




Improving Classification of Tetanus Severity for Patients in Low-Middle Income Countries Wearing ECG Sensors by Using a CNN-Transformer Network

Ping Lu , Chenyang Wang , Jannis Hagenah, Shadi Ghiasi, VITAL consortium, Tingting Zhu , Louise Thwaites, and David A. Clifton

Abstract—Tetanus is a life-threatening infectious disease, which is still common in low- and middle-income countries, including in Vietnam. This disease is characterized by muscle spasm and in severe cases is complicated by autonomic dysfunction. Ideally continuous vital sign monitoring using bedside monitors allows the prompt detection of the onset of autonomic nervous system dysfunction or avoiding rapid deterioration. Detection can be improved using heart rate variability analysis from ECG signals. Recently, characteristic ECG and heart rate variability features have been shown to be of value in classifying tetanus severity. However, conventional manual analysis of ECG is time-consuming. The traditional convolutional neural network (CNN) has limitations in extracting the global context information, due to its fixed-sized kernel filters. In this work, we propose a novel hybrid CNN-Transformer model to automatically classify tetanus severity using tetanus monitoring from low-cost wearable sensors. This model can capture the local features from the CNN and the global features from the Transformer. The time series imaging - spectrogram - is transformed from one-dimensional ECG signal and input to the proposed model. The CNN-Transformer model outperforms state-of-the-art methods in tetanus classification, achieves results with a F1 score of 0.82 ± 0.03 , precision of 0.94 ± 0.03 , recall of 0.73 ± 0.07 , specificity of 0.97 ± 0.02 , accuracy of 0.88 ± 0.01 and AUC of 0.85 ± 0.03 . In addition, we found that Random Forest with enough manually selected features can be comparable with the proposed CNN-Transformer model.

Index Terms—Classification, CNN, transformer, electrocardiogram, tetanus, spectrogram.

I. INTRODUCTION

TETANUS is a vaccine-preventable infectious disease, caused by a neurotoxin produced by the *Clostridium tetani* bacterium [1]. Tetanus is estimated to cause around 213 000–293 000 deaths in the world each year, including 5–7% of all neonatal deaths and 5% of maternal deaths globally [2]. Although tetanus is rare in high-income countries, it is still common in many low- and middle-income countries (LMIC) [3], [4], [5]. In 2015, 79% of deaths due to tetanus (44612 of 56743) were estimated to occur in south Asia and sub-Saharan Africa [6].

Tetanus is caused by a powerful neurotoxin which inhibits transmission at central nervous system synapses, resulting in muscle stiffness and spasms. In severe cases, cardiovascular system instability occurs. Over a period of 2–5 days, approximately half of all patients will progress to severe disease where mechanical ventilation is needed. Around 25% of all patients experience autonomic nervous system (ANS) dysfunction, affecting heart rate and blood pressure. This is the leading cause of death for tetanus patients. The early detection of severe tetanus is highly valuable, because it allows timely intervention and allows more appropriate resource utilization [7]. The Ablett score is the simplest classification system for tetanus severity, ranging from 1 to 4 [5]. In grades 1 and 2, (mild or moderate disease), patients' clinical conditions can be managed without the need for invasive Intensive Care Unit (ICU) intervention such as mechanical ventilation. In grades 3 and 4 of disease, patients have severe disease requiring mechanical ventilation and, in the case of grade 4 disease, may require additional organ support to manage ANS dysfunction [4], [8]. Conventionally Ablett grading relies on a combination of clinical features, many of which may occur in other co-existing conditions (e.g., fever, hypertension and tachycardia). In busy clinical settings or those with limited clinical staff experience accurate classification may be difficult.

Providing ICU care is expensive in all countries, including LMICs. For diseases, such as tetanus, where patients may deteriorate rapidly, all patients require hospital admission for

Manuscript received 24 May 2022; revised 16 September 2022 and 17 October 2022; accepted 18 October 2022. Date of publication 21 October 2022; date of current version 21 March 2023. This work was supported by the Wellcome Trust under Grant 217650/Z/19/Z. The work of David A. Clifton was supported by NIHR Oxford Biomedical Research Centre, the InnoHK Hong Kong Centre for Cerebro-cardiovascular Health Engineering, and the Pandemic Sciences Institute at the University of Oxford. (Corresponding author: Ping Lu.)

Ping Lu is with the Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, U.K. (e-mail: ping.lu@eng.ox.ac.uk).

Chenyang Wang, Jannis Hagenah, Shadi Ghiasi, and Tingting Zhu are with the Department of Engineering Science, University of Oxford, U.K.

Louise Thwaites is with the Oxford University Clinical Research Unit, Vietnam.

David A. Clifton is with the Department of Engineering Science, University of Oxford, U.K., and also with the Oxford Suzhou Centre for Advanced Research, University of Oxford, China.

Digital Object Identifier 10.1109/TBME.2022.3216383

careful observation (and if necessary rapid emergency treatment). In most low-resource settings this means admission to high-dependency or intensive care units (ICUs) as these are the only places with sufficient staff and equipment to do this. This large burden of additional cases results in suboptimal use of already scarce resources and likely worsens outcomes for those who do require ICU level care [9], [10], [11]. Additionally in countries like Vietnam where many patients pay for care out-of-pocket, the extra costs of ICU care, compared to normal ward care, are significant. There are the information about direct medical costs of tetanus, dengue, and sepsis patients in an ICU in Vietnam from previous research [9], [10], [11]. If patients do not require mechanical ventilation, the median total ICU cost per patient varied between US\$64.40 and US\$675 for the different diseases [9]. If patients required mechanical ventilation, the costs were higher, and the median total ICU cost per patient for the different diseases varied between US\$2,590 and US\$4,250 [9]. The main cost drivers varied depending on disease and its related severity [9].

In high-income countries, complex continuous monitoring systems and high staff-to-patient ratios facilitate improved the tetanus outcomes [11]. However, in LMICs, inexperienced staff, lack of equipment with limited time are commonly cited impediments to providing high quality care for patients with tetanus. Low-cost wearable sensors have been proposed as an alternative solution for tetanus in resource-limited settings. The wearable sensors are small, lightweight and wireless. They can continuously monitor vital signs in real-time, in order to help in the early identification of patient deterioration [11], [12]. One challenge of utilizing low-cost wearable sensors is that the recorded continuous physiological data can be less precise, due to the large amount of noise (caused by muscle movement and the monitors electrical source) and missing data [11]. The ultimate aim of our work is to develop a tetanus severity warning tool, which can improve the clinical outcomes and disease incidence. This warning tool will classify tetanus disease severity based on the patient's electrocardiogram (ECG) data using wearable sensors. This tool would be appropriate for low-resource settings where lack of equipment and staff impacts patient care, but also high income settings where limited numbers of cases of tetanus means staff are inexperienced in tetanus management. The tool could assist clinical decision making, avoiding unnecessary ICU admissions (for mild cases) and reducing treatment delays (for severe cases).

In this paper, we use ECG data collected with wearable sensors in a Vietnamese ICU and propose a warning tool, created using a deep learning approach, in order to classify tetanus severity indicated by Ablett score. In our recent study [13], we used 2D Convolutional Neural Network (CNN) with a channel-wise attention mechanism to classify the severity of tetanus using wearable monitors in a resource-limited setting. Our method outperforms 1D CNN, 2D CNN and 2D CNN with attention mechanism by combining either the gating function or sequential techniques. We also discuss how the window length of log-spectrograms of ECG signals influences the performance of the proposed method. Our channel-wise based method employs the 2D convolution for the feature extraction. Because the CNN

captures the local structure with a fixed size of convolution kernel, the pixels which are far away from the receptive field will not affect the value of the feature calculated by the convolution kernel. Hence, it is unable to extract the global information of the input image. Hence, we add the global information of the log-spectrograms in this work. The contributions of this work are as follows:

- We propose a novel hybrid CNN-Transformer network for classifying the severity of tetanus, which captures both rich local features and global context information of the 2-dimensional (2D) logarithmic spectrogram of ECG signals. This completely data-driven network could be transferable to similar infectious diseases.
- To the best of our knowledge, this is first transformer-based network in tetanus diseases classification for capturing global context information, which cannot be obtained from CNNs.
- The proposed network outperforms CNN and state-of-art vision transformer on the 2D logarithmic spectrogram of the ECG signal acquired from tetanus patients.

The paper is structured as follows: Section II introduces related work in the diagnosis of tetanus diseases in LMIC, time series imaging and deep learning approaches. Section III describes the proposed network for classifying the severity of tetanus in intensive-care settings. Section IV provides the details of our collected tetanus dataset, implementation details, a comparison of baseline methods and the evaluations of the classification results with several performance metrics. Section V presents and discusses the experimental results. Finally, Section VII provides the conclusion of our work.

II. RELATED WORK

The diagnosis of lethal infectious diseases plays a significant role in patient treatment. In tetanus, disease severity is associated with autonomic nervous system (ANS) activity [13]. Heart rate variability (HRV) represents the variation of beat-to-beat in RR intervals. This variation is controlled by the autonomic nervous system (ANS) and it indicates ANS activity [13]. The changes in conventional HRV parameters measured from ECG have been shown to correlate with tetanus disease severity. In HRV-based methods for classifying tetanus severity, an extra pre-processing step is needed, after which features - RR intervals and QRS complex - are then extracted [14], [15], [16], [17]. Duong et al. [14] demonstrated the utility of ANSD. However, conventional methods of HRV detection require expensive equipment and expertise which is usually not available in ICU or low-resource settings. Van et al. [11] used wearable devices to extract RR intervals from tetanus ECG recordings. However, it is still a challenge to robustly extract RR intervals [18].

Artificial intelligence, including machine learning (ML) and deep learning methods, has revolutionized healthcare for diagnosing and classifying the severity of infectious diseases. Traditional ML approaches require the feature engineering process for manually extracting features such as RR intervals from the dataset [19]. Tadesse et al. [20] applied support vector machines (SVM) to automatically detect the ANS dysfunction

level for tetanus and demonstrated SVM outperforms HRV on infectious diseases detection. Pathological photoplethysmogram (PPG) signals have been used to improve diagnosis performance for hand, foot and mouth disease (HFMD) and tetanus [21]. Time series imaging - spectrograms - have been employed to classify the severity of two infectious diseases - Tetanus and HFMD - using ECG and PPG with transfer learning in [18]. Deep learning methods have been approved outperforming traditional machine learning methods (e.g., SVM) [18]. In previous research, the synchronised ECG and PPG data from 10 tetanus patients were used in [18], [20], and PPG data from 19 tetanus patients were utilised in [21]. Because of the small datasets used in the previous work, the experimental results were limited.

Time series imaging is a popular technology which transforms time series data into images, such as recurrence plot, gramian angular field and spectrogram. Time series imaging is widely used in 2D CNNs for classification tasks, e.g., the 2D representation of time-frequency analysis - spectrogram, log spectrogram, mel spectrogram, and scalogram - is used in 2D CNN [18], [22], [23], [24], [25]. The 2D CNN has a promising performance in image classification, according to the recent literature work. Although the 1-dimensional (1D) CNNs have been employed in signal processing applications, such as biomedical data classification and early diagnosis [26], [27], an image-based ECG signal classification structure using 2D spectrograms achieves a better performance than the 1D CNN [28].

Transformers [29] were first introduced for natural language processing (NLP) dealing with sequential input data. Transformers can capture global/long range dependencies using parallel self-attention mechanisms in various NLP tasks. A standard Transformer layer [29] includes a multi-head attention mechanism modelling global relationships between sequence tokens, and a feed-forward network (FFN) learning wider representations.

Inspired by the novel architecture of transformers in NLP, transformers are recently applied to computer vision tasks [30]. Vision Transformer (ViT) [31] achieved the state-of-the-art performance on image classification. The ViT is good at capturing long-range dependencies between patches. Firstly, images are split into 16×16 non-overlapping patches. These patches combined with positional encoding input into transformer blocks to model global relations for classification. Several variants of ViT have been suggested to improve the performance on vision tasks. Data-efficient image Transformer (DeiT) [32] is a type of ViT for image classification using knowledge distillation [33]. Swin Transformer is a hierarchical ViT using shifted windows [34]. TNT [35] chooses an inner transformer block to process the relationship between sub-patches and an outer transformer block to model the relationship among patch-level embeddings. Transformers have been raised in computer vision and image analysis [36], [37], [38], [39], [40]. Inspired by transformers on audio spectrograms [41], [42], [43], transformers have great potential on tetanus ECG spectrograms for improving the performance of classifying the severity of tetanus in intensive-care settings.

III. METHOD

A. Data Preprocessing

ECG signal denoising is a crucial pre-processing step. Low band frequency noise [44] and high band frequency noise [44] are primarily two types of noise that disturb the ECG signal analysis. They are caused by patient muscle movement and the electrical source operating the ECG monitor, respectively. In this work, we choose one-lead ECG signals acquired from the low-cost wearable monitor. Then we remove the background noise and clean the data using the Butterworth filter. We set a cutoff point of 0.05 Hz and 100 Hz, for the high-pass filter and the low-pass filter respectively. The implementation is performed utilizing the SciPy package [45].

B. Time Series Imaging

In the proposed method named 2D-CNN-Transformer, the input data is required to be a 2D type of image. We can use time series imaging converted from the one-dimensional ECG to be the input. Spectrograms are one of the most widely used 2D images as the representations for signal. Spectrograms are described by the time series spectra along one axis and frequency along the other axis. The logarithmic spectrogram is a log-scaled spectrogram based on the consecutive Fourier transform and this scale gives more attention on lower frequencies. Hence, we choose logarithmic spectrograms \tilde{I} as input in our proposed method. We normalise the spectrograms I by their maximum value and scale the value in the range 0 to 255, and scale the normalised spectrograms as a logarithm (see (1)).

$$\tilde{I} = \log \left(\frac{I}{\max(I)} * 255 \right) \quad (1)$$

We resize and stitch on the 2D logarithmic spectrogram. In our work, we choose 60 s ECGs based on previous experiments in Lu et al. [13], [18].

C. CNN-Transformer Network

In general, Vision Transformer (ViT) splits a raw image into patches. Instead, the proposed CNN-Transformer hybrid model splits a feature map from CNN. The CNN encoder extracts middle-level features from a logarithmic spectrogram. These fixed-size split feature patches are then linearly embedded. Next, position embeddings are added to these patch embeddings to keep positional information.

1) CNN Encoder: We employ three 2D convolutional blocks to extract mid-level features. A logarithmic spectrogram is input to the convolutional blocks. These convolutional blocks extract rich local spatial features in each 2D spectrogram. The architecture of each block was inspired by Lu et al. [13] and Zihlmann et al. [23]. Each convolutional block contains the 2D convolutional layers (3×3 kernel size), exponential linear unit (ELU) and 2D batch normalization. The second convolutional block is followed by a 2D max pooling layer (2×2 window).

2) Transformer Encoder: Given a feature map - the output of the three 2D CNN blocks - $\mathbf{x} \in \mathbb{R}^{W \times H \times C}$, where the C

denotes the number of channels, and W and H are the width and height of the feature map. We first split the \mathbf{x} into flattened non-overlapping patches $\tilde{\mathbf{x}}_p \in \mathbb{R}^{N \times (P^2 \times C)}$, where the N is the total number of the patches ($N = \frac{H}{P} \times \frac{W}{P}$) and the P is the patch size. Then we convert these patches into a D -dimensional embedding space with a trainable linear projection. We concatenate position embeddings and patched embeddings for keeping the spatial information of these extracted patches, which can be described as follows:

$$\mathbf{m}_0 = [\tilde{\mathbf{x}}_p^1 \mathbf{E}; \tilde{\mathbf{x}}_p^2 \mathbf{E}; \dots; \tilde{\mathbf{x}}_p^N \mathbf{E}] + \mathbf{E}_{pos}, \quad (2)$$

where $\mathbf{E} \in \mathbb{R}^{(P^2 \times C) \times D}$ represents the projected patch embedding, $\mathbf{E}_{pos} \in \mathbb{R}^{N \times D}$ stands for the learnable position embedding.

After the embeddings, we employ L transformer layers. In each transformer layer [29], [31], there are three main components: Multi-head Self-Attention (MSA), Multi-Layer Perceptron (MLP) and Layer Normalization (LN). The output of the l -th layer is as follows:

$$\mathbf{m}'_l = MSA(LN(\mathbf{m}_{l-1})) + \mathbf{m}_{l-1}, l = 1, \dots, L, \quad (3)$$

$$\mathbf{m}_l = MLP(LN(\mathbf{m}'_l)) + \mathbf{m}'_l, l = 1, \dots, L. \quad (4)$$

MSA: The inputs $\mathbf{m} \in \mathbb{R}^{n \times d}$ are transformed into three vectors: queries $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$, keys $\mathbf{K} \in \mathbb{R}^{n \times d_k}$ and values $\mathbf{V} \in \mathbb{R}^{n \times d_v}$, where d_k are the dimensions of the queries and keys, and d_v are the dimensions of the values. The scaled dot-product attention can be described as [29]

$$Att(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V}, \quad (5)$$

where $\frac{1}{\sqrt{d_k}}$ is a scale factor that leads to stable gradients by avoiding the softmax function, which falls into regions, resulting in exceedingly small gradients.

The MSA is a core module of the transformer, which consists of n parallel self-attention (SA) heads. It splits the \mathbf{Q} , \mathbf{K} and \mathbf{V} into different subspaces and performs the scaled dot-product attention function in parallel. Next, the outputs of each head are concatenated and produce a final output of the MSA via a linear projection. The formula is as follows

$$MSA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(Head_1, \dots, Head_n) W^o, \quad (6)$$

$$Head_i = Att(\mathbf{Q}W_i^Q, \mathbf{K}W_i^K, \mathbf{V}W_i^V), \quad (7)$$

where W^o denotes the multi-headed trainable parameter weights.

MLP: The MLP can be obtained as

$$MLP(X) = FC(\sigma(FC(X))) \quad (8)$$

where the FC represents a fully-connected layer, $\sigma(\cdot)$ means an activation function GELU [46].

LN: Layer normalization [47] improves the stability of hidden state dynamics within the training network and enables faster training time and convergence. The formula is as follows

$$LN(x) = \gamma \circ \frac{x - \mu}{\varrho} + \beta, \quad (9)$$

where γ and β are learnable parameters, \circ represents the element-wise dot, μ and ϱ are the mean and standard deviation of the elements in x .

IV. EXPERIMENTS

A. ECG Acquisition for Tetanus Patients

The tetanus data collection has received the approval from both the Ethics Committee of the Hospital for Tropical Diseases and the Oxford Tropical Research Ethics Committee. This dataset is collected from 110 patients at the Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam. Recently, this tetanus dataset has been published [11].

To obtain ECG data from tetanus patients, we chose the low-cost wearable monitor ePatch (ePatch V.1.0, BioTelemetry, USA) [48] (see Fig. 1). The lightweight ePatch¹ was stuck firmly to the patient's chest skin. The ePatch records ECG in two channels with a sampling rate of 256 Hz. Channel 1 and channel 2 of the ePatch do not relate to lead 1 and lead 2 in conventional ECG from the bedside monitor. The continuous ECG was stored in the device and exported after the recording period. Tetanus patients ≥ 16 years old, admitted to the ICU at the Hospital for Tropical Diseases Ho Chi Minh City, were enrolled for the vital sign monitoring data collection. The first 24-hour ECG data were recorded on this 1st day at ICU. 24-hour ECG recordings were taken on the 1st and 5th day of hospitalization. ECG signals from channel 1 of ePatch were used in our experiment. In order to get stable signals, the first and last five minutes of each ECG recording were trimmed [11].

B. Implementation Details

1) Pre-Processing.: The dataset consists of 178 time series ECG waveform example files from 110 patients on days 1 and 5. We split our data into the Training/Validation/Test datasets with a 141/19/18 ratio. The same patient data are not in Training/Validation/Test datasets at the same time. The time series ECG waveform is divided into a sequence of ECG samples without overlapped windows. We set the duration of the window length as 60 seconds. We choose 30 60-second ECG samples from each ECG example file. There are 4230 (141 * 30) ECG log spectrograms in the training set, including 2370 samples of mild disease and 1860 samples of severe disease; 540 (18 * 30) ECG log spectrograms in the validation set, including 270 samples of mild disease and 270 samples of severe disease; 570 (19 * 30) ECG log spectrograms in the test set, including 360 samples of the mild disease and 210 samples of severe disease (see Table I). The mild and severe tetanus are labelled by clinician at the Hospital for Tropical Diseases.

First, we remove the noise from these split ECG samples. The implementation is performed utilizing the SciPy package [45]. Next, spectrograms are computed by `scipy.signal.spectrogram` in SciPy [45]. We choose the Tukey window width to be 25% of a window's length overlap. We set the `nperseg` - length of each segment - as 64, and the `noverlap` - numbers of points to overlap

¹ePatch. <https://www.cardiologic.co.uk/epatch-2-0sensor>

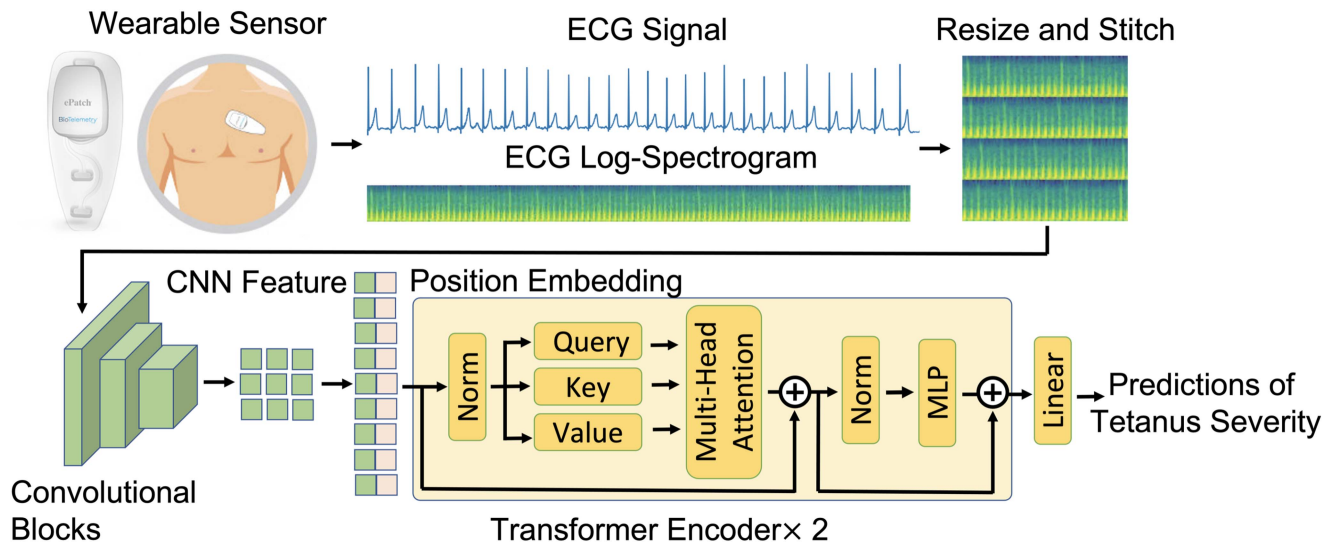


Fig. 1. Overview of the proposed framework for the classification of tetanus severity. The raw ECG data is collected by an ePatch wearable sensor. The resized and stitched 60 s window length Log-Spectrogram of raw ECG data is the input of the proposed method called 2D-CNN-Transformer. The output of the proposed method is the label classification, label 0 - mild tetanus and label 1 - severe tetanus.

TABLE I

DEFINITION OF TRAIN-VALID-TEST SPLIT FOR THE TETANUS DATASET

Data Set	30 ECG samples from each ECG example		
	Total Number	Mild Tetanus	Severe Tetanus
Training	4230 (141*30)	2370	1860
Validation	540 (18*30)	270	270
Test	570 (19*30)	360	210

TABLE II

EMPLOYED PARAMETERS OF THE TRANSFORMER ENCODER IN THE PROPOSED 2D-CNN-TRANSFORMER

Parameters		
img_size	112	the size (resolution) of each image
in_chans	128	the number of input channels
patch_size	8	the size (resolution) of each patch
n_classes	2	the number of classes
embed_dim	386	the embedding dimension
depth	2	the number of the transformer block
n_heads	2	the number of the heads
qkv_bias	True	the bias of the queries, keys and values
mlp_ratio	4	the MLP ratio

between segments - as 32. There are $15360 = 256 \text{ Hz} * 60 \text{ s}$ sampling points in a window of length which are used to compute a spectrogram; these are based on 60 seconds at the sampling rate 256 Hz of the ECG data. We then apply normalization and logarithmic scale to the spectrogram. The spectrogram is saved as a PNG format image with the default 'viridis' colourmap. Finally, the rectangular picture of the spectrogram (479×33 pixels of the Log-Spectrograms on every 60 seconds of ECG) is ready for the proposed deep learning approach.

2) Experimental Setup: Based on experiments, the Transformer Encoder with the selected hyperparameters of the proposed method achieves optimal results (see Table II). The model is trained over 100 epochs using the Adam optimizer

with a learning rate 0.001 and a batch size of 32. We choose `torch.nn.BCEWithLogitsLoss` for the loss function. The proposed network was implemented using Python 3.7 with PyTorch. Experiments are run with computational hardware NVidia GeForce GTX 1080 Ti 10 GB, NVidia GeForce RTX 3060 12 GB and NVIDIA RTX A6000 48 GB.

C. Baseline Methods

In our work, we compare the proposed method - 2D-CNN - Transformer - with four different baseline methods. The baseline methods include three 2D deep learning methods (2D-CNN, 2D-CNN + Dual Attention, 2D-CNN + Channel-wise Attention) and the 1D deep learning method 1D-CNN.

D. Evaluation Metrics

We choose widely used metrics to evaluate the performance of the binary classification, including F1-score, precision, recall, specificity and accuracy [18].

F1-score is the harmonic mean of precision and recall. The formula is as follows

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \quad (10)$$

Precision evaluates how precisely a method predicts the positive labels. The formula is as follows

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (11)$$

Recall measures the percentage of true positives that a method correctly detects. The formula is as follows

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (12)$$

TABLE III
ABLATION STUDIES OF THE PROPOSED METHOD - 2D-CNN-TRANSFORMER - USING RESIZED AND STITCHED 60 S WINDOW LENGTH LOG-SPECTROGRAMS AS INPUT

Method	The resized and stitched 60 second window length Log-Spectrogram					
	F1 score	precision	recall	specificity	accuracy	AUC
2D-CNN part only	0.55±0.12	0.57±0.13	0.63±0.28	0.62±0.32	0.64±0.12	0.65±0.08
ViT	0.76±0.07	0.94±0.05	0.69±0.13	0.97±0.03	0.85±0.03	0.81±0.05
Proposed 2D-CNN-Transformer/8	0.82±0.03	0.94±0.03	0.73±0.07	0.97±0.02	0.88±0.01	0.85 ±0.03

The results are presented as mean ± standard deviation. The best performance is indicated in bold.

TABLE IV
QUANTITATIVE COMPARISON ON PATCH SIZE OF THE TRANSFORMER ENCODER OF THE PROPOSED METHOD - 2D-CNN -TRANSFORMER

Patch Size of the Transformer	The proposed method					
	F1 score	precision	recall	specificity	accuracy	AUC
Patch 16 * 16	0.81±0.05	0.90±0.06	0.74±0.11	0.95±0.04	0.87±0.02	0.85±0.04
Patch 14 * 14	0.71±0.02	0.89±0.07	0.60±0.05	0.95±0.04	0.82±0.01	0.77±0.01
Patch 12 * 12	0.77±0.06	0.86±0.06	0.71±0.10	0.93±0.04	0.85±0.03	0.82±0.04
Patch 10 * 10	0.82±0.03	0.90±0.03	0.76±0.05	0.95±0.02	0.88±0.02	0.86±0.02
Patch 8 * 8	0.82±0.03	0.94±0.03	0.73±0.07	0.97±0.02	0.88±0.01	0.85 ±0.03
Patch 6 * 6	0.70±0.07	0.95±0.02	0.56±0.09	0.98±0.01	0.83±0.03	0.77±0.04
Patch 4 * 4	0.72±0.09	0.94±0.06	0.59±0.13	0.97±0.03	0.83±0.04	0.78±0.05

The results are presented as mean ± standard deviation. The best performance is indicated in bold.

Specificity evaluates the percentage of true negative classification that a method correctly detects. The formula is as follows

$$\text{Specificity} = \frac{TN}{TN + FP}. \quad (13)$$

Accuracy measures the total number of classifications that a method gets correctly generates. The formula is as follows

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (14)$$

Here the terms TP (true positive) and TN (true negative) represent accurately predicted numbers of severe tetanus and mild tetanus, respectively. The FP (false positive) means the mild tetanus has been incorrectly identified as severe tetanus, whilst the FN (false negative) means severe tetanus has been incorrectly identified as mild tetanus.

We also use the area under the curve (AUC) metric [49]. The higher the AUC, the better the proposed method distinguishes severe tetanus from mild tetanus. We run each model 5 times and calculate the mean and the standard deviation of the performance metrics on the test dataset.

V. RESULTS AND DISCUSSION

In this section, we evaluate the proposed 2D-CNN-Transformer method and show how it works. Firstly, we measure the performance of the two main components of the 2D-CNN-Transformer. We discuss the patch size of the transformer, and two types of resized Log-Spectrograms. Then, we compare the proposed 2D-CNN-Transformer to the 1D-CNN; vanilla 2D-CNN, 2D-CNN + Dual Attention, and 2D-CNN + Channel-wise Attention with two types of input images [13]. In addition, we compare the proposed method to the traditional machine learning method of Random Forest. In our experiments, we run each model five times with the same split training / validation / test datasets. We perform the splitting of the dataset into training,

validation and test based on unique ECG samples. After splitting, we apply windowing on ECG time series to split each signal into 60 s time series. Therefore, we made sure that ECG samples in each split of the dataset are unique.

A. Ablation Study

We investigate the effects of two main components - CNN encoder, transformer encoder - within our proposed 2D-CNN-Transformer. As shown in Table III, The proposed 2D-CNN-Transformer clearly outperforms a pure ViT model (vision transformer) on the given task. We also investigate the optimal batch size of the transformer encoder to generate image patches.

1) Patch Size: A patch sequence represents the feature map obtained from the CNN encoder. The different path size influences the performance of the proposed method. Table IV reports the performance for the proposed 2D-CNN-Transformer method using different patch sizes. We observe that the patch size is an important factor for the tetanus severity prediction. From a patch size 4×4 to 16×16 , the path size 8×8 and 10×10 achieve the optimal performance. The 2D-CNN-Transformer/8 represents the proposed method using the path size 8×8 . The 2D-CNN-Transformer/10 represents the proposed method using the path size 10×10 . Fig. 2 shows the examples of Grad-CAM visual explanations of the features for the label 1 - severe tetanus - in the 2nd transformer layer of the proposed method.

2) Resized and Stitched Spectrograms: We have explored different types of resized spectrograms as inputs of the proposed 2D-CNN-Transformer. As shown in Fig. 3, the resized spectrograms in (c) failed and are not suitable for the proposed method. Because too much resize into 224×224 pixels makes the image information loss. Also, Fig. 3(a) can be used in the proposed 2D-CNN-Transformer. Based on experiments, we find the original Log-Spectrograms only have 33 pixels in the width. The short width does not perform well in the patch embeddings.

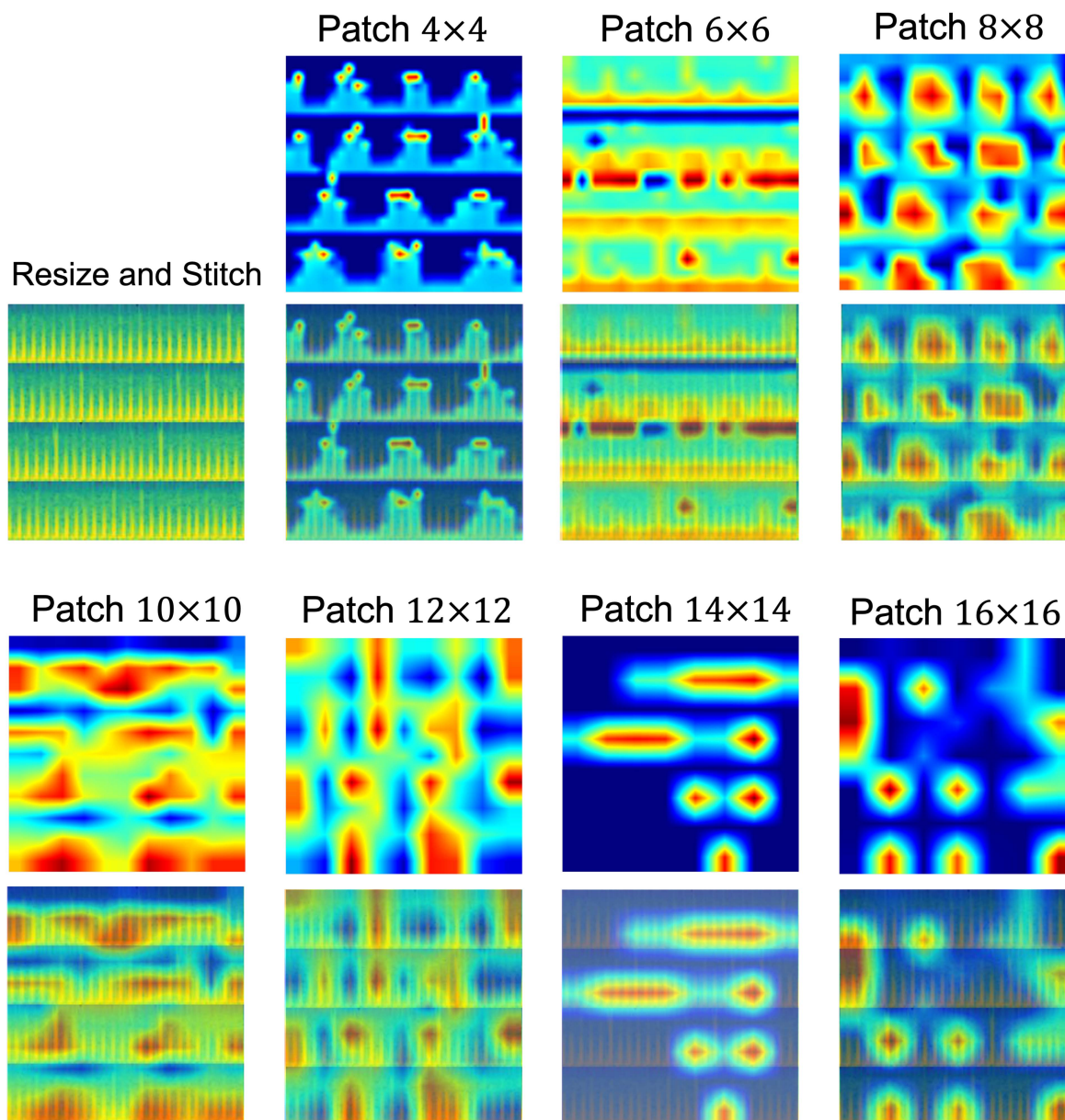


Fig. 2. Examples of visual explanations of the features in the 2nd transformer layer of the proposed method 2D-CNN-Transformer. The resized and stitched 60 s window length Log-Spectrogram of raw ECG data is the input of the proposed method. The performance of the proposed method using different patch sizes from 16×16 to 4×4 in the transformer encoder are compared.

However, the information of the image still maintains well in Fig. 3(b), which is the best option for the proposed method.

3) CNN Blocks Selection: The exploration of the number of the CNN blocks in the CNN encoder is meaningful. Because the low-, mid- and high-level local features effect the performance of the transformer encoder. We do experiments on two layers, three layers and four layers of the CNN encoder of the proposed 2D-CNN-Transformer. At the end, we find that three layers of the CNN encoder achieve the better performance. We also consider adding channel-wise into each CNN block. From our experiments, the channel-wise layers lead to an unstable CNN-Transformer model which has the gradients explosion. Fig. 4 shows the examples of Grad-CAM visual

explanations of the features for the label 1 - severe tetanus - in the different blocks of the proposed method. The resized and stitched Log-Spectrograms are the inputs of the proposed method. The red colours emphasise the most important areas the 2D-CNN-Transformer focuses on for classification.

B. Comparisons

We compare the proposed 2D-CNN-Transformer with four different deep learning methods. Based on the experimental results in Table V, the image-based method 2D-CNN-Transformer boosts the performance of diagnosing tetanus and outperforms other methods.

TABLE V

QUANTITATIVE COMPARISON ON THE PROPOSED METHOD - 2D-CNN-TRANSFORMER/8 USING RESIZED AND STITCHED 60 s WINDOW LENGTH LOG-SPECTROGRAMS AS INPUT AND THE BASELINE METHODS USING ORIGINAL 60 s WINDOW LENGTH ECG AS INPUT

Method	60 second window length Log-Spectrogram					
	F1 score	precision	recall	specificity	accuracy	AUC
2D-CNN	0.61±0.14	0.68±0.07	0.57±0.19	0.85±0.02	0.75±0.07	0.72±0.09
2D-CNN + Dual Attention	0.65±0.19	0.71±0.17	0.61±0.21	0.86±0.09	0.76±0.11	0.74±0.13
2D-CNN + Channel-wise Attention	0.79±0.03	0.78±0.08	0.82±0.05	0.85±0.08	0.84±0.04	0.84±0.03
Proposed 2D-CNN-Transformer/8	0.82±0.03	0.94±0.03	0.73±0.07	0.97±0.02	0.88±0.01	0.85±0.03

Method	No time series Images					
	F1 score	precision	recall	specificity	accuracy	AUC
1D-CNN	0.65±0.14	0.61±0.05	0.77±0.25	0.70±0.13	0.73±0.05	0.74±0.08

The results are presented as mean ± standard deviation. The best performance is indicated in bold.

TABLE VI

QUANTITATIVE COMPARISON OF THE PROPOSED METHOD (2D-CNN-TRANSFORMER/8) AND THE BASELINE METHODS (TRADITIONAL MACHINE LEARNING), USING ORIGINAL 60 s WINDOW LENGTH ECG AS INPUT. THE RESULTS ARE PRESENTED AS MEAN ± STANDARD DEVIATION. THE BEST PERFORMANCE IS INDICATED IN BOLD

Method	60 second window length Log-Spectrogram					
	F1 score	precision	recall	specificity	accuracy	AUC
Proposed 2D-CNN-Transformer/8	0.82±0.03	0.94±0.03	0.73±0.07	0.97±0.02	0.88±0.01	0.85±0.03

Method	No time series Images					
	F1 score	precision	recall	specificity	accuracy	AUC
Random Forest (HRV time domain features (Set 1))	0.81±0.00	0.77±0.00	0.85±0.01	0.85 ±0.00	0.85±0.00	0.80±0.00
Random Forest (HRV time domain features (Set 2))	0.82±0.00	0.76±0.01	0.89±0.00	0.86±0.00	0.84±0.01	0.86±0.03

60-second window length Log-Spectrogram

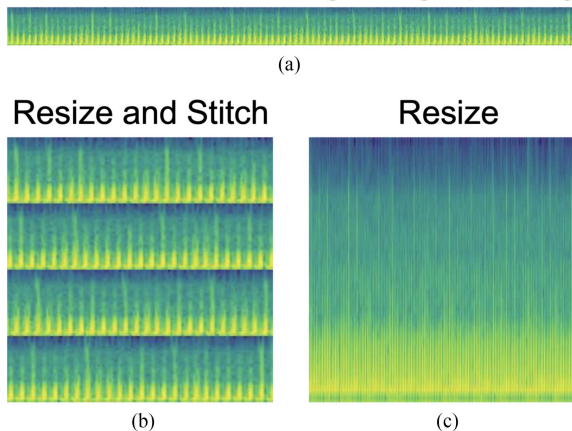


Fig. 3. The different types of Log-Spectrograms: (a) 60-second window length Log-Spectrograms in 479 pixels × 33 pixels; (b) Log-Spectrograms after resize and stitch from (a), in 224 pixels × 224 pixels; (c) Log-Spectrograms after resize from (a), in 224 pixels × 224 pixels.

We also compare the proposed 2D-CNN-Transformer with the traditional machine learning method Random Forest [50], [51], [52] (see Table VI). The extracted features for Random Forest are HRV time domain features (see Table VII). In our work, we detect r peaks of ECG using the open-source packages py-ecg-detectors 1.3.2 [53] and extract features using the open-source packages hrv-analysis 1.0.4 [54]. In Random Forest, the HRV time domain features (Set 2) as input produce better performance than the HRV time domain features (Set 1) as input.

VI. DISCUSSION

In this work, we proposed a novel end-to-end deep learning method - 2D-CNN-Transformer - to classify the severity of tetanus using wearable monitors in a resource-limited setting. The low cost of obtaining ECG outputs makes this method of vital sign data collection affordable in LMICs. We are able to reliably use this low-quality data and classify tetanus symptoms as mild or severe tetanus. Despite this, there are limitations to this method. Because of the small tetanus dataset, we make a classification of tetanus severity using ECG data recorded on day 1 and day 5. In future, we will extend the tetanus dataset. With a larger dataset, we will be able to classify the severity of tetanus on day 5 using the ECG data from only day 1.

ECG wearable devices are indeed increasingly available in LMICs. The device used in this study is the ePatch, which is reusable and is used with a single disposable electrode (\$5 approximately). We appreciate that the potential of such techniques will be greater as more lower cost devices become available, but in comparison to a daily ICU cost for monitoring, or a conventional ICU monitor capable of providing waveform data (\$16,000), the ePatch is low cost.

The proposed method uses patterns of the series imaging - spectrogram - to classify the severity level of tetanus. In our experiments, the spectrogram parameters does not impact performance very much. We will explore time series imaging further in future work, which will aim to find the optimal range of time windows and parameters.

If tetanus patients suffer from heart diseases, their ECG signals are different from the tetanus patients without any heart

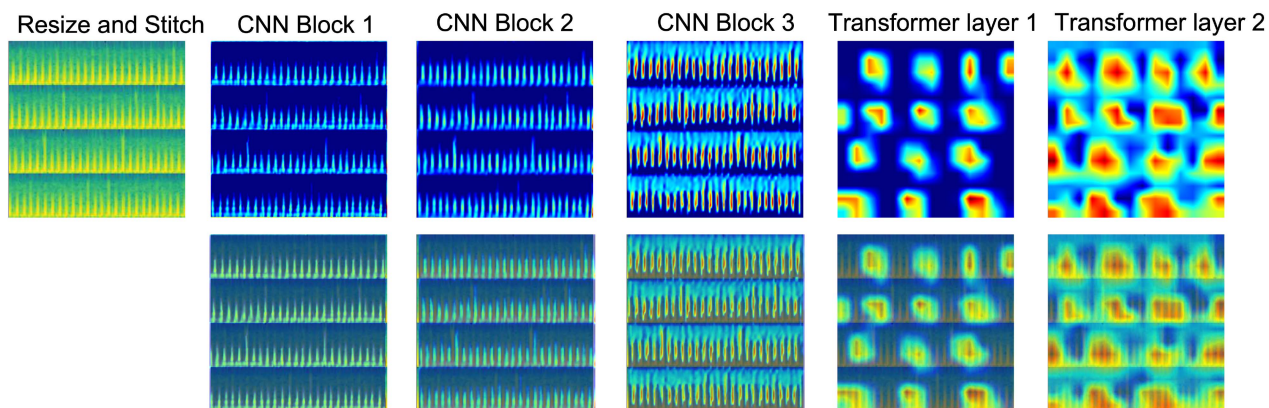


Fig. 4. Example of visual explanations of the features in different layers of the proposed method 2D-CNN-Transformer/8 (patch size 8×8 of the Transformer part). The resized and stitched 60 s window length Log-Spectrogram of raw ECG data is the input of the proposed method.

TABLE VII

LIST OF EXTRACTED HEART RATE VARIABILITY (HRV) FEATURES IN TRADITIONAL MACHINE LEARNING

Parameters	
HRV time domain features (Set1)	
mean_nni	mean of RR-intervals
sdnn	standard deviation of RR-intervals
sdsd	standard deviation of differences between adjacent RR-intervals
rmssd	square root of the mean of the sum of the squares of differences between adjacent NN-intervals
mean_hr	mean Heart Rate
max_hr	max heart rate
min_hr	min heart rate
std_hr	standard deviation of heart rate
HRV time domain features (Set2)	
mean_nni	mean of RR-intervals
sdnn	standard deviation of RR-intervals
sdsd	standard deviation of differences between adjacent RR-intervals
nni_50	number of interval differences of successive RR-intervals greater than 50 ms
pnni_50	proportion derived by dividing nni_50
nni_20	number of interval differences of successive RR-intervals greater than 20 ms
pnni_20	proportion derived by dividing nni_20
rmssd	square root of the mean of the sum of the squares of differences between adjacent NN-intervals
median_nni	median Absolute values of the successive differences between the RR-intervals
range_nni	difference between the maximum and minimum nn_interval.
cvsd	coefficient of variation of successive differences equal to the rmssd divided by mean_nni
cvnni	coefficient of variation equal to the ratio of sdnn divided by mean _{nni} .
mean_hr	mean Heart Rate
max_hr	max heart rate
min_hr	min heart rate
std_hr	standard deviation of heart rate

diseases. This requires further research. Some severity tetanus patients need ventilators which influence HRV. We will consider how various HRV related to the severity tetanus in the future work.

In our experiments, we compared the proposed 2D-CNN-Transformer with the Random Forest model. We applied the Random Forest model with two feature sets containing HRV time domain features. The first feature set (Set 1) contains 8 standard HRV time domain features while the second feature set (Set 2) contains 8 additional time domain features. We found that Random Forest with enough manually selected features can be comparable with the proposed 2D-CNN-Transformer model in classifying the severity of Tetanus disease. Moreover, Random Forest with these two HRV time domain features sets have higher recall values, which indicates that Random forest yields a better prediction of severe tetanus. Here, recall measures the ability of the model in correctly predicting the percentage of the severe tetanus (TP) condition.

VII. CONCLUSION

The proposed 2D-CNN-Transformer method captures both the local spatial information from the CNN features and the global context information from Transformers. Experimental results demonstrate that the proposed deep learning method outperforms other state-of-the-art methods in tetanus classification. The proposed deep learning framework can help clinical care decision-making and assist in the allocation of limited healthcare resources in LMICs, and could be applied to similar infectious diseases such as sepsis. In future work, we will integrate multi-modal physiological data with the current work to further improve tetanus severity classification.

ACKNOWLEDGMENT

The authors wish to thank the patients and staff in the ICU of the Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam.

APPENDIX

VIETNAM ICU TRANSLATIONAL APPLICATIONS LABORATORY (VITAL) INVESTIGATORS

OUCRU inclusive authorship list in Vietnam (alphabetic order by surname): Dang Phuong Thao, Dang Trung Kien, Doan Bui Xuan Thy, Dong Huu Khanh Trinh, Du Hong Duc, Ronald Geskus, Ho Bich Hai, Ho Quang Chanh, Ho Van Hien, Huynh

Trung Trieu, Evelyne Kestelyn, Lam Minh Yen, Le Dinh Van Khoa, Le Thanh Phuong, Le Thuy Thuy Khanh, Luu Hoai Bao Tran, Luu Phuoc An, Angela McBride, Nguyen Lam Vuong, Nguyen Quang Huy, Nguyen Than Ha Quyen, Nguyen Thanh Ngoc, Nguyen Thi Giang, Nguyen Thi Diem Trinh, Nguyen Thi Le Thanh, Nguyen Thi Phuong Dung, Nguyen Thi Phuong Thao, Ninh Thi Thanh Van, Pham Tieu Kieu, Phan Nguyen Quoc Khanh, Phung Khanh Lam, Phung Tran Huy Nhat, Guy Thwaites, Louise Thwaites, Tran Minh Duc, Trinh Manh Hung, Hugo Turner, Jennifer Ilo Van Nuil, Vo Tan Hoang, Vu Ngo Thanh Huyen, Sophie Yacoub

Hospital for Tropical Diseases, Ho Chi Minh City (alphabetic order by surname): Cao Thi Tam, Duong Bich Thuy, Ha Thi Hai Duong, Ho Dang Trung Nghia, Le Buu Chau, Le Mau Toan, Le Ngoc Minh Thu, Le Thi Mai Thao, Luong Thi Hue Tai, Nguyen Hoan Phu, Nguyen Quoc Viet, Nguyen Thanh Dung, Nguyen Thanh Nguyen, Nguyen Thanh Phong, Nguyen Thi Kim Anh, Nguyen Van Hao, Nguyen Van Thanh Duoc, Pham Kieu Nguyet Oanh, Phan Thi Hong Van, Phan Tu Qui, Phan Vinh Tho, Truong Thi Phuong Thao

University of Oxford (alphabetic order by surname): Natasha Ali, David Clifton, Mike English, Shadi Ghiasi, Heloise Greff, Jannis Hagenah, Ping Lu, Jacob McKnight, Chris Paton, Tingting Zhu

Imperial College London (alphabetic order by surname): Pantelis Georgiou, Bernard Hernandez Perez, Kerri Hill-Cawthorne, Alison Holmes, Stefan Karolcik, Damien Ming, Nicolas Moser, Jesus Rodriguez Manzano

King's College London (alphabetic order by surname): Liane Canas, Alberto Gomez, Hamideh Kerdegari, Andrew King, Marc Modat, Reza Razavi, Miguel Xochicale

University of Ulm (alphabetic order by surname): Walter Karlen

The University of Melbourne (alphabetic order by surname): Linda Denehy, Thomas Rollinson

Mahidol Oxford Tropical Medicine Research Unit (MORU) (alphabetic order by surname): Luigi Pisani, Marcus Schultz

REFERENCES

- [1] C. Thwaites, "Botulism and tetanus," *Medicine*, vol. 45, no. 12, pp. 739–742, 2017.
- [2] Disease factsheet about tetanus. 2021. [Online]. Available: <https://www.ecdc.europa.eu/en/tetanus/facts>
- [3] D. B. Thuy et al., "Tetanus in southern Vietnam: Current situation," *Amer. J. Trop. Med. Hyg.*, vol. 96, no. 1, 2017, Art. no. 93.
- [4] C. Thwaites et al., "Predicting the clinical outcome of tetanus: The tetanus severity score," *Trop. Med. Int. Health*, vol. 11, no. 3, pp. 279–287, 2006.
- [5] L. M. Yen and C. L. Thwaites, "Tetanus," *Lancet*, vol. 393, no. 10181, pp. 1657–1668, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0140673618331313>
- [6] H. H. Kyu et al., "Mortality from tetanus between 1990 and 2015: Findings from the global burden of disease study 2015," *BMC Public Health*, vol. 17, no. 1, pp. 1–17, 2017.
- [7] The importance of diagnostic tests in fighting infectious diseases. 2021. [Online]. Available: <https://www.lifechanginginnovation.org/medtech-facts/importance-diagnostic-tests-fighting-infectious-diseases.html>
- [8] M. Afshar et al., "Narrative review: Tetanus—A health threat after natural disasters in developing countries," *Ann. Intern. Med.*, vol. 154, no. 5, pp. 329–335, 2011.
- [9] T. M. Hung et al., "Direct medical costs of tetanus, dengue, and sepsis patients in an intensive care unit in Vietnam," *Front. Public Health*, vol. 10, 2022, Art. no. 893200.
- [10] T. M. Hung et al., "The estimates of the health and economic burden of dengue in Vietnam," *Trends Parasitol.*, vol. 34, no. 10, pp. 904–918, 2018.
- [11] H. M. T. Van et al., "Vital sign monitoring using wearable devices in a vietnamese intensive care unit," *BMJ Innovations*, vol. 7, no. Suppl 1, pp. s1–s5, 2021.
- [12] M. Joshi et al., "Wearable sensors to improve detection of patient deterioration," *Expert Rev. Med. Devices*, vol. 16, no. 2, pp. 145–154, 2019.
- [13] P. Lu et al., "Classification of tetanus severity in intensive-care settings for low-income countries using wearable sensing," *Sensors*, vol. 22, no. 17, 2022, Art. no. 6554.
- [14] H. T. H. Duong et al., "Heart rate variability as an indicator of autonomic nervous system disturbance in tetanus," *Amer. J. Trop. Med. Hyg.*, vol. 102, no. 2, 2020, Art. no. 403.
- [15] I. Cygankiewicz and W. Zareba, "Heart rate variability," *Handbook Clin. Neurol.*, vol. 117, pp. 379–393, 2013.
- [16] Electrophysiology, Task Force of the European Society of Cardiology the North American Society of Pacing, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.
- [17] M. Bolanos, H. Nazeran, and E. Haltiwanger, "Comparison of heart rate variability signal features derived from electrocardiography and photoplethysmography in healthy individuals," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2006, pp. 4289–4294.
- [18] G. A. Tadesse et al., "Multi-modal diagnosis of infectious diseases in the developing world," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 7, pp. 2131–2141, Jul. 2020.
- [19] S. Ghiasi et al., "Sepsis mortality prediction using wearable monitoring in low–middle income countries," *Sensors*, vol. 22, no. 10, 2022, Art. no. 3866. [Online]. Available: <https://www.mdpi.com/1424-8220/22/10/3866>
- [20] G. A. Tadesse et al., "Severity detection tool for patients with infectious disease," *Healthcare Technol. Lett.*, vol. 7, no. 2, pp. 45–50, 2020.
- [21] D. Kiyasseh et al., "Plethaugment: GAN-based PPG augmentation for medical diagnosis in low-resource settings," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 11, pp. 3226–3235, Nov. 2020.
- [22] A. Ullah et al., "Classification of arrhythmia by using deep learning with 2-D ECG spectral image representation," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1685.
- [23] M. Zihlmann, D. Perekrestenko, and M. Tschannen, "Convolutional recurrent neural networks for electrocardiogram classification," in *Proc. Comput. Cardiol.*, 2017, pp. 1–4.
- [24] A. Diker et al., "A novel application based on spectrogram and convolutional neural network for eeg classification," in *Proc. 1st Int. Informat. Softw. Eng. Conf.*, 2019, pp. 1–6.
- [25] G. Liu et al., "ECG quality assessment based on hand-crafted statistics and deep-learned s-transform spectrogram features," *Comput. Methods Programs Biomed.*, vol. 208, 2021, Art. no. 106269.
- [26] B. Tutuko et al., "AFibNet: An implementation of atrial fibrillation detection with convolutional neural network," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, pp. 1–17, 2021.
- [27] S. Kiranyaz et al., "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, 2021, Art. no. 107398.
- [28] Y. Wu et al., "A comparison of 1-D and 2-D deep convolutional neural networks in ECG classification," 2018, *arXiv:1810.07088*.
- [29] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
- [30] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, doi: [10.1109/TPAMI.2022.3152247](https://doi.org/10.1109/TPAMI.2022.3152247).
- [31] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [32] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
- [33] G. Hinton et al., "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
- [34] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
- [35] K. Han et al., "Transformer in transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 15908–15919, 2021.
- [36] A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 574–584.
- [37] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

- [38] C. Zhao et al., "Visual-assisted probe movement guidance for obstetric ultrasound scanning using landmark retrieval," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2021, pp. 670–679.
- [39] J. Zhang et al., "A CNN-transformer hybrid approach for decoding visual neural activity into text," *Comput. Methods Programs Biomed.*, vol. 214, 2022, Art. no. 106586.
- [40] H. Wu et al., "FAT-Net: Feature adaptive transformers for automated skin lesion segmentation," *Med. Image Anal.*, vol. 76, 2022, Art. no. 102327.
- [41] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," 2021, *arXiv:2104.01778*.
- [42] S. Park, Y. Jeong, and T. Lee, "Many-to-many audio spectrogram transformer: Transformer for sound event localization and detection," *DCASE*, pp. 105–109, 2021.
- [43] Q. Kong et al., "Sound event detection of weakly labelled data with CNN-transformer and automatic threshold optimization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2450–2460, Aug. 2020.
- [44] Y.-H. Byeon and K.-C. Kwak, "Pre-configured deep convolutional neural networks with various time-frequency representations for biometrics from ECG signals," *Appl. Sci.*, vol. 9, no. 22, 2019, Art. no. 4810.
- [45] P. Virtanen et al., "SciPy 1.0: Fundamental algorithms for scientific computing in python," *Nature Methods*, vol. 17, pp. 261–272, 2020.
- [46] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
- [47] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
- [48] epatch the world's most wearable holter monitor. 2021. [Online]. Available: <https://www.gobio.com/clinical-research/cardiac-safety/epatch/>
- [49] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.
- [50] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [51] P. Lu et al., "Highly accurate facial nerve segmentation refinement from CBCT/CT imaging using a super-resolution classification approach," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 1, pp. 178–188, Jan. 2018.
- [52] P. Lu et al., "Facial nerve image enhancement from CBCT using supervised learning technique," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2015, pp. 2964–2967.
- [53] Seven ECG heartbeat detection algorithms and heartrate variability analysis. 2022. [Online]. Available: <https://www.ecdc.europa.eu/en/tetanus/facts>
- [54] Heart rate variability analysis. 2022. [Online]. Available: <https://pypi.org/project/hrv-analysis/>