Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 9, No. 6, December, 2025

Feature Transformation on Big Data for Species Classification in Machine Learning

Li Wen Yow [1], Lee Yeng Ong [1, 2], Joon Liang Tan [1, 3*]

[1] *Faculty of Information Science and Technology, Multimedia University, Melaka 75450, Malaysia.*

[2] *Center for Advanced Analytics, CoE for Artificial Intelligence Multimedia University, Melaka 75450, Malaysia.*

[3] *Center for Intelligent Cloud Computing, CoE for Advanced Cloud, Multimedia University, Melaka 75450, Malaysia.*

Abstract

Classification of bacterial species, particularly for closely related taxa, remains a major challenge in many areas, e.g., public health, food industries, and many others. The issues are mainly caused by overlapping genetic features of organisms and data complexities. In this study, a bacterial taxonomic identification framework that integrates genome-derived motif sequences with machine learning was introduced. Two hundred and forty genome sequences from *Salmonella enterica*, representing six subspecies and ten serovars, were used for modelling. Sequence motifs were predicted from single-copy orthologous core genes of the downloaded genomes. Single nucleotide polymorphisms (SNPs) within these motifs were extracted and numerically encoded as machine learning features. The 20 top-most informative predictors from feature selections were used for model training in Random Forest and Support Vector Machine. Comparing the output from multiple analyses, the Random Forest model achieved the highest accuracy of 97.92%, demonstrating reliable differentiation of *Salmonella* at both subspecies and serovar levels. This research presents two key innovations: i) the use of sequence motifs as molecular signatures for bacterial classification; ii) a novel feature engineering method that transforms genome-derived data into machine learning-readable features. The proposed framework offers a practical and scalable solution for fine-level bacterial classification and has high potential to be applied for other microbial taxa.

Keywords:

Big Data;
Bioinformatics;
Feature Selection;
Machine Learning;
Sequence Motifs.

Article History:

Received:	23	May	2025
Revised:	21	November	2025
Accepted:	28	November	2025
Published:	01	December	2025

1- Introduction

The exponential growth of data has introduced new challenges in resolving bacterial taxonomy, revealing complexities in bacterial relationships that challenge the traditional classification frameworks. The early classification systems relied on phenotypic characteristics and biochemical tests as fundamental approaches [1, 2]. Although the methods are useful for basic differentiation, they lack the resolution to distinguish genetically similar bacterial strains. Advances in molecular biology later introduced targeted DNA analysis methods such as 16S ribosomal RNA (rRNA) sequencing [3-7] and Multi-Locus Sequence Typing (MLST) [8-10], which significantly improved taxonomic accuracy at broader levels. However, these techniques still fall short of resolving fine-scale relationships, such as those at the subspecies or serovar level [3, 4, 8, 11-13]. The increasing number of reports on taxonomic misclassification and proposals on taxonomic revision indicate the limitations in conventional classification systems, particularly in distinguishing closely related bacterial strains that may exhibit distinct pathogenic or ecological properties [14-17]. The limitations have driven an urgent need for more sophisticated classification approaches in clinical and epidemiological contexts. Whole genome sequencing (WGS), that is, achieving the genome of organisms, marked a pivotal advancement in microbial genomics [18-20]. While WGS provides unprecedented detail for taxonomic analysis, the volume, size, and complexity of genomic data present significant computational challenges [21, 22]. Moreover, contemporary methods

* **CONTACT**: jltan@mmu.edu.my

DOI: http://dx.doi.org/10.28991/ESJ-2025-09-06-09

often struggle to distinguish meaningful taxonomic signals from genomic noises [23, 24], resulting in ambiguous relationships among closely related species. However, the risks can be reduced with proper classification modeling and with the use of features sampled from genomic data.

Machine learning has emerged as a tool to address the stated challenges in genomics. By leveraging algorithms that can learn from data, machine learning methods can analyze large genomic datasets, identify patterns, and improve classification accuracy. These algorithms can manage high-dimensional data and extract meaningful features from complex genome sequences [25-28], making the data well-suited for modern microbial taxonomy. Despite its potential, most current applications of machine learning in bacterial classification focus either on entire genomes or on generic features like k-mers, 16S, or MLST, which may not fully capture the evolutionary signals for high-resolution taxonomy classification. In this context, we developed a classification framework using *Salmonella enterica* as model data. *Salmonella enterica* was selected for its complex taxonomic hierarchies and its impact on the current public health concerns [17, 29]. The genus encompasses numerous serovars with distinct pathogenic profiles, presenting a fine-scale classification challenge that available methods struggle to resolve [30-32].

Our framework introduces a biologically grounded approach by extracting conserved sequence motifs from single-copy core orthologs as molecular markers. Single nucleotide polymorphisms (SNPs) within the predicted motifs were identified and transformed into machine-readable formats through numerical encoding. These motif-specific SNP patterns serve as features for model training. This approach enables the classification of closely related bacterial strains using a smaller yet informative subset of genomic data. By focusing on evolutionarily and functionally relevant regions, the method balances interpretability and accuracy. The framework design and principles are expandable across different bacterial genera, offering a solution for modern taxonomic research. It contributes to microbial taxonomy by converting complex genomic information into structured, interpretable features suitable for machine learning applications.

2- Methodology

The research framework implemented in this study consisted of two major phases: bioinformatics processing and machine learning implementation (Figure 1). This workflow enables systematic analysis from raw genomic data to final classification.

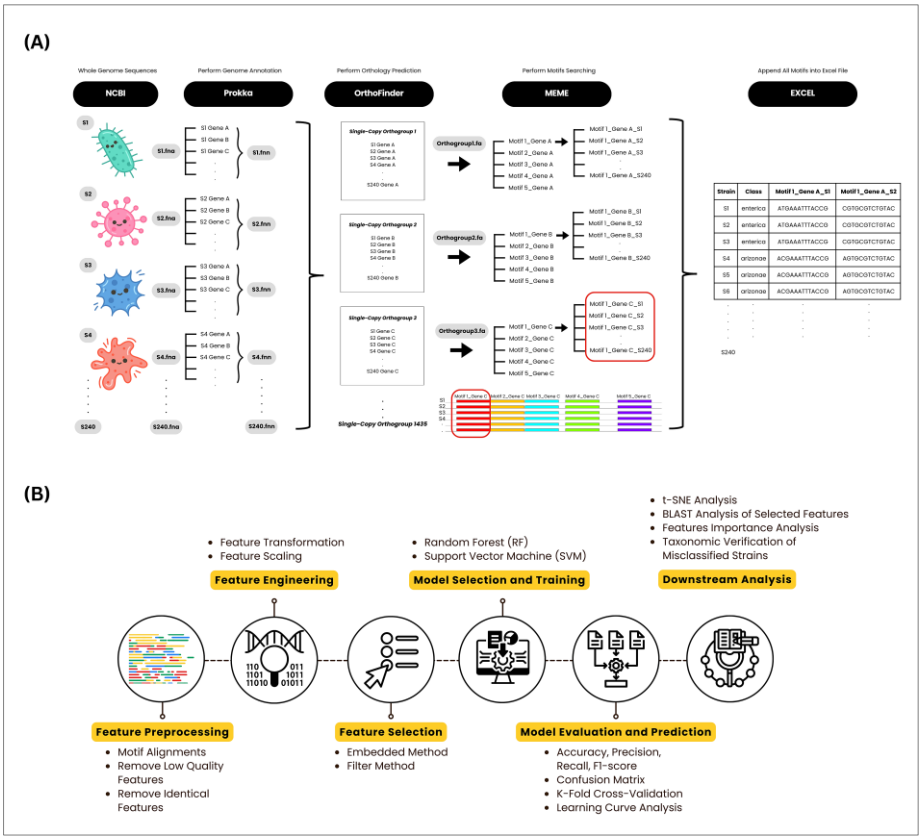

Figure 1. Workflow for bacteria species classification integrating bioinformatics processing and machine learning analysis. (A) Bioinformatics workflow for whole genome sequence analysis encompasses five key steps: (1) NCBI: retrieval of whole genome sequences; (2) Prokka: genome annotation; (3) OrthoFinder: orthology prediction and grouping; (4) MEME: motif searching within orthogroups; and (5) Excel: compilation of motif information across strains. This process transforms raw genomic data into structured feature sets. (B) Machine learning pipeline builds upon the extracted features through six key stages: Feature Preprocessing, Feature Engineering, Feature Selection, Model Selection and Training, Model Evaluation and Prediction, and Downstream Analysis. Each stage incorporates specific methodologies crucial for developing an accurate classification model. Together, these workflows demonstrate the complete process from raw genomic data acquisition to final classification model development, providing a framework for bacterial species classification using whole genome sequence data.

2-1- Data Acquisition

In this study, genomes from the members of *Salmonella enterica* were used for the framework development. Genome sequences representing diverse *Salmonella enterica* serovars were obtained from the Reference Sequence (RefSeq) [33] and GenBank databases [34] hosted by the National Center for Biotechnology Information (NCBI).

The dataset was carefully curated to ensure the inclusion of a diverse representation of *Salmonella*'s diversity and to avoid potential clonal bias. Strains from multiple sources, such as data submitted by the Centers for Disease Control and Prevention (CDC), the Food and Drug Administration (FDA), and disease control centers worldwide, were downloaded for the modelling purposes. Complete genomes were prioritized and downloaded in FASTA format.

DNA instead of protein was chosen as the genetic material for this research for its advantageous properties: i) DNA can effectively preserve genetic information, ensuring the integrity of the data [35]. ii) DNA is directly connected to genetic makeup for evolutionary studies, making it a suitable choice for this bacterial classification research [36]. iii) DNA has the capacity to precisely capture mutational events [37], which is crucial for understanding the genetic diversity within the *Salmonella enterica* species.

2-2- Feature Extraction: Orthology Search and Motifs Prediction

All the downloaded genomes were subjected for annotation in Prokka [38]. The annotated gene files were subsequently used for orthology prediction in OrthoFinder [39]. OrthoFinder utilizes sequence similarity-based searching to define orthologous genes across the predicted pangenome. Single-copy core sequences were identified and extracted for motifs prediction. Motifs are short, specific sequences of DNA or protein that often serve important functions. MEME [40] was used to identify recurring patterns or motifs within a dataset. The parameters for motif prediction in MEME were set as follows: -dna, -nmotifs 5, -minw 8, -maxw 50, -minsites 2. The predicted motifs were extracted from the single-copy orthologous sequences from each of *Salmonella enterica* strain. These motifs served as the features for further processing and analyses.

2-3- Feature Preprocessing

Feature preprocessing is an important step to prepare raw data into an appropriate data matrix for analyses. In this study, the extracted motifs underwent a series of preprocessing steps using an inhouse Python script. The steps include alignment of motifs for gap filling, removal of features with high missing values, and elimination non-informative features.

Motifs were first extracted from output files yielded from MEME. An E-value lower than 0.05 was set to filter the motifs. Each motif from individual orthologous sets were treated independently and aligned in the MUSCLE algorithm [41]. The alignment created motifs of equal lengths by adding "-" for INDEL events. The aligned motifs were then organized into a 2-dimensional matrix where rows represent different samples and columns represent the features (motifs sequences). To achieve consistency, the "-" symbols were replaced with "N" to represent nucleotide deletions or gaps. Individual feature was further inspected for "N" frequency greater than 5% and identical features in single column were filtered. A class label column was added to represent the different classes of subspecies and serovar.

2-4- Feature Engineering

Feature engineering involves the step in transforming data into a machine learning readable format. The transformation first tokenized the feature sequences that was originally represented in ATCG format as "A"=1, "T"=2, "C"=3 and "G"=4. Alignment gaps "N" were treated as the fifth character ("N"=5) in tokenization step. Subsequently, the "Subspecies" column was encoded using the *LabelEncoder* function (Appendix I: Figure S1).

The process was preceded by screening for Single Nucleotide Polymorphisms (hereafter referred to as SNPs) or informative sites from the raw feature matrix. Each column in the matrix represents a motif feature, and each row corresponds to a sample (strain). The SNPs extraction process was performed by detecting non-identical numerical by comparing different rows of individual columns. To standardize the process, a representative sequence was selected as a reference. Each row (sample) in a given motif column was compared against this reference to identify SNP positions. Once the variant positions were determined, the nucleotides at these positions were extracted and concatenated into a string, generating a unique numeric SNP pattern for each sample. For example, in OG0001278_motif_2 column, SNP positions were found at indices 1, 4, 13, 26, and 43. The resulting SNP pattern for the class "0" is 24332, while for the class "11", had a pattern of 21332, clearly indicating genetic differentiation between classes (Figure 2). This process was iterated for each of the subsequent columns until all features were processed. By extracting the unique feature patterns from the various motif features, the code could capture the distinctive genetic variations associated with each label class. This information is invaluable for further analysis, as it aids in understanding the underlying genomic differences among the diverse label groups represented in the dataset.

The approach is designed with a modular structure involving sequence alignment, SNP position identification, and numerical encoding, which makes it theoretically scalable to longer sequences or more complex organisms. By focusing on informative regions (motifs) rather than entire genomes, the method reduces data dimensionality and computational resources. The methodology can be adapted and extended to support broader applications beyond bacterial genomes.

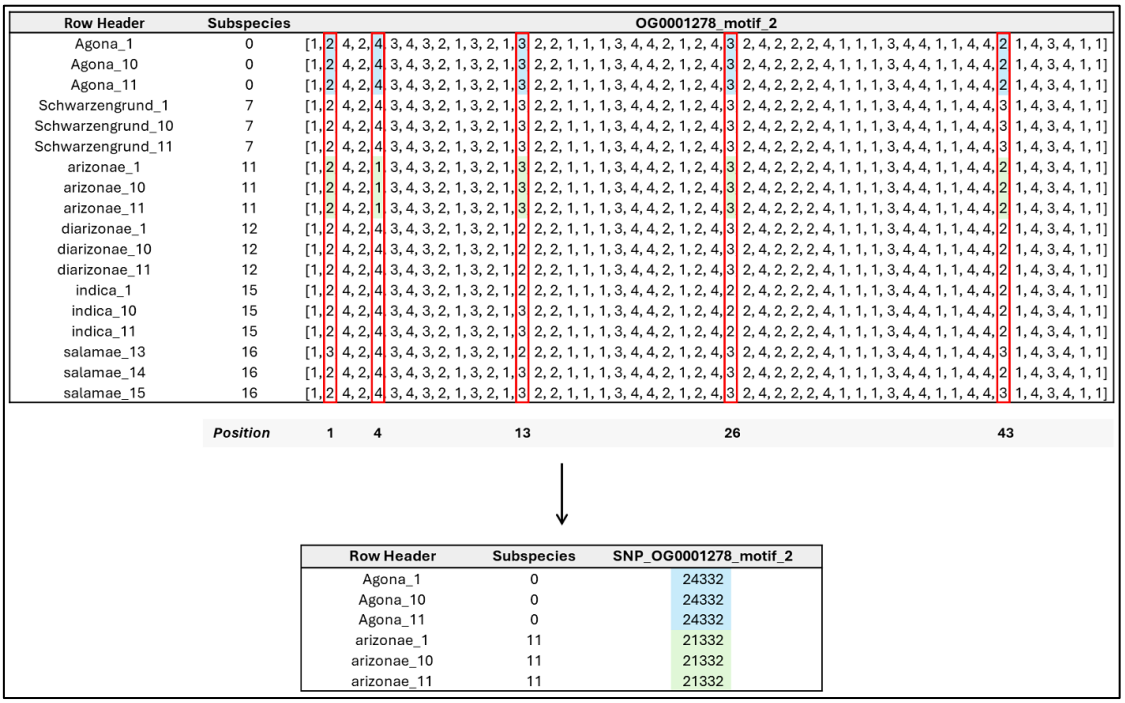

Figure 2. Illustration of the SNP pattern extraction process, showing the transformation of nucleotide sequences into numerical format and subsequent identification of informative sites. Example shows distinct SNP patterns (24332 vs 21332) for different classes.

A streamlined DataFrame was created based on the targeted features. The DataFrame included the 'Row Header' (sample label), 'Subspecies' (class label), and all columns starting with 'SNP_', which contained the extracted SNPs feature. The raw feature was excluded from this consolidated view of the data (Appendix II: Table S1). The transformed features were treated as numeric data and *StandardScaler* from scikit-learn was used to convert the features into a standardized range for the feature selection stage. This step is crucial to ensure that the features are scaled to a comparable range because many machine learning algorithms are sensitive to the relative scale of the input features.

2-5- Feature Selection

Feature selection was performed to identify informative predictors for the bacterial species classification tasks. In this study, filter and embedded feature selection methods were evaluated for their capability of selecting suitable features for classification purposes.

I. Embedded Method

The *SelectFromModel* method, an embedded feature selection technique, was employed in this study. This approach integrates feature selection into the model training process, leveraging the Random Forest Classifier to assess feature importance and select the most significant predictors.

The feature selection process involved running the Random Forest classifier over 100 iterations. During each iteration, the dataset was split into training and testing sets, and the Random Forest model was trained on the training data. The *SelectFromModel* method was then applied, using the mean feature importance as the selection threshold. The indices of the chosen features were recorded, and their counts were incremented accordingly. The embedded feature selection approach resulted in six distinct sets of selected features: Top 100, Top 50, Top 30, Top 20, Top 15 and Top 10. These feature sets were used to train machine learning models for the classification of *Salmonella enterica*, with the aim of identifying the set that yields the best machine learning's evaluation metrics.

II. Filter Method

On the other hand, a filter-based feature selection approach was also employed, using the *SelectKBest* method. Three different score functions were utilized: ANOVA F-value (*f_classif*), Mutual Information (*mutual_info_classif*), and Chi-Squared Test (chi^2). These score functions were selected to leverage their distinct capabilities in identifying relevant features for the classification task.

The ANOVA F-value is suitable for continuous features when the target variable is categorical, as it measures the statistical significance of the relationship between the feature and the target. The Mutual Information score can be applied to both categorical and continuous features, capturing the amount of information gained about one variable through the other, including non-linear relationships. The Chi-Squared Test is particularly effective for categorical features, assessing the independence between the feature and the target variable [42].

A range of k values representing the number of top-ranked features to be selected was evaluated. For each k value, the *SelectKBest* method was used to identify the most significant features based on the respective score function. Subsequently, a Random Forest classifier and a Support Vector Machine were initialized, and a 10-fold cross-validation was performed to assess the accuracy of the models on the selected features. The mean cross-validation accuracy scores were recorded for each k value and score function, enabling analysis of the optimal number of features for the final classification model. The best-performing k value for each score function was identified, and the corresponding top features were selected for training the machine learning models to classify *Salmonella enterica*.

This multi-faceted feature selection approach, leveraging the strengths of different score functions, provided a robust framework for identifying the most relevant features for the classification task, ultimately enhancing the performance of the machine learning models.

2-6- Model Selection and Training

The selected feature dataset was split into training and testing sets using the *train_test_split* function. A test size of 48 was selected, maintaining 80:20 training-to-testing ratio. To ensure the relative proportions of each *Salmonella enterica* subspecies and serovar class were retained in both the training and testing sets, the stratify=y parameter was used during the split. The y variable, which represents the 'Subspecies' target label, served as a basis for the stratified split. This approach helped to avoid skewed representation of the classes, which could have negatively impacted the model's ability to generalize and learn the underlying patterns.

Two machine learning algorithms, Random Forest (RF) and Support Vector Machine (SVM), were evaluated for their effectiveness in the species classification modelling. To optimize the model hyperparameters, the *GridSearchCV* and *RandomizedSearchCV* functions from scikit-learn were employed. These techniques systematically explored a range of hyperparameter values, identifying the best-performing configurations for each model.

2-7- Model Evaluation and Prediction

The trained models were evaluated for their effectiveness using various metrics: accuracy, precision, recall, F1-score, confusion matrix, K-Fold Cross-Validation, and learning curve analysis. These metrics provided insights into the models' performance and their ability to classify the bacterial species.

The accuracy score on the testing set was calculated to assess the models' predictive capabilities. The confusion matrix was then used to analyze the distribution of true and false positives and negatives, offering deeper insights into the models' classification patterns. To evaluate the robustness and generalization of the models, K-Fold Cross-Validation was performed, which functionally to validate the consistency of the models' performance across multiple iterations. Furthermore, learning curve analysis was conducted to understand the models' learning behaviour and identify any potential issues, such as overfitting or underfitting. This analysis provided valuable information about the models' ability to learn from the training data and generalize to new, future samples.

2-8- Selected Features Analysis

A T-distributed Stochastic Neighbor Embedding (t-SNE) was employed to visualize the clustering based on the top feature set. This technique uncovers patterns and outliers within the dataset, enabling the identification of trends and relationships that clarify the data's variations. Additionally, a feature importance analysis has been conducted, ranking the features based on their significance. The Basic Local Alignment Search Tool (BLAST) has also been utilized to analyse the genes associated with the top-ranked features. This analysis predicts identity of genes and their encoded proteins.

3- Result

3-1- Input Matrices

A total of 240 *Salmonella enterica* genomes were selected (Appendix II: Table S2), consisting of 15 samples from each of the six selected subspecies and top ten commonly reported serovars from *S. enterica* subsp. *enterica*. The classes for the machine learning modelling made up of the 16 subspecies and serovars:

- *Salmonella enterica* subsp. *arizonae*,
- *Salmonella enterica* subsp. *diarizonae*,

- *Salmonella enterica* subsp. *enterica*,
- *Salmonella enterica* subsp. *houtenae*,
- *Salmonella enterica* subsp. *indica*,
- *Salmonella enterica* subsp. *salamae*,
- *Salmonella enterica* subsp. *enterica* serovar Typhimurium,
- *Salmonella enterica* subsp. *enterica* serovar Enteritidis,
- *Salmonella enterica* subsp. *enterica* serovar Newport,
- *Salmonella enterica* subsp. *enterica* serovar Typhi,
- *Salmonella enterica* subsp. *enterica* serovar Infantis,
- *Salmonella enterica* subsp. *enterica* serovar Agona,
- *Salmonella enterica* subsp. *enterica* serovar Heidelberg,
- *Salmonella enterica* subsp. *enterica* serovar Dublin,
- *Salmonella enterica* subsp. *enterica* serovar Montevideo, and
- *Salmonella enterica* subsp. *enterica* serovar Schwarzengrund.

Each *Salmonella* genome has an approximate size of 5 million base pairs ($5x10^6$ characters). The annotation predicted 1,131,569 genes in the 240 genomes (16 classes x 15 sample in each class) and further clustered into 22,864 families. A total of 1435 single-copy core genes were predicted and resulted in 1,244,719 nucleotides per strain. From the 1435 families, 7122 motifs were identified, generating a 240 x 7122 matrix. The preprocessing steps further reduced the data to a 240 x 5256 matrix.

3-2- Model Performance Metrics

The performances of Random Forest (RF) and Support Vector Machine (SVM) classifiers indicated distinct metrics across different feature selection methods (Table 1). Both models demonstrated varying levels of effectiveness in classifying *Salmonella enterica*.

Table 1. Random Forest VS SVM: Performance Comparison

Classifier	Feature Selection Method	#Features Selected	Accuracy (Untuned)	Accuracy (Tuned)	Precision	Recall	F1-Score
Random Forest	Embedded - Top 100	100	0.9792	0.9792	0.98	0.98	0.98
Random Forest	Embedded - Top 50	50	0.9792	0.9792	0.98	0.98	0.98
Random Forest	Embedded - Top 30	30	0.9792	0.9792	0.98	0.98	0.98
Random Forest	Embedded - Top 20	20	0.9792	0.9792	0.98	0.98	0.98
Random Forest	Embedded - Top 15	15	0.9167	0.9167	0.88	0.92	0.89
Random Forest	Embedded - Top 10	10	0.6875	0.6875	0.70	0.69	0.65
Random Forest	ANOVA F-value	200	0.9792	0.9792	0.98	0.98	0.98
Random Forest	Mutual Information	100	0.9792	0.9583	0.98	0.98	0.98
Random Forest	Chi-Squared	100	0.9583	0.9375	0.97	0.96	0.96
SVM	Embedded - Top 100	100	0.9375	0.9792	0.95	0.94	0.94
SVM	Embedded - Top 50	50	0.8958	0.9167	0.92	0.90	0.90
SVM	Embedded - Top 30	30	0.8125	0.9375	0.77	0.81	0.78
SVM	Embedded - Top 20	20	0.8542	0.8958	0.80	0.85	0.82
SVM	Embedded - Top 15	15	0.6458	0.7708	0.67	0.65	0.61
SVM	Embedded - Top 10	10	0.4583	0.5417	0.52	0.46	0.43
SVM	ANOVA F-value	200	0.8542	0.8542	0.81	0.85	0.82
SVM	Mutual Information	100	0.8958	0.9792	0.87	0.90	0.87
SVM	Chi-Squared	100	0.6667	0.7292	0.59	0.67	0.60

The Random Forest classifier exhibited consistent performance across most feature selection methods. Using Embedded feature selection methods (Top-ranked 100, 50, 30, and 20 features, hereafter referred to as Top *<#selected features>*), the RF model maintained an accuracy of 0.9792, with corresponding precision, recall, and F1-scores of 0.98. Similar performance was achieved with ANOVA F-value (0.9792) and Mutual Information (0.9792) feature selection methods. However, performance began to decline with Embedded - Top 15 features (accuracy 0.9167), and a significant

drop was observed with Embedded - Top 10, where accuracy decreased to 0.6875. The decline emphasizes the importance of maintaining an adequate feature set size, as excessive feature reduction can lead to information loss and reduced classification performances.

The SVM model showed improvements when comparing untuned versus tuned configurations across various feature selection methods. With the Embedded - Top 100 method, accuracy improved from 0.9375 to 0.9792, while the Mutual Information method showed a significant improvement from 0.8958 to 0.9792 after tuning. Similar improvements were observed with other feature sets: Top 50 improved from 0.8958 to 0.9167, and Top 30 increased from 0.8125 to 0.9375. However, the SVM model struggled with reduced feature sets, particularly with Embedded - Top 10, where accuracy remained low at 0.5417 even after tuning. This pattern highlights the model's sensitivity to feature selection, indicating that a reduced number of features may not provide sufficient information for effective predictions.

The comparisons on the frameworks highlighted the influence of feature selection strategies on the performance of the machine learning. The performance of Random Forest model was stable across various sizes of feature sets, as compared to the SVM classifier, which has shown higher sensitivity on the choice of features. Both models reinforced the importance of feature selection in bioinformatics classification tasks, where the type and number of features greatly influenced taxonomy predictions. To assess the models' stability and generalization capability, K-Fold Cross-Validation was implemented across all feature selection methods (Table 2). The validation approach involved splitting the dataset into subsets for iterative training and validation, providing a more reliable assessment than a single train-test split. A 10-fold cross-validation was applied to both untuned and tuned Random Forest and SVM models to reduce bias and mitigate overfitting. The Random Forest model consistently achieved high accuracy across all feature selection methods, with scores between 0.95 and 0.97 for the Top 20 to Top 100 features. In contrast, the SVM model showed greater improvement after tuning, particularly with smaller feature sets, boosting accuracy from 0.78 to 0.91 with the Top 20 features. However, performance dropped significantly with very small feature sets, indicating reduced robustness. Overall, both models demonstrated minimal performance variance across folds, supporting the reliability of the proposed classification framework.

Table 2. Ten-Fold Cross Validation: Random Forest VS SVM Performance

Model	Feature Selection Method	Average Accuracy (10-Fold CV Untuned)	Average Accuracy (10-Fold CV Tuned)
Random Forest	Embedded - Top 100	0.95	0.95
Random Forest	Embedded - Top 50	0.95	0.96
Random Forest	Embedded - Top 30	0.97	0.95
Random Forest	Embedded - Top 20	0.96	0.96
Random Forest	Embedded - Top 15	0.91	0.92
Random Forest	Embedded - Top 10	0.69	0.71
Random Forest	ANOVA F-value	0.94	0.92
Random Forest	Mutual Information	0.94	0.94
Random Forest	Chi-Squared	0.94	0.93
SVM	Embedded - Top 100	0.92	0.95
SVM	Embedded - Top 50	0.89	0.92
SVM	Embedded - Top 30	0.84	0.89
SVM	Embedded - Top 20	0.78	0.91
SVM	Embedded - Top 15	0.58	0.78
SVM	Embedded - Top 10	0.38	0.55
SVM	ANOVA F-value	0.50	0.79
SVM	Mutual Information	0.91	0.93
SVM	Chi-Squared	0.37	0.63

3-3- Confusion Matrix

The Random Forest model with Embedded - Top 20 emerges as the best choice for further analysis, showing high accuracy score and consistent performance metrics in classifying *Salmonella enterica* subspecies and serovar classes with a minimal set of features. This model was selected for detailed analysis using confusion matrices to examine its classification patterns across different *Salmonella enterica* subspecies and serovars.

The confusion matrix (Figure 3) provides insights into the model's classification performance, revealing the distribution of predictions across all classes. With three samples per class in the testing dataset, the matrix's diagonal elements represent successful classifications. The model demonstrated a near-perfect classification accuracy, correctly identifying all three samples in most categories. However, one misclassification occurred where a sample from class 2 (serovar Enteritidis) was incorrectly assigned to class 10 (serovar Typhimurium). Despite this single error, the overall performance remained strong, with an accuracy of 0.9792, validating the model's effectiveness for bacterial species classification tasks.

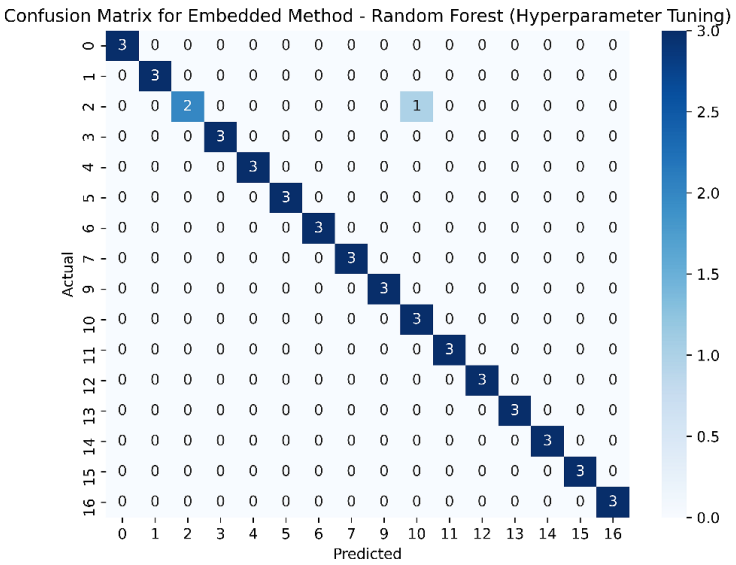

Figure 3. Confusion matrix of Random Forest model for *Salmonella* identification using Embedded-Top 20 features, showing high accuracy with one misclassification

3-4- Learning Curve Analysis

Learning curve analysis was conducted to evaluate model generalization capabilities and identify potential overfitting or underfitting patterns in the classification models. The analysis compared the performance characteristics of two optimized classifiers: Random Forest and Support Vector Machine (SVM), both implementing Embedded - Top 20 feature selection (Figure 4).

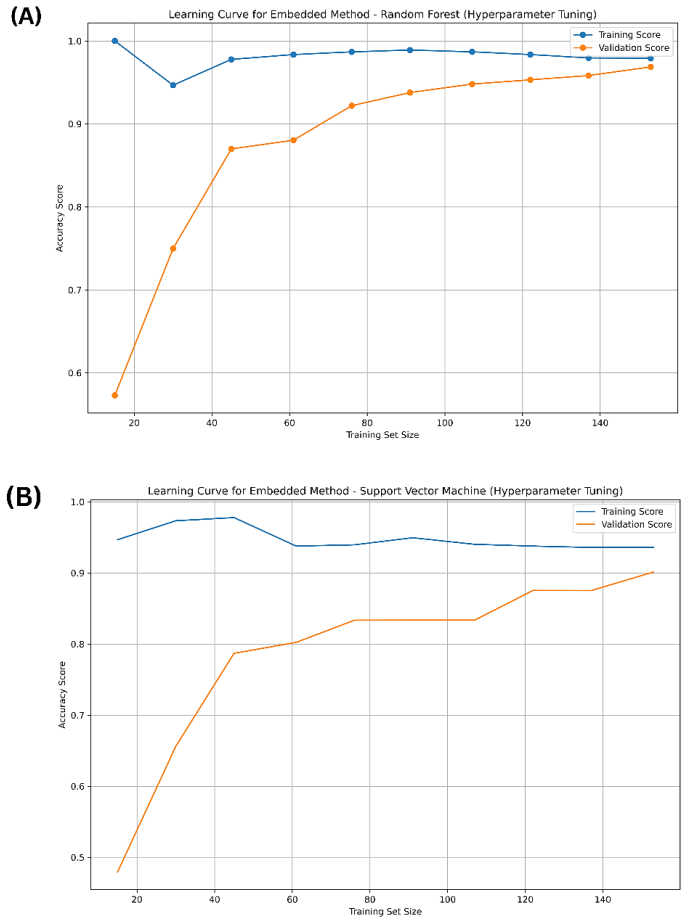

Figure 4. Learning curves of the optimized Random Forest and Support Vector Machine models with Embedded-Top 20 features. (A) Random Forest model demonstrates high training scores (~1.0) and improving validation performance (0.57 to 0.95), with convergence between training and validation scores indicating effective learning and good generalization capabilities. (B) SVM model exhibits high training scores (~0.98) but shows slower improvement in validation performance (0.48 to 0.9), characterized by a persistent gap between training and validation scores, suggesting different learning dynamics compared to Random Forest.

The Random Forest model demonstrated strong learning behaviour, achieving consistent training scores (approximately 1.0) throughout the learning process. More importantly, its validation performance showed improvement from 0.57 to 0.95 as training data increased, with training and validation scores ultimately converging. This pattern, particularly evident at larger training sizes, indicates effective generalization capabilities in handling complex bacterial genomic data.

In contrast, although the SVM model maintained high training scores (approximately 0.98), its learning dynamics indicated potential limitations. The model exhibited a more gradual improvement in validation performance (from 0.48 to 0.90) and, notably, maintained a persistent gap between training and validation scores. This gap, was pronounced with smaller training sets, suggests suboptimal generalization capabilities and potential underfitting issues.

The comparative analysis of these learning patterns provides quantitative evidence supporting the selection of the Random Forest model as the optimal classifier for bacterial species identification, demonstrating adaptability to the complexity of biological sequence data.

3-5- Selected Features and Functional Analysis

The Random Forest (RF) model with embedded feature selection identified the top 20 most discriminative features for bacterial classification (Appendix II: Table S3). These features showed predictive power in distinguishing bacteria into their respective classes. Of the 20 motifs selected, they were extracted from 17 genes. Orthogroup OG0002551 contributed four distinct motifs to the feature set. This observation suggests that single genes can contain multiple informative variation sites, each potentially influencing various aspects of protein function or regulation.

BLAST analysis was performed using the complete gene sequences of the 17 orthogroups rather than individual motifs because full-length genes provide more sequence context and yield statistically robust matches in homology searches. The analysis revealed sequence conservation across different bacterial strains (Table 3), with 14 out of 17 proteins demonstrating 100% sequence identity to known bacterial proteins. The remaining three proteins showed lower identity values (98.85-99.48%) with e-values ranged from 0.0 to 6e-111.

Table 3. BLAST Analysis Results of Top Selected Features

Selected Features	Protein	Gene	E value	Percent Identity	Accession
OG0001969	putative MutT-like protein	*ymfB*	6e-111	100.00%	CCF88280.1
OG0001324	bacterioferritin-associated ferredoxin	-	2e-36	100.00%	EAN8508535.1
OG0002646	protease modulator HflC	*hflC*	0.0	100.00%	ECL6753627.1
OG0001679	ATP-dependent RNA helicase SrmB	*srmB*	0.0	100.00%	WP_000219174.1
OG0001322	Small ribosomal subunit protein uS7 * 30S ribosomal protein S7	*rpsG*	3e-108	100.00%	Q0SZX6.1
OG0002551 * 4	30S ribosomal protein S20	*rpsT*	3e-52	98.85%	WP_001518655.1
OG0002500	isoleucine tRNA synthetase	*ileS*	0.0	100.00%	AAX63946.1
OG0002819	F0F1 ATP synthase subunit alpha	*atpA*	0.0	100.00%	WP_001176751.1
OG0001645	glutamate--cysteine ligase	-	0.0	100.00%	ESF02326.1
OG0002023	L,D-transpeptidase family protein	-	0.0	100.00%	WP_000817060.1
OG0002227	very short patch repair endonuclease	-	3e-111	100.00%	WP_000785978.1
OG0002173	zinc ABC transporter permease subunit ZnuB	*znuB*	1e-134	98.85%	EAA7245923.1
OG0002326	DUF1479 domain-containing protein	-	0.0	100.00%	WP_000210948.1
OG0002110	cytochrome bd-II oxidase subunit 1	*appC*	0.0	100.00%	WP_000263612.1
OG0002528	3-isopropylmalate dehydratase small subunit	*leuD*	3e-147	100.00%	WP_000818267.1
OG0002228	DNA cytosine methyltransferase	-	0.0	100.00%	WP_001157300.1
OG0001434	divisome-associated lipoprotein YraP	*yraP*	4e-119	99.48%	ENG9187551.1

Feature importance analysis revealed patterns in the predictive power of these motifs. Among the 20 selected features (Appendix II: Table S4), SNP_OG0002326_motif_3 emerged as the most significant predictor with an importance score of 0.1123, followed by SNP_OG0002228_motif_5 (0.1064) and SNP_OG0002528_motif_2 (0.0961). Despite having multiple motifs selected by the model, features from OG0002551 showed lower impact on predictions, with SNP_OG0002551_motif_1 (0.0043), SNP_OG0002551_motif_2 (0.0047), and SNP_OG0002551_motif_5 (0.0050) displaying the lowest importance scores. This suggests that while a gene may contain multiple informative motifs, their contributions to classification accuracy are varies.

Evolution theory highlights the concept of nucleotide mutation for adaptation. Functional characterization of the high-ranked features from this study showed the possible association of motif sequence variations with their biological adaptation in different subspecies/serovars. The highest-ranked feature, SNP_OG0002326_motif_3, corresponds to a DUF1479 domain-containing protein, important for regulating bacterial physiology and biofilm formation [43]. The second most influential feature, SNP_OG0002228_motif_5, was identified within the DNA cytosine methyltransferase gene, indicating its involvement in epigenetic modifications and gene regulation [44, 45]. The third-ranked feature, SNP_OG0002528_motif_2, mapped to the 3-isopropylmalate dehydratase small subunit, highlighting the importance of amino acid biosynthesis pathways [46] in distinguishing bacterial strains.

3-6- t-SNE Visualization of Subspecies and Serovar Clustering based on Selected Features

To analyze the discriminative power of the selected features, a t-SNE analysis was performed, revealing distinct genetic boundaries among Salmonella taxa and supporting the feature selection approach through a clear separation at the subspecies level (Figure 5). On the other hand, the clustering patterns among serovars provide insights into their close evolutionary relationships. The overlapping distributions between serovars indicate shared ancestral lineages or convergent evolution under similar selective pressures. These patterns are relevant for understanding virulence mechanisms and host specificity among closely related strains. This visualization demonstrates the genetic relationships within *Salmonella* and highlights areas for future research. Further investigation of the genetic elements driving these clustering patterns could reveal mechanisms underlying pathogenicity and host adaptation, advancing our understanding of *Salmonella* diversity and evolution.

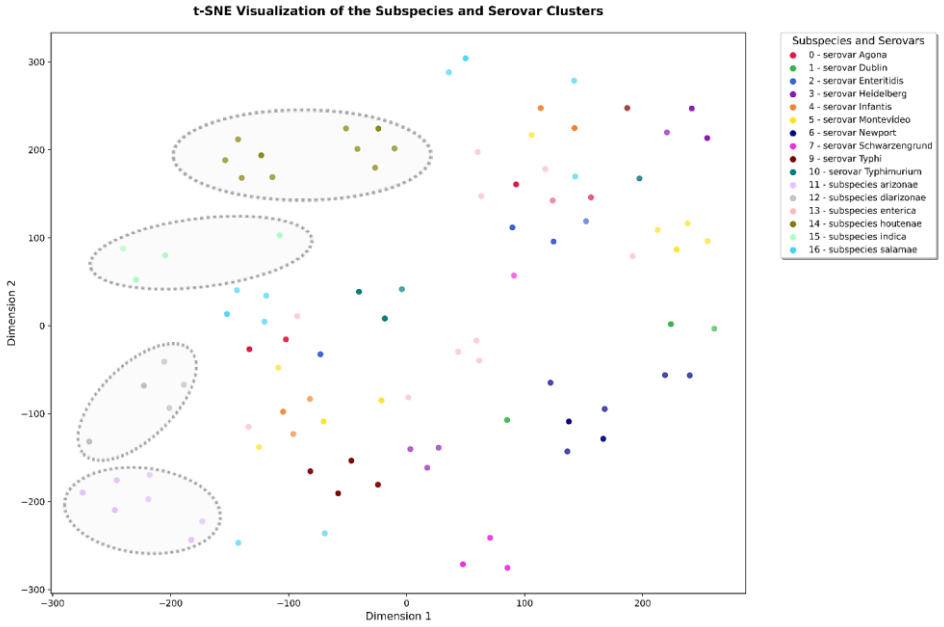

Figure 5. t-SNE visualization of *Salmonella* subspecies and serovar clustering patterns based on Embedded-Top20 dataset. Individual points represent samples, color-coded by subspecies/serovar classification. Subspecies *arizonae*, *diarizonae*, *houtenae*, and *indica* form distinct clusters (circled with grey), while *salamae* shows dispersed distribution. Subspecies *enterica* displays widespread distribution due to its diverse serovar composition (Panama, Pullorum, Blockley etc). Among serovars, Newport forms a distinct cluster, while Infantis-Montevideo and Typhi-Schwarzengrund show overlapping distributions.

4- Discussion

4-1- Identification of Bacterial Species

The emergence of bacteria can be dated back to approximate 3.5 billion years ago [47]. Through the course of evolution, bacteria speciate and have vital roles in diverse ecosystems, ranging from human pathogens and probiotics to environment sustainability. Hence, it is important to identify the species to determine its biological potentials and threats.

Traditional biochemical methods for bacterial identification provide inexpensive, quantitative, and qualitative insights into microbial diversity. However, these techniques are labour-intensive and time-consuming, involving multiple steps such as media preparation, dilution, plating, incubation, counting, isolation, and characterization, often require days to obtain results. Most importantly, traditional methods tends to yield false positives results and may fail to deliver consistent and reliable identification of microorganisms at the species or strain level [1, 2].

The 16S ribosomal RNA (rRNA) gene has been widely used in bacterial phylogenetics and microbiome studies. However, research within the recent decades have identified significant limitations associated with this approach and overruled some previously defined bacterial taxonomy [3, 4, 48]. Hassler et al. (2022) [12] reported that the *16S rRNA* can exhibit considerable intragenomic heterogeneity, recombination, and horizontal gene transfer, all of which can

compromise the reliability of speciation signals. The complexity can also be observed in some bacterial species with multiple copies of *16S rRNA* predicted originate from different identities, such as from the members of *Streptococcus* and *Mycobacterium* [5]. In addition, although *16S rRNA* serves as a gold standard in bacterial species identification, however it could only be suitable to resolve taxonomy level of majority but not all of bacteria genera. These concept suggest that relying solely on the *16S rRNA* gene may not be sufficient for accurate bacterial identification and phylogenetic analysis, particularly at the species or strain level [49].

Multi-Locus Sequence Typing (MLST) provides valuable insights into the phylogenetic relationships among bacteria for species identification purpose. The concept of MLST is to maximize speciation signal through combination of multiple genes in the bacteria for species identification purpose. However, it may be lacked of the necessary resolution for epidemiological tracing. Research by Yan et al. (2021) [8] found that MLST often struggles to differentiate the serotypes of bacterial isolates. This is evidenced by instances of unexpected identification of novel sequence types or the absence of specific serovar designations within the MLST database.

Multiple reports showed genome-based identification provides a clearer resolution and separate taxonomies that were impossible to achieve with the classical approaches [50-52]. Genome provides an advantage on screening for regions with genuine species signals along genomes, such as genes and intergenic regions. For this study, we focused on motifs predicted from genomes and evaluated its usefulness for species identification. Motifs are short, specific sequences of DNA or protein that play essential roles in various biological functions. They are crucial for understanding the structure, function, and evolution of biological molecules [40]. Motifs can represent conserved elements that interact with specific proteins, such as transcription factors. Identifying these motifs allows researchers to gain insights into the regulatory mechanisms governing gene expression [53]. Conserved motifs can indicate the presence of promoters, regulatory elements, and splice junctions between protein-coding exons [54]. Given its significance, motif is hypothesised to retain the memory of speciation (evolution) and not to be influenced by random mutation and recombination.

4-2- Previous Related Methods and Advantages of Current Research

The use of genome able to maximize the speciation signals. However, the signal can be disrupted with the amplification of noise from within the genome and lead to species misclassification. The risk of the issue can be minimized with a proper methodology and algorithms. The techniques in machine learning have been increasingly applied to evaluate signals for bacterial classification purposes. Previous studies have primarily utilized imaging techniques for bacterial classification. However, these methods pose significant challenges, including issues such as image noise, low resolution, and size of objects. Additionally, the limited availability of high-quality datasets, along with the small size and imbalanced nature of existing datasets, can lead to overfitting, particularly in deep learning models [55-59]. In addition, others machine learning studies have also focused on genes markers such as *16S rRNA* [60, 61], *16S rDNA* [62] or MLST genes [63, 64]. However, these approaches are still unable to differentiate closely related taxonomic levels (Table 4). Besides image and sequence datasets, spectral datasets are also commonly used in classifying bacterial species [65-68].

Table 4. Summary of Previous Related Methods (Sequence Datasets) and Their Performance

Research Citation	Dataset	Taxonomic Classification Level	Reported Accuracy
Fiannaca et al. (2018) [60]	Organism: Bacteria (*Proteobacteria* phylum), using 57,788 16S gene sequences	From class to genus level	**Highest Accuracy:** Class level: Close to 100% Amplicon (AMP) dataset at genus level: 91.3% (CNN and DBN) **Lowest Accuracy:** Shotgun (SG) dataset at class level: Around 20% Shotgun (SG) dataset at genus level: Around 85.5%
Cserhati et al. (2019) [69]	Organism: Insect (*Anopheles* genus, *Drosophila* genus, *Glossina* genus), using k-mer sequences (specifically k-mers of lengths 7-9 base pairs)	Genus	*Anopheles* genus: Mean correlation coefficient of 0.948 *Drosophila* genus: Mean correlation coefficient of 0.869 *Glossina* genus: Mean correlation coefficient of 0.978
Liang et al. (2020) [70]	Organism: Bacteria, 2,505 bacterial genomes from human gut microbiome, with simulated metagenomic reads	Species and Genus	**Species Classification:** Read-level Precision: Mean of 0.942 Read-level Recall: Mean of 0.428 **Genus Classification:** Read-level Precision: 0.969 on average Read-level Recall: 0.866 on average
Helaly et al. (2021) [61]	Organism: Bacteria, 1,000 sequences from the *16S rRNA* gene barcode (three bacterial phyla: *Actinobacteria*, *Firmicutes*, and *Proteobacteria*)	Phylum, Class, Order, Family, Genus	Maximum accuracy of 91.7% (achieved with a deeper CNN using more representative sequence representation) and accuracy of 90.6% (achieved with a wide and shallow CNN using less representative representations)
Mock et al. (2022) [71]	Organisms: Four superkingdoms - Archaea, Bacteria, Eukaryotes, and Viruses (full DNA sequences form NCBI RefSeq)	Superkingdom (Archaea, Bacteria, Eukaryota, Viruses) Phylum Genus	For Final Model Dataset: Superkingdom: 98.62% Phylum: 95.10% Genus: 66.92%
Meharunnisa et al. (2024) [62]	Organism: Bacteria, 12,508 sequences from *16S ribosomal DNA (rDNA)* gene (371 genera)	Genus	Proposed Integrated CNN-RF Model: 98.93% Standalone CNN Model: 91.95% Standalone Random Forest Model: 68.78%

Among all other related studies, Cserhati et al. [69] utilized DNA k-mers derived from whole genome sequences, making this study particularly relevant to our research. Different from our study, the authors developed a k-mer analysis algorithm to examine not only whole-genome sequences but also specific genomic regions, such as introns, 5' UTRs, and 3' UTRs, across 58 insect species from the *Anopheles*, *Drosophila*, and *Glossina* genera. The methodology involved enumerating and scoring all possible k-mers (7-9 bp) in both whole-genome and subgenomic regions for each species. The authors calculated Pearson correlation coefficients between the whole-genome k-mer signatures (WGKS) of all species pairs and visualized the correlation matrix as a heatmap to identify clusters of closely related species within each genus. The classification was based on the overall k-mer signatures of the genomes. Although both studies employ similar genomic features - the authors' k-mers and the present study's motifs derived from whole genome sequences, Cserhati et al.s' [69] approach is confined to genus-level classification, whereas this research machine learning-based framework extend the classification capability to the serovar level.

4-3- Salmonella as a Model Genus

We challenged the performance of the developed framework by using genomes from 16 subspecies/serovars of *Salmonella*. Making accurate classification on pathogens is crucial for public health and food safety [72] and *Salmonella* is a significant pathogen known for its potential to cause several health imapcts. Most importantly, the genetic relatedness in *Salmonella* has caused many misclassifications through laboratory works and bioinformatics algorithms [17, 73].

Traditionally, classifying *Salmonella* down to the serovar level requires a labor-intensive and time-consuming serotyping process [6], which must be performed in a well-equipped laboratory by experienced personnel. This research presents an efficient framework for classifying *Salmonella* into respective subspecies or serovar classes using genome data.

4-4- Novelty of Research

Clustering based on whole genome phylogenies were generated to illustrate the effect of noise on *Salmonella* classification (Appendix I: Supp S1). Both Neighbour-Joining (distance based) and Maximum-Likelihood (character based) that were inferred using whole genome sequences showed incongruent classification for several *Salmonella* strains, including: *Salmonella enterica* subsp. *enterica* serovar Enteritidis str. 92-0392 (Enteritidis_14), *Salmonella enterica* subsp. *salamae* serovar 40:c:e,n,z15 str. 2013K-0524 (Salamae_12), *Salmonella enterica* subsp. *salamae* serovar 48:z81:z39 str. 2015K-0023 (Salamae_13), *Salmonella enterica* subsp. *salamae* serovar 6,7:m,t:- str. CFSAN028548 (Salamae_15), *Salmonella enterica* subsp. *enterica* serovar Montevideo str. ATCC 8387 (Montevideo_1), *Salmonella enterica* subsp. *enterica* serovar Montevideo str. USDA-ARS-USMARC-1913 (Montevideo_6), and *Salmonella enterica* subsp. *enterica* serovar Montevideo str. 4441 H (Montevideo_13). The clustering further proven the amplification of noise able to mask genuine signal for species classification when there is no proper screening, at least in the *Salmonella* genus.

One of the main challenges in Bioinformatics is data transformation into a machine learning readable format. In this research, although the use of motifs significantly reduced sequence data into a more manageable size in machine learning, the dimension of the data matrix is still huge. Apart from data size, compression of data into feature vectors without loss of species information was a greater challenge. This research employed tokenization and label encoding methods, which provide a direct and efficient way to represent both the labels and features of the data. Furthermore, Single Nucleotide Polymorphism patterns extracted from the motif features enhance the ability to distinguish between closely related bacteria, down to the *Salmonella* serovars level.

Commonly used techniques for taxonomy classification are one-hot encoding or k-mer representation [53, 60, 61, 74-76] and position weight matrix (PWM) models [77, 78]. The systematic approach to reduce the data size and feature engineering marks this research as distinctive compared to previous studies, which often relied on partial genomic information or less robust data transformation methods. The techniques applied in transforming features to a machine-learning-readable format also effectively reduce data dimensionality.

Particularly noteworthy was the finding that Random Forest (RF) is particularly suitable for classification tasks using motif sequences. Comparative analysis revealed that the Random Forest model maintained consistent performance across various feature selection methods compared to the Support Vector Machine (SVM) model, especially when processing larger feature sets. This performance highlights RF's robustness in managing high-dimensional data. These results align with existing literature indicating that Random Forest consistently delivers strong classification outcomes, attributed to its ensemble nature, which mitigates overfitting and enhances generalization capabilities [79-81]. Despite the strong performance of Random Forest, both models emphasized the importance of feature selection, as reducing the number of selected features resulted in performance degradation.

The machine learning classification model identified a single instance of misclassification in the dataset. *Salmonella enterica* subsp. *enterica* serovar Enteritidis str. 92-0392 (GCF_002761055.1), originally labelled as serovar Enteritidis (class 2), was classified by the model as serovar Typhimurium (class 10). This discrepancy prompted further

investigation into the cause of misclassification. The *16S rRNA* and single-copy core genes phylogenies grouped the strain to the Typhimurium clade (Appendix I: Supp S2, Figure S2; Appendix II: Table S5). This classification was also consistent with the NCBI database identification of *Salmonella enterica subsp. enterica* serovar Typhimurium str LT2 as the best match type strain for the misclassified isolate (*https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002761055.1/*).

Apart from the systematic feature engineering from this study, the literature search indicated limited use of motifs for classification purposes. Moreover, there were no reports on the use of sequence motifs in machine learning. The findings from this study further highlight the utility of evolutionary signals from sequence motifs within orthologous genes. These motifs enable precise discrimination of bacterial taxa at finer taxonomic levels. This is in contrast with other approaches, such as full-genome comparisons that often introduce noise, and traditional 16S rRNA marker-based methods that frequently lack taxonomic resolution.

4-5- Selected Feature Analysis: Implications for Bacterial Classification

The current feature selection approach showed that our framework could be useful in predicting new taxonomic markers. The current findings align with and extend previous research on established taxonomic markers. The identification of *atpA* as a significant feature corroborates earlier studies [82, 83] that demonstrated its discriminatory power as compared to the *16S rRNA* sequences, particularly in distinguishing closely related *Enterococcus* species. Additionally, ribosomal proteins (*rpsG* and *rpsT*) were among the selected features. Research has shown that phylogenetic analyses based on ribosomal proteins often provide higher resolution than traditional 16S rRNA-based methods [84, 85]. The application of *rpsG* in studying Phl-producing *Pseudomonas* strains [86] further supports the predictive power of the framework. The analysis has also identified several candidate markers that extend beyond the traditionally used genes. The identification of *HflC* protein from this study, with its unique single-amino-acid repeat patterns [87] suggests new possibilities for strain-level identification. Similarly, the strong performance of *ileS* in the analysis aligns with recent findings demonstrating its effectiveness in classifying dairy-associated *Pseudomonas* [88] indicating its broader potential as a taxonomic marker.

The varying importance levels among the identified features suggest potential utility in a hierarchical classification approach. These features could serve as complementary markers alongside existing classification systems, enhancing the accuracy of bacterial taxonomy. Further research and validation studies would be valuable to fully establish their effectiveness as taxonomic markers.

5- Conclusion

This study presents a computational framework that integrates motif-based genomic analysis with machine learning techniques to improve bacterial taxonomy classification. The framework centres on three key components: the extraction of evolutionary sequence motifs from single-copy orthologous genes, the engineering and selection of informative SNP-based features, and the application of machine learning algorithms - particularly Random Forest - for accurate classification. The use of motifs not only reduces the dimension of whole-genome data but also preserves biologically meaningful variations. This approach successfully differentiates *Salmonella* strains at both subspecies and serovar levels, achieving high accuracy using a minimal feature set. Among the classifiers tested, the Random Forest model with embedded feature selection (Top 20 features) demonstrated the most consistent performance. Although the study focused on the *Salmonella*, the capability of the framework to resolve identity of strains down to the serovar level highlight its applicability potential onto other genetically or genomically complex taxa, especially in addressing public health concerns.

6- Declarations

6-1- Author Contributions

Conceptualization, L.W.Y. and J.L.T.; methodology, L.W.Y., L.Y.O., and J.L.T.; validation, L.W.Y., L.Y.O., and J.L.T.; formal analysis, L.W.Y. and J.L.T.; data curation, L.W.Y.; writing—original draft preparation, L.W.Y., L.Y.O., and J.L.T.; writing—review and editing, L.W.Y., L.Y.O., and J.L.T.; visualization, L.W.Y., L.Y.O., and J.L.T. All authors have read and agreed to the published version of the manuscript.

6-2- Data Availability Statement

Data and source code available in a publicly accessible repository: The data presented in this study are openly available in GitHub at https://github.com/libunn/Bacteria-Species-Classification-ML-Framework.

6-3- Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6-4- Institutional Review Board Statement

Not applicable.

6-5- Informed Consent Statement

Not applicable.

6-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7- References

[1] Edet, U., Antai, S., Brooks, A., Asitok, A., Enya, O., & Japhet, F. (2017). An Overview of Cultural, Molecular and Metagenomic Techniques in Description of Microbial Diversity. Journal of Advances in Microbiology, 7(2), 1–19. doi:10.9734/jamb/2017/37951.

[2] Franco-Duarte, R., Černáková, L., Kadam, S., Kaushik, K. S., Salehi, B., Bevilacqua, A., Corbo, M. R., Antolak, H., Dybka-Stępień, K., Leszczewicz, M., Tintino, S. R., de Souza, V. C. A., Sharifi-Rad, J., Coutinho, H. D. M., Martins, N., & Rodrigues, C. F. (2019). Advances in chemical and biological methods to identify microorganisms—from past to present. Microorganisms, 7(5), 130. doi:10.3390/microorganisms7050130.

[3] Srinivasan, R., Karaoz, U., Volegova, M., MacKichan, J., Kato-Maeda, M., Miller, S., Nadarajan, R., Brodie, E. L., & Lynch, S. V. (2015). Use of 16S rRNA gene for identification of a broad range of clinically relevant bacterial pathogens. PLoS ONE, 10(2), 117617. doi:10.1371/journal.pone.0117617.

[4] Mishra, A., Nam, G. H., Gim, J. A., Seong, M., Choe, Y., Lee, H. E., Jo, A., Kim, S., Kim, D. H., Cha, H. J., Kang, H. Y., Choi, Y. H., & Kim, H. S. (2017). Comparative evaluation of 16S rRNA gene in worldwide strains of Streptococcus *iniae* and Streptococcus parauberis for early diagnostic marker. Genes and Genomics, 39(7), 779–791. doi:10.1007/s13258-017-0542-7.

[5] Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clinical Microbiology Reviews, 17(4), 840–862. doi:10.1128/CMR.17.4.840-862.2004.

[6] Grinevich, D., Harden, L., Thakur, S., & Callahan, B. (2024). Serovar-level identification of bacterial foodborne pathogens from full-length 16S rRNA gene sequencing. MSystems, 9(3), 00757–23. doi:10.1128/msystems.00757-23.

[7] Bertolo, A., Valido, E., & Stoyanov, J. (2024). Optimized bacterial community characterization through full-length 16S rRNA gene sequencing utilizing MinION nanopore technology. BMC Microbiology, 24(1), 58. doi:10.1186/s12866-024-03208-5.

[8] Yan, S., Zhang, W., Li, C., Liu, X., Zhu, L., Chen, L., & Yang, B. (2021). Serotyping, MLST, and Core Genome MLST Analysis of Salmonella enterica From Different Sources in China During 2004–2019. Frontiers in Microbiology, 12. doi:10.3389/fmicb.2021.688614.

[9] Jacob, J. J., Rachel, T., Shankar, B. A., Gunasekaran, K., Iyadurai, R., Anandan, S., & Veeraraghavan, B. (2020). MLST based serotype prediction for the accurate identification of non typhoidal Salmonella serovars. Molecular Biology Reports, 47(10), 7797–7803. doi:10.1007/s11033-020-05856-y.

[10] Floridia-Yapur, N., Rusman, F., Diosque, P., & Tomasini, N. (2021). Genome data vs MLST for exploring intraspecific evolutionary history in bacteria: Much is not always better. Infection, Genetics and Evolution, 93, 104990. doi:10.1016/j.meegid.2021.104990.

[11] Georgiades, K., & Raoult, D. (2011). Defining pathogenic bacterial species in the genomic era. Frontiers in Microbiology, 1, 151. doi:10.3389/fmicb.2010.00151.

[12] Hassler, H. B., Probert, B., Moore, C., Lawson, E., Jackson, R. W., Russell, B. T., & Richards, V. P. (2022). Phylogenies of the 16S rRNA gene and its hypervariable regions lack concordance with core genome phylogenies. Microbiome, 10(1), 104. doi:10.1186/s40168-022-01295-y.

[13] Alikhan, N. F., Zhou, Z., Sergeant, M. J., & Achtman, M. (2018). A genomic overview of the population structure of Salmonella. PLoS Genetics, 14(4), 1007261. doi:10.1371/journal.pgen.1007261.

[14] Yates, J. R., & Osterman, A. L. (2007). Introduction: Advances in genomics and proteomics. Chemical Reviews, 107(8), 3363-3366. doi:10.1021/cr068201u.

[15] Chan, J. Z. M., Halachev, M. R., Loman, N. J., Constantinidou, C., & Pallen, M. J. (2012). Defining bacterial species in the genomic era: Insights from the genus Acinetobacter. BMC Microbiology, 12(1), 302. doi:10.1186/1471-2180-12-302.

[16] Hugenholtz, P., Chuvochina, M., Oren, A., Parks, D. H., & Soo, R. M. (2021). Prokaryotic taxonomy and nomenclature in the age of big sequence data. ISME Journal, 15(7), 1879–1892. doi:10.1038/s41396-021-00941-x.

[17] Trees, E., Carleton, H. A., Folster, J. P., Gieraltowski, L., Hise, K., Leeper, M., Nguyen, T. A., Poates, A., Sabol, A., Tagg, K. A., Tolar, B., Vasser, M., Webb, H. E., Wise, M., & Lindsey, R. L. (2024). Genetic Diversity in Salmonella enterica in Outbreaks of Foodborne and Zoonotic Origin in the USA in 2006–2017. Microorganisms, 12(8), 1563. doi:10.3390/microorganisms12081563.

[18] Uelze, L., Grützke, J., Borowiak, M., Hammerl, J. A., Juraschek, K., Deneke, C., Tausch, S. H., & Malorny, B. (2020). Typing methods based on whole genome sequencing data. One Health Outlook, 2(1), 3. doi:10.1186/s42522-020-0010-1.

[19] Pightling, A. W., Pettengill, J. B., Luo, Y., Baugher, J. D., Rand, H., & Strain, E. (2018). Interpreting whole-genome sequence analyses of foodborne bacteria for regulatory applications and outbreak investigations. Frontiers in Microbiology, 9, 1482. doi:10.3389/fmicb.2018.01482.

[20] Jin, Y., Li, Y., Huang, S., Hong, C., Feng, X., Cai, H., Xia, Y., Li, S., Zhang, L., Lou, Y., & Guan, W. (2024). Whole-Genome Sequencing Analysis of Antimicrobial Resistance, Virulence Factors, and Genetic Diversity of Salmonella from Wenzhou, China. Microorganisms, 12(11), 2166. doi:10.3390/microorganisms12112166.

[21] Jiang, M., Bu, C., Zeng, J., Du, Z., & Xiao, J. (2021). Applications and challenges of high performance computing in genomics. CCF Transactions on High Performance Computing, 3(4), 344–352. doi:10.1007/s42514-021-00081-w.

[22] Bagger, F. O., Borgwardt, L., Jespersen, A. S., Hansen, A. R., Bertelsen, B., Kodama, M., & Nielsen, F. C. (2024). Whole genome sequencing in clinical practice. BMC Medical Genomics, 17(1), 39. doi:10.1186/s12920-024-01795-w.

[23] Qin, Y., Wu, L., Zhang, Q., Wen, C., Van Nostrand, J. D., Ning, D., Raskin, L., Pinto, A., & Zhou, J. (2023). Effects of error, chimera, bias, and GC content on the accuracy of amplicon sequencing. MSystems, 8(6), 01025–23. doi:10.1128/msystems.01025-23.

[24] Jia, H., Tan, S., & Zhang, Y. E. (2024). Chasing Sequencing Perfection: Marching Toward Higher Accuracy and Lower Costs. Genomics, Proteomics and Bioinformatics, 22(2), 24. doi:10.1093/gpbjnl/qzae024.

[25] Nguembang Fadja, A., Riguzzi, F., Bertorelle, G., & Trucchi, E. (2021). Identification of natural selection in genomic data with deep convolutional neural network. BioData Mining, 14(1), 51. doi:10.1186/s13040-021-00280-9.

[26] Hamed, B. A., Ibrahim, O. A. S., & Abd El-Hafeez, T. (2023). Optimizing classification efficiency with machine learning techniques for pattern matching. Journal of Big Data, 10(1), 124. doi:10.1186/s40537-023-00804-6.

[27] Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics, 2, 927312. doi:10.3389/fbinf.2022.927312.

[28] Yin, L., Zhang, H., Zhou, X., Yuan, X., Zhao, S., Li, X., & Liu, X. (2020). KAML: Improving genomic prediction accuracy of complex traits using machine learning determined parameters. Genome Biology, 21(1), 146. doi:10.1186/s13059-020-02052-w.

[29] Ashton, P. M., Nair, S., Peters, T. M., Bale, J. A., Powell, D. G., Painset, A., Tewolde, R., Schaefer, U., Jenkins, C., Dallman, T. J., De Pinna, E. M., & Grant, K. A. (2016). Identification of Salmonella for public health surveillance using whole genome sequencing. PeerJ, 2016(4), 1752. doi:10.7717/peerj.1752.

[30] Chattaway, M. A., Langridge, G. C., & Wain, J. (2021). Salmonella nomenclature in the genomic era: a time for change. Scientific Reports, 11(1), 7494. doi:10.1038/s41598-021-86243-w.

[31] Chen, S. H., Parker, C. H., Croley, T. R., & McFarland, M. A. (2021). Genus, species, and subspecies classification of salmonella isolates by proteomics. Applied Sciences (Switzerland), 11(9), 4264. doi:10.3390/app11094264.

[32] Pearce, M. E., Langridge, G. C., Lauer, A. C., Grant, K., Maiden, M. C. J., & Chattaway, M. A. (2021). An evaluation of the species and subspecies of the genus Salmonella with whole genome sequence data: Proposal of type strains and epithets for novel S. enterica subspecies VII, VIII, IX, X and XI. Genomics, 113(5), 3152–3162. doi:10.1016/j.ygeno.2021.07.003.

[33] O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Research, 44(D1), D733–D745. doi:10.1093/nar/gkv1189.

[34] Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. Nucleic Acids Research, 41(D1), 36– 42. doi:10.1093/nar/gks1195.

[35] Dong, Y., Sun, F., Ping, Z., Ouyang, Q., & Qian, L. (2020). DNA storage: Research landscape and future prospects. National Science Review, 7(6), 1092–1107. doi:10.1093/nsr/nwaa007.

[36] Emerson, D., Agulto, L., Liu, H., & Liu, L. (2008). Identifying and characterizing bacteria in an era of genomics and proteomics. BioScience, 58(10), 925–936. doi:10.1641/B581006.

[37] Cotter, D. J., Webster, T. H., & Wilson, M. A. (2023). Genomic and demographic processes differentially influence genetic variation across the human X chromosome. PLoS ONE, 18(11 November), 287609. doi:10.1371/journal.pone.0287609.

[38] Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30(14), 2068-2069. doi:10.1093/bioinformatics/btu153.

[39] Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology, 20(1), 238. doi:10.1186/s13059-019-1832-y.

[40] Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME suite. Nucleic acids research, 43(W1), W39-W49. doi:10.1093/nar/gkv416.

[41] Edgar, R. (2024). rcedgar/muscle: C++. Available online: https://github.com/rcedgar/muscle (accessed on November 2025).

[42] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825-2830.

[43] Horng, Y. T., Dewi Panjaitan, N. S., Chang, H. J., Wei, Y. H., Chien, C. C., Yang, H. C., Chang, H. Y., & Soo, P. C. (2022). A protein containing the DUF1471 domain regulates biofilm formation and capsule production in Klebsiella pneumoniae. Journal of Microbiology, Immunology and Infection, 55(6P2), 1246–1254. doi:10.1016/j.jmii.2021.11.005.

[44] Gromova, E. S., & Khoroshaev, A. V. (2003). Prokaryotic DNA Methyltransferases: The Structure and the Mechanism of Interaction with DNA. Molecular Biology, 37(2), 260–272. doi:10.1023/A:1023301923025.

[45] Lyko, F. (2018). The DNA methyltransferase family: A versatile toolkit for epigenetic regulation. Nature Reviews Genetics, 19(2), 81–92. doi:10.1038/nrg.2017.80.

[46] Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T. G., Fan, J., Garmiri, P., da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., … Zhang, J. (2023). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research, 51(D1), D523–D531. doi:10.1093/nar/gkac1052.

[47] Amyes, S. G. B. (2013). 1. Origins. Bacteria, 1–6. doi:10.1093/actrade/9780199578764.003.0001.

[48] Winand, R., Bogaerts, B., Hoffman, S., Lefevre, L., Delvoye, M., Van Braekel, J., Fu, Q., Roosens, N. H. C., De Keersmaecker, S. C. J., & Vanneste, K. (2020). Targeting the 16s rRNA gene for bacterial identification in complex mixed samples: Comparative evaluation of second (illumina) and third (oxford nanopore technologies) generation sequencing technologies. International Journal of Molecular Sciences, 21(1), 298. doi:10.3390/ijms21010298.

[49] Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nature Communications, 10(1), 5029. doi:10.1038/s41467-019-13036-1.

[50] Aanensen, D. M., Feil, E. J., Holden, M. T. G., Dordel, J., Yeats, C. A., Fedosejev, A., Goater, R., Castillo-Ramírez, S., Corander, J., Colijn, C., Chlebowicz, M. A., Schouls, L., Heck, M., Pluister, G., Ruimy, R., Kahlmeter, G., Åhman, J., Matuschek, E., Friedrich, A. W., … Kearns, A. (2016). Whole-genome sequencing for routine pathogen surveillance in public health: A population snapshot of invasive Staphylococcus aureus in Europe. MBio, 7(3), 10 1128 00444–16. doi:10.1128/mBio.00444-16.

[51] Nouioui, I., Carro, L., García-López, M., Meier-Kolthoff, J. P., Woyke, T., Kyrpides, N. C., Pukall, R., Klenk, H. P., Goodfellow, M., & Göker, M. (2018). Genome-based taxonomic classification of the phylum actinobacteria. Frontiers in Microbiology, 9(AUG), 355158. doi:10.3389/fmicb.2018.02007.

[52] Xu, X., He, M., Xue, Q., Li, X., & Liu, A. (2024). Genome-based taxonomic classification of the genus Sulfitobacter along with the proposal of a new genus Parasulfitobacter gen. nov. and exploring the gene clusters associated with sulfur oxidation. BMC Genomics, 25(1), 389. doi:10.1186/s12864-024-10269-3.

[53] He, Y., Shen, Z., Zhang, Q., Wang, S., & Huang, D. S. (2021). A survey on deep learning in DNA/RNA motif mining. Briefings in Bioinformatics, 22(4), 229. doi:10.1093/bib/bbaa229.

[54] Vens, C., Rosso, M. N., & Danchin, E. G. J. (2011). Identifying discriminative classification-based motifs in biological sequences. Bioinformatics, 27(9), 1231–1238. doi:10.1093/bioinformatics/btr110.

[55] Majchrowska, S., Pawłowski, J., Guła, G., Bonus, T., Hanas, A., Loch, A., ... & Drulis-Kawa, Z. (2021). AGAR a microbial colony dataset for deep learning detection. arXiv Preprint, arXiv:2108.01234. doi:10.48550/arXiv.2108.01234.

[56] Kotwal, S., Rani, P., Arif, T., Manhas, J., & Sharma, S. (2022). Automated Bacterial Classifications Using Machine Learning Based Computational Techniques: Architectures, Challenges and Open Research Issues. Archives of Computational Methods in Engineering, 29(4), 2469–2490. doi:10.1007/s11831-021-09660-0.

[57] Wu, Y., & Gadsden, S. A. (2023). Machine learning algorithms in microbial classification: a comparative analysis. Frontiers in Artificial Intelligence, 6. doi:10.3389/frai.2023.1200994.

[58] Khasim, S., Ghosh, H., Rahat, I. S., Shaik, K., & Yesubabu, M. (2024). Deciphering Microorganisms through Intelligent Image Recognition: Machine Learning and Deep Learning Approaches, Challenges, and Advancements. EAI Endorsed Transactions on Internet of Things, 10. doi:10.4108/eetiot.4484.

[59] Ramos-Briceño, D. A., Flammia-D'Aleo, A., Fernández-López, G., Carrión-Nessi, F. S., & Forero-Peña, D. A. (2025). Deep learning-based malaria parasite detection: convolutional neural networks model for accurate species identification of Plasmodium falciparum and Plasmodium vivax. Scientific Reports, 15(1), 3746. doi:10.1038/s41598-025-87979-5.

[60] Fiannaca, A., La Paglia, L., La Rosa, M., Lo Bosco, G., Renda, G., Rizzo, R., Gaglio, S., & Urso, A. (2018). Deep learning models for bacteria taxonomic classification of metagenomic data. BMC Bioinformatics, 19(7), 198. doi:10.1186/s12859-018-2182-6.

[61] Helaly, M. A., Rady, S., & Aref, M. M. (2021). Deep Learning for Taxonomic Classification of Biological Bacterial Sequences. Studies in Big Data, 77, 393–413. doi:10.1007/978-3-030-59338-4_20.

[62] Meharunnisa, M., Sornam, M., & Ramesh, B. (2024). An Optimized Hybrid Model for Classifying Bacterial Genus using an Integrated CNN-RF Approach on 16S rDNA Sequences. Journal of Scientific and Industrial Research, 83(4), 392–404. doi:10.56042/jsir.v83i4.2670.

[63] Arning, N., Sheppard, S. K., Bayliss, S., Clifton, D. A., & Wilson, D. J. (2021). Machine learning to predict the source of campylobacteriosis using whole genome data. PLoS Genetics, 17(10), 1009436. doi:10.1371/journal.pgen.1009436.

[64] Cohen, S., Rokach, L., Motro, Y., Moran-Gilad, J., & Veksler-Lublinsky, I. (2021). minMLST: machine learning for optimization of bacterial strain typing. Bioinformatics, 37(3), 303–311. doi:10.1093/bioinformatics/btaa724.

[65] Wang, L., Tang, J.-W., Li, F., Usman, M., Wu, C.-Y., Liu, Q.-H., Kang, H.-Q., Liu, W., & Gu, B. (2022). Identification of Bacterial Pathogens at Genus and Species Levels through Combination of Raman Spectrometry and Deep-Learning Algorithms. Microbiology Spectrum, 10(6), 258022. doi:10.1128/spectrum.02580-22.

[66] Ren, Y., Zheng, Y., Wang, X., Qu, S., Sun, L., Song, C., Ding, J., Ji, Y., Wang, G., Zhu, P., & Cheng, L. (2024). Rapid identification of lactic acid bacteria at species/subspecies level via ensemble learning of Ramanomes. Frontiers in Microbiology, 15. doi:10.3389/fmicb.2024.1361180.

[67] Kim, E., Yang, S. M., Ham, J. H., Lee, W., Jung, D. H., & Kim, H. Y. (2025). Integration of MALDI-TOF MS and machine learning to classify enterococci: A comparative analysis of supervised learning algorithms for species prediction. Food Chemistry, 462, 140931. doi:10.1016/j.foodchem.2024.140931.

[68] Jeon, Y., Lee, S., Jeon, Y. J., Kim, D., Ham, J. H., Jung, D. H., Kim, H. Y., & You, J. (2025). Rapid identification of pathogenic bacteria using data preprocessing and machine learning-augmented label-free surface-enhanced Raman scattering. Sensors and Actuators B: Chemical, 425, 136963. doi:10.1016/j.snb.2024.136963.

[69] Cserhati, M., Xiao, P., & Guda, C. (2019). K-mer-Based Motif Analysis in Insect Species across Anopheles, Drosophila, and Glossina Genera and Its Application to Species Classification. Computational and Mathematical Methods in Medicine, 2019(1), 4259479. doi:10.1155/2019/4259479.

[70] Liang, Q., Bible, P. W., Liu, Y., Zou, B., & Wei, L. (2020). DeepMicrobes: taxonomic classification for metagenomics with deep learning. NAR Genomics and Bioinformatics, 2(1), 9. doi:10.1093/nargab/lqaa009.

[71] Mock, F., Kretschmer, F., Kriese, A., Böcker, S., & Marz, M. (2022). Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. Proceedings of the National Academy of Sciences of the United States of America, 119(35), 2122636119. doi:10.1073/pnas.2122636119.

[72] W.H.O. (2024). Salmonella (non-typhoidal). World Health Organization (W.H.O.), Geneva, Switzerland. Available online: https://www.who.int/news-room/fact-sheets/detail/salmonella-(non-typhoidal) (accessed on November 2025).

[73] Radomski, N., Cadel-Six, S., Cherchame, E., Felten, A., Barbet, P., Palma, F., Mallet, L., Le Hello, S., Weill, F. X., Guillier, L., & Mistou, M. Y. (2019). A simple and robust statistical method to define genetic relatedness of samples related to outbreaks at the genomic scale – application to retrospective salmonella foodborne outbreak investigations. Frontiers in Microbiology, 10(OCT), 2413. doi:10.3389/fmicb.2019.02413.

[74] Rizzo, R., Fiannaca, A., La Rosa, M., & Urso, A. (2016). A deep learning approach to DNA sequence classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 9874 LNCS, 129–140. doi:10.1007/978-3-319-44332-4_10.

[75] Bhandari, N., Khare, S., Walambe, R., & Kotecha, K. (2021). Comparison of machine learning and deep learning techniques in promoter prediction across diverse species. PeerJ Computer Science, 7, 1–17. doi:10.7717/PEERJ-CS.365.

[76] Zou, X., Nguyen, M., Overbeek, J., Cao, B., & Davis, J. J. (2022). Classification of bacterial plasmid and chromosome derived sequences using machine learning. PLoS ONE, 17(12 December), 279280. doi:10.1371/journal.pone.0279280.

[77] Sharan, R., & Myers, E. W. (2005). A motif-based framework for recognizing sequence families. Bioinformatics, 21(SUPPL. 1), 387– 393. doi:10.1093/bioinformatics/bti1002.

[78] Xiong, H., Capurso, D., Sen, Ś., & Segal, M. R. (2011). Sequence-based classification using discriminatory motif feature selection. PLoS ONE, 6(11), 27382. doi:10.1371/journal.pone.0027382.

[79] Parmar, A., Katariya, R., & Patel, V. (2019). A Review on Random Forest: An Ensemble Classifier. Lecture Notes on Data Engineering and Communications Technologies, 26, 758–763. doi:10.1007/978-3-030-03146-6_86.

[80] Deng, X., Milligan, K., Ali-Adeeb, R., Shreeves, P., Brolo, A., Lum, J. J., Andrews, J. L., & Jirasek, A. (2022). Group and Basis Restricted Non-Negative Matrix Factorization and Random Forest for Molecular Histotype Classification and Raman Biomarker Monitoring in Breast Cancer. Applied Spectroscopy, 76(4), 462–474. doi:10.1177/00037028211035398.

[81] Antonio Eng Lim, P., & Hee Park, C. (2024). A collaborative ensemble construction method for federated random forest. Expert Systems with Applications, 255, 124742. doi:10.1016/j.eswa.2024.124742.

[82] Naser, S., Thompson, F. L., Hoste, B., Gevers, D., Vandemeulebroecke, K., Cleenwerck, I., Thompson, C. C., Vancanneyt, M., & Swings, J. (2005). Phylogeny and identification of enterococci by atpA gene sequence analysis. Journal of Clinical Microbiology, 43(5), 2224–2230. doi:10.1128/JCM.43.5.2224-2230.2005.

[83] Thompson, C. C., Thompson, F. L., Vicente, A. C. P., & Swings, J. (2007). Phylogenetic analysis of vibrios and related species by means of atpA gene sequences. International Journal of Systematic and Evolutionary Microbiology, 57(11), 2480–2484. doi:10.1099/ijs.0.65223-0.

[84] Evseev, P., Lukianova, A., Tarakanov, R., Tokmakova, A., Shneider, M., Ignatov, A., & Miroshnikov, K. (2022). Curtobacterium spp. and Curtobacterium flaccumfaciens: Phylogeny, Genomics-Based Taxonomy, Pathogenicity, and Diagnostics. Current Issues in Molecular Biology, 44(2), 889–927. doi:10.3390/cimb44020060.

[85] Jolley, K. A., Bliss, C. M., Bennett, J. S., Bratcher, H. B., Brehony, C., Colles, F. M., Wimalarathna, H., Harrison, O. B., Sheppard, S. K., Cody, A. J., & Maiden, M. C. J. (2012). Ribosomal multilocus sequence typing: Universal characterization of bacteria from domain to strain. Microbiology, 158(4), 1005–1015. doi:10.1099/mic.0.055459-0.

[86] Frapolli, M., Défago, G., & Moënne-Loccoz, Y. (2007). Multilocus sequence analysis of biocontrol fluorescent Pseudomonas spp. producing the antifungal compound 2,4-diacetylphloroglucinol. Environmental Microbiology, 9(8), 1939–1955. doi:10.1111/j.1462-2920.2007.01310.x.

[87] Kim, C., Oh, K. K., Jothi, R., & Park, D. S. (2024). An innovative approach to decoding genetic variability in Pseudomonas aeruginosa via amino acid repeats and gene structure profiles. Scientific Reports, 14(1), 22610. doi:10.1038/s41598-024-73031-5.

[88] Reichler, S. J., Murphy, S. I., Martin, N. H., & Wiedmann, M. (2021). Identification, subtyping, and tracking of dairy spoilage-associated Pseudomonas by sequencing the ileS gene. Journal of Dairy Science, 104(3), 2668–2683. doi:10.3168/jds.2020-19283.

Appendix I: Supplementary Figures

Figure S1_Sequence_Encoding; The figure displays the transformed format of DNA sequences into numerical values (A=1, T=2, C=3, G=4, N=5) alongside the label-encoded 'Subspecies' column data. Both transformations illustrate the conversion of categorical data into numerical format required for machine learning analysis.

	Row Header	Subspecies	OG0001278_motif_2	OG0001278_motif_3	OG0001278_motif_4	OG0001278_motif_5	OG0001279_motif_1	OG0001279_motif_3
0	Agona_1	0	[1, 2, 4, 2, 4, 3, 4, 3, 2, 1, 3, 2, 1, 3, 2, ...	[3, 4, 2, 4, 3, 4, 2, 3, 2, 4, 2, 1, 3, 4, 2, ...	[1, 4, 3, 3, 1, 3, 2, 3, 4, 4, 2, 4, 1, 2, 2, ...	[1, 3, 3, 4, 4, 3, 1, 3, 3, 4, 1, 2, 3, 2, 2, ...	[2, 4, 3, 2, 3, 1, 2, 2, 1, 1, 4, 2, 2, 2, 2, ...	[2, 2, 1, 1, 2, 4, 1, 2, 2, 3, 1, 4, 4, 3, 4, ...
1	Agona_10	0	[1, 2, 4, 2, 4, 3, 4, 3, 2, 1, 3, 2, 1, 3, 2, ...	[3, 4, 2, 4, 3, 4, 2, 3, 2, 4, 2, 1, 3, 4, 2, ...	[1, 4, 3, 3, 1, 3, 2, 3, 4, 4, 2, 4, 1, 2, 2, ...	[1, 3, 3, 4, 4, 3, 1, 3, 3, 4, 1, 2, 3, 2, 2, ...	[2, 4, 3, 2, 3, 1, 2, 2, 1, 1, 4, 2, 2, 2, 2, ...	[2, 2, 1, 1, 2, 4, 1, 2, 2, 3, 1, 4, 4, 3, 4, ...
2	Agona_11	0	[1, 2, 4, 2, 4, 3, 4, 3, 2, 1, 3, 2, 1, 3, 2, ...	[3, 4, 2, 4, 3, 4, 2, 3, 2, 4, 2, 1, 3, 4, 2, ...	[1, 4, 3, 3, 1, 3, 2, 3, 4, 4, 2, 4, 1, 2, 2, ...	[1, 3, 3, 4, 4, 3, 1, 3, 3, 4, 1, 2, 3, 2, 2, ...	[2, 4, 3, 2, 3, 1, 2, 2, 1, 1, 4, 2, 2, 2, 2, ...	[2, 2, 1, 1, 2, 4, 1, 2, 2, 3, 1, 4, 4, 3, 4, ...
3	Agona_12	0	[1, 2, 4, 2, 4, 3, 4, 3, 2, 1, 3, 2, 1, 3, 2, ...	[3, 4, 2, 4, 3, 4, 2, 3, 2, 4, 2, 1, 3, 4, 2, ...	[1, 4, 3, 3, 1, 3, 2, 3, 4, 4, 2, 4, 1, 2, 2, ...	[1, 3, 3, 4, 4, 3, 1, 3, 3, 4, 1, 2, 3, 2, 2, ...	[2, 4, 3, 2, 3, 1, 2, 2, 1, 1, 4, 2, 2, 2, 2, ...	[2, 2, 1, 1, 2, 4, 1, 2, 2, 3, 1, 4, 4, 3, 4, ...
4	Agona_13	0	[1, 2, 4, 2, 4, 3, 4, 3, 2, 1, 3, 2, 1, 3, 2, ...	[3, 4, 2, 4, 3, 4, 2, 3, 2, 4, 2, 1, 3, 4, 2, ...	[1, 4, 3, 3, 1, 3, 2, 3, 4, 4, 2, 4, 1, 2, 2, ...	[1, 3, 3, 4, 4, 3, 1, 3, 3, 4, 1, 2, 3, 2, 2, ...	[2, 4, 3, 2, 3, 1, 2, 2, 1, 1, 4, 2, 2, 2, 2, ...	[2, 2, 1, 1, 2, 4, 1, 2, 2, 3, 1, 4, 4, 3, 4, ...

Figure S1. Nucleotide sequence encoding scheme and subspecies label encoding representation.

Figure S2_16S_rRNA_Phylogenetic_Tree; This figure illustrates the phylogenetic relationships based on *16S rRNA* analysis. The circular tree visualization demonstrates the genetic diversity and evolutionary distances among *Salmonella enterica* serovars. The clade containing strain 92-0392 (highlighted in red) shows close genetic relationship with Typhimurium strains, providing additional genomic evidence for serovar classification.

Figure S2. Phylogenetic Tree Analysis Based on 16S Ribosomal RNA Sequences

Supp S1_Whole Genome Phylogenetic Analyses of Salmonella Genomes

The Supp S1 contains the phylogenetic trees generated using both Neighbour-Joining (distance-based) and Maximum-Likelihood (character-based) methods for the 240 complete *Salmonella enterica* genomes used in this research. The phylogenetic analysis revealed incongruent classification for several *Salmonella* strains, including Enteritidis_14, Salamae_12, Salamae_13, Salamae_15, Montevideo_1, Montevideo_6, and Montevideo_13 (labelled in red) between the two tree-building approaches. These results illustrate the effect of noise or inconsistencies in the genomic data on the *Salmonella* strain classification.

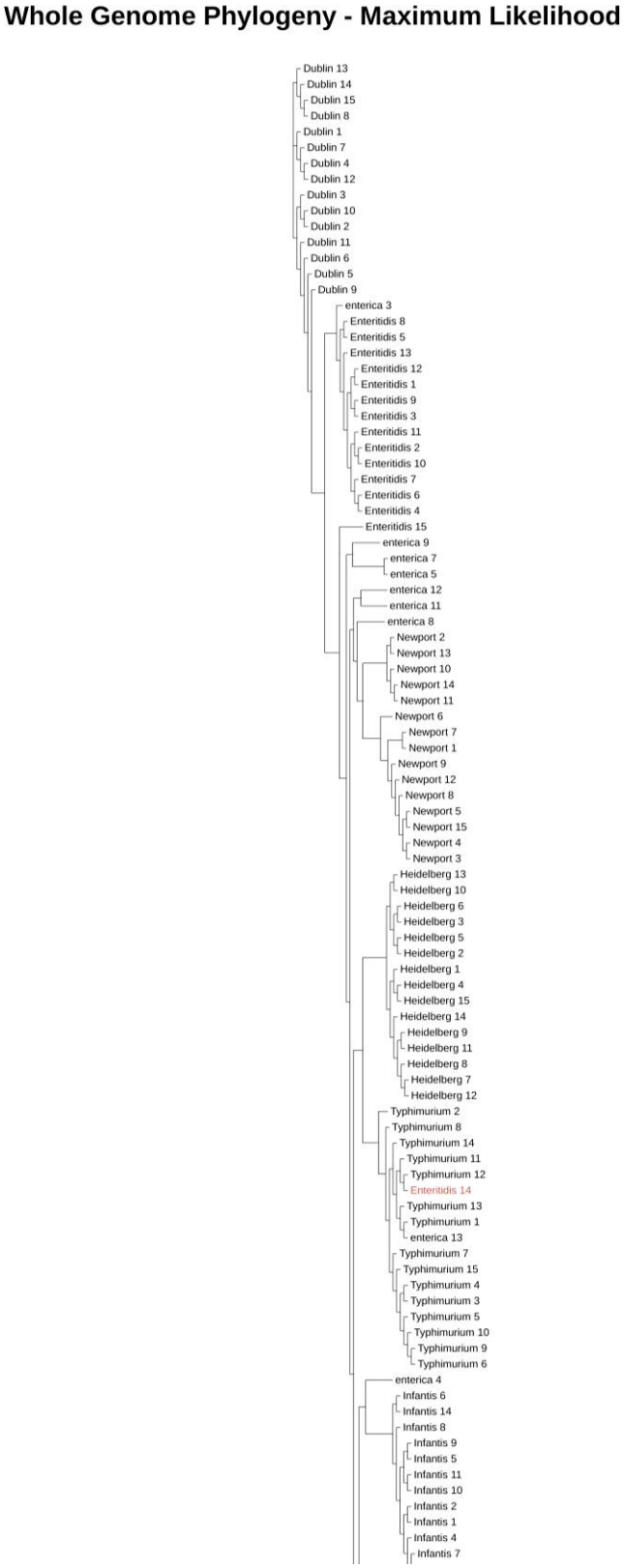

Figure Supp S1 1. Whole Genome Phylogenetic Tree by using Maximum-Likelihood

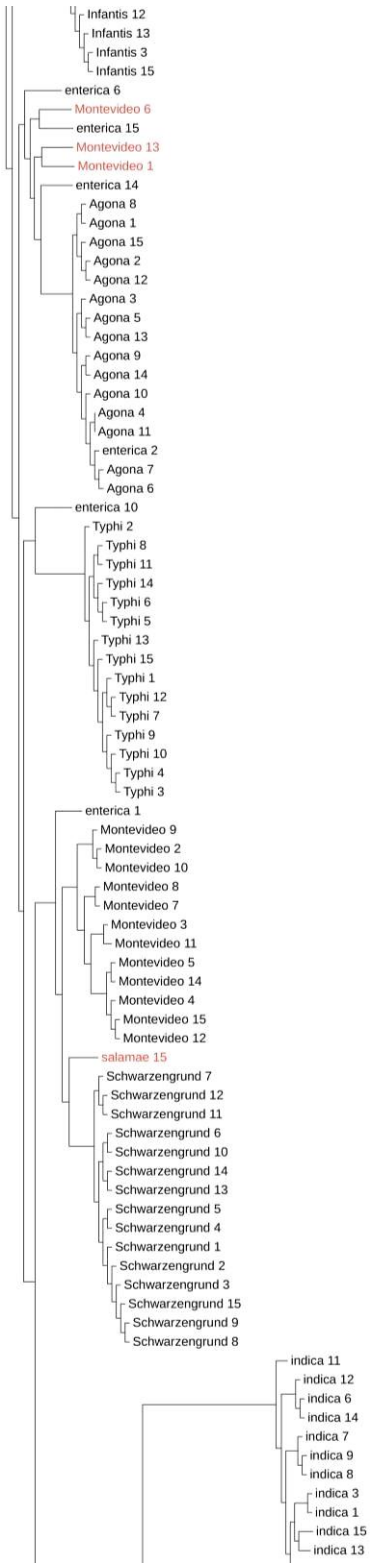

Figure Supp S1 1. Whole Genome Phylogenetic Tree by using Maximum-Likelihood (Continued)

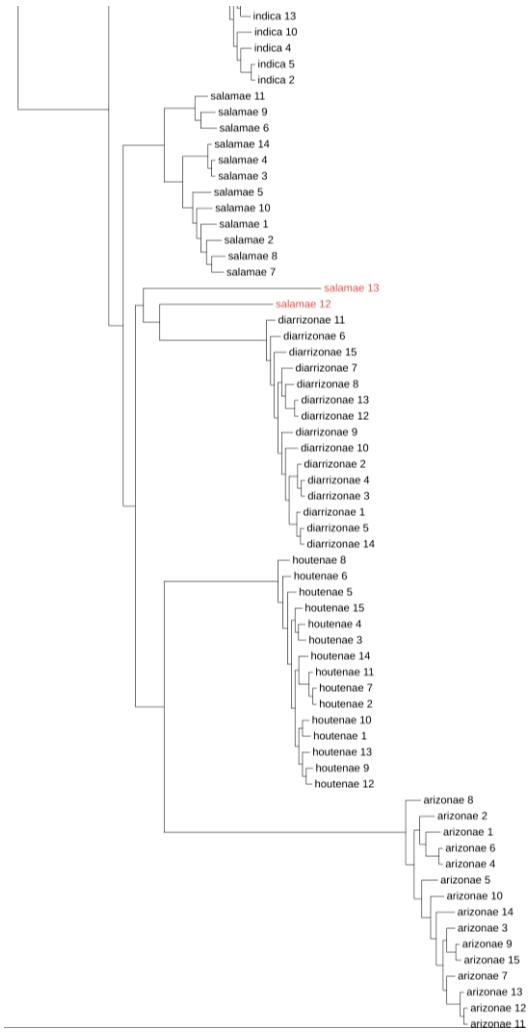

Figure Supp S1 1. Whole Genome Phylogenetic Tree by using Maximum-Likelihood (Continued)

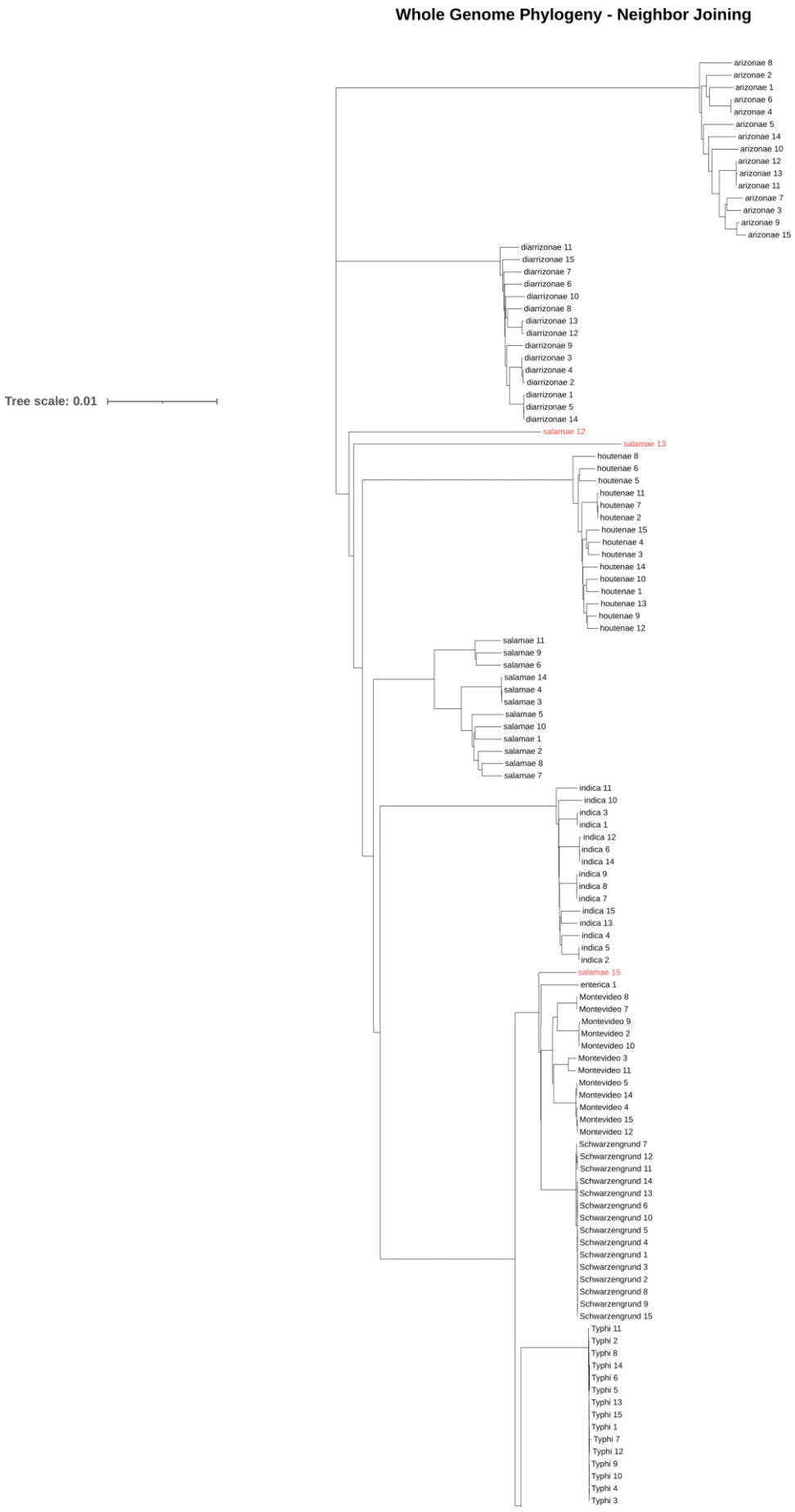

Figure Supp S1 2. Whole Genome Phylogenetic Tree by using Neighbour-Joining

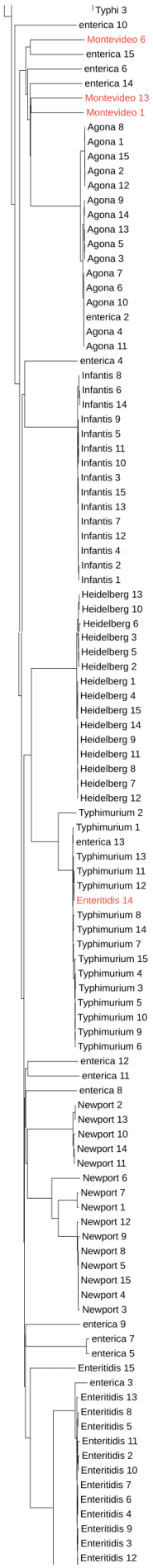

Figure Supp S1 2. Whole Genome Phylogenetic Tree by using Neighbour-Joining (Continued)

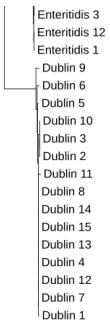

Figure Supp S1 2. Whole Genome Phylogenetic Tree by using Neighbour-Joining (Continued)

Supp S2_Taxonomic Verification of Misclassified Strains

To gain a more comprehensive understanding of the classification, a detailed *16S rRNA* analysis of all samples included in this research was performed. The genomic data from all analysed samples were combined into a single file, which was then converted into a searchable database using the BLAST function *makeblastdb*. Reference *16S rRNA* sequences from the *Salmonella* Typhimurium LT2 strain were downloaded (NCBI Reference Sequence: NR_074910.1) and used to conduct a BLAST search against this database. This process facilitated the extraction of the *16S rRNA* sequences from all samples involved in the study. Subsequently, a phylogenetic tree was constructed using W-IQ-TREE based on the obtained *16S rRNA* sequences. The resulting tree was visualized using the Interactive Tree of Life (ITOL) tool, providing a clear representation of the genetic relationships among the strains. This structured methodology enabled a thorough examination of the genetic relationships between the misclassified strain and other *Salmonella* strains in the study, offering valuable insights into their taxonomic classifications. This phylogenetic analysis provided a clear representation of the genetic relationships among the strains, further supporting the classification of strain 92- 0392 within the Typhimurium clade. Additionally, the BLAST analysis of the 16S ribosomal RNA region (genome coordinates 1113643...1115184) of strain 92-0392 showed the strongest matches to serovar Typhimurium sequences. The comprehensive genetic analyses presented in the Supplementary Materials section, in conjunction with the insights discussed in the main text, provide a robust validation of the machine learning-based classification of the misclassified strain.

Appendix II: Supplementary Tables

You can download this Supplementary **Here**.