IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 May 2023, accepted 6 July 2023, date of publication 17 July 2023, date of current version 26 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3296221

RESEARCH ARTICLE

TransAMR: An Interpretable Transformer Model for Accurate Prediction of Antimicrobial Resistance Using Antibiotic Administration Data

MUKUNTHAN THARMAKULASINGAM[1], (Member, IEEE), WENWU WANG[1], (Senior Member, IEEE), MICHAEL KERBY[2], ROBERTO LA RAGIONE[3,4], AND ANIL FERNANDO[5], (Senior Member, IEEE)

[1]Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH Guildford, U.K.
[2]Synergy Farm Health Ltd., Maiden Newton, DT2 0HS Dorset, U.K.
[3]School of Veterinary Medicine, University of Surrey, GU2 7AL Guildford, U.K.
[4]School of Biosciences, University of Surrey, GU2 7XH Guildford, U.K.
[5]Department of Computer and Information Sciences, University of Strathclyde, G1 1XQ Glasgow, U.K.

Corresponding author: Mukunthan Tharmakulasingam (m.tharmakulasingam@surrey.ac.uk)

This work was supported in part by the University of Surrey, and in part by Zoetis through the vHive Initiative.

ABSTRACT Antimicrobial Resistance (AMR) is a growing public and veterinary health concern, and the ability to accurately predict AMR from antibiotics administration data is crucial for effectively treating and managing infections. While genomics-based approaches can provide better results, sequencing, assembling, and applying Machine Learning (ML) methods can take several hours. Therefore, alternative approaches are required. This study focused on using ML for antimicrobial stewardship by utilising data extracted from hospital electronic health records, which can be done in real-time, and developing an interpretable 1D-Transformer model for predicting AMR. A multi-baseline Integrated Gradient pipeline was also incorporated to interpret the model, and quantitative validation metrics were introduced to validate the model. The performance of the proposed 1D-Transformer model was evaluated using a dataset of urinary tract infection (UTI) patients with four antibiotics. The proposed 1D-Transformer model achieved 10% higher area under curve (AUC) in predicting AMR and outperformed traditional ML models. The Explainable Artificial Intelligence (XAI) pipeline also provided interpretable results, identifying the signatures contributing to the predictions. This could be used as a decision support tool for personalised treatment, introducing AMR-aware food and management of AMR, and it could also be used to identify signatures for targeted interventions.

INDEX TERMS Transformer, multi-drug AMR, antimicrobial stewardship, missing labels, XAI, multi-label prediction.

I. INTRODUCTION

Antimicrobial Resistance (AMR) is a major global health concern associated with the inappropriate use of antimicrobial drugs. In particular, some empirical treatments used for humans and livestock may be inappropriate [1], and can lead to the emergence of AMR. One solution to combat AMR is personalised antibiotic treatments and effective antimicrobial stewardship for humans and animals.

In order to combat AMR, it is crucial to develop accurate and interpretable models for predicting AMR. This can aid in selecting the appropriate antibiotic treatment for patients and animals, and guiding antibiotic prescribing decisions. Traditionally, laboratory testing has provided accurate information about how patients will respond to different treatment options. However, using this approach can

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno M. Garcia.

take several days to obtain results and may not be feasible for health professionals who need immediate treatment decisions. In order to obtain a comprehensive understanding of the response to all relevant treatments, it is vital to test the isolated microorganisms for various resistances, not just those used for current treatments. Another approach is to use a genomics-based approach using Machine Learning (ML) to analyse genomic data and predict AMR [2], [3]. While this approach is faster than laboratory-based methods, it still requires the extraction, annotation, and prediction of AMR from genomic data, which is not feasible in real-time.

An alternative approach is to use ML to predict AMR based on antibiotic usage data collected from Electronic Health Records (EHR) and farm records. This approach enables real-time AMR prediction at the initial treatment decision point and helps identify trade-offs between multiple factors, such as treatment effectiveness and harmful side effects. However, previous studies on predicting AMR from EHR and farm data have been limited and have used black-box models that do not provide any interpretability of the results, do not support missing labels and complex data and cannot predict multiple AMRs simultaneously.

The availability of more electronic health data and increasing Graphical Processing Unit (GPU) capabilities have created a use case for rapid prediction of AMR with Deep Learning (DL) models. However, DL models are commonly referred to as ''black boxes'' because their internal mechanisms are not easily understandable. This lack of interpretability hinders trust in the model among healthcare stakeholders and limits the ability to identify and address systematic bias.

To address these issues, this study proposed to use interpretable 1D-Transformer models for predicting AMR from antibiotics administration data and identify the features that contribute to their predictions and validate them. The study aimed to assess how DL models can be used with 1D signals to classify antibiotic resistance status and to translate the DL model output into recommendations to help select the antibiotic of the narrowest possible spectrum. The novel contributions of this article can be summarised as follows:

1) To propose a 1D Transformer model to extract features from antibiotics administration data collected to predict the antibiotic resistance status and recommend therapies.
2) To propose approaches to handle missing labels in the antibiotics administration dataset.
3) To propose a multi-baseline Integrated Gradient pipeline to identify significant features contributing to the decisions overcoming the limitations of baselines.
4) To quantitatively validate the proposed model by comparing its performance with those from other models in the literature and demonstrate the effectiveness of the proposed Transformer and XAI approaches.

The paper is structured as follows. Section II covers the background of this study. Section III describes the proposed model, Explainable AI (XAI) pipeline and metrics to measure the performance. Section IV describes the experimental setup to validate the methodology proposed. Section V reports the test results. Section VI discusses the results and Section VII finishes with concluding remarks.

II. BACKGROUND

Antibiotics are commonly used in the treatment of infections. However, the widespread and inappropriate antibiotic usage has resulted in the emergence of AMR. The problem has further been exacerbated by the empirical administration of antibiotics. To combat this issue, applying ML models to the data on antibiotics usage collected from EHRs and farm records are gaining prominence as a solution to expanding antimicrobial stewardship efforts while providing personalised treatment [4], [5]. By analysing antibiotics administration data with metadata and ML algorithms, healthcare professionals can identify patterns in the usage of antibiotics and improve prescription practices, leading to the more effective and responsible use of antibiotics and, ultimately, the reduction of AMR.

EHR contains crucial information that can be used to predict AMR, such as prior infections, resistance profiles, visits to hospitals, antibiotic treatments and treatment response history. From these data, more significant features can be identified using feature selection algorithms, and those features can be used to predict AMR and their potential influence on AMR.

A. ML-BASED AMR PREDICTION WITH ANTIBIOTICS ADMINISTRATION DATA

ML-based prediction of AMR is becoming more prevalent due to the growth of experimental and clinical data, advancements in technology, and the need for innovative solutions to reduce the impact of the disease. The ML algorithms can predict the AMR patterns for humans and livestock, allowing for a more personalised approach to antibiotics treatment using EHR and antibiotics usage data. Yet, antibiotic treatments are empirically prescribed for suspected infections before identifying the causative pathogen using biological or phenotypic approaches, which will take more than a few hours. A few of them may be inappropriate treatments that may drive the problem of antibiotic-resistant infections [1].

There have only been a limited number of studies that have aimed at predicting AMR from EHR, despite the fact that the EHR data contains crucial information that can be used to determine AMR risks, such as past infections and resistance patterns, exposure to hospitals and antibiotics, laboratory results, pathology reports, and previous treatment outcomes [6]. In another study, ML models were applied to forecast antibiotic susceptibility for various pathogen-antibiotic combinations in patients suffering from sepsis in an

intensive care unit [7]. Another similar study was published on predicting AMR for intensive care unit (ICU) patients in Greece [8].

Another system was developed to predict resistance in outpatients with Urinary tract infections (UTI) using antibiotics administration data and a small set of clinical metadata [9]. A 30% reduction in the selection of inappropriate antibiotic therapy was achieved using logistic regression and gradient-boosted decision trees [9]. A similar study focused on outpatients with uncomplicated UTIs [4].

Another study presented a decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated UTIs by applying a Gradient Boosted Trees-based ML algorithm to predict resistance in outpatient UTI patients and evaluated the impact of a treatment decision algorithm [5]. The algorithm considered both the susceptibility of the patient's urine culture and the patient's demographic information. The study found that the algorithm reduced inappropriate antibiotic selection by 30%, improving patient outcomes and reducing healthcare costs [5]. This study highlighted the importance of incorporating patient demographic information, in addition to susceptibility information, in AMR prediction algorithms. However, this study only focused on uncomplicated UTI treatment data, a common scenario. The limitations of these ML-based models in predicting AMR in complicated cases reveal the requirement for more robust and interpretable ML models that can effectively handle complex datasets and identify the contributing factors to the decisions made. This lack of ability to predict AMR in complicated cases highlights the need for more robust and interpretable ML models that can effectively handle complex datasets.

In another study, eight ML models were explored to predict resistance in Cambodian children with bacteraemia [10]. These recent studies indicate the trend of designing ML applications to predict AMR from clinical data and the need for near-time AMR prediction. Furthermore, the ML algorithm must evaluate numerous potential treatment decisions to assess the effect of various treatment options simultaneously and determine all potential treatments. This necessitates using multi-label models to predict all possible AMR susceptibility outcomes.

As far as we are aware, there is currently a lack of published studies focusing on predicting AMR, specifically from farm data. However, some literature-based frameworks are available that aim to analyse the various environmental factors contributing to AMR development [11].

DL models have improved significantly in recent years due to the growth in computing power, allowing them to achieve state-of-the-art results in many areas. Despite a few studies published on predicting AMR from clinical data [1], [4], [5], [6], [10], they are based on traditional ML approaches, and the DL approach to handling missing label scenarios has not been widely explored. Even though methods proposed in previous studies address missing labels for genomics data, these approaches are not often applied to clinical data and tend to perform poorly in imbalanced and missing data scenarios. A few studies have been published predicting AMR using DL approaches [12], but these published approaches did not consider missing labels or imbalanced data scenarios.

B. EXPLAINABLE MODELS FOR AMR PREDICTION

Though DL models perform better, DL models are often considered black box models that do not allow for a straightforward interpretation of the results. This lack of interpretability makes DL models less popular in critical fields such as medicine, where interpretability is essential. Therefore, there is a need to develop DL models that can provide interpretable results, as it is vital to identify biomarkers and use dimensionality reduction techniques in predicting AMR [7].

Existing DL models are deficient in returning the feature set and weights contributing to the classification decision, making these models non-interpretable. The opacity of DL models makes them difficult to interpret; hence they are not widely adopted in critical fields such as medicine. Increasing autonomy, complexity, and ambiguity in AI methods increases the need for interpretability, transparency, understandability, and explainability of AI output. Even though there are few kinds of research done on interpreting the results using ML approaches [2], [3], [14], [15], [16], [17], there are only very limited researches to interpret the results obtained by using clinical data [5], [18].

An ensemble-based ML approach was developed to predict antibiotic resistance of bacterial infections of hospitalised patients using the patients' EHRs [18]. This study reported the risk factors highly associated with results using the Shapely approach. In another study, a logistic regression-based decision algorithm was proposed, and the risk factors related to the models were explained using the logistic regression weights. This study uses a model-specific explanation, and the feature weights are subject to the feature value range.

Another study demonstrated the insights about results using interpretability analysis of data-driven models [19]. The post-hoc analysis of black-box models using Shapley Additive exPlanations (SHAP) was applied to interpret the model outputs which may be computationally complex in larger number of features and size of models.

The lack of interpretability in DL models is a significant limitation in fields such as medicine, where trust in the system is crucial. While DL models can achieve high accuracy, their inability to explain the decision-making process hinders their adoption. There is a balance between the accuracy and interpretability of AI models. The most accurate models, like Convolutional Neural Networks, do not provide any explanation, while more interpretable methods, like rule-based systems, are often less accurate. XAI aims to bridge this

gap by providing insight into the reasoning behind a system's predictions and decisions [20].

Hence, there is a need for a comprehensive method to overcome the issues of missing labels and to identify signatures that contribute to the prediction of multi-AMR from antibiotics administration data. To address this gap, a comprehensive method is proposed that combines the Transformer model with XAI to predict multi-drug resistance from antibiotics administration data with missing labels while also identifying the features contributing to the prediction. This approach aims to improve model interpretability, enabling healthcare stakeholders to trust the model and assess and fix any systematic bias without compromising its performance. Adding an explainable component to the DL pipeline provides model interpretability and enables healthcare stakeholders to trust the model or assess and fix any systematic bias without compromising the model's output and performance.

III. METHODOLOGY

A Transformer [21] is a type of neural network architecture using the attention concept. An attention layer is a neural network layer used to weigh and combine different features of the input data to allow the model to pay attention to other data features at different times. The Transformer architecture has been highly successful and has achieved state-of-the-art results on several natural language processing tasks. It has also been extended and modified for use in other domains, such as computer vision and speech recognition.

A. 1D TRANSFORMER

This study utilises a 1D Transformer architecture, which is well-suited for handling sequential data such as time series or text data. The model includes an attention layer to extract features from the input data, a classifier layer to make predictions, and a loss function designed to handle missing labels. A Transformer model extracts features from the input data using a multi-head attention layer, and a classifier layer made of two fully connected dense layers with dropouts is used to predict multiple AMR phenotypes using the antibiotics administration data automatically.

Different attention mechanisms can be used in Transformers, including dot-product attention, multiplicative attention, and additive attention [21]. These mechanisms typically involve calculating attention weights based on the input data and some form of context or query, then using the attention weights to weigh the input data before passing it on to the next layer of the model. The dot-product attention mechanism is computationally efficient compared to other attention mechanisms. This is because the dot-product attention is based on the dot product of the query and key matrices, which can be computed quickly and efficiently using matrix multiplication and easily parallelised for large-scale computations, as the dot-product calculation can be done independently for each position in the sequence. These advantages make dot-product attention the better choice.

B. MULTI-BASELINE INTEGRATED GRADIENT

Combining a Transformer model with an XAI pipeline has the potential to improve the accuracy of predicting antibiotic resistance and aid in developing personalised treatment plans. Identifying features and contributions informative to the classification decisions based on different feature selection approaches [16], [17], [22] is challenging since many layers are involved, and backtracking the contribution is almost impossible. There is a trade-off between AI accuracy and explainability: Normally, DL models provide no explanations; understandable ways, such as rule-based, tend to be less accurate.

Numerous XAI techniques have been developed to interpret DL model outcomes without compromising accuracy. These methods can be broadly categorised into forward-pass-based attribution and backwards-pass-based attribution. Forward-pass models, known as perturbation-based methods, are independent of any particular model and can be utilised after training on any such model. SHAP [23] and Local interpretable model-agnostic explanations (LIME) [24] are examples of this approach. Backwards-pass-based attribution methods compute the attributions for all input features in a single pass through the network. These methods are generally faster than perturbation-based methods, but the outcome may not be directly related to an output variation. The saliency method [25], Gradient $*$ Input [26] and Integrated Gradient [27] are examples of this approach.

Due to the large number of features in this study, using Shapely for feature importance calculation can be computationally expensive, as it involves calculating the contribution of all possible coalitions of feature values [28]. Additionally, the correct usage of neighbourhood selection is not well defined when using LIME with tabular data [28] and becomes even more complex with a large number of features. Model-agnostic perturbation-based methods, such as SHAP and LIME, are also more prone to instability compared to gradient-based approaches [29].

To address these issues, this study employs gradient-based approaches for interpretability. Gradient-based XAI pipelines for 1D multi-label signals refer to a method for interpreting the predictions made by an ML model that uses 1D multi-label signals as input. The gradients-based feature attribution method assigns importance scores to each feature in the input data, indicating how much each feature contributes to the model's final prediction. By combining this method with a Transformer model, the model's predictions can be made more interpretable, allowing users to understand the reasoning behind the model's decisions and identify any potential biases or errors in the model. This can be useful in applications such as predicting AMR in clinical data, where understanding and trusting the model's predictions is critical.

Integrated Gradient is a popular gradient-based method for interpreting the predictions of DL models. It generates a linear interpolation between the baseline and the original data, calculates gradients to measure the relationship between changes to a feature and changes in the model's predictions, computes the numerical approximation through averaging gradients and scales Integrated Gradient to Input. However, it has some practical issues that may affect its performance and usability:

- Baseline choice: The choice of baseline input for Integrated Gradient can significantly impact the resulting attribution values, and there is no universally agreed-upon best choice for the baseline. Using all zeros as a base for tabular data is not always appropriate because many features can take zero as the values, and those features cannot be captured through Integrated Gradient.
- Assumptions about the model: Integrated Gradient assumes that the model is linear in the input space, which may not be true for complex models such as deep neural networks.
- Sensitivity to input noise: Integrated Gradient is sensitive to small perturbations in the input, which can lead to noisy attribution maps.

To address these issues, some possible fixes, such as applying multiple baselines to assess the attribution results' robustness and mitigate the baseline choice's impact and applying methods such as SmoothGrad [30] to smooth the attribution maps and reduce the effect of input noise. This paper proposes a multi-baseline Integrated Gradient approach where all zero, all 0.5 and random value baselines are used to calculate the attributes and the average attributions are returned to overcome the baseline effect.

C. MASKED LOSS

In order to train a Transformer model with antibiotics administration data that have a large number of missing labels, a masked loss function is proposed to mask out the missing target values by introducing a Boolean mask matrix where each row represents each sample, and each column indicates whether it is a missing label value [2].

D. METRICS

During model training, metrics are utilised to monitor and evaluate its performance. However, when label values are absent and replaced with a default value, metrics based on the predicted and target values become unreliable and may generate inaccurate results. To address this, masked accuracy, masked sensitivity, masked specificity and masked f1 score are proposed to obtain justifiable values for a significant amount of missing labels [2].

By masking missing target values, the masked F1 score enables the calculation of precision and recall scores based only on available labels, conveying a balance between the two metrics. It can serve as an average score representing precision and recall [2].

The area under the curve (AUC) is another metric used to measure performance, representing the test's ability to distinguish between positive and negative cases [31]. In general, a higher AUC value indicates a better-performing model. As this study has missing values, the Masked AUC is calculated using the receiver operating characteristic (ROC) curve, which plots masked sensitivity against (1 - masked specificity) with different threshold values.

E. QUANTITATIVE VALIDATION OF XAI RESULTS AND METRICS

Quantitative validation is crucial for ensuring the reliability and generalisability of XAI methods, and it can help build trust in the models and results produced by these methods.

The Add-Delete metrics approach is used to evaluate the stability and robustness of different Integrated Gradient models. This approach assesses the impact of adding or removing a single feature from the input on the model's output. Commonly, this metric is applied to a single label. This study computed this metric for each label's top N positive and negative contributing features. The Add-Delete metrics approach aims to examine the impact of single feature changes on the model's output, thus providing insights into the stability and robustness of the Integrated Gradient models. The approach starts by computing the Integrated Gradient for each feature and selecting the top N positive and negative contributing features. Then, the Add metric was calculated by selecting only the selected feature from the input, marking other features by the baseline values, and observing the impact of this change on the model's output. Similarly, the Delete metric was calculated by removing selected features by marking them with baseline and observing the impact of this change on the model's output. These metrics were computed for each of the four labels' top N positive and negative contributing features.

AUC of the Model Relevance N-features (AUC-MRNF) is another metric proposed to validate the selected top N quantitively features through the XAI pipeline by training the model with only the best N features and calculating the AUC using those subsets of features. N features from each label are selected, and AUC-MRNF is calculated for the multi-label in this study.

IV. EXPERIMENTS

This section describes our experiment setup and results for AMR in the UTI dataset (AMR-UTI) [32], publicly available on Physionet [33]. The proposed model was implemented in Python using the Scikit-learn [34] library and TensorFlow framework [35]; our source code, the genome IDs we used for these experiments, and the pre-processed datasets are made available on GitHub.[1]

[1] https://github.com/mukunthan/TransAMR

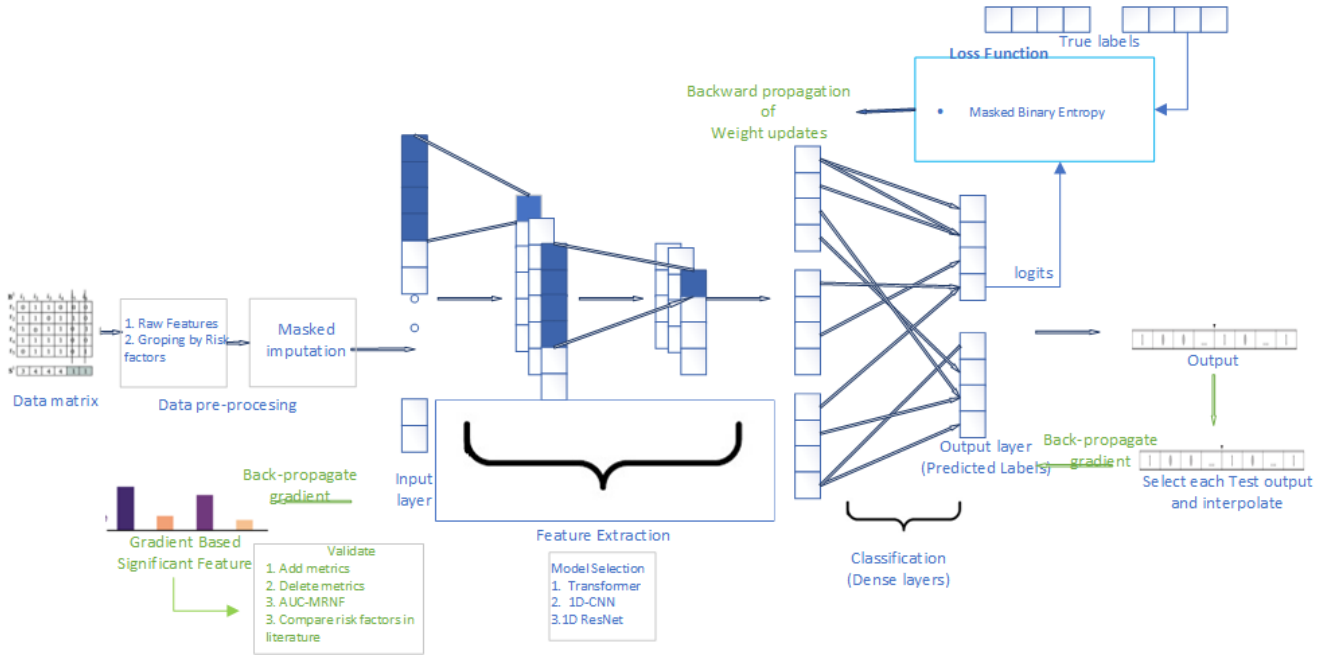

FIGURE 1. Visualisation of the steps carried out for this chapter. First, different approaches are applied to group the features and then imputed with the masking approach. Then, different models were implemented with masked-loss function to perform feature extraction and classification. Using the best model, training data and test data gradient-based explainable AI model is implemented to get the interpretation for the results.

As shown in Fig. 1, the Transformer model and a few other models with masked loss functions were applied for the pre-processed dataset with the XAI pipeline. The results from applying models have been analysed on the benchmark AMR-UTI dataset, the uncomplicated and whole dataset. Then, the best model based on the above result was compared against similar works in the literature [5].

In addition, the results from the best model were further analysed to get an explanation for the results. The key signatures contributing to the decision were reported using this model's proposed XAI pipeline, and those results are validated by comparing results in the literature.

A. DATASET AND PRE-PROCESSING

The AMR-UTI [32] dataset, publicly available on the Physionet [33] website, includes patient demographic information, antibiotic prescriptions, medical procedures, previous resistance test results or infections, comorbidities, and lab test results as ground-truth antibiotic resistance profiles for each patient. The data includes antibiotic exposures, prior antibiotic resistance, prior infections, comorbidities, and procedures over the 7, 14, 30, 90, and 180-day periods. A population-level feature called colonisation pressure is also included, defined as the proportion of resistant specimens to a given antibiotic within a specified location and time window. The dataset contains 788 features after some indicative fields such as test data, uncomplicated status, and example ID are removed.

The dataset is labelled for four antibiotics, Nitrofurantoin (NIT), Trimethoprim-sulfamethoxazole (SXT), Ciprofloxacin (CIP), and Levofloxacin (LVX), with the first two being considered first-line antibiotics and the latter two being second-line antibiotics. There are 80962 samples collected from the 2007-2013 year period and were assigned for the training and internal cross-validation, while 11865 samples out of them were collected with uncomplicated cases. 35940 samples collected from the 2014-2016 year period were designated as the independent test set, and 3941 samples out of them were collected with uncomplicated cases.

As illustrated in Supplementary Fig. 1, nearly 65 % of the labels are Susceptible, while the other 35% counts for resistant and missing values out of 80962 training data with complicated and uncomplicated instances.

The study focused on predicting AMR and developing better treatment strategies rather than tracking clinician prescriptions and aimed to support automated prescriptions in the future. The data was pre-processed to remove missing values and outliers. The categorical variables were one-hot encoded, and the data were normalised to 0 to 1. The missing labels are marked as -1 to enable masked loss and metric calculation.

In addition to EHR data, antibiotic usage data were also collected from July 2019 to August 2020 from a farm. Those data were used to start this work before obtaining the above public dataset and contained different antibiotics usage for the given period and metadata related to cattle on a farm. These data are weakly labelled as AMR and could not be identified through laboratory-based testing. Therefore, these farm antibiotics usage data were not included in reporting the results but referred to in the conclusion section to highlight

the future work needed. Ethical approval for collecting the farm data was sought and obtained via the University of Surrey – NASPA 514292-514283-58798367.

B. MODEL SELECTION

This study evaluated several models, and XAI approaches to select the best one for the task at hand. A 1D-Transformer model was developed to predict multiple AMR phenotypes using clinical data, as shown in Supplementary Fig. 2. The model comprises an input layer, a multi-head attention layer to select contributing features, and two fully connected dense layers with dropouts. Considering the fast and efficient manner of dot-product attention compared to other attention mechanisms, dot-product attention was applied to the model. As the data was in tabular format, temporal dependency was not crucial, while spatial dependence was vital in this problem. Transformers were used to capture spatial dependencies between features, and positional encoding was added to the input of the Transformer encoder to capture these dependencies. The encoded features were then passed through a fully connected layer and the output layer for classification.

Another model was a 1D-CNN, which includes an input layer, two convolution layers, two down-sampling layers, and four fully connected dense layers. The first convolution layer used 64 1D convolution kernels with a length of 7 sampling points, while the second convolution layer used 32 1D convolution kernels with 7 sampling points. The output of the convolution layers was then passed through a pooling layer to compress the selected features.

The third model was compared with a ResNet architecture with 3×3 convolution and a fixed feature map dimension of [64, 128]. This architecture bypassed the input every two convolutions, allowing the gradients to flow directly from later layers to the initial filters, mitigating the vanishing gradient problem, as shown in Supplementary Fig. 3.

Dropout regularisation is applied to the hidden layers of all models to mitigate potential overfitting. The objective of the model is to minimise the loss, and the prediction error is used to update the network parameters through back propagation using the Adaptive Moment Estimation (Adam) optimiser. The masked loss function is designed to handle missing labels and is crucial in measuring the model's error and defining how to fit the data to achieve optimal results.

C. HYPER-PARAMETER TUNING

For each tested DL model, parameters were optimised from preliminary experiments, and 0.0001 learning rate, Adam optimiser and 100 EPOCHS with early stop by monitoring loss were used as the optimal parameters based on the empirical study. Every fully connected layer of each model has been used with the Relu activation function except for the output layer, where the sigmoid activation function has been applied. The multi-head attention has been used in Transformer with two heads and an embedding dimension of 100.

D. XAI MODEL PARAMETERS

The Integrated Gradient approach establishes a baseline to compare a data instance that is typically all zero. This baseline will help the model gauge each feature's influence on the input data with respect to the prediction [28]. The gradients are summed at small intervals along the path between the baseline and original input. The method involves selecting specific points to calculate gradients, which are then summed using a number of interpolated steps (represented by ''k'') to approximate the integral of the gradients. In this study, 50 interpolated states are selected for the Integrated Gradient approach with the absence of all features as the base dataset through an empirical study. As part of multiple baselines, all zero baseline, all 0.5 value baseline and random normalised baseline were applied as a baseline to overcome baseline effects and to get robust results.

E. OPTIMAL DECISION THRESHOLD VALUE SELECTION

The decision threshold is the point at which a decision is made based on the outcome of the DL models. In predicting AMR, the optimal decision threshold is the point at which model results are used to determine whether or not a sample is susceptible or resistant, which can impact the accuracy and other metrics. Determining the optimal decision threshold value requires careful consideration of the trade-offs between sensitivity and specificity and the consequences of false positive and false negative test results. This study determined the optimal decision threshold values based on the desired balance between sensitivity and specificity. The masked geometric mean of sensitivity and specificity was calculated for different threshold values to calculate the balance between sensitivity and specificity.

Another key metric for the imbalanced dataset is the F1 score, which measures the balance between Precision and Recall. The threshold with better F1 scores can be selected as the optimum threshold. As part of this research, it was observed that thresholds for better F1 scores and the geometric mean of sensitivity and specificity were different. Therefore, a new metric combining the F1 score with the geometric mean of sensitivity and specificity was proposed to improve the overall performance. The optimal threshold was selected to maximise this metric.

V. RESULTS

This section presents the results of the evaluation experiments that were conducted. As part of this experiment, 1D-CNN, modified ResNet to support 1D-CNN and Transformer were trained with the uncomplicated AMR-UTI and whole AMR-UTI datasets separately to measure their performance and select one model to predict multi-label AMR from clinical data collected from EHR. To validate the model, the independent 3941 uncomplicated samples and 35,940

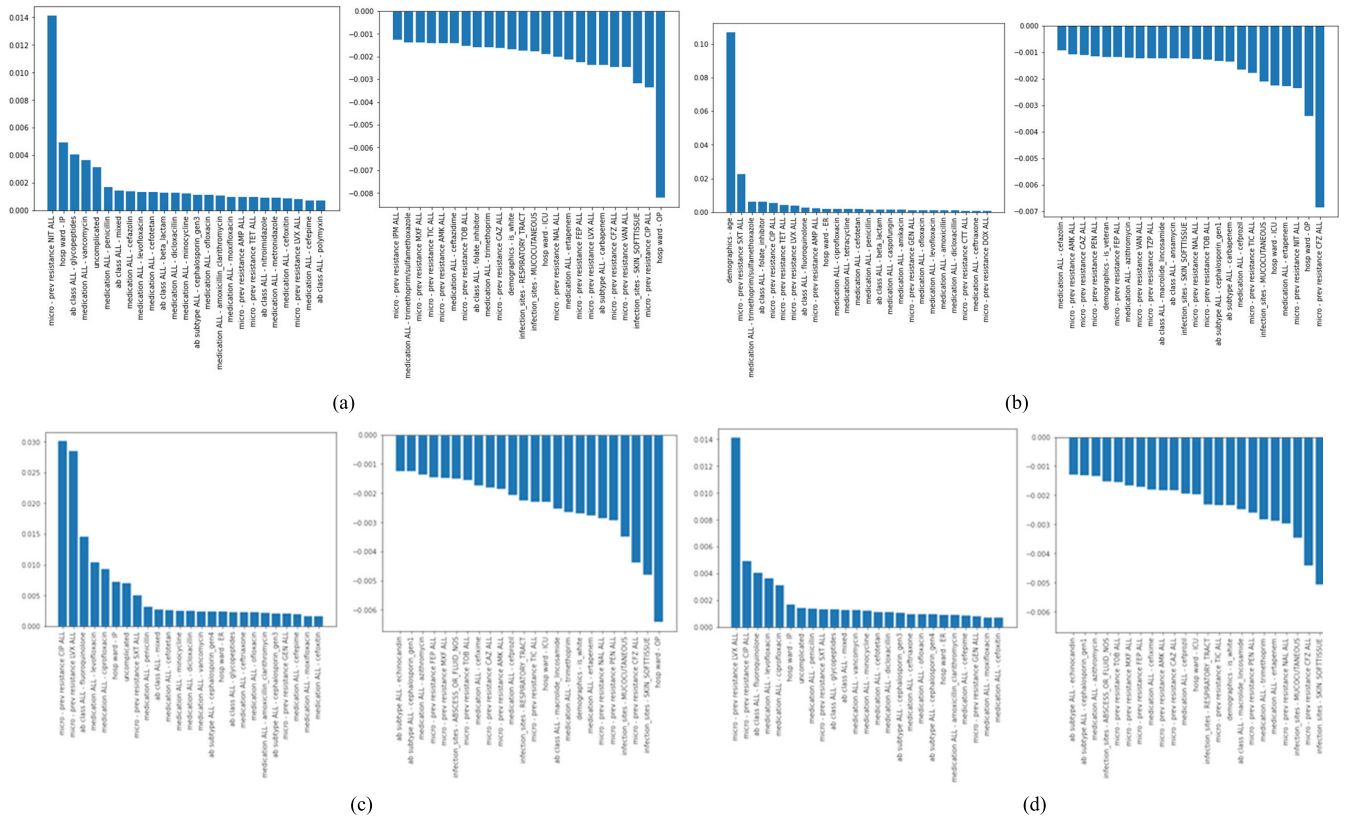

FIGURE 2. A: Features contributing to NIT prediction; the left side plot represents the top 25 features positively contributing, while the right-side plot represents the top 25 features negatively contributing to the prediction of NIT. B: Features contributing to SXT prediction; the left side plot represents the top 25 features positively contributing while the right-side table represents the top 25 features negatively contributing to the prediction of SXT. C: Features contributing to CIP prediction; the left side plot represents the top 25 features positively contributing while the right-side plot represents the top 25 features negatively contributing to the prediction of CIP. D: Features contributing to LVX prediction; the left side plot represents the top 25 features positively contributing while the right-side plot represents top 25 features negatively contributing to the prediction of LVX.

whole samples, including complicated cases, were used for all models, and the results are listed in Table 1.

Table 1 compares the performance of different DL models trained to predict the susceptibility and resistance to antibiotics used to treat uncomplicated UTIs. The models include 1D-CNN, modified ResNet, and Transformer. The models were trained using two different datasets: the uncomplicated AMR-UTI dataset and the full AMR-UTI dataset. Table 1 presents the results of 1D-CNN, modified ResNet, and Transformer models that were trained and evaluated for predicting the susceptibility and resistance to antibiotics used to treat uncomplicated UTIs. Table 1 shows the performance metrics for each model, including masked accuracy, masked sensitivity, masked specificity, masked AUC, and masked F1 score. The results were reported for the same independent 3941 uncomplicated samples and independent 35940 full test samples, including complicated cases. It should also be noted that all the models were trained with masked Entropy loss unless mentioned otherwise. Also, 15 % of the training data sets were used for internal validation to control the training.

From the results, the Transformer model performed the best overall, with the highest AUC and F1-score in most configurations. The 1D-CNN model performed relatively poorly in comparison when a larger training dataset was used. The ResNet model has poor performance in most of the datasets. It was observed that the performance of the models varied when applied to uncomplicated and full datasets, suggesting that the characteristics of the data and the amount of data may impact the models' performance. In addition, it is important to note that using more layers in a ResNet led to longer training time. Furthermore, the loss change over epoch in the training step of the proposed 1D Transformer model can be observed in Supplementary Fig. 4, and the optimum thresholds were 0.164 for NIT, 0.196 for SXT, 0.196 for CIP and 0.216 LVX and identified using an index that combined the F1-score and G-mean of specificity and sensitivity, as shown in Supplementary Fig. 5.

As Transformer gave the best result, it is compared with the best-performing linear regression model in the literature [5], as shown in Table 2. Table 2 compares the performance of the proposed Transformer model with the model in the literature using linear regression on independent test sets of 3941 uncomplicated UTI cases. The performance metrics reported in the table are the same as in Table 1. As the Linear

TABLE 1. Performance of different models. All the models were trained with Masked Entropy loss unless mentioned otherwise. 11865 samples were used for the uncomplicated AMR-UTI dataset, and 80962 samples were used with the full dataset. Results were reported for the same independent 3941 uncomplicated samples and independent 35940 full test samples, including complicated cases.

Model	Description	Masked Accuracy	Masked Sensitivity	Masked Specificity	Masked AUC	Masked F1 Score
1D-CNN	Trained with Uncomplicated Dataset and Results for Uncomplicated test set	59.97	79.25	63.44	69.72	**24.29**
ResNet	Trained with Uncomplicated Dataset and Results for Uncomplicated test set	97.93	98.89	50.56	65.98	19.73
Transformer	Trained with Uncomplicated Dataset and Results for Uncomplicated test set	84.09	92.25	54.63	**70.35**	21.15
1D-CNN	Trained with full Dataset and Results for Uncomplicated test set	19.69	41.15	85.41	70.95	29.24
ResNet	Trained with full Dataset and Results for Uncomplicated test set	52.27	75.22	66.90	65.86	25.87
Transformer	Trained with full Dataset and Results for Uncomplicated test set	15.42	36.41	88.61	**71.20**	**30.09**
1D-CNN	Trained with full Dataset and Results for full test set	46.90	73.39	72.61	75.52	47.17
ResNet	Trained with full Dataset and Results for full test set	71.47	85.94	60.35	57.68	40.75
Transformer	Trained with full Dataset and Results for full test set	49.61	76.27	71.47	**76.13**	**47.18**

The table presents the performance of three models: 1D-CNN (1-Dimensional Convolution Neural Network), ResNet (Residual Neural Network), and Transformer. The models were trained and evaluated using different configurations: trained with the uncomplicated dataset and tested on the uncomplicated test set, trained with the full dataset and tested on the uncomplicated test set, and trained with the full dataset and tested on the full test set.

TABLE 2. Comparison of performances of the proposed model and the models in the literature. 11 865 uncomplicated samples from the AMR-UTI dataset were used to train the models, and results were reported for the same independent 3,941 uncomplicated samples.

Model	Masked Accuracy	Masked Sensitivity	Masked Specificity	Masked AUC	Masked F1 score
Proposed Transformer	84.09	92.25	54.63	70.35	21.15
Linear Regression [5]	80.63	27.98	85.01	60.38	25.72

TABLE 3. Comparison of performances with the proposed Transformer model with full features and the grouped features. 11,865 uncomplicated samples from the AMR-UTI datasets were used for training, and results were reported for the same independent 35,940 full test samples, including complicated samples.

	Masked Accuracy	Masked Sensitivity	Masked Specificity	Masked AUC	Masked F1 score
with Full features	49.61	76.27	71.47	76.13	47.18
Grouped features	57.08	79.08	67.22	72.95	44.30

Regression model in the literature calculated performance for each label separately [5], results were reported as average values of those scores.

The results show that when trained with the uncomplicated dataset, the proposed Transformer model performed better than the linear regression model in the literature on masked accuracy, masked sensitivity and AUC. While the Transformer has a lower masked specificity compared to Linear Regression, it is important to consider the trade-off between sensitivity and specificity. The higher sensitivity of the Transformer model indicates its ability to accurately detect positive cases, even if it comes at the cost of reduced specificity. This can be advantageous in the context of AMR prediction, as it prioritises correctly identifying individuals at risk, potentially leading to more effective interventions. Though the Linear regression model in the literature outperformed in F1 score, the F1 score of the proposed model improved significantly and surpassed the F1 score when the full dataset was used to train, as listed in Table 2.

In summary, the proposed Transformer model outperformed the linear regression model in the literature on all performance metrics. This suggests that the Transformer model is better suited for predicting antibiotic resistance status and recommending first-line and second-line therapies to treat UTI patients.

The selected Transformer model trained with the uncomplicated AMR-UTI dataset was chosen for further experiments to identify the significant features contributing to the prediction using the Integrated Gradient XAI pipeline. The Integrated Gradient XAI pipeline was chosen because it performed well in initial experiments and is considered uncomplicated, meaning it is relatively easy to understand and interpret.

Based on the results obtained from the Integrated Gradient pipeline, it can be observed that the micro-previous resistance of each type of antibiotic plays a significant role in predicting AMR. For example, the top three features predicting AMR for Nitrofurantoin (NIT) were the micro-previous resistance of NIT, the micro-prev organism *Escherichia* 180, and the selected micro colonisation pressure CLI 90 -granular level. Similarly, the top three features predicting AMR for Trimethoprim-sulfamethoxazole (SXT) were those of a white race, Micro-previous Resistance of SXT All, and Micro-previous Resistance of CFZ all. It is also worth noting that other factors, such as antibiotics usage, demographic information, complicated treatment status, and hospital admission, also contributed to the decision-making process. This highlights the importance of considering multiple factors when predicting AMR and developing personalised patient treatment plans.

In addition, it is observed that the top three features predicting AMR of Ciprofloxacin (CIP) were Micro-previous Resistance of CIP all, Micro-previous Resistance of Levofloxacin (LVX) all, and prior fluoroquinolone use. Also, the top three features predicting AMR of LVX were the same as those predicting AMR of CIP, which shows that these antibiotics are closely related in terms of resistance. This further underlines the importance of considering multiple factors and the relationship between different antibiotics when predicting AMR and developing treatment plans. It can be noted that micro-previous resistance of the drug and previous exposures of all periods were contributed rather than 7, 14, 30, 90, and 180 days periods.

The use of grouped features was also employed to obtain a more consistent interpretation of the results, and these findings are compared with those obtained from other ML approaches reported in the literature [5]. The features were grouped based on their known risk factor domains associated with resistance. This included combining columns with different time frames for prior antibiotic resistance, prior antibiotic exposures, colonisation pressure, and hospital antibiotic usage. As a result, a new dataset was created with only 127 features, reducing the number of inconsistent and insignificant features.

Table 3 compares the performance of the proposed Transformer model when trained with the full set of features and when trained with a reduced set of features grouped based on their known risk factor domains associated with resistance. The model was evaluated using masked accuracy, sensitivity, specificity, AUC, and F1 score metrics.

TABLE 4. Comparison of Add and Delete metrics for the N top-performing features with the different integrated gradient pipelines and model trained with the grouped features.

Metrics	Integrated Gradient	InputX Gradient	Multi baseline IG	IG with Smooth Grad
Full features	69.69	69.69	69.69	69.69
Top 10 Add	67.21	67.47	68.85	68.86
Top 25 Add	67.49	67.35	69.08	69.08
Top 50 Add	68.11	67.76	69.07	69.07
Top 100 Add	68.26	68.11	69.31	69.31
Top 10 Delete	66.53	66.20	66.60	66.60
Top 25 Delete	65.76	65.80	66.09	66.09
Top 50 Delete	65.22	65.27	65.94	65.94
Top 100 Delete	65.17	65.42	65.30	65.30

All scores are reported for the independent validation set only; Independent 15% of the training data sets were used for internal validation. IG stands for Integrated Gradient. InputXGradient was calculated by multiplying the gradient with the Input.

The results show that when the Transformer model was trained with the grouped features, it achieved an overall improvement in performance compared to when it was trained with the complete set of features. Specifically, the masked accuracy increased by 7.4%, the masked sensitivity increased by 2.81%, and the masked specificity decreased by 4.25%. These results indicate that the proposed Transformer model performance did not change too much by grouping correlated features. This suggests that using a reduced set of features more relevant to the task at hand may lead to better performance in predictive models in addition to the model complexity and advantages in interpretability.

As the next steps, different interpretation methods were applied to a model's predictions, and interpretations were quantitively validated. Table 4 compares the Add and Delete metrics of the N top-performing features using different Integrated Gradient pipelines, namely Integrated Gradient, InputXGradient, multi-baseline Integrated Gradient and Integrated Gradient with SmoothedGrad. The results were validated using the model's performance on the dataset and an independent validation set, which constituted 15% of the training data set. The table's first row indicates the model's AUC when all input features were used. The remaining rows show the impact on the AUC when adding or deleting the top N important features from each label, with N being 10, 25, 50, and 100. The results indicate that the interpretation method impacted on the model's accuracy, as evidenced by the varying AUC scores between pipelines.

When trained on subsets of the top N features, the performance varies somewhat between the different methods, but the differences are not large overall. When adding

the top 25 important features, the AUC scores ranged from 67.35% to 69.07%. Similarly, when deleting the top 10 important features, the AUC scores ranged from 66.53% to 66.60%. The results suggest that the multi-baseline Integrated Gradient and Integrated Gradient with Smoothed gradients pipelines provided more consistent results across different feature addition and deletion scenarios, making them more robust and reliable in interpreting the model's predictions.

It is worth noting that the impact of Add and Delete metric was not the only factor to consider when evaluating the accuracy of the model. The complexity of the pipeline also played a role. In this case, the multi-baseline Integrated Gradient pipeline performed well as it had fewer computations than the SmoothGrad [30] approach. Therefore, this study proposes using the multi-baseline Integrated Gradient method to identify the signatures contributing to the proposed 1D Transformer model.

Based on the results obtained from the multi-baselines Integrated Gradient pipeline, it can be observed that the micro-previous resistance of each type of antibiotic played a significant role in predicting AMR, as shown in Fig. 2A, 2B, 2C and 2D. Microbial resistance to specific antibiotics and the use of certain drugs were the most significant factors in predicting AMR. Microbial resistance to second-line antibiotics, such as LVX and CIP, played a significant role in the prediction of AMR for both antibiotics as they are both fluoroquinolones. Additionally, factors such as complicated treatment status and hospital admission (inpatient or emergency room) had a positive impact, while factors such as hospital admission (outpatient), ICU admission, and being white had a negative impact. The ten most positive and negative significant features for four labels were extracted during this experiment and listed in Supplementary Table 1. Comparing these results with those in the literature [5], the features selected in the current experiment align with the previous study's findings regarding the impact of prior exposure to antibiotics and prior antibiotic resistance on AMR prediction. However, the current investigation includes additional features, such as hospital admission and uncomplicated status, which were not included in the previous study.

As having zero values for the non-significant features may be due to possible zero values in the real dataset, the AUC of the Model Relevance N-features (AUC-MRNF) was calculated to validate the selected approach. The Transformer model was trained using the most significant 25 positive and negative features extracted from each of the four labels in this approach. This resulted in a model with a total of 53 features, and the performance was evaluated using the AUC metric. The results showed that the model achieved an AUC score of 68.49%, which was only 0.20% lower than when all the features were used. This result suggests that the extracted 25 significant positive and negative features represent the overall trend in the data and that the model can still perform well with fewer features.

VI. DISCUSSION

This study clearly demonstrated that the Transformer model outperformed ResNet and 1D-CNN on Masked Accuracy, Masked Sensitivity, Masked AUC, and Masked F1 score when the uncomplicated dataset was used for training. Although the ResNet has more layers, the Transformer outperformed due to the ability of the Transformer to extract the features and the lack of sufficient data in the training phase. Generally, the proposed models achieved good performance in predicting AMR, especially with the Transformer model, which appeared to be robust to different data types. However, it also seems that there is still room for improvement in precision and F1 score. Further research should be conducted to optimise the models and apply them to more varied and representative data to improve the results.

This study applied the multi-baseline Integrated Gradient XAI pipeline to the 1D data from the AMR-UTI dataset, which contained demographic, antibiotic prescriptions, medical procedures and lab tests. The goal was to identify which features are most important for predicting antibiotic resistance and to understand how these features related to the risk factor domains associated with resistance. The Integrated Gradient pipeline returned the features contributing to the prediction of each label. It could be observed that each type's micro-previous resistance played a significant role while other factors, such as antibiotics usage, demographic information, complicated treatment status, and hospital admission, contributed next. This information can be used to control AMR spread and introduce robust antimicrobial stewardship for a common outpatient diagnosis.

The significant features include prior exposure to antibiotics, prior antibiotic resistance, and various clinical factors such as hospital admission, age, and uncomplicated status. The significant features identified in this study can have major implications for antibiotic resistance management. The findings indicate that previous resistance patterns significantly predict antibiotic resistance for UTI. Healthcare providers can use this information better to understand the likelihood of resistance in a given patient and to make more informed decisions about which antibiotics to prescribe. In addition, the discovery of the contribution of factors such as demographic information, complicated treatment status, and hospital admission highlights the need for a holistic approach to managing antibiotic resistance. These findings suggest that strategies to prevent the spread of resistance must not only focus on the responsible use of antibiotics but also on other factors that may contribute to resistance, such as patient demographics and healthcare practices.

The proposed 1D Transformer-based model offers several advantages over traditional CNN and ResNet models. The 1D Transformer-based model has been designed to capture spatial dependencies between features, which is important for correctly classifying the data. The addition of positional encoding enables the model to capture these dependencies

and perform well in predicting multiple AMR phenotypes. The 1D Transformer-based model has a higher complexity than traditional CNN and ResNet models due to the inclusion of the attention mechanism, but the improved performance of the model and future possibilities to use time series data justify the choice.

An Integrated Gradient was used in this study to identify the significant features contributing to the model's decision. Multi-baselines were used to avoid baseline effects and ensure the robustness of the feature importance estimates. This approach enabled the researchers to obtain a more accurate and reliable estimate of feature importance and gain valuable insights into the genomic features contributing to AMR prediction. While the use of Integrated Gradients can increase the computational complexity of the model, it is an invaluable tool for XAI and can provide valuable insights into the decision-making process of complex models.

Overall, these results support the importance of considering both prior exposure to antibiotics and prior antibiotic resistance and various clinical factors. These findings can inform the development of decision algorithms for promoting outpatient antimicrobial stewardship in uncomplicated UTI.

Moreover, the findings also highlight the importance of considering patients' treatment and admission status in predicting AMR. Patients undergoing complicated treatments or hospital admission, especially inpatients or emergency rooms, are more likely to have AMR. In addition, the impact of race on AMR is also an essential factor to consider.

According to the findings, being white is negatively correlated with AMR, suggesting that race may play a role in the development of AMR. This highlights the need for further research to understand the impact of race and other demographic factors on AMR.

Although colonisation pressure, an indirect measure of local resistance rates, was included as a feature in the model, it was not selected as a significant predictor. Various factors can influence local resistance rates, including antibiotic prescribing practices and regional variations in resistance patterns, which can introduce noise and make it difficult to isolate their direct impact on individual predictions. Furthermore, the model may indirectly capture the effect of local resistance rates through other correlated features in the dataset.

Furthermore, the ability of the Integrated Gradient XAI pipeline to identify the most important features contributing to the prediction of antibiotic resistance has the potential to improve the accuracy and precision of existing models for predicting resistance. By focusing on the most relevant features, the models can be optimised for better prediction and reduced false positive and false negative rates. This can lead to more efficient use of resources and a reduced risk of antibiotic resistance spreading. The results demonstrate that the proposed multi-baseline Integrated Gradient approach can effectively identify the most important features in a DL model and can be a valuable tool for feature selection and model interpretation.

VII. CONCLUSION

In this study, data-driven prescription strategies are proposed to identify AMR and prescribe optimal treatment to treat UTIs, which will help to reduce broad-spectrum antibiotic use. The results of this paper suggest that the proposed Transformer-based model and Integrated Gradient XAI pipeline are effective tools for predicting antibiotic resistance and providing personalised treatment recommendations for patients with UTIs. The study highlights the potential of using DL models and XAI techniques for predicting antibiotic resistance and improving patient outcomes. In addition, interpretable models were able to identify the features contributing to the decision, which may support the decision as a second opinion. This reduces the risk of choosing unnecessarily broad antibiotics and increases the chances of successfully treating the infection.

However, there are also limitations to this study. The limited representation of different patient populations and demographics is one limitation. Further research is needed to validate the findings with diverse patient populations and cultural backgrounds to ensure the model is generalisable and applicable to a wide range of patients and antibiotics that enhance its clinical usefulness. In addition, this study only focused on predicting the antibiotic resistance status of UTI patients and recommending first-line and second-line therapies. However, in practice, a patient may be diagnosed with multiple conditions, and antibiotics may affect the resistance status of other bacterial infections. Therefore, future research should explore the potential of using this approach to predict antibiotic resistance for multiple bacterial infections and develop a comprehensive treatment plan that considers the potential interactions between different antibiotics.

In conclusion, this study has demonstrated the potential for using advanced ML techniques to predict antibiotic resistance and provide personalised treatment recommendations. The proposed models effectively identify significant features, reduce the complexity of feature selection, and provide interpretable predictions that can help clinicians make informed decisions. While there are limitations to the approach, such as the need for high-quality clinical data and significant investment in resources, the potential benefits are substantial. By supporting antimicrobial stewardship efforts and improving human and animal patient outcomes, this approach casn significantly impact the fight against antimicrobial resistance.

ACKNOWLEDGMENT

The authors would like to thank the farms for providing pilot data for the study.

REFERENCES

[1] M. Paul, S. Andreassen, E. Tacconelli, A. D. Nielsen, N. Almanasreh, U. Frank, R. Cauda, and L. Leibovici, "Improving empirical antibiotic treatment using TREAT, a computerized decision support system: Cluster randomized trial," *J. Antimicrobial Chemotherapy*, vol. 58, no. 6, pp. 1238–1245, Dec. 2006, doi: 10.1093/jac/dkl372.

[2] M. Tharmakulasingam, B. Gardner, R. L. Ragione, and A. Fernando, "Explainable deep learning approach for multilabel classification of antimicrobial resistance with missing labels," *IEEE Access*, vol. 10, pp. 113073–113085, 2022, doi: 10.1109/ACCESS.2022.3216896.

[3] M. Tharmakulasingam, B. Gardner, R. La Ragione, and A. Fernando, "Rectified classifier chains for prediction of antibiotic resistance from multi-labelled data with missing labels," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 1, pp. 625–636, Jan. 2023, doi: 10.1109/TCBB.2022.3148577.

[4] S. Boominathan, M. Oberst, H. Zhou, S. Kanjilal, and D. Sontag, "Treatment policy learning in multiobjective settings with fully observed outcomes," 2020, *arXiv:2006.00927*.

[5] S. Kanjilal, M. Oberst, S. Boominathan, H. Zhou, D. C. Hooper, and D. Sontag, "A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection," *Sci. Transl. Med.*, vol. 12, no. 568, pp. 1–10, Nov. 2020, doi: 10.1126/scitranslmed.aay5067.

[6] M. N. Anahtar, J. H. Yang, and S. Kanjilal, "Applications of machine learning to the problem of antimicrobial resistance: An emerging model for translational research," *J. Clin. Microbiol.*, vol. 59, no. 7, pp. 1–22, Jun. 2021, doi: 10.1128/JCM.01260-20.

[7] D. Ghosh, S. Sharma, E. Hasan, S. Ashraf, V. Singh, D. Tewari, S. Singh, M. Kapoor, and D. Sengupta, "Machine learning based prediction of antibiotic sensitivity in patients with critical illness," medRxiv, Tech. Rep. 19007153, 2019. [Online]. Available: http://medrxiv.org/content/early/2019/09/20/19007153.abstract, doi: 10.1101/19007153.

[8] G. Feretzakis, E. Loupelis, A. Sakagianni, D. Kalles, M. Martsoukou, M. Lada, N. Skarmoutsou, C. Christopoulos, K. Valakis, A. Velentza, S. Petropoulou, S. Michelidou, and K. Alexiou, "Using machine learning techniques to aid empirical antibiotic therapy decisions in the intensive care unit of a general hospital in Greece," *Antibiotics*, vol. 9, no. 2, p. 50, Jan. 2020, doi: 10.3390/antibiotics9020050.

[9] I. Yelin, O. Snitser, G. Novich, R. Katz, O. Tal, M. Parizade, G. Chodick, G. Koren, V. Shalev, and R. Kishony, "Personal clinical history predicts antibiotic resistance of urinary tract infections," *Nature Med.*, vol. 25, no. 7, pp. 1143–1152, Jul. 2019, doi: 10.1038/s41591-019-0503-6.

[10] M. Oonsivilai, Y. Mo, N. Luangasanatip, Y. Lubell, T. Miliya, P. Tan, L. Loeuk, P. Turner, and B. S. Cooper, "Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in Cambodia," *Wellcome Open Res.*, vol. 3, p. 131, Oct. 2018, doi: 10.12688/wellcomeopenres.14847.1.

[11] B. Gardner et al., "Mapping the evidence of the effects of environmental factors on the prevalence of antibiotic resistance in the non-built environment: Protocol for a systematic evidence map," *Environ. Int.*, vol. 171, Jan. 2023, Art. no. 107707, doi: 10.1016/j.envint.2022.107707.

[12] X. Kuang, F. Wang, K. M. Hernandez, Z. Zhang, and R. L. Grossman, "Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN," *Sci. Rep.*, vol. 12, no. 1, p. 2427, Feb. 2022, doi: 10.1038/s41598-022-06449-4.

[13] A. Drouin, S. Giguère, M. Déraspe, M. Marchand, M. Tyers, V. G. Loo, A.-M. Bourgault, F. Laviolette, and J. Corbeil, "Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons," *BMC Genomics*, vol. 17, no. 1, p. 754, Dec. 2016, doi: 10.1186/s12864-016-2889-6.

[14] J. Kim, D. E. Greenberg, R. Pifer, S. Jiang, G. Xiao, S. A. Shelburne, A. Koh, Y. Xie, and X. Zhan, "VAMPr: Variant mapping and prediction of antibiotic resistance via explainable features and machine learning," *PLOS Comput. Biol.*, vol. 16, no. 1, Jan. 2020, Art. no. e1007511, doi: 10.1371/journal.pcbi.1007511.

[15] E. S. Kavvas, L. Yang, J. M. Monk, D. Heckmann, and B. O. Palsson, "A biochemically-interpretable machine learning classifier for microbial GWAS," *Nature Commun.*, vol. 11, no. 1, p. 2580, May 2020, doi: 10.1038/s41467-020-16310-9.

[16] M. Tharmakulasingam, C. Topal, A. Fernando, and R. La Ragione, "Improved pathogen recognition using non-Euclidean distance metrics and weighted kNN," in *Proc. 6th Int. Conf. Biomed. Bioinf. Eng.*, Nov. 2019, pp. 118–124.

[17] M. Tharmakulasingam, C. Topal, A. Fernando, and R. L. Ragione, "Backward feature elimination for accurate pathogen recognition using portable electronic nose," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2020, pp. 1–5.

[18] O. Lewin-Epstein, S. Baruch, L. Hadany, G. Y. Stein, and U. Obolski, "Predicting antibiotic resistance in hospitalized patients by applying machine learning to electronic medical records," *Clin. Infectious Diseases*, vol. 72, no. 11, pp. e848–e855, Jun. 2021, doi: 10.1093/cid/ciaa1576.

[19] S. Martínez-Agüero, C. Soguero-Ruiz, J. M. Alonso-Moral, I. Mora-Jiménez, J. Álvarez-Rodríguez, and A. G. Marques, "Interpretable clinical time-series modeling with intelligent feature selection for early prediction of antimicrobial multidrug resistance," *Future Gener. Comput. Syst.*, vol. 133, pp. 68–83, Aug. 2022, doi: 10.1016/j.future.2022.02.021.

[20] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 11, nos. 3–4, pp. 24:1–24:45, Aug. 2021, doi: 10.1145/3387166.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[22] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[23] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, *arXiv:1705.07874*.

[24] M. Tulio Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," 2016, *arXiv:1602.04938*.

[25] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[26] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," 2016, *arXiv:1605.01713*.

[27] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, Jul. 2017, pp. 3319–3328. [Online]. Available: https://proceedings.mlr.press/v70/sundararajan17a.html

[28] C. Molnar. *Interpretable Machine Learning*. Accessed: Jun. 20, 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[29] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," 2018, *arXiv:1806.08049*.

[30] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[31] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[32] M. Oberst, S. Boominathan, H. Zhou, S. Kanjilal, and D. Sontag, "AMR-UTI: Antimicrobial resistance in urinary tract infections," PhysioNet, Tech. Rep., 2020. [Online]. Available: https://physionet.org/content/antimicrobial-resistance-uti/1.0.0/, doi: 10.13026/SE6W-F455.

[33] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000, doi: 10.1161/01.cir.101.23.e215.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2012.

[35] B. Pang, E. Nijkamp, and Y. N. Wu, "Deep learning with TensorFlow: A review," *J. Educ. Behav. Statist.*, vol. 45, no. 2, pp. 227–248, Apr. 2020, doi: 10.3102/1076998619872761.

MUKUNTHAN THARMAKULASINGAM (Member, IEEE) received the B.Sc. degree (Hons.) in electronics and telecommunications engineering from the University of Moratuwa, Sri Lanka, in 2014. He is currently pursuing the Ph.D. degree with the Centre for Vision, Speech, and Signal Processing (CVSSP), University of Surrey, U.K. He is a Researcher with King's College London on interpretable multi-modal lung cancer prediction. Before the Ph.D. research, he was a Lecturer (Probationary) with the University of Jaffna, Sri Lanka, and a Software Engineer with LSEG Technology. His current research interest includes applying explainable artificial intelligence techniques in healthcare.

WENWU WANG (Senior Member, IEEE) received the B.Sc., M.E., and Ph.D. degrees from Harbin Engineering University, China, in 1997, 2000, and 2002, respectively. He was with King's College London, from 2002 to 2003, Cardiff University, from 2004 to 2005, Tao Group Ltd. (now Antix Labs Ltd.), from 2005 to 2006, and Creative Labs, from 2006 to 2007, before joining the University of Surrey, in May 2007. He is currently a Professor of signal processing and machine learning and the Co-Director of the Machine Audition Laboratory, Centre for Vision Speech and Signal Processing, University of Surrey, U.K. He is also an AI Fellow with the Surrey Institute for People-Centred Artificial Intelligence. He has been involved as a Principal or Co-Investigator in more than 30 research projects, funded by the U.K. and EU research councils, and industry (e.g., BBC, NPL, Samsung, Tencent, Huawei, Saab, Atlas, and Kaon), with a total grant portfolio of over £30M. His current research interests include audio-visual signal processing, machine learning, and perception. He has (co)authored over 300 papers in these areas. He is the (co)author or (co-)recipient of over 15 awards, including the 2022 IEEE Signal Processing Society Young Author Best Paper Award, the ICAUS 2021 Best Paper Award, the DCASE 2020 Judge's Award, the DCASE 2019 and 2020 Reproducible System Award, and the LVA/ICA 2018 Best Student Paper Award. He is the elected Chair of the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee, the Vice Chair of the EURASIP Technical Area Committee on Acoustic Speech and Music Signal Processing, an elected member of the IEEE Signal Processing Theory and Methods Technical Committee, and an elected member of the International Steering Committee of Latent Variable Analysis and Signal Separation. He was the Satellite Workshop Co-Chair of INTERSPEECH 2022, the Publication Co-Chair of IEEE ICASSP 2019, the Local Arrangement Co-Chair of IEEE MLSP 2013, and the Publicity Co-Chair of IEEE SSP 2009. He is a Senior Area Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING, an Associate Editor of IEEE/ACM TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING, an Associate Editor of *Scientific Report* (Nature), and a Specialty Editor-in-Chief of *Frontiers in Signal Processing*.

MICHAEL KERBY received the degree from the University of Bristol, in 1985. He has been in farm animal practice, since 1985. He started in Malmesbury, Wiltshire, for four years, then via North Dorset joined the Delaware Veterinary Group, Castle Cary, Somerset, in 1992, becoming a Partner, in 1994. He has done consultancy and presented extensively at home and abroad. He joined Synergy Farm Health Ltd., Evershot, Dorset, in June 2020. He was an Honorary Lecturer with the University of Bristol. He is currently an Honorary Lecturer with The University of Liverpool. He is also a member of the Nottingham Dairy Innovation Forum, Nottingham University. His main research interests include dairy practice fertility, nutrition, and production, and especially all bovine surgery whilst at the same time developing the next generation of cattle vets. He received the DBR from The University of Liverpool, in 1994.

ROBERTO LA RAGIONE received the degree in 1995, the master's degree in veterinary microbiology from the RVC, and the Ph.D. degree from Royal Holloway. He is currently a Professor of veterinary microbiology and pathology with the School of Veterinary Medicine and the Head of the School of Biosciences, University of Surrey. His current research interests include AMR and understanding the pathogenesis of food-borne pathogens, with a particular interest in developing control and intervention strategies.

ANIL FERNANDO (Senior Member, IEEE) received the B.Sc. degree (Hons.) in electronics and telecommunication engineering from the University of Moratuwa, Sri Lanka, in 1995, the M.Sc. degree (Hons.) in communications from the Asian Institute of Technology, Bangkok, Thailand, in 1997, and the Ph.D. degree in computer science (video coding and communications) from the University of Bristol, U.K., in 2001. He is currently a Professor of video coding and communications with the Department of Computer and Information Sciences, University of Strathclyde, U.K. He leads the Video Coding and Communication Research Team, University of Strathclyde. He has been working with all major EU broadcasters, BBC, and major European media companies/SMEs in the last decade in providing innovative media technologies for British and EU citizens. He has graduated more than 110 Ph.D. students and is also supervising 20 Ph.D. students. He has worked on major national and international multidisciplinary research projects and led most of them. He has published over 420 papers in international journals and conference proceedings and published a book on 3D video broadcasting. His main research interests include video coding and communications, machine learning (ML), artificial intelligence (AI), semantic communications, signal processing, networking and communications, interactive systems, resource optimizations in 6G, distributed technologies, media broadcasting, and quality of experience (QoE).