Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 1 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i01.52521

PAPER

ARGai 2.0: A Feature Engineering Enabled Deep Network Model for Antibiotic Resistance Gene and Strain Identification in *E. coli*

Debasish Swapnesh Kumar Nayak[1](✉), Arpita Priyadarshini[2], Sweta Padma Routray[3], Santanu Kumar Sahoo[4], Tripti Swarnkar[5]

[1]Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India

[2]Department of Statistics, Utkal University, Bhubaneswar, Odisha, India

[3]Centre for Biotechnology, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India

[4]Department of Electronics and Communication Engineering, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India

[5]Department of Computer Application, National Institute of Technology, Raipur, Chhattisgarh, India

debasish.nayak@cutm.ac.in

ABSTRACT

Escherichia coli (E. coli) is a group of bacteria that cause infections in the gastrointestinal (GI) tract and urinary tract (UTIs). The rise in antimicrobial resistance (AMR) due to antibiotic-resistant genes (ARGs) linked to *E. coli* strains that cause UTIs poses a significant threat. Identifying ARGs and resistant strains is crucial for effective treatment. To identify ARGs and classify resistant strains in *E. coli* utilizing gene expression (GE) data with advanced computational techniques such as feature engineering and transfer learning (TL). In TL, knowledge acquired by the baseline model is transferred to the target domain (BI-LSTM-GRU). ARGai 2.0 utilizes the Synthetic Minority Over-sampling Technique (SMOTE) for over-sampling the GE dataset to evaluate the effectiveness of the proposed TL framework. Our proposed ARGai 2.0 model achieved a higher classification accuracy of 14% compared to 1D CNN and 11% compared to BI-LSTM-GRU individually. The analysis revealed that genes associated with the nitrate reductase operon (narU, narV, narW, narY, and narZ) exhibit high connectivity and interaction scores, indicating their central role in nitrate metabolism. This aligns with the high enrichment FDR of 3.07E-10 and fold enrichment of 229.28 for pathways related to nitrate reductase complex and nitrite transmembrane transporter activity. ARGai 2.0 successfully detected the significant genes responsible for antibiotic resistance and classified the resistant strains. The gene network analysis highlights the central role of nitrate metabolism genes, while peripheral genes like ansP and yncG are involved in more specialized functions.

KEYWORDS

antimicrobial resistance, antibiotic resistant genes, deep learning, transfer learning, SMOTE, urine tract infections (UTIs)

Nayak, D.S.K., Priyadarshini, A., Routray, S.P., Sahoo, S.K., Swarnkar, T. (2025). ARGai 2.0: A Feature Engineering Enabled Deep Network Model for Antibiotic Resistance Gene and Strain Identification in *E. coli*. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(1), pp. 76–96. https://doi.org/10.3991/ijoe.v21i01.52521

Article submitted 2024-09-28. Revision uploaded 2024-11-05. Final acceptance 2024-11-05.

1 INTRODUCTION

Infectious diseases have consistently posed significant risks to global health, financial stability, and social structures. In terms of mortality and morbidity, infectious diseases have the most apparent and immediate impacts. The threat is exacerbated by antimicrobial resistance (AMR) in pathogenic bacteria, leading to prolonged hospitalizations, elevated medical expenses, and increasing mortality rates, while complicating the treatment of prevalent ailments [1, 2]. More than 1.2 million fatalities were directly related to AMR in 2019 [3, 4]. An estimated 10 million individuals could lose their lives to AMR), sometimes called the "Silent Pandemic," by 2050 if current trends continue. If this occurs, it could surpass all other global killers in terms of sheer volume [5]. World health is at threat because of antibiotic resistance in several bacteria, including *Escherichia coli (E. coli)*, which causes urinary tract infections (UTIs). The early identification of resistant strains and antibiotic-resistant genes (ARGs) needs to be addressed more effectively. In a study it was found that *E. coli* was the predominant causal agent, accounting for 64.29% of urinary tract infections in the entire study cohort [6]. In non-*E. coli* urinary tract infections, the predominant causal bacteria were *Klebsiella pneumoniae* (16.43%), succeeded by *Pseudomonas aeruginosa*, *Enterobacter*, *Proteus mirabilis*, and *Enterococcus* [7–9].

The traditional approaches for AMR analysis, particularly the ARG identification, are complex, time-consuming, and prone to errors [10]. To increase antimicrobial resistance detection through better genomic data processing, different machine learning (ML) and deep learning (DL) models are employed [11, 12]. To uncover resistance genes in *E. coli* and gain knowledge about the behavior and variability of ARGs, aiGeneR 1.0, an AI model, is being employed [11]. This advancement contributes to the development of efficient diagnostic tools and individualized therapies, which considerably improve the early diagnosis and management of resistant strains.

In a recent study, it was found that using ML with whole-genome sequencing and gene-sharing network analysis may efficiently find broad networks of ARGs shared by *E. coli* populations across animals, people, and the environment in livestock farming [13]. In recent years, ML and DL models have been effectively applied to gene expression data (GE) to predict outcomes, particularly in cancer; however, research into discovering ARGs using advanced methods is restricted. These developments, together with AI models, can help in the identification of gene and strain classification of infectious diseases.

In this work, we aim to use effective feature selection strategies (ensemble approach) to identify ARGs and adopt the potential of transfer learning techniques to classify the resistant strains utilizing the NGS gene expression data. We hypothesized that data augmentation, ensemble feature selection, and TL improve classification accuracy. We also hypothesized that fine-tuning TL architecture along with ensemble features reduces the computational cost. In addition to this, we aim that the genes identified by our proposed model (ARGai 2.0) have more biological significance towards AMR. Following is a summary of the paper's structure and its main points:

The relevant AMR analysis work is presented in Section 2. Section 3 delves into the materials and processes, while Section 4 lays out the blueprints for our projected ARGai 2.0. Section 5 showcases the assessments of performance. Section 6

delves into the results of our proposed model, while section 7 showcases the ROC of the models we examined. Our ARGai 2.0 model's biological validation and the training size effect are covered in Sections 8 and 9, respectively. The limits, benchmarking, and special notes are contained in Section 10. In Section 11, we go over our concluding remarks.

2 RELATED WORK

In recent years, significant progress has been achieved through AMR research using ML and DL approaches, including the identification of ARGs using pre-trained protein language models such as PLM-ARG [9]. This technique refines ARG identification by utilizing large protein datasets, increasing precision for both known and unknown resistance genes, and aligning with uniting-centered pan-genome strategies. Applying ML to a pan-genome graph improves the prediction of antibiotic resistance and the identification of new genes in infectious diseases [14]. DL algorithms, such as DeepARG, increase the detection of ARGs in metagenomic data by providing higher accuracy, precision, and recall than conventional approaches [15]. DeepARG broadens the detection range of ARGs by using specialized models for whole gene sequences. Advanced computational techniques are used to combat antibiotic resistance with the help of developments in models like Hyper VR and ARGNet [16, 17]. In this section, we discussed a few other existing AI models and tools for ARG identification and resistant strain classification.

Pei et al. [17] developed ARGNet, a deep neural network model that utilizes gene sequence (GS) data to identify antibiotic-resistant genes (ARG) without sequence alignment. With a remarkable 95% accuracy rate, ARGNet performed better than DeepARG and HMD-ARG. Tharmakulasingam et al. [18] analyzed the GS data on antibiotic resistance in urinary tract infection (AMR-UTI) and proposed a 1D-Transformer model that performed better than traditional ML models, such as logistic regression (LR), ResNet, and 1-Dimensional convolutional neural network (1-D CNN), with a 10% higher AUC, to identify features responsible for AMR.

Nsubuga et al. [19] used data on *E. coli* strains from Africa and England to examine several ML models, including LR, random forest (RF), gradient boost (GB), support vector machine (SVM), Cat Boost (CB), eXtreme gradient boosting (XG Boost), feed-forward neural network (FF-NN), and Light GBM (LGBM). It was discovered that the specific antibiotic used had a substantial impact on the model's performance. With 87% accuracy and 81% precision for ciprofloxacin (CIP), Cat Boost performed exceptionally well, but SVM performed best for cefotaxime (CTX) with 100% accuracy and 92% precision, indicating that the model's efficacy is extremely antibiotic-specific.

Ji et al. [16] utilized GS data and proposed Hyper VR, a hybrid deep ensemble learning approach that simultaneously predicts virulence factors and ARGs. It was observed that HyperVR performed better than the existing tools to improve pathogen monitoring and epidemic identification by increasing accuracy and reliability without depending on rigid cutoff levels.

Chung et al. [20] analyzed mass spectrometry data and proposed a matrix-assisted laser desorption ionization–time-of-flight mass spectrometry (MALDI-TOF MS) framework for the identification of ARGs in *E. coli* by utilizing different ML models like LR, RF, XG Boost, and SVM. It was observed that XGBoost performed better than other models with an AUC value ranging from 62% to 87% for different antibiotics.

Creating a prediction model that can adequately cover all isolates is a major challenge due to the tremendous variety of the *E. coli* population.

Al-Shaebi et al. [21] analyzed the Stanford data and proposed a U-net model for accurately identifying bacteria and ARG. It was found that the U-net model, when combined with Raman spectroscopy, achieved an accuracy of 95% compared to the multi-scale and ResNet models. The limitation includes low accuracy because of information loss during model training.

A considerable gap exists through the assessment of ML and DL models for the identification of genes, especially concerning computing cost, where advanced technologies have a higher cost than traditional approaches, model complexity, data availability, effective feature selection techniques, and accessibility of annotated ARGs, and prediction algorithms' accuracy in finding resistant bacteria needs improving. The gene expression data are complex and non-linear; thus, a novel feature selection approach is very essential and needs to be generalized and robust. Incorporating these limitations, our main objective is to design a complete AI pipeline (ARGai 2.0) that will identify the ARGs and resistant strains. In addition to this, our ARGai 2.0 integrates advanced computational intelligence techniques, especially to address the efficacy of the ensemble feature selection strategies with DL classification models.

3 MATERIALS AND METHODS

In this section, we discussed the datasets utilized in our study, their characteristics, and the selected AI models that are deployed on the dataset for the identification of resistant strains in *E. coli*.

3.1 Dataset description

The studied datasets belong to *E. coli*, high-throughput (NGS) GE data. The dataset used in our study was collected from the National Centre for Biotechnology Information (NCBI) [22]. Our studied dataset with an accession number GSE96706 has a total of 4501 genes, with 73 cases as resistant and 11 susceptible. In our preprocessing phase (normalization and imputation), the final data is of 4493 genes and 84 strains. We adopt min-max normalization and mean data imputation techniques in which 9 genes are removed due to more than 30% null values. After applying the SMOTE over-sampling techniques, we made the final dataset with 350 over-sampled cases in the training data; a detailed summary is in Table 1.

Table 1. Description of the studied dataset

Description	#Susceptible	#Resistant	Total
Samples	11	73	48
Genes	–	–	4493
Training data	7	58	65
Train augmentation	150	200	350
Testing data	04	15	19

3.2 SMOTE

The synthetic minority over-sampling technique (SMOTE) [23] is a method for data augmentation that generates synthetic samples for the minority class to mitigate class imbalance. Conventional SMOTE generates synthetic instances through interpolation of existing samples; however, due to the high-dimensional nature of GE data, careful consideration is necessary when implementing this method. SMOTE employs a carefully designed algorithm to ensure that synthetic samples are strategically located within the feature space, thereby preserving the essential expression patterns and interactions among genes.

3.3 Ensemble feature selection

Ensemble feature selection improves accuracy and resilience by integrating various approaches for detecting varied gene patterns in high-dimensional GE data. It reduces overfitting, resulting in more accurate and thorough feature selection. Ensemble approaches produce more consistent and effective outcomes by identifying physiologically significant genes and dealing with nonlinear patterns. This method is very useful for GE data, which is typically complicated and noisy. Finally, ensemble feature selection enhances the performance of classification and prediction tasks in such datasets.

In our previous work, we have adopted a single feature selection technique, and our proposed approach identifies various significant genes with stand-alone feature selection models [11, 24]. However, in this work, we aim to explore the efficacy of the ensemble feature selection approach against stand-alone feature selection strategies to identify significant genes, ARGs, and contributions towards model matrices. We also aim to perform the biological validation on our model-selected genes.

3.4 Random forest

Random forest produces reliable significance ratings and can handle high-dimensional, complicated datasets [25]. Its capacity to capture nonlinear links and interactions makes it useful for detecting key traits. The feature importance score $I(f_i)$ for a feature f_i is calculated by summing the reduction in the Gini impurity (or another impurity measure) across all the nodes in the forest where that feature is used to split the data. Eq. 1 represents the importance of the feature f_i as below,

$$I(f_i) = \frac{1}{T}\sum_{t=1}^{T}\sum_{n \epsilon N_t}\Delta I_n(f_i) \tag{1}$$

Where:

- T is the total number of trees in the random forest.
- N_t is the set of all nodes in tree t.
- $\Delta I_n(f_i)$ is the decrease in impurity at node n caused by splitting on the feature f_i.
- $I(f_i)$ is the importance of feature.

3.5 eXtreme gradient boosting (XGBoost)

XG Boost has strong significance metrics and can handle non-linear connections effectively [11, 26]. It is resistant to overfitting and ranks features based on their contribution to model performance. The Eq. 2 represents gain-based feature importance as below,

$$I_{XGB}(f_i) = \sum_{t=1}^{T} \sum_{n \epsilon N_t} \Delta Gain_n(f_i) \tag{2}$$

Where:

- T is the total number of trees in the XG Boost model.
- N_t is the set of all nodes in the tth tree.
- $\Delta Gain_n$ is the improvement in the loss function (or decrease in error) at node n due to splitting on feature f_i.
- $I_{XGB}(f_i)$ gain-based feature importance.

3.6 Polynomial features

Polynomial features (PF) are used to represent non-linear correlations and interactions between features [27]. GE data are non-linear, and finding a correlation among those features is very complex. So, PF helps to extract those correlations among those features. By broadening the feature space to incorporate polynomial and interaction terms, they enable models to detect complicated patterns that linear features alone may miss, improving the overall capacity to choose relevant features in conjunction with other techniques [28]. Given an original feature vector $X = x_1, x_2, \ldots\ldots, x_n$ and a polynomial degree d (for our experiment it is 2), the PF transformation $P(X)$ is shown in Eq. 3,

$$P(X) = \left[1 + x_1, x_2, \ldots\ldots x_n, x_1^2 x_1 x_2, \ldots x_n^d\right] \tag{3}$$

To assess the importance of polynomial features, use a model (like linear regression or decision trees) and evaluate feature importance based on model coefficients or feature importance.

For a linear regression model, the importance of a polynomial feature f_i can be computed as shown in Eq. 4,

$$I_p(f_i) = \left|\beta_i\right| \tag{4}$$

Where:

- β_i is the coefficient of the PF in the regression model.
- $|\beta_i|$ represents the absolute value of the coefficient, indicating the feature's f_i contribution to the model.

We ensure a thorough and robust selection process for optimal model performance with our ensemble feature selection strategy, which uses polynomial feature expansion to capture non-linear relationships in the data and combines feature

importance ranking from RF and XG Boost to determine and give preference to the most relevant features. The feature relevance was assessed using XG Boost with gain-based metrics and RF with Gini impurity. An ensemble technique used these relevance scores to rank and choose the most significant characteristics. The chosen characteristics were then utilized to train and evaluate final models, and performance was measured using common metrics including accuracy, precision, and recall.

4 PROPOSED METHODOLOGY

In this section, we discussed the architecture of our proposed pipeline and the functionality of individual steps associated with ARGai 2.0. Transfer learning is the enhancement of learning in a new model by the transfer of information from a previously learned model [29]. In our proposed architecture, ARGai 2.0, we adopted the TL methodology with the two most widely used models, namely 1-D CNN and modified LSTM, for GE data analysis [30, 31].

4.1 Data preparation

The raw dataset comprises a substantial quantity of labeled data that is utilized to train a DL algorithm. In our data preparation step, SMOTE was used to oversample seven positive instances to 150 and 58 positive cases to 200, handling duplicates to increase dataset variety without duplication. Oversampling, such as from seven to 150 cases, is generally appropriate, but model performance must be monitored to ensure that these additional samples improve learning and generalization rather than introduce noise or overfitting. There are 200 resistant and 150 susceptible strains in the over-sampled training set, which is used for our model training. The description of raw data and SMOTE over-sampled data in the training set of our studied model can be visualized in Figure 1.

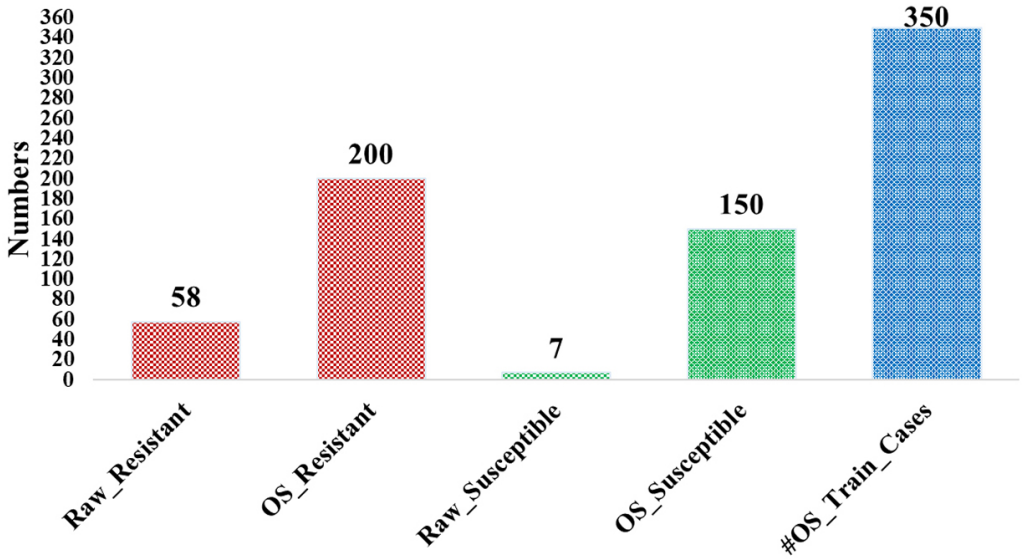

Fig. 1. Description of GSE96706 dataset

Notes: OS: oversampling; #: total.

4.2 Proposed ARGai 2.0 model

The proposed ARGai 2.0 model includes a robust pipeline that uses GE data to improve the analysis and discovery of resistant genes. The process begins with dataset preparation, which includes data augmentation and feature engineering, followed by feature selection using techniques RF, XG Boost, and PF for the identification of significant genes as shown in Figure 2. The processed data is separated into three subsets: training, validation, and testing. A 1D convolutional neural network (1D-CNN) is initially trained on the training data and is validated. In 1D-CNN, many convolutional layers collect discriminative features layer by layer, whereas fully connected layers connect these characteristics to their source labels. After the source model (1-D CNN) is trained, a portion of its architecture, including the learned weights, is frozen and transferred to the target domain (BI-LSTM-GRU). To adjust the base model to the target labels, another DL model (BI-LSTM-GRU) is added to the target model, which is frequently made up of thick layers [32]. The information from this base model is then transferred to the BI-LSTM-GRU model adopting the TL concept, which is fine-tuned on the target dataset. Finally, the ARGai 2.0 model's performance is evaluated by the performance matrices to ensure that it accurately identifies significant genes.

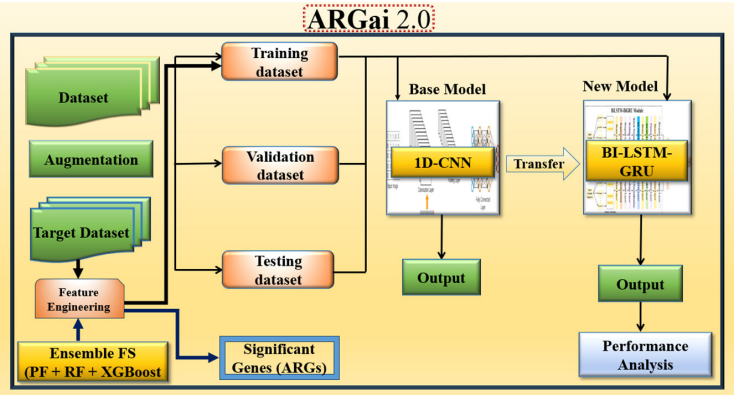

Fig. 2. The pipeline of our proposed **ARGai** 2.0 model

Our ARGai 2.0 model is the feature selection and classifies sequences using a 1D CNN and a bidirectional LSTM-GRU network to discover antibiotic resistance strains. Preprocessing and feature engineering begin with 'Standard Scaler' normalizing features, PF capturing second-degree interactions, and our proposed ensemble feature selection approach reducing dimensionality while maintaining 95% variance. The reshaped data is fed into a 1D CNN with four convolutional layers, 'ReLU' activation, 'Batch Normalization,' and 'max-pooling' for 'down sampling.'

The 1D CNN model consists of four Conv1D layers with filter sizes of 32, 64, 64, and 64, and a kernel size of 3. This is followed by Batch Normalization and MaxPooling1D layers. The weights of the third Conv1D layer, which contains 64 filters, utilizes ReLU activation, and employs 'same' padding, have been frozen. Freezing this layer preserves its pre-learned feature extraction capabilities, enabling it to maintain the capture of intermediate-level features without undergoing updates during training. This method enhances the stability of the feature extraction process from this layer, minimizes the number of trainable parameters, and directs training efforts towards the remaining layers, especially the final Conv1D layer and the Bidirectional LSTM-GRU classifier. This strategy is designed to improve training efficiency and may enhance generalization by utilizing stable intermediate representations while adjusting higher-level features to align with the specific dataset.

To avoid overfitting, the CNN flattens hierarchical features and passes them through a fully connected 128-unit dense layer with dropout. Reshaped features are supplied into a 64-unit bidirectional LSTM to capture sequential dependencies, followed by a 32-unit GRU layer for temporal modeling. A dense 64-unit layer with ReLU activation precedes the output, and a sigmoidal activation layer classifies binary sequences. The Adam optimizer with a 0.001 learning rate and binary 'cross-entropy loss' function compiles the model. Early halting and learning rates decrease callbacks and increase training efficiency and prevent overfitting. The experimental setup of our proposed models with several parameter fine-tunings can be visualized in Table 2.

Table 2. The key parameters tuned in our ARGai 2.0 model

Components of ARGai 2.0	Parameter	Value
CNN Layers	Number of Conv1D Layers	4
	Conv1D Filters	32, 64, 64, 64
	First Conv 1D Layer	Kernel = 3, Weight range = min 1, max 1, min_bias 1, max_bias 1
	First Conv 1D Layer	Kernel = 3, Weight range = min 2, max 2, min_bias 2, max_bias 2
	Kernel Size	3
	Activation Function	ReLU
	Pooling	MaxPooling (Pool Size = 2)
	Dropout Rate (CNN)	0.5
Dense Layer (CNN)	Number of Units	128
	Dropout Rate	0.5
LSTM-GRU Layers	Bidirectional LSTM Units	64
	GRU Units	32
	Dropout Rate (LSTM-GRU)	0.5
Dense Layer (LSTM-GRU)	Number of Units	64
	Activation Function	ReLU
Output Layer	Activation Function	Sigmoid
Optimizer	Optimizer	Adam
	Learning Rate	0.001
Loss Function		Binary Crossentropy
Early Stopping	Patience	5
Reduce LR on Plateau	Factor	0.5
	Patience	3
Training	Epochs	100
	Batch Size	32

5 PERFORMANCE EVALUATION

Several of the criteria that we used to evaluate performance included the utilization of true positive (TP): an antibiotic-resistant strain is correctly classified, true negative (TN): a non-antibiotic-resistant strain is correctly classified, false

positive (FP): an antibiotic-resistant strain is incorrectly classified as susceptible, and false negative (FN): a susceptible strain is incorrectly classified into a resistant strain. For this study, the evaluation metrics that were applied were the F1-score (Ƒ), specificity (SP), sensitivity (SN), Matthew's correlation coefficient (MCC), and accuracy (ɳ) as shown in Eq. 5–9. The diagnostic potential was evaluated by calculating the area under the curve (AUC) and doing an analysis of the receiver operating characteristic curve (ROC).

$$\eta = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$SP = \frac{TN}{TN + FP} \tag{6}$$

$$SN = \frac{TP}{TP + FN} \tag{7}$$

$$F = 2 * \frac{SN * \frac{TP}{TP + FN}}{SN + \frac{TP}{TP + FN}} \tag{8}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{9}$$

6 RESULTS AND DISCUSSION

All the studied models' testing and training were carried out on a high-end computational system using Python [11]. All our experiments were carried out with a system having an Ubuntu 20.04 operating system, 32 GB of RAM, and a 1 TB SSD.

Our result section is categorized into three important parts: (i) Model evaluation with raw data, (ii) Effect of SMOTE over-sampling, and (iii) SMOTE and ensemble feature engineering (Stand-alone DL model vs. transfer learning).

i) **Model evaluation with raw data:** We evaluate our studied models on raw data to assess the efficacy of the baseline models and our proposed ARGai 2.0 model. Due to the null values, non-linear data, and very high-value range, the performance of the models is impacted and results in minimal model metrics. In our experimental environment, we observed the highest classification accuracy of 73% with our proposed ARGai 2.0, followed by BI-LSTM-GRU and 1-D CNN. Both these models achieve a classification accuracy of 72%. Notably, the sensitivity of these two models compared to ARGai 2.0 is high, whereas the specificity is very poor, as shown in Table 3 and can be visualized in Figure 3.

Table 3. Performance metrics of studied models on raw data

Models	Sen	Spe	F-1	MCC	AUC	Acc
1-D CNN	96	08	83	10	58	72
BI-LSTM-GRU	90	25	82	19	44	72
ARGai 2.0	81	40	83	20	64	73

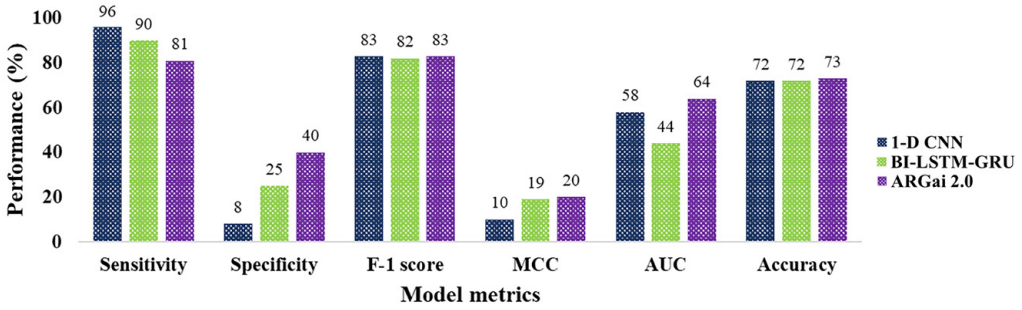

Fig. 3. Performance metrics of all the studied models on raw data

ii) Effect of SMOTE: We assess our examined models on SMOTE-augmented data to evaluate the effectiveness of the baseline models and our proposed ARGai 2.0 model. Following the application of SMOTE to rectify the unbalanced dataset and the elimination of null values in preprocessing, the models' performance was markedly enhanced. In our experimental setting, we recorded a peak classification accuracy of 79% with our suggested ARGai 2.0 model, succeeded by BI-LSTM-GRU at 76% and 1-D CNN at 78%. Although BI-LSTM-GRU and 1-D CNN had similar sensitivity, their specificities were inferior to those of ARGai 2.0, as shown in Table 4.

Table 4. Performance metrics of studied models on SMOTE augmented data

Models	Sen	Spe	F-1	MCC	AUC	Acc
1-D CNN	89	40	81	11	65	78
BI-LSTM-GRU	90	45	84	39	67	76
ARGai 2.0	92	67	92	59	79	79

iii) SMOTE and ensemble feature engineering: The outcome of our complete pipeline setup is described in this section. The ensemble feature engineering approach adopted in our ARGai 2.0 pipeline boosts the performance of the models significantly. The classification accuracy obtained by our ARGai 2.0 is 97%, which is 24% and 18% more compared to its operation over raw data and SMOTE-augmented data without feature selection, respectively. Similarly, the stand-alone models, 1-D CNN and BI_LSTM-GRU, achieve an increment in classification accuracy of 5% and 10% from SMOTE-augmented data and 11% and 14% from raw data, respectively. The sensitivity of ARGai 2.0 with feature engineering and SMOTE is significant and reached 100%, which is a remarkable achievement during our experiment as shown in Table 5. Both the sensitivity and specificity of all the studied models are improved with our proposed pipeline, as shown in Figure 4; this proves the robustness and effectiveness of our approach (ARGai 2.0).

Table 5. Performance metrics of studied models on SMOTE augmented with selected features dataset

Models	Sen	Spe	F-1	MCC	AUC	Acc
1-D CNN	95	56	88	58	95	83
BI-LSTM-GRU	95	62	91	63	92	86
ARGai 2.0	**100**	**86**	**98**	**93**	**99**	**97**

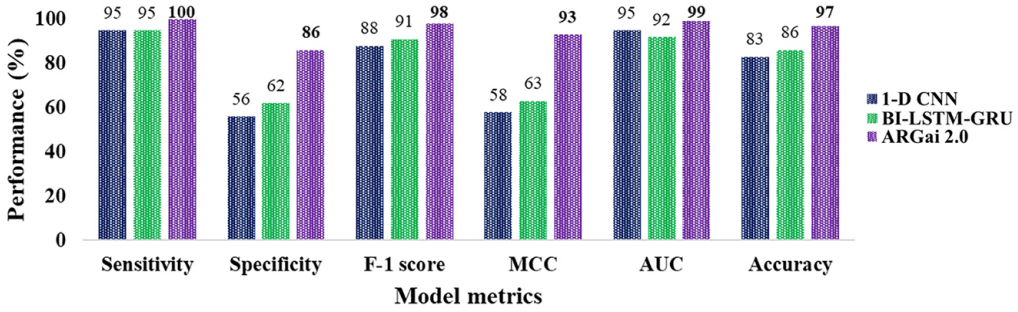

Fig. 4. Performance metrics of studied models on SMOTE and top 20 selected features

Our suggested approach, ARGai 2.0, identifies the ARGs to classify the resistant-susceptible strains with higher model metrics and utilizes minimal computational time. The learning rate of ARGai 2.0 during the training phase is significant with the non-linear data; this proves the robustness of our model. We observed an average increment of 12.5% in classification accuracy with ARGai 2.0 compared to the studied stand-alone DL models together. The genes identified by ARGai 2.0 are validated in our biological validation section (Section 8). This proves our hypothesis; the ensemble and effective feature selection minimize the computational cost and provide significant insight (higher model metrics) into the important gene identification and resistant strain classification with ARGai 2.0. We also compare our proposed ensemble feature selection with the frequently used feature selection technique principal component analysis (PCA) to assess the efficacy of ARGai 2.0. It is observed that all the studied models perform less well with the features selected by PCA. In our experiment, we observe a classification accuracy gain of 6% for ARGai 2.0 with an ensemble feature selection technique compared to PCA as shown in Figure 5.

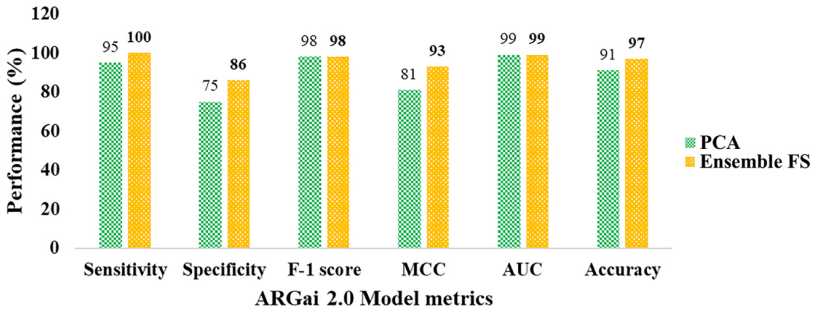

Fig. 5. Performance of ARGai 2.0 on different feature selection techniques

7 RECEIVER OPERATING CURVES

The SMOTE model significantly improves the ARGai 2.0 model's performance by providing synthetic values. Figure 6 shows the ROC performance of all the classification models. With a p-value < 0.001, our proposed model, ARGai 2.0, attained a remarkable AUC of 99.31%. The classification model does, however, reach an AUC of over 90% when trained on the SMOTE-enhanced dataset with 1-D CNN and BI-LSTM-GRU. Using SMOTE-supplemented data improved the classification performance of all the models, including the one we proposed, ARGai 2.0. The complex non-linear dataset posed challenges; however, ARGai 2.0 was able to acquire the best area under the curve (AUC) value while analyzing high-throughput GE data.

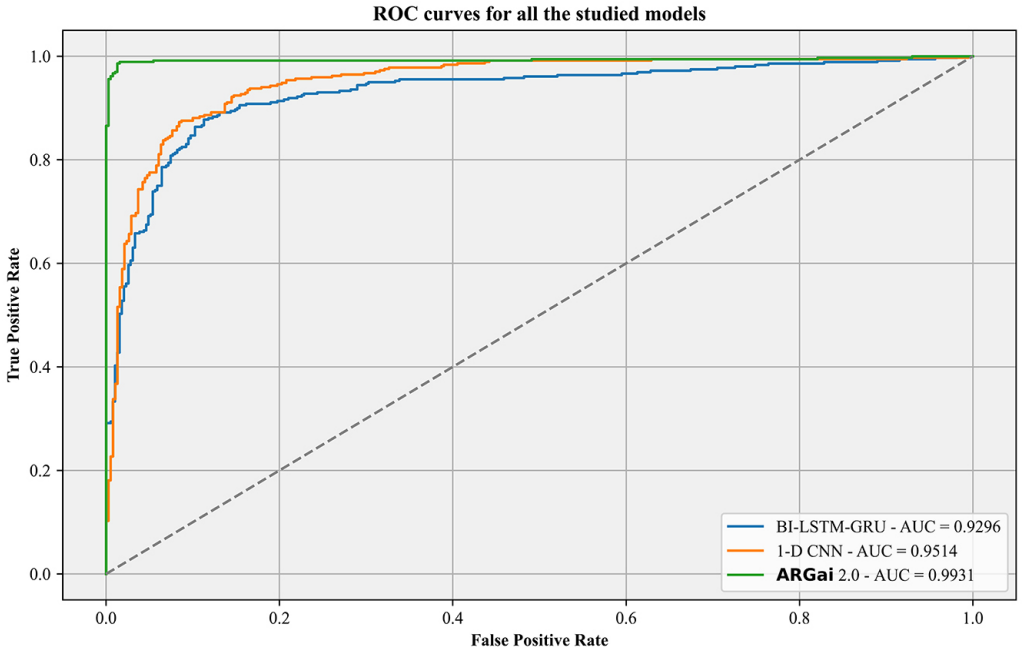

Fig. 6. ROC-AUC curve of all the studied models

8 EFFECT OF TRAINING SIZE

We assess the impact of various train-test splits on the models we examined. To test how well the models worked, we trained them on 350 enhanced strains using the top 20 features. Both the 1-D CNN and the BI-LSTM-GRU models require at least 280 data points (strains) to achieve generalization. Alternatively, as seen in Figure 7, our suggested ARGai 2.0 only needs 210 strains to achieve the substantial label of model metrics, which is 25% less than both of the studied stand-alone DL models. This indicates that our ARGai 2.0 model is both versatile and reliable for making predictions.

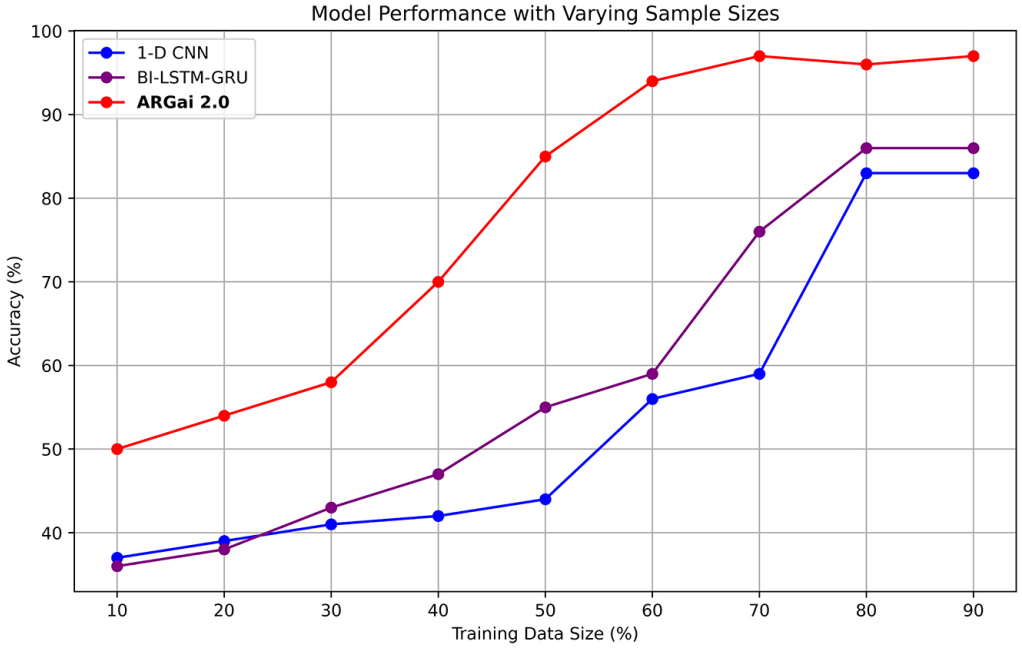

Fig. 7. Effect of various training data sizes on our studied models

9 BIOLOGICAL VALIDATION

In our study, we successfully mapped 18 out of 20 gene identifiers from the *E. coli* K12 strain to their respective gene symbols using the STRING database. This mapping is important for understanding the linkages and functional activities of these genes within the bacterial network.

Our analysis revealed that genes associated with the nitrate reductase operon (narU, narV, narW, narY, and narZ) exhibit the highest connectivity and interaction scores. This finding underscores their essential roles in nitrate metabolism under anaerobic conditions. The central positioning of narV within the network highlights its role as a hub gene, coordinating the nitrate reduction process as shown in Figure 8. This is consistent with previous studies that have identified the nitrate reductase operon as crucial for *E. coli's* adaptation to anaerobic environments [33]. The enrichment analysis further supports this, showing significant fold enrichment for pathways related to the nitrate reductase complex and nitrite transmembrane transporter activity. For instance, the pathway "Nitrate reductase complex and nitrite transmembrane transporter activity" has an enrichment FDR of 3.07E-10 with a fold enrichment of 229.28, involving genes narV, narW, narY, narZ, and narU. This high enrichment indicates the critical role these genes play in nitrate metabolism [34].

In contrast, genes such as ansP (L-asparagine permease) and yncG (glutathione S-transferase homolog) displayed weaker interactions, suggesting more peripheral roles within the network. These genes, while still important, appear to be involved in more specialized functions such as amino acid transport and detoxification. The yddE gene, although moderately involved, interacts with several nitrate reductase genes, albeit with lower confidence, indicating a potential but less central role in nitrate metabolism [35]. The enrichment analysis for these peripheral genes shows lower fold enrichment values. For example, the pathway "Mixed, incl. family of unknown function (duf5445), and protein of unknown function (duf805)" involving ansP, yncG, and yncH has an enrichment FDR of 2.75E-05 and a fold enrichment of 137.57. This suggests that while these genes are involved in specific pathways, their roles are not as central as those of the nitrate-reductase genes [36].

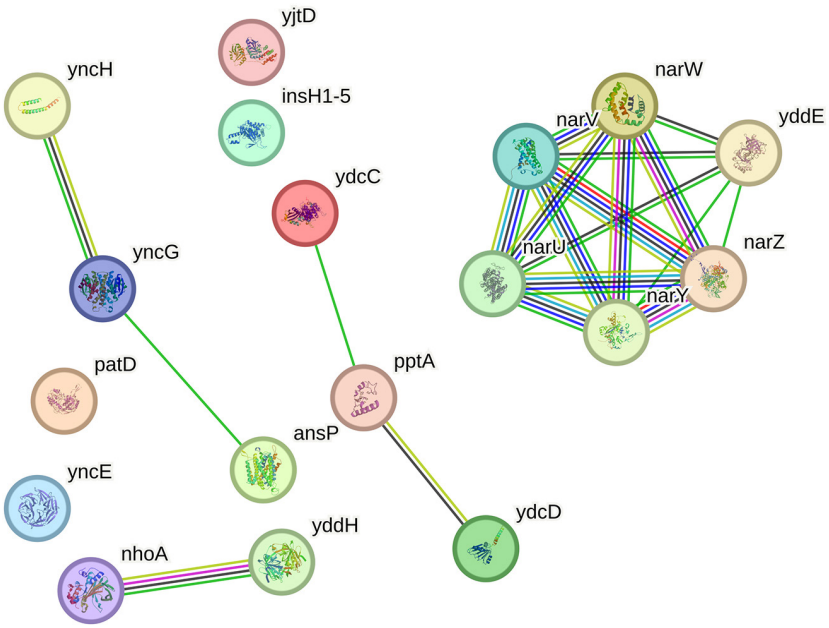

Fig. 8. Gene network of ARGai 2.0 identified genes

The STRING database results provide a comprehensive view of gene interactions in a 2D network layout. The network illustrates that nitrate metabolism genes are clustered centrally, reflecting their critical role in anaerobic respiration. In contrast, genes involved in detoxification (nhoA), amino acid transport (ansP), and other specialized functions are more peripherally located. This spatial organization within the network highlights the functional landscape of *E. coli*, where central genes are pivotal for core metabolic processes, while peripheral genes contribute to a range of metabolic and regulatory functions [33] as shown in Figure 9.

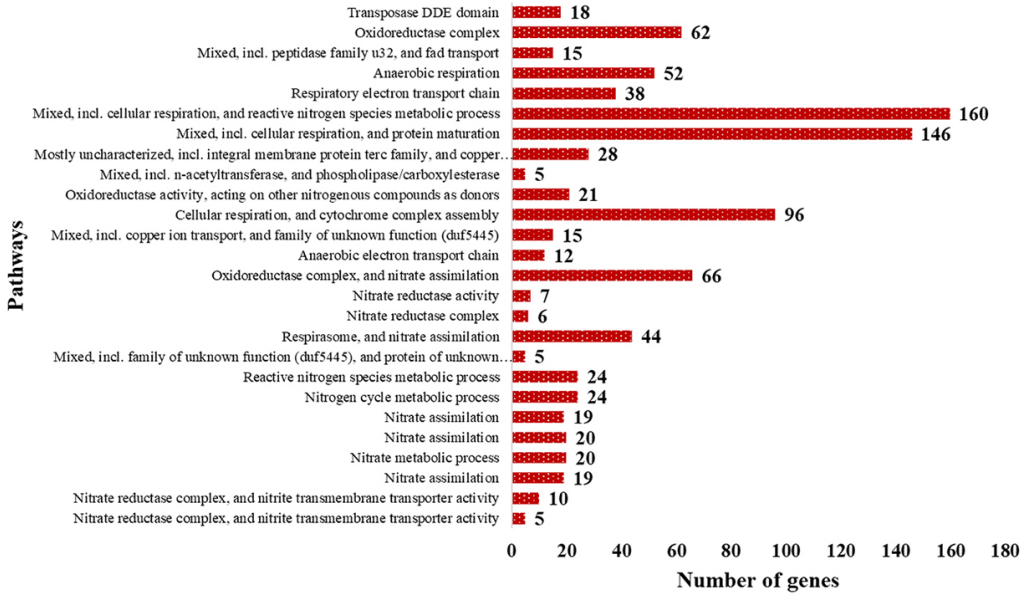

Fig. 9. Path ways analysis of ARGai 2.0 identified genes

Our gene network analysis indicates varying levels of node degrees, with high node degree genes such as nar operon members being central players. This hierarchical organization suggests that genes with high connectivity are essential for maintaining network stability and functionality. Conversely, genes with low or zero node degrees may represent specialized or independent functions, contributing to the bacterium's adaptability and resilience [34]. The detailed annotation and interaction mapping provided by this study offer valuable insights into the functional organization and adaptability of *E. coli*. By understanding the central and peripheral roles of various genes, we can better appreciate the complex gene interactions that underpin bacterial survival and adaptation.

10 SPECIAL NOTES, BENCHMARKING, AND LIMITATIONS

Access to diverse datasets on infections, drugs, and resistance mechanisms is critical for AMR studies; additionally, data complexity and scarcity provide obstacles, especially for unusual or emergent patterns. We proposed a novel ARGai 2.0 model that utilizes transfer learning for the identification of ARG and resistant strains. The key features of our proposed ARGai 2.0 model are listed below,

- Effective feature selection strategies (ensemble approach) of ARGai 2.0 boost the classification model performance by 6% compared to single feature selection methods.

- The top 20 genes considered from the outcome of ARGai 2.0 are significant and may possess AMR.
- Integrated and fine-tuned model design of ARGai 2.0 reduces the computational time.
- The pathways analysis reveals the characteristics of the top 20 genes towards AMR and thus their chances to become ARGs.
- The average classification accuracy of ARGai 2.0 is 12.5% higher than the two studied DL models together.
- For analysis of GE, especially resistant strain identification, our model ARGai 2.0 achieves a sensitivity and specificity of 98% and 94% respectively.
- The evaluations demonstrated that the model is dependable, generalizable, robust, and consistent.

In the literature, as we observed, little work is done for ARG identification using GE data. We benchmark our proposed ARGai 2.0 with several state-of-the-art AI models that utilize GE and GS data for ARG, resistant strain identification, and other disease gene identification (especially biomarkers in cancer), as shown in Table 6. Our model excels in several state-of-the-art AI models for gene identification (ARGs) and resistant strain classification. With a notable model metrics accuracy of 97%, sensitivity of 100%, and F1 of 98%, our ARGai 2.0 is an effective tool for real-time implementation.

Table 6. Benchmarking of our ARGai 2.0 with other state-of-the-art AI models

Authors	Dataset	Techniques	Acc (%)	Pre (%)	Sen (%)	Spe (%)	F1 (%)	MCC (%)
Wu et al. [37]	GS	ML	X	X	X	X	X	98.3
Nayak et al. [11]	GE	ML/DL	93	100	87	100	93	X
Argoty et al. [15]	GS	DL	X	97	90	X	X	X
Ji et al. [16]	GS	ML/DL	X	99.8	99	X	99.4	X
Moradigaravand et al. [38]	GS	ML	91	X	X	X	X	X
Babichev et al. [39]	GE	DL	97.8	X	X	X	X	X
Amniouel et al. [40]	GE	ML	96	X	89	92	X	X
Liu et al. [41]	GE	ML	X	X	X	X	X	X
Peng et al. [13]	GS	ML	97.7	X	98.2	95.8	X	X
Proposed **ARGai** 2.0	GE	DL	97	X	100	86	98	93

Notes: GS: Gene sequence; GE: Gene expression.

We observed that training the model with synthetic data might lead to better model metrics. In addition to this, our study has a few limitations; physicians do not recommend this method (the enhancement of medical data) because it is medically inaccurate. Additional research could rectify some biases in our model, including (i) the existence of relevant but smaller studies, (ii) the use of data augmentation, and (iii) comparisons with other well-known ML and DL models, such as deep networks with attention mechanisms. (iv) During our experiments we observe very little specificity compared to sensitivity; this is due to the high volume of susceptible (positive) data augmentation, and (vi) there are no comments on the real-time biological validation.

11 CONCLUSION

In our work, the advanced DL model, namely ARGai 2.0, has shown remarkable effectiveness in detecting antibiotic resistance genes (ARGs) and resistant strains in *E. coli* through the analysis of NGS GE data. The integration of advanced techniques, including SMOTE oversampling, ensemble feature selection, and TL, resulted in a model that attained high accuracy (97%) and sensitivity (100%), alongside a ROC value of 99.31. The identification of significant genes, such as pptA, patD, ydcC, yncH, narY, and insH1-5, highlights the strength of our methodology. The reduced computational time significantly improves its applicability for extensive analyses. The real-time deployment of ARGai 2.0 can boost the diagnosis and drug design for patients suffering from infectious diseases, especially UTIs. Considering the limitations posed by restricted GE data, our objective is to broaden the use of ARGai 2.0 to encompass whole genome sequence data. This approach has the potential to uncover more significant ARGs and enhance the model's generalizability and effectiveness in addressing antibiotic resistance.

12 DATA AVAILABILITY STATEMENT

The datasets analyzed for this study are freely available and can be found on NCBI (https://www.ncbi.nlm.nih.gov/gds/?term=GSE96706).

13 CONFLICT OF INTEREST

We declare that there is no conflict of interest.

14 FUNDING

No external funding is received for this study.

15 REFERENCES

[1] D. E. Bloom and D. Cadarette, "Infectious disease threats in the twenty-first century: Strengthening the global response," *Frontiers in Immunology*, vol. 10, 2019. https://doi.org/10.3389/fimmu.2019.00549

[2] M. Abavisani, A. Khoshrou, S. K. Foroushan, and A. Sahebkar, "Chatting with artificial intelligence to combat antibiotic resistance: Opportunities and challenges," *Current Research in Biotechnology*, vol. 17, p. 100197, 2024. https://doi.org/10.1016/j.crbiot.2024.100197

[3] P. Dadgostar, "Antimicrobial resistance: Implications and costs," *Infection and Drug Resistance*, vol. 12, pp. 3903–3910, 2019. https://doi.org/10.2147/IDR.S234610

[4] T. R. Walsh, A. C. Gales, R. Laxminarayan, and P. C. Dodd, "Antimicrobial resistance: Addressing a global threat to humanity," *PLoS Med.*, vol. 20, no. 7, p. e1004264, 2023. https://doi.org/10.1371/journal.pmed.1004264

[5] M. Paneri and P. Sevta, "Overview of antimicrobial resistance: An emerging silent pandemic," *Glob J of Med Pharm Biomed Update*, vol. 18, 2023. https://doi.org/10.25259/GJMPBU_153_2022

[6] P. Phungoen, J. Sarunyaparit, K. Apiratwarakul, L. Wonglakorn, A. Meesing, and K. Sawanyawisuth, "The association of ESBL Escherichia coli with mortality in patients with Escherichia coli bacteremia at the emergency department," *Drug Target Insights*, vol. 16, no. 1, pp. 12–16, 2022. https://doi.org/10.33393/dti.2022.2422

[7] S. Sarshar, R. Mirnejad, and E. Babapour, "Frequency of blaCTX-M and blaTEM virulence genes and antibiotic resistance profiles among Klebsiella pneumoniae isolates in urinary tract infection (UTI) samples from Hashtgerd, Iran," *Reports of Biochemistry & Molecular Biology*, vol. 10, no. 3, pp. 412–419, 2021. https://doi.org/10.52547/rbmb.10.3.412

[8] M. Navidinia *et al.*, "Study prevalence of verotoxigenic E. coli isolated from urinary tract infections (UTIs) in an Iranian children hospital," *The Open Microbiology Journal*, vol. 6, pp. 1–4, 2012. https://doi.org/10.2174/1874285801206010001

[9] V. Niranjan and A. Malini, "Antimicrobial resistance pattern in Escherichia coli causing urinary tract infection among inpatients," *Indian Journal of Medical Research*, vol. 139, no. 6, pp. 945–948, 2014.

[10] M. Boolchandani, A. W. D'Souza, and G. Dantas, "Sequencing-based methods and resources to study antimicrobial resistance," *Nat. Rev. Genet.*, vol. 20, pp. 356–370, 2019. https://doi.org/10.1038/s41576-019-0108-4

[11] D. S. K. Nayak *et al.*, "aiGeneR 1.0: An artificial intelligence technique for the revelation of informative and antibiotic resistant genes in Escherichia coli," *Frontiers in Bioscience-Landmark*, vol. 29, no. 2, p. 82, 2024. https://doi.org/10.31083/j.fbl2902082

[12] D. S. K. Nayak, S. Mohapatra, D. Al-Dabass, and T. Swarnkar, "Deep learning approaches for high dimension cancer microarray data feature prediction: A review," in *Computational Intelligence in Cancer Diagnosis*, 2023, pp. 13–41. https://doi.org/10.1016/B978-0-323-85240-1.00018-3

[13] Z. Peng *et al.*, "Whole-genome sequencing and gene sharing network analysis powered by machine learning identifies antibiotic resistance sharing between animals, humans and environment in livestock farming," *PLoS Computational Biology*, vol. 18, no. 3, p. e1010018, 2022. https://doi.org/10.1371/journal.pcbi.1010018

[14] D. T. Do, M.-R. Yang, T. N. S. Vo, N. Q. K. Le, and Y.-W. Wu, "Unitig-centered pan-genome machine learning approach for predicting antibiotic resistance and discovering novel resistance genes in bacterial strains," *Computational and Structural Biotechnology Journal*, vol. 23, pp. 1864–1876, 2024. https://doi.org/10.1016/j.csbj.2024.04.035

[15] G. Arango-Argoty, E. Garner, A. Pruden, L. S. Heath, P. Vikesland, and L. Zhang, "DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data," *Microbiome*, vol. 6, pp. 1–15, 2018. https://doi.org/10.1186/s40168-018-0401-z

[16] B. Ji *et al.*, "HyperVR: A hybrid deep ensemble learning approach for simultaneously predicting virulence factors and antibiotic resistance genes," *NAR Genomics and Bioinformatics*, vol. 5, no. 1, p. lqad012, 2023. https://doi.org/10.1093/nargab/lqad012

[17] Y. Pei *et al.*, "ARGNet: Using deep neural networks for robust identification and classification of antibiotic resistance genes from sequences," *Microbiome*, vol. 12, 2024. https://doi.org/10.1186/s40168-024-01805-0

[18] M. Tharmakulasingam, W. Wang, M. Kerby, R. La Ragione, and A. Fernando, "TransAMR: An interpretable transformer model for accurate prediction of antimicrobial resistance using antibiotic administration data," *IEEE Access*, vol. 11, pp. 75337–75350, 2023. https://doi.org/10.1109/ACCESS.2023.3296221

[19] M. Nsubuga, R. Galiwango, D. Jjingo, and G. Mboowa, "Generalizability of machine learning in predicting antimicrobial resistance in *E. coli*: A multi-country case study in Africa," *BMC Genomics*, vol. 25, 2024. https://doi.org/10.1186/s12864-024-10214-4

[20] C.-R. Chung *et al.*, "Data-driven two-stage framework for identification and characterization of different antibiotic-resistant escherichia coli isolates based on mass spectrometry data," *Microbiology Spectrum*, vol. 11, no. 3, pp. e03479–22, 2023. https://doi.org/10.1128/spectrum.03479-22

[21] Z. Al-Shaebi, F. Uysal Ciloglu, M. Nasser, and O. Aydin, "Highly accurate identification of bacteria's antibiotic resistance based on raman spectroscopy and U-net deep learning algorithms," *ACS Omega*, vol. 7, no. 33, pp. 29443–29451, 2022. https://doi.org/10.1021/acsomega.2c03856

[22] https://www.ncbi.nlm.nih.gov/gds/?term=GSE96706 [Accessed: 05/08, 2024].

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. https://doi.org/10.1613/jair.953

[24] D. S. K. Nayak, S. P. Routray, S. Sahooo, S. K. Sahoo, and T. Swarnkar, "A comparative study using next generation sequencing data and machine learning approach for Crohn's disease (CD) identification," in *2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS)*, 2022, pp. 17–21. https://doi.org/10.1109/MLCSS57186.2022.00012

[25] B. F. Darst, K. C. Malecki, and C. D. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," *BMC Genetics*, vol. 19, pp. 1–6, 2018. https://doi.org/10.1186/s12863-018-0633-8

[26] D. S. K. Nayak, A. Pati, A. Panigrahi, S. Sahoo, and T. Swarnkar, "ReCuRandom: A hybrid machine learning model for significant gene identification," *AIP Conference Proceedings*, vol. 2819, no. 1, 2023. https://doi.org/10.1063/5.0137029

[27] W. Zhou, Z. Yan, and L. Zhang, "A comparative study of 11 non-linear regression models highlighting autoencoder, DBN, and SVR, enhanced by SHAP importance analysis in soybean branching prediction," *Scientific Reports*, vol. 14, 2024. https://doi.org/10.1038/s41598-024-55243-x

[28] L. Huang, J. Jia, B. Yu, B.-G. Chun, P. Maniatis, and M. Naik, "Predicting execution time of computer programs using sparse polynomial regression," *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[29] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, pp. 1–40, 2016. https://doi.org/10.1186/s40537-016-0043-6

[30] S. A. B. Parisapogu, C. S. R. Annavarapu, and M. Elloumi, "1-Dimensional convolution neural network classification technique for gene expression data," in *Deep Learning for Biomedical Data Analysis: Techniques, Approaches, and Applications*, M. Elloumi, Eds., Springer, Cham. 2021, pp. 3–26. https://doi.org/10.1007/978-3-030-71676-9_1

[31] H. Wang, C. Li, J. Zhang, J. Wang, Y. Ma, and Y. Lian, "A new LSTM-based gene expression prediction model: L-GEPM," *Journal of Bioinformatics and Computational Biology*, vol. 17, no. 4, p. 1950022, 2019. https://doi.org/10.1142/S0219720019500227

[32] F. M. Alotaibi and Y. D. Khan, "A framework for prediction of oncogenomic progression aiding personalized treatment of gastric cancer," *Diagnostics*, vol. 13, no. 13, p. 2291, 2023. https://doi.org/10.3390/diagnostics13132291

[33] M. Babu *et al.*, "Genetic interaction maps in Escherichia coli reveal functional crosstalk among cell envelope biogenesis pathways," *PLoS Genetics*, vol. 7, no. 11, p. e1002377, 2011. https://doi.org/10.1371/journal.pgen.1002377

[34] D. Zhang, S. H.-J. Li, C. G. King, N. S. Wingreen, Z. Gitai, and Z. Li, "Global and gene-specific translational regulation in Escherichia coli across different conditions," *PLoS Computational Biology*, vol. 18, no. 10, p. e1010641, 2022. https://doi.org/10.1371/journal.pcbi.1010641

[35] S. Cardinale and G. Cambray, "Genome-wide analysis of E. coli cell-gene interactions," *BMC Systems Biology*, vol. 11, pp. 1–8, 2017. https://doi.org/10.1186/s12918-017-0494-1

[36] K. Syal, "Evaluation of interplay of gene expression and chromosome structure in E. coli growth: Regulatory insights," *Current Microbiology*, vol. 81, 2024. https://doi.org/10.1007/s00284-024-03773-y

[37] J. Wu *et al.*, "PLM-ARG: Antibiotic resistance gene identification using a pretrained protein language model," *Bioinformatics*, vol. 39, no. 11, p. btad690, 2023. https://doi.org/10.1093/bioinformatics/btad690

[38] D. Moradigaravand, M. Palm, A. Farewell, V. Mustonen, J. Warringer, and L. Parts, "Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data," *PLoS Computational Biology*, vol. 14, no. 12, p. e1006258, 2018. https://doi.org/10.1371/journal.pcbi.1006258

[39] S. Babichev, I. Liakh, and I. Kalinina, "Applying the deep learning techniques to solve classification tasks using gene expression data," *IEEE Access*, vol. 12, pp. 28437–28448, 2024. https://doi.org/10.1109/ACCESS.2024.3368070

[40] S. Amniouel and M. S. Jafri, "High-accuracy prediction of colorectal cancer chemotherapy efficacy using machine learning applied to gene expression data," *Frontiers in Physiology*, vol. 14, 2024. https://doi.org/10.3389/fphys.2023.1272206

[41] L. Liu *et al.*, "Integrated bioinformatics combined with machine learning to analyze shared biomarkers and pathways in psoriasis and cervical squamous cell carcinoma," *Frontiers in Immunology*, vol. 15, 2024. https://doi.org/10.3389/fimmu.2024.1351908

16 AUTHORS

Debasish Swapnesh Kumar Nayak is currently an Assistant professor at the Department of Computer Science and Engineering, SOET, Centurion University of Management and Technology, Bhubaneswar, India. He is continuing his Ph.D. at the Department of Computer Science and Engineering, FET-ITER, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, India. He obtained his M. Tech in Computer Science and Data Processing from Siksha 'O' Anusandhan (Deemed to be) University in 2015. He also obtained a Master of Computer Science and Application from Orissa University of Agriculture and Technology in 2009. He has a total expertise of 13 years in the field of Teaching, Research and Development, and Software Development. He has over 50 publications in SCIE, Scopus journals, and conferences. He served as a core committee member and chair in IEEE RMC-2024 and ICAIHC-2025.His research interests include AI for Biomedical Research, Deep Learning, Infectious Disease, Antimicrobial Resistance Analysis, Cancer Biology, Biomedical Engineering, IoT, and Data Mining. He can be contacted at email: swapnesh.nayak@gmail.com and debasish.nayak@cutm.ac.in.

Arpita Priyadarshini is a Master's (M. Tech) student at Department of Statistics, Utkal University, India. She holds a B. Tech degree with specialization in Information Technology from Silicon Institute of Technology, India. Her areas of interest in research are computational modeling and gene expression profiling, as well as the use of statistical and machine learning approaches in data analysis. She is skilled in statistical modeling, feature selection, R, Python, and sophisticated machine learning algorithms. She can be contacted at email: pri.arpita@gmail.com.

Sweta Padma Routray completed her B.Sc. and M.Sc. degrees in Bioinformatics at Buxi Jagabandhu Bidyadhar Autonomous College, India, in 2017 and 2019, respectively. Currently, she is pursuing her Ph.D. in Biotechnology at the Center of Biotechnology, Siksha O Anusandhan Deemed to be University. Her primary research interests lie in the fields of bioinformatics, microbial genomics, and transcriptomics. She can be contacted at email: sweta.routray6@gmail.com.

Santanu Kumar Sahoo is currently working as Associate Professor, department of Electronics and Communication Engineering, FET-ITER, Siksha 'O' Anusandhan University. He received his B.Tech. degree in electronics and communication

engineering from Utkal University, Odisha, India in 2004 and Doctoral degree in communication system engineering from Siksha O Anusandhan University, Odisha, India in 2018. His areas of interest are biomedical signal and image processing. He can be contacted at email: santanusahoo@soa.ac.in.

Tripti Swarnkar received the Ph.D. degree in Computer Science & Engineering from IIT Kharagpur WB India. She is currently a Professor in the department of Computer Application, National Institute of Technology, Raipur. She has more than two decades of teaching experience in the field of Computer Science & Engineering. Dr. Swarnkar's principal research interest is Machine learning, Omics data analysis and Medical image analysis. Her aspiration is to work at the interface of these different fields or Multidisciplinary Environment. She is IEEE senior member and IEEE EMBS & GRSS member, she is the founder chair for IEEE Bhubaneswar Subsection WIE Affinity group. She was also a Principal Investigator of Multidisciplinary Project on "Validation of Artificial Intelligence (AI) based models in screening and diagnosis of diseases in routine clinical practices", sponsored by Intel India. She can be contacted at email: tswarnkar.mca@nitrr.ac.in.