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

RESULTS AND DISCUSSION

2.1. 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].

2.2. Unsupervised Learning Results

2.2.1. 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.

2.2.1.1. Ward's Linkage Method

Ward's minimum variance method was employed as the linkage criterion [9]. 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 1: 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 1, 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.

2.2.1.2. 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 2: 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 2, 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.

2.2.2. 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 [10].

Table 3: 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 3, 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 [11], [12]. 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 4: 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 4 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.

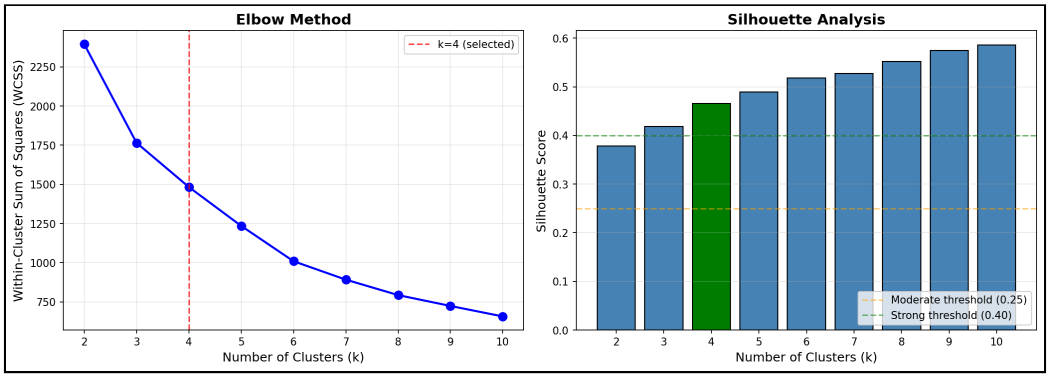

Figure 1: Elbow method (left) and silhouette analysis (right) for cluster validation. The WCSS curve shows the elbow point at k=4, while the silhouette plot confirms moderate-to-strong structure at this configuration.

2.2.3. Cluster Characteristics

The four identified clusters exhibited distinct resistance phenotype profiles:

Table 5: 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 5, 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.

2.2.3.1. 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.

2.2.3.2. 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.

2.2.3.3. 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 [13]—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.

2.2.3.4. 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.

2.2.4. Visualizations of Cluster Structure

The cluster resistance profiles, hierarchical structure, and geographic distribution are visualized in the following figures. Figure 2 presents the mean resistance scores for each cluster across all antibiotics, revealing distinct phenotypic signatures.

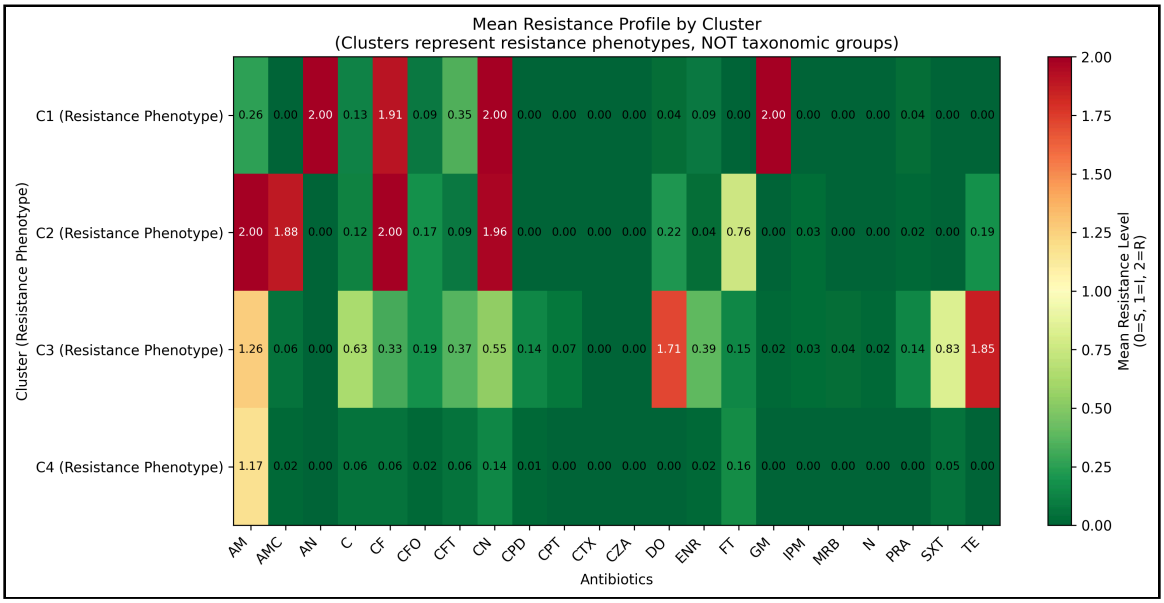

Figure 2: Cluster resistance profiles showing mean resistance scores (0–2 scale) per antibiotic for each of the four clusters. C1 (Salmonella-Aminoglycoside) shows elevated aminoglycoside resistance; C2 (Enterobacter-Penicillin) exhibits β-lactam resistance; C3 (MDR Archetype) displays broad resistance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

The hierarchical structure of the clustering solution is visualized in Figure 3, which links the dendrogram with a resistance heatmap to show the relationship between isolate groupings and their resistance patterns.

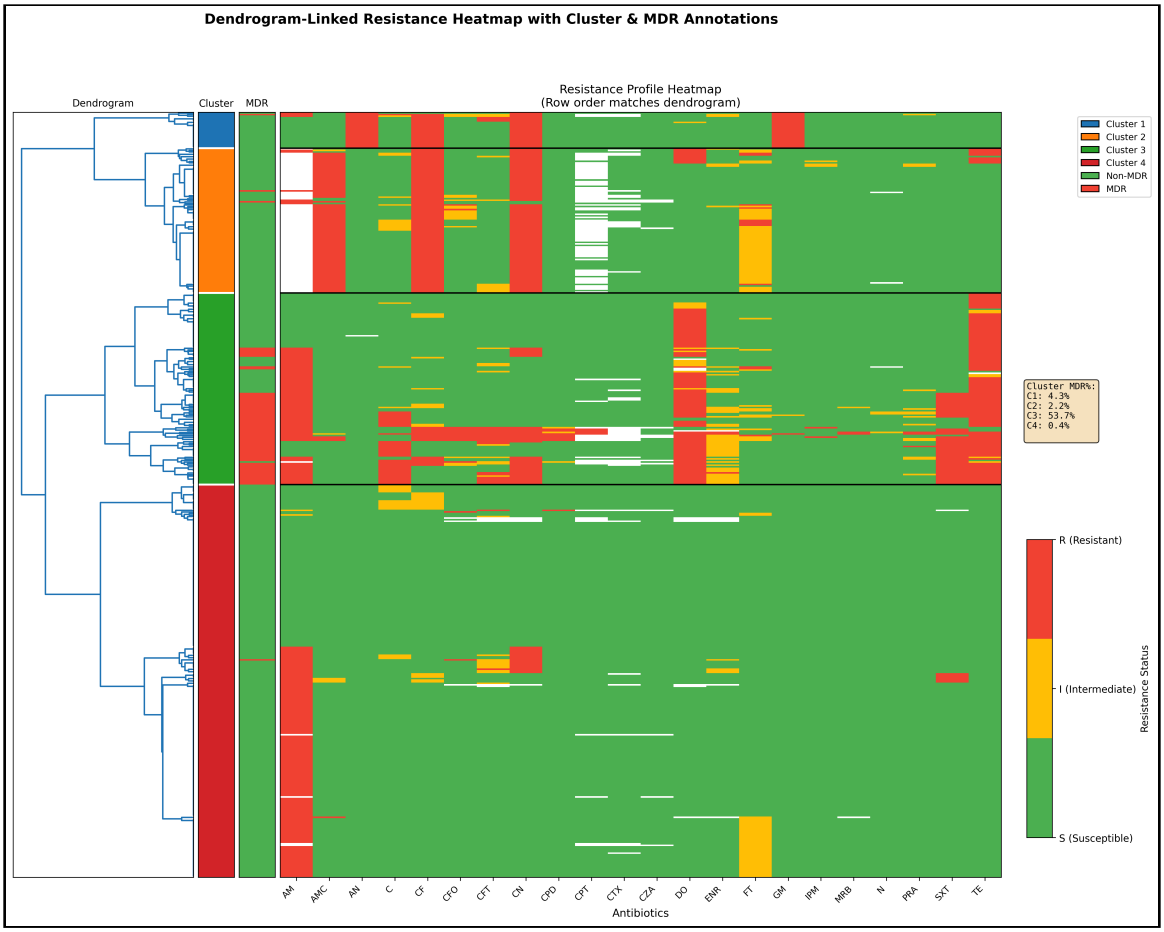

Figure 3: Dendrogram-linked resistance heatmap showing hierarchical clustering structure. Rows represent isolates ordered by dendrogram position; columns represent antibiotics. Color intensity indicates resistance level (blue=susceptible, red=resistant). The four main clusters are visible as distinct blocks with characteristic resistance patterns.

The detailed dendrogram in Figure 4 illustrates the complete hierarchical structure with cluster assignments.

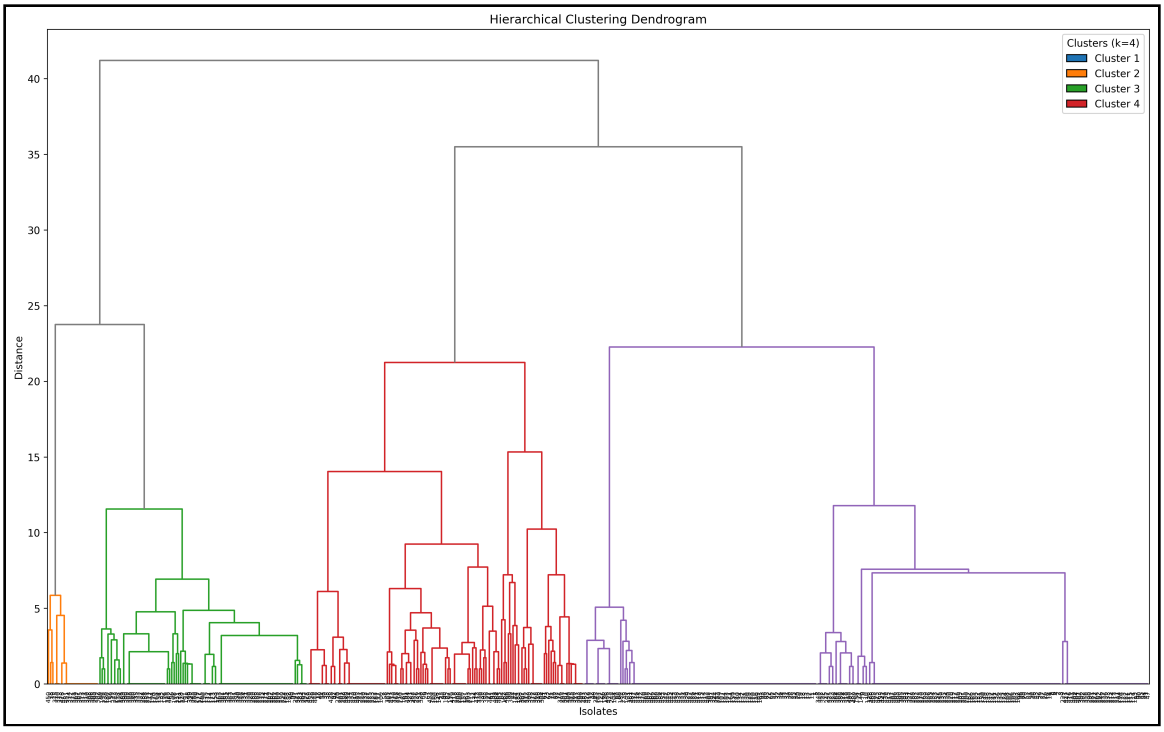

Figure 4: High-resolution dendrogram showing hierarchical agglomerative clustering of 491 isolates using Ward's linkage. The horizontal dashed line indicates the cut point for k=4 clusters. Distinct color branches represent the four identified phenotype clusters.

Geographic and environmental distributions of cluster assignments are shown in Figure 5 and Figure 6 respectively.

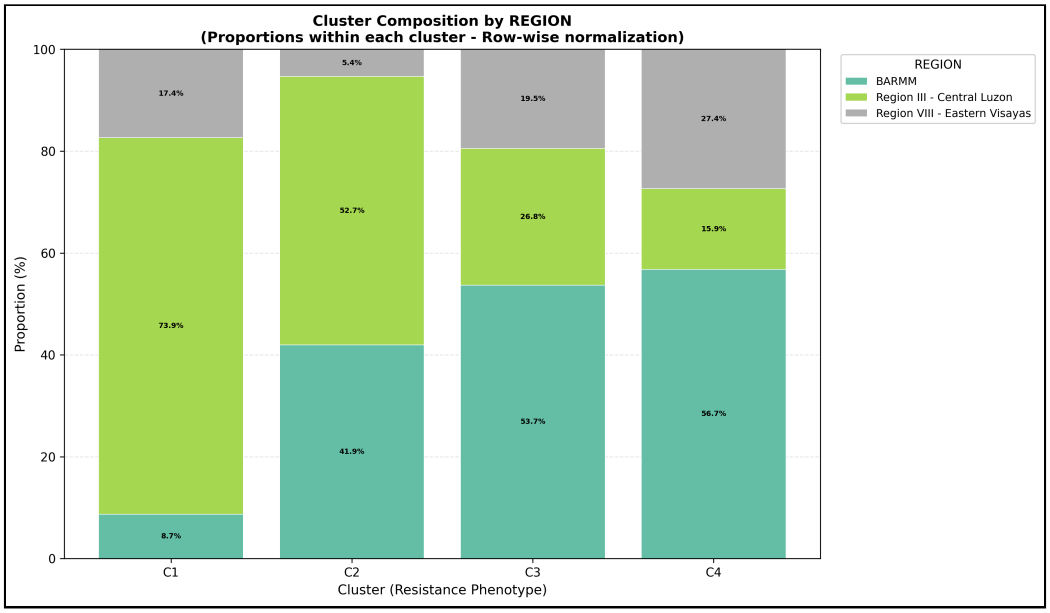

Figure 5: Cluster composition by geographic region. The stacked bar chart shows the proportion of each cluster originating from BARMM, Region III (Central Luzon), and Region VIII (Eastern Visayas). C3 (MDR Archetype) shows a strong association with BARMM, while C1 (Salmonella-Aminoglycoside) is predominantly from Region III.

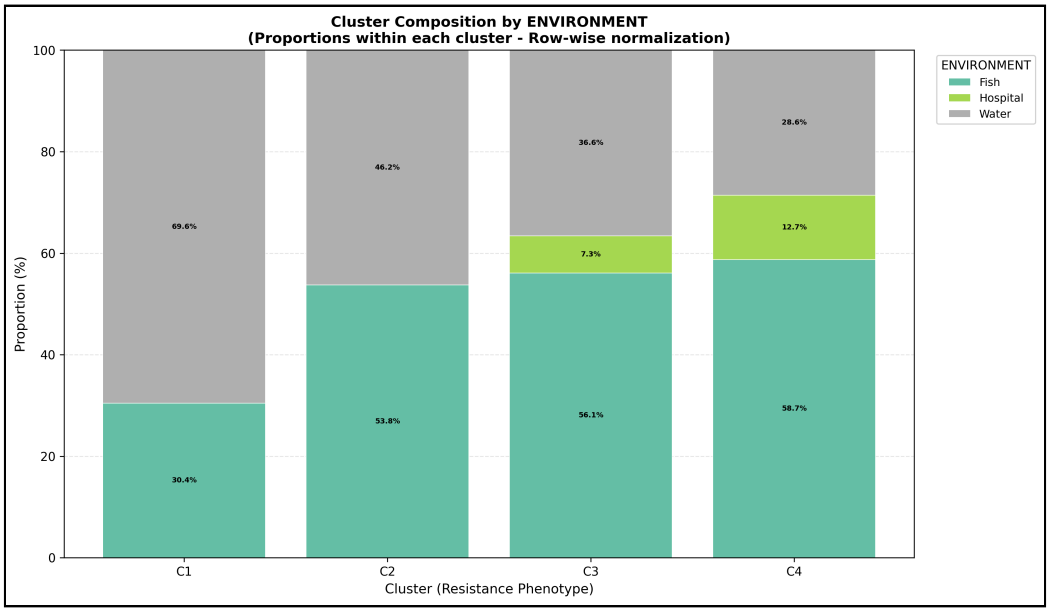

Figure 6: Cluster composition by environmental source. The stacked bar chart shows the distribution of fish, water, and hospital-derived isolates across clusters. C3 (MDR Archetype) contains the majority of hospital-environment isolates, while C4 (Susceptible Majority) is predominantly from aquatic environments.

The resistance heatmap in Figure 7 provides a comprehensive view of resistance patterns across all isolates and antibiotics.

Figure 7: 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.

2.3. 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.

2.3.1. Random Forest Classification

A Random Forest classifier was trained to predict cluster membership using the 22-dimensional encoded resistance data as input features [14]. The model was evaluated on a held-out test set (20%) after stratified splitting to ensure robust performance estimates across all four clusters.

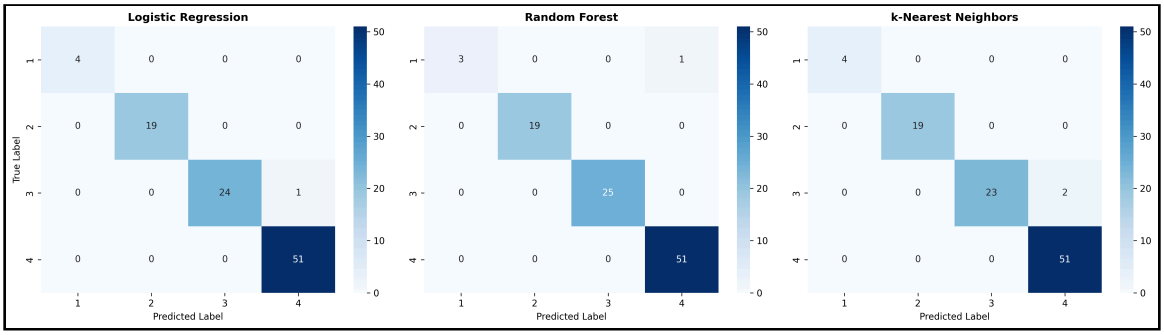

Figure 8: Confusion Matrices for Supervised Classifiers. Comparison of confusion matrices for Logistic Regression, Random Forest, and k-Nearest Neighbors. The diagonal dominance across all models confirms robust cluster separability.

Table 6 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 6: 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).

2.3.2. Feature Importance

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

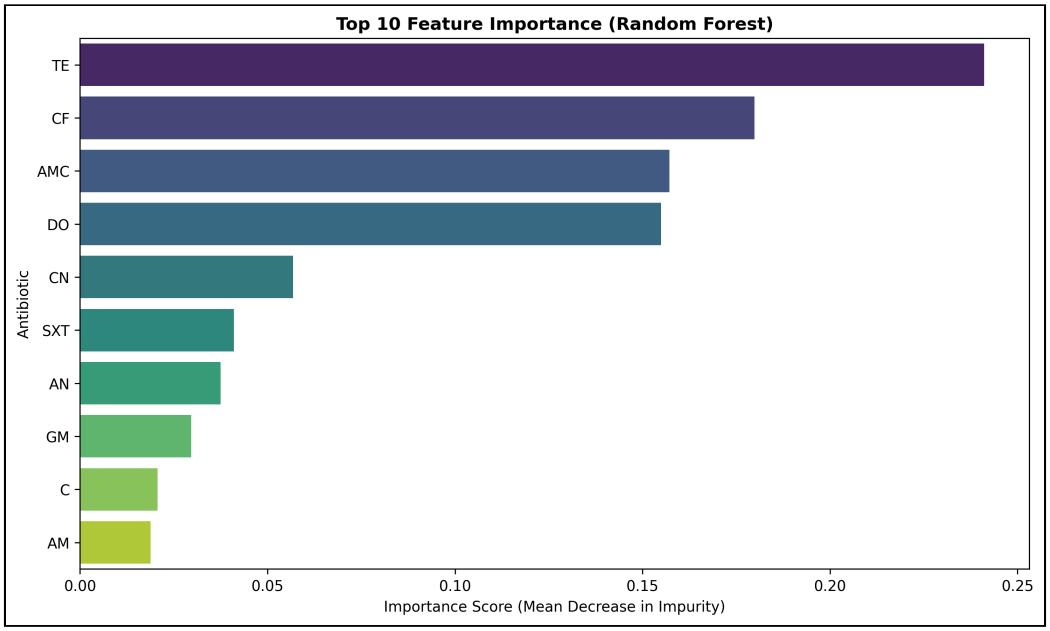

Figure 9: Feature Importance for Random Forest Classifier. Tetracycline (TE) and Doxycycline (DO) are the most discriminatory features, driving the separation of the MDR Archetype cluster. Importance scores represent the mean decrease in impurity.

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.

2.3.3. 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.

2.3.3.1. Split Ratio Comparison

Table 7: 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

2.3.3.2. Cross-Validation Comparison

Table 8: 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.

2.3.4. 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.

2.3.5. 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 10 shows the variance explained by each principal component.

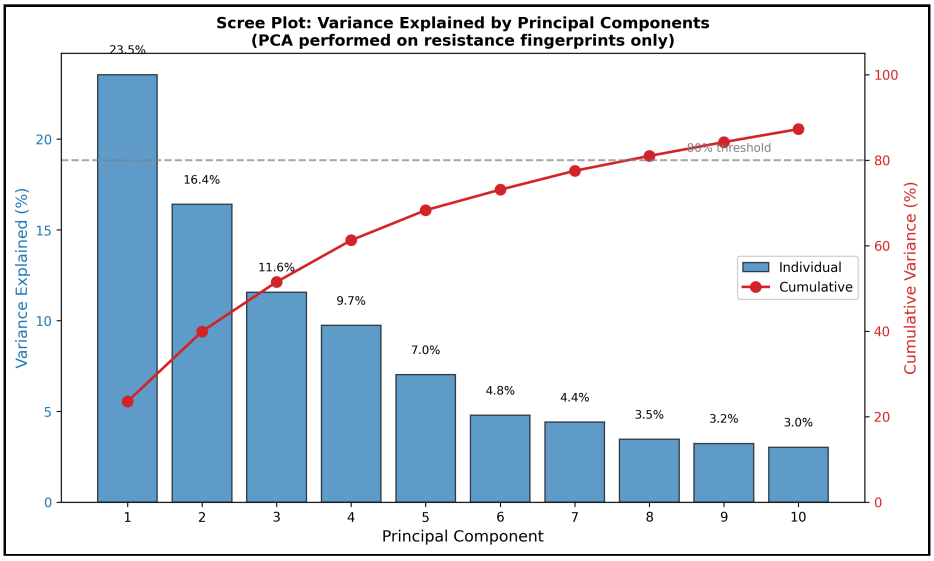

Figure 10: PCA scree plot showing cumulative variance explained. The first two principal components capture substantial variance, enabling meaningful 2D visualization of the resistance data structure.

The PCA biplot in Figure 11 reveals the contribution of individual antibiotics to the principal component space.

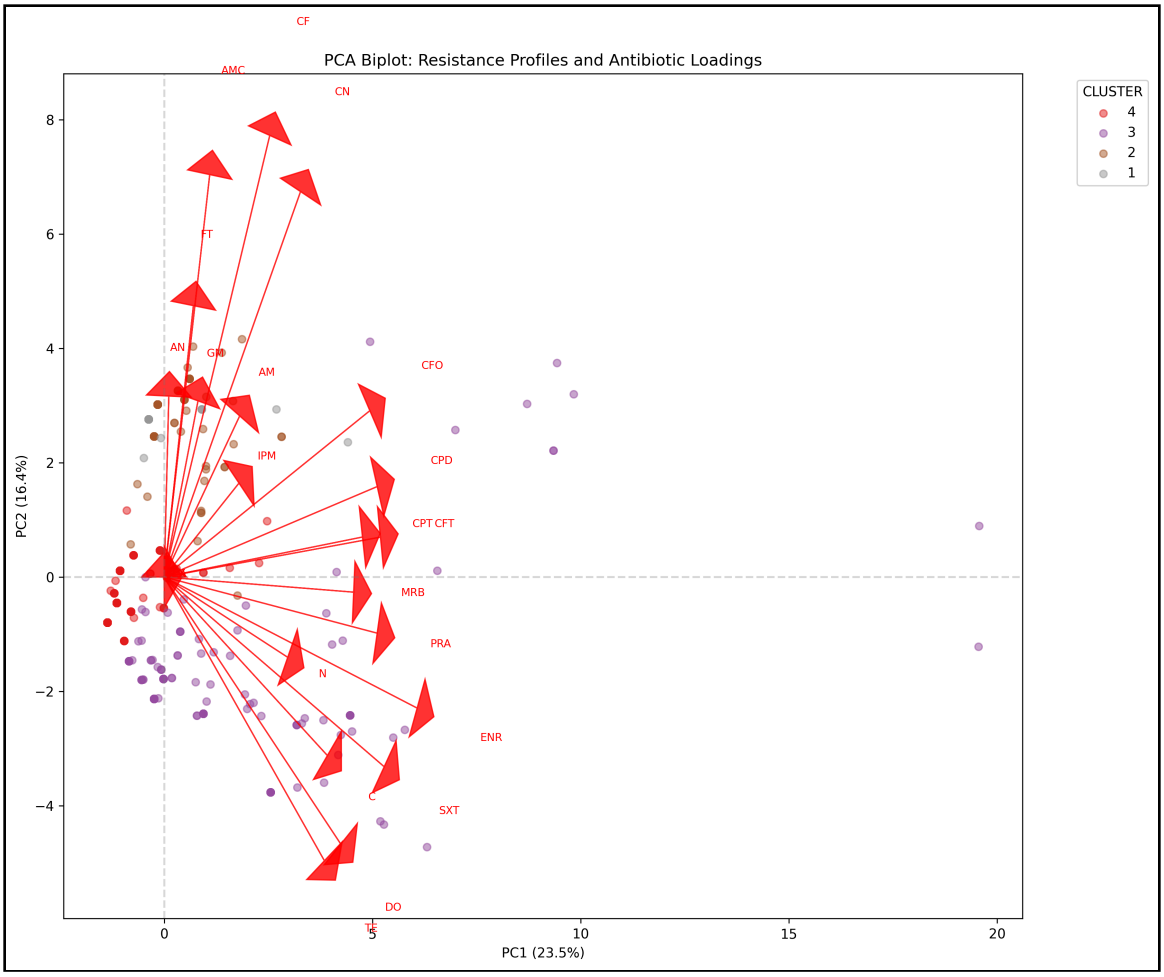

Figure 11: PCA biplot showing isolates (points) and antibiotic loadings (vectors) in the first two principal components. Vector directions indicate antibiotic contributions to cluster separation. Tetracycline-class antibiotics (TE, DO) show strong loadings consistent with their importance in defining the MDR Archetype cluster.

Figure 12 visualizes the cluster assignments in PCA space, demonstrating the separability of the four phenotype groups.

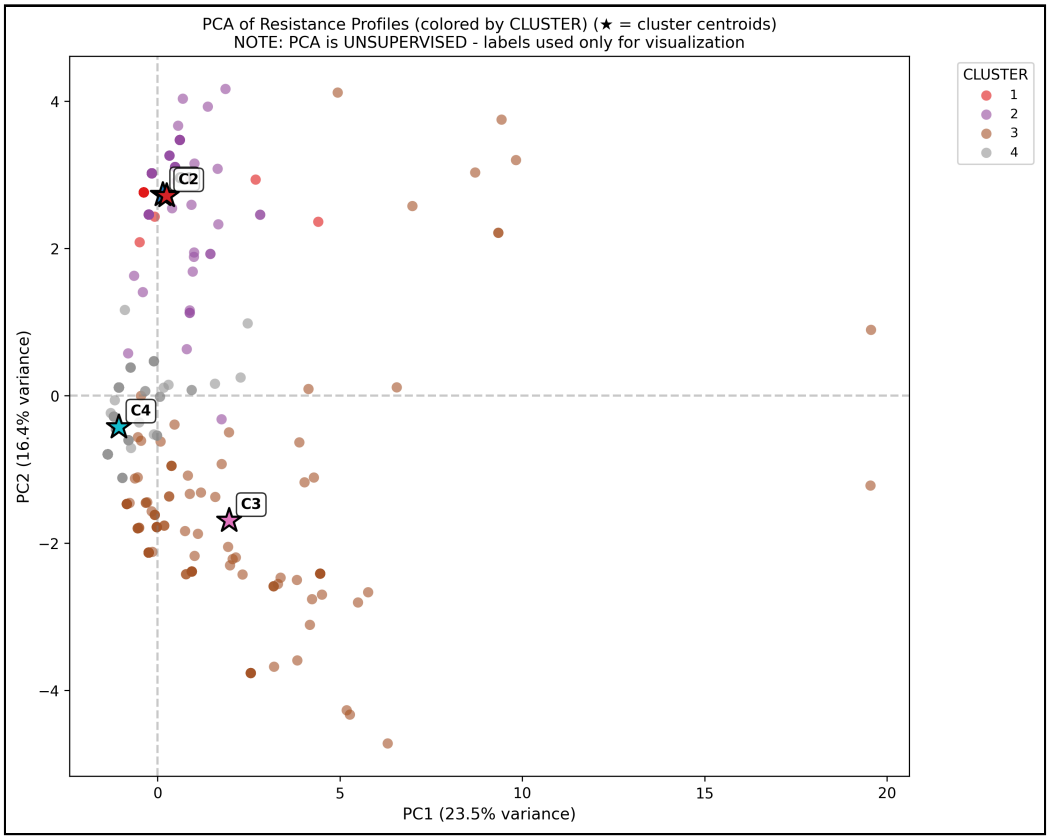

Figure 12: PCA visualization colored by cluster assignment. The four clusters show distinct spatial distributions in the reduced-dimensional space, confirming that the clustering solution captures meaningful phenotypic structure. C3 (MDR Archetype) and C4 (Susceptible Majority) form clearly separated regions.

The relationship between cluster structure and MDR status is visualized in Figure 13.

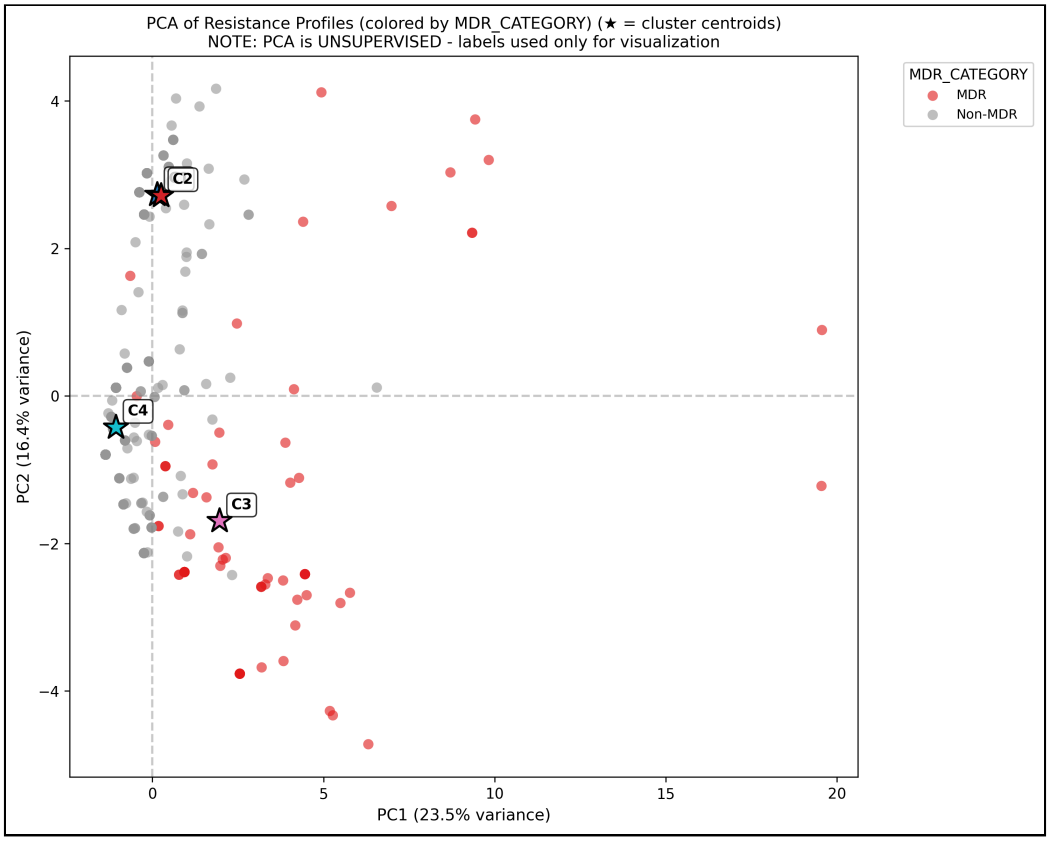

Figure 13: PCA visualization colored by MDR status. MDR isolates (red) cluster distinctly from susceptible isolates (blue), with the majority concentrated in the C3 (MDR Archetype) region of the plot. This confirms the phenotypic coherence of the MDR classification.

2.3.6. Silhouette Analysis Detail

The detailed silhouette plot for the k=4 solution is presented in Figure 14, showing cluster cohesion and separation for each isolate.

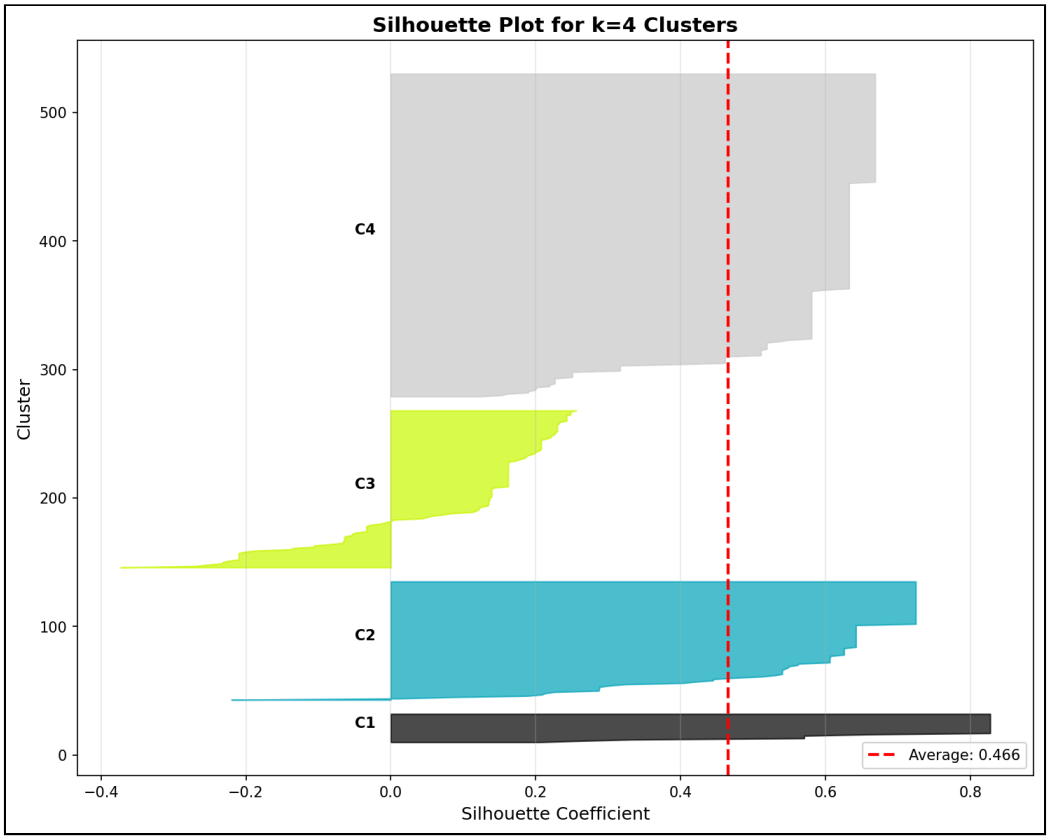

Figure 14: Silhouette plot for k=4 cluster solution. Each bar represents an isolate's silhouette coefficient; longer bars indicate better cluster fit. All four clusters show predominantly positive silhouette values, with C3 and C4 demonstrating the strongest internal cohesion (mean silhouette = 0.466).

2.4. 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.

2.4.1. 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 9 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 9: 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 [15].

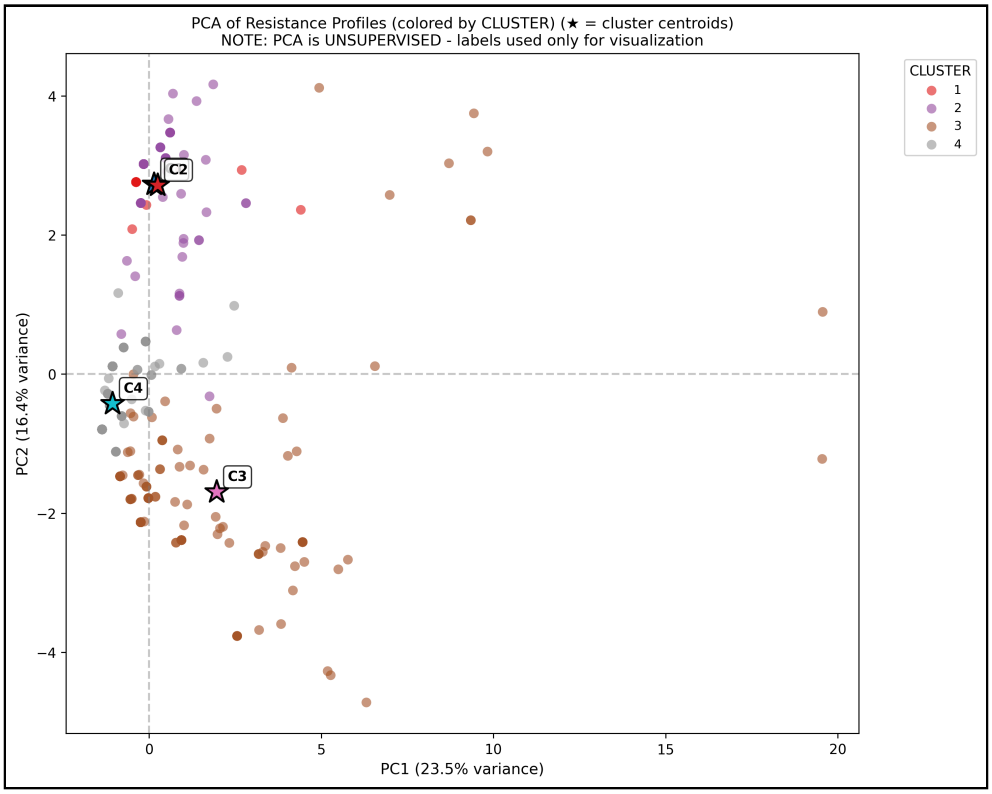

Figure 15: PCA projection of 491 isolates colored by cluster assignment. The scatter plot visualizes the separation of the four distinct resistance phenotypes along the first two principal components (PC1 and PC2).

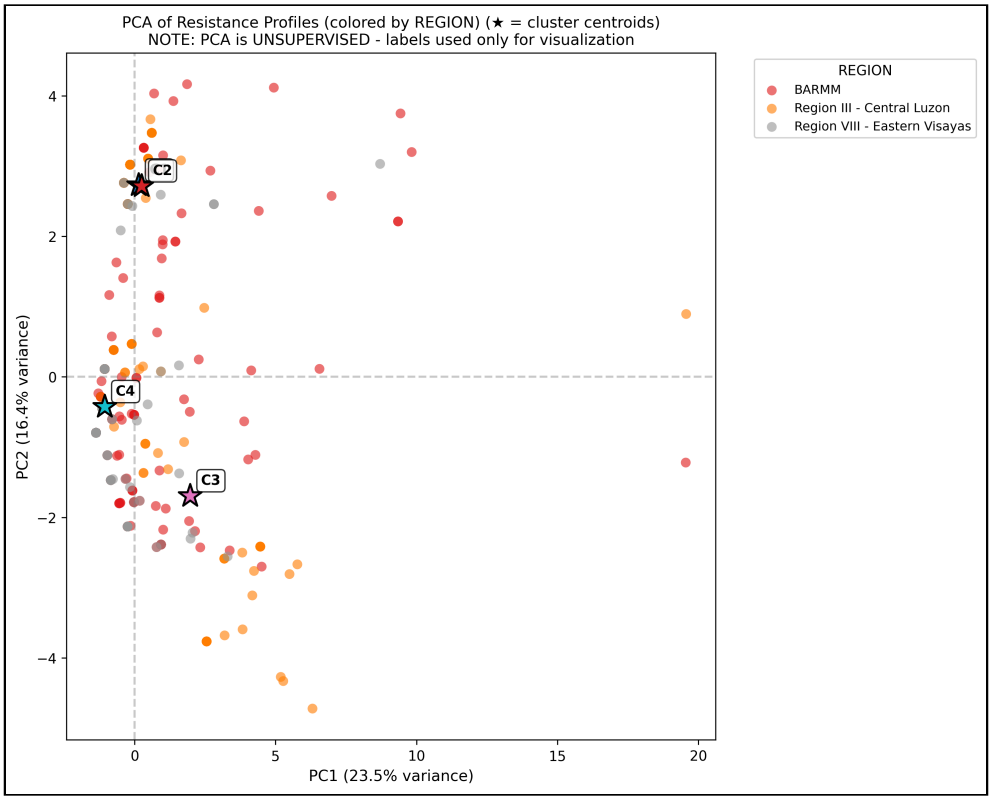

Figure 16: PCA projection colored by geographic region. Isolates from BARMM, Region III (Central Luzon), and Region VIII (Eastern Visayas) show overlapping distributions with subtle regional clustering tendencies, particularly along PC2.

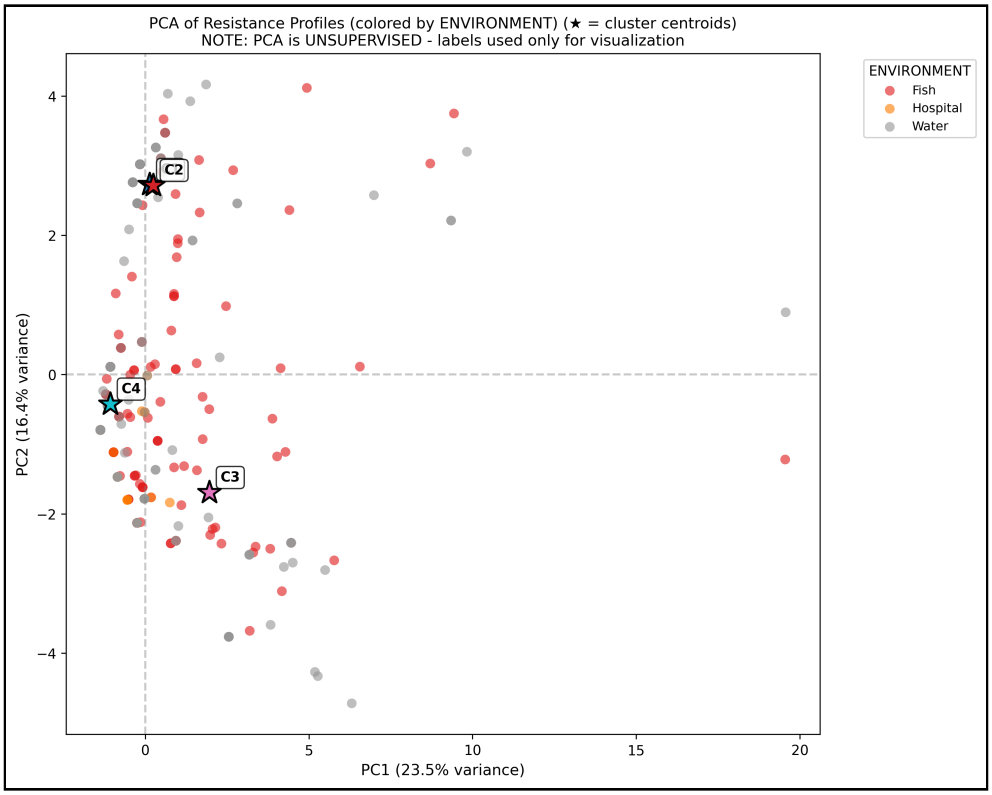

Figure 17: PCA projection colored by environmental source. Fish, water, and hospital-derived isolates are distributed across the phenotypic space, with hospital isolates showing a tendency toward the susceptible region (C4 territory) on PC1.

2.4.2. Regional Distribution Patterns

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

Table 10: 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.

2.4.3. Environmental Niche Associations

Table 11: 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].

2.4.4. 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 18 shows the MAR index distribution across all isolates, while Figure 19 illustrates MDR prevalence.

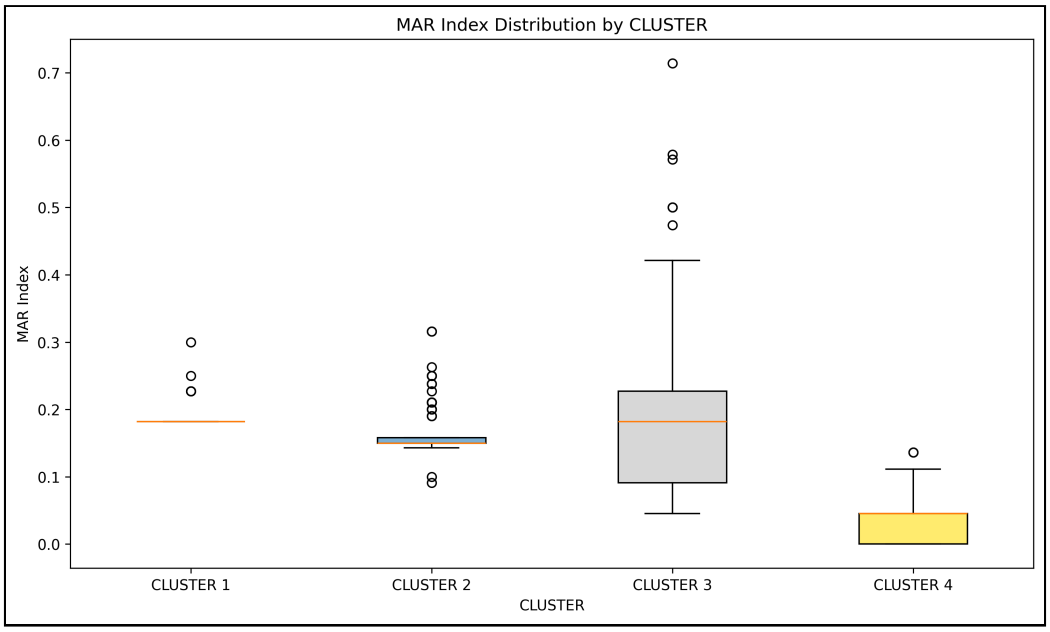

Figure 18: Distribution of Multiple Antibiotic Resistance (MAR) index across 491 isolates. The histogram shows the proportion of antibiotics to which each isolate is resistant. Most isolates exhibit low MAR values (< 0.2), with a distinct high-MAR subpopulation corresponding to the MDR Archetype cluster (C3).

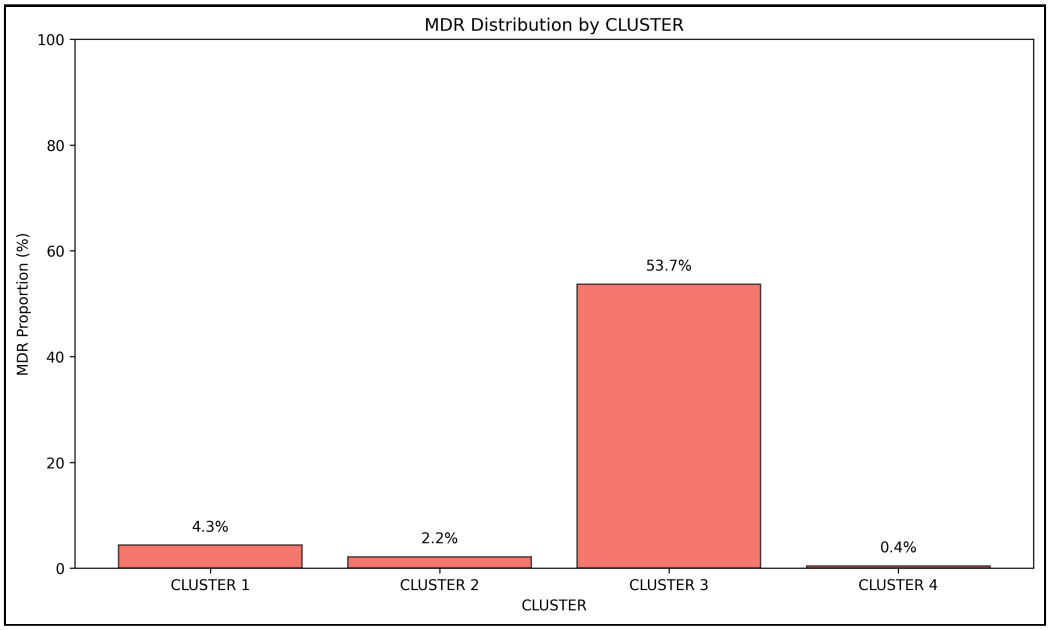

Figure 19: Multi-Drug Resistance (MDR) status distribution across clusters. The bar chart shows the proportion of MDR (≥3 resistant classes) and non-MDR isolates within each cluster. C3 (MDR Archetype) contains 53.7% MDR isolates, while other clusters exhibit less than 5% MDR prevalence.

2.4.5. Antibiotic Clustering Analysis

Hierarchical clustering was also applied to antibiotics to identify groups with correlated resistance patterns. Figure 20 presents the antibiotic dendrogram, revealing clusters of antibiotics that tend to co-occur in resistance profiles.

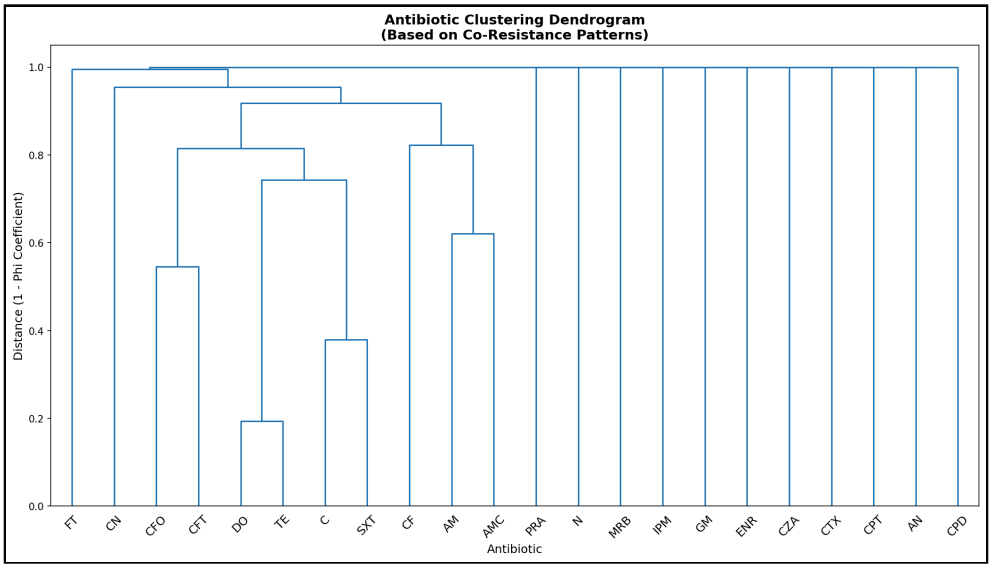

Figure 20: Dendrogram of antibiotic clustering based on resistance co-occurrence patterns. Antibiotics that cluster together exhibit similar resistance profiles across isolates. The tight grouping of tetracycline-class antibiotics (TE, DO) and fluoroquinolones (ENR, MRB, PRA) reflects mechanistically related resistance mechanisms.

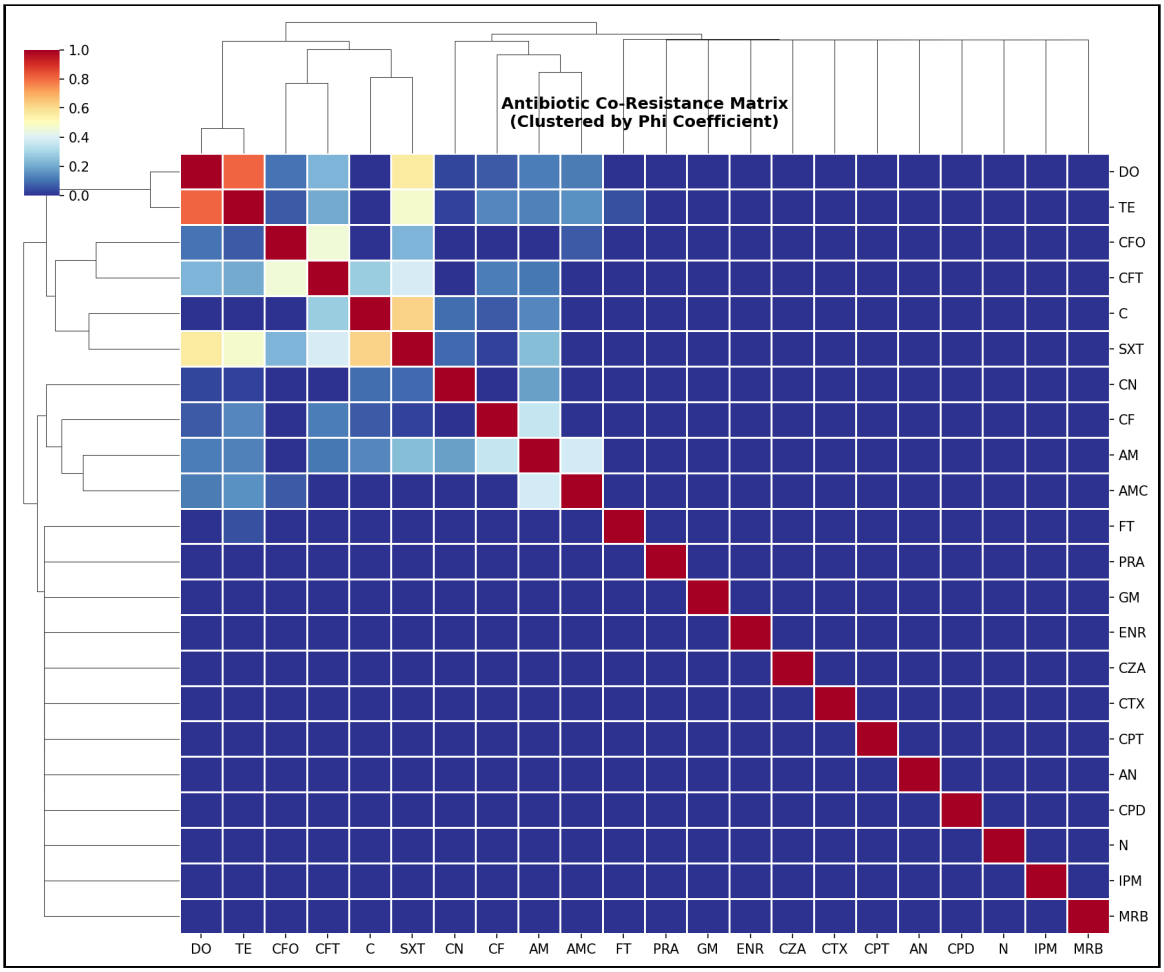

Figure 21: Clustered heatmap of antibiotic resistance correlations. Rows and columns represent antibiotics ordered by hierarchical clustering. Color intensity indicates correlation strength between resistance to antibiotic pairs. Strong positive correlations (red) identify potential co-selection targets, while negative correlations (blue) suggest inversely related resistance mechanisms.

2.5. Co-resistance Pattern Analysis

2.5.1. 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 [16]. Pairs exhibiting Phi > 0.3 and $p < 0.001$ were considered statistically significant co-resistance associations.

Table 12: 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

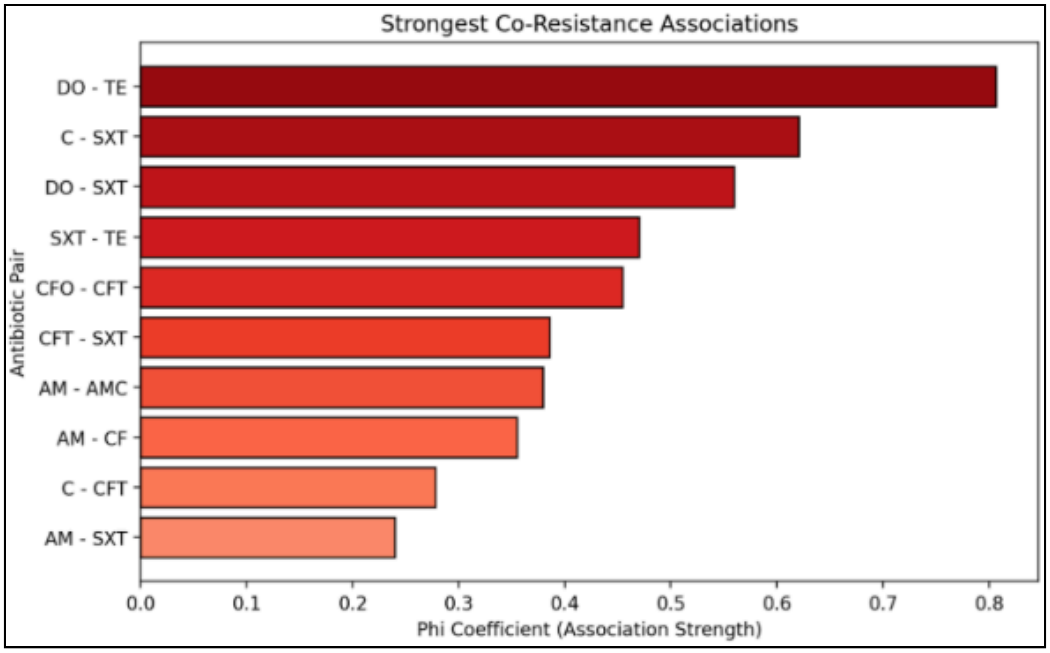

Figure 22: Top 5 Significant Co-resistance Pairs ($p < 0.001$). Doxycycline-Tetracycline shows the strongest association (Phi=0.81), followed by Sulfamethoxazole/Trimethoprim pairs. Interpreted based on Cohen's conventions.

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 [17].

2.5.2. 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 23, where nodes represent antibiotics and edges indicate statistically significant co-resistance relationships.

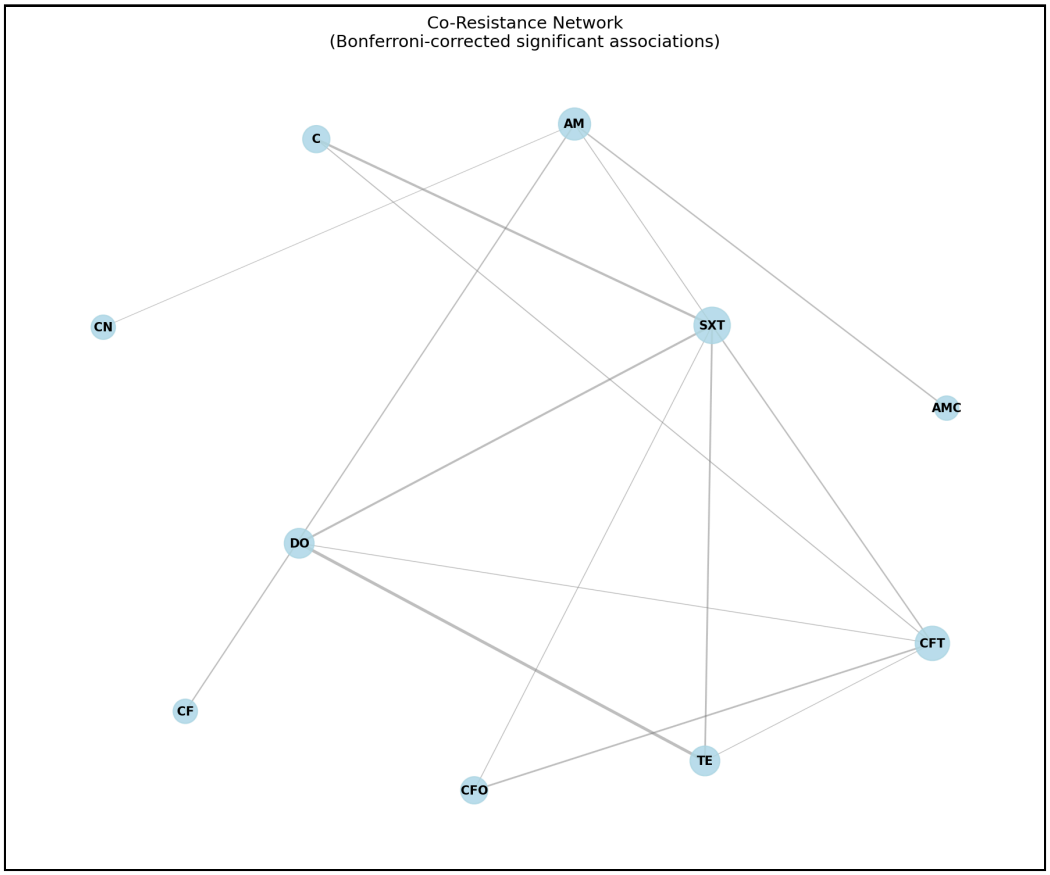

Figure 23: Co-resistance network graph. Nodes represent antibiotics; edges connect pairs with significant co-resistance (Phi > 0.3, p < 0.001). Edge thickness reflects correlation strength. Hub antibiotics (Ampicillin, Tetracycline) show high connectivity, indicating frequent co-occurrence with resistance to multiple agents. The tight clustering of fluoroquinolones suggests shared efflux-mediated resistance mechanisms.

2.5.3. 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 [18].

2.6. Discussion of Results

2.6.1. 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 [19].

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 [20].

2.6.2. 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 [14].

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.

2.6.3. 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 13 details the similarities and key methodological advancements distinguishing this thesis from the primary surveillance reports.

Table 13: 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.

2.6.4. 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.

2.7. 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].

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] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.