August 2022

Applications of Machine Learning in Medical Prognosis Using Electronic Medical Records

Farnaz HajiahmadiFoomani
University of Wisconsin-Milwaukee

APPLICATIONS OF MACHINE LEARNING IN MEDICAL PROGNOSIS USING ELECTRONIC MEDICAL RECORDS

by

Farnaz H. Foomani

A Dissertation Submitted in
Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
in Electrical Engineering

at

The University of Wisconsin-Milwaukee

August 2022

ABSTRACT

APPLICATIONS OF MACHINE LEARNING IN MEDICAL PROGNOSIS USING ELECTRONIC MEDICAL RECORDS

by

Farnaz H. Foomani

The University of Wisconsin-Milwaukee, 2022
Under the Supervision of Professor Zeyun Yu and Professor Sandeep Gopalakrishnan

Approximately 84 % of hospitals are adopting electronic medical records (EMR) In the United States. EMR is a vital resource to help clinicians diagnose the onset or predict the future condition of a specific disease. With machine learning advances, many research projects attempt to extract medically relevant and actionable data from massive EMR databases using machine learning algorithms. However, collecting patients' prognosis factors from Electronic EMR is challenging due to privacy, sensitivity, and confidentiality. In this study, we developed medical generative adversarial networks (GANs) to generate synthetic EMR prognosis factors using minimal information collected during routine care in specialized healthcare facilities. The generated prognosis variables used in developing predictive models for (1) chronic wound healing in patients diagnosed with Venous Leg Ulcers (VLUs) and (2) antibiotic resistance in patients diagnosed with Skin and soft tissue infections (SSTIs). Our proposed medical GANs, EMR-TCWGAN and DermaGAN, can produce both continuous and categorical features from EMR. We utilized conditional training strategies to enhance training and generate classified data

regarding healing vs. non-healing in EMR-TCWGAN and susceptibility vs. resistance in DermGAN. The ability of the proposed GAN models to generate realistic EMR data was evaluated by TSTR (test on the synthetic, train on the real), discriminative accuracy, and visualization. We analyzed the synthetic data augmentation technique's practicality in improving the wound healing prognostic model and antibiotic resistance classifier. We achieved the area under the curve (AUC) of 0.875 in the wound healing prognosis model and an average AUC of 0.830 in the antibiotic resistance classifier by using the synthetic samples generated by GANs in the training process. These results suggest that GANs can be considered a data augmentation method to generate realistic EMR data.

To my beloved husband,

Benny

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank the following people, without whom I would not have been able to complete this dissertation and without whom I would not have made it through my Ph.D. degree:

First, I would like to thank my supervisor, Professor Zeyun Yu, for letting me be a part of his incredible research team and for all his constant support, guidance, and patience during my program. Thank you for providing me with opportunities to work on some fascinating projects in the Big Data Analytics and Visualization Lab. You encouraged me to grab every opportunity and move ahead, and I will always be thankful for it.

Further, I would like to express my sincere gratitude to Professor Sandeep Gopalakrishnan for his continuous guidance, motivation, and energy. I am honored to have the opportunity to work with you during my program. This thesis would not have been possible without your insights and support.

I am grateful to all those with whom I have had the pleasure to work during these projects. I would like to extend a special thanks to Dr. Jeffrey Niezgoda, the AZH Wound and Vascular Centers, Dr. Shahzad Mizra, and Dr. Aayush Gupta for providing me with valuable datasets and supporting me during the projects.

I would like to offer my gratitude to my thesis committee members: Professor Hu, Professor Zhang, and Professor Dabagh, for their insightful and important comments.

And my biggest thanks go to my parents and my sister, Paniz. Your love and support made me strong and resilient to pursue my dreams.

Finally, I cannot begin to express my unfailing gratitude and love to my husband, Benyamin. You supported me throughout this process and have constantly encouraged me when the situations seemed arduous and insurmountable. I will now stop working until midnight and we can watch movies again!

Chapter 1

Introduction

1.1 AI and Machine Learning in Healthcare

Artificial intelligence (AI) and its well-known branch, Machine Learning (ML), are attracting researchers in different fields such as economics and finance [1], marketing [2], risk management [3], power systems [4], medicine, and health [5, 6] to the area of computer science. AI in medicine has two main disciplines: virtual and physical AI. The virtual branch incorporates informatics approaches using deep learning and electronic medical records to control health management systems and assist physicians in diagnosis and treatment recommendations. The physical area includes robotics that can serve the elderly and disabled patients or surgeons in the operation rooms [7]. In the past decade, ML techniques have been extensively used for disease diagnoses, such as kidney disease [8], skin cancer [9], breast cancer [10], heart disease [11], retinal layer segmentation, diagnosis of Alzheimer's disease [12], prostate cancer [13], and chronic wound healing prediction [14, 15]. Moreover, the Prediction of treatment efficiency using ML techniques has been discussed in oral cancer [16], epileptic seizure [17], neurodegenerative diseases [18], and depression [19].

Conventional ML algorithms such as logistic regression, support vector machines, decision trees, and random forests are highly dependent on feature representations, where predictive variables are extracted from medical data carefully before feeding to the model for training. This process is required an intensive human effort for feature engineering. Deep Learning (DL)

techniques address this problem, using an end-to-end learning architecture to take raw patient data as an input and map it to outcomes using many layers of nonlinear processing units. This method reduces human involvement in high-level feature engineering. On the other hand, humans still need to develop effective DL model architectures and fine-tune optimal model parameters. The field's continuous challenge is reducing the human intervention required to construct these architectures [20].

1.2 Wound Prognosis Models

More than 6 million people in the United States are suffering from various types of chronic wounds such as Venous Leg Ulcer (VLU), Arterial Ulcer (AU), Diabetic Foot Ulcer (DFU), and Pressure Ulcer (PU). About 0.15% to 0.3% of people are suffering from active VLU worldwide, and annually more than $25 billion is spent on wound management and Medicare cost [8]. Although there is no consensus about wound healing time, a wound is considered chronic if it has not healed in 4-12 weeks or has shown less than 20 % reduction in its area after a maximum of four weeks of treatment [21]. An accurate estimate of healing time could assist clinicians in making better decisions about therapies and interventions.

Patients with four weeks of prognostic information provided to a clinic are more likely to heal than patients without predictive records. Shanu K. Kurd et al. [22] suggested that the essential factor associated with a healed wound is a change in wound size after four weeks of care. Skene et al. [23] stated smaller initial ulcer area, shorter duration of ulceration, younger age, and no deep vein involvement as the most significant healing predictors of VLU. Franks et al. [24] reported ulcer size and duration, general mobility, and limb joint mobility as the critical predictors of leg ulceration healing. Vesna Karanikolic et al. [25] declared that factors associated

with delayed VLU healing are infection, number of ulcers, and larger ulcer surface area. Ankle-brachial pressure index and lipodermatosclerosis are essential positive factors for the healing of VLU. Factors that do not appear to have significant roles in healing are age, sex, obesity, condition of the surface, the deep venous system, and chronic diseases. Wound size and duration are the most critical factors in developing a predictive model for VLU. In addition to these two factors, ulcer grade and wound number are important characteristics for developing the model [26]. The most significant predictors for delayed healing of neuropathic diabetic foot ulcer (DFU), as reported by Margolis et al. [27], are patients' wound size, duration, and ethnicity. Significant predictors mentioned by the same authors for VLU include wound area, wound duration, ankle-brachial index, ethnicity, limb ulcer, history of stripping or venous ligation, inability to walk (1 block), wound margin, lipodermatosclerosis, fibrin-covered wound, and history of surgical wound debridement [28]. Khachemoune et al. [29] asserted that wound size and duration, lipodermatosclerosis, and history of failed prior split-thickness skin grafts are the main factors impacting healing chronic VLUs and is treated with Cryopreserved Epidermal Cultures (CEC). Ulcer duration and area are the most significant factors influencing the healing of VLUs; where patient sex, age, race, skin condition, and infection have no prognostic significance [30]. In summary, the most common significant wound healing predictors are wound size (length, surface area), depth, grade, duration, distance, color, and numbers; ankle-brachial index; ethnicity; lipodermatosclerosis; and previous wound treatment history. The most common non-significant wound healing predictors are the patient's age and sex, body mass index (obesity), the deep venous system, and infection.

In this thesis, we aim to predict if a wound heals within 12 weeks from receiving the first treatment in patients with venous leg ulcers. We develop a prognosis deep learning model based

on patients' demographic and clinical characteristics collected from their Electronic Medical records (EMRs).

1.3 Prediction of Antibiotic Resistance

Skin and soft-tissue infections (SSTIs) remain the most frequently observed infections in ambulatory and hospital settings and involve microbial invasion of the skin's layers and soft tissues [31]. The number of SSTI episodes with a culture-confirmed pathogen has increased dramatically in the United States in recent decades [32]. This infection constitutes around 30% of all infections in 2019 in India [33]. Depending on their strain characteristic, the bacteria causing SSTIs are classified as Gram-Negative Bacilli (GNB) or Gram-Positive Cocci (GPC). SSTIs range from mild infections such as pyoderma to severe life-threatening infections involving necrotizing fasciitis and extensive cellulitis. The distinction between severe SSTIs that need immediate intervention from mild infections is still challenging [31]. Patients' comorbidities such as diabetes mellitus and ischemia can advance a mild infection and result in treatment failure [34].

Most mild infections can be treated empirically by antibiotic prescriptions [35]. However, the evolution of antibiotic resistance is surging in SSTIs with the advent of new antibiotics. Mayo Clinic reports that more than 2 million infections from antibiotic-resistant bacteria occur annually in the United States, resulting in 35,000 deaths. Misuse of antibiotics promotes antibiotic resistance, leading to serious illnesses, longer recovery, longer hospital stays, and excessive medical expenses [36]. Hence, antimicrobial resistance surveillance is essential to estimate antibiotic resistance and monitor the results of medical interventions [37]. Conventional detection methods of bacterial resistance are standardized and widely used. In 2019,

Ramakrishna et al. conducted antimicrobial susceptibility testing using the Kirby Bauer disk diffusion method and E-strip method on 3570 samples suspected of SSTI in India to provide a predictable bacterial profile of the wound infections for clinicians [38]. However, the results may take more than 48hrs to be prepared, leading to overuse or misuse of antibiotics [39]. The Indian Council of Medical Research (ICMR) has been researching antimicrobial resistance through the Antimicrobial Resistance Research & Surveillance Network (AMRSN) annually. Their study aims to track resistance trends and understand resistance mechanisms in pathogens using molecular characterization techniques and whole-genome sequencing (WGS) [40]. In the United States, the Multidrug-Resistant Organism Repository and Surveillance Network (MRSN) is a unique entity that serves as the primary surveillance organization for antibiotic-resistant bacteria across the Army, Navy, and Air Force. MRSN focuses on the most common pathogens associated with antibiotic resistance, including *methicillin-resistant Staphylococcus aureus (MRSA), Klebsiella pneumonia, Acinetobacter baumannii, Pseudomonas aeruginosa, Enterobacter spp. and Escherichia coli [41]*. Moreover, National Antimicrobial Resistance Monitoring System (NARMS) tracks changes in the antimicrobial susceptibility of these common pathogens and provides information about emerging bacterial resistance [42].

The antibiotic susceptibility and sensitivity tests are time-consuming and costly. Hence, we developed ML models to predict antimicrobial resistance using antibiotic susceptibility testing (ABST) data (as ground truth labels) collected from patients diagnosed with SSTIs over one year. The networks learn from patients' clinical and demographic information such as age, gender, diagnoses, and bacterial pathogens involved with the skin infections collected from EMRs.

1.4 Electronic Medical Record (EMR)

In the United States, approximately 84 % of hospitals adopt electronic medical records (EMR). EMR is a vital resource to help clinicians diagnose the onset or predict the future condition of a specific disease [43]. With machine learning advances, many research projects attempt to extract medically relevant and actionable data from massive EMR databases using machine learning algorithms [44].

EMR of a patient includes structured information such as coded diagnoses, interventions, and treatments and unstructured data such as text documents from physicians and nurses that usually contain precious clinical data about the specific visit [45]. ML techniques can analyze structured data. By representing patient EMRs as longitudinal matrices with one dimension corresponding to the features and the other dimension corresponding to the time, ML models can be developed to analyze the medical data [20]. For instance, in [46], Sahni et al. have proposed a prognosis model based on the random forest to predict 1-year death risk using factors such as metabolic panel, demographic information, and ICD codes from the EMR data available at the end of hospitalization in multi-condition Patients. Yeh et al. have introduced Xception architecture, a CNN-based neural network to predict lung cancer within one year from diagnosis and medication codes obtained from the EMRs [47]. In [48], deep learning was deployed to predict knee osteoarthritis within a year using the previous three years of demographic characteristics and diagnosis codes from EMR.

Analyzing unstructured information mainly involves natural language processing (NLP). For example, in [49], Kaur et al. have developed NLP algorithms that automatically extract patients who meet Asthma Predictive Index (API) criteria from the EMR. Sung et al. have

introduced an NLP algorithm to help clinicians determine eligibility for intravenous thrombolysis in patients with stroke from clinical notes [50].

Despite the improvements that EMR has brought to the healthcare system, its adaptation among healthcare professionals is still controversial due to privacy and security concerns. Although EMRs were established in the 1970s, only 41% of U.S. hospitals had implemented a basic EMR system by 2005 due to privacy concerns. To resolve this tension, researchers proposed de-identification techniques such as k-anonymity, l-diversity, and t-closeness to generate anonymized data. However, there is still a risk of re-identification attacks in these procedures [51-53]. An alternative technique for overcoming these obstacles is to generate synthetic data that looks realistic. There is no direct mapping between real and synthetic data in these techniques; therefore, synthetic data, unlike de-identified data, is immune from re-identification cyber-attacks. Suppose synthetic data can have properties that are similar to real data. In that case, it can help researchers and companies access data by minimizing the privacy challenges of collecting EMR data [54].

Machine learning researchers have increasingly focused on developing generative models that can automatically extract underlying knowledge of data and synthesize new samples with characteristics similar to the original record. Generative adversarial networks (GANs) have recently demonstrated an impressive capacity to generate synthetic data with realistic features. This thesis has also overcome the challenges against EMR data collection by producing synthetic EMR using GANs. In the next section, we will talk about GANs in detail.

1.5 Generative Adversarial Networks in Medicine

GANs received much attention recently because of their ability to produce high-quality synthetic images. In GANs, two neural networks are deployed; the first network is a generator that trains to create realistic instances from the latent space, which can mislead the second network (discriminator) into identifying them as the original data. GANs have been successfully employed in producing high-quality synthetic images in an adversarial manner that may be indistinguishable from original images. Maayan et al. generated synthetic computed tomography (CT) images of liver lesions using the GAN model and showed that it improved liver lesion classification performance [55]. Christopher et al. introduced synthetic data to the training set of brain segmentation tasks using the GAN model and showed a 1-5% improvement in segmentation results [56]. Changhee et al. generated Magnetic brain Resonance (MR) images with the GAN model, which defeated an expert physician in the visual Turing Test [57]. Using GAN, Jyoti et al. generated Positron Emission Tomography (PET) images for three different Alzheimer's stages. They showed that these synthesized images are close to the real brain PET images through qualitative and quantitative evaluations [58]. Shin et al. generated synthetic MR images with brain tumors using the GAN network and showed that these images improved the segmentation performance and helped patients' record preservation [59]. Ren et al. generated synthetic gastric X-ray images using their proposed loss function-based conditional progressive, growing generative adversarial network (LC-PGGAN) [60].

1.6 Contributions and Goals

The contributions and goals of the wound prognosis model are as follows:

I. We demonstrate that the development of deep learning techniques can predict the healing process of venous leg ulcers with high accuracy based on patients' demographic and clinical characteristics. The purpose is to predict if a VLU heals within 12 weeks after receiving the first treatment.

II. We show that Generative Adversarial Networks are not limited to generating image datasets. Our EMR-CWGAN can successfully generate synthetic wound prognostic factors with characteristics similar to the original dataset.

The contributions and goals of the antibiotic resistance classifier are as follows:

I. We develop antibiotic resistance classifiers to classify the susceptibility or resistance of well-known bacteria to twelve different antibiotics in patients diagnosed with skin and soft tissue infections.

II. We investigate each pathogen's effect and predictive power in the Prediction of antibiotic resistance.

III. We develop a DermaGAN network to generate synthetic SSTI samples with characteristics similar to the original dataset.

IV. We deploy the generated samples as an augmented dataset to improve the antibiotic resistance classification accuracy.

The rest of this thesis is organized as follows: in chapter 2, we will discuss the machine learning algorithms used in this thesis and also the performance metrics used to evaluate their performance. Chapter 3 will explain the wound healing prognosis EMR data, the evaluation metrics, our proposed time-series medical GAN model, the deep prognosis model to predict wound healing of patients with VLU, and the experimental results. Chapter 4 will introduce the SSTI EMR data, the evaluation metrics, our proposed DermaGAN, the resistance classifier to

predict antibiotic resistance in patients diagnosed with SSTI, and the experimental results. Finally, this thesis is included in Chapter 5 and discusses the future directions.

Chapter 2

Conventional and Deep Machine Learning Algorithms

2.1 Introduction

Conventional models such as Random Forest, Logistic Regression, Support Vector Machines, etc., cannot simulate the complexity of decision-making in the human neuronal system [13]. Deep models (essentially a three-layer neural network, Convolutional Neural Networks, Long Short Term Memories, etc.) are inspired by the multi-level cognition of the human brain. Deep learning algorithms have proven to model the nonlinearity and complexity of human thinking. Deep neural network models have the potential for automatic feature extraction and can abstract high-level representations from low-level information [61].

This thesis performs both conventional and deep algorithms in medical prognosis applications. This chapter summarizes the generative adversarial networks, their application in generating electronic medical records, conventional ML models, deep learning algorithms, and performance metrics used in this study.

2.2 Machine Learning Algorithms

2.2.1 Feed Forward Neural Networks

Feed Forward Neural Networks (FNN), or in other words, Multi-Layer Perceptrons (MLP), is an artificial neural network model that maps input data to a set of suitable outputs through multiple nonlinear or linear functions. MLP is a supervised learning technique in which, during training, the weights are updated by the backpropagation algorithm and by propagating the errors

backward from the output layer to the input layer. The performance of MLP is highly dependent on hyperparameters: number of neurons, number of hidden layers, learning rate, and momentum [62]. MLP has been widely used in medical applications such as heart disease diagnosis [63], thyroid disease diagnosis [64], breast cancer classification [65], and chronic kidney disease prediction [66].

2.2.2 Convolutional Neural Networks

ConvNets or CNNs can process data in multiple arrays, such as speech, text, image, and video. ConvNets are constructed in various stages. The first stage is the convolutional layer, in which a set of weights called a filter bank is convolved with the input vector. This locally weighted sum is then passed through a nonlinearity such as a ReLU called activation function. Two or three stages of convolution, activation functions, and pooling layers are stacked, followed by fully-connected layers. Backpropagating gradients through a ConvNet is done to train all the weights in all the filter banks. This hierarchy structure will allow the higher-level features (details) to be obtained by composing a lower-level one. The convolutional and pooling layers in ConvNets are inspired by the cells in visual neuroscience [67]. There have been numerous applications of convolutional networks in medicine, such as medical image segmentation [68-70], wound image classification [71, 72], breast cancer classification and diagnosis [73-75], and medical image denoising and enhancement [76, 77].

2.2.3 Generative Adversarial Networks

Ian J. Goodfellow introduced GANs in 2014 [78]. The main idea is to simultaneously train two networks, generator "G" and discriminator "D". The generator learns the distribution of the data and outputs a sample that looks like real data. On the other hand, the discriminator, a binary classifier, needs to classify the sample as real or fake. In this minimax two-player game, G maps

an input noise, p_zTo data space, D learns to maximize the probability of accurate classification of real and fake samples: $\log D(x)$. G trains to minimize the difference between the discriminator output and real labels to produce more realistic samples: $\log(1 - D(G(z))$. Formally, the game between G and D is represented by the following equation in which p_{data} is data distribution, and $p_z(z)$ is a random Gaussian distribution:

$$\min_G \max_D V(D,G) = E_x \sim p_{data}(x)[\log[D(x)] + E_z \sim p_z(z)[\log(1 - D(x)] \quad (1)$$

Later, Arjovsky et al. [79] later introduced WGAN, in which the Jensen-Shannon (JS) divergence was replaced with Wasserstein divergence. WGAN can overcome the challenges of maintaining balance in training the generator and discriminator, dependency of the network's architecture, mode drop (failure in generating all the underlying distribution of the original data), and mode collapse (generation of the same output from different inputs). The Wasserstein GAN value function is as follows:

$$\min_G \max_D \mathrm{E}\,[D(x)]_{x \sim P_r} - \mathrm{E}\,[D(\tilde{x})]_{\tilde{x} \sim P_g} \quad (2)$$

Where D is the set of 1-Lipschitz functions, P_r is real data distribution and P_g is the distribution of fake samples defined by $\tilde{x} = G(z),\ z \sim P_z$. Under an optimal critic, minimizing the value function with respect to the generator parameters minimizes W(P_r, P_g). The WGAN value function creates a critic function with a better gradient with respect to its input than its GAN version. This feature will make generator optimization easier. Arjovsky et al. also proposed weight clipping of the critic to lie within a specific range [-c,c] [79].

In [80], Gulrajani et al. Introduced an alternative to clipping weights that penalizes the norm of the gradient of the critic with respect to its input. This method stabilizes the training of various GAN architectures with almost no hyperparameter tuning.

Recently, a few studies have been performed on medical records to produce synthetic structured and categorical data. MedGAN was introduced by Choi et al. [81] in 2017 to generate high-dimensional discrete variables by incorporating an autoencoder in generative adversarial networks. Diagnosis, medication, and procedure codes in EMR data were expressed as a vector, where the i^{th} array indicates the number of occurrences of the i^{th} variable in a patient (counts). Moreover, a binary vector representation of the EMR data was performed, in which the i^{th} dimension can be represented by 0 or 1, indicating the absence or presence of the i^{th} variable in a patient's record. In MedGAN, autoencoders learn from the count and binary discrete input vectors to map them to a lower-dimensional space and reconstruct the original input in the output by mapping them back to the original dimension. The generator needs to learn this low dimensional representation of the original data; therefore, the same decoder was used to reconstruct the original dimension after the generator. Later, in 2018, Baowaly et al. [54] integrated the idea of the Wasserstein GAN with gradient penalty and boundary-seeking GAN to generate more realistic synthetic patient records by introducing medWGAN and medBGAN. They performed the K–S similarity test and reported a 3% improvement in the generated data's quality using medBGAN compared to the medGAN. Zhang et al. [51] said EMR-WGAN and EMR-CWGAN in which the autoencoder was removed due to the model bias. Moreover, they introduced a conditional training strategy and incorporated the concept of labels as part of the generator and discriminator.

2.2.4 Random Forests

Random forest, developed by Leo Breiman, is an ensemble learner that generates multiple classifiers and aggregates their results. RF classifier is a set of decision trees created from a randomly selected training set. Each tree in RF will vote for its input, and then the output is determined by the majority voting of trees [82]. In RFs, each tree is grown using a subset of training samples, and some variables not used to grow the corresponding trees are known as out-of-bag (OOB) samples. One of the properties of OOBs is the estimation of variable importance, which quantifies the degree of contribution of a given variable in providing classification accuracy [83]. This is done by measuring the misclassification rate when the OOB examples for a variable, x_i, are randomly permuted and passed through the corresponding tree to vote for x_i. If the classification accuracy decreases significantly, it suggests a substantial contribution of variable x_i in the classification result. On the other hand, if it doesn't affect the predictive performance, then x_i is considered unimportant [82, 84].

2.2.5 Logistic Regression

Logistic regression is an efficient technique to predict outcomes and analyze the unique contribution of a group of independent variables to a binary outcome. Logistic regression Detects the effect of independent variables with the following equation [85]:

$$ln\frac{\hat{Y}}{1-\hat{Y}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i \qquad (3)$$

The equation transforms the linear regression equation to the natural log of the odds of being in one outcome category ($\hat{Y}$) over the other ($1 - \hat{Y}$). This equation calculates a linear combination of independent variables that can increase the likelihood of predicting the outcome

through iterative cycles. This process is known as the maximum likelihood estimation. The model structure ensures that logistic regression produces an accurate model [85].

2.2.6 Gradient Boosted Trees

Gradient boosted decision tree (GBDT) is an ML technique widely used due to its high accuracy and fast training in medicine. GBDT uses decision trees as a base learner. One decision tree is optimized at each iteration to minimize an aggregated loss function calculated from the previous decision trees [86]. If F(x) is the classification function that maps an input set of x to y, this function is optimized in a way to minimize a given loss function (L) as follows:

$$F^* = argmin \sum L\big(y, F(x)\big) \qquad (4)$$

Gradient boosting considers the estimated classification function as a sum of each function optimized by a decision tree, where T is the number of the decision trees [86]:

$$F(x) = \sum_{t=1}^{T} f_t(x) \qquad (5)$$

2.3 Performance Metrics

In this section, we provide a brief overview of metrics used in this study to investigate the performance of the classifiers.

2.3.1 Sensitivity and Specificity

Sensitivity and Specificity measure the validity of a diagnostic test for a binary outcome against a gold standard [87]. The sensitivity and specificity are dependent on the cut-off value above or below which the test is positive. The higher the sensitivity, the lower the specificity, and vice versa [88].

$$Sensitivity = \frac{TP}{TP + FN} \quad (6)$$

$$Specificty = \frac{TN}{TN + FP} \quad (7)$$

TP is the number of true positives, FN is the number of false negatives, TN is the number of true negatives, and FP is the number of false positives.

2.3.2 AUC-ROC

The Receiver Operating Characteristic (ROC) curve is a performance measurement for the classification problems that shows the trade-off between Sensitivity and 1-Specificity at various threshold settings. The Area Under the Curve (AUC) measures the ability of a classifier to distinguish between classes. AUC is mainly used for medical classification since a highly imbalanced dataset is usually involved in medical research. In some medical applications, Sensitivity is more critical than Specificity. The ROC curve helps find the optimum sensitivity value at a fixed specificity value [87].

Chapter 3

Wound Healing Prognosis Model

3.1 Problem Statement

This chapter introduces an EMR-Timeseries Conditional Wasserstein Generative Adversarial Network (EMR-TCWGAN) to generate synthetic wound prognosis factors. We then evaluate the ability of the suggested GAN to produce realistic data by training a wound healing prognosis model based on CNN. This model learns from the wound prognosis factors (both real and synthetic) collected from the first three visits. The goal is to predict if a patient with a venous leg ulcer heals within 12 weeks after receiving the first treatment.

3.2 Related Works

In 2020, Cho et al. [14] reported a wound healing predictive model based on logistic regression and classification tree models. They achieved an area under the curve (AUC) of 0.712 and 0.717, respectively, by training the models using the dataset collected from the first intake visit. Their dataset included AU (Arterial Ulcer), DFU (Diabetic Foot Ulcer), PU (Pressure Ulcer), and VLU (Venous Leg Ulcer), their relative wound measurements, and patient clinical and demographic characteristics. To avoid model complexity and overfitting, they added variables stepwise to evaluate their contribution to the model performance, considering the AUC as their metric. Jung et al. [89] developed their proposed model using logistic regression, random forest, and gradient boosted tree models and reported the AUC between 0.834 and 0.847. The training dataset consisted of patient age, sex, insurance, zip codes, wound information including 40 different wound types and 37 wound locations, and wound assessments such as

dimension, edema, erythema, and rubor. Cukjati et al. [15] performed a classification decision tree to predict the wound healing rate after one to six weeks of follow-up. The data included three categories:

1) Wound characteristics (length, width, depth, grade, date of appearance, date of treatment beginning, etiology, and location),

2) Patient characteristics (sex, age, number of wounds, diagnosis, date of spinal cord injury, and degree of spasticity),

3) Treatment/management (type of treatment, daily duration of therapy, course of treatment).

They reported 62% classification accuracy by training the models using two weeks of data. The classification accuracy increased to 80% when three weeks of data were available to the model. Margolis et al. [90] built their predictive model using logistic regressions based on different DFU prognosis variables to predict the wound status by the 20th week of care. Using variables such as age, sex, and the number, duration, size, and grade of wounds, they achieved a maximum AUC of 0.70.

As discussed above, although there are valuable studies in this field with outstanding results, all the wound prognosis models have been developed based on the traditional ML techniques. Therefore, in this thesis, we aim to predict the wound healing status of VLU patients based on deep learning models. We will compare deep models' results to traditional networks to assess the efficiency of deep learning in predicting wound healing.

3.3 Dataset and Data Processing

The data in this study is derived from the EMR of patients diagnosed with VLU in AZH Wound and Vascular Centers, Milwaukee, WI. The data has been carefully de-identified and includes patients' general information such as age, sex, ethnicity, their wound-related information such as wound length, width, area, location, and duration, as well as their clinical information, including the history of any vascular diseases, systolic and diastolic blood pressure, BMI, etc. The data contains both categorical and continuous values. Some features such as age, BMI, and systolic and diastolic blood pressures were discretized and converted to the categorical form. However, wound measurements remained continuous, including wound length, width, and area. Table 3-1 represents the prognosis factors included in this study, along with their detailed categories and statistics.

Each patient went through a weekly follow-up for up to 12 weeks, and their wounds were evaluated at the end of each visit and labeled as healed or not healed by the expert physician. Generally, a wound is considered cured if the ulcer has zero measurements [14]. However, the ground truth labels are the status of a wound by week 12 of the first visit. Our initial data includes the medical records of 70 patients. Patients with less than three weeks of wound assessments were excluded from the dataset. Those who stopped follow-ups before 12 weeks were considered not healed unless the termination was due to their healing in less than 12 weeks. Since not all the patients followed a weekly visit regularly, we applied this irregularity by defining a new parameter, separator, to indicate the time gaps between each visit. To handle the missing data, we used polynomial regressions in continuous and averaging in the categorical variables. The final dataset included the medical records of 60 patients (55% healed- 45% not healed), in which 75% of data (46 patients: 26 healed, 20 not healed) were randomly selected as

the training dataset, and the remaining data (14 patients: 7 healed, 7 not healed) were used as the test set.

Previous studies have reported that wound level factors such as wound dimensions and locations would substantially enhance predictive accuracy [14, 89]. Besides the wound measurements, clinical variables that show a weekly changing rate, such as wound fibrin and eschar percentage, have more prognostic values in time series analysis compared to those fixed between each visit, such as sex and ethnicity. In this study, the analysis of the relative importance of variables has been conducted based on the Random Forest (RF) classifier to identify the factors with more prognostic information. Random forest (RF) classifiers have been widely used in medicine and have shown outstanding performance as feature selector tools [83, 91, 92]. They provide good predictive performance, low overfitting, and easy interpretability. This interpretability makes it possible to study the interaction of variables that provide predictive accuracy [82]. This characteristic makes RFs more interesting to be used as a prognosis and diagnosis model in medicine and a feature selector tool. RF also can be applied to a mixture of continuous and categorical predictors. We explained the details of the RF classifier in section 2.1.4. Note that the RF model has been trained using the data collected after the first assessment for each patient. The second and third visit information was not used to conduct the relative importance of features analysis.

3.4 Models

3.4.1 Proposed EMR- Time-series Conditional Wasserstein GAN: EMR-TCWGAN

In the proposed EMR-TCWGAN, illustrated in Figure 3-1, we employed WGAN-GP introduced by Ishaan Gulrajani et al. to minimize the optimization difficulties that occasionally occur in weight clipping by penalizing the norm of the gradient of the critic with respect to its input [80]. We utilized the conditional training strategy, in which the GAN learns to generate prognosis labels, healed vs. not healed, along with the data. We can further employ the labeled synthetic data in training our prognosis network as an augmented dataset. We incorporated the prognosis labels into the generator and critic to design a conditional GAN.

The generator network, G, is a CNN, which takes the input noise and the desired label (healed:1 vs. not healed:0) and outputs a time series signal which is a T_x by n_x matrix in which T_x is the number of successive visits, and n_x is the number of prognosis variables. In our model, $T_x = 3$, and $n_x = 14$. This transformation is done through a dense layer with 128 neurons, two deconvolution layers with 64 and 128, 4 by four filters, LeakyReLU activation functions, batch normalization, and dropout, and four dense layers with the network structure of (128,128,128,42). The LeakyReLU activation functions, batch normalization, dropout layer, and a reshape layer are used to convert the data to a T_x by n_x dimension. The activation function for the last dense layer is tanh.

Table 3-1- Summary statistics of EMR dataset used in wound healing prognosis model.

Prognosis factor	Percentage of prognosis factors			Prognosis factor	Percentage of prognosis factors		
	1st	2nd visit	3rd visit		1st visit	2nd visit	3rd visit

	visit						
Age	8.69	8.69	8.69	**Systolic blood pressure**	6.52	10.87	8.69
<55	15.22	15.22	15.22	<120	13.04	2.17	8.69
56-64	32.60	32.60	30.43	121-129	26.08	17.39	21.73
65-74	43.48	43.48	45.65	130-139	41.30	58.69	54.35
>75				140-179	13.04	10.87	6.52
				>180			
Sex		54.34		**Diastolic blood pressure**	76.09	86.95	84.78
Female		45.65		<80	15.22	8.69	10.86
Male				81-89	8.69	4.35	4.34
				90-119	0.0	0.0	0.0
				>120			
Ethnicity		15.22		**Wound location**		67.39	
Black		84.78		Mid leg		19.56	
White		0.0		Distal		4.34	
others				Anterior leg		6.52	
				Lateral leg		2.17	
				others			
Smoking status		8.69		**Wound duration**	47.82	39.13	23.91
Yes		43.47		Less than four weeks	23.91	32.61	47.82
No		47.82		1-3 months	28.26	28.26	28.26
Reformed				Greater than three months			
BMI	0.0	0.0	0.0	**Prior wound infection**		10.87	
<18.5	4.35	4.35	6.52	Yes		82.61	
18.5-24	34.78	34.78	32.60	No		6.52	
25-29	60.87	60.87	60.87	Unknown			
>30							
History of DVT		13.04		**Prior ulcer grafting**		0.0	
Yes		84.78		Yes		97.83	
No		2.17		No		2.17	
Unknown				Unknown			
History of VD		8.69		**Percentage of wounds**	50.0	58.69	58.69
Yes		91.30		**covered with fibrin**	8.69	6.52	4.35
No				<25%	13.04	13.04	13.04
				25-50%	6.52	15.21	17.39
				50-75%	21.74	6.52	6.52
				75-100%			
				100%			
Diabetes Type		0.0		**Percentage of wound**	86.95	82.61	86.95
Insulin		54.35		**covered with eschar**	4.35	6.52	2.17
Oral medications or diet		45.65		<25%	2.17	4.35	4.34
				25-50%	0.0	0.0	0.0
No diabetes				50-75%	6.52	6.52	6.52
				75-100%			
				100%			
Drainage	4.35	2.17	21.74	**Doppler pulses**	56.52	54.35	54.35
None	36.96	60.87	43.48	**None**	21.74	23.91	23.91
Mild	56.52	34.78	32.61	Monophasic	17.40	17.39	17.39
Moderate	2.17	2.17	2.17	Biphasic	4.35	4.35	4.35
Heavy				Triphasic			
Edema		2.17		**Doppler evidence**	69.56	43.48	30.43
None		28.26		**insufficiency**	26.09	52.17	65.22
Mild		60.86		No	0.0	0.0	0.0
Moderate		8.69		GSV/SSV	97.83	4.35	4.35
Heavy				Deep			
				Both			

Treatment	82.61	84.78	82.61	**CEAP wound stage**	0.0	0.0	0.0
20-30/30-40 mmHg	80.43	80.43	73.91	No visual or palpable	0.0	0.0	0.0
Compression	13.04	19.56	21.74	signs of CVD	0.0	0.0	0.0
Stockings	0.0	0.0	0.0	Telangiectasia or reticular	0.0	0.0	0.0
3-4- Layer	0.02	0.06	0.0	veins	2.17	0.0	0.0
Compression	0.0	0.04	0.04	Varicose veins	0.0	0.0	19.56
Edema ware/Farrow				Edema	97.83	100	80.43
wrap/Spandagrips				Pigmentation: skin			
and Short Stretch				changes - hemosiderin			
Bandage				staining			
Pneumatic Pump				Healed ulcer			
Sharp/ultrasonic				Active ulcer			
Debridement							
Central Venous							
treatment/Peripheral							
Venous Ablation							
Sharp	8.69	13.04	8.69	**Endogenous**	10.87	15.22	17.39
debridement	91.31	86.95	91.31	**intervention**	89.13	84.78	82.61
Yes				Yes			
No				No			
DermaPace		0.0		**HBOT**	2.17	2.17	0
Yes		100		Yes	97.82	97.82	100
No				No			
Separator	NAN	91.30	86.95	**Lipodermatosclerosis**		39.13	
1week		2.17	10.86	Yes		60.87	
2weeks		6.52	2.17	No			
>3weeks							

The critic, C, is a CNN, receiving the real or generated data and their associated ground truth labels as inputs. The ground truth labels have the same dimension as the real and generated data. C outputs a score, representing whether the data is real or generated. The critic network comprises three layers of convolution and three dense layers. Each of the three convolution layers has 128 filters of size 3 by 3 with a LeakyReLU activation function followed by a dropout layer. The two dense layers have 256 and 128 units, followed by a layer normalization, a LeakyReLU activation function, and a dropout layer. The last dense layer has one unit with a linear activation function.

We have compared our proposed model to a baseline model, EMR- Conditional Wasserstein GAN: EMR-CWGAN, reported in [51]. The architecture of this network is shown in Figure 3-2. The generator and critic in EMR-TCWGAN are dense layers with a network structure of (128, 128, 128, 42) and (42, 256, 128,1), respectively. Each dense layer in G is followed by batch

normalization, and in C is followed by a layer normalization. The activation function is LeakyReLU, followed by a dropout layer. The activation function for the last dense layer in G is tanh followed by a reshape layer to convert the data to a T_x by n_x dimension. The activation function for the last dense layer in C is linear.

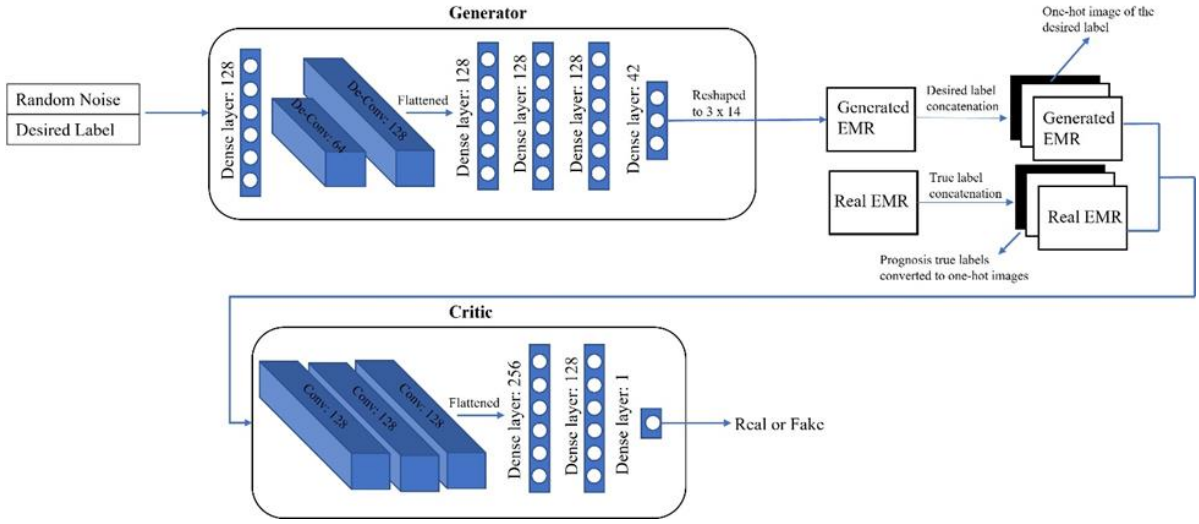


Figure 3-1- Architecture of our proposed GAN; EMR-TCWGAN

Since there is a limited number of data available, K-fold cross-validation with four folds was applied to estimate the performance of GANs. Each EMR-TCWGAN and EMR-CWGAN were trained using the four different training sets and evaluated by the remaining data as test sets. We summarized the results as the average performance of trained models.

3.4.2 Wound prognosis classifier

This classifier is a simple CNN with two 1D convolution layers followed by a dropout layer. There are 16 filters of size 3 x 3 for each layer. In the end, there are two fully connected layers with 5 and 1 units having a sigmoid activation function.

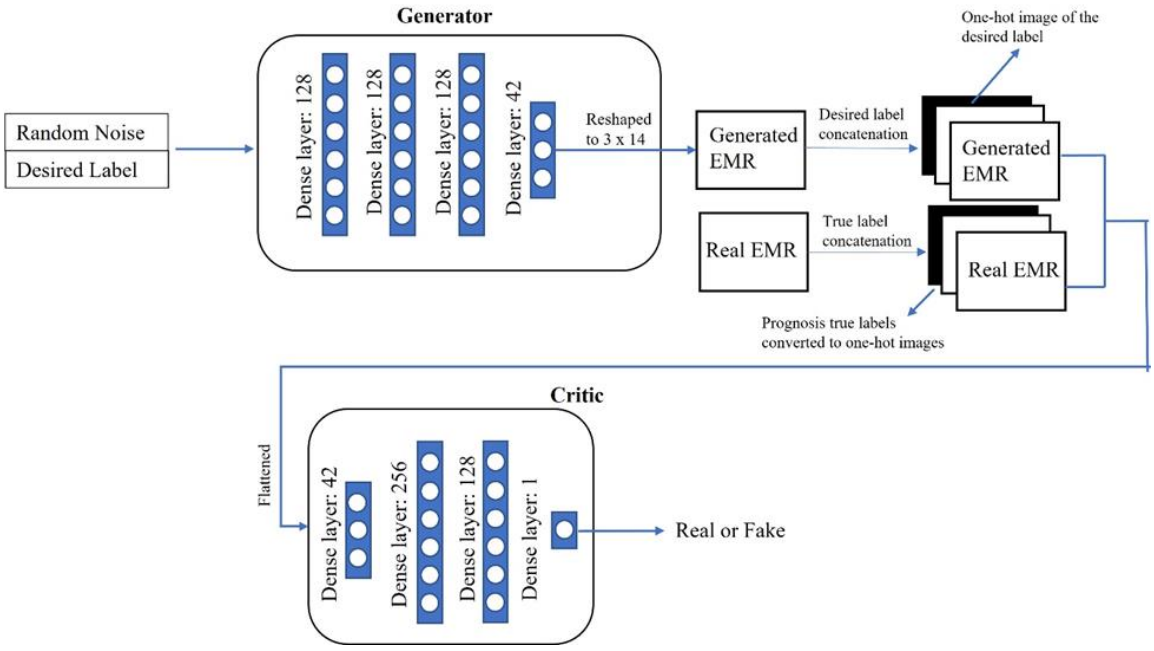


Figure 3-2- Baseline GAN model; EMR-CWGAN

3.5 Results

3.5.1 Evaluation of EMR-TCWGAN

We summarized the relative importance of variables based on the random forest regression model in Figure 3-3. Wound length had the highest variable importance in predicting wound healing, and the importance of the other variables is presented, respectively. All three wound measurements (wound length, width, and area) are considered the most critical variables for healing prediction, with scores of 1.00, 0.83, and 0.67, respectively. Doppler evidence and fibrin percentage are listed as the second and the third essential variables in Prediction. Surprisingly, the least important predictors were separator, which represents the irregular visits of a patient. After that, DermaPACE and CEAP wound is listed as the less important variables. Although having more prognosis factors could increase the model's predictive power, it may cause complexity and overfitting [14]. Therefore, to improve the predictive accuracy, we

disregarded variables with a relative importance of less than 0.3. Hence, we ended up with a total of 14 predictive features. The selected prognostic factors are wound length, width, area, Doppler evidence, percentage of fibrin, the number of Doppler pulses, systolic blood pressure, age, duration of the wound, history of DVT, wound location, drainage, diabetes type, and edema. We used these 14 predictive features to train EMR-TCWGAN and EMR-CWGAN.

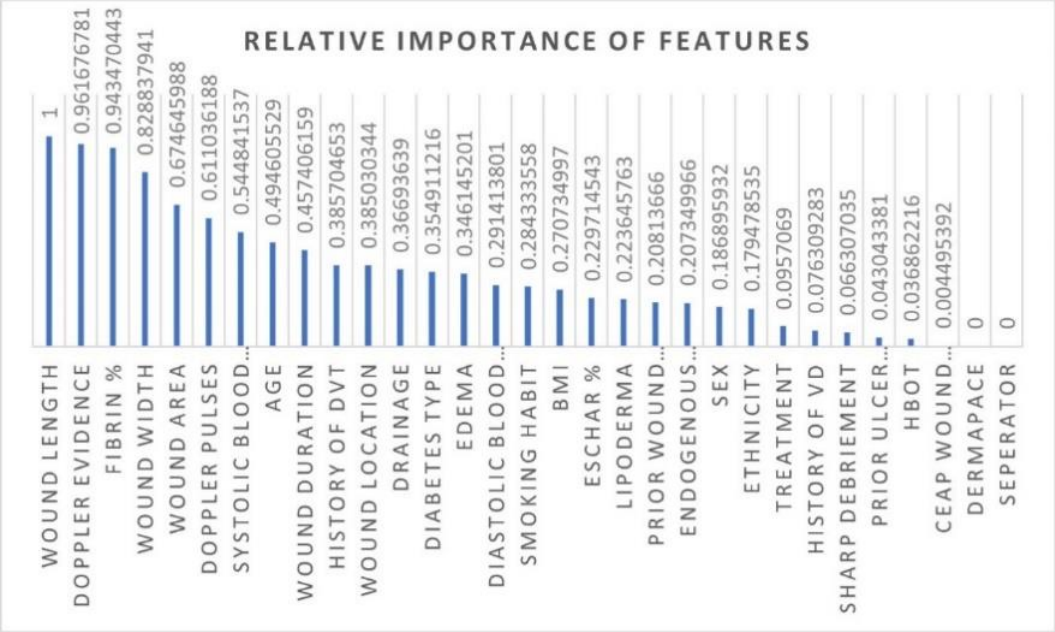


Figure 3-3- Relative importance of the VLU prognosis variables

We first examined the performance of our proposed GAN model (Electronic Medical Record- Time-series Conditional Wasserstein Generative Adversarial Network: EMR-TCWGAN) and compared it to the most recent and related method (Electronic Medical Record-Conditional Wasserstein Generative Adversarial Network: EMR-CWGAN [54]). To assess the quality of the generated data, we considered three criteria:

I. The distribution of the generated data should match the original data [93]

II. Samples should be as valuable as the original data in real-life applications [94].

III. The generated samples should not be distinguishable from the original data [93].

To evaluate the GAN models, we have considered four datasets;

1) The real training data,
2) The real test data,
3) The synthetic dataset generated by our proposed model, EMR-TCWGAN,
4) The synthetic dataset generated by the state-of-the-art model, EMR-CWGAN.

We started by visualizing the synthetic samples in two dimensions. We used Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP), a dimension reduction technique for visualization and nonlinear dimension reduction. UMAP is constructed from a theoretical framework based on Riemannian geometry and algebraic topology [95]. We applied U-map on both synthetic and original data (Test and Train dataset) for visualizations in two dimensions. Before using U-map, we flattened the temporal dimension. This evaluation can show the similarity of the original and synthetic data distribution. We also applied U-map on the test dataset to assess the ability of EMR-TCWGAN to generate synthetic data that can cover the distribution of unseen data points.

We applied a k-nearest neighbors classifier (KNN) with five neighbors on top of the 2D compressed features to classify healed samples vs. non-healed samples. This metric can quantitatively assess the performance of EMR-TCWGAN in generating realistic labels and compare it to the baseline model. We trained the KNN using the real training dataset and tested it by the real test dataset and synthetic datasets generated by EMR-TCWGAN and EMR-CWGAN. To evaluate the performance of GANs, we tested the KNN using 100 different randomly

generated synthetic datasets by GANs. Each dataset's length and ground truth label vary and are generated by random numbers. The classification accuracy is reported as the average accuracy of all 100 datasets.

Figure 3-4 illustrates the two-dimensional distribution of the synthetic (red dots), train (blue dots), and test (green dots) dataset. We performed dimension reduction using U-map and compared the synthetic labels' distribution to the ground truth labels in Figure 3-4b. In Figure 3-4c, healed and not healed original samples were mapped into two-dimensional space. Comparing the distribution of the synthetic instances generated by EMR-TCWGAN (first row) to the ones generated by the baseline model, EMR-CWGAN (second row), a comparable performance is observed in samples generated by our proposed model. Comparing the test dataset (green dots) vs. synthetic instances (red dots) in Figure 3-4a, we observe that both networks could cover unseen data distribution. To assess the performance of the two models quantitatively, we reported the average of the KNN classification model in Table 3-2 for both EMR-TCWGAN and the baseline model. The classification accuracy is slightly higher in samples generated by our proposed model. KNN classified healed vs. non-healed real samples with 66.66% accuracy; however, the classification accuracy has increased to 77.98% and 76.57% by involving EMR-TCWGAN and EMR-CWGAN synthetic samples in training, respectively. This observation suggests that GANs produce more uncomplicated and distinctive labeled samples than the actual data. Further analysis is required to investigate the performance of the generated labels. Therefore, in the next chapter, we conduct TSTR and TRTR evaluation methods to assess the applicability of the generated samples.

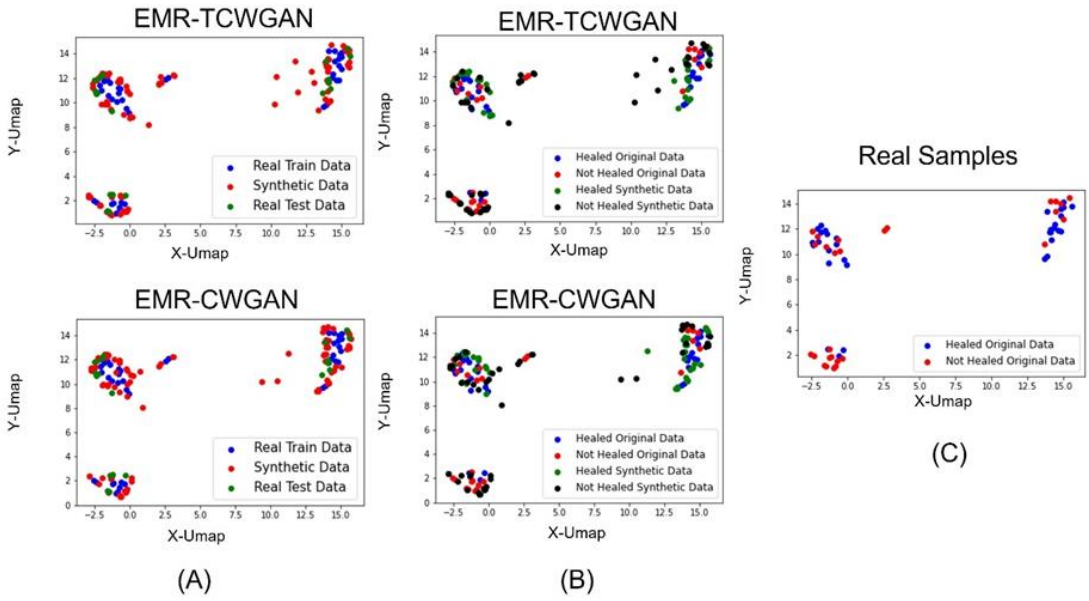


Figure 3-4- U-map visualization of time-series EMR data generated by the proposed EMR-TCWGAN (first row) and the baseline model EMR-CWGAN (second row). (a) Synthetic and real data distribution, red denotes synthetic, blue represents original train, and green denotes original test data mapped into two-dimensional space. (B) healed vs. not healed distribution in synthetic and real data. Blue indicates real healed data. Red denotes real not healed data. Green represents generated healed class, and black denotes generated not healed class. (C) Real train and test data mapped into two-dimensional space. Blue represents healed samples, and red indicates not healed samples.

Table 3-2-The results of the KNN classifier trained on a 2D dataset transformed by Umap. KNN was trained by the original training dataset and tested by 100 different randomly generated synthetic datasets from EMR-TCWGAN and EMR-CWGAN.

KNN Classifier	**Accuracy**
EMR-TCWGAN	77.98%
EMR-CWGAN	76.57%
Real Data	66.66%

We applied the Kolmogorov-Smirnov (K-S) test, a Goodness-of-fit statistic that tests if a sample comes from a population with a specific distribution [96]. Therefore, we employed the K-

S statistics to test the null hypothesis that synthetic and original samples come from populations with the same distribution. K-S statistics are calculated by finding the maximum absolute value between the two cumulative distribution functions. Comparing two datasets with cumulative distribution functions F(x) and G(x), the statistic is defined as [96]:

$$D_{KS} = \max|F(x) - P(x)| \qquad (6)$$

The null hypothesis is accepted if the p-value > 0.05 or otherwise rejected with a 95% confidence level. To compare the distribution of synthetic and real samples, we applied K-S tests followed by Mann–Whitney tests as post hoc comparisons to evaluate whether GAN models learned the distribution of the real samples. If GANs produce realistic examples, we expect the null hypothesis not to be rejected in favor of the alternative [97].

The distribution, $\hat{\chi}$ of the continuous features, including wound length, width, and area, and their temporal variations generated by EMR-TCWGAN were compared to the distribution, χ, of those in real samples using K-S tests. To compare EMR-TCWGAN to the baseline model (EMR-CWGAN), the distribution of the continuous features generated by EMR-CWGAN, $\hat{\chi}_b$ was also compared to the distribution of the continuous training samples, χ. Figure 3-5 compares the distribution of $\hat{\chi}$ and $\hat{\chi}_b$ to χ for three successive visits. From Figure 3-5, we can observe that the distribution of $\hat{\chi}$ is slightly closer to the distribution of χ in general, however, statistical analysis is required to compare the three distributions quantitively. Thus, we applied K-S statistics on each K-fold cross-validation GANs to measure the similarity of the distribution of $\hat{\chi}$ and $\hat{\chi}_b$ to χ. We reported the average results in Table 3-3. We reject the null hypothesis for the wound areas generated by EMR-CWGAN, suggesting that the generated wound areas could not follow the distribution of the real ones. We reported the p-values between K-S statistics from

EMR-TCWGAN and EMR-CWGAN calculated by Mann–Whitney tests in Table 3-4. Results suggest that our proposed GAN significantly outperformed the baseline model in generating wound area variable (p-value=0.001). However, the performance of the two models in generating synthetic wound length and width features are not significantly different.

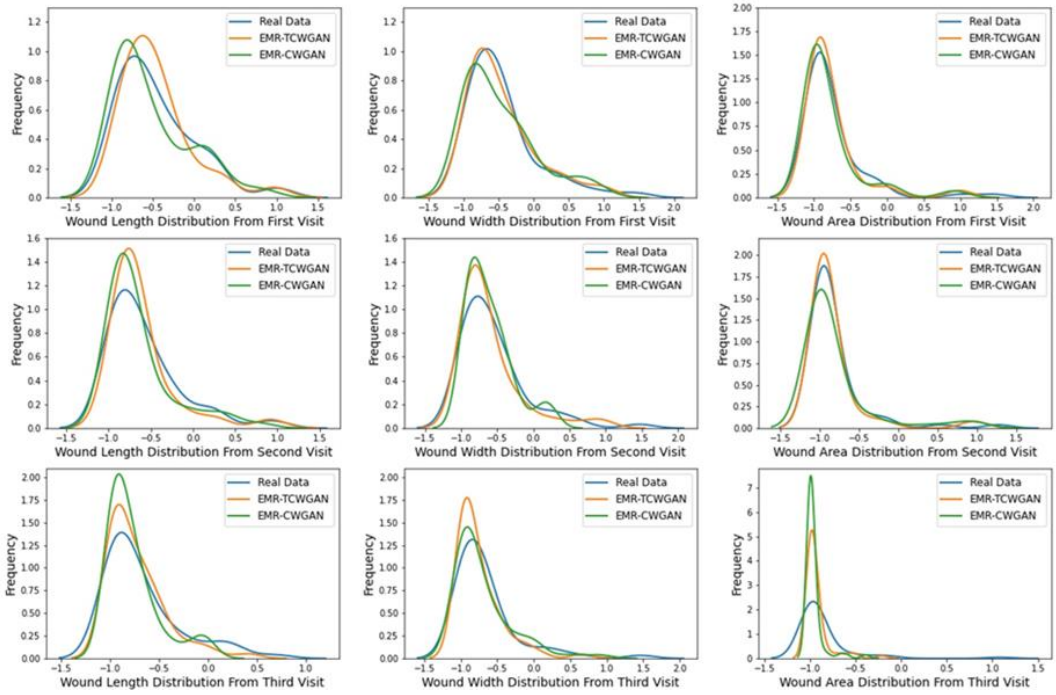


Figure 3-5- Probability density function of the continuous features (wound length, wound width, and wound area) for real samples, synthetic samples by EMR-TCEGAN, and synthetic samples by EMR-CWGAN in three successive visits. The three rows represent the results from the first, second, and third visits from top to bottom.

The generator should produce samples that are indistinguishable from the real data. Therefore, we trained a post-hoc classifier to classify real and fake samples. The CNN classifier with two convolutional layers was trained on an equal number of real and synthetic instances and tested on the synthetic data. The classifier must classify a given sample as real or fake. The classifier should achieve less than 50% accuracy for an excellent generator at this task [98].

Table 3-3- Kolmogorov-Smirnov statistical analysis to compare the probability distribution functions of the continuous prognosis factors in real and synthetic datasets generated by the proposed EMR-TCWGAN and the baseline model in three successive visits. Results represent the average of K fold cross-validation networks.

		EMR-TCWGAN			**EMR-CWGAN**		
Feature	**Visit**	P-value	Statistic	Accept/Reject	P-value	Statistic	Accept/Reject
Wound Length	First	0.847	0.108	Accept	0.639	0.137	Accept
	Second	0.291	0.183	Accept	0.356	0.174	Accept
	Third	0.209	0.199	Accept	0.283	0.203	Accept
Wound Width	First	0.539	0.166	Accept	0.686	0.129	Accept
	Second	0.312	0.191	Accept	0.331	0.195	Accept
	Third	0.170	0.212	Accept	0.192	0.208	Accept
Wound Area	First	0.630	0.137	Accept	0.151	0.254	Accept
	Second	0.313	0.179	Accept	0.049	0.304	Reject
	Third	0.301	0.182	Accept	0.131	0.254	Accept

Table 3-4- follow-up post hoc Mann–Whitney tests to compare Kolmogorov-Smirnov statistics in EMR-TCWGAN and EMR-CWGAN.

Prognostic Factor	**P-value**
Wound Length	0.977
Wound Width	0.670
Wound Area	0.001

We reported the discriminative accuracies in Table 3-5. The discriminative accuracy in samples generated by EMR-TCWGAN is relatively lower than those generated by EMR-CWGAN by 45.03%. As was mentioned before, the discriminative accuracy represents how well a classifier can distinguish between real and fake instances. We tested the discriminative classifier only on the synthetic data. Therefore, 23.97% accuracy means that 76.03% of the EMR-TCWGAN samples were realistic enough to be mistakenly classified as real by the

classifier. However, this number was reduced to only 31% in EMR-CWGAN samples. Our proposed EMR-TCWGAN produces more realistic synthetic data than the baseline model.

Table 3-5- Discriminative accuracy of the post-hoc classifier to classify real vs. fake on samples generated by EMR-TCWGAN and EMR-CWGAN.

Metrics	**EMR-TCWGAN**	**EMR-CWGAN**
Discriminative Accuracy (Lower the better)	23.97%	69.00%

3.5.2 Evaluation of Wound Healing Prognosis Model

We proposed a wound healing predictive model to predict if a VLU heals within 12 weeks from the first visit. We used the synthetic samples generated by GANs to train the predictive model and tested it on the original dataset. With this evaluation, we can assess the performance of the GANs in producing realistic instances that can be useful in real-life applications. We also trained the wound healing prognosis model on the original dataset and tested it on the original dataset. Comparing the two results, we can investigate the power of GANs as EMR augmentation tools. Moreover, we compared the result from CNN to the models widely employed for medical purposes, such as Random Forest, Logistic Regression, and Gradient Boosted Trees, to estimate the power of deep learning in medical prognosis.

The generated samples should be realistic and practical in real-life applications. Therefore, we have trained a wound healing prognosis model based on CNN with two different approaches to test this statement.

(1) We trained the first network using the synthetic dataset and tested it on the real test dataset (TSTR).

(2) We trained the second network using the real training dataset and tested it on the real test dataset (TRTR).

If the GANs are good enough to generate realistic samples, we expect the prediction performance in TSTR to be close to the accuracy of the TRTR technique.

We used the Area Under the ROC Curve (AUROC or AUC) metric to evaluate the performance of TSTR-CNN and TRTR-CNN. An advantage of AUC over accuracy is that AUC is not a function of threshold. It evaluates the classifier as the threshold varies over all possible values. Hence it can be used when highly imbalanced classes are involved [99]. Moreover, AUC is a helpful metric for comparing two diagnostic models since it enables comparing the entire ROC curve rather than at a particular point [100]. A random guess result is an AUC of 0.5, while a prefect model will achieve an AUC of 1.0. Generally, models with AUC above 0.7 are considered an acceptable predictive model fit [14]. To test if the EMR-TCWGAN is not copying the training samples, we reported the classification AUC on the real test dataset (unseen by GANs). Moreover, we trained a random forest, a logistic regression, and a gradient boosted tree model with TSTR and TRTR approaches to compare the performance of the CNN to the state-of-the-art prognosis models mentioned in chapter 2.

In the TSTR evaluation approach, we trained each network 30 times using randomly generated synthetic datasets by GANs and reported the average AUC with a 95% confidence interval. Since the number of the training data is limited, AUCs can validate the hypothesis that the generated instances by EMR-TCWGAN can be applicable enough to train a prognosis model and predict the healing status of new patients using the trained network.

The classification AUC of the proposed model is indicated in Table 3-6. T indicates the number of successive visits we used to train the prognosis CNN. Generally, the prognosis model trained by the EMR-TCWGAN samples shows a higher AUC than those trained by samples generated by the EMR-CWGAN. Using the synthetic samples generated by our proposed EMR-TCWGAN, we trained the prognosis model using the factors from the first three, the first two, and the first visit. The classification AUC decreased from 0.875 to 0.810 and then to 0.647 due to less temporal information available to the network. The classification AUCs were relatively lower when we trained the prognosis CNN model with synthetic samples generated by EMR-CWGAN. For EMR-CWGAN, the AUC decreased from 0.836 for T=3 to 0.751 for T=2 and then to 0.590 for T=1.

Table 3-6- The area under the curve (AUC) of the prognosis model (Prog-CNN) was trained using data generated by EMR-TCWGAN and EMR-CWGAN. T indicates the number of follow-up visits.

	EMR-TCWGAN	**EMR-CWGAN**
Visits/metrics	**AUC, 95% CI**	**AUC, 95% CI**
T=1	0.647, [0.531 – 0.720]	0.590, [0.520 – 0.651]
T=2	0.810, [0.719 – 0.872]	0.751, [0.653 – 0.817]
T=3	0.875, [0.822 – 0.912]	0.836, [0.797 – 0.878]

Table 3-7 compares the classification AUC of the prognosis CNN model to the other state-of-the-art models; random forest, logistic regression, and gradient boosted trees. We trained the networks by TSTR and TRTR methods. In the TSTR method, we trained each network ten times using ten different synthetic datasets and reported the average AUC and the confidence interval of all AUCs. In the TRTR method, we used k-fold cross-validation with k=10 and trained each network ten times using a different distribution of the real dataset. We also reported the average

AUC and the confidence interval of all AUCs in the TRTR metric method in Table 3-7. Generally, networks trained by synthetic EMR-TCWGAN datasets achieved a higher AUC than those trained by synthetic EMR-CWGAN. This observation suggests that our proposed GAN model could generate more efficient samples than the baseline.

Moreover, training the networks on the real data results in a lower AUC and wider confidence interval than EMR-TCWGAN. This wide confidence interval is due to the limited number of available real training datasets and their sparse distribution. The AUC's confidence intervals decreased by training the networks with the generated samples, and the average AUC increased significantly. This result suggests that EMR-TCWGAN can act as a data augmentation tool to create new datasets which can improve the ML models' prediction accuracy.

Table 3-7- The area under the curve (AUC) of the prognosis models (CNN, Random Forest, Logistic Regression, and Gradient Boosted Tree) trained with TSTR and TRTR approaches. GANs generated synthetic datasets used in the TSTR method for three follow-up visits. The average AUC with 95% confidence intervals is reported.

	EMR-TCWGAN	**EMR-CWGAN**	**TRTR**
Model	**AUC, 95% CI**	**AUC, 95% CI**	**AUC, 95% CI**
CNN	0.875, [0.822 – 0.912]	0.836, [0.797 – 0.878]	0.884
Random Forest	0.806, [0.732 – 0.869]	0.775, [0.681 – 0.841]	0.750
Logistic Regression	0.736, [0.701 – 0.778]	0.732, [0.691 – 0.784]	0.828
Gradient Boosted Tree	0.836, [0.754 – 0.913]	0.723, [0.586 – 0.822]	0.766

3.6 Discussion and Concolusion

Statistical analysis shows our proposed model has a relatively close performance to the baseline, if not slightly better. Using the K-nearest neighbor classification model, we could classify healed vs. not healed synthetic samples generated by our proposed model with 77.97%

accuracy. Comparing the continuous prognostic factors, such as wound length, width, and area, both models performed equally in generating the synthetic factors. However, our proposed model learned to generate the distribution of wound areas better than the baseline model.

We utilized the GANs as data augmentation techniques to increase the number of our training datasets. We developed a wound prognosis model based on deep learning to predict the healing of chronic VLUs within 12 weeks of the initial intake exam. We used the wound factors from the first three visits to train the prognosis model. Factors include wound length, width, area, Doppler evidence, percentage of fibrin, the number of Doppler pulses, systolic blood pressure, age, duration of the wound, history of DVT, wound location, drainage, diabetes type, and edema.

Training deep prognosis model with the real available dataset, we could achieve the AUC of 0.828 with a wide confidence interval of 0.523-0.963. However, training the network using the generated samples improved the AUC to 0.875, and the confidence interval was reduced to [0.822-0.912]. This result indicates that EMR-TCWGAN can help augment the EMR dataset, which solves the challenges against EMR data accessibility.

Comparing the deep CNN prognosis model to the random forest, logistic regression, and gradient boosted trees, our results show higher AUC in the deep model. Although deep learning proved its prediction power once again compared to the conventional method, the only disadvantage of deep learning is that the feature importance of each risk factor in predicting the healing status is not evident. Further analysis, such as neural network weight-based and Breiman's perturbation feature ranking algorithms, can rank the feature vectors on their relative importance to the model's accuracy [101].

Chapter 4

Prediction of Antibiotic Resistance

4.1 Problem Statement

Since the introduction of antibiotics, an alarming increase in the resistance of bacterial pathogens has been observed [5]. The prevalence of multidrug resistance (MDR) has increased significantly among many pathogens due to the overuse and misuse of antimicrobial agents [6]. Such a misuse increases the cost of medical care, exposes patients to potential adverse effects, and significantly risks the development and spread of antimicrobial resistance in healthcare facilities. Hence, antimicrobial resistance surveillance is essential for estimating the magnitude and trends of antibiotic resistance and monitoring the results of medical interventions [7]. Local surveillance data are critical and must be used to direct clinical management, formulate treatment guidelines, and guide infection control policies [8]. Although there is a relative dearth of local antibiotic susceptibility data in patients with SSTI in India, our data was collected from a health care center in Western Maharashtra. This project aims to assess the magnitude and clinical pattern of SSTIs, their causative microorganisms, and the antibiotic resistance patterns using machine learning algorithms.

4.2 Related Works

In this section, we reviewed the previous research focused on the prediction of antibiotic resistance using ML. Generally, we could find a few studies conducted on this topic. Studies can be categorized into two groups: 1. Those are focusing on predicting antibiotic resistance using genome sequences, 2. Those are developing the resistance classifiers using EMR data. Hicks et

al. used a random forest classifier to predict the resistance of *gonococci* genome sequence to Ciprofloxacin and Azithromycin. They reported a balanced accuracy of 76%-87% and 73%-83% for CIP and AZM, depending on their dataset [102]. They also developed classifiers to predict CIP non-susceptibility to *K. pneumoniae* and *A. baumannii* and achieved a significantly lower accuracy than the *gonococci* dataset. Kavvas et al. developed an SVM classifier to predict the resistance of the *Mycobacterium tuberculosis* genome to five different antibiotics listed in Table *4-1*. They achieved an average AUC of 0.8 depending on the antibiotic family. Kim et al. performed Gradient boosting trees to study the antibiotic resistance pathway of *Enterobacter cloacae*, *Escherichia coli*, *Klebsiella pneumonia*, and *Pseudomonas aeruginosa* whole genome sequences to Cefepime, Meropenem, and Ceftazidime with AUC in range of [0.7-1], [0.98-1] and [0.88-0.99], depending on the species, respectively [103].

We only found three studies incorporating EMR data in training the resistance classifiers, similar to our research. Feretzakis et al. used information from samples gram strains (GPC or GNB), site of infections (blood, tracheobronchial aspirates, urine, skin/wounds/soft tissue), and patient demographics in training various machine learning models listed in Table 4-1. They reported achieving a Maximum AUC of 0.726 and F-measure of 0.663 for the MLP algorithm [104]. Lewin-Epstein et al. dataset contain patient demographics data such as age and sex, and clinical data, including duration of hospitalization. By incorporating the three most common bacteria species information, they achieved the highest AUC in the range of [0.8-0.88], depending on the antibiotics. They reached the best performance using an ensemble model with a combination of Lasso logistic regression, neural networks, and gradient-boosted trees. They reported *Escherichia coli* as the most common bacterial species resistant to Ceftazidime, Gentamicin, Imipenem, Ofloxacin, Sulfamethoxazole-trimethoprim. *Klebsiella pneumonia* is the

second most common species resistant to Ceftazidime, Gentamicin, and Sulfamethoxazole-trimethoprim. *Pseudomonas aeruginosa* and *Staphylococcus coagulase-negative* group are the second most frequent resistant species to Imipenem and Ofloxacin, respectively [105]. Ayyıldız et al. studied the resistance of *Escherichia coli* to 15 antibiotics listed in Table 4-1. They analyzed biochemical parameters such as complete blood count, urinalysis, and C-Reactive protein with machine learning models without using an antibiogram. They reported a classification accuracy of [62%-98%] depending on the antibiotic family [106].

Table 4-1- summary of studies conducted on antibiotic resistance using machine learning algorithms.

Research	Dataset	Metrics	Antibiotics	performance	Methods
Hicks et al. [102]	*gonococci* datasets with whole-genome sequence data	Balanced accuracy	Ciprofloxacin, Azithromycin	76–87% 73–83%	random forest
Kavvas et al. [107]	*Mycobacterium tuberculosis* pan-genome	Average AUC	Isoniazid, Rifampicin, Ethambutol, Pyrazinamide, Streptomycin, Ofloxacin, 4-Aminosalicylic acid Ethionamide	0.8	support vector machine
Kim et al. [103]	*Enterobacter cloacae*, *Escherichia coli*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa* dataset with whole-genome sequencing data	AUC	Cefepime, Meropenem, Ceftazidime	0.7-1.0, 0.98-1.0, 0.88-0.99	gradient boosting tree
Feretzakis et al. [104]	Sample's Gram stain, Site of infection including blood, tracheobronchial aspirates/ bronchoalveolar lavage fluid, urine, skin/wounds/soft tissue, and patient demographics	The weighted average of F-measure and AUC	Amikacin, Aztreonam, Cefepime, Ceftazidime, ciprofloxacin, Colistin, Gentamicin, Imipenem, Meropenem, doripenem, piperacillin/tazobactam, Tobramycin, and Levofloxacin	Maximum AUC of 0.726 and F-measure of 0.663	Support Vector Machine, Sequential Minimal Optimization, k-Nearest Neighbors, Random Forest, Multilayer Perceptron
Lewin-Epstein et al. [105]	Age, Sex, Most common bacterial species,	AUC	Ceftazidime Gentamicin Imipenem Ofloxacin	0.8-0.88	Lasso logistic regression Neural networks Gradient boosted

	Second-most common bacterial species, Third-most common bacteria species, Latest hospitalization duration,		Sulfamethoxazole-trimethoprim		trees Ensemble learning
Ayyıldız et al. [106]	Complete blood count, Urinalysis, C-Reactive Protein	Accuracy	Amikacin, Ampicillin, Ceftazidime, Cefixime, Cefotaxime, Colistin, Ciprofloxacin, Cefepime, Ertapenem, Nitrofurantoin, Phosphomycin, Gentamicin, Levofloxacin, Piperacillin-Tazobactam, Trimethoprim-Sulfadiazine	96.0%, 77%, 62%, 63%, 68%, 95%, 76%, 70%, 96%, 90%, 98%, 84%, 98%, 92%, 79%	K-Nearest Neighbors, Artificial Neural Networks (ANN), Support Vector Machine, and Decision Tree Learning

4.3 Methodology

4.3.1 Dataset

We analyzed clinically relevant data of patients diagnosed with SSTIs collected from the Departments of Dermatology and Microbiology of a tertiary care center in Pune, India, for over one year. The dataset of 103 patients with GPC bacteria contains the variables of age in years, gender, MRSA screening test, inducible Clindamycin resistance, organism, and diagnoses. The class attribute, which is antibiotic non-susceptibility, has binary values for six antibiotics, including Gentamicin (GEN), Cotrimoxazole (COT), Cefoxitin (CEF), Erythromycin (ERY), Clindamycin (CLIN), and Ciprofloxacin (CIP). The dataset of 107 patients with GNB bacteria consists of age in years, gender, Extended-spectrum β-lactamases (ESBL) test, Carbapenem-resistant Enterobacteriaceae (CRE), organism, and diagnoses. The class labels that indicate antibiotic resistance are binary for six antibiotics, including Ceftazidime (CEFT), Ceftazidime-

Clavulanic Acid (CEFT+CLAV), Imipenem (IMP), Piperacillin-Tazobactam (PIP+TAZO), Ofloxacin (OFL), and Meropenem (MERO). Table 4-2 includes the summary statistics of the dataset and the distribution of each class (R: Resistance, S: Susceptible) in each antibiotic family for GPC and GNB bacteria. Data Processing includes scaling for age, one-hot encoding for the organism, and diagnosis and categorization for the other factors.

4.3.2 Models

DermaGAN

Generator. The inputs to the generator are a noise vector (z) with a normal distribution and a dimension of Z_{dim}=128 and label information. The label information is converted to a dense vector of size Z_{dim} using an embedding layer. Then the embedded label is multiplied by the noise vector. The resulting vector is then fed to a generator with the structure as follows:

Table 4-2- Summary statistics of the dataset and the distribution of resistance and susceptible class in each antibiotic family for GPC and GNB bacteria.

GPC (Gram Positive Cocci Bacteria)				Distribution of antibiotic resistance (GPC)
Feature	**Type**	**Feature**	**Type**	▪ Gentamycin (R: 38.55%, S: 61.45%) ▪ Cotrimoxazole (R: 33.66%, S: 66.34%) ▪ Cefoxitin (R: 49.47%, S: 50.53%) ▪ Erythromycin (R: 74.75%, S: 25.25%) ▪ Clindamycin (R: 54.45%, S: 45.55%) ▪ Ciprofloxacin (R: 87.37%, S: 12.63%)
Age (Years)	Mean: 44.34 Std: 15.74 Range: 95	**Organism**	**Categorical:** ▪ Staphylococcus Aureus (82.52%) ▪ Enterococcus SPP (1.94%) ▪ Streptococcus Pyogenes (5.8%) ▪ Staphylococcus, coagulase negative (9.7%)	
Sex	Male (65%) Female (35%)	**Diagnosis**	**Categorical:** ▪ Psoriasis (0.97%) ▪ Erythema (0.97%) ▪ Erythrasma (1.94%) ▪ Folliculitis (4.85%) ▪ Furuncle (1.94%) ▪ Hansen (15.53%) ▪ Infected Ulcer (0.97%) ▪ Impetigo (6.79%) ▪ Lichen (0.97%) ▪ Lupus (0.97%) ▪ Cellulitis (0.97%) ▪ Stasis ulcer (0.97%) ▪ Trophic ulcer (0.97%) ▪ Traumatic ulcer (0.97%) ▪ Mycetoma (1.94%) ▪ Pemphigus (6.79%) ▪ Pyoderma (5.82%) Gangrenosum ▪ Pyoderma (18.44%) ▪ Secondary infected eczema (4.85%) ▪ Sclerosis (0.97%) ▪ Toxic Necrolysis (0.97%) ▪ Abscess (3.88%) ▪ Burn (2.91%) ▪ Ecthyma (1.94%) ▪ Sebaceous cyst (1.94%) ▪ Vascular ulcer (0.97%) ▪ Eczema (8.73%)	
Methicillin-Resistant Staphylococcus aureus (MRSA screening test)	▪ Positive (38.83%) ▪ Negative (43.68%) ▪ Not applicable (17.46%)			
Inducible clindamycin resistance	▪ Positive (25.24%) ▪ Negative (74.76%)			
GNB (Gram Negative Bacilli Bacteria)				**Distribution of antibiotic resistance (GNB)**

Age (Years)	Mean: 44.13 Std: 14.94 Range: (11-89)	**Organism**	**Categorical:** ▪ Citrobacter (5.61%) ▪ Acinetobacter SPP (7.48%) ▪ Enterobacter SPP (1.87%) ▪ Escherichia Coli (6.54%)	▪ Klebsiella (28.04%) ▪ Proteus (4.67%) ▪ Pseudomonas SPP (9.34%) ▪ Pseudomonas Aeruginosa (34.58%) ▪ maltophilia (1.87%)	▪ Ceftazidime (R: 55.14%, S: 44.85%) ▪ Ceftazidime and Clavulanic Acid (R: 80.55%, S: 19.45%) ▪ Imipenem (R: 88.57%, S: 11.43%) ▪ Piperacillin and Tazobactam (R: 85.71%, S: 14.28%) ▪ Ofloxacin (R: 73.73%, S: 26.26%) ▪ Meropenem (R: 87.85%, S: 12.14%)
Sex	Male (54%) Female (46%)	**Diagnosis**	**Categorical:** ▪ Nodosum (0.93%) ▪ Hansen (35.51%) ▪ Mycetoma (0.93%) ▪ Pemphigus Vulgaris (5.61%) ▪ Gangrenosum (5.61%) ▪ Perianal ulcers (0.93%) ▪ Scrofuloderma (6.54%) ▪ Lupus Erythematosus (1.87%) ▪ Vascular ulcer (2.80%)	▪ Necrolysis (2.80%) ▪ Abscess (0.93%) ▪ Burn (1.87%) ▪ Carbuncle (0.93%) ▪ Cellulitis (0.93%) ▪ Diabetic ulcer (1.87%) ▪ Ecthyma (1.87%) ▪ Eczema (2.80%) ▪ Furuncle (0.93%) ▪ Stasis ulcer (11.21%) ▪ SCLEROSIS (0.93%) ▪ Ulcer (3.74%)	
ESBL test	▪ Positive (26.17%) ▪ Negative (73.83%)				
carbapenem-resistant Enterobacteriaceae (CRE)	▪ Positive (10.28%) ▪ Negative (89.71%)				

- Layer 1: A fully connected layer of 128*3*3 hidden units with a ReLU activation function (Rectified Linear Unit). This is followed by a reshape layer to reshape the information to the size of (3,3,128) and an up-sampling layer to convert it to the size of (6,6,128).
- Layer 2: A convolutional layer containing 128 filters with a size of 4. This is followed by a batch normalization layer and a ReLU activation function.
- Layer 3: A convolutional layer containing one filter size 4. This is followed by a Tanh activation function.

Critic. The inputs to the critic are the original and generated samples and their label information. The label information is converted into a dense vector of size 6*6*1=36 using an embedding layer. The input sample is also flattened and multiplied by the embedded labels. The resulting vector goes through a structure as follows:

- Layer 1: A fully connected layer of 128*3*3 hidden units with a Leaky ReLU activation function (α=0.2). This is followed by a reshape layer to reshape the information to the size of (3,3,128).
- Layer 2: A convolutional layer containing 16 filters with a size of 3 and strides of 2. This is followed by a Leaky ReLU activation function (α=0.2) and a dropout layer with a rate of 0.25.
- Layer 3: A convolutional layer containing 32 filters with a size of 3 and strides of 2. This is followed by a batch normalization layer, a Leaky ReLU activation function (α=0.2), and a dropout layer with a rate of 0.25. The result is then flattened.
- Layer 4: A fully connected linear layer of 1 hidden unit.

Training procedure. A DermaGAN is trained to synthesize SSTI data for susceptible and resistant classes per antibiotic family. Therefore, 12 DermaGANs are trained for CEFT, CEFT+CLAV, IMP, PIP+TAZO, OFL, MERO, GEN, COT, CEF, ERY, CLIN, and CIP antibiotics. Data preprocessing involved zero padding and resizing (6 * 6 * 1). The following hyperparameters are used for training DermaGAN: optimizer = RMSprop, batch_size = 8, learning_rate = 0.0002, number of epochs = 5000. The critic gets optimized using two loss functions: The Wasserstein loss and the Gradient penalty loss. The gradient penalty loss function is a soft version of the Lipschitz constraint used to avoid gradient vanishing/explosion. The generator's weights get optimized using the Wasserstein loss function. Figure 4-1 illustrates the general schematic of the DermaGAN.

Antibiotic Resistance Classifier

FNN. We trained post-hoc fully connected classifiers to predict the nonsusceptibility of different bacteria to antibiotics used to treat SSTIs. The classifiers are constructed of two

fully connected layers of 16 and 8 hidden units and ReLU activation functions. This is followed by batch normalization and a dropout layer with a rate of 0.2. the last layer is a fully-connected layer of 1 unit and a sigmoid activation function. For each antibiotic family, we trained three classifiers:

- To evaluate the performance of GANs in producing realistic samples.
- To assess the effect of synthetic data augmentation in improving nonsusceptibility classification.

CNN. Our CNN classifiers comprise a 1-D Convolution layer with 64 filters of size 3, a batch normalization layer, and a dropout layer with a rate of 0.2. This is followed by four fully-connected layers with 16, 8, and 1 units. The activation functions are ReLU except for the last layer, which is a sigmoid function.

4.4 Results

We Initially used a supervised ML approach to classify each sample as either susceptible or resistant to twelve antibiotics. We trained an FNN, a CNN, and an RF classifier for each antibiotic. We examined the success of each model in predicting antibiotic resistance in three feature combinations:

I. Networks were trained and tested only on the bacteria information listed in Table 4-2. Four GPC bacteria species, including *Staphylococcus Aureus, Enterococcus SPP, Streptococcus Pyogenes, Staphylococcus,* and *coagulase-negative,* were used to train the resistance classifier for GEN, COT, CEF, ERY, CLIN, and CIP. Nine GNB bacteria species, including *Citrobacter, Acinetobacter SPP, Enterobacter SPP, Escherichia Coli,*

Klebsiella, *Proteus*, *Pseudomonas SPP*, *Pseudomonas Aeruginosa*, and *maltophilia*, were used to train resistance classifiers for CEFT, CEFT+CLAV, IMP, PIP+TAZO, OFL, and MERO.

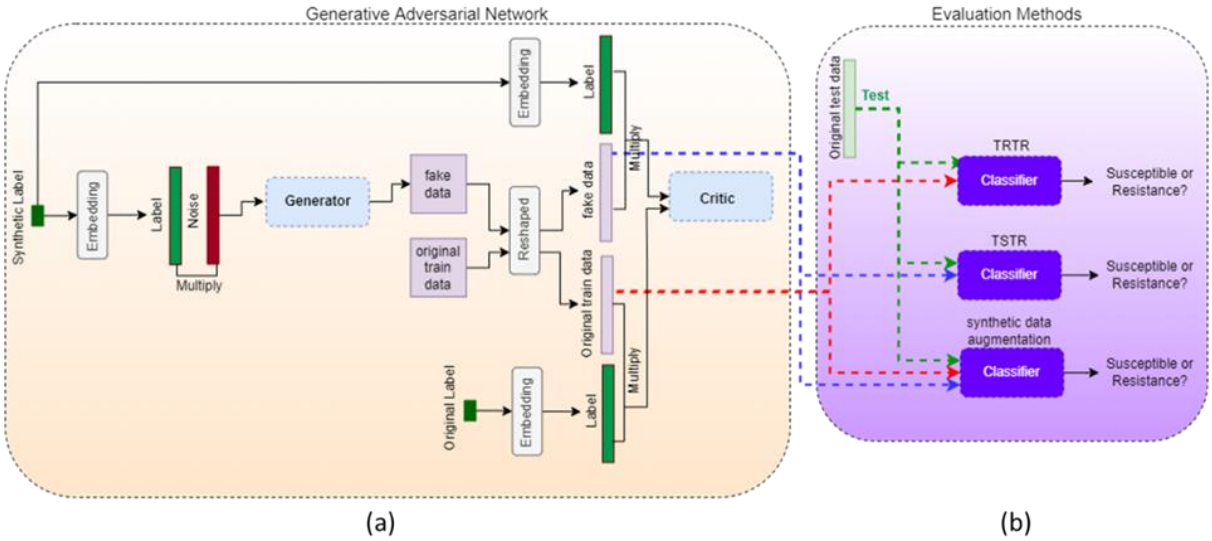


Figure 4-1- General schematic of (a) DermaGAN and (b) evaluation methods.

II. Networks were trained and tested on demographic and clinical data listed in Table 4-2. Bacterial species information was excluded for this dataset. Age, Sex, MRSA test, ICR test, and diagnoses were employed in training resistance classifiers for GEN, COT, CEF, ERY, CLIN, and CIP. Age, Sex, ESBL, CRE, and diagnoses were used in training resistance classifiers for CEFT, CEFT+CLAV, IMP, PIP+TAZO, OFL, and MERO.

III. The bacteria species information and patient demographic and clinical data were employed in training the classifiers.

We then analyzed the performance of the DermaGAN in producing realistic synthetic data. We synthesized data from patients diagnosed with SSTIs using DermaGAN. Initially, to identify the resistance of bacteria to antibiotics, we trained an FNN-based baseline resistant classifier on

the original training set per antibiotic and tested it on the original test dataset. This evaluation method is called TRTR, Train on Real, Test on Real. We also train a secondary resistant classifier on a generated dataset and test it on a similar original test dataset as the TRTR approach. This approach is called the TSTR evaluation method, Train on Synthetic, Test on Real [93, 108]. The evaluation methods are illustrated in Figure 4-1-b. Because of the limited data, for each evaluation, the performance of the models is estimated by 5-fold cross-validation. The last training fold is used as a validation set for the hyperparameter tuning. We reported the mean per fold classification AUC as a classification metric. We kept the training size similar in both approaches. The TSTR evaluation method assesses the assumption that the generated samples by DermaGAN are realistic to be employed in training a machine learning classifier with a real-life application. To accept this assumption, the classification AUC in the TSTR method should correlate with the TRTR AUC.

We also applied t-distributed stochastic neighbor embedding (t-SNE) analyses to both the original and synthetic datasets to reduce the dimension of feature space for visualization. This 2D visualization helps to understand how close the distribution of the generated samples is to the original samples.

Further, in this section, we analyze the effect of the synthetic data augmentation technique in improving antibiotic resistance prediction. We trained a resistant classifier per antibiotic on the augmented dataset and tested them on the original real test dataset, as depicted in Figure 4-1-b.

4.4.1 Performance of antibiotic resistance classifiers

The CNN-based and FNN-based classifiers achieved a similar classification success in terms of AUC. However, their performance outperforms the RF model in all three conformations of

features significantly. Generally, higher classification success was achieved when networks were trained on demographic, diagnoses, bacterial information, and clinical test results. Prediction of OFL, COT, and CIP resistance is dependent mainly on the identity of the bacterial species. Excluding this information from the dataset decreased the classification AUC in these families by 25%. This result is consistent with the correlation analysis depicted in Figure 4-3. *Enterococcus SPP*, *Coagulase-negative staphylococci*, and *Streptococcus pyogenes* are significantly correlated to the prediction of CIP resistance with positive correlation coefficients and zero P-values. *Enterococcus SPP* is associated substantially with predicting CIP susceptibility by a negative coefficient of 0.59 and P-value < 0.05. *Coagulase-negative staphylococci* and *Streptococcus*

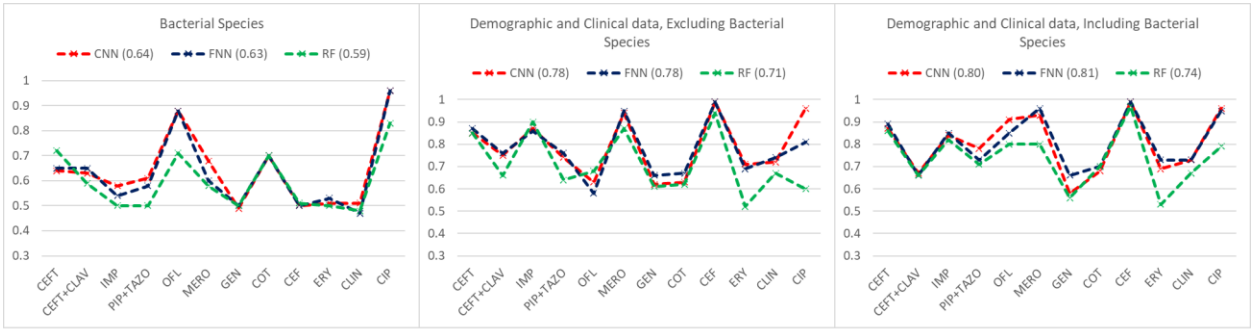


Figure 4-2- Classification AUC of the resistance classifiers trained on (a) bacterial species information, (b) basic demographic information, diagnoses, and clinical test results, (c) all the predictive variables.

pyogenes have shown to be significantly contributed to the prediction of COT susceptibility (coef<0), and Enterococcus SPP is highly correlated to the prognosis of COT resistance (coef>0). Acinetobacter SPP, Proteus mirabilis, and Pseudomonas SPP contributed to

the prediction of resistance class, and Stenotrophomonas maltophilia helped predict susceptibility class in OFL.

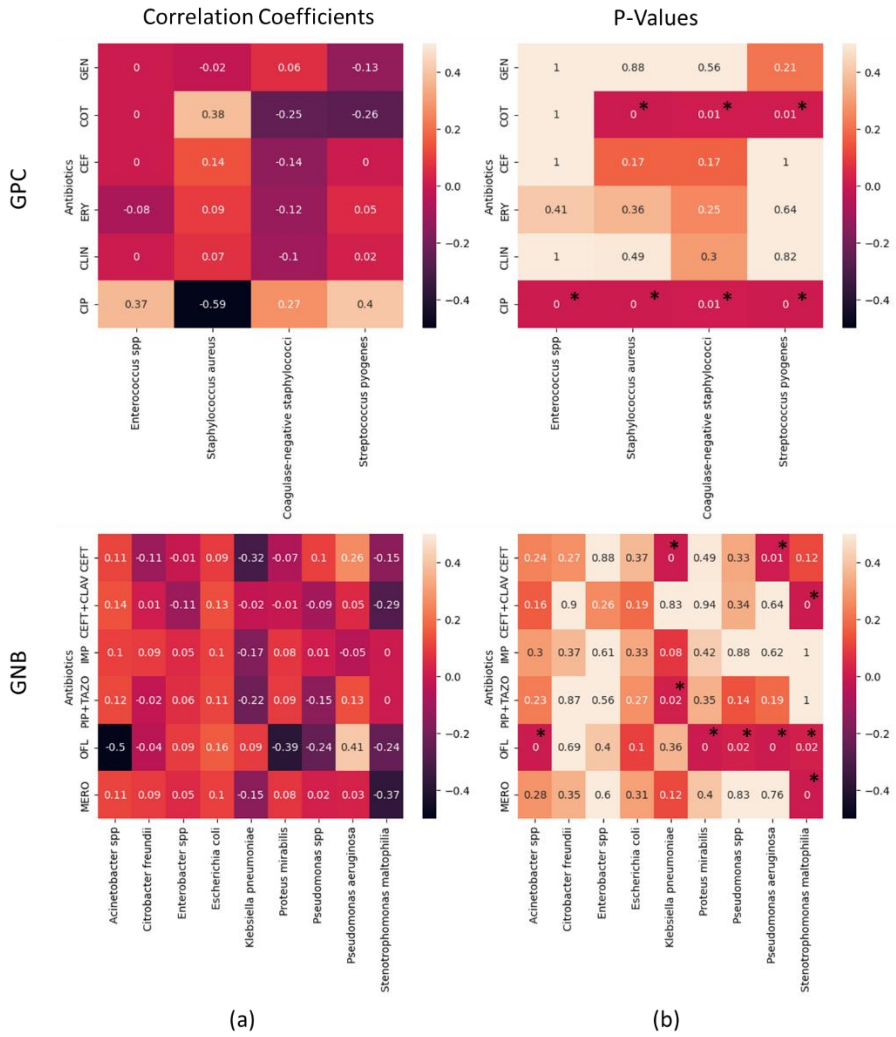


Figure 4-3- Correlation analysis of GNB and GPC bacteria with antibiotic resistance. (a) correlation coefficients. Bacteria with Positive coefficients are directly correlated with antibiotic resistance. Negative coefficients show a direct correlation with antibiotic susceptibility. (b) p-values. Stars represent a significant linear correlation.

Prediction of resistance in MERO and CEF has shown to be highly dependent on the patients' demographic and clinical data. We have performed variable importance analysis using Random Forest to investigate variables' predictive power in predicting each antibiotic resistance.

In contrast to neural networks, the contribution of each factor to an outcome can be obtained using an RF classifier. The result is shown in Figure 4-4. For the diagnosis, we reported the average coefficient for all the diagnoses listed in Table 4-2. This analysis revealed that the factors with the highest effect across all the families were Age, MRSA test, and CRE test. For MERO, age and CRE test, and for CEF, MRSA test and age are the two crucial predictive factors contributing to the outcome.

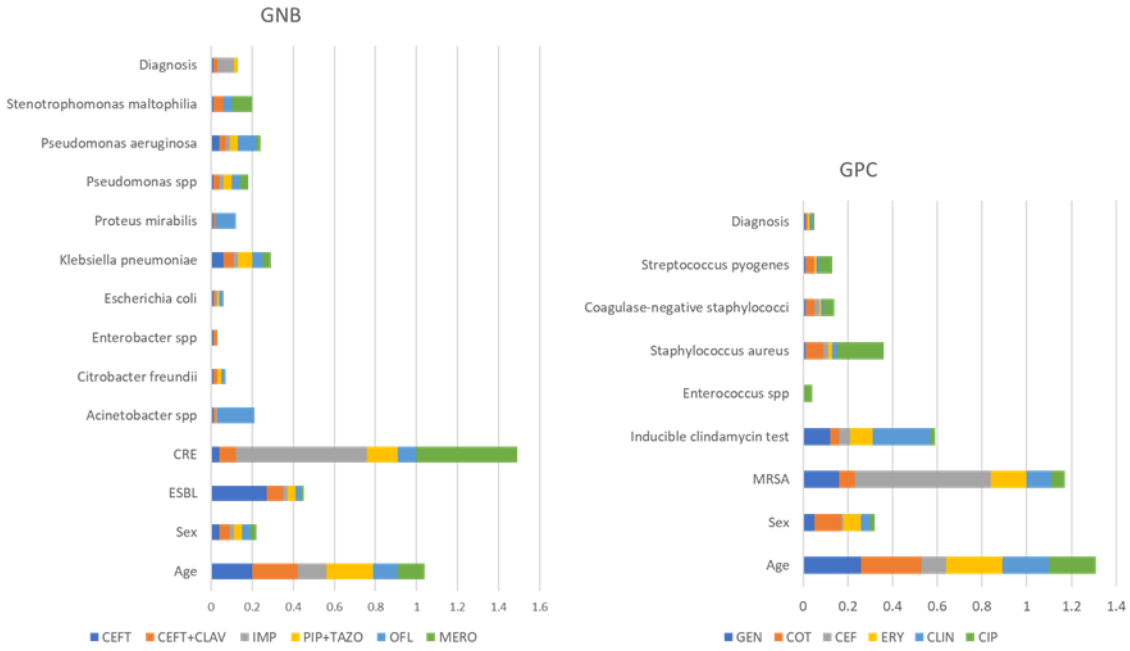


Figure 4-4- Variable importance analysis using Random Forest (RF).

4.4.2 Evaluation of DermaGAN

The summary result of classification AUC for TSTR and TRTR techniques is shown in Figure 4-5. Out of the twelve antibiotics under study, the classification AUC decreased by less than 5% in eight families while tied in one family in TSTR method compared to TRTR. The classification accuracy dropped in PIP+TAZO, MERO, and GEN by 11%, 24%, and 24%, respectively. A possible reason for this significant diminish in the AUC of TSTR is mode

collapsing, as reported in [108]. A possible solution to mode collapsing might be hyperparameter tuning, which needs to be considered in future research. Moreover, the AUC of the TSTR method is more than 0.6 in most families. A random guess results in an AUC of 0.5 and models above 0.7 are considered a good fit [14]. With this result, we can conclude that the generated samples have meaningful features closely related to most families' original data.

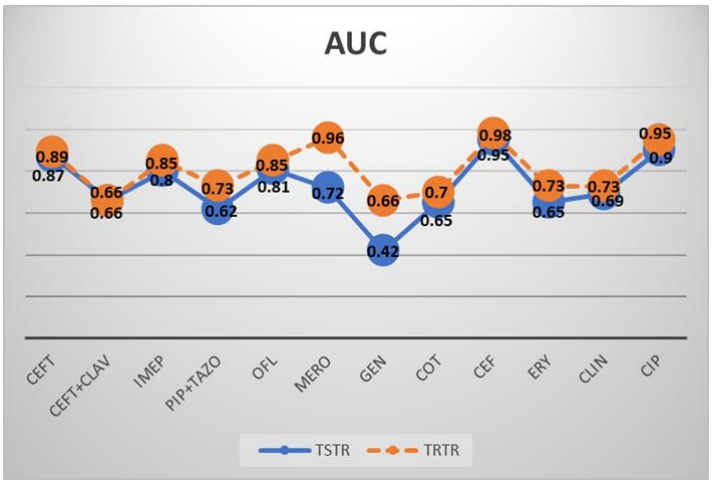


Figure 4-5- Performance of the classifier trained by only synthetic dataset (TSTR, solid lines) compared to the baseline (TRTR, dash lines) in twelve antibiotic families.

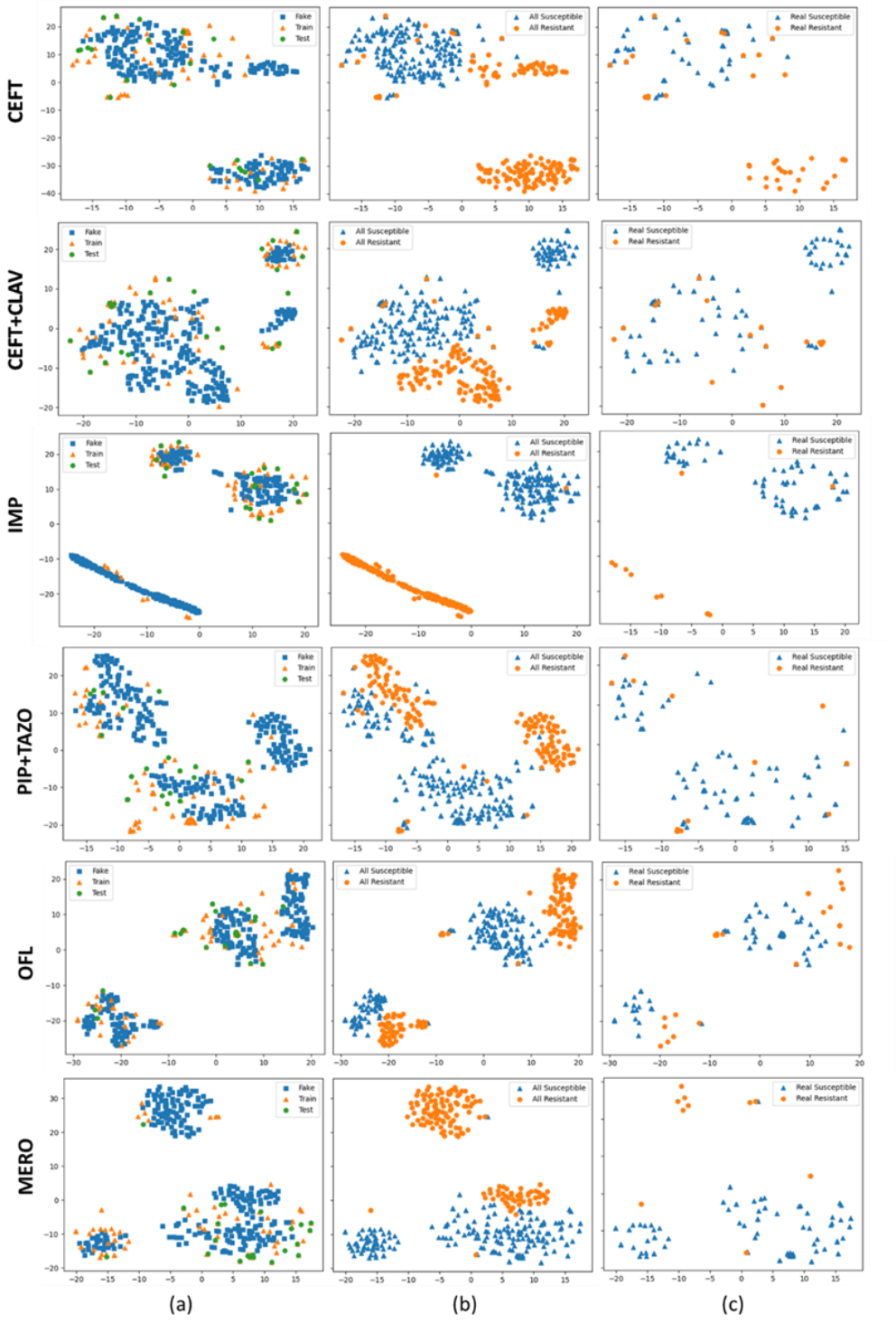
To go one step further, we applied t-SNE analyses to both the original and synthetic datasets to reduce the dimension of feature space for visualization. This 2D visualization helps to understand how close the distribution of the generated samples is to the original samples. Also, we visualized the distribution of the resistance and susceptible labels to assess the performance of the conditional GAN in producing ground-truth labels. Generally, with this evaluation, we can ensure that GANs are not suffering from mode drop and generate all the underlying distribution of the original data. Figure 4-6 demonstrates a 2D t-SNE visualization for twelve families under study. Figure 4-6-a shows the distribution of train, test, and fake samples. Figure 4-6-b depicts the susceptible and resistance distribution in the train, test, and fake instances.

Figure 4-6-c shows the distribution of the ground truth labels only for the original train and test dataset. Firstly, GANs could successfully generate all the possible distributions in the original dataset. Secondly, the distribution of the synthetic data is close to the distribution of the original data. Thirdly, synthetic ground-truth labels generated by the conditional GANs also successfully follow the distribution of labels in the original dataset; however, in some families, including GEN, COT, ERY, and CLIN, it is difficult to cluster the resistance and susceptible class in the latent space.

4.4.3 Performace of synthetic data augmentation

In this section, we analyzed the effect of the synthetic data augmentation technique in the improvement of antibiotic resistance prediction. We trained a resistant classifier per antibiotic on the combination of the real and synthetic dataset and tested it on the original real test dataset, as depicted in Figure 4-1-b. The mean per fold classification AUC is reported in Table 4-3. Generally, the classifier's performance depends on the proportion of the additional synthetic data. The amount of the augmented synthetic data is proportion to the number of the original train dataset. 0x represents no appending synthetic data, and Nx (0<N<=5) represents N times the number of the original training set. We achieved the mean AUC of 0.80 for the baseline model. In general, the mean AUC of the classifiers trained on the augmented dataset is slightly higher than the baseline except for the data size of 5x. Appending synthetic data with the amount of 3 and 4 times the original train dataset achieved the highest mean AUC of 0.82. The classification AUC enhanced up to 11% in data size of 3x depending on the antibiotic family. We reported the percentage improvement rate in the mean AUC of each antibiotic family in Table 4-3. We defined an improvement rate as the number of families with an enhancement in their classification AUC over the total number of families with an improvement or deterioration in

their classification AUC. With the increment of the synthetic data by 3x, the AUC was enhanced in 6 out of 12 (improvement rate = 60.00%) antibiotic families (CEFT+CLAV, IMP, PIP+TAZO, OFL, CLIN, CIP) while tied in two families (CEFT and CEF).



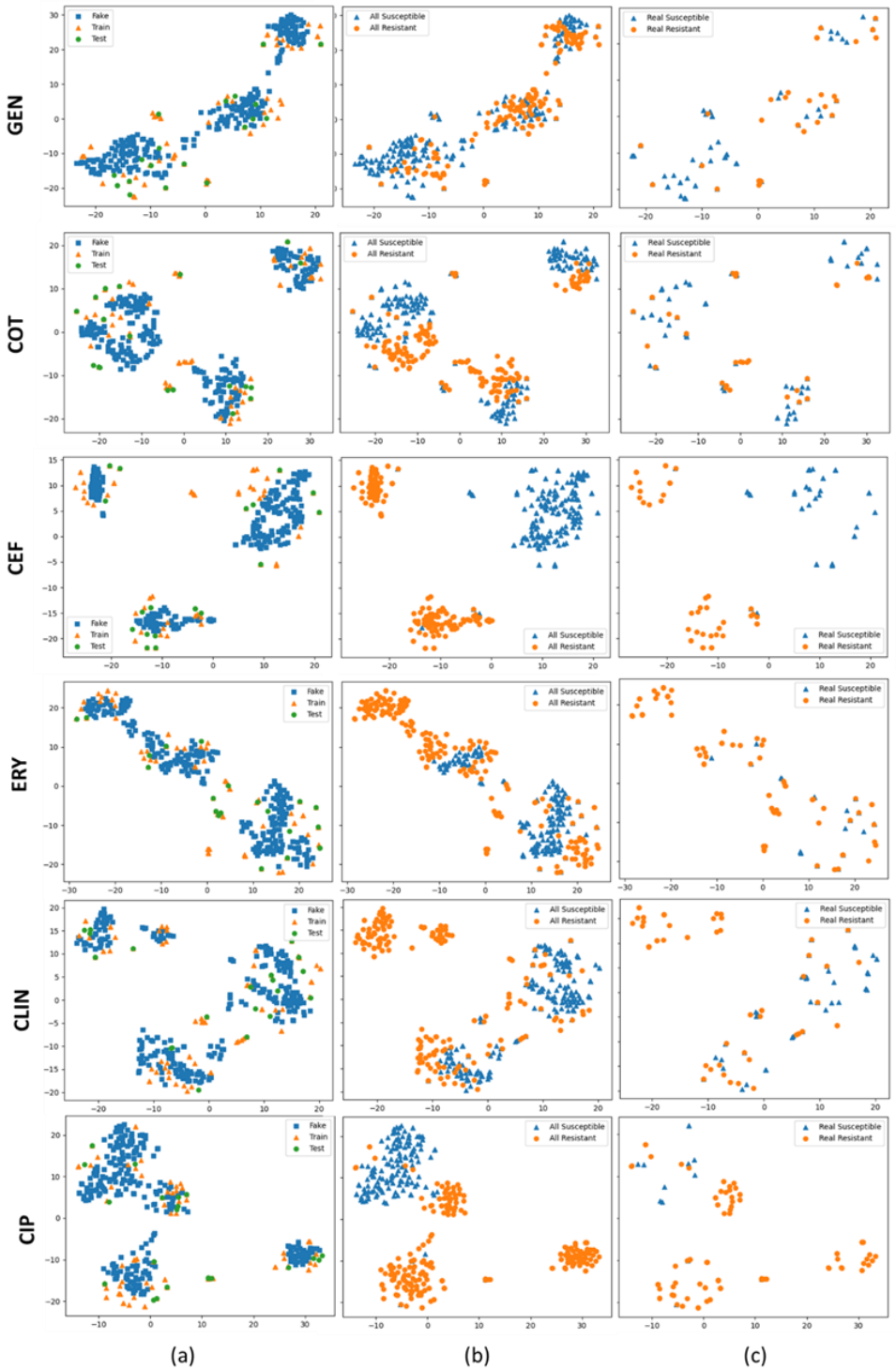


Figure 4-6- Two-dimensional visualizations of the real and generated dataset. (a) Data points with blue, orange and green colors represent the synthetic, train, and test data. (b) Data points with blue, and orange represent susceptible and resistant samples in fake, train, and test data. (c) Blue and orange data points represent the original susceptible and resistant labels.

This result confirms that synthetic data augmentation can improve the model's performance depending on the antibiotic family.

Table 4-3- Performance of the classifiers trained on the combination of the synthetic and original dataset compared to the baseline model (0x). Nx (N times the size of the original dataset, N=1:5) indicates the amount of the appendant synthetic dataset to the original training set. The AUCs are the mean of classification AUC per fold. The percentage of improvement compared to the baseline, the average AUC of all antibiotic families, and the rate of improvement are reported.

Bacteria Type	Appendant rate	0x	1x		2x		3x		4x		5x	
	Antibiotics|Metrics	AUC	AUC	%Improvement	AUC	%Improvement	AUC	%Improvement	AUC	%Improvement	AUC	%Improvement
GNB	CEFT	0.89	0.87	-2%	0.87	-2%	**0.89**	**0%**	0.87	-2%	0.90	+1%
	CEFT+CLAV	0.66	0.70	+4%	0.65	-1%	**0.70**	**+4%**	0.69	+3%	0.69	+3%
	IMP	0.85	0.93	+8%	0.91	+6%	**0.96**	**+11%**	0.93	+8%	0.89	+4%
	PIP+TAZO	0.73	0.79	+6%	0.79	+6%	**0.84**	**+11%**	0.85	+12%	0.81	+8%
	OFL	0.85	0.95	+10%	0.93	+8%	**0.92**	**+7%**	0.94	+9%	0.96	+11%
	MERO	0.96	0.93	-3%	0.90	-6%	0.90	-6%	0.92	-4%	0.91	-5%
GPC	GEN	0.66	0.58	-8%	0.55	-11%	0.60	-6%	0.59	-7%	0.52	-14%
	COT	0.70	0.66	-4%	0.74	+4%	0.66	-4%	0.70	0%	0.64	-6%
	CEF	0.99	0.99	0%	0.99	0%	**0.99**	**0%**	0.99	0%	0.99	0%
	ERY	0.73	0.68	-5%	0.73	0%	0.67	-6%	0.71	-2%	0.69	-4%
	CLIN	0.73	0.73	0%	0.75	+2%	**0.76**	**+3%**	0.74	+1%	0.75	+2%
	CIP	0.95	0.95	0%	0.94	-1%	**0.98**	**+3%**	0.94	-1%	0.95	0%
Average of AUC | improvement rate		0.80	0.81	44.44%	0.81	50.00%	**0.82**	**60.00%**	0.82	50.00%	0.80	60.00%

We reported TPR, FNR, FPR, and TNR in the form of a confusion matrix for each antibiotic in Figure 4-7. TPR or sensitivity shows the ability of the classifiers to classify resistant samples correctly. TNR or specificity, on the other hand, is the ability of the classifiers to identify susceptible samples accurately. In antibiotic resistance studies, sensitivity is a more serious concern than specificity. In other words, misclassifying resistant samples as susceptible can put a patient's life in danger. Figure 4-7-a and Figure 4-7-b include metrics for the baseline classifier (0x), and the augmented classifier (3x), respectively. Generally, synthetic data augmentation enhanced the sensitivity (TPR) by 5%-19%, especially in cases where the baseline model achieved a low sensitivity and high specificity, such as in CEFT+CLAV, PIP+TAZO,

OFL, and COT. This result suggests that the data augmentation compensates for the dataset's imbalanced problem.

Sensitivity decreased in GEN, ERY, CLIN, and CIP between 1%-6%; however, this looks like a compromise between sensitivity and specificity. A possible reason for this significant decrease in sensitivity in GEN and ERY is the difficulty in clustering the resistant and susceptible classes in the latent space, as illustrated in Figure 4-6. As observed from this 2-D visualization, the boundary between resistance and susceptible samples in these antibiotics is unclear. A possible solution to improve the accuracy is feature selection, which must be considered in future studies. The worst result is for MERO, in which sensitivity and specificity decreased.

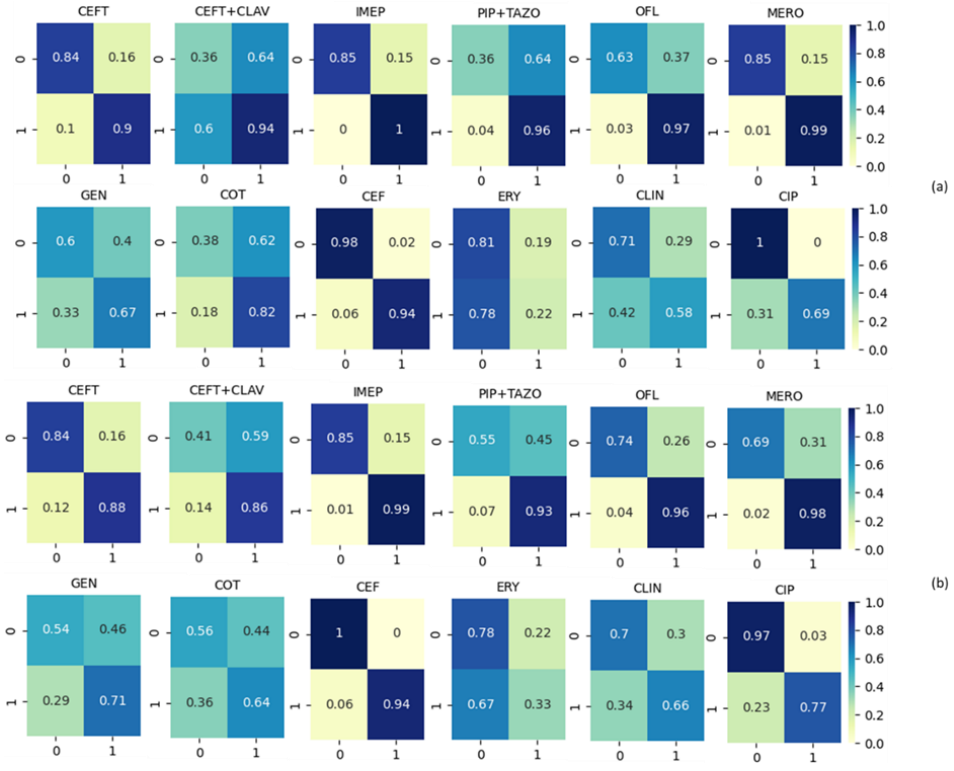


Figure 4-7- TPR, FNR, FPR, and TNR of the classifiers trained by (a) original training set and (b) original + synthetic training set (3x) for 12 antibiotic families.

4.4.4 Further improvement in antibiotic resistance classifiers

GANs are not perfect in generating synthetic samples. The predictive variables used in this study are categorical integer variables with values of -1, 0, and 1. However, the synthetic factors generated by GANs are continuous float variables in a range of [-1,1]. To compensate for these differences, we applied a post-processing cleaning process to round the synthetic float variables to the nearest integer values in categorical factors. For "age," which is a continuous float variable, we removed the synthetic data with a negative value and rounded the positive values to two decimal places. The cleaning process of each variable is shown in Figure 4-8. The 2D visualization by t-SNE is shown in Figure 4-9 for twelve families. Figure 4-9-a shows the distribution of synthetic and original samples without a cleaning process. Figure 4-9-b illustrates the distribution of the generated and original samples after the cleaning process. We obtained fewer instances with a more concentrated distribution after cleaning. With this process, we can preserve the samples with a closer distribution similarity to the original dataset and remove the duplicated samples.

The generator should produce samples that are indistinguishable from the original dataset. Therefore, we trained a post-hoc MLP-based classifier to classify between the real and fake examples. We trained the discriminative classifier on an equal number of original and synthetic instances and tested it on the synthetic data. The classifier must classify a given sample as real or fake. The classifier should achieve 50% (or less) accuracy for an excellent generator at this task. The discriminator accuracy for samples generated by GANs is shown in Figure 4-10. We achieved the average discriminator accuracy of 76.05% before and 72.99% after cleaning. The result suggests that the generators could not produce completely indistinguishable synthetic

samples from the original real data. The discriminative accuracy decreased by approximately 3% by applying the cleaning process, meaning that cleaning helped create a more realistic dataset.

We applied a further post-processing analysis to select the discriminative classifier's synthetic samples chosen as realistic. The discriminative classifier tested each synthetic sample produced by the generator. If the sample is classified as “real” by mistake, it is selected to be a part of the synthetic dataset. Otherwise, it is disregarded and removed from the set. We called this processing “selection.”

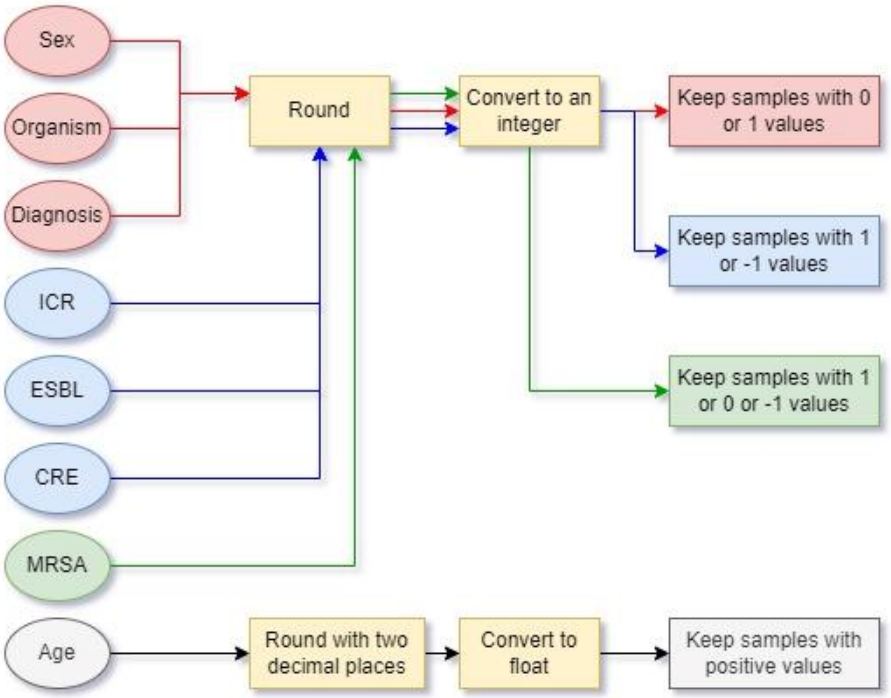


Figure 4-8- The cleaning method for each predictive variable used in this study. Except for age, which became a continuous variable after normalization, the other factors are rounded to the nearest integer value.

As depicted in Figure 4-7, the sensitivity is still relatively low for some families such as CEFT+CLAV, PIP+TAZO, GEN, and COT even after the data augmentation. Since we are

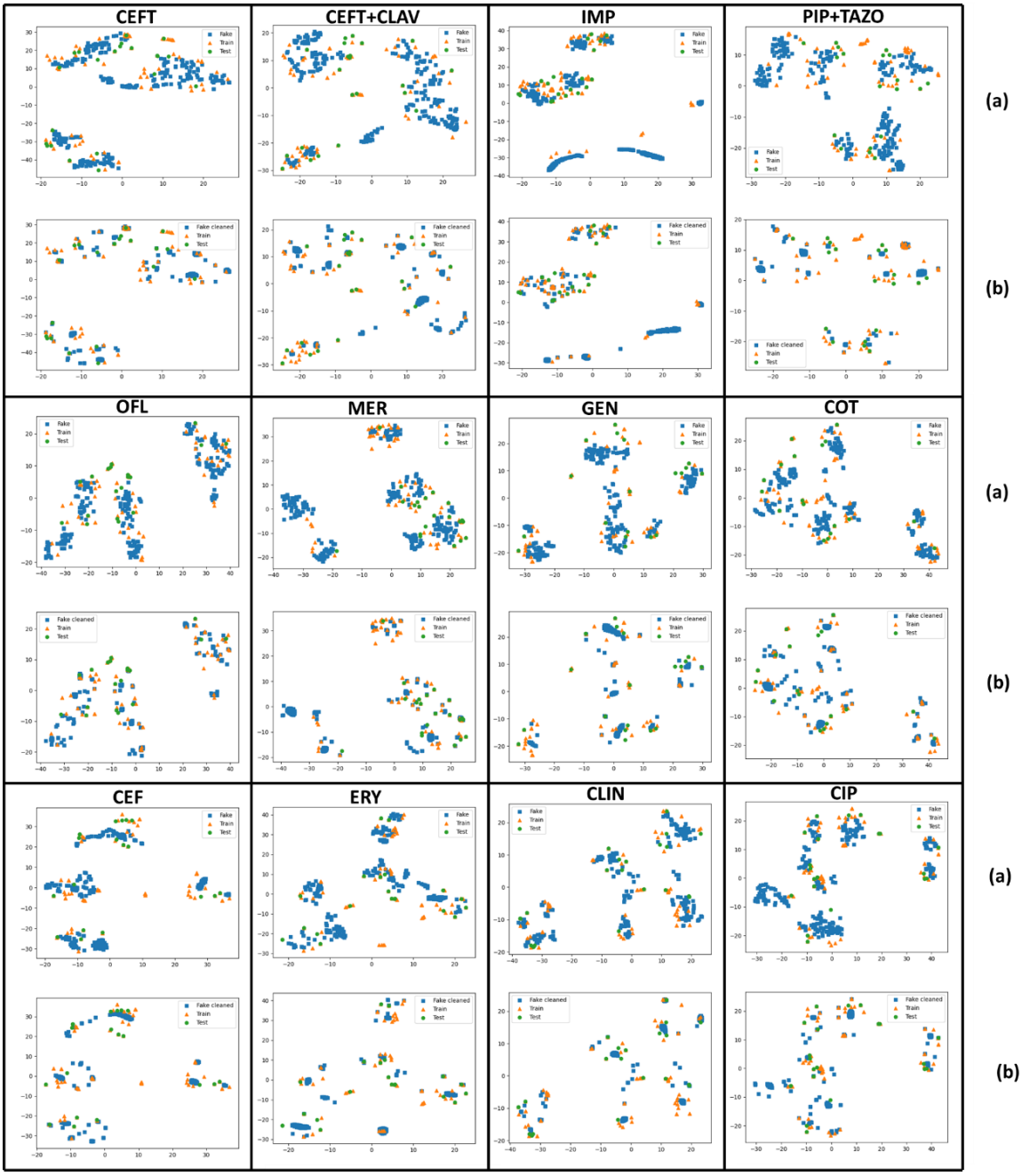


Figure 4-9- Two-dimensional visualizations of the real and generated dataset. (a) before cleaning. (b) after cleaning. Data points with blue, orange and green colors represent the synthetic, train, and test data.

dealing with a very limited and highly imbalanced dataset, usually, we have a maximum of two resistance samples in our test set in some cross-validation folds. With this limited number of

resistance test samples, the sensitivity will be either 0%, 50%, or 100%. Therefore, we cannot evaluate theperformance of the classifiers properly unless we have more samples in our test set. We have generated thousands of synthetic samples for each family to overcome this problem. We performed four types of analysis on this big synthetic dataset as follows:

I. We applied both cleaning and selection to the dataset.

II. We applied just cleaning to the dataset (no selection).

III. We applied just selection to the dataset (no cleaning).

IV. Baseline dataset (no cleaning, no selection).

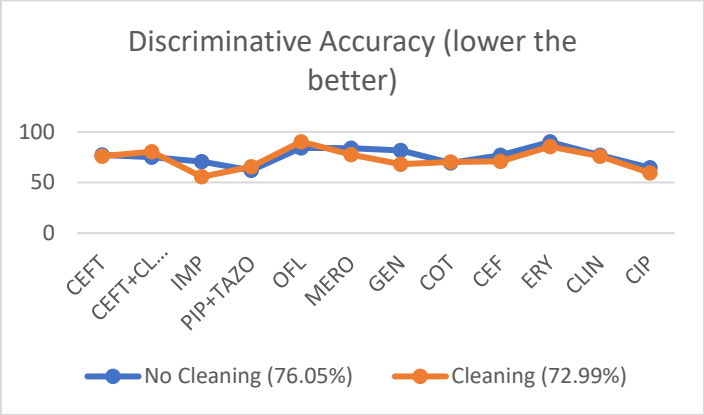


Figure 4-10- Discriminative accuracy of the synthetic samples before and after the cleaning process.

We trained the resistance classifiers using a big synthetic dataset and tested them on the original dataset. Our test set now includes the real training and the real testing set to create a bigger test set for better evaluation. Please note that the training set is only used in training the GANs and is completely unseen by the resistance classifiers. Figure 4-11 represents the AUC, recall, and specificity for the four analyses described above. We achieved the average AUC of 0.83 by applying the cleaning process. The performance improved for antibiotics with relatively low

recall, including CEFT+CLAV, PIP+TAZO, GEN, and COT, with an increase to 51%, 70.66%, 63.24%, and 72.94%, respectively. Surprisingly, applying selection to the cleaned synthetic dataset showed the minimum average AUC of 0.78 among the four analyses. As shown in Figure 4-7, the sensitivity of ERY, which has the lowest performance of 33%, increased to 49.23% with cleaning. In general, selecting more realistic samples did not help with the enhancement of the performance; however, cleaning the synthetic samples improved the performance significantly.

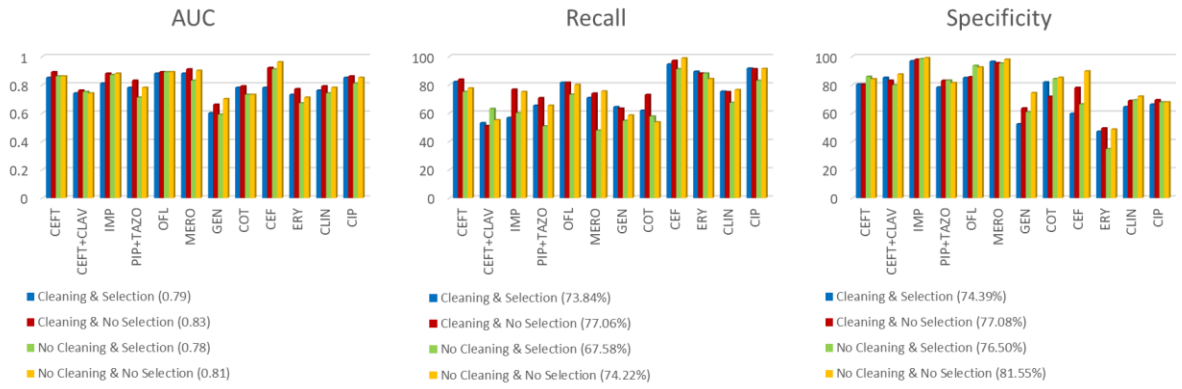


Figure 4-11- Performance metrics of the TSTR method trained and tested on a larger dataset. Each metric is reported for the four types of analysis (cleaning with selection, cleaning with no selection, selection with no cleaning, and the baseline (no selection, no cleaning).

4.5 Discussion and Conclusion

Traditional microbiological susceptibility testing result requires 24-48 hours to be processed; therefore, machine learning techniques could be used as clinical decision support tools to predict antimicrobial resistance and select appropriate empirical antibiotic treatment [109]. This chapter discussed a CWGAN-based model called DermaGAN that generates synthetic samples for SSTI to enlarge the dataset and improve the performance of the antibiotic-resistant classifiers. The research is implemented on a dermatological dataset from 103 patients with GPC and 107 patients with GNB bacteria. Our limited dataset highlights medical data's

scarcity and accessibility challenges in the research communities. Initially, we trained a baseline resistant classifier on the original dataset.

Further, the performance of the DermaGAN in producing realistic features is investigated by training the resistant classifier only on a synthetic dataset. Classifiers trained on synthetic data have shown similar performance to the baseline in most families. We have also visualized the feature space in 2D using a dimension reduction technique (t-SNE). The 2D visualization helps us ensure DermaGANs do not suffer from mode dropping. We also trained the resistant classifiers by enlarging the original training set by adding synthetic data. Up to 11% and 19% of improvement are observed in classification AUC and sensitivity depending on the antibiotic family, respectively. We applied post-synthetic data processing to create more realistic samples. We improved the resistance classifiers' performance by generating a big synthetic dataset, cleaning the samples, and employing them in training the resistance classifiers.

Staphylococcus aureus showed a significant correlation in predicting CIP and COT resistance. These antibiotics are often prescribed in clinical practice as a treatment option for MRSA; however, there is reported that there are limited data to support the effectiveness of these antibiotics in the treatment of methicillin-resistant *Staphylococcus aureus* (MRSA) [110]. The RF classifier has also confirmed this correlation. Previously, the emergence of resistance to CIP and COT in these species has also been discussed in [111, 112].

The carbapenem-resistant *Enterobacteriaceae* (CRE) can cause various infections such as pneumonia, wound, and urinary tract infections. The carbapenem class of antibiotics, including MERO and IMP, represent a valuable option and are considered last-resort antibiotics for treating infections caused by resistant Gram-negative bacteria, including *Enterobacteriaceae*

[113, 114]. The Random Forest classifier also shows the importance of CRE in predicting outcomes.

Our result from the RF classifier confirms that ESBLs can mediate resistance to CEFT, as reported by Jonathan et al. [115]. Age was also a decisive factor in predicting resistance to most antibiotic families. Garcia et al. investigated the correlation between age and antibiotic resistance in patients with positive MRSA. They found that the antibiotics that target DNA syntheses, such as OFL and CIP, show a significant correlation between older patients and antibiotic resistance. However, antibiotics targeting ribosomal functions or cell wall synthesis, such as aminoglycosides, cephalosporins, etc., showed consistent resistance across all age classes [116].

It is reported that any model with AUC> 0.7 can be regarded as a good fit [14]. We achieved the AUC> 0.75 in all the families except for GEN. Generally, it is more challenging to predict resistance for antibiotics associated with unknown or multifactorial resistance mechanisms than those in which resistance is significantly related to a single variable, such as IMP and MERO [102]. Our result shows that the Gram-positive group had slightly worse performance than the Gram-negative group due to their distinctive structure. Gram-negative bacteria are more resistant than Gram-positive bacteria and cause significant morbidity and mortality worldwide [117].

Generally, antibiotics resistance studies have focused on the resistance of single species to different antibiotics and developed an individual machine learning model for each bacterium. For example, Ayyıldız et al. detected the resistance of *Escherichia coli* to a wide range of antibiotics. Those in common with our study are Ceftazidime (acc: 62%, tpr: 75%, tnr:92%), Ciprofloxacin (acc: 76%, tpr: 82%, tnr: 69%), Gentamicin (acc: 84%, tpr: 92%, tnr: 59%), and Piperacillin-Tazobactam (acc: 92%, tpr: 97%, tnr: 44%) [106]. Kim et. al studied the resistance of

Enterobacter cloacae, *Escherichia coli*, *Klebsiella pneumonia*, and *Pseudomonas aeruginosa* to Cefepime (AUC: 1, 0.7, 0.87, 1 for each species, respectively), Meropenem (AUC: n/a, 1, 1, 0.98 for each bacterium, respectively) and Ceftazidime (AUC: n/a, 0.88, 0.99, 0.88 for each bacterium, respectively). In this study, the existence of each species forms a binary vector determined by a binary value of either 1 (existence) or 0 (nonexistence) in the dataset; therefore, there is no need to train a separate network per species.

In conclusion, we proposed a way to enhance the accuracy of antibiotic-resistant detection with minimal data by generating synthetic samples using GANs. Despite the good performance of the resistance classifiers, they are not intended to compete with laboratory testing. Limited data and sampling bias, such as sampling from limited patient demographics within a certain age or gender, present a substantial challenge in any predictive modeling. Our result showed a significant variation in the performance of classifiers trained on a different data distribution in each cross-validation fold due to data imbalance. In highly imbalanced datasets, models trained on datasets with more resistance samples, such as IMP, CIP, OFL, and MERO, showed a higher predictive performance with AUC > 0.90. For optimizing the sensitivity of predictive models for drugs with a low prevalence of resistance samples, we enrich the resistance class by generating synthetic samples using GANs. However, physician expertise remains crucial in prescription choices.

Chapter 5

Contribution and Discussion

5.1 Wound Healing Prognosis Model

Although medical generative adversarial networks have recently shown acceptable performance in generating synthetic patients' records, there have been some limitations that have been taken into account in this study:

1. The previous studies have only focused on binary features, and the input dataset only included count and binary representation of the medical record. Moreover, the input vector size equals the number of the diagnosis, medication, and procedure codes in the available EMR data. This representation creates a very high dimensional input vector and requires having access to a large EMR dataset. It is almost impossible to include all the current diagnoses, medications, and procedure codes in the dataset; therefore, having a small dataset, the GAN learns to generate only a minimal distribution of patients' records and can not cover all the possible distributions.

2. Our dataset consists of both categorical and continuous features. The input data size equals the number of wound prognosis factors listed in Table 3-1. To reduce the input data size, instead of training an autoencoder, we trained a random forest classification model to select variables that significantly contribute to the healing process. With this method, the generator must produce only the informative features to make the model less complex. This representation of the medical

records does not require an extensive dataset; thus, it can overcome the challenges of access to massive EMR data.

3. Previous studies have used a statical EMR dataset. In these studies, temporal information, representing a disease evolution, has not been considered. For example, in wound assessments, the variations in the wound dimensions are deemed essential in predicting the healing status. Temporal information was included and modeled accordingly in the current study to generate a realistic synthetic medical record.

4. The data generated by previous medGANs has never been applied in a real-life application. This evaluation is vital to assess the quality of the generated data as well as their practicality. Hence, a wound prognosis model was trained in the present work to predict a chronic wound healing status within 12 weeks of the initial intake exam. The model was prepared using the generated instances and tested on the original EMR to investigate the functionality of our synthetic medical records.

5. Using GANs techniques in time series medical data, we have provided a pathway to overcoming the challenges of accessing patients' electronic medical records with VLU. This limited access is due to privacy, security, and difficulties extracting useful information.

6. Combining the conditional training with Wasserstein divergence allowed the medical GAN to generate more realistic EMR data with a limited number of train data than the simple training strategy [41].

By training a prognosis model as a real-life application for augmented data by GAN, we could validate the power of the suggested EMR-TCWGAN in producing synthetic time-series EMR. The performance of EMR-TCWGAN in prognosis classification AUC and discriminative accuracy was relatively higher than the baseline model, implying that CNN will perform better than feedforward networks in generating EMR data.

Our prognosis CNN model achieved higher performance than the other state-of-the-art models. The proposed prognosis CNN may help the clinicians with treatment decision-making by alerting them if the wound has a low probability of healing during the early stages.

Based on the definition of a chronic wound, a wound that has not healed in 4-12 weeks, it would be reasonable to consider the critical prognosis factors collected from the first three visits. Moreover, it was reported that although four weeks is a short period to predict a wound healing status clinically, a shorter time for Prediction would help clinicians decide and modify treatment strategies [15]. Our study suggests that it may not be feasible to accurately predict the wound healing rate using only the first-week data (in this case, we achieved AUC=0.647). We believe data from three weeks of follow-up will provide enough information to calculate a strong prediction of healing potential (with AUC=0.875). This early Prediction of wound healing will enhance clinical outcomes and provide efficiencies in care.

Despite the success of the suggested prognosis model in wound healing prediction, the interpretability of the model needs to be considered [118]. Moreover, by reporting the area under the curve, we have compared the overall quality of the prognosis models. However, an optimum decision criterion (threshold) needs to be defined based on the cost of the screening, the prevalence of a disease, and its mortality rate [119].

In conclusion, we developed the pipeline of the medical GANs by representing patients' EMR data based on their prognosis factors and applying deep learning models in medGAN to generate time-series continuous and categorical EMR data. We utilized samples generated by medGAN in training a real-world prognosis classifier to predict the wound healing status within 12 weeks of the first visit. Our experimental results illustrated that the proposed EMR-TCWGAN outperforms the previous EMR-GAN. Moreover, prognosis accuracy has shown a promising development for clinical decision-making.

5.2 Antibiotic Resistance Classifier

This study improves antibiotic resistance classifiers using data augmentation techniques based on generative adversarial networks. The key findings of this project are as follows:

1. Previous studies have mostly focused on the genomic information as input features and investigated if a bacteria genome is resistant to various types of antibiotics. Therefore, they mostly reported performance metrics for the combination of bacteria species and antibiotics. This study, however, investigates if a patient with a specific bacteria and diagnosis is resistant to particular antibiotics by knowing basic patient demographics, diagnosis, bacteria species, and some relevant clinical tests. We have a single dataset to train resistance classifiers for each antibiotic. A binary value determines the existence of 13 bacteria species. If a bacterium exists in the dermatology sample, it is determined by 1; otherwise, 0. With this approach, there is no need to train multiple classifiers for each species-antibiotic combination.

2. To the best of our knowledge, it is the first time that the Generative Adversarial Networks are used as a data augmentation tool for non-image datasets. Despite the

recent development of EMR-GANs, there is no evidence of their application in real-life studies.

3. We achieved a slightly better classification of AUC in most families compared to the previous studies [104, 106]. However, comparing the results to prior studies is still difficult. The reasons are due to differences in objectives, infection types, size of the test data, and the class distributions in the test data.

In conclusion, we developed a pipeline for data augmentation using GANs for non-image data. We employed the generated samples to enlarge our current dataset to improve the performance of the antibiotic resistance classifiers. We hope that these findings lead to developing GANs on non-image datasets and incorporating ML algorithms in real-life applications.

5.3 Future Directions

The future directions for this thesis are:

- Since very limited data was used in this study, adding more real data to the dataset will help create robust and more reliable prognosis algorithms.
- Hyperparameter tuning of the GANs across different cross-validation folds is required to avoid poor performance in some folds.
- The wound healing prognosis model can be improved by incorporating image data and performing hybrid machine learning algorithms.
- Enriching the antibiotic resistance dataset by including further clinical information such as recorded symptoms and patient comorbidities may increase the classifiers' performance.

Bibliography

1. Chen, S.-H., *Computational intelligence in economics and finance: Carrying on the legacy of Herbert Simon.* Information Sciences, 2005. **170**(1): p. 121-131.
2. Ma, L. and B. Sun, *Machine learning and AI in marketing–Connecting computing power to human insights.* International Journal of Research in Marketing, 2020. **37**(3): p. 481-504.
3. Aziz, S. and M. Dowling, *Machine learning and AI for risk management*, in *Disrupting finance*. 2019, Palgrave Pivot, Cham. p. 33-50.
4. Sun, Y., et al., *Battery-based energy storage transportation for enhancing power system economics and security.* IEEE Transactions on Smart Grid, 2015. **6**(5): p. 2395-2402.
5. Topol, E.J., *High-performance medicine: the convergence of human and artificial intelligence.* Nature medicine, 2019. **25**(1): p. 44-56.
6. Toh, T.S., F. Dondelinger, and D. Wang, *Looking beyond the hype: Applied AI and machine learning in translational medicine.* EBioMedicine, 2019. **47**: p. 607-615.
7. Hamet, P. and J. Tremblay, *Artificial intelligence in medicine.* Metabolism, 2017. **69**: p. S36-S40.
8. Han, G. and R. Ceilley, *Chronic wound healing: a review of current management and treatments.* Advances in therapy, 2017. **34**(3): p. 599-610.
9. Wang, H.-H., et al., *Assessment of deep learning using nonimaging information and sequential medical records to develop a prediction model for nonmelanoma skin cancer.* JAMA dermatology, 2019. **155**(11): p. 1277-1283.
10. Ayer, T., et al., *Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration.* Cancer, 2010. **116**(14): p. 3310-3321.
11. Chokwijitkul, T., et al. *Identifying risk factors for heart disease in electronic medical records: A deep learning approach*. in *Proceedings of the BioNLP 2018 workshop*. 2018.
12. Gheshlaghi, S.H., et al. *Efficient Oct Image Segmentation Using Neural Architecture Search*. in *2020 IEEE International Conference on Image Processing (ICIP)*. 2020. IEEE.
13. Liang, Z., et al. *Deep learning for healthcare decision making with EMRs*. in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2014. IEEE.
14. Cho, S.K., et al., *Development of a model to predict healing of chronic wounds within 12 weeks.* Advances in Wound Care, 2020.
15. Cukjati, D., et al., *Prognostic factors in the prediction of chronic wound healing by electrical stimulation.* Medical and Biological Engineering and Computing, 2001. **39**(5): p. 542-550.
16. Chu, C.S., et al., *Machine learning and treatment outcome prediction for oral cancer.* Journal of Oral Pathology & Medicine, 2020. **49**(10): p. 977-985.
17. Shoeb, A.H., *Application of machine learning to epileptic seizure onset detection and treatment*. 2009, Massachusetts Institute of Technology.
18. Myszczynska, M.A., et al., *Applications of machine learning to diagnosis and treatment of neurodegenerative diseases.* Nature Reviews Neurology, 2020. **16**(8): p. 440-456.
19. Chekroud, A.M., et al., *Cross-trial prediction of treatment outcome in depression: a machine learning approach.* The Lancet Psychiatry, 2016. **3**(3): p. 243-250.

20. Wang, F. and A. Preininger, *AI in health: state of the art, challenges, and future directions.* Yearbook of medical informatics, 2019. **28**(01): p. 016-026.
21. Siddiqui, A.R. and J.M. Bernstein, *Chronic wound infection: facts and controversies.* Clinics in dermatology, 2010. **28**(5): p. 519-526.
22. Kurd, S.K., et al., *Evaluation of the use of prognostic information for the care of individuals with venous leg ulcers or diabetic neuropathic foot ulcers.* Wound Repair and Regeneration, 2009. **17**(3): p. 318-325.
23. Skene, A., et al., *Venous leg ulcers: a prognostic index to predict time to healing.* British Medical Journal, 1992. **305**(6862): p. 1119-1121.
24. FRANKS, P.J., et al., *Factors associated with healing leg ulceration with high compression.* Age and ageing, 1995. **24**(5): p. 407-410.
25. Karanikolic, V., et al., *Prognostic factors related to delayed healing of venous leg ulcer treated with compression therapy.* Dermatologica sinica, 2015. **33**(4): p. 206-209.
26. Margolis, D.J., et al., *The accuracy of venous leg ulcer prognostic models in a wound care system.* Wound Repair and Regeneration, 2004. **12**(2): p. 163-168.
27. Margolis, D.J., et al., *Risk factors for delayed healing of neuropathic diabetic foot ulcers: a pooled analysis.* Archives of dermatology, 2000. **136**(12): p. 1531-1535.
28. Margolis, D.J., J.A. Berlin, and B.L. Strom, *Risk factors associated with the failure of a venous leg ulcer to heal.* Archives of dermatology, 1999. **135**(8): p. 920-926.
29. Khachemoune, A., Y.M. Bello, and T.J. Phillips, *Factors that influence healing in chronic venous ulcers treated with cryopreserved human epidermal cultures.* Dermatologic surgery, 2002. **28**(3): p. 274-280.
30. Phillips, T.J., et al., *Prognostic indicators in venous ulcers.* Journal of the American Academy of Dermatology, 2000. **43**(4): p. 627-630.
31. Ki, V. and C. Rotstein, *Bacterial skin and soft tissue infections in adults: a review of their epidemiology, pathogenesis, diagnosis, treatment and site of care.* Canadian Journal of Infectious Diseases and Medical Microbiology, 2008. **19**(2): p. 173-184.
32. Kaye, K.S., et al., *Current epidemiology, etiology, and burden of acute skin infections in the United States.* Clinical Infectious Diseases, 2019. **68**(Supplement_3): p. S193-S199.
33. Ramakrishna, M.S., et al., *Microbial Profile and Antibiogram Pattern Analysis of Skin and Soft Tissue Infections at A Tertiary Care Center in South India.* Journal of Pure and Applied Microbiology, 2021.
34. Dryden, M.S., *Skin and soft tissue infection: microbiology and epidemiology.* International journal of antimicrobial agents, 2009. **34**: p. S2-S7.
35. Keil, F., et al., *Use of daptomycin for Gram-positive infections in neutropenic patients: Clinical experience from a European Outcomes Registry.* Transplantation. **124**: p. 28.
36. Mohareb, A.M., et al. *Addressing antibiotic overuse in the outpatient setting: lessons from behavioral economics*. in *Mayo Clinic Proceedings*. 2021. Elsevier.
37. Gandhi, N.R., et al., *Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis.* The Lancet, 2010. **375**(9728): p. 1830-1843.
38. Ramakrishna, M.S., et al., *Microbial Profile and Antibiogram Pattern Analysis of Skin and Soft Tissue Infections at A Tertiary Care Center in South India.* J Pure Appl Microbiol, 2021. **15**(2): p. 915-925.
39. Benkova, M., O. Soukup, and J. Marek, *Antimicrobial susceptibility testing: currently used methods and devices and the near future in clinical practice.* Journal of Applied Microbiology, 2020. **129**(4): p. 806-822.

40. Sanchez, G.V., et al., *Antibiotic resistance among urinary isolates from female outpatients in the United States in 2003 and 2012.* Antimicrobial agents and chemotherapy, 2016. **60**(5): p. 2680-2683.
41. Reed, W. *MRSN*. 2022.
42. CDC. *National Antimicrobial Resistance Monitoring System for Enteric Bacteria (NARMS)*. 2022.
43. Nguyen, P., et al., *$\mathtt {Deepr} $: a convolutional net for medical records.* IEEE journal of biomedical and health informatics, 2016. **21**(1): p. 22-30.
44. Ho, L.V., et al. *The dependence of machine learning on electronic medical record quality*. in *AMIA Annual Symposium Proceedings*. 2017. American Medical Informatics Association.
45. Kim, Y.J., et al., *Highrisk prediction from electronic medical records via deep attention networks.* arXiv preprint arXiv:1712.00010, 2017.
46. Sahni, N., G. Simon, and R. Arora, *Development and validation of machine learning models for prediction of 1-year mortality utilizing electronic medical record data available at the end of hospitalization in multicondition patients: a proof-of-concept study.* Journal of general internal medicine, 2018. **33**(6): p. 921-928.
47. Yang, H.-C., et al., *Artificial Intelligence–Based Prediction of Lung Cancer Risk Using Nonimaging Electronic Medical Records: Deep Learning Approach.* Journal of medical Internet research, 2021. **23**(8): p. e26256.
48. Ningrum, D.N.A., et al., *A Deep Learning Model to Predict Knee Osteoarthritis Based on Nonimage Longitudinal Medical Record.* Journal of Multidisciplinary Healthcare, 2021. **14**: p. 2477.
49. Kaur, H., et al., *Automated chart review utilizing natural language processing algorithm for asthma predictive index.* BMC pulmonary medicine, 2018. **18**(1): p. 1-9.
50. Sung, S.-F., et al., *Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: a feasibility study.* International journal of medical informatics, 2018. **112**: p. 149-157.
51. Zhang, Z., et al., *Ensuring electronic medical record simulation through better training, modeling, and evaluation.* Journal of the American Medical Informatics Association, 2020. **27**(1): p. 99-108.
52. Miller, A.R. and C. Tucker, *Privacy protection and technology diffusion: The case of electronic medical records.* Management science, 2009. **55**(7): p. 1077-1093.
53. Enaizan, O., et al., *Effects of privacy and security on the acceptance and usage of EMR: the mediating role of trust on the basis of multiple perspectives.* Informatics in Medicine Unlocked, 2020. **21**: p. 100450.
54. Baowaly, M.K., et al., *Synthesizing electronic health records using improved generative adversarial networks.* Journal of the American Medical Informatics Association, 2019. **26**(3): p. 228-241.
55. Frid-Adar, M., et al., *GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification.* Neurocomputing, 2018. **321**: p. 321-331.
56. Bowles, C., et al., *Gan augmentation: Augmenting training data using generative adversarial networks.* arXiv preprint arXiv:1810.10863, 2018.
57. Han, C., et al. *GAN-based synthetic brain MR image generation*. in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018. IEEE.

58. Islam, J. and Y. Zhang, *GAN-based synthetic brain PET image generation.* Brain informatics, 2020. **7**(1): p. 1-12.
59. Shin, H.-C., et al. *Medical image synthesis for data augmentation and anonymization using generative adversarial networks*. in *International workshop on simulation and synthesis in medical imaging*. 2018. Springer.
60. Togo, R., T. Ogawa, and M. Haseyama, *Synthetic gastritis image generation via loss function-based conditional pggan.* IEEE access, 2019. **7**: p. 87448-87457.
61. Hung, C.-Y., et al. *Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database*. in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2017. IEEE.
62. Taud, H. and J. Mas, *Multilayer perceptron (MLP)*, in *Geomatic Approaches for Modeling Land Change Scenarios*. 2018, Springer. p. 451-455.
63. Yan, H., et al., *A multilayer perceptron-based medical decision support system for heart disease diagnosis.* Expert Systems with Applications, 2006. **30**(2): p. 272-281.
64. Hosseinzadeh, M., et al., *A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things.* The Journal of Supercomputing, 2021. **77**(4): p. 3616-3637.
65. Ting, F. and K. Sim. *Self-regulated multilayer perceptron neural network for breast cancer classification*. in *2017 International Conference on Robotics, Automation and Sciences (ICORAS)*. 2017. IEEE.
66. Yildirim, P. *Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction*. in *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*. 2017. IEEE.
67. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436-444.
68. Teng, L., H. Li, and S. Karim, *DMCNN: A deep multiscale convolutional neural network model for medical image segmentation.* Journal of Healthcare Engineering, 2019. **2019**.
69. Wang, C., et al., *Fully automatic wound segmentation with deep convolutional neural networks.* Scientific Reports, 2020. **10**(1): p. 1-9.
70. Guo, Z., et al. *Medical image segmentation based on multi-modal convolutional neural network: Study on image fusion schemes*. in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018. IEEE.
71. Rostami, B., et al., *Multiclass Wound Image Classification using an Ensemble Deep CNN-based Classifier.* Computers in Biology and Medicine, 2021: p. 104536.
72. Anisuzzaman, D., et al., *Multi-modal Wound Classification using Wound Image and Location by Deep Neural Network.* arXiv preprint arXiv:2109.06969, 2021.
73. Ting, F.F., Y.J. Tan, and K.S. Sim, *Convolutional neural network improvement for breast cancer classification.* Expert Systems with Applications, 2019. **120**: p. 103-115.
74. Alom, M.Z., et al., *Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network.* Journal of digital imaging, 2019. **32**(4): p. 605-617.
75. Zou, L., et al., *A technical review of convolutional neural network-based mammographic breast cancer diagnosis.* Computational and mathematical methods in medicine, 2019. **2019**.

76. Jifara, W., et al., *Medical image denoising using convolutional neural network: a residual learning approach.* The Journal of Supercomputing, 2019. **75**(2): p. 704-718.
77. Zhang, Y. and H. Yu, *Convolutional neural network based metal artifact reduction in x-ray computed tomography.* IEEE transactions on medical imaging, 2018. **37**(6): p. 1370-1381.
78. Goodfellow, I.J., et al., *Generative adversarial networks.* arXiv preprint arXiv:1406.2661, 2014.
79. Arjovsky, M., S. Chintala, and L. Bottou. *Wasserstein generative adversarial networks*. in *International conference on machine learning*. 2017. PMLR.
80. Gulrajani, I., et al., *Improved training of wasserstein gans.* arXiv preprint arXiv:1704.00028, 2017.
81. Choi, E., et al. *Generating multi-label discrete patient records using generative adversarial networks*. in *Machine learning for healthcare conference*. 2017. PMLR.
82. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
83. Khalilia, M., S. Chakraborty, and M. Popescu, *Predicting disease risks from highly imbalanced data using random forest.* BMC medical informatics and decision making, 2011. **11**(1): p. 1-13.
84. Sage, A., *Random forest robustness, variable importance, and tree aggregation.* 2018.
85. Stoltzfus, J.C., *Logistic regression: a brief primer.* Academic Emergency Medicine, 2011. **18**(10): p. 1099-1104.
86. Si, S., et al. *Gradient boosted decision trees for high dimensional sparse output*. in *International conference on machine learning*. 2017. PMLR.
87. Kumar, R. and A. Indrayan, *Receiver operating characteristic (ROC) curve for medical researchers.* Indian pediatrics, 2011. **48**(4): p. 277-287.
88. Lalkhen, A.G. and A. McCluskey, *Clinical tests: sensitivity and specificity.* Continuing Education in Anaesthesia Critical Care & Pain, 2008. **8**(6): p. 221-223.
89. Jung, K., et al., *Rapid identification of slow healing wounds.* Wound Repair and Regeneration, 2016. **24**(1): p. 181-188.
90. Margolis, D.J., et al., *Diabetic neuropathic foot ulcers: predicting which ones will not heal.* The American journal of medicine, 2003. **115**(8): p. 627-631.
91. Kaur, P., R. Kumar, and M. Kumar, *A healthcare monitoring system using random forest and internet of things (IoT).* Multimedia Tools and Applications, 2019. **78**(14): p. 19905-19916.
92. Alam, M.Z., M.S. Rahman, and M.S. Rahman, *A Random Forest based predictor for medical data classification using feature ranking.* Informatics in Medicine Unlocked, 2019. **15**: p. 100180.
93. Yoon, J., D. Jarrett, and M. Van der Schaar, *Time-series generative adversarial networks.* Advances in neural information processing systems, 2019. **32**.
94. Esteban, C., S.L. Hyland, and G. Rätsch, *Real-valued (medical) time series generation with recurrent conditional gans.* arXiv preprint arXiv:1706.02633, 2017.
95. McInnes, L., J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction.* arXiv preprint arXiv:1802.03426, 2018.
96. Massey Jr, F.J., *The Kolmogorov-Smirnov test for goodness of fit.* Journal of the American statistical Association, 1951. **46**(253): p. 68-78.
97. Zoufal, C., A. Lucchi, and S. Woerner, *Quantum generative adversarial networks for learning and loading random distributions.* npj Quantum Information, 2019. **5**(1): p. 1-9.

98. Aviñó, L., M. Ruffini, and R. Gavaldà, *Generating synthetic but plausible healthcare record datasets.* arXiv preprint arXiv:1807.01514, 2018.
99. Bradley, A.P., *The use of the area under the ROC curve in the evaluation of machine learning algorithms.* Pattern recognition, 1997. **30**(7): p. 1145-1159.
100. Hajian-Tilaki, K., *Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation.* Caspian journal of internal medicine, 2013. **4**(2): p. 627.
101. Heaton, J., et al. *Early stabilizing feature importance for TensorFlow deep neural networks*. in *2017 International Joint Conference on Neural Networks (IJCNN)*. 2017. IEEE.
102. Hicks, A.L., et al., *Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data.* PLoS computational biology, 2019. **15**(9): p. e1007349.
103. Kim, J., et al., *VAMPr: VA riant M apping and P rediction of antibiotic r esistance via explainable features and machine learning.* PLoS computational biology, 2020. **16**(1): p. e1007511.
104. Feretzakis, G., et al., *Using machine learning techniques to aid empirical antibiotic therapy decisions in the intensive care unit of a general hospital in Greece.* Antibiotics, 2020. **9**(2): p. 50.
105. Lewin-Epstein, O., et al., *Predicting Antibiotic Resistance in Hospitalized Patients by Applying Machine Learning to Electronic Medical Records.* Clinical Infectious Diseases, 2020. **72**(11): p. e848-e855.
106. Ayyıldız, H. and S.A. Tuncer, *Is it possible to determine antibiotic resistance of E. coli by analyzing laboratory data with machine learning?* Turkish Journal of Biochemistry, 2021. **46**(6): p. 623-630.
107. Kavvas, E.S., et al., *Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance.* Nature communications, 2018. **9**(1): p. 1-9.
108. Smith, K.E. and A.O. Smith, *Conditional GAN for timeseries generation.* arXiv preprint arXiv:2006.16477, 2020.
109. Feretzakis, G., et al., *Machine learning for antibiotic resistance prediction: A prototype using off-the-shelf techniques and entry-level data to guide empiric antimicrobial therapy.* Healthcare Informatics Research, 2021. **27**(3): p. 214-221.
110. Hong, J., M.H. Ensom, and T.T. Lau, *What Is the Evidence for Co-trimoxazole, Clindamycin, Doxycycline, and Minocycline in the Treatment of Methicillin-Resistant Staphylococcus aureus (MRSA) Pneumonia?* Annals of Pharmacotherapy, 2019. **53**(11): p. 1153-1161.
111. Ball, P., *Emergent resistance to ciprofloxacin amongst Pseudomonas aeruginosa and Staphylococcus aureus: clinical significance and therapeutic approaches.* Journal of Antimicrobial Chemotherapy, 1990. **26**(suppl_F): p. 165-179.
112. Blumberg, H.M., et al., *Rapid development of ciprofloxacin resistance in methicillin-susceptible and-resistant Staphylococcus aureus.* Journal of Infectious Diseases, 1991. **163**(6): p. 1279-1285.
113. Pascale, R., et al., *Use of meropenem in treating carbapenem-resistant Enterobacteriaceae infections.* Expert Review of Anti-infective Therapy, 2019. **17**(10): p. 819-827.

114. Sheu, C.-C., et al., *Infections caused by carbapenem-resistant Enterobacteriaceae: an update on therapeutic options.* Frontiers in microbiology, 2019. **10**: p. 80.
115. Jonathan, N., *Screening for extended-spectrum beta-lactamase-producing pathogenic enterobacteria in district general hospitals.* Journal of clinical microbiology, 2005. **43**(3): p. 1488-1490.
116. Garcia, A., T. Delorme, and P. Nasr, *Patient age as a factor of antibiotic resistance in methicillin-resistant Staphylococcus aureus.* Journal of medical microbiology, 2017. **66**(12): p. 1782-1789.
117. Breijyeh, Z., B. Jubeh, and R. Karaman, *Resistance of gram-negative bacteria to current antibacterial agents and approaches to resolve it.* Molecules, 2020. **25**(6): p. 1340.
118. Holzinger, A., et al., *Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI.* Information Fusion, 2021. **71**: p. 28-37.
119. Safari, S., et al., *Evidence based emergency medicine; part 5 receiver operating curve and area under the curve.* Emergency, 2016. **4**(2): p. 111.