from BARMM, Region III (Central Luzon), and Region VIII (Eastern Visayas) show overlapping distributions with subtle regional clustering tendencies, particularly along PC2.

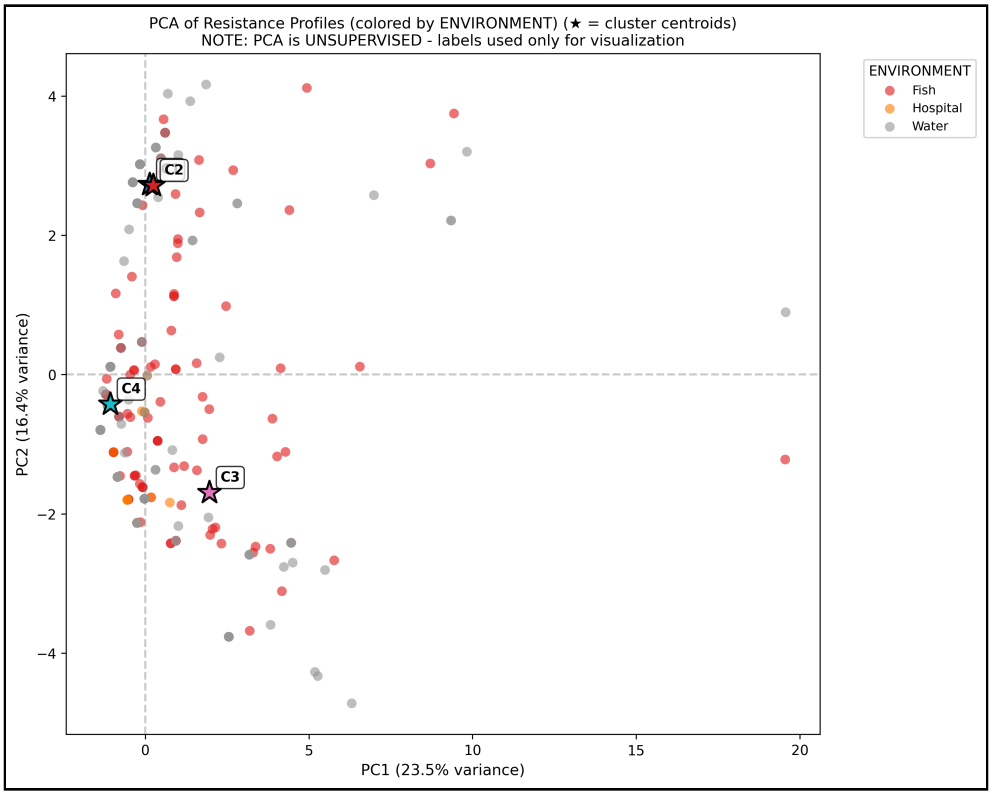

Figure 24: PCA projection colored by environmental source. Fish, water, and hospital-derived isolates are distributed across the phenotypic space, with hospital isolates showing a tendency toward the susceptible region (C4 territory) on PC1.

6.4.2. Regional Distribution Patterns

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

Table 29: Regional distribution of resistance phenotype clusters (percentage of each cluster by region)

Cluster	Region			Total
	BARMM	Central Luzon	Eastern Visayas	
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%

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.

6.4.3. Environmental Niche Associations

Table 30: Environmental distribution of resistance phenotype clusters

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%

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].

6.4.4. Resistance Distribution Analysis

The distribution of resistance levels across the dataset is characterized by the Multiple Antibiotic Resistance (MAR) index and Multi-Drug Resistance (MDR) classification.

Figure 25 shows the MAR index distribution across all isolates, while Figure 26 illustrates MDR prevalence.

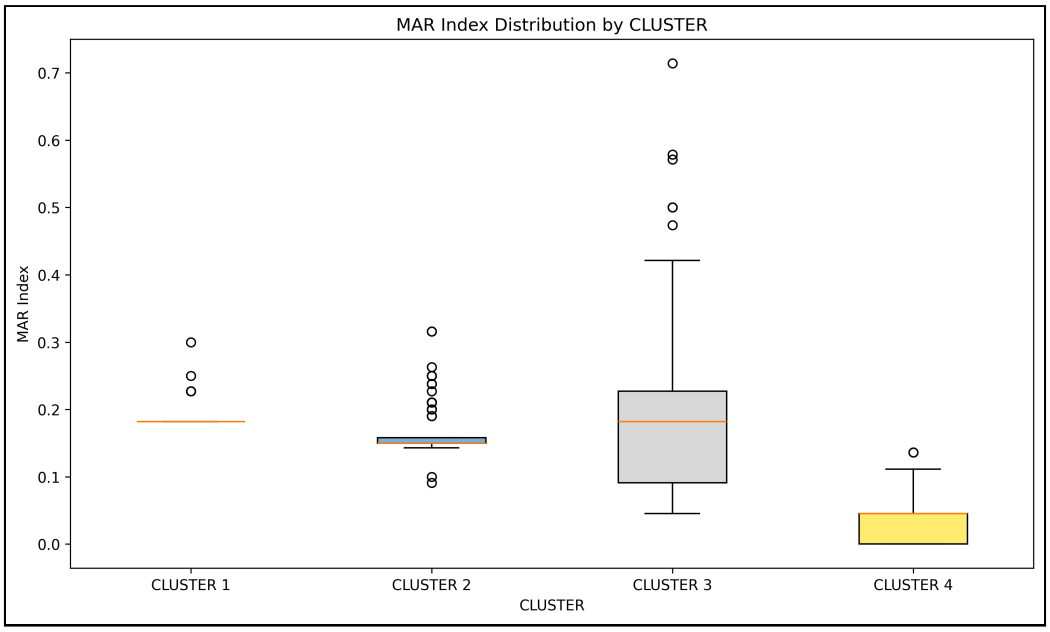

Figure 25: Distribution of Multiple Antibiotic Resistance (MAR) index across 491 isolates. The histogram shows the proportion of antibiotics to which each isolate is resistant. Most isolates exhibit low MAR values (< 0.2), with a distinct high-MAR subpopulation corresponding to the MDR Archetype cluster (C3).

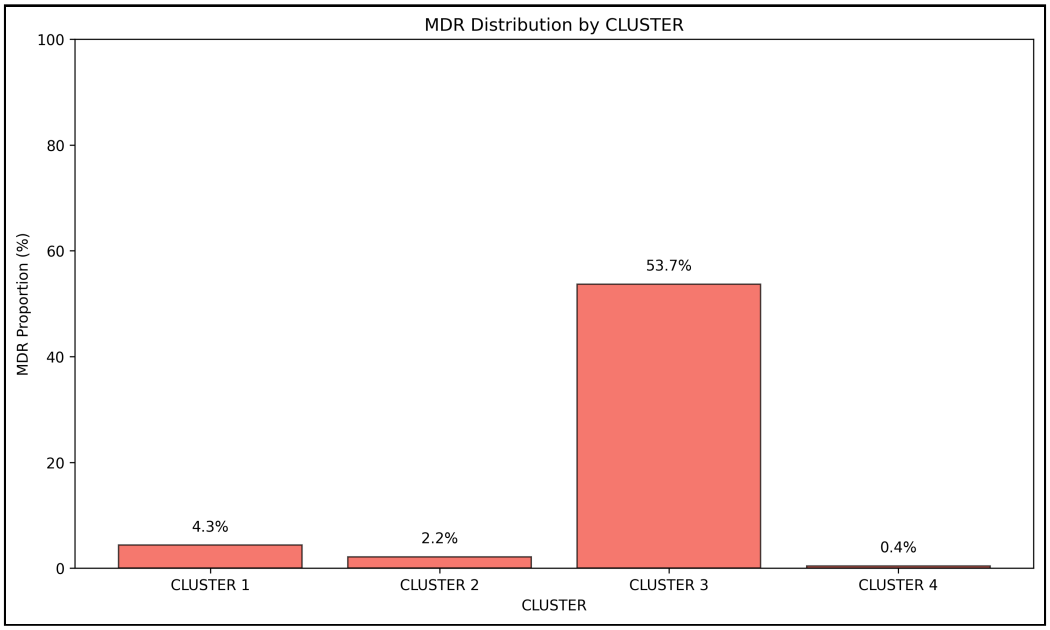

Figure 26: Multi-Drug Resistance (MDR) status distribution across clusters. The bar chart shows the proportion of MDR (≥3 resistant classes) and non-MDR isolates within each cluster. C3 (MDR Archetype) contains 53.7% MDR isolates, while other clusters exhibit less than 5% MDR prevalence.

6.4.5. Antibiotic Clustering Analysis

Hierarchical clustering was also applied to antibiotics to identify groups with correlated resistance patterns. Figure 27 presents the antibiotic dendrogram, revealing clusters of antibiotics that tend to co-occur in resistance profiles.

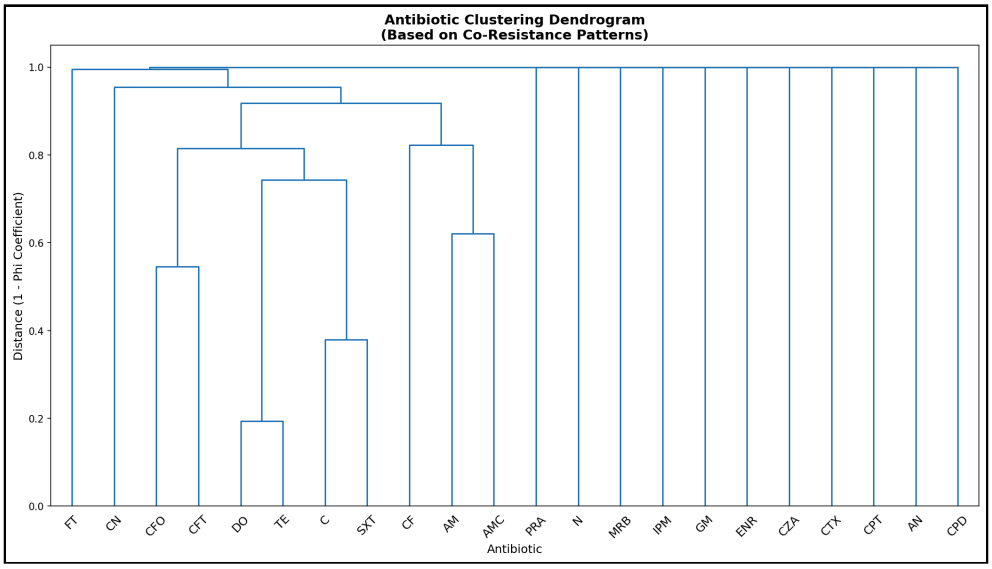

Figure 27: Dendrogram of antibiotic clustering based on resistance co-occurrence patterns. Antibiotics that cluster together exhibit similar resistance profiles across isolates. The tight grouping of tetracycline-class antibiotics (TE, DO) and fluoroquinolones (ENR, MRB, PRA) reflects mechanistically related resistance mechanisms.

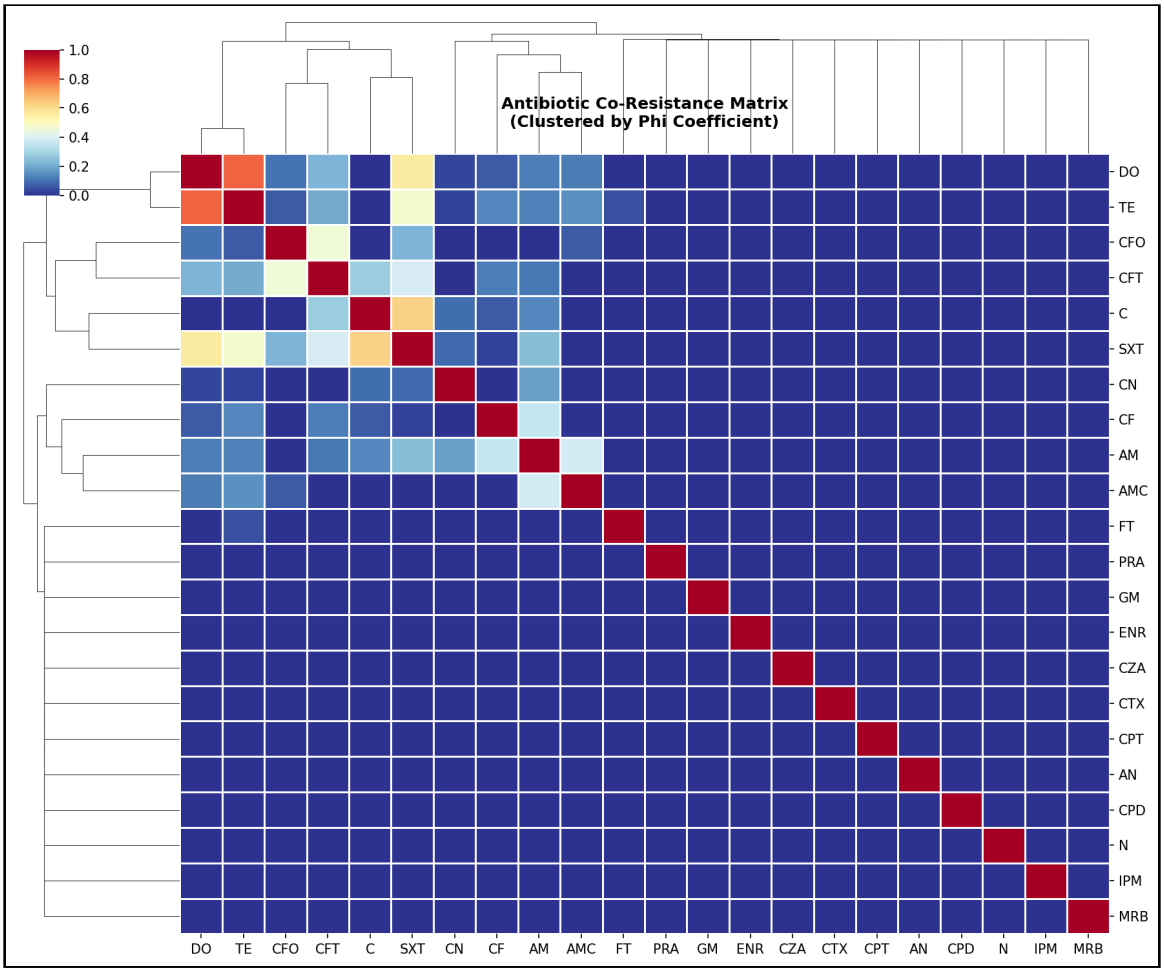

Figure 28: Clustered heatmap of antibiotic resistance correlations. Rows and columns represent antibiotics ordered by hierarchical clustering. Color intensity indicates correlation strength between resistance to antibiotic pairs. Strong positive correlations (red) identify potential co-selection targets, while negative correlations (blue) suggest inversely related resistance mechanisms.

6.5. Co-resistance Pattern Analysis

6.5.1. Phi Coefficient Analysis

To investigate the complex interactions between resistances, pairwise co-occurrence patterns of resistance profiles were analyzed. This analysis aims to uncover significant

associations that may reflect shared genetic mechanisms, co-selection pressures, or cross-resistance phenomena within the isolate population.

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.

Table 31: Top Significant Co-resistance Pairs

Antibiotic Pair	Phi Coefficient	p-value
Doxycycline – Tetracycline	0.806	< 0.001
Chloramphenicol – Trimethoprim-Sulfamethoxazole	0.621	< 0.001
Doxycycline – Trimethoprim-Sulfamethoxazole	0.559	< 0.001
Trimethoprim-Sulfamethoxazole – Tetracycline	0.470	< 0.001
Cefoxitin – Ceftiofur	0.454	< 0.001

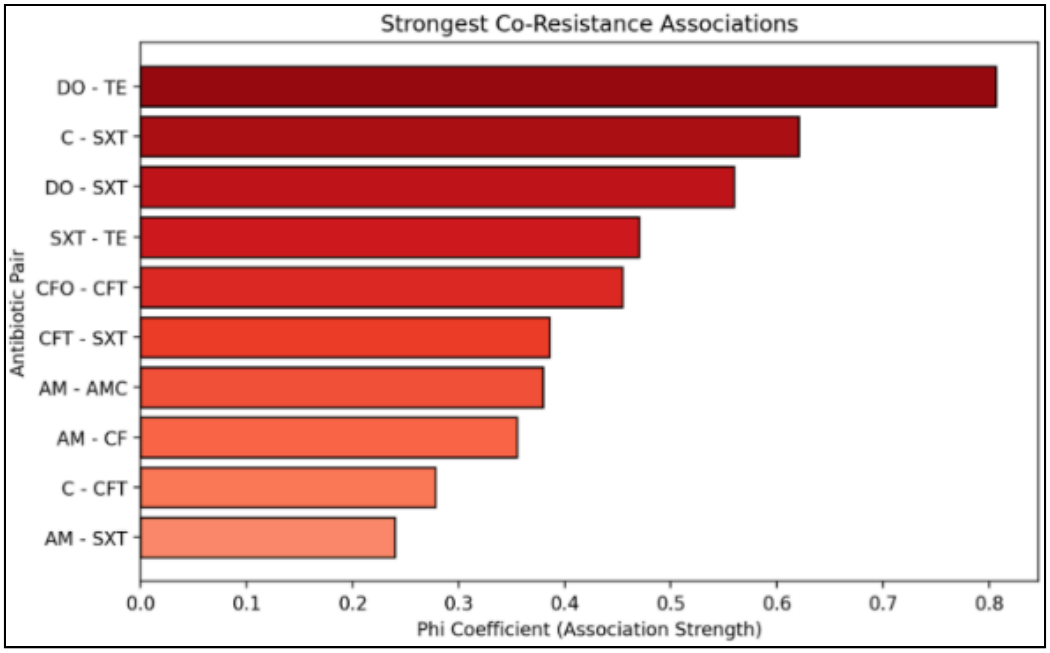

Figure 29: Top 5 Significant Co-resistance Pairs ($p < 0.001$). Doxycycline-Tetracycline shows the strongest association (Phi=0.81), followed by Sulfamethoxazole/Trimethoprim pairs. Interpreted based on Cohen's conventions.

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 [41].

6.5.2. 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 (enrofloxacin, marbofloxacin) formed a tightly connected subnetwork

The co-resistance network is visualized in Figure 30, where nodes represent antibiotics and edges indicate statistically significant co-resistance relationships.

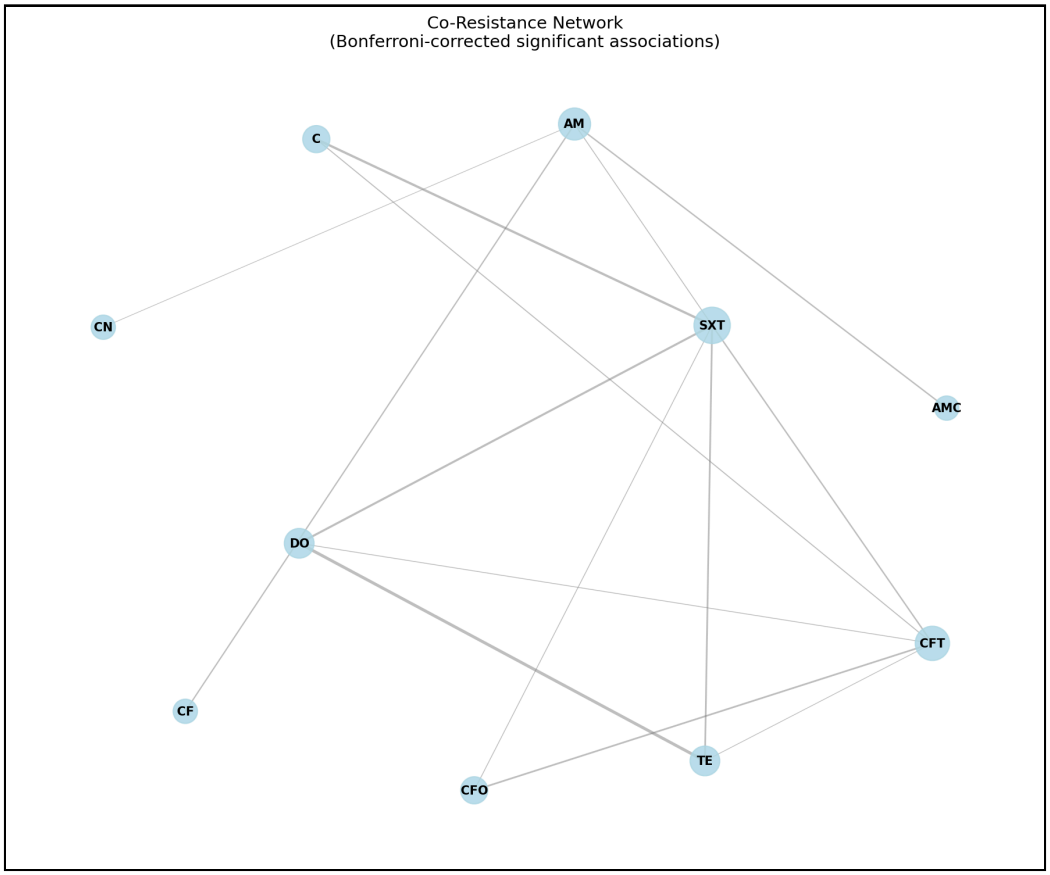

Figure 30: Co-resistance network graph. Nodes represent antibiotics; edges connect pairs with significant co-resistance (Phi > 0.3, p < 0.001). Edge thickness reflects correlation strength. Hub antibiotics (Ampicillin, Tetracycline) show high connectivity, indicating frequent co-occurrence with resistance to multiple agents. The tight clustering of fluoroquinolones suggests shared efflux-mediated resistance mechanisms.

6.5.3. 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 [42].

6.6. Discussion of Results

6.6.1. 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 [43].

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 [44].

6.6.2. 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 [24].

The high macro F1-Score (0.96) 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.

6.6.3. Comparison with Parent Project Data

This study builds upon the extensive surveillance data collected by the parent project (INOHAC Project 2), shifting the analytical focus from descriptive statistics to multivariate pattern recognition. Table 32 details the similarities and key methodological advancements distinguishing this thesis from the primary surveillance reports.

Table 32: Comparative analysis between Parent Project surveillance data and Thesis Clustering results

Comparison Aspect	Parent Project (Surveillance)	Current Study (Clustering)	Synthesis
Scope & Population	Surveillance of >1,300 presumptive isolates across BARMM & other regions.	Analytical subset of 491 confirmed isolates with complete profiles.	Focuses on high-integrity data to ensure robust pattern recognition, filtering surveillance noise.
Methodology	Univariate analysis: prevalence rates & species-specific MDR counts.	Multivariate clustering (HAC) & Random Forest (RF) validation.	Parent project identifies where resistance exists; this study explains how resistance traits cluster.
MDR Findings	Identified BARMM & hospitals as MDR hotspots (MDR vs Non-MDR).	Defined 'Cluster 3' (MDR Archetype) linked to BARMM (53.7%) & tetracycline.	Validates geographical risks by defining the specific antibiotic signature (TE-DO-beta-lactam).
Species vs. Phenotype	Analyzed resistance by species (separate tables).	Cluster 3 (MDR) spans multiple species; Cluster 1 is species-specific (Salmonella).	Demonstrates MDR as a convergent phenotype across species in high-risk environments.

The integration of data from the parent project [7] provides the necessary volume to detect these patterns, while the clustering approach elucidates the underlying structure of resistance that descriptive counts alone cannot reveal. Specifically, the "MDR Archetype" (Cluster 3) unifies the high MDR counts observed in BARMM E. coli and K. pneumoniae into a single, trackable phenotypic entity.

6.6.4. 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.

6.7. 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 99.0% test set accuracy (macro F1 = 0.96), 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 7

CONCLUSION AND RECOMMENDATION

7.1. Conclusion

This study 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:

7.1.1. 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.

7.1.2. 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 99.0% test set accuracy (macro F1 = 0.96), confirming that cluster assignments represent reproducible, learnable patterns rather than algorithmic artifacts [24].

7.1.3. 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].

7.1.4. 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 [41]. These patterns have direct implications for empirical therapy selection and resistance prediction.

7.1.5. 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.

7.2. Recommendations

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

7.2.1. For Public Health Authorities

1. Prioritization of 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 warrant prioritization 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.

7.2.2. 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.

7.2.3. For Surveillance Programs

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

2. Leverage Machine Learning: The demonstrated 99.0% 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.

7.2.4. For Aquaculture Management

1. Reduction of 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 [45].
2. Water Quality Monitoring: The water-associated *Salmonella* cluster (C1) suggests that water quality improvements could reduce environmental resistance transmission.

7.3. Future Research Directions

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

7.3.1. 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.

7.3.2. 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 [26].

7.3.3. 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.

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

[2] The Review on Antimicrobial Resistance, "Tackling Drug-Resistant Infections Globally: Final Report and Recommendations." [Online]. Available: https://amr-review.org/sites/default/files/160525_Final%20paper_with%20cover.pdf

[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] 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.

[10] 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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[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] R. Gouareb, A. Bornet, D. Proios, S. G. Pereira, and D. Teodoro, "Detection of Patients at Risk of Multidrug-Resistant Enterobacteriaceae Infection Using Graph Neural Networks: A Retrospective Study," *Health Data Science*, vol. 3, 2023, doi: 10.34133/hds.0099.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[31] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[32] Centers for Disease Control and Prevention, "One Health." [Online]. Available: https://www.cdc.gov/one-health/index.html

[33] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in Data Mining: Formulation, Detection, and Avoidance," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 4, pp. 1–21, 2012, doi: 10.1145/2382577.2382579.

[34] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946.

[35] J. Podani, "Extending Gower's general coefficient of similarity to ordinal characters," *Taxon*, vol. 48, no. 2, pp. 331–340, 1999.

[36] 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.

[37] 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.

[38] 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

[39] 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.

[40] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, 1988.

[41] 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.

[42] 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.

[43] W. Zhao, P. Sun, W. Li, and L. Shang, "Machine Learning-Based Prediction Model for Multidrug-Resistant Organisms Infections: Performance Evaluation and Inter-

pretability Analysis," *Infection and Drug Resistance*, vol. 18, pp. 2255–2269, 2025, doi: 10.2147/idr.s459830.

[44] 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.

[45] 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.

APPENDICES

Appendix A. Supplementary Figures

This appendix contains additional figures supporting the sensitivity analysis and detailed cluster profiles discussed in the methodology and results sections.

Sensitivity Analysis: k=5 and k=6

The silhouette plots below illustrate the cluster validity metrics for alternative k-values (k=5 and k=6), which were compared against the selected optimal k=4 solution.

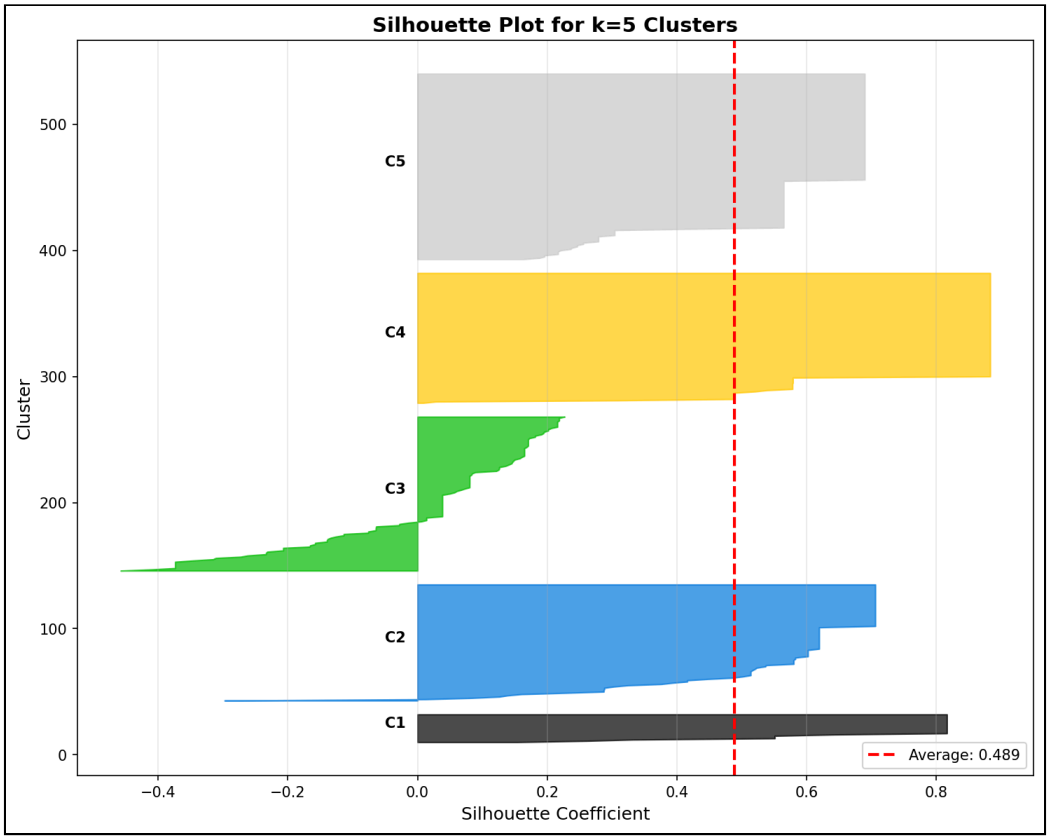

Figure 31: Silhouette Analysis for k=5. Cluster cohesion remains high, but separation decreases compared to k=4.

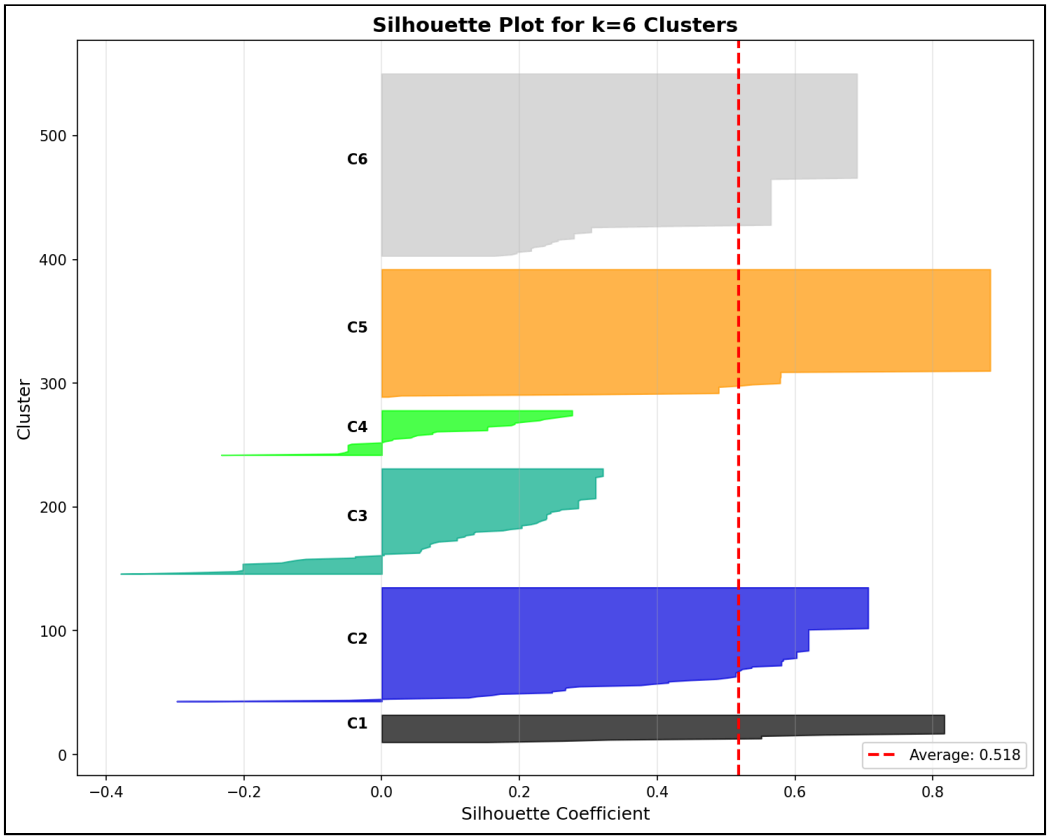

Figure 32: Silhouette Analysis for k=6. The emergence of smaller, less distinct clusters indicates over-segmentation.