Detection of Patients at Risk of Enterobacteriaceae Infection Using Graph Neural Networks: a Retrospective Study

Racha Gouareb[1,†] (racha.gouareb@unige.ch), Alban Bornet[1,2,†] (alban.bornet@unige.ch), Dimitrios Proios[1,2] (dimitrios.proios@unige.ch), Sónia Gonçalves Pereira[3] (sonia.pereira@ipleiria.pt), Douglas Teodoro[1,2,4] (douglas.teodoro@unige.ch)

[1] Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

[2] HES-SO University of Applied Arts Sciences and Arts of Western Switzerland, Geneva, Switzerland

[3] Center for Innovative Care and Health Technology, Polytechnic of Leiria, Leiria, Portugal

[4] Swiss Institute of Bioinformatics, Lausanne, Switzerland

[†] Authors contributed equally.

Abstract

While Enterobacteriaceae bacteria are commonly found in healthy human gut, their colonisation of other body parts can potentially evolve into serious infections and health threats. We aim to design a graph-based machine learning model to assess risks of inpatient colonisation by multi-drug resistant (MDR) Enterobacteriaceae. The colonisation prediction problem was defined as a binary classification task, where the goal is to predict whether a patient is colonised by MDR Enterobacteriaceae in an undesirable body part during their hospital stay. To capture topological features, interactions among patients and healthcare workers were modelled using a graph structure, where patients are described by nodes and their interactions by edges. Then, a graph neural network (GNN) model was trained to learn colonisation patterns from the patient network enriched with clinical and spatiotemporal features. The GNN model predicts colonisation risk with an AUROC of 0.93 (95% CI: 0.92-0.94), 7% above a logistic regression baseline (0.86 [0.85-0.87]). Comparing different graph topologies, the configuration that considers only in-ward edges (0.93 [0.92-0.94]) outperforms the configurations that include only out-ward edges (0.86 [0.85-0.87]) and both edges (0.90 [0.89-0.91]). For the top-3 most prevalent MDR Enterobacteriaceae, the AUROC varies from 0.92 (0.90-0.93) for *Escherichia coli* up to 0.95 (0.92-0.98) for *Enterobacter cloacae*, using the GNN – in-ward model. Topological features via graph modelling improves the performance of machine learning models for Enterobacteriaceae colonisation prediction. GNNs could be used to support infection prevention and control programmes to detect patients at risk of colonisation by MDR Enterobacteriaceae and other bacteria families.

Introduction

Healthcare-associated infection (HAI) is a severe health problem for patients, health professionals and visitors in a healthcare facility[1,2]. The World Health Organization estimates that one in every ten patients develops an HAI[3] and, in US hospitals alone, the Centers for Disease Control estimate that HAIs account for 1.7 million infections and 99'000 associated deaths each year[4]. Among these infections, more than one third are caused by Enterobacteriaceae[5], a family of bacteria that includes the most prevalent human pathogenic species and the leading causes of nosocomial infections, such as *Escherichia coli*, *Salmonella enterica*, and *Klebsiella pneumoniae*. Given that these infections are acquired in environments under high antimicrobial pressure, they are often caused by antimicrobial resistant (AMR) and multidrug resistant (MDR) bacteria.

MDR Enterobacteriaceae infections have augmented drastically over the last two decades, especially with the rise of carbapenemase-producing Enterobacteriaceae[6]. These pathogens are able to resist not only to the action of all available beta-lactams (except aztreonam), but also to other available antimicrobial classes like fluoroquinolones and aminoglycosides, leaving physicians with few treatment options[7]. This leads to more expensive treatments, longer hospital stays, increased risks of complication, and higher risks of death[8].

The continuous rise of these pathogens in healthcare settings is multifactorial, with their ability to spread and persist in the environment and asymptomatically in patients and healthcare workers accounting as main contributors[9]. The risk of colonisation, subsequent infection and mortality due to Enterobacteriaceae increases exponentially with age, health history, and length of hospital stay[10]. Colonisation can be defined as the asymptomatic presence of a pathogen in the human body. It is not only the first step towards an overt disease of the colonised patient, with more or less severity, but also one of the main contributors to infection outbreaks in healthcare settings[11]. Indeed, some studies showed that between 36% and 39% of patients colonised by AMR Enterobacteriaceae develop a subsequent infection[12,13]. Asymptomatic infections, specially by MDR bacteria, pose also a prominent public health issue as the pathogen that the colonised patient carries can inadvertently be transmitted to other patients, which can become only colonised or, more concerningly, symptomatically infected, with increased risks of complications and even death[6]. Infection prevention and control (IPC) programs provides critical measures for preventing disease transmission in healthcare settings, with the potential to lower HAI rates by at least 30%, being sometimes the only solution to prevent and avoid these MDR colonisations and infections.

Leveraging the availability of large-scale healthcare data[14–16], routinely collected and stored in electronic health records (EHRs), machine learning models have been proposed for early detection of patients at risk of infection and to support IPC programs[17–21]. Classic machine learning methods, such as decision trees and random forest, have demonstrated good performance to predict patients at risk of HAI[22–25]. For methicillin-resistant *Staphylococcus aureus*[22] and *Clostridioides difficile*[25], these algorithms were shown to provide warnings as early as five days before diagnosis. Machine learning methods for colonisation prediction was also explored in very recent studies[26–28]. Tree-based machine learning methods, such as decision trees, random forest and extreme gradient boosting, achieved sensitivity and specificity above 80% for detecting MDR species from different pathogenic families[27], while the use of spatiotemporal features to identify patients colonised by vancomycin-resistant Enterococcus resulted in area under the receiver operating characteristic curve (AUROC) performance above 88%[26].

While classic machine learning models and hand-crafted features might show effective results in limited use cases, they often fail to generalize to large-scale and longitudinal EHR data[29,30]. Another limitation of previous approaches for Enterobacteriaceae colonisation prediction is that key interactions between patients and healthcare workers are neglected, hindering their application to complex care networks. To address these gaps, we propose a deep-learning approach based on a graph neural network (GNN) architecture[31]. This approach aims to incorporate interactions between patients and healthcare workers, inside and outside the wards, as well as other clinical and spatiotemporal features, to predict risks of

Enterobacteriaceae colonisation for inpatients. Our models were trained and evaluated using the Medical Information Mart for Intensive Care (MIMIC-III) dataset[32] and compared with classic machine learning baselines. Interestingly, the GNN models provide stronger predictive performance for early detection of AMR and MDR Enterobacteriaceae, compared to models trained on data without the patient network information. Our main contributions can be summarized as follows:

- We propose a graph-based colonisation model that considers spatiotemporal features in addition demographic and clinical condition. To avoid adding biases to the model due to information leakage, we deliberately did not use antimicrobial information.
- We design a new machine learning architecture for colonisation prediction using a GNN architecture that learns transmission network patterns from spatiotemporal and patient data. Different network configurations and transmission paths were proposed and evaluated.
- We evaluate our model against classic state-of-the-art machine learning baselines and show that it achieves superior performance, both for the original dataset and for an alternate version of the dataset that is free of class imbalance. We also conducted an explainability study to demonstrate the capacity of the model to automatically identify features associated with colonisation risk factors.
- There have been many studies investigating HAI prediction. To the best of our knowledge, this is the first attempt to explore the problem of predicting risks of AMR and MDR Enterobacteriaceae colonisation for undesirable body parts using graph models and provide data-driven hypothesis for transmission.

Methods

Study design and data sources

To train and evaluate our colonisation risk prediction models, we used laboratory, clinical, and administrative data from patients who stayed in critical care units of the Beth Israel Deaconess Medical Center (Massachusetts, USA). These data were recorded between 2001 and 2012 and made publicly available through the MIMIC-III dataset[32]. MIMIC-III is a freely available and deidentified healthcare dataset that consists of 26 tables and includes static and dynamic patient information, such as demographics, medical history and records, clinical measures, laboratory tests, and interventions. The database contains data from 46’520 unique patients aged 16 years or older and associated to 58’976 admissions. Patients can be admitted to the hospital more than once and moved between 50 different wards and seven care units during their stays. Additionally, activities from 7’567 unique healthcare workers - a nurse or a medical doctor - are recorded.

In the MIMIC-III dataset, we observed that 17% of inpatients had a positive result for Enterobacteriaceae screen. In total, 14 different bacterial species of the Enterobacteriaceae family were found from a total of 30 unique specimen types collected from inpatients. Figure 1 shows their distribution for different sample types (Figure 1-left) and different resistant profiles (Figure 1-right). *E. coli* was the most frequently found positive culture (50%), while *Citrobacter amalonaticus* or *Salmonella enterica* (not shown), were rarely found. A bacterial isolate was considered AMR if showing resistance to at least one agent in only one or two antimicrobial categories, and MDR if it was resistant to at least one agent in three or more antimicrobial categories. Otherwise, it was classified as antimicrobial susceptible (AMS). As shown in Figure 1, *C. koseri*, *E. coli*, and *K. pneumonia* were the species with highest levels of resistance (>50%), the latter two showing MDR profiles in more than 25% of cases.

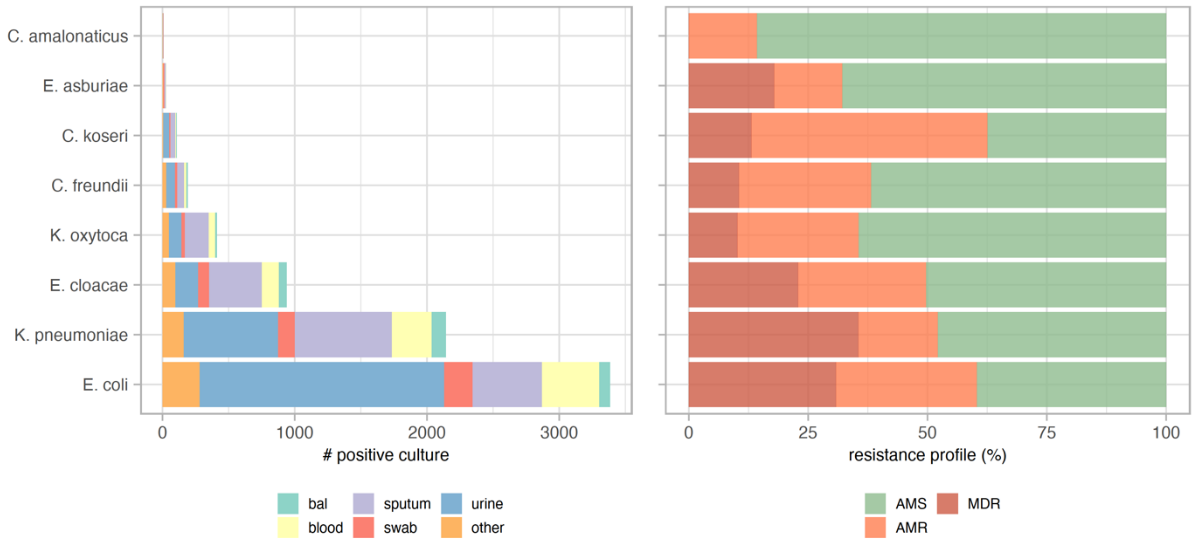

Figure 1: Frequency of positive culture and resistant profile for each Enterobacteriaceae family. Only species with more than 5 positive cultures are shown. Bal: bronchoalveolar lavage.

The training and evaluation dataset used in this study was created using the cohort selection criteria described in Figure 2. The *Microbiology Events* table from MIMIC-III was used to detect positively colonised patients. The table contains bacterial identification and antimicrobial testing results, and consists of 631'726 events related to 46'520 patients. A list of Enterobacteriaceae species was selected using the National Center for Biotechnology Information terminology[33] and used to select the microbiology events of patients colonised by Enterobacteriaceae. This first step resulted in 109'318 events related to 4'868 colonised patients. Then, a list of abnormal specimens (or uncommon body parts) where these species were found was identified by two clinical microbiologist experts and categorized into six specimen categories: blood, gastric-related, respiratory, skin, tissue, and urine. This list defined the set of positive colonisation events that were relevant to our study, i.e., presence of Enterobacteriaceae in abnormal body parts, resulting in 107'313 microbiology events and 4'838 colonised patients. Finally, the *Admissions* table provided information regarding every unique hospitalization for every patient in the database. The table was used to define the remaining non-colonised patients. Amongst all admitted patients, the ones that were not found in the filtered *Microbiology Events* table, in addition to those with Enterobacteriaceae in regular specimens (i.e., stool samples), were considered non-colonised. Lastly, the table *Transfers,* which contains patient location information and their transfers between wards, was used to assign patients to wards. The final study dataset contained 46'520 unique patients from 58'976 admissions, and a total of 274'316 patient-ward instances. If during a whole stay in a ward, there was no positive abnormal Enterobacteriaceae culture for a patient, the patient-ward instance was labelled as non-colonised; otherwise, as colonised. This resulted in 7'216 positive Enterobacteriaceae colonisations (2.6%) and 267'100 negative specimens (97.4%). The dataset was randomly divided into train (60%), dev (20%) and test (20%) sets to train the machine learning model parameters, optimize the hyper parameters and evaluate the performance, respectively. Each set contained around 2.5-3% of colonised patients.

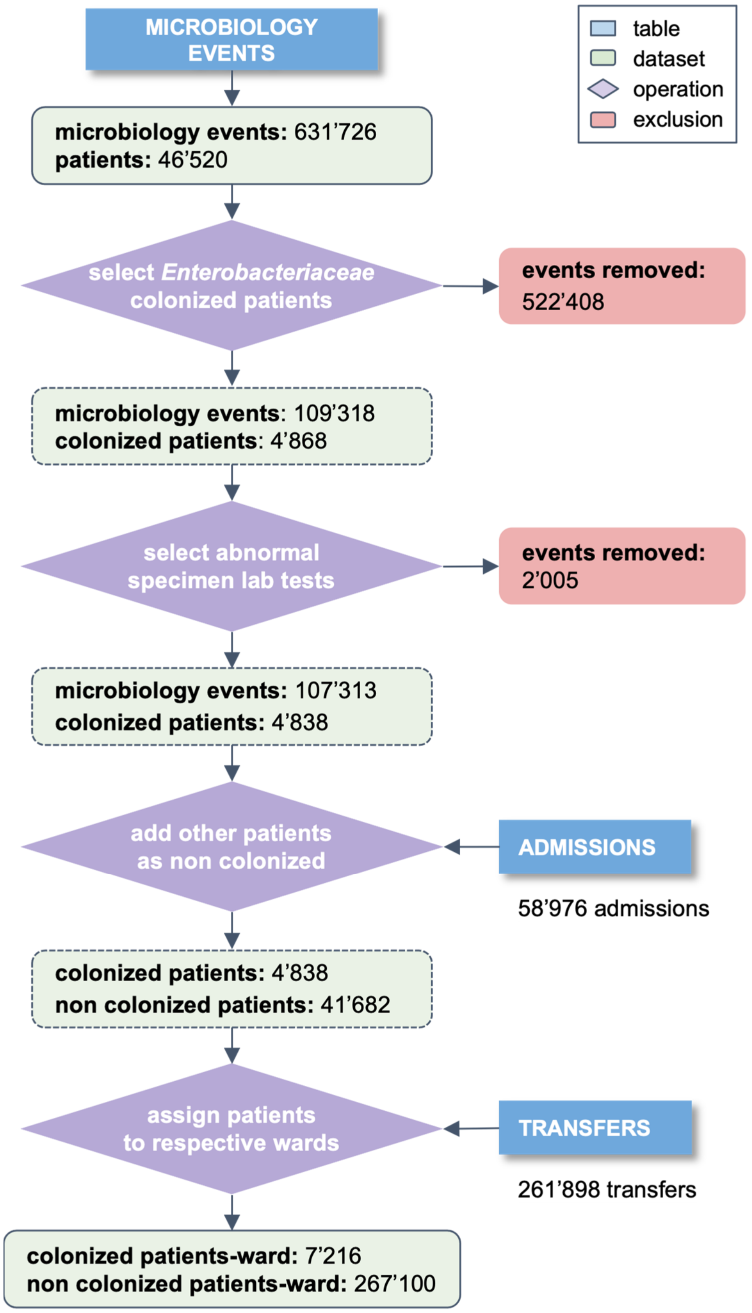

Figure 2: Cohort selection criteria. Starting from the *Microbiology Events* table of MIMIC-III, lab results were filtered for the existence of bacteria belonging to the Enterobacteriaceae family in unusual body parts to define colonised patients. *Admissions* and *Transfers* tables were used to complement the remaining patients and to label all patients.

Feature selection and data pre-processing

The feature selection process was performed iteratively. The MIMIC-III dataset was first analysed to pre-identify the set of features we considered relevant to the colonisation risk prediction problem. Then, based on model's performance computed on the dev set, less significant features, such as the death time and the discharge status of the patient, were eliminated. The final feature set can be grouped into two types: *i)* spatiotemporal features (current and previous ward, current and previous care unit, length of stay in each ward and in the hospital) and *ii)* patient features (gender and diagnosis at admission). To complement this set, we computed three new features from the data: the number of colonised patients, the total number of patients per ward, and the colonisation pressure[34]. The latter was calculated as the ratio of colonised and the total number of patients in a ward per day. Finally, the features were normalized using the *robust scaler* method of scikit-learn[35], version 1.1.2. The statistics of the resulting dataset are shown in Table 1.

Table 1: Statistics of the cohort used for model training and evaluation.

	Non-colonised (n = 267'100)	Colonised (n = 7'216)
Sex		
female	116'886 (43.8%)	3'631 (50.3%)
male	150'214 (56.2%)	3'585 (49.7%)
Age (years)		
0-17	35'446 (13.3%)	152 (2.1%)
18-25	6'325 (2.4%)	130 (1.8%)
26-45	29'023 (10.9%)	754 (10.4%)
46-65	83'912 (31.4%)	2'436 (33.7%)
66-88	101'629 (38.0%)	3'461 (48.0%)
≥ 89	10'765 (4.0%)	283 (4.0%)
Reason for admission		
newborn	34'162 (12.8%)	122 (1.7%)
pneumonia	6'736 (2.5%)	140 (1.9%)
sepsis	4'603 (1.7%)	222 (3.1%)
coronary artery disease	4'449 (1.7%)	52 (0.7%)
congestive heart failure	4'236 (1.6%)	121 (1.7%)
chest pain	3'905 (1.5%)	75 (1.0%)
other	209'009 (78.2%)	6'484 (89.9%)
Resistance profile		
AMS	–	3288 (45.6%)
AMR	–	3928 (54.4%)
MDR	–	2099 (29.1%)
Length of stay (days)		
0-4	57'511 (21.5%)	407 (5.6%)
5-10	105'531 (39.5%)	1'566 (21.7%)
11-50	93'651 (35.1%)	4'284 (59.4%)
51-100	8'183 (3.1%)	738 (10.2%)
≥ 100	2'224 (0.8%)	221 (3.1%)

Specimen		
urine	–	2’954 (40.9%)
sputum	–	1’944 (26.9%)
blood	–	934 (12.9%)
swab	–	483 (6.7%)
bronchoalveolar lavage	–	288 (4.1%)
other	–	613 (8.5%)
Enterobacteriaceae species		
E. coli	–	3’385 (47.0%)
K. pneumoniae	–	2’142 (29.7%)
E. cloacae	–	939 (13.0%)
K. oxytoca	–	410 (5.7%)
C. freundii	–	191 (2.6%)
C. koseri	–	107 (1.5%)
other	–	42 (0.5%)

Colonisation network model

We propose a homogeneous graph to model interactions between patients and healthcare workers. A graph can be defined as $G = (V, E)$, where $V = v_1, \ldots, v_{|V|}$ denotes a set of nodes and E denotes a set of edges connecting pairs of nodes $v_i, v_j \in V$. In our case, a node represents a patient and edges represent potential connections between them, either via contact with the same healthcare worker or via a common location within the hospital. As shown in Figure 3a, we considered three network configurations: *i) in-ward links* (left), where two patients are linked only if they stay in the same ward at the same time, *ii) out-ward links* (middle), where two patients are connected only if they are visited by the same healthcare worker on the same day, and *iii) all links* (right), where both ward and healthcare worker links are considered. Nodes represent a patient in a ward and their features are created using the selected feature set described in the previous section. When a patient is transferred, a new node is added to the graph with its corresponding new edges according to the different network configurations previously described (i.e., in-ward, out-ward and all links).

Graph neural network architecture

An elegant deep learning architecture for modelling graph-like data structures is GNN[31,36–38] and learn topological features, i.e., properties of the transmission network in our case. GNNs can learn complex relationships and interdependencies in graph-like data via optimizable transformations on attributes (nodes, edges, etc.) that preserve graph symmetries (i.e., permutation invariance/equivariance). Hence, in theory GNNs can make more informed predictions about entities in a network and their interactions, as compared to models that consider entities in isolation. To solve graph representation learning tasks, different GNN network architectures and algorithms have been proposed, such as graph convolutional network (GCN)[31], graph attention networks[39], and GraphSAGE[40]. These approaches use various graph feature aggregation and data sampling strategies to learn dense representations of graph components (i.e., nodes and edges), often called embeddings, that can be later used in downstream prediction tasks, such as node classification.

In our experiments, we used the GCN architecture, which employs convolutional aggregations to create graph features. GCNs are invariant to node permutations, which means that isomorphic graphs result in the same learned representation. In GCNs, node representations are learned based on neighbouring node features, which are propagated across the graph using the message passing algorithm[41]. At each layer of the GCN, every node of the graph is represented by a hidden state $h_u^{(l)}$, where u indexes the node and l the GCN layer that encodes the node features. An aggregation function A is used to send information from the immediate neighbourhood $v \in N(u)$ to every node u. Finally, to update each node representation $h_u^{(l+1)}$ in the subsequent layer $l + 1$ of the network, an update function U is used, i.e., $h_u^{(l+1)} = U(h_u^{(l)}, A(h_{v \in N(u)}^{(l)})$. In our case, the aggregation function is the sum (the hidden states of neighbouring nodes are summed over the node dimension), and the update function is a multi-layer perceptron.

A high-level view of the graph-based prediction pipeline is shown in Figure 3. Using laboratory, clinical and administrative data, patient features at the ward level (i.e., a node represented a patient in a ward) were extracted and modelled in different network colonisation models (Figure 3a). The colonisation graph was fed to a two-layer GCN ($L = 2$), with embedding size set 32, followed by a sigmoid layer for classification (Figure 3b). A binary cross-entropy with logits loss was used to train the models. To account for data imbalance, the loss coming from the non-colonised class was weighted using the ratio of non-colonised patient to the total number of patients during training. The decision threshold was set to 0.9, i.e., a patient was classified as colonised if the inferred colonisation risk was over 0.9, and otherwise classified as non-colonised. We trained all models for 400 epochs, using the Adam optimizer. The number of layers, embedding size, number of epochs and decision threshold were tuned using the dev set.

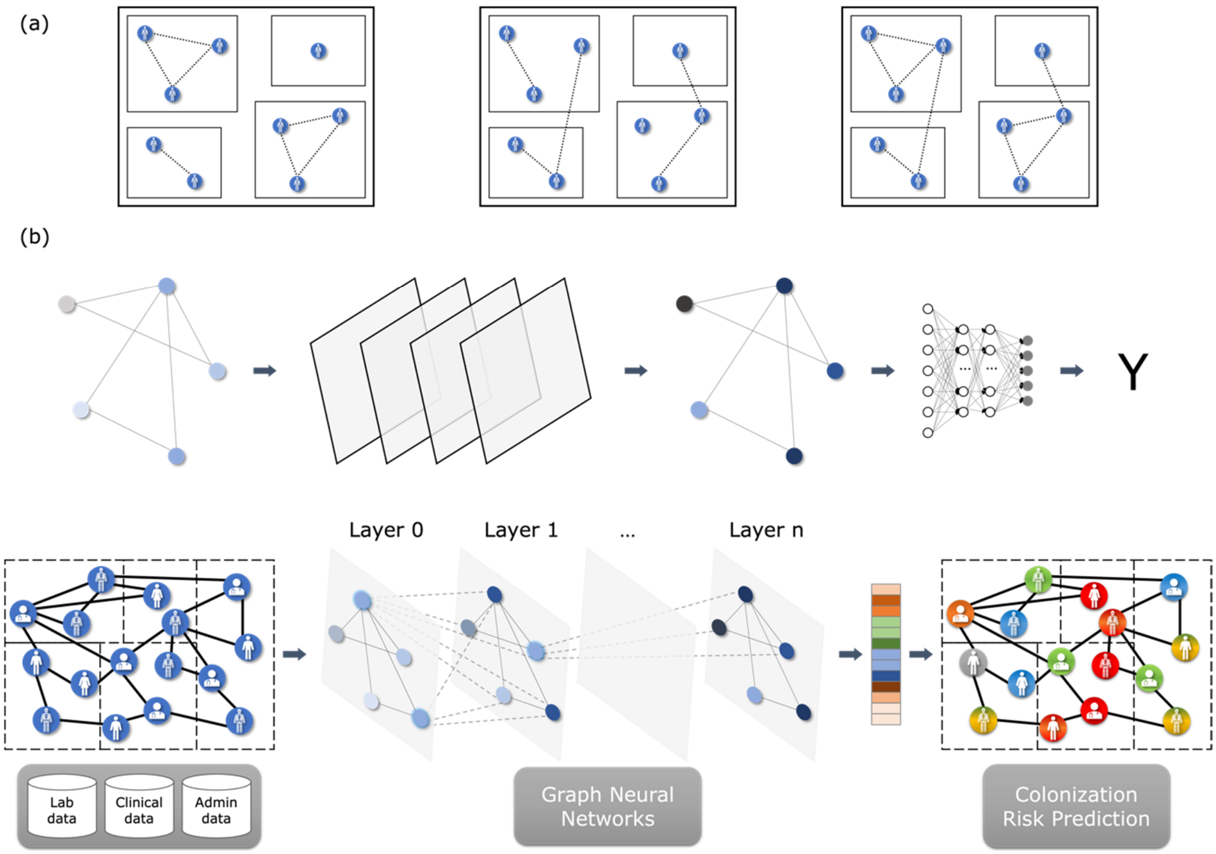

Figure 3: (a) Colonisation models. We constructed 3 different graphs, in which links were created between patients only if they were in the same ward (left), only if they were visited by the same healthcare worker (centre) or both (right). (b) Graph-based machine learning pipeline for colonisation risk prediction.

Statistical analysis

To evaluate the performance of the colonisation risk prediction models, standard binary classification metrics were computed: accuracy, macro-averaged precision, recall, F1-score, and AUROC. The GNN models were compared to classic machine learning baselines: k-nearest neighbours (kNN)[42], logistic regression[43], random forest[44] and CatBoost[45]. Student t-test was used to compare model performance. Results were deemed statistically significant for *p*-value smaller than 0.05. Shapley values were used to measure the importance of each feature to the model's predictions.

Role of funding source

Funders were not involved in the study design, data pre-processing, data analysis, interpretation, or report writing.

Results

Results obtained with the different colonisation risk prediction models are presented in Table 2. In addition to the individual classic and graph-based models (Table 2A), we created three types of ensemble models

(Table 2B): *i*) *ensemble - classic*, which combines the results of classic machine learning models; *ii*) *ensemble - GNN*, which combines the results of GNN models; and *iii*) *ensemble - all*, which combines the results of all models. For each ensemble type, we applied three voting strategies: *i*) unanimity vote [25], where a prediction is considered only if all model participants vote for the same class, and rejected otherwise; *ii*) majority vote, where a prediction is considered only if it reaches the majority amongst model participants, and rejected if there is an equality; *iii*) average probability, where the predicted class probabilities are averaged over all models to generate a prediction. Comparing the individual models, the GNN-based models show strong AUROC performance, all above 86% and outperform the classic machine learning models (*p*-value < 0.001) for this metric. Particularly, the GNN model that uses *in-ward* topology only (*GNN - in-ward*) achieves the highest AUROC (92.59% [95% confidence interval (CI): 91.67-93.51]), outperforming all individual classic models (*p*-value < 0.001). Within the graph-based models, the *out-ward* topology shows the weakest performance (86.20% [95% CI: 85.01-87.40]) followed by the *all-links* topology (89.58% [95% CI: 88.51-90.65]) (*p*-value < 0.001). These results suggest that network features enhance the predictive power of machine learning models for colonisation risk prediction, and that transmission patterns within the same ward are more useful features.

For the metrics with a decision threshold, kNN achieves the highest accuracy, with a performance of 97.82% (*p*-value < 0.001). Apart from logistic regression (82.70%), all models achieve strong accuracy, with values of 96% or higher. Due to the large class imbalance, this strong performance is expected for the accuracy metric. It is particularly noteworthy though that the *ensemble - all* model, following the unanimity vote strategy, obtained an accuracy of 99.47% when classifying 80% of the test set (the remaining 20% were discarded due to the lack of convergence between the individual classifiers). For the macro-average metrics, random forest achieves the highest F1-score among the individual models, with a performance of 73.74% (*p*-value < 0.001), nearly 6% above the best GNN model. This is likely due to an overfitting of the decision threshold for GNNs (which was trained using the dev set). Ensemble models also improve significantly upon individual models for the macro-average metrics, with an F1-score of up to 82.75% for the *ensemble - classic* configuration following the unanimity vote strategy, while classifying 82% of the samples. Lastly, as shown by the macro-average metrics (precision, recall and F1-score), the ensemble models achieve a more balanced predictive performance between colonised and non-colonised patients, which together with a high accuracy may foster better practical applications (at the expense of a reduced assessment set).

Table 2: Performance of the individual colonisation prediction models (A) and of the ensemble models (B), based on three voting strategies: *i)* unanimity vote (uv), *ii)* majority vote (mv) and *iii)* average probability (ap). Precision, recall and F1-score are macro-averaged.

A

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUROC(%) (95% CI)
kNN	97.82	83.11	66.90	72.22	80.82 (79.48-82.16)
Logistic regression	82.70	54.74	77.61	54.25	85.79 (84.58-86.99)
Random forest	97.71	79.33	70.03	73.74	85.88 (84.68-87.09)
CatBoost	97.71	81.17	65.41	70.45	84.97 (83.74-86.20)

GNN - out-ward	96.98	68.08	60.48	63.17	86.20 (85.01-87.40)
GNN - in-ward	95.84	64.94	72.20	67.76	92.59 (91.67-93.51)
GNN - all	96.45	65.52	65.12	65.32	89.58 (88.51-90.65)

B

Ensemble	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUROC(%) (95% CI)	Support (%)
Classic - uv	99.09	88.77	78.38	82.75	89.43 (87.88-90.99)	82
GNN - uv	98.22	75.80	63.36	67.50	91.10 (89.87-92.32)	96
All - uv	99.47	92.91	72.16	79.17	91.39 (89.45-93.33)	80
Classic - mv	98.05	85.09	67.20	72.92	89.29 (88.16-90.41)	99
GNN - mv	96.62	66.57	64.92	65.70	92.19 (91.24-93.13)	100
All - mv	97.60	78.12	67.34	71.36	93.62 (92.76-94.48)	100
Classic - ap	97.65	78.68	68.73	72.59	89.81 (88.75-90.87)	100
GNN - ap	82.25	55.67	84.77	55.38	92.19 (91.24-93.13)	100
All - ap	96.57	69.01	75.46	71.72	93.62 (92.76-94.48)	100

Stratified performance analysis for the GNN model

Figure 4 presents the AUROC performance of the best individual model - *GCN - in-ward* - stratified by species, specimen type, length of stay and resistance profile. The results show that the model provides consistent performance across different bacteria species, with AUROC above 89% for species that have at least 7 examples in the training dataset. The best performance is seen for *E. cloacae* (94.08% [95% CI: 91.70-96.45]) (939 examples in the training set) while the worse is for *C. amalonaticus* (95% CI: 79.60% [95% CI: 27.02-100.00]) (7 examples in the training set). Similarly, consistent performance is observed across specimens, with AUROC varying from 90.58% (95% CI: 87.85-93.31) for *blood* culture to 95.00% (95% CI: 93.53-96.47) for *sputum*. The results of Figure 4c show a decreasing trend in performance as patients stay longer in the hospital (R^2=0.8347), with performance as high as 94.14% (95% CI: 90.99-97.28) for patients that stay 4 days or less and as low as 85.60% (95% CI: 78.62-92.59) for patients that stay more than 100 days. Lastly, the model achieves similar predictive performance for different resistance profiles with the lowest score of AUROC at 92.26% (95% CI: 90.90-93.63) for AMS Enterobacteriaceae and the highest score at 93.25% (95% CI: 91.58-94.92) for MDR Enterobacteriaceae.

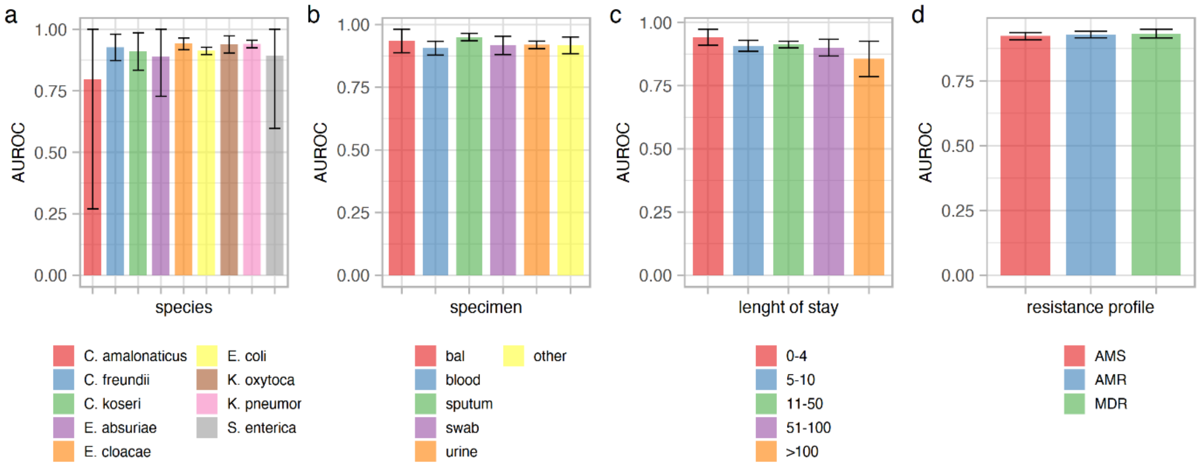

Figure 4: Performance results per species, specimen type, length of stay and resistance profile. bal: bronchoalveolar lavage; AMS: antimicrobial susceptible; AMR: antimicrobial resistant; MDR: multi-drug resistant.

Predictive performance for AMR and MDR resistance profiles

The predictive performance according to AMS, AMR and MDR resistance profiles and for the three most frequent MDR Enterobacteriaceae species is shown in Figure 5. Similar to the overall case, the *GNN – in-ward* and *Ensemble – all* models show robust performance across the difference resistance profiles and species, outperforming all the respective individual and ensemble models. For the AMR and MDR resistance profiles, the *GNN – in-ward* model achieved an AUROC of 92.89% (95% CI: 92.13-93.65) and 93.25% (95% CI: 92.28-94.21) respectively, which were slightly outperformed by the *Ensemble – all* model (93.87% [95% CI: 93.10-94.64] and 94.33% [95% CI: 93.32-95.34], respectively). For the top-3 most prevalent MDR Enterobacteriaceae, the AUROC varied from 91.74% (95% CI: 90.25-93.23) for *E. coli* up to 95.16% (95% CI: 91.92-98.41) for *E. cloacae* using the *GNN – in-ward* model, and from 92.63% (95% CI: 90.99-94.28) for *E. coli* up to 96.33% (95% CI: 95.16-97.50) for *K. pneumoniae* using the *Ensemble – all* model. Results for the classic models are slightly less consistent, with logistic regression (AMR, MDR, MDR *E. coli* and MDR *E. cloacae*), random forest (MDR *K. pneumoniae*) and CatBoost (AMS) claiming the best performance depending on the test set strata.

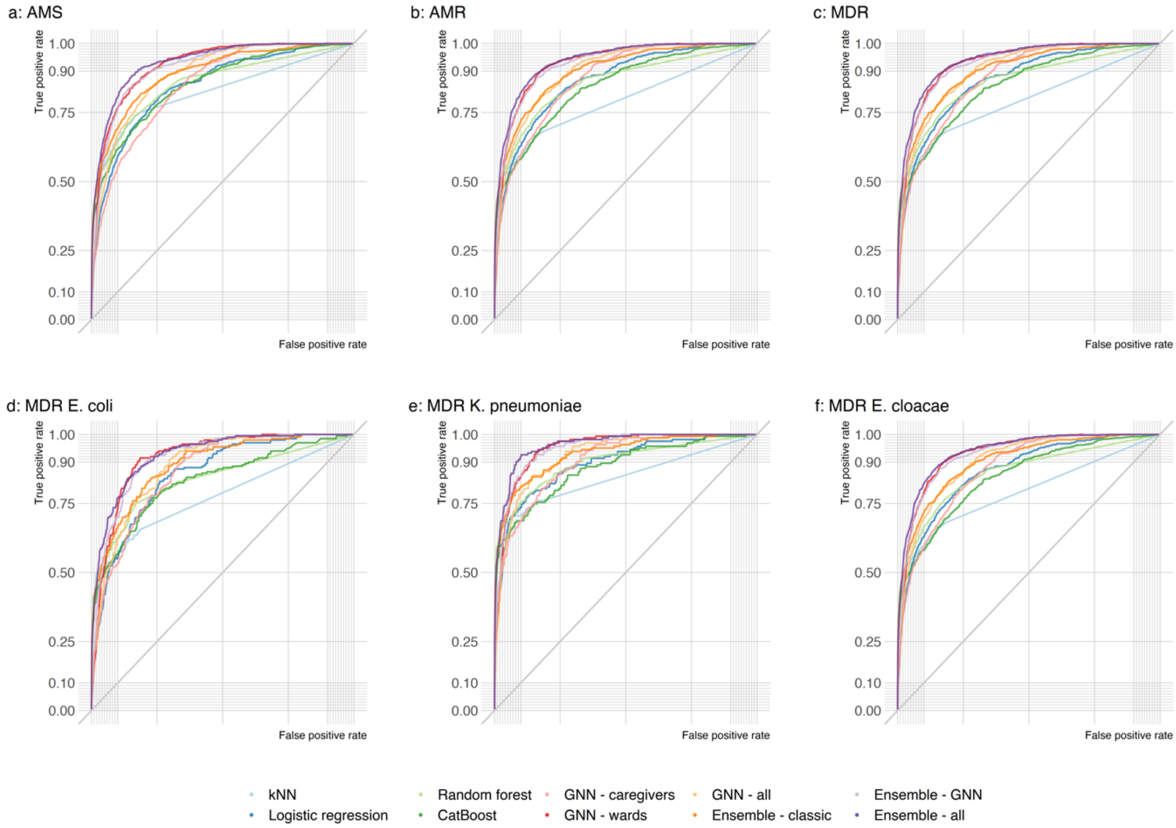

Figure 5: Model performance for antimicrobial susceptible (AMS), resistant (AMR) and multi-drug resistant (MDR) Enterobacteriaceae.

Balanced dataset scenario

As the distribution of colonised and non-colonised patients in our dataset is highly imbalanced, we evaluated the proposed models on a balanced dataset, representing an optimal-case scenario for machine learning methods. In our experiments, the balanced dataset was generated by under-sampling the original database, resulting in 8658 samples for the training set and 2887 samples for the test set, including 1473 non-colonised (51%) and 1414 colonised (49%) examples. As shown in Table 3, the results of the balanced scenario followed a similar pattern as for the original dataset. Amongst individual models, the best performance in terms of AUROC was again achieved by the GNN models, more specifically, *GNN in-ward*, with an of AUROC 92.08% (95% CI: 91.03-93.12) (p-value < 0.001). Similar to the original data scenario, for the metrics with a decision threshold, the *GNN - in-ward* model reached the highest performance in the balanced setup, with an accuracy of 80.04% (p-value < 0.001) and an F1-score of 79.61% (p-value < 0.001), outperforming even the ensemble approach based on the average probability vote strategy.

Table 3: Performance of the colonisation prediction models using a balanced dataset. Precision, recall and F1-score are macro-averaged. The ensemble model was built based on the average probability strategy.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUROC (%) (95% CI)
kNN	72.73	72.72	72.72	72.72	81.57 (80.00-83.13)

Logistic regression	71.70	71.90	71.58	71.55	78.70 (77.03-80.36)
Random forest	77.76	77.77	77.78	77.76	86.62 (85.27-87.97)
CatBoost	78.38	78.42	78.42	78.38	86.06 (84.69-87.44)
GNN - out-ward	71.14	79.73	71.68	69.22	85.76 (84.37-87.15)
GNN - in-ward	80.04	83.67	80.38	79.61	92.08 (91.03-93.12)
GNN - all	73.92	79.76	74.36	72.79	88.76 (87.52-90.01)
Ensemble - classic	78.14	78.14	78.15	78.14	86.56 (85.21-87.91)
Ensemble - GNN	65.01	79.06	65.71	60.77	91.44 (90.36-92.53)
Ensemble - all	76.06	82.19	76.50	75.06	89.06 (87.83-90.28)

Feature impact on model predictions

To explain the importance and impact of the features used in our colonisation risk prediction models, we calculated their Shapley values using the SHAP method[46]. For simplicity, we used the results of the random forest model as the one with the highest F1-score in the original dataset. Figure 6a shows the importance of the top-11 features sorted by their predictive impact, with the most significant features on top and the least important ones at the bottom. Figure 6b shows the mean absolute value of every feature presented in Figure 6a, computed over all data samples. As expected, *length of stay* in the ward and in the hospital have the highest impact on the predictions. The Shapley analysis results showed that the longer the stay in a ward or hospital, the more likely it is for a patient to be classified by the model as colonised. Conversely, the shorter the stay in a ward or the hospital, the more likely to be classified as non-colonised. The number of patients in a ward also has an important impact on model predictions. The higher the number of patients in the ward, the more probable the model output to be positive (colonised). Despite its lower impact, *female* gender influenced the model output in the positive (colonised) direction compared to *male*, which has the opposite effect. This could be explained by the fact that that most prevalent bacteria in the dataset were *E. coli* and that urinary tract infections are more common among women than men[47]. Similarly, the *neonatal intensive care unit* (NICU) was less important to the model decisions than the *medical intensive care unit* (MICU) and *surgical intensive care unit* (SICU). A patient in SICU and MICU will more likely drive the model towards a positive output (colonised), while a patient in NICU will more likely drive the model towards a negative output (non-colonised). These findings are aligned with previous risk factor analysis studies for nosocomial infections in adult intensive-care units[48].

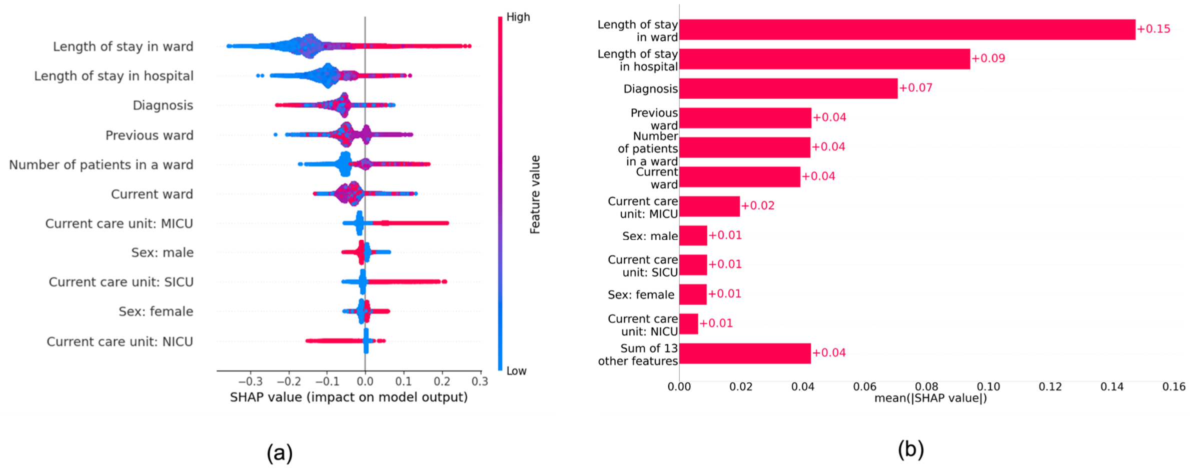

Figure 6: Feature contribution to colonisation risk prediction. a) Shapley values for top-11 features, sorted by their impact on model predictions. b) Mean absolute value of every feature presented in a).

Colonisation path analysis

A major advantage of using graph models and GNNs to predict colonisation risks is that they naturally provide possible transmission paths via graph edges. In Figure 7, we show three examples of patients that were classified correctly as colonised by the *GNN - all* model: nodes 57627 (top left), 154208 (top right) and 211904 (bottom). Nodes in green represent non-colonised patients and nodes in red represent positive culture for Enterobacteriaceae. Filled colours represent colonised patients. In the scenario of Figure 7 - top left, patient 57627 (focus patient hereafter), who was colonised by *K. pneumoniae*, stayed in the hospital for 9 days and was directly linked to four patients: two in the same room (one non-colonised and one colonised) and two in different rooms (both non-colonised). Similar to the focus patient, patient 119123 was colonised by *K. pneumoniae* and had the longest hospital stay in this subnetwork (11 days). Thus, if both bacteria strains were genetically identical (or derived phylogenetically), a possible transmission route could have been from patient 119123 to the focus patient or vice-versa, or from a common source within the ward (e.g., door handle). In Figure 7 - top right, patient 154208 (focus patient hereafter) stayed for 24 days in the hospital and had an immediate link to patient 23117 (non-colonised) from a different ward via a healthcare worker, and a second-degree connection to patient 111558 (colonised) from another ward. The latter patient and the focus patient were both colonised by *K. pneumoniae*, like in the previous scenario. Hence, the path 111558-23117-154208 could be one of the possible transmission routes within the hospital. For the third scenario, Figure 7 - bottom, patient 211904 (focus patient hereafter), male, stayed for 10 days and had a direct connection to patient 36255 via the same ward, both colonised, but by different bacteria. Moreover, these patients had a second-degree connection to patient 158476, female, via a healthcare worker link, who was colonised by *E. coli*, as the focus patient. Since patient 158476 was hospitalised for 7 days, she may have been colonised by the same strain as the one of the focus patient (or vice-versa), who may have been previously colonised. Thus, the undirected path 211904-36255-158476 could be a possible transmission route. Nevertheless, exact identification of transmission routes for such scenarios would require detailed phylogenetic analysis of bacterial samples[49].

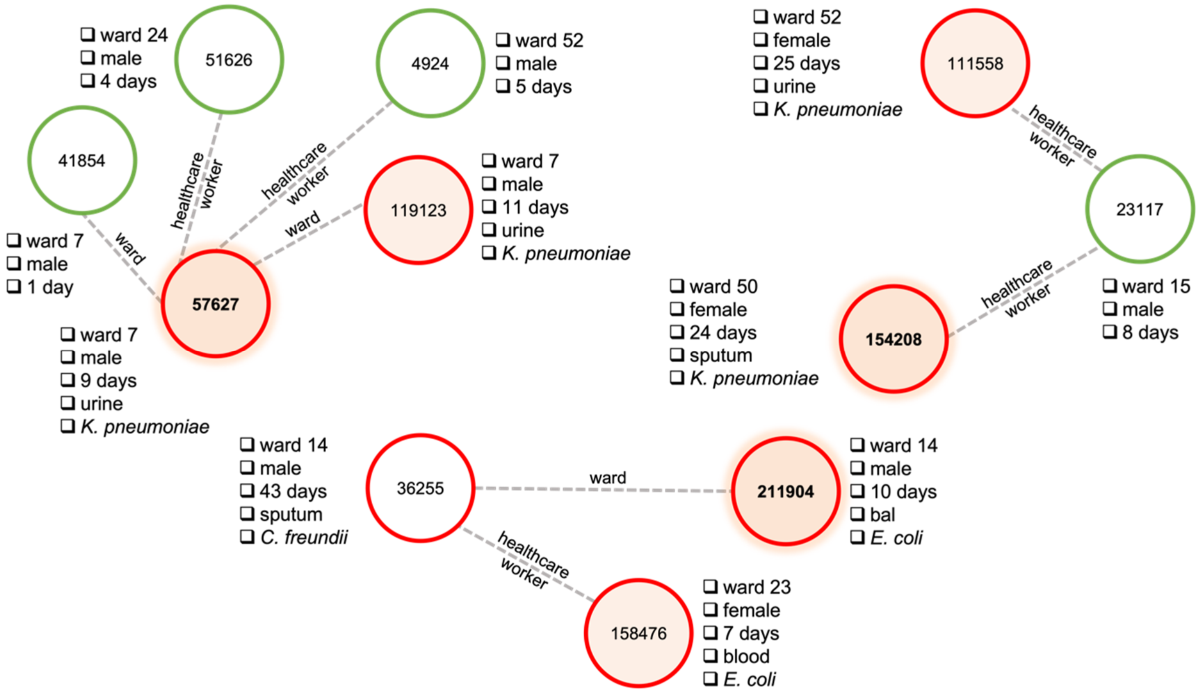

Figure 7: Bacteria transmission scenarios via graph paths. Green nodes: non-colonised patients; red nodes: colonised patients.

Discussion

This study describes a machine learning model based on graph neural networks to predict patients at risk of colonisation by AMR and MDR Enterobacteriaceae. We model the data as a graph to represent possible connections and interactions between patients and healthcare workers inside the healthcare facility. Different graph topologies were proposed based on geographic location and interaction with healthcare workers. We considered spatiotemporal features, such as length of stay and ward movement, in addition to clinical and laboratory information, to encode patients via node features in different graph topologies. Performance analyses showed that GNN models provide robust predictive performance, often above AUROC of 92%, outperforming all classic machine learning baselines used in our experiments. These results demonstrate the importance of incorporating topological features to learn patterns of patient profiles that are more likely to be colonised by MDR Enterobacteriaceae.

Other recent studies investigated the use of machine learning to predict colonisation risk of AMR species of the Enterobacteriaceae[27,28], Enterococcaceae[26] and Staphylococcaceae[18] families, achieving robust predictive performance with an AUROC between 88% and 89%. Our study is the first to consider the colonisation risk for AMR and MDR Enterobacteriaceae family, which are responsible for the highest incidence of nosocomial infections and HAI-related mortality[50], using a transmission network approach and spatiotemporal information. Moreover, in contrast to previous studies, which were based on ensemble of tree methods such as random forest, our proposed methodology used a deep learning approach and showed superior predictive power for the colonisation prediction problem of Enterobacteriaceae (in our experiments, 86% for random forest vs. 93% for GNN). Another advantage of the graph-based modelling, as opposed to tabular data used in previous studies, is that possible transmission routes can be inherently extracted from the model, opening an avenue for data-driven transmission route hypothesis generation.

Following IPC guidelines, when an AMR Enterobacteriaceae outbreak occurs in a hospital or in a long-term care facility, colonised patients are initially isolated. Then, the contact group, i.e., patients potentially colonised by the outbreak strain, is identified to determine the magnitude of the outbreak and, if required, additional IPC measures are applied[51]. Using administrative information from the EHR system, contact tracing information can be obtained and used to determine other patients potentially at risk, which will ultimately go through a screening process to duly confirm colonisation by the AMR strain. This process is reactive and can be uncomfortable for patients, as well as very costly and time consuming, preventing thus corrective actions to be taken in due time[52]. The predictive model proposed in this study could help improve IPC measures against Enterobacteriaceae, and other pathogens, in several ways. First, it could help to estimate the contact group with high accuracy, which in turn could lead to more effective measures to curb transmission and infection. Second, possible transmission paths could be automatically derived from the graph model, providing hypotheses for transmission routes. Lastly, and more importantly, if deployed in a surveillance mode, it could support early identification of potential patients at risk of AMR and MDR colonisation and enable outbreak forewarning, which could have an even bigger positive impact on live-saving and financial costs.

Despite the black-box nature of neural networks, explainable artificial intelligence methods, such as the Shapley values used to analyse our results, can provide an effective approach to interpret the model decisions and support identification of risk factors associated to colonisation risks. Among the features having the highest impact on model predictions, features such as *length of stay*, *previous ward* and *gender* have also been identified as relevant by previous epidemiological studies that investigated risk factors for HAI colonisation and infection. For example, Patel *et al.*[53] showed that carbapenem-resistant *K. pneumoniae* infection was independently associated with longer length of stay before infection. McHaney-Lindstrom *et al.*[54] showed that unit transfer increases the odds of contracting an infection by 7%. For the case of gender, the model not only identified this feature as a risk factor but also showed that being a female is associated with higher risk of Enterobacteriaceae colonisation. This result was found in previous risk analysis studies, which identified higher incidence rates of *E. coli* in females as compared to males[55].

Applying machine learning algorithms to solve the task of colonisation risk prediction is challenging due to the imbalanced nature of the data. Machine learning models are often biased towards the majority class (i.e., non-colonised in our case), and in the worst-case scenario, they will ignore the minority group entirely. In such cases, accuracy and other micro-average metrics are not optimal to evaluate model performance. Even if the model fails to predict the minority class, i.e., *colonised* in our case, accuracy might still be high due to the high percentage of non-colonised patients. To provide a more comprehensive view of our results, we reported macro-averaged metrics, which assign equal weights to positive (colonised) and negative (non-colonised) classes. Moreover, we reported these same metrics in a balanced scenario, using an undersampling technique[56]. In both cases, results showed that the models learned colonisation patterns, with macro F1-score well above the 50% threshold indicating that some colonisation patterns were indeed learned by the model.

Our study has several limitations, both in terms of data and modelling. First, the model might not be able to generalise to other hospitals as it was only evaluated in a single hospital unit dataset. Indeed, it is known that the epidemiology of HAI varies within different units and geographies[57]. Investigations of generalization performance for this type of models will warrant specific future research. Second, while we avoided using predictors that might overlap with the dependent variable, such as antimicrobial consumption (e.g., trimethoprim-sulfamethoxazole antimicrobial medication could be a predictor for *E. coli* infection[58]), other predictors, such as diagnosis at admission, could still have caused prediction bias. Nevertheless, given the distribution of diagnoses in the dataset, we expect that this bias is limited, if any. Third, our graph topology does not include environmental transmission, while it is known that indirect transmission via the environment is an important part of HAI routes[59]. Due to the lack of fine-grained contact and sampling data in the MIMIC-III dataset, environment-related transmission pathways were ignored in our models as this scenario could not be realistically captured. Understanding the impact of environmental transmission on model performance could be another research direction. Lastly, due to the anonymisation strategy of MIMIC-III and, more specifically, to the time shift, the data used in our experiments could be better regarded as a synthetic data (generated from real data) rather than as real hospital data[60,61].

To conclude, this study shows that encoding topological information about patient-healthcare worker interactions using GNNs can improve predictive performance of AMR/MDR Enterobacteriaceae colonisation models and support identification of patients potentially at risk of infection. Hence, these models could be used to enhance IPC programmes and reduce HAI burden. Given the data-driven approach of our method, we expect that it could be expanded to other pathogens with similar transmission dynamics and to other healthcare settings.

Competing Interests

The Authors declare no Competing Financial or Non-Financial Interests.

Data sharing

The MIMIC-III database is freely and publicly available through PhysioNet. The code will be shared on GitHub upon acceptance.

Authors and contributors

RG designed and implemented the models, ran the experiments and analyses. RG and DT wrote the manuscript draft. DT and SGP conceptualised the experiments and acquired funding. RG, DP and SGP curated the data. RG, AB, DP and DT analysed the data. All authors reviewed and approved the manuscript.

References

(1) Allegranzi, B.; Nejad, S. B.; Combescure, C.; Graafmans, W.; Attar, H.; Donaldson, L.; Pittet, D. Burden of Endemic Health-Care-Associated Infection in Developing Countries: Systematic Review and Meta-Analysis. *The Lancet* **2011**, *377* (9761), 228–241.

(2) World Health Organization. *Charter: Health Worker Safety: A Priority for Patient Safety*; World Health Organization, 2020.

(3) Organization, W. H. Report on the Burden of Endemic Health Care-Associated Infection Worldwide. **2011**.

(4) Klevens, R. M.; Edwards, J. R.; Richards Jr, C. L.; Horan, T. C.; Gaynes, R. P.; Pollock, D. A.; Cardo, D. M. Estimating Health Care-Associated Infections and Deaths in US Hospitals, 2002. *Public health reports* **2007**, *122* (2), 160–166.

(5) Carelink, P. *Healthcare-Acquired Infections (HAIs). 2016*; 2018.

(6) Tzouvelekis, L. S.; Markogiannakis, A.; Piperaki, E.; Souli, M.; Daikos, G. L. Treating Infections Caused by Carbapenemase-Producing Enterobacteriaceae. *Clinical Microbiology and Infection* **2014**, *20* (9), 862–872.

(7) Fritzenwanker, M.; Imirzalioglu, C.; Herold, S.; Wagenlehner, F. M.; Zimmer, K.-P.; Chakraborty, T. Treatment Options for Carbapenem-Resistant Gram-Negative Infections. *Deutsches Ärzteblatt International* **2018**, *115* (20–21), 345.

(8) Marchetti, A.; Rossiter, R. Economic Burden of Healthcare-Associated Infection in US Acute Care Hospitals: Societal Perspective. *Journal of medical economics* **2013**, *16* (12), 1399–1404.

(9) Dalton, K. R.; Rock, C.; Carroll, K. C.; Davis, M. F. One Health in Hospitals: How Understanding the Dynamics of People, Animals, and the Hospital Built-Environment Can Be Used to Better Inform Interventions for Antimicrobial-Resistant Gram-Positive Infections. *Antimicrob Resist Infect Control* **2020**, *9* (1), 78. https://doi.org/10.1186/s13756-020-00737-2.

(10) Denton, M. Enterobacteriaceae. *International journal of antimicrobial agents* **2007**, *29*, S9–S22.

(11) Jamrozik, E.; Selgelid, M. J. Invisible Epidemics: Ethics and Asymptomatic Infection. *Monash Bioeth. Rev.* **2020**, *38* (S1), 1–16. https://doi.org/10.1007/s40592-020-00123-z.

(12) Gao, Y.; Chen, M.; Cai, M.; Liu, K.; Wang, Y.; Zhou, C.; Chang, Z.; Zou, Q.; Xiao, S.; Cao, Y.; Wang, W.; Liu, Z.; Lv, L.; Luo, Y.; Wu, Y. An Analysis of Risk Factors for Carbapenem-Resistant Enterobacteriaceae Infection. *J Glob Antimicrob Resist* **2022**, *30*, 191–198. https://doi.org/10.1016/j.jgar.2022.04.005.

(13) Akturk, H.; Sutcu, M.; Somer, A.; Aydın, D.; Cihan, R.; Ozdemir, A.; Coban, A.; Ince, Z.; Citak, A.; Salman, N. Carbapenem-Resistant Klebsiella Pneumoniae Colonization in Pediatric and Neonatal Intensive Care Units: Risk Factors for Progression to Infection. *Braz J Infect Dis* **2016**, *20* (2), 134–140. https://doi.org/10.1016/j.bjid.2015.12.004.

(14) Liu, Q.; Yang, J.; Zhang, J.; Zhao, F.; Feng, X.; Wang, X.; Lyu, J. Description of Clinical Characteristics of VAP Patients in MIMIC Database. *Frontiers in pharmacology* **2019**, *10*, 62.

(15) Lin, J.; Gu, C.; Zhang, S.; Tian, L.; Ren, K.; Cao, Z.; Han, X. Sites and Causes of Infection in Patients with Sepsis-Associated Liver Dysfunction: A Population Study from the Medical Information Mart for Intensive Care III. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research* **2021**, *27*, e928928-1.

(16) Zhao, L.; Gao, Y.; Guo, S.; Lu, X.; Yu, S.; Ge, Z.; Zhu, H.; Li, Y. Prognosis of Patients with Sepsis and Non-Hepatic Hyperammonemia: A Cohort Study. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research* **2020**, *26*, e928573-1.

(17) Peiffer-Smadja, N.; Rawson, T. M.; Ahmad, R.; Buchard, A.; Georgiou, P.; Lescure, F.-X.; Birgand, G.; Holmes, A. H. Machine Learning for Clinical Decision Support in Infectious Diseases: A Narrative Review of Current Applications. *Clinical Microbiology and Infection* **2020**, *26* (5), 584–595.

(18) Hirano, Y.; Shinmoto, K.; Okada, Y.; Suga, K.; Bombard, J.; Murahata, S.; Shrestha, M.; Ocheja, P.; Tanaka, A. Machine Learning Approach to Predict Positive Screening of Methicillin-Resistant Staphylococcus Aureus During Mechanical Ventilation Using Synthetic Dataset From MIMIC-IV Database. *Frontiers in medicine* **2021**, 2222.

(19) Baldominos, A.; Puello, A.; Oğul, H.; Aşuroğlu, T.; Colomo-Palacios, R. Predicting Infections Using Computational Intelligence–a Systematic Review. *IEEE Access* **2020**, *8*, 31083–31102.

(20) Teodoro, D.; Lovis, C. Empirical Mode Decomposition and K-Nearest Embedding Vectors for Timely Analyses of Antibiotic Resistance Trends. *PloS one* **2013**, *8* (4), e61180.

(21) Teodoro, D.; Pasche, E.; Gobeill, J.; Emonet, S.; Ruch, P.; Lovis, C. Building a Transnational Biosurveillance Network Using Semantic Web Technologies: Requirements, Design, and Preliminary Evaluation. *Journal of medical Internet research* **2012**, *14* (3), e2043.

(22) Hartvigsen, T.; Sen, C.; Brownell, S.; Teeple, E.; Kong, X.; Rundensteiner, E. A. Early Prediction of MRSA Infections Using Electronic Health Records. In *HEALTHINF*; 2018; pp 156–167.

(23) Jeng, S.-L.; Huang, Z.-J.; Yang, D.-C.; Teng, C.-H.; Wang, M.-C. Machine Learning to Predict the Development of Recurrent Urinary Tract Infection Related to Single Uropathogen, Escherichia Coli. *Sci Rep* **2022**, *12* (1), 17216. https://doi.org/10.1038/s41598-022-18920-3.

(24) Yang, D.; Xie, Z.; Xin, X.; Xue, W.; Zhang, M. A Model for Predicting Nosocomial Carbapenem-Resistant Klebsiella Pneumoniae Infection. *Biomed Rep* **2016**, *5* (4), 501–505. https://doi.org/10.3892/br.2016.752.

(25) Sen, C.; Hartvigsen, T.; Rundensteiner, E.; Claypool, K. Crest-Risk Prediction for Clostridium Difficile Infection Using Multimodal Data Mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer, 2017; pp 52–63.

(26) van Niekerk, J. M.; Lokate, M.; Braakman-Jansen, L. M. A.; van Gemert-Pijnen, J.; Stein, A. Spatiotemporal Prediction of Vancomycin-Resistant Enterococcus Colonisation. *BMC infectious diseases* **2022**, *22* (1), 1–12.

(27) Çağlayan, Ç.; Barnes, S. L.; Pineles, L. L.; Harris, A. D.; Klein, E. Y. A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized With Multidrug-Resistant Organisms. *Front Public Health* **2022**, *10*, 853757. https://doi.org/10.3389/fpubh.2022.853757.

(28) Goodman, K. E.; Simner, P. J.; Klein, E. Y.; Kazmi, A. Q.; Gadala, A.; Toerper, M. F.; Levin, S.; Tamma, P. D.; Rock, C.; Cosgrove, S. E.; Maragakis, L. L.; Milstone, A. M. Predicting Probability of Perirectal Colonization with Carbapenem-Resistant Enterobacteriaceae (CRE) and Other Carbapenem-Resistant Organisms (CROs) at Hospital Unit Admission. *Infect Control Hosp Epidemiol* **2019**, *40* (5), 541–550. https://doi.org/10.1017/ice.2019.42.

(29) Kawaguchi, K.; Kaelbling, L. P.; Bengio, Y. Generalization in Deep Learning. *arXiv preprint arXiv:1710.05468* **2017**.

(30) Si, Y.; Du, J.; Li, Z.; Jiang, X.; Miller, T.; Wang, F.; Zheng, W. J.; Roberts, K. Deep Representation Learning of Patient Data from Electronic Health Records (EHR): A Systematic Review. *Journal of Biomedical Informatics* **2021**, *115*, 103671.

(31) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907* **2016**.

(32) Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; Mark, R. G. MIMIC-III, a Freely Accessible Critical Care Database. *Scientific data* **2016**, *3* (1), 1–9.

(33) Schoch, C. L.; Ciufo, S.; Domrachev, M.; Hotton, C. L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; Mcveigh, R.; O'Neill, K.; Robbertse, B. NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools. *Database* **2020**, *2020*.

(34) Bonten, M. J.; Gaillard, C. A.; Johanson Jr, W. G.; van Tiel, F. H.; Smeets, H. G.; Van Der Geest, S.; Stobberingh, E. E. Colonization in Patients Receiving and Not Receiving Topical Antimicrobial Prophylaxis. *American journal of respiratory and critical care medicine* **1994**, *150* (5), 1332–1340.

(35) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-Learn: Machine Learning in Python. *the Journal of machine Learning research* **2011**, *12*, 2825–2830.

(36) Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE transactions on neural networks* **2008**, *20* (1), 61–80.

(37) Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R. Relational Inductive Biases, Deep Learning, and Graph Networks. *arXiv preprint arXiv:1806.01261* **2018**.

(38) Bronstein, M. M.; Bruna, J.; Cohen, T.; Veličković, P. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv preprint arXiv:2104.13478* **2021**.

(39) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph Attention Networks. *arXiv preprint arXiv:1710.10903* **2017**.

(40) Hamilton, W.; Ying, Z.; Leskovec, J. Inductive Representation Learning on Large Graphs. *Advances in neural information processing systems* **2017**, *30*.

(41) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. In *International conference on machine learning*; PMLR, 2017; pp 1263–1272.

(42) Cunningham, P.; Delany, S. J. K-Nearest Neighbour Classifiers-a Tutorial. *ACM Computing Surveys (CSUR)* **2021**, *54* (6), 1–25.

(43) Wright, R. E. Logistic Regression. **1995.**

(44) Breiman, L. Random Forests. *Machine learning* **2001**, *45* (1), 5–32.

(45) Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. *Advances in neural information processing systems* **2018**, *31*.
(46) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in neural information processing systems* **2017**, *30*.
(47) Harrington, R. D.; Hooton, T. M. Urinary Tract Infection Risk Factors and Gender. *The journal of gender-specific medicine: JGSM: the official journal of the Partnership for Women's Health at Columbia* **2000**, *3* (8), 27–34.
(48) Vincent, J.-L. Nosocomial Infections in Adult Intensive-Care Units. *The Lancet* **2003**, *361* (9374), 2068–2077. https://doi.org/10.1016/S0140-6736(03)13644-6.
(49) Abbas, M.; Robalo Nunes, T.; Martischang, R.; Zingg, W.; Iten, A.; Pittet, D.; Harbarth, S. Nosocomial Transmission and Outbreaks of Coronavirus Disease 2019: The Need to Protect Both Patients and Healthcare Workers. *Antimicrobial Resistance & Infection Control* **2021**, *10* (1), 1–13.
(50) Cassini, A.; Högberg, L. D.; Plachouras, D.; Quattrocchi, A.; Hoxha, A.; Simonsen, G. S.; Colomb-Cotinat, M.; Kretzschmar, M. E.; Devleesschauwer, B.; Cecchini, M. Attributable Deaths and Disability-Adjusted Life-Years Caused by Infections with Antibiotic-Resistant Bacteria in the EU and the European Economic Area in 2015: A Population-Level Modelling Analysis. *The Lancet infectious diseases* **2019**, *19* (1), 56–66.
(51) Boonstra, M. B.; Spijkerman, D. C.; Voor, A. F.; van der Laan, R. J.; Bode, L. G.; van Vianen, W.; Klaassen, C. H.; Vos, M. C.; Severin, J. A. An Outbreak of ST307 Extended-Spectrum Beta-Lactamase (ESBL)–Producing Klebsiella Pneumoniae in a Rehabilitation Center: An Unusual Source and Route of Transmission. *Infection Control & Hospital Epidemiology* **2020**, *41* (1), 31–36.
(52) Dik, J.-W. H.; Hendrix, R.; Poelman, R.; Niesters, H. G.; Postma, M. J.; Sinha, B.; Friedrich, A. W. Measuring the Impact of Antimicrobial Stewardship Programs. *Expert review of anti-Infective therapy* **2016**, *14* (6), 569–575.
(53) Patel, G.; Huprikar, S.; Factor, S. H.; Jenkins, S. G.; Calfee, D. P. Outcomes of Carbapenem-Resistant Klebsiella Pneumoniae Infection and the Impact of Antimicrobial and Adjunctive Therapies. *Infection Control & Hospital Epidemiology* **2008**, *29* (12), 1099–1106.
(54) McHaney-Lindstrom, M.; Hebert, C.; Flaherty, J.; Mangino, J. E.; Moffatt-Bruce, S.; Root, E. D. Analysis of Intra-Hospital Transfers and Hospital-Onset Clostridium Difficile Infection. *Journal of Hospital Infection* **2019**, *102* (2), 168–169.
(55) Uslan, D. Z.; Crane, S. J.; Steckelberg, J. M.; Cockerill, F. R.; Sauver, J. L. S.; Wilson, W. R.; Baddour, L. M. Age-and Sex-Associated Trends in Bloodstream Infection: A Population-Based Study in Olmsted County, Minnesota. *Archives of internal medicine* **2007**, *167* (8), 834–839.
(56) Yap, B. W.; Rani, K. A.; Rahman, H. A. A.; Fong, S.; Khairudin, Z.; Abdullah, N. N. An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. In *Proceedings of the first international conference on advanced data and information engineering (DaEng-2013)*; Springer, 2014; pp 13–22.
(57) Livermore, D. M.; Pearson, A. Antibiotic Resistance: Location, Location, Location. *Clinical Microbiology and Infection* **2007**, *13*, 7–16.
(58) Brown, P. D.; Freeman, A.; Foxman, B. Prevalence and Predictors of Trimethoprim-Sulfamethoxazole Resistance among Uropathogenic Escherichia Coli Isolates in Michigan. *Clinical infectious diseases* **2002**, *34* (8), 1061–1066.
(59) Blanco, N.; O'Hara, L. M.; Harris, A. D. Transmission Pathways of Multidrug-Resistant Organisms in the Hospital Setting: A Scoping Review. *Infection Control & Hospital Epidemiology* **2019**, *40* (4), 447–456.
(60) Hittmeir, M.; Ekelhart, A.; Mayer, R. Utility and Privacy Assessments of Synthetic Data for Regression Tasks. In *2019 IEEE International Conference on Big Data (Big Data)*; IEEE, 2019; pp 5763–5772.
(61) Hittmeir, M.; Ekelhart, A.; Mayer, R. On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*; 2019; pp 1–6.