IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 September 2022, accepted 9 October 2022, date of publication 25 October 2022, date of current version 2 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3216896

RESEARCH ARTICLE

Explainable Deep Learning Approach for Multilabel Classification of Antimicrobial Resistance With Missing Labels

MUKUNTHAN THARMAKULASINGAM[1], **(Member, IEEE), BRIAN GARDNER**[2], **ROBERTO LA RAGIONE**[2,3], **AND ANIL FERNANDO**[4], **(Senior Member, IEEE)**

[1]Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH Guildford, U.K.
[2]School of Veterinary Medicine, University of Surrey, GU2 7XH Guildford, U.K.
[3]School of Biosciences and Medicine, University of Surrey, GU2 7XH Guildford, U.K.
[4]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 by the University of Surrey through its Library Open Access funds and Zoetis through the vHive Initiative.

ABSTRACT Predicting Antimicrobial Resistance (AMR) from genomic sequence data has become a significant component of overcoming the AMR challenge, especially given its potential for facilitating more rapid diagnostics and personalised antibiotic treatments. With the recent advances in sequencing technologies and computing power, deep learning models for genomic sequence data have been widely adopted to predict AMR more reliably and error-free. There are many different types of AMR; therefore, any practical AMR prediction system must be able to identify multiple AMRs present in a genomic sequence. Unfortunately, most genomic sequence datasets do not have all the labels marked, thereby making a deep learning modelling approach challenging owing to its reliance on labels for reliability and accuracy. This paper addresses this issue by presenting an effective deep learning solution, Mask-Loss 1D convolution neural network (ML-ConvNet), for AMR prediction on datasets with many missing labels. The core component of ML- ConvNet utilises a masked loss function that overcomes the effect of missing labels in predicting AMR. The proposed ML-ConvNet is demonstrated to outperform state-of-the-art methods in the literature by 10.5%, according to the F1 score. The proposed model's performance is evaluated using different degrees of the missing label and is found to outperform the conventional approach by 76% in the F1 score when 86.68% of labels are missing. Furthermore, the ML-ConvNet was established with an explainable artificial intelligence (XAI) pipeline, thereby making it ideally suited for hospital and healthcare settings, where model interpretability is an essential requirement.

INDEX TERMS Multilabel classification, deep neural network, multi-drug AMR, missing labels, explainable AI.

I. INTRODUCTION

Antimicrobial resistance (AMR) is a critical issue for global health, food security and economics [1]. The World Health Organization (WHO) has listed AMR as one of the three most critical health issues of the 21st century [2]. Estimates indicate that there will be 10 million deaths per year globally, costing \$100 trillion by 2050, if no actions are taken to tackle the rising trend in AMR [1]. To tackle AMR, antibiotic usage must be appropriately managed. For this, it is vital to identify AMR at an earlier stage. Conventional antimicrobial susceptibility testing based on microbiological culture is widely used in clinical practice. A caveat of this approach is that it requires professional facilities, skilled expertise and is viable only for cultivable bacteria [3].

Fortunately, promising solutions to address these concerns include developing novel strategies to identify AMR presence in bacteria using machine learning trained on

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

genomic sequences. This approach is faster than the conventional lab-based approach [4] and facilitates a personalised treatment plan, thus avoiding unnecessary antibiotic use that would otherwise exert selective pressure for resistance emergence. The development of Next-Generation Sequencing (NGS) technologies and increasing Graphical Processing Unit (GPU) accelerated data processing capability have enabled rapid and more cost-friendly prediction of AMR with machine learning [5], [6], [7], [8].

A single bacterium can be resistant to many antibiotic drugs simultaneously; therefore, for AMR prediction methods to be practical, they must identify multiple AMRs present in a single genomic sequence. This multi-resistance for a single genomic sequence makes it a multilabel dataset. Even though increasing genomic data availability and machine capabilities make a case for machine learning-based prediction, many ground truth AMR phenotypes are missing in multi-labelled datasets. They need to be manually labelled with complex and time-consuming microbiological experiments. As the genomes were sequenced at different times and laboratory-based experiments were done only for the resistance of selected antibiotics, the resistances of other antibiotics are unknown and treated as missing labels [8]. This can result in poor performance as current deep learning models depend on correctly labelled data. Another concern regarding deep learning models is that they act as black-box models and provide no interpretability of the results derived from the models. In this paper, we propose the Mask-Loss 1D convolution neural network (ML-ConvNet) model to overcome the issues arising from missing labels. To overcome the interpretability issues, an explainable artificial intelligence (XAI) framework is applied with the proposed model to predict multiple AMR phenotypes so that domain experts can analyse the various features learned by these models. A few validation experiments are performed to ensure the proposed model has overcome the missing label issue. The proposed XAI pipeline with the proposed ML-ConvNet is also validated by identifying significant features from the pipeline, using them to make predictions and comparing the performance. In summary, the major contributions of this article can be identified as follows:

1) Proposing a novel deep learning-based model for the multilabel AMR classification of genomic sequences with missing labels.
2) Establishing an XAI framework for AMR multilabel classification and validating the framework.
3) Validating the proposed model and XAI pipeline by performing experiments with different levels of missing labels and evaluating the selected features in the XAI pipeline.

The remainder of the paper is structured as follows. Section 2 covers the background of this study, section 3 describes the proposed loss function and metrics to measure the performance, section 4 describes the models used to measure the performance, section 5 reports the test results, section 6 discusses the results and section 7 finishes with concluding remarks.

II. BACKGROUND

A. DEEP LEARNING FOR AMR PREDICTION

Genes are functional units of genome sequences that contribute to an organism's phenotypic traits. Therefore, identifying genes associated with AMR from the genomic sequence data supports phenotype prediction in pathogens. As part of gene identification, existing annotated genome sequences available from public databases[1,2] are used to annotate genes in a new sequence [9] with the help of tools such as the Basic Local Alignment Search Tool (BLAST) [10]. Once genes are extracted from a genomic sequence, this data is treated as input features for a machine learning approach to enable phenotypic prediction [6]; Several studies have been undertaken using machine learning algorithms to predict AMR from annotated genes [8], [11], [12], [13]. More genomics data is now available with the advancement in next-generation genomic sequencing. This availability of large-scale genomic sequences and advances in GPU processing capability has opened a path for deep learning applications using genomic sequences to predict AMR [14].

A deep learning system was proposed to predict genes related to AMR, with two different implementations for short-read sequences and long-read sequences [14]. A high precision and Recall of 97% and 91%, respectively, were achieved using this system for the data collected from the Comprehensive Antibiotic Resistance Database (CARD), Antibiotic Resistance Genes Database (ARDB), and UNIversal PROTein Resource (UNIPROT) database [14]. Even though these models identify different antibiotic resistance genes (ARG) from genome reads, they do not explicitly predict AMR types as gene presence does not directly imply resistance to any specific type of antibiotic.

A Wide and Deep Neural Network (WDNN) incorporating logistic regression and a deep multilayer perceptron (MLP) was proposed to improve Tuberculosis (TB) prediction with better sensitivity and accuracy compared to regularised Logistic Regression and Random Forest methods [15]. This paper used a multitask deep learning architecture that can classify resistance across multiple drugs and share information across different anti-tubercular drugs and genes to provide more accurate phenotypic predictions. Despite this, it can only identify ten different anti-tuberculosis drugs from whole-genome sequencing data, and low reproducibility and high variance results were observed for this model.

Multidrug resistance is becoming a critical issue in human and animal health systems [14]. In the context of the AMR multilabel dataset, each label is treated as a binary classification since it is represented as one of two values: 'Resistance' or 'Susceptible.' Other values such as 'Intermediate' and 'Susceptible-dose dependent' are typically converted

[1] ftp://ftp.patricbrc.org/
[2] https://ftp.ncbi.nih.gov/genomes/

to 'Resistant' and 'Susceptible', respectively. The standard approach in the literature trains a separate binary classifier for each AMR label by splitting the original multilabel problem into many single-label problems. Therefore, multilabel classification models must be considered to predict whether genome sequences are susceptible or resistant to many types of antibiotics. To our best knowledge, there has been little effort in applying multilabel classification models to predict multi-AMR types using a single model.

Comparatively, few studies have been published on applying multilabel classification methods to genomics data for predicting multiple types of AMR [8], [16], [17], [18, p.], [19], [20], [21]. DeepGo [21] and DeepGoplus [22] are two popular methods applied to predicting multilabel protein classes from genomic sequences using deep learning methodologies; however, they do not predict AMR specifically nor handle missing labels as part of their operation.

B. MULTILABEL CLASSIFICATION WITH MISSING LABELS

In multilabel learning, each data sample is assigned multiple class labels simultaneously, and only a partial label set can be observed for some real applications, especially for a genomic dataset. The performance of multilabel learning approaches is significantly influenced by label incompleteness, as models are built assuming all labels are present [23]. The approaches used to overcome this problem can be categorised as follows [23].

- **Preprocessing approaches**: These approaches impute the missing label first and then use the new complete set of labels to train the multilabel classification problem.
- **Transductive approaches**: They impute the missing label matrix by creating a matrix with all the features and labels from all the data and then filling the missing entries of this matrix by applying matrix completion methods.
- **Synchronised approaches**: The missing labels are recovered while simultaneously training a multilabel classifier by determining the correlations between labels.

The label imputation is achieved based on label consistency, label smoothness and statistical means in preprocessing approaches. If too many missing labels exist, these approaches may result in noisy labels and consequently poor performance. Transductive approaches are incapable of inductive reasoning, causing practical challenges in applying these approaches [23]. The synchronised approaches impute the missing labels by utilising label correlations derived from incomplete labels [23].

A joint approach using both positive and negative label correlations and the locality of data information was proposed to impute the missing labels to overcome the above issue and ensure that similar data instances will have identical class labels [23]. Another method named LSML was also described to learn label-specific features by creating a new supplementary label matrix augmented from the incomplete label matrix and learning high-order label correlations. Hence, a label-specific data representation for each class label can be found, and the learned high-order label correlations are incorporated to impute the missing labels [24].

A multilabel classification probabilistic model with label correlations and missing labels (LCML) was proposed to deal with missing labels effectively and automatically exploit the label correlations. This approach was inspired by the label transformation approach, but expressed in the original label space rather than the transformed label space [25, p.]. This modification allowed flexibility in handling both label dependencies and missing labels. A limitation of this model is that it only considers pairwise, symmetric, and positive label correlations.

When there are a larger number of missing labels, correlation-based approaches tend to be incorrect, and these approaches can then result in poor performance. Therefore, novel approaches are required to handle the erroneous correlation-based approach when there are many missing labels.

Despite existing studies on handling missing labels on multilabel prediction, missing label scenarios for multidrug AMRs are considered in only a few research studies [8]. In this study, Rectified Classifier Chain (RCC) method for predicting multidrug resistance was proposed by internally modifying the existing classifier chain approach to handle missing labels. Each label classifier in that model is trained using only existing labels, and missing labels are internally predicted. Those predicted labels are used as features and an existing label for other label predictions [8]. Even though this method internally handles missing labels, this approach cannot be applied to deep learning approaches. A few studies predicting AMR using deep learning approaches have been published [26]. However, missing label scenarios were not considered in the published approaches.

C. EXPLAINABLE MODELS FOR AMR PREDICTION

Despite a number of challenges, deep learning models and other machine learning techniques remain appealing tools to identify AMR. Identifying biomarkers from genomic sequences contributing to the predictions and applying dimensional reduction techniques for this data is crucial for achieving higher accuracy in predicting AMR from genome sequence data [7]. Although certain machine learning models give distinct feature sets that domain experts can further interpret [28], [29], deep learning models act as black-box models whose results cannot be easily interpreted. Existing deep learning models are deficient in returning the feature set and weights contributing to the classification decision, thereby making hard to interpret the model and results. The opacity of deep learning 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 Machine learning approaches [8], [30], [31], only a limited number of research studies have sought to interpret the results obtained by deep learning to our knowledge [32].

A platform consisting of a deep convolutional neural network (DCNN) model for resistance diagnosis and a support vector machine (SVM) model as a surrogate to identify resistance genes and mutations was developed [32]. These studies applied the deep convolutional neural network model for the prediction and SVM models to get the significant features separately. Even though they identified significant features for the SVM results, it cannot be proven that those features contributed to the results obtained by the DCNN. Otherwise, this work was only conducted for Mycobacterium tuberculosis to predict resistance to the Pyrazinamide drug.

Hence, there is a need for a comprehensive method to overcome the issues arising from missing labels and to identify biomarkers contributing to the multi-AMR prediction from deep learning models and metrics that are used to measure performance on an imbalanced dataset. Therefore, we propose a Mask-Loss convolution neural network (ML-ConvNet) model to predict multidrug resistance with missing labels along with the explainable AI pipelines with RAST [33], [34] based annotated *Escherichia coli* (*E. coli*) genomic data to improve classifier accuracy and interpretability.

III. PROPOSED MODEL

Deep neural networks' stacked and hierarchical learning system efficiently captures complex relationships between high-dimensional, spatial or consequential features [35]. A deep neural network is a network of nodes constituting multiple layers, where nodes with different layers are connected with the weight of edges and biases.

A convolution neural network is one type of deep neural network, consisting of convolutional and down sampling layers, mainly performing two tasks, feature extraction and classification. Convolution layers can be considered as fuzzy filters, which enhance features while reducing noise. The down sampling layer reduces genomic matrix feature data's dimension and preserves useful features. Therefore, the convolution and down sampling hidden layers automatically extract compelling features from the feature matrix, while the dense final layer classifies the data accurately by using the extracted features as part of the classification.

The convolution neural network is conventionally applied to two-dimensional data, especially images. As genomic sequence features are one-dimensional data, convolution layer and down sampling should be changed to one dimensional

Deep neural networks are trained by estimating the optimal values of the biases and edge weights, minimising the difference between the true and predicted values of the labels. The function used to minimise this difference is termed the loss function [36], and the model's performance is estimated based on the loss value. Cross-entropy is a common choice of loss function for deep neural networks, which measures the difference between two probability values of true labels and predicted labels, as shown in Eq. (1).

$$\text{Loss}\,(e_i) = -\frac{1}{n}\sum_{j\in n} y_{i,j}\log\left(\widehat{y_{ij}}\right) + \left(1-y_{i,j}\right)\log\left(1-\widehat{y_{i,j}}\right) \tag{1}$$

where n is the number of scalar values in the prediction, $(\widehat{y_{i,j}})$ is the j^{th} scalar value in the i^{th} predicted label and $y_{i,j}$ is the j^{th} scalar value of the i^{th} true label. When the model is trained using batches of samples, the averaged loss function is defined as in Eq. (2)

$$\text{Average Loss} = \frac{1}{m}\sum_{i\in m} e_i \tag{2}$$

where m is the size of the batch and e_i is the loss calculated for the i^{th} instance in the batch based on Eq. (1).

The cross-entropy loss tends to be zero when the predicted probability of the data classification approaches that of the actual class [37] and increases if the data classification is erroneous. The binary cross-entropy is a type of cross-entropy where the labels can take only one of two values: 1 or 0 [38]. Yet, when the label values are missing and imputed with a default value, the loss between the target value and the predicted value will be affected. Therefore, the model performance may be impacted. In this work, we propose to use a masked loss function, including a selection of alternative evaluation and monitoring metrics to overcome the effect of missing values.

A. MASKED LOSS FUNCTION

In order to train a deep learning model with genomic samples that have a large number of missing labels, we propose a masked loss function to mask out the missing target values in Eq. (1) by introducing a boolean mask matrix where each column indicates whether it is a missing a particular label value. Hence, this mask matrix is created by setting its elements as -1 which corresponds to entries with missing labels in the multilabel dataset. Accordingly, the mask is defined for each sample as follows:

$$\text{Mask}(m_{i,j}) = \begin{cases} 0 & y\,(i,j) == -1 \\ 1 & y\,(i,j)! = -1 \end{cases} \tag{3}$$

where *y(m,n)* is the n^{th} true label value of the m^{th} sample. Once the mask matrix is calculated, it masks out the predicted and true values at the same index with a missing target value in the ground truth values.

As a result, the masked predicted value $(\widehat{my_{i,j}})$ as defined in Eq. (4) and masked-true value $(my_{i,j})$ as defined in Eq.(5) for the missing labels will be assigned to zero as $m_{i,j}$ will be zero. Here, $*$ represents the element-wise multiplication

$$\text{Masked Predicted label}\ (\widehat{my_{i,j}}) = \widehat{y_{i,j}} * m_{i,j} \tag{4}$$

$$\text{Masked True label}\,(my_{i,j}) = y_{i,j} * m_{i,j} \tag{5}$$

Therefore, the error between the target and predicted values for the missing value indices will be zero in the masked loss defined in Eq. (6) and will not affect the masked binary cross-entropy value.

$$\text{Masked loss} = -\frac{1}{n}\sum_{j\in n} my_{i,j}\log\left(\widehat{my_{ij}}\right) + \left(1 - my_{i,j}\right)\log\left(1 - \widehat{my_{i,j}}\right) \quad (6)$$

B. MASKED METRICS

Metrics are used to monitor and measure the performance of a model during training. When the label values are missing and imputed with a default value, metrics based on the target and predicted values are incorrect and may report incorrect metrics values. In order to get correct metrics values for a large number of missing labels, we propose masked metrics where missing values in the true value and relevant predicted value will be omitted in the metrics calculations.

The Masked Accuracy (MA) is calculated as shown in Eq. (8), where N refers to the total number of samples and M the total number of available labels as it avoids comparing the missing labels with the predicted values in measuring the performance.

$$I_{A==B} = \begin{cases} 1 & A = B \\ 0 & A \neq B \end{cases} \quad (7)$$

$$\text{MA} = \frac{1}{N}\sum_{i\in sample}\frac{1}{M}\sum_{j\in label} I_{(\widehat{y_{i,j}}*m_{i,j}==y_{i,j}*m_{i,j})} \quad (8)$$

The predicted values are elementary multiplied with the masked matrix and then with the true labels to derive the Masked True Positive (MTrP) metric by counting the number of correctly predicted positive labels (1) in that multiplied matrix as defined in Eq. (9). The Masked-Total Positives (MToP) is the total number of true positive labels (1), and it is calculated by counting the number of positive labels (1) in the matrix derived by multiplying the true label and masking matrices to mask out the missing label as defined in Eq. (10). The Masked-Predicted Positives (MPrP) is the total number of predicted positive labels (1), and it is calculated by counting the number of positive labels (1) in the metrix derived by multiplying the predicted label and masking matrices to mask out the missing label as defined in Eq. (11). Masked Precision (MPr) is the proportion of Masked True Positive (MTrP) and Masked-Predicted Positives (MPrP) as defined in Eq. (12), and Masked Recall (MRe) is the proportion of Masked True Positive (MTrP) and Masked-Total Positives (MToP) as defined in Eq. (13). The Masked F1 score conveys the balance between the Precision and the Recall, as shown in Eq. (14). It can be used as a score that can be used as an average of both precision and recall scores [47]. These metrics were calculated by considering only the available labels and masking out the missing ones.

$$\text{MTrP}_\text{i} = \sum_{j\in labels} I_{(\widehat{y_{i,j}}*y_{i,j}*m_{i,j}==1)} \quad (9)$$

$$\text{MTotP}_\text{i} = \sum_{j\in labels} I_{(y_{i,j}*m_{i,j}==1)} \quad (10)$$

$$\text{MPrP}_\text{i} = \sum_{j\in labels} I_{(\widehat{y_{i,j}}*m_{i,j}==1)} \quad (11)$$

$$\text{MPr}_\text{i} = \frac{\text{MTrP}_\text{i}}{\text{MPrP}_\text{i}} \quad (12)$$

$$\text{MRe}_\text{i} = \frac{\text{TrP}_\text{i}}{\text{MTotP}_\text{i}} \quad (13)$$

$$\text{Masked F1 score} = \sum_{i\in samples}\left(\frac{2\times\text{MPr}_\text{i}\times\text{MRe}_\text{i}}{\text{MPr}_\text{i} + \text{MRe}_\text{i}}\right) \quad (14)$$

C. EXPLAINABLE AI PIPELINE

Identifying features and contributions informative to the classification decisions is challenging since many layers are involved, and backtracking the contribution is substantially difficult. There is a trade-off between AI accuracy and explainability: frequently, deep learning methods provide limited explanations; interpretable methods, such as rule-based schemes, tend to be less accurate.

Explainable AI (XAI) seeks to resolve this issue. An XAI system is a self-explanatory intelligent system that describes the reasoning behind its decisions and predictions [39]. Adding an explainable component to our deep learning pipeline provides model interpretability and enables the stakeholders to gain trust in a model or may be used to assess and fix the systematic bias in our model without compromising the model output and performance.

Many XAI methods are introduced to interpret the results without reducing the accuracy. These methods can generally be divided into two main categories: forward-pass-based attribution and backwards-pass-based attribution. Forward-pass attribution is model-agnostic, and can be applied to any machine learning model after training. The following approaches are taken on the forward pass approach, which can be referred to as the input-based attribution method.

- Taking input data.
- Making some adjustments to the input (such as partial occlusion or perturbing some values).
- Observing the effect on the predictions.

Shapley Additive exPlanations (SHAP) [40] and Local interpretable model-agnostic explanations (LIME) [41] are examples of this approach. The SHAP values measure the contributions of each feature in the model and interpret the predicted values using the Shap values of each input feature. SHAP values can be used as global interpretability since SHAP captures how much each predictor contributes to the target variable, either positively or negatively [40], [42]. The SHAP approach requires much computing time and memory when there are many features as it tries different combinations of features to get the contribution of each feature.

LIME measures the changes to the predictions when you give variations of your data into the machine learning model. It generates a new dataset consisting of perturbed samples, gets the corresponding black-box model's corresponding

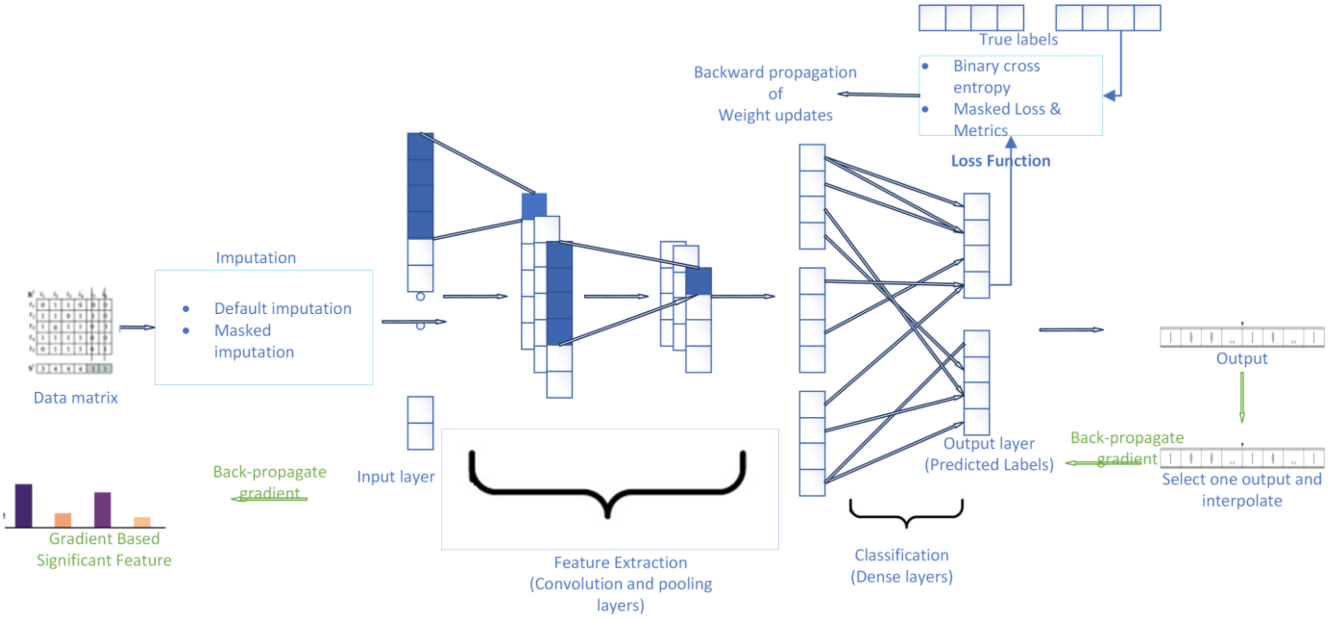

True labels
Backward propagation of Weight updates
Binary cross entropy
Masked Loss & Metrics
Loss Function
Imputation
Default imputation
Masked imputation
Data matrix
Output
Back-propagate gradient
Select one output and interpolate
Output layer (Predicted Labels)
Input layer
Back-propagate gradient
Gradient Based Significant Feature
Feature Extraction (Convolution and pooling layers)
Classification (Dense layers)

FIGURE 1. Visualisation of the steps carried out for this research. First, Missing labels are imputed in different ways. Then, different loss functions are defined including masking loss function and then 1D-CNN model is implemented to carry out feature extraction and classification. Using the trained model, training data and test data gradient based explainable AI model is implemented to get the interpretation for the results.

predictions, and uses those to interpret the models [1]. It normally interprets individual predictions.

Another approach is Backpropagation-based methods which compute the attributions for all input features in a single forward and backwards pass through the network. While these methods are generally faster than perturbation-based methods, their outcome can hardly be directly related to an output variation. Saliency method [41], gradient * input [42] and integrated gradient [43] are examples of this approach. The saliency method interprets the results by calculating the gradient of the output with respect to the input, and gradient * input interprets the results by performing an element-wise product of the input and the gradient. The integrated gradient is defined as the integral of the gradients along the straight-line path from baseline data and the input [43].

D. VALIDATING EXPLAINABLE AI PIPELINE

Considering the 16345 features in this study, Shapely is computationally expensive because 2^{16345} possible coalitions of the feature values are there for SHAP calculations [43]. Correct usage of the neighbourhood is not well defined when using LIME with tabular data [43] and is complex when many features exist. The model-agnostic perturbation-based methods, such as SHAP and LIME, are more prone to instability than the gradient-based approaches [44].

Therefore, gradient-based approaches are applied for this study. The saliency map, gradient * input and integrated gradient approaches are evaluated to understand the proposed ML-ConvNet model decisions and select the topmost significant features. The variation of gradient values within samples and the percentage of performance reduction with the first 500 significant features are introduced as the metrics to validate the results.

IV. EXPERIMENTATIONS WITH THE PROPOSED METHODOLOGY

This section describes our experiment setup and results for *Escherichia coli* (*E. coli)* and *Salmonella* annotated genomic datasets publicly available on PATRIC. The proposed model was implemented in Python using the Scikit-learn [45] library and TensorFlow framework [46]; our source code, the genome IDs we used for these experiments, and the preprocessed datasets are made available on GitHub.[3]

As shown in Fig. 1, deep learning models with modified loss functions were applied for the preprocessed dataset with the explainable AI pipeline. The results from applying different loss functions and models were analysed on the benchmark PATRIC dataset with different levels of missing labels. Then, the best model based on the above result was compared against similar works in multilabel classification and multilabel AMR prediction from Protein ID in the literature [15], [30], [33].

Following these steps, our model with the best performing base classifier was further analysed to get an explanation for the results. The key biomarkers contributing to the decision were reported using this model's proposed Explainable AI pipeline.

A. DATASET

The PATRIC database [47] is one of the most comprehensive for antibiotic resistance where genomes are

[3] https://github.com/mukunthan/ML-ConvNet

annotated using RAST [34]: the Rapid Annotations using Subsystems Technology. AMR genes may not be suitable for use when genomes are incomplete [13]. Therefore, protein genus-specific families that were identified using the RAST annotations for 2775 genomic feature files marked against 32 AMR [8] were used in our experiments. At the end of this preprocessing, 16345 Protein genus-specific families (PLfams) were extracted as input features, and a binary-valued matrix was created by indicating the presence/absence of protein genus-specific families for each genome sequence [8].

In the data preprocessing step, strains labelled with intermediate levels of resistance and susceptible-dose dependent labels were considered as the missing values in the masked approach.

Two approaches were followed to analyse the impact of the missing label ratio.

1. Select a specific set of a dataset and measure the performance of the model
2. Labels were randomly removed to make them NaN, and experiments were conducted with different ratios of labels

B. MODEL SELECTION

As part of this study, a few models were explored based on similar work done in the literature [14], [26]. Those models were then modified to suit this study through empirical studies. The first model is a 1D-CNN, consisting of an input layer, two convolution layers, two down sampling layers and four fully connected dense layers. 64 1D convolution kernels with a length of 7 sampling points are used in the first convolution layer, while the second convolution layer is built with 32 1D convolution kernels with a length of 7 sampling points. The outcome of the convolution layers was sent through the pooling layer to compress selected features.

Another model is an ANN consisting of an input layer and four fully connected dense layers with dropouts to predict multiple AMR phenotypes using all the features given in the input.

If the deep neural network has several layers, then the training process takes much time and needs a larger number of datasets. In addition, the prediction performance becomes saturated with an increasing number of hidden layers due to the gradient degradation problem [48]. As our study has only a limited number of data instances, we limited our models to smaller layer models rather than modifying other larger deep learning models to make them support our data.

C. HYPER-PARAMETER TUNING

For each tested deep learning model, parameters were optimised from preliminary experiments@comm and optimal parameters were selected, as shown in Table 1. In the integrated gradient approach, a baseline is established 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 [43]. The gradients are summed at small intervals along the path between the baseline and original input. This provides the points at which the gradients are found and then summed with k as the number of interpolated steps 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.

TABLE 1. Hyper-parameter overview of two deep learning architectures explored in this study.

Hyper parameter	1D CNN	ANN
Learning rate	0.001	0.001
Optimizers	Adam	Adam
Batch Size	32	32
Total Training Epoch	50	50
Early-Stop monitor	Masked Accuracy	Masked Accuracy
Early-Stop Patience	5	5

D. EVALUATION METRICS

Accuracy, Precision, Recall, and F1 scores are evaluation metrics commonly used for multilabel classification. Since AMR data is imbalanced with respect to their labels, it is vital to measure the recall and precision metrics to determine the performance of a predictive algorithm. As defined in section III B, masked Accuracy, masked Precision, masked Recall and masked F1 scores are used as the evaluation metrics in this experiment.

V. RESULTS

This section presents the results of the evaluation experiments that were conducted. The masked Accuracy, masked Precision, masked Recall and masked F1 score for the different models were analysed, and the results are summarised in Table 2.

The proposed masked loss function performed best regarding both models' masked Accuracy, masked Recall and masked F1 score. The binary entropy loss function performed better in terms of masked Precision in both models. Moreover, the proposed ML-ConvNet outperformed the default 1D-CNN F1 score by 7%, Recall by 14% and Accuracy by 3%, when the missing ratio is 68.91%, as given in Supplementary File 1. Here, the missing ratio is calculated by dividing the total number of missing labels by total label counts.

It is observed that Precision was reduced by a small margin, which was mainly caused due to larger false positives prediction returned from the proposed ML-ConvNet model than the models with default imputation. As that default imputation approach had more negative samples in training, there was less chance of creating a False Positive (FP) than the proposed method. At the same time, the True Positive (TP) increment in the proposed method was small compared

TABLE 2. Masked Accuracy for different Deep learning models with different loss function are reported for dataset with different ratio of missing labels. Original E. coli dataset with 32 labels has 68.91% missing label and other were created by randomly removing labels to test the effect of the models on dataset with higher number of missing labels.

Method	Percentage of Missing labels (%)	ANN				1D-CNN			
		Masked Accuracy	Masked Precision	Masked Recall	Masked F1 score	Masked Accuracy	Masked Precision	Masked Recall	Masked F1 score
Default imputation + Binary entropy	68.91	73.83±1.66	**87.20±3.56**	16.34±4.08	27.16±0.55	83.53±0.89	83.58±0.18	64.11±2.30	72.20±1.63
Impute mask value +Masked Loss	68.91	82.75±1.41	72.60±3.13	69.26±3.83	70.48±2.39	**88.67±0.69**	82.55±1.62	**79.45±1.70**	**80.87±1.08**
Default imputation + Binary entropy	78.21	70.32±1.34	36.24±41.41	2.07±2.97	3.84±5.43	79.98±1.54	**88.72±2.21**	38.77±4.30	53.22±4.20
Impute mask value +Masked Loss	78.21	82.41±1.43	73.18±3.11	67.75±6.68	69.74±3.67	**88.11±0.62**	82.61±2.00	**77.58±2.05**	**79.74±1.32**
Default imputation + Binary entropy	87.68	69.51±1.73	0.37±1.38	9.5e-05 ±0.04	0.02±0.07	69.56±1.76	5.56±16.60	0.18±0.57	0.35±1.10
Impute mask value +Masked Loss	87.68	80.76±1.778	71.67±3.70	60.63±8.07	64.87±5.46	**85.98±0.84**	**77.88±2.92**	**75.22±2.49**	**76.06±1.50**
Default imputation + Binary entropy	93.88	70.28±2.15	0	0	0	70.28±2.15	0	0	0
Impute mask value +Masked Loss	93.88	79.89±2.26	71.55±4.44	57.59±8.27	62.71±6.35	**84.70±1.26**	**76.22±4.28**	**73.30±3.57**	**73.78±2.47**
Default imputation + Binary entropy	96.83	68.55±2.07	0.37±1.38	0.02±0.09	0.04±0.16	68.54±2.07	0	0	0
Impute mask value +Masked Loss	96.83	75.37±2.34	64.39±6.56	40.16±5.08	47.61±4.62	**81.79±1.62**	**72.17±3.93**	**65.64±4.8**	**66.91±3.64**

ANN –Artificial Neural Network, 1D-CNN- 1 Dimensional Convolution Neural Network
Percentage of Missing labels is calculated by diving total number of missing labels by total number of labels (Total same* number of labels per sample)
All scores are reported by mean of the 3 times repeated k-fold (5-fold) experiment with standard deviation as the error. (I.e., mean ± standard deviation).

to the default approach. Therefore, Total positives were also increasing, and Masked Precision, as defined in Eq (12), was not increasing due to the proposed approach; instead, decreasing when the missing label ratio was nominal. Yet, when the missing ratio had increased, the TP increment was significant in the proposed method compared to the default approach. Therefore, the Precision of the proposed method started to increase compared to the default method when the missing label ratio increased.

In terms of 1D-CNN over the ANN, the 1D-CNN model outperformed the ANN model by 9% in masked F1-score, 12% in masked Recall, 5% in masked Precision, and 5% in masked Accuracy.

Furthermore, different missing label ratios were synthetically achieved by randomly removing the label, and the performance of ML-ConvNet became more apparent when the missing ratio increased. For the synthetically created dataset with an 87.68% missing ratio, as given in Supplementary File 2, the ML-ConvNet outperformed the 1D-CNN with default imputation and loss function by more than 75% in F1 and Recall, by 72% in Precision and by 16% in Accuracy. The proposed ML-ConvNet model returned 85.98% accuracy, 77.88% precision, 73.30% recall and 73.78% F1-score for 93.88% of missing ratio data, while the default imputation-based 1D-CNN and ANN gave 0% Precision, Recall and F1-score while 71.67% Accuracy. The ML-ConvNet returned

TABLE 3. Masked Accuracy for different Deep learning models with different loss function are reported for dataset with different ratio of missing labels. Original Salmonella dataset with 22 labels has 48.66% missing label and other were created by randomly removing labels to test the effect of the models on dataset with higher number of missing labels.

Method	Percentage of Missing labels (%)	ANN				1D-CNN			
		Masked Accuracy	Masked Precision	Masked Recall	Masked F1 score	Masked Accuracy	Masked Precision	Masked Recall	Masked F1 score
Default imputation + Binary entropy	48.66	87.42±1.04	**81.15**±1.83	58.65±4.62	67.60±3.29	93.13±0.46	87.50±1.37	81.76±2.14	84.33±1.11
Impute mask value +Masked Loss	48.66	88.01±0.83	79.15±2.21	64.41±5.57	70.59±3.61	93.42±0.52	87.92±1.45	82.84±2.40	85.09±1.10
Default imputation + Binary entropy	64.21	81.65±1.71	87.59±2.32	24.49±6.71	37.33±7.80	90.34±0.62	91.08±1.24	64.76±2.28	75.37±1.56
Impute mask value +Masked Loss	64.21	87.29±0.94	79.01±2.09	60.86±6.09	68.15±3.64	92.92±0.54	86.66±1.99	82.14±2.39	84.09±1.19
Default imputation + Binary entropy	79.49	77.15±0.83	0.53±1.36	0.02±0.06	0.04±0.11	77.42±0.88	26.31±27.6 6	1.23±1.36	2.33±2.56
Impute mask value +Masked Loss	79.49	85.96±1.54	77.91±3.91	54.66±7.54	63.37±5.15	92.33±0.72	85.81±1.75	80.22±2.62	82.56±1.49
Default imputation + Binary entropy	89.78	77.05±1.08	0	0	0	77.05±1.08	0	0	0
Impute mask value +Masked Loss	89.78	84.83±1.19	76.59±3.63	46.04±5.94	56.14±5.10	91.04±0.75	82.44±2.57	76.17±1.94	78.40±1.53

ANN –Artificial Neural Network, 1D-CNN- 1 Dimensional Convolution Neural Network
Percentage of Missing labels is calculated by diving total number of missing labels by total number of labels (Total same* number of labels per sample)
All scores are reported by mean of the 3 times repeated k-fold (5-fold) experiment with standard deviation as the error. (I.e., mean ± standard deviation).

81.79% Accuracy, 72.17% Precision, 65.64% Recall and 66.91% F1-score for the synthetically created dataset with 96.83% missing ratio data. These results illustrate the significant improvement in the performance of the proposed model over the standard approach when there are a large number of missing labels. The default imputation-based 1D-CNN and ANN returned almost all the predictions as zero, which led to nearly 70% accuracy and very minimal Precision, Recall and F1-score due to almost 70% zero label data due to default imputation.

These models were also tested with the Salmonella dataset with 22 labels, and similar performance improvement was observed, as listed in Table 3. Smaller improvement was observed in terms of Accuracy, Recall and F1-score with the ML-ConvNet with the 48.66% missing ratio data set, while considerable improvements were observed when the missing ratio is high. As observed with the *E.coli* dataset, Precision was reduced when the missing percentages were 48.66% and 64.21%, while improvement was observed for larger percentage of missing labels.

TABLE 4. Comparison of masked Accuracy and F1 Score accuracy with state of art different multi-label method with Genus protein genes.

Method	Our Proposed ML-ConvNet		XGBoost based RCC (MR)[8]	
	Masked Accuracy	Masked F1 Score	Masked Accuracy	Masked F1 Score
E. coli dataset	88.14±0.74	**80.41±1.49**	**90.70±0.70**	69.76±0.69

RCC (MR) – Rectified Classifier chain with Missing label, ML-1D CNN- Masked loss based 1d CNN . These performances are measured as mean values in each fold in 5-fold validation steps.

Since ML-ConvNet provided the best result, this model was compared with other models in the literature. In this experiment, our proposed method outperformed the best performing semi supervised method in the literature, Rectified Classifier chain with Missing label (RCC (MR)) [8]

TABLE 5. Comparison of different explainable AI models performance.

Metrics	*Number of Non-Zero Features*	*Mean of Variance*	*Masked Accuracy with all features*	*Masked Accuracy with top 500 Significant features*	*Masked F1-Score with all features*	*Masked F1-Score with top 500 Significant features*
Integrated gradient [42]	**10466**	3.26e-06	88.67±0.69	**88.88±0.77**	**80.87±1.08**	**80.80±1.31**
Saliency method (Gradient) [40]	14436	4.70e-06	88.67±0.69	88.38±0.82	80.87±1.08	80.07±1.30
Gradient Input [41]*	11749	**1.45e-06**	88.67±0.69	88.56±0.72	80.87±1.08	80.17±1.12

by 10.7% in the F1 score for the *E. coli* dataset, as shown in Table 4.

The proposed ML-ConvNet model and *E. coli* dataset were selected for further experiments to ascertain the significant features of classification with the saliency method, gradient*input and integrated gradient XAI pipelines. During this experiment, significant features were extracted based on different pipelines and validated by measuring the model's performance with the dataset with the 500 top-performing features in each pipeline, as shown in Table 5. It can be observed that all three pipelines perform nearly similar, though the integrated gradient's masked Accuracy was little improved. The integrated gradient pipeline returned a smaller number of significant features and considerably low variance of the values during different experiments. Even though all three performed well, the integrated gradient pipeline was selected as it gave consistent features with lower variance and a little better performance with the chosen features than other pipelines. Therefore, the Integrated gradient pipeline was implemented as the XAI pipeline and the significant features contributing are identified as shown in Fig. 2 and Supplementary Table 3.

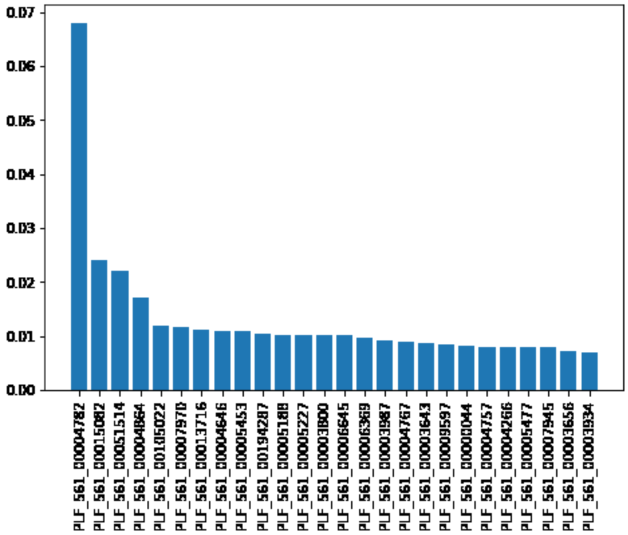


FIGURE 2. Significant Features contributing to ML- ConvNet model decision were identified by calculating integrated gradient values for each test and finding the mean for all the test cases. Clear description of each feature is given in Supplementary Table 3.

VI. DISCUSSION

As reported in the results section, the proposed ML-ConvNet model outperformed default imputation and the best-performing model [8] with the same data in the literature. During the training, the proposed masked loss approach ignored the missing values in calculating loss, while the default approach imputed a default value and used this value for the loss function. Therefore, the masked loss approach performed better when there are a larger number of missing labels in predicting labels on this *E. coli* and *Salmonella* AMR dataset, and the masking approach performed well when there were many missing labels. When comparing the models' performance, 1D-CNN performed better than the ANN approach as 1D-CNN use convolution layers and max-pooling layers to select more suitable features.

Few methodologies have been proposed for predicting labels in the literature on predicting AMR from the Genus Protein dataset. Yet, better F1 scores were not achieved with those methodologies. Therefore, we introduced a deep learning approach to gain a better F1 score without compromising Accuracy. As reported in Table 4, our ML-ConvNet model outperformed the next best model for the *E. coli* dataset in terms of F1 score, which is more meaningful for an imbalanced dataset. The performance improvement was significantly great with the increase of the missing ratio until it reached a considerable number of labels. As our study had only a limited number of data instances and the performance of larger models may get reduced compared to small layer models, we limited our models to smaller layer models.

In the explainable pipeline, gradient approaches were selected since forward-pass-based approaches are unsuitable due to the higher number of features, consequently making them more unstable. Even though saliency map, gradient*input and integrated gradient gave a similar performance, the integrated gradient approach gave somewhat more consistent results compared to other approaches since getting the mean of the gradient with the interpolated dataset stabilised the selection of features in the integrated gradient approach.

VII. CONCLUSION

Predicting multiple AMR phenotypes has recently received significant research interest. However, a more comprehensive explainable deep learning model to predict different multilabel methods with many missing labels is still lacking in the literature. Here, we proposed an explainable ML-ConvNet model to handle multilabel classification for genus-specific protein feature data with many missing labels and evaluated its performance with an appropriate selection of metrics.

This study demonstrated that the proposed ML-ConvNet model could provide a high F1 score compared with other models in the literature. Furthermore, we have extensively worked on different explainable AI approaches and identified a suitable XAI framework for this study. To the best of our knowledge, this study is the first to apply multilabel deep learning methods to handling missing labels, analyse explainable AI pipelines and report the most significant features for protein annotated datasets. The identified features will help reduce the annotation complexity used to identify AMR and provide new knowledge on biomarkers identifying multiple AMRs. As there are thousands of features in the genomic feature dataset, it is essential to identify important features to avoid overfitting and improve the AMR prediction results. Newly identified features allow scientists to analyse the contributions of those genome subsets in AMR prediction. In addition, we have explored a few metrics to measure explainable AI performance, which can also be utilised to measure the performance of explainable AI models in other fields.

Even though the proposed explainable ML-ConvNet approach performed better, it has a few limitations in this study. These experiments were conducted with binary classification models; however, this approach can also be extended to address multiclass classifications by changing the binary-entropy loss function to categorical entropy, which can support multiclass prediction. Our study focused on predicting AMR from annotated Protein genus-specific families (PLfams); however, these annotations require high computational power and laboratory-based experiments to obtain reference genomes. Therefore, identifying and analysing the genomes using the k-mer approach [33] should help mitigate these limitations. Other than this, point mutations which are not associated with any proteins or genes may also cause AMR. Further studies are needed to capture point mutation for AMR prediction.

Our study only explores the effect of introducing masked-based loss function with the improved deep learning models used in the literature. Further exploration is needed to identify the impact of the proposed approach on the larger deep learning models such as ResNet, AlexNet, VGGNet and InceptionNet by modifying those architectures to support 1-D data and retraining with genomic data.

It must be noted that only a few explainable AI models were analysed, and significant features contributing to the overall decision were identified. In the future, further approaches might be explored to identify significant features contributing to each label and the inter-feature interaction effect.

ACKNOWLEDGMENT

The authors would like to thank Dr. Arnoud van Vliet and Dr. Guido Cordoni from the School of Veterinary Medicine, University of Surrey, for their guidance in gaining bioinformatics domain knowledge.

REFERENCES

[1] N. R. Naylor, R. Atun, N. Zhu, K. Kulasabanathan, S. Silva, A. Chatterjee, G. M. Knight, and J. V. Robotham, "Estimating the burden of antimicrobial resistance: A systematic literature review," *Antimicrobial Resistance Infection Control*, vol. 7, no. 1, p. 58, Apr. 2018, doi: 10.1186/s13756-018-0336-y.

[2] *Antimicrobial Resistance: Global Report on Surveillance*, World Health Org., Geneva, Switzerland, 2014.

[3] M. Boolchandani, A. W. D'Souza, and G. Dantas, "Sequencing-based methods and resources to study antimicrobial resistance," *Nature Rev. Genet.*, vol. 20, no. 6, pp. 356–370, Jun. 2019, doi: 10.1038/s41576-019-0108-4.

[4] P.-J. Van Camp, D. B. Haslam, and A. Porollo, "Bioinformatics approaches to the understanding of molecular mechanisms in antimicrobial resistance," *Int. J. Mol. Sci.*, vol. 21, no. 4, p. 1363, Feb. 2020, doi: 10.3390/ijms21041363.

[5] A. Drouin, F. Raymond, G. L. St-Pierre, M. Marchand, J. Corbeil, and F. Laviolette, "Large scale modeling of antimicrobial resistance with interpretable classifiers," Dec. 2016, *arXiv:1612.01030*. Accessed: Apr. 13, 2019.

[6] E. Avershina et al., "AMR-Diag: Neural network based genotype-to-phenotype prediction of resistance towards β-lactams in *Escherichia coli* and *Klebsiella pneumoniae*," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 1896–1906, Jan. 2021, doi: 10.1016/j.csbj.2021.03.027.

[7] Y. Ren, T. Chakraborty, S. Doijad, L. Falgenhauer, J. Falgenhauer, A. Goesmann, A.-C. Hauschild, O. Schwengers, and D. Heider, "Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning," *Bioinformatics*, vol. 38, no. 2, pp. 325–334, Jan. 2022, doi: 10.1093/bioinformatics/btab681.

[8] 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.*, early access, Feb. 7, 2022, doi: 10.1109/TCBB.2022.3148577.

[9] D. Wheeler and M. Bhagwat, *BLAST QuickStart*. Totowa, NJ, USA: Humana Press, 2007. Accessed: Aug. 11, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK1734/

[10] S. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.

[11] E. S. Kavvas, E. Catoiu, N. Mih, J. T. Yurkovich, Y. Seif, N. Dillon, D. Heckmann, A. Anand, L. Yang, V. Nizet, J. M. Monk, and B. O. Palsson, "Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance," *Nature Commun.*, vol. 9, no. 1, p. 4306, Oct. 2018, doi: 10.1038/s41467-018-06634-y.

[12] H.-L. Her and Y.-W. Wu, "A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the *Escherichia coli* strains," *Bioinformatics*, vol. 34, no. 13, pp. i89–i95, Jul. 2018, doi: 10.1093/bioinformatics/bty276.

[13] M. Nguyen, R. Olson, M. Shukla, M. VanOeffelen, and J. J. Davis, "Predicting antimicrobial resistance using conserved genes," *PLOS Comput. Biol.*, vol. 16, no. 10, Oct. 2020, Art. no. e1008319, doi: 10.1371/journal.pcbi.1008319.

[14] G. Arango-Argoty, E. Garner, A. Pruden, L. S. Heath, P. Vikesland, and L. Zhang, "DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data," *Microbiome*, vol. 6, no. 1, pp. 1–15, Dec. 2018, doi: 10.1186/s40168-018-0401-z.

[15] M. L. Chen, A. Doddi, J. Royer, L. Freschi, M. Schito, M. Ezewudo, I. S. Kohane, A. Beam, and M. Farhat, "Deep learning predicts tuberculosis drug resistance status from genome sequencing data," *bioRxiv*, Jun. 2018, Art. no. 275628, doi: 10.1101/275628.

[16] D. Heider, R. Senge, W. Cheng, and E. Hüllermeier, "Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction," *Bioinformatics*, vol. 29, no. 16, pp. 1946–1952, Aug. 2013, doi: 10.1093/bioinformatics/btt331.

[17] F. Javed and M. Hayat, "Predicting subcellular localization of multi-label proteins by incorporating the sequence features into Chou's PseAAC," *Genomics*, vol. 111, no. 6, pp. 1325–1332, Dec. 2019, doi: 10.1016/j.ygeno.2018.09.004.

[18] S. Kouchaki, Y. Yang, A. Lachapelle, T. M. Walker, A. S. Walker, T. E. A. Peto, D. W. Crook, D. A. Clifton, and C. Consortium, "Multi-label random forest model for tuberculosis drug resistance classification and mutation ranking," *Frontiers Microbiol.*, vol. 11, p. 667, Apr. 2020.

[19] X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, and K.-C. Chou, "IAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Anal. Biochem.*, vol. 436, no. 2, pp. 168–177, May 2013, doi: 10.1016/j.ab.2013.01.019.

[20] Y. Ren, T. Chakraborty, S. Doijad, L. Falgenhauer, J. Falgenhauer, A. Goesmann, O. Schwengers, and D. Heider, "Multi-label classification for multi-drug resistance prediction of *Escherichia coli*," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 1264–1270, Jan. 2022, doi: 10.1016/j.csbj.2022.03.007.

[21] M. Kulmanov, M. A. Khan, and R. Hoehndorf, "DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier," *Bioinformatics*, vol. 34, no. 4, pp. 660–668, Feb. 2018, doi: 10.1093/bioinformatics/btx624.

[22] M. Kulmanov and R. Hoehndorf, "DeepGOPlus: Improved protein function prediction from sequence," *Bioinformatics*, vol. 36, no. 2, pp. 422–429, Jul. 2019, doi: 10.1093/bioinformatics/btz595.

[23] R. Rastogi and S. Mortaza, "Multi-label classification with missing labels using label correlation and robust structural learning," *Knowl.-Based Syst.*, vol. 229, Oct. 2021, Art. no. 107336, doi: 10.1016/j.knosys.2021.107336.

[24] J. Huang, F. Qin, X. Zheng, Z. Cheng, Z. Yuan, W. Zhang, and Q. Huang, "Improving multi-label classification with missing labels by learning label-specific features," *Inf. Sci.*, vol. 492, pp. 124–146, Aug. 2019, doi: 10.1016/j.ins.2019.04.021.

[25] W. Bi and J. Kwok, "Multilabel classification with label correlations and missing labels," in *Proc. AAAI Conf. Artif. Intell.*, 2014, vol. 28, no. 1, pp. 1–7.

[26] 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, Feb. 2022, Art. no. 1, doi: 10.1038/s41598-022-06449-4.

[27] 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, pp. 1–15, Dec. 2016, doi: 10.1186/s12864-016-2889-6.

[28] 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.

[29] 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.

[30] 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.

[31] 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, pp. 1–11, Dec. 2020, doi: 10.1038/s41467-020-16310-9.

[32] A. Zhang, L. Teng, and G. Alterovitz, "An explainable machine learning platform for pyrazinamide resistance prediction and genetic feature identification of *Mycobacterium tuberculosis*," *J. Amer. Med. Inf. Assoc.*, vol. 28, no. 3, pp. 533–540, Mar. 2021, doi: 10.1093/jamia/ocaa233.

[33] J. J. Davis, S. Boisvert, T. Brettin, R. W. Kenyon, C. Mao, R. Olson, R. Overbeek, J. Santerre, M. Shukla, A. R. Wattam, R. Will, F. Xia, and R. Stevens, "Antimicrobial resistance prediction in PATRIC and RAST," *Sci. Rep.*, vol. 6, no. 1, p. 27930, Jun. 2016, doi: 10.1038/srep27930.

[34] R. K. Aziz et al., "The RAST server: Rapid annotations using subsystems technology," *BMC Genomics*, vol. 9, no. 1, p. 75, Feb. 2008, doi: 10.1186/1471-2164-9-75.

[35] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[36] Q. Wu, A. Boueiz, A. Bozkurt, A. Masoomi, A. Wang, D. L. DeMeo, S. T. Weiss, and W. Qiu, "Deep learning methods for predicting disease status using genomic data," *J. Biometrics Biostatistics*, vol. 9, no. 5, p. 417, 2018.

[37] J. Cao, Z. Su, L. Yu, D. Chang, X. Li, and Z. Ma, "Softmax cross entropy loss with unbiased decision boundary for image classification," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2018, pp. 2028–2032, doi: 10.1109/CAC.2018.8623242.

[38] U. Ruby and V. Yendapalli, "Binary cross entropy with deep learning technique for image classification," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 5393–5397, Aug. 2020.

[39] 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.

[40] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Nov. 2017, *arXiv:1705.07874*. Accessed: Oct. 21, 2021.

[41] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?: Explaining the predictions of any classifier," Aug. 2016, *arXiv:1602.04938*. Accessed: Jun. 20, 2022.

[42] S. Chen, "Interpretation of multi-label classification models using Shapley values," Apr. 2021, *arXiv:2104.10505*. Accessed: Oct. 21, 2021.

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

[44] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," Jun. 2018, *arXiv:1806.08049*. Accessed: Jun. 20, 2022.

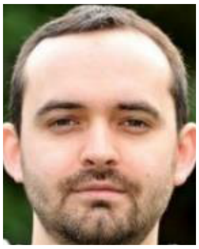
[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

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

[47] A. R. Wattam et al., "PATRIC, the bacterial bioinformatics database and analysis resource," *Nucleic Acids Res.*, vol. 42, pp. D581–D591, Jan. 2014, doi: 10.1093/nar/gkt1099.

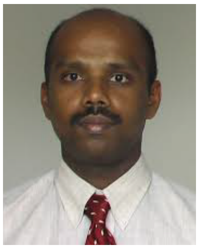
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

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. Before his Ph.D. research, he worked as a Lecturer (Probationary) at the University of Jaffna, Sri Lanka, and a Software Engineer at LSEG Technology. His current interest includes applying explainable artificial intelligence techniques to the healthcare domain.



BRIAN GARDNER received the M.Phys. degree (Hons.) in physics from the University of Exeter, U.K., in 2011, and the Ph.D. degree in computational neuroscience from the Department of Computer Science, University of Surrey, U.K., in 2016. He is currently a Research Fellow with the Department of Pathology and Infectious Diseases, School of Veterinary Medicine, University of Surrey. His research interests include modeling biological data using mechanistic and machine learning-based approaches.

ROBERTO LA RAGIONE received the B.Sc. degree (Hons.) in animal biology, in 1995, the master's degree in veterinary microbiology from the RVC (University of London), in 1996, and the Ph.D. degree from the Royal Holloway (University of London), in 2000. He is currently a Professor in veterinary microbiology and pathology with the School of Veterinary Medicine and the Head of the School of Biosciences and Medicine, University of Surrey. His current research interests include AMR and understanding the pathogenesis of food-borne pathogens with a particular interest in the development of control and intervention strategies.



ANIL FERNANDO (Senior Member, IEEE) received the B.Sc. degree (Hons.) in electronics and telecommunications engineering from the University of Moratuwa, Sri Lanka, in 1995, the M.Eng. degree (Hons.) in telecommunications from the Asian Institute of Technology, Thailand, in 1997, and the Ph.D. degree in video coding from the Department of Electrical and Electronic Engineering, University of Bristol, U.K., in 2001. He is currently a Professor with the Department of Computer Science, University of Strathclyde, Glasgow, U.K. His research interests include video processing/coding, artificial intelligence and machine learning, resource optimization, quality of experience, intelligent video encoding for wireless systems, and video communication in 5G/6G. He has published over 360 international journals and conference proceedings on these domains.