# Supplementary Materials for Multimodal Fusion Learning with Dual Attention for Medical Imaging

Joy Dhar[1] Nayyar Zaidi[2] Maryam Haghighat[3] Sudipta Roy[4] Puneet Goyal[1,6]
Azadeh Alavi[5] Vikas Kumar[1]

[1]Indian Institute of Technology Ropar, India [2]Deakin University, Australia [3]Queensland University of Technology, Australia
[4]Jio Institute, India [5] RMIT University, Australia [6]NIMS University Rajasthan, India

## A. Overview

This document provides supplementary material for the dual robust information fusion attention `DRIFA` mechanism integrated within a deep neural network named `DRIFA-Net`. This approach adopts a multimodal fusion learning `(MFL)` strategy and incorporates two attention mechanisms: a multi-branch fusion attention `(MFA)` module to learn enhanced diverse local representations for each modality, and a multimodal information fusion attention `(MIFA)` module to enhance multimodal representations. These modules enhance multimodal representation learning, as evidenced by improved performance in the results.

Section B presents the pseudo-codes in Algorithms 1(a)-1(d), outlining the `DRIFA-Net` approach.

Section C explains the optimization of all learnable weights, including $\omega_d$, $\omega_l$, and $\omega_c$ for the `MFA` module and $\omega_{d_m}$, $\omega_{l_m}$, and $\omega_{c_m}$ for the `MIFA` module, using a back-propagation strategy.

Section D details the datasets used in this study, while Section E presents additional experimental results, including the confusion matrix and training-validation loss and accuracy curves for our model trained on the `HAM10000 (D1)` [9] and `SIPaKMeD (D2)` [8] datasets.

Section F provides a qualitative analysis to evaluate the efficacy of our proposed method using the `Grad-CAM` technique. This visualization highlights the regions of highest importance in the `D1` [9] and `D3 (Nickparvar MRI)` [7] datasets, as shown in Fig. 3. The attention maps validate our model's decisions by emphasizing areas crucial to prediction scores across these datasets in this study.

## B. Brief Explanation for `DRIFA-Net` as Algorithm 1

In this section, we exhibit Algorithm 1, detailing each of the most important modules of `DRIFA-Net`.

- **Algorithm 1(a)**: This algorithm describes the residual robust attention `(RRA)` block, which enables enhanced representation learning. Detailed steps and operations of this block are elaborated in Section 3.2.1 of the main paper.
- **Algorithm 1(b)**: This algorithm outlines the `MFA` module, designed for diverse representation learning of local information. The `MFA` module employs multiple branches of attention mechanisms to capture intricate details across different modalities.
- **Algorithm 1(c)**: This algorithm outlines the `MIFA` module (see Section 3.2.2 of the main paper), which is designed to learn enhanced multimodal representations across various modalities, thereby boosting the performance of the learning network.
- **Algorithm 1(d)**: This algorithm covers the multitask learning `(MTL)` module (ref. Section 3.2.3 of the main paper), facilitating the classification of multiple disease tasks. The `MTL` module leverages shared representations to perform classification tasks across different disease categories efficiently.

Each algorithm is integral to the functioning of `DRIFA-Net`, contributing to its robust performance, feature learning capabilities, and effectiveness of the multimodal medical image classification tasks. Detailed pseudo-code and step-by-step operations for these modules are provided to offer a comprehensive understanding of their implementation and integration within the network.

## C. Optimized Learnable Weights using Back-propagation

In this study, for the `MFA` module, all learnable weights—such as $\omega_d$, $\omega_l$, and $\omega_c$—are used to adjust the importance of each learned information along with each chan-

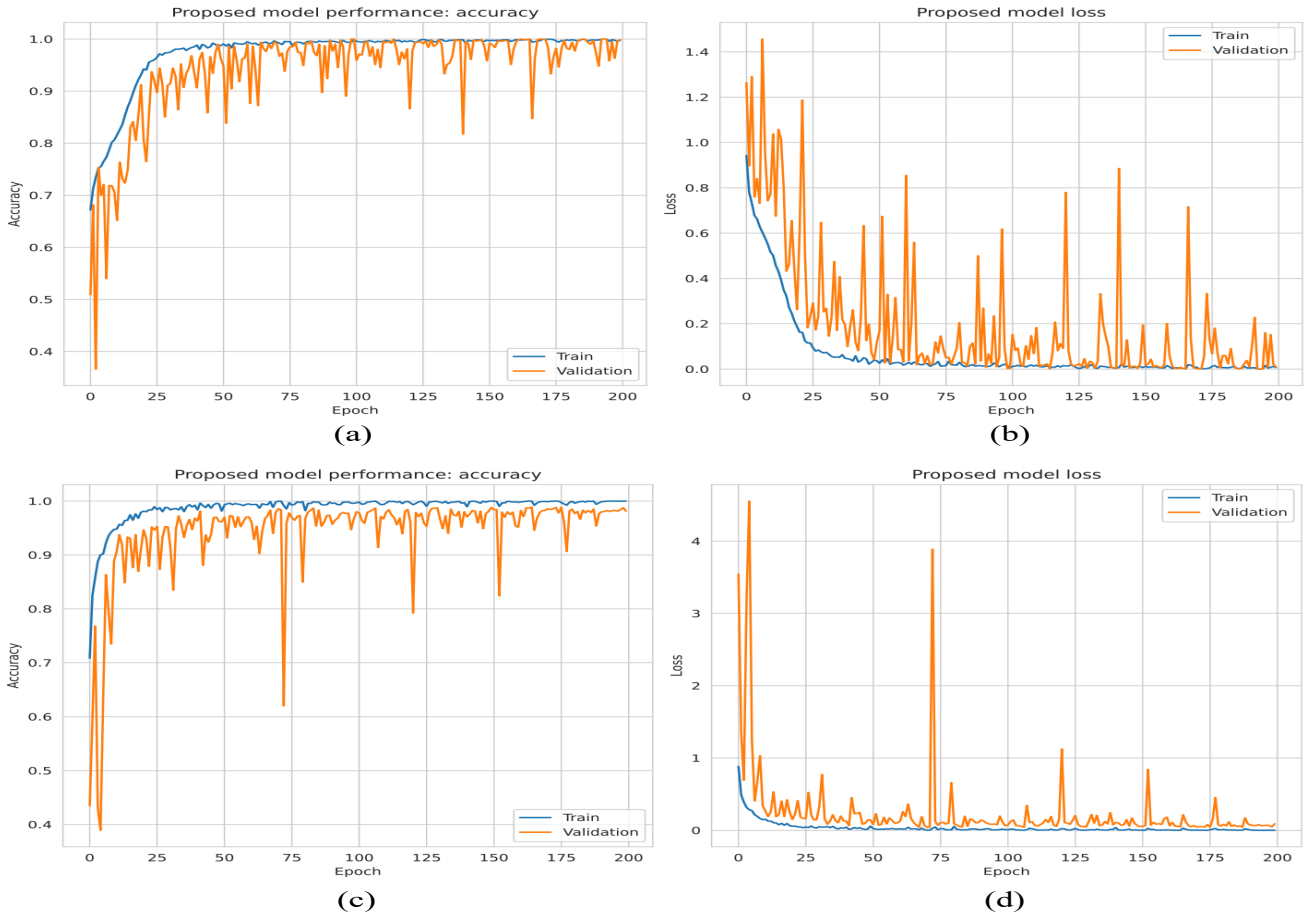


Figure 1. Visualization of training and validation accuracy and loss curves per epoch while training our proposed DRIFA-Net models with the D1 [9] and D2 [8] datasets. (a) and (b) show accuracy and corresponding loss for the D1 dataset [9], while (c) and (d) depict accuracy and corresponding loss for the D2 dataset [8].

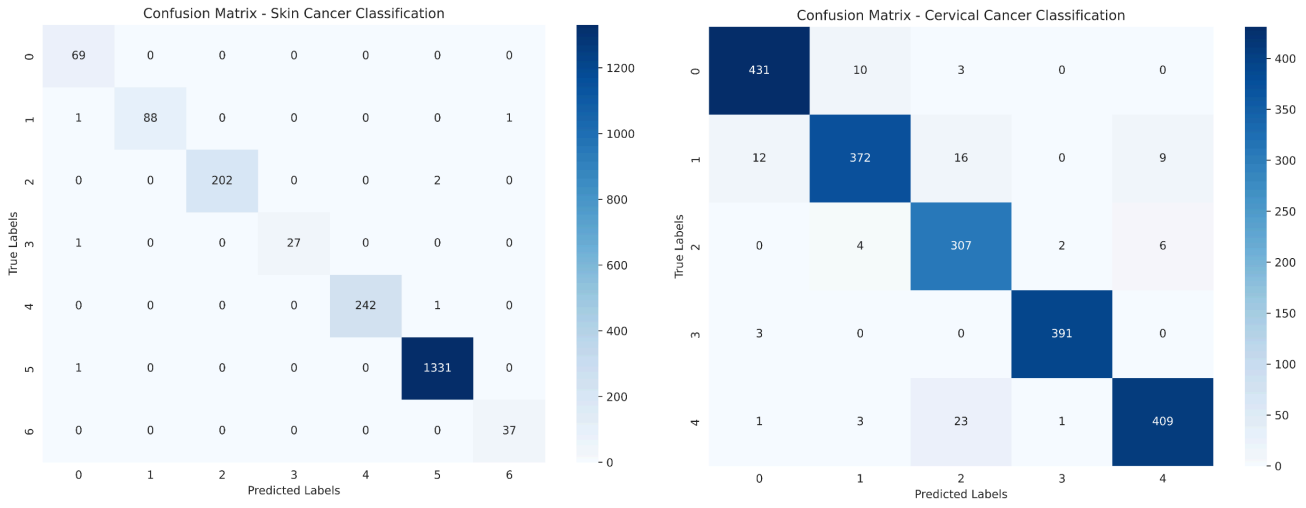


Figure 2. Confusion matrices illustrating the performance of our proposed DRIFA-Net, for (**left**) skin cancer classification in the D1 dataset [9], and (**right**) cervical cancer classification in the D2 dataset [8].

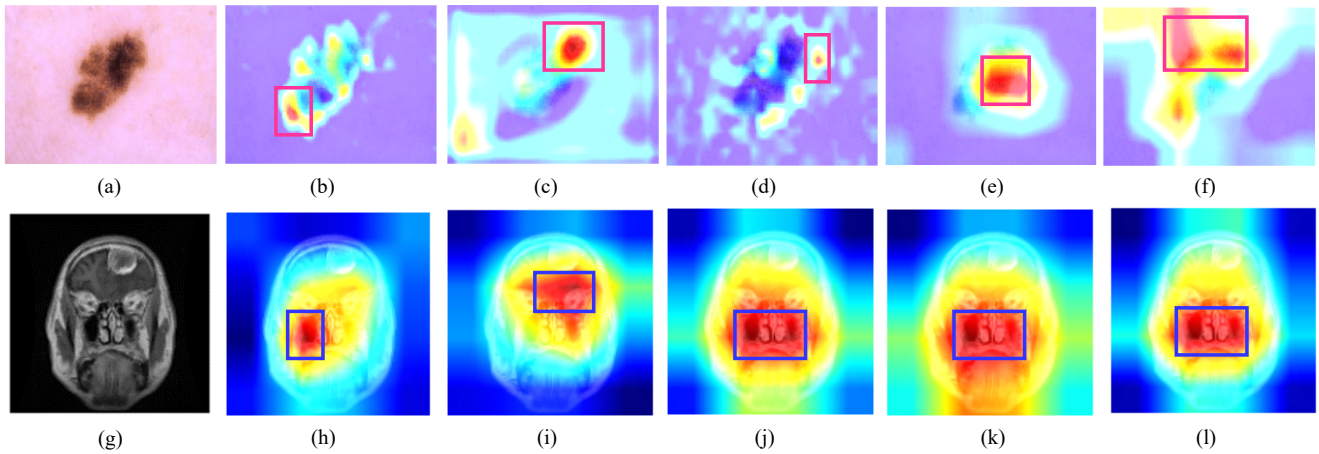


Figure 3. Visualization of important regions highlighted using red or blue rectangular boxes by our proposed DRIFA-Net and four existing methods [2–5] using the Grad-CAM technique on the D1 [9] and D3 [7] datasets. Figure (a) and (g) display the original images, while (b) and (h) present results for Gloria [5], (c) and (i) for MTF with MA [2], (d) and (j) for CAF [4], (e) and (k) for MTTU-Net [3], and (f) and (l) for our DRIFA-Net.

nel during training. This process helps refine both the diverse local information $\hat{d}$ and the channel-wise local information $\hat{l}$. These learnable weights are optimized using a back-propagation strategy.

Specifically, we compute the gradient of the loss, denoted as $\nabla\partial_{MTL}$ (from Eq. 9 of the main paper), with respect to $\omega_d$, $\omega_l$, and $\omega_c$, as specified below:

$$\nabla_{\omega_d}\partial_{MTL} = (\nabla_a\partial_{MTL} \times \nabla_{\omega_d}a) \quad (1)$$

$$\nabla_{\omega_l}\partial_{MTL} = (\nabla_a\partial_{MTL} \times \nabla_{\omega_l}a) \quad (2)$$

where $a$ represents learned attention maps from MFA module, $\nabla\omega_d a$ and $\nabla\omega_l a$ are back-propagation of the gradients. From Eq. 4 of the main paper, we can derive that $\nabla_{\omega_d}a = a \times (1-a)$ and $\nabla_{\omega_l}a = a \times (1-a)$. Now, we can compute the gradient loss with respect to learnable $\omega_c$ parameter as follows.

$$\nabla_{\omega_c}\partial_{MTL} = \zeta \times \nabla_a\partial_{MTL} \times x \times (a \times (1-a)) \quad (3)$$

where $\nabla_{\omega_c}\partial_{MTL}$ is the back-propagation of the gradient, and $\zeta$ is the constant in back-propagation.

Throughout the optimization process, our proposed approach iteratively updates the parameters. $\omega_d$, $\omega_l$, and $\omega_c$ based on their computed gradients, aiming to minimize the overall loss of our proposed DRIFA-Net model. These learnable parameters undergo updates at step $t+1$ for layer $u$ as follows.

$$\begin{aligned}\omega_d^{t+1} &= \omega_d^t - u_r \times \nabla_{\omega_d}\partial_{MTL}^t,\\ \omega_l^{t+1} &= \omega_l^t - u_r \times \nabla_{\omega_l}\partial_{MTL}^t,\\ \omega_c^{t+1} &= \omega_c^t - u_r \times \nabla_{\omega_c}\partial_{MTL}^t\end{aligned} \quad (4)$$

Similarly, all employed learnable weights—such as $\omega_{d_m}$, $\omega_{l_m}$, and $\omega_{c_m}$ for the MIFA module—are optimized using a back-propagation strategy, as outlined in Eqs. 5 and 8 of the main paper, as follows:

$$\begin{aligned}\omega_{d_m}^{t+1} &= \omega_{d_m}^t - u_{m_r} \times \nabla_{\omega_{d_m}}\partial_{MTL}^t,\\ \omega_{l_m}^{t+1} &= \omega_{l_m}^t - u_{m_r} \times \nabla_{\omega_{l_m}}\partial_{MTL}^t,\\ \omega_{c_m}^{t+1} &= \omega_{c_m}^t - u_{m_r} \times \nabla_{\omega_c}\partial_{MTL}^t\end{aligned} \quad (5)$$

where layer $u_m$ represents layers for $m$ heterogeneous modalities.

## D. Dataset

Our experiments utilized five medical imaging datasets: HAM10000 [9], SIPaKMeD [8], Nickparvar MRI [7], IQ-OTHNCCD lung cancer [1], and BraTS2020 [6] (denoted as D1, D2, D3, D4, and D5 respectively). HAM10000 comprises 10,015 images across seven classes: Actinic keratoses and intraepithelial carcinoma/Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc). SIPaKMeD includes 4,049 images distributed over five classes: Dyskeratotic, Koilocytotic, Metaplastic, Parabasal, and Superficial-Intermediate.

**Algorithm 1** DRIFA-Net(X)

**Input:** Input features, $[x_1, x_2, \ldots, x_m] \in X$; where $[x_1, x_2, \ldots, x_m] \in \mathbb{R}^{H \times W \times C}$
**Output:** Enhanced learned multimodal features, $X^S$
Perform multitask operation based on $X^S$
**Procedure:**
*/* For the development of the DRIFA mechanism, integrated with neural networks such as ResNet18, referred to as DRIFA-Net. */*
**if** phase == TMFL **then**
  **for** each modality $m$ **do**
    **for** each filter size in {64, 128, 256, 512} **do**
      **if** filter size == 64 **then**
        $x \leftarrow$ Use a convolution block
      **end if**
      $x' \leftarrow$ Call RRA($x$) block
      **if** filter size in {128, 256} **then**
        $x' \leftarrow$ Use dropout (for no UQ) else use MCD
      **end if**
      $x' \leftarrow$ Call MFA($x'$) module for each $m$
    **end for**
  **end for**
  $X^S \leftarrow$ Call MIFA($[x'_1, x'_2, \ldots, x'_m]$) module for all $m$
  Call MTL($X^S$, $y_t$) module to classify multiple diseases based on each $m$
**else**
  Generate $n$ ensemble learning models
  Perform Monte Carlo equation to estimate the uncertainty of the DRIFA-Net as per equation 10
**end if**

The brain tumor dataset [7] combines images from Figshare, SARTAJ, and BrH35, comprising 7,023 MRI scans categorized into Glioma (324 test and 1,297 training images, labeled as 0), Meningioma (329 test and 1,316 training images, labeled as 1), No Tumor (400 test and 1,600 training images, labeled as 2), and Pituitary (351 test and 1,406 training images, labeled as 3). The IQ-OTHNCCD lung cancer dataset [1] consists of 1,098 images across three classes: normal, benign, and malignant.

**Algorithm 1(a)** RRA(x)

**for** each layer **do**
  $x \leftarrow$ Use a convolution layer
  $x' \leftarrow$ Call MFA() module
  Use skip connection strategy
**end for**
**return** $x'$

Data augmentation techniques, including rotation and transformation, were applied to ensure consistent sample sizes for training across modalities. All images were standardized to $128 \times 128 \times 3$ pixels, with an 80% training, 10% validation, and 10% testing split.

## E. Experimental Results

In this section, we present the generated confusion matrices along with training-validation results to demonstrate the performance of our proposed model during the training, validation, and testing stages. Figs. 1 and 2 illustrate these results for the D1 [9] and D2 [8] datasets, respectively.

## F. Impact of Qualitative Analysis

To facilitate qualitative analysis and compare the effectiveness of our proposed method with other approaches, we show results using Grad-CAM technique. Grad-CAM visualizations highlighted regions of highest importance in the D1 [9] and D3 [7] datasets, as depicted in Fig. 3. These attention maps validate the decisions made by our proposed DRIFA-Net by emphasizing critical areas for prediction scores across the D1 and D3 datasets [7, 9], providing insights into the model's decision-making process.

**Algorithm 1(b)** MFA(x)

**Procedure:** */* For developing Multi-branch fusion attention MFA module */*
**HIFA(x):** */* HIFA module: To learn diverse enhanced local information $\hat{d}$ */*
**for** all $p$ **do**
  $l_p \leftarrow$ Follow Eq. 1 of the main paper;
**end for**
*/* Use hierarchical fusion strategy to fuse all learned local information for each ith index */*
**for** all $l_p$ **do**
  $\hat{d} = f(\forall_{l_p}[\varphi\{\phi(l_p, l_{p+1}), \phi(l_{p+2}, l_{p+3})\}])$;
**end for**
**CLIA(x):** */* CLIA module: To learn channel-wise local information $\hat{l}$ */*
**for** each $q$ **do**
  $\hat{l} \leftarrow$ Follow Eq. 3 of the main paper;
**end for**
*/* Modulation Strategy: To learn local attention map to highlight important features while suppressing less significant ones followed by fusion with sigmoid to learn both learned versions of capturing local information */*
*/* use learnable weights $\omega_d$ and $\omega_l$ to enhance diverse local information */*
$a = \sigma((\hat{d} \times \omega_d) \oplus (\hat{l} \times \omega_l))$
*/* To learn enhanced local representation learning using MFA */*
$x' = x \times a \times \omega_c$
**return** $x'$

**Algorithm 1(c)** `MIFA`$([x'_1, x'_2, \ldots, x'_m])$

**Procedure:** */* For developing Multimodal Information Fusion Attention (`MIFA`) module */*
**`MGIFA`**$([x'_1, x'_2, \ldots, x'_m])$**:** */* `MGIFA` module: To learn diverse enhanced multimodal global information $g'$ (refer to Eq. 6 of the main paper) */*
**for** each pool $\in \{\max, \text{avg}, \min\}$ **do**
    $g' = \phi\left(f_{pool}\left(\sum_{i=1}^{m} G_{pool,i}(X)\right)\right)$
**end for**
**`MLIFA`**$([x'_1, x'_2, \ldots, x'_m])$**:** */* `MLIFA` module: To learn diverse enhanced multimodal local information $l'$ (refer to Eq. 7 fo the main paper) */*
**for** each pool $\in \{\max, \text{avg}, \min\}$ **do**
    $l' = \phi\left(f_{pool}\left(\sum_{i=1}^{m} L_{pool,i}(X)\right)\right)$
**end for**
*/* Modulation Strategy: To learn global-local attention map using modulation strategy to highlight important features while suppressing less significant ones followed by fusion with sigmoid to learn both information */*
*/* use learnable weights $\omega_{d_m}$ and $\omega_{l_m}$ to enhance diverse global-local information */*
$A = \sigma((g' \otimes \omega_{d_m}) + (l' \otimes \omega_{l_m}))$
*/* To learn enhanced multimodal representation learning using `MIFA` module */*
$X^S = X \times A \times \omega_{c_m}$
**return** $X^S$

**Algorithm 1(d)** `MTL`$(X^S, y_t)$

**Procedure:** */* For designing multitask learning */*
**for** each task $t$ **do**
    $\theta(X^S, y_t) = [x_1, \ldots, x_m] \rightarrow [y_1, \ldots, y_t]$, and $\omega_t^m$
    $\partial_{\text{MTL}} = \sum_t \omega_t^m \times \partial_t^m(\theta(X^S, y_t))$
**end for**
**return** $\partial_{\text{MTL}}$

## References


[1] Hamdalla Alyasriy and A Muayed. The iq-othnccd lung cancer dataset. *Mendeley Data*, 1(1):1–13, 2020. 3, 4

[2] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, and D. Yang. A multimodal transformer to fuse images and metadata for skin disease classification. *The Visual Computer*, 39(7):2781–2793, 2023. 3

[3] J. Cheng, J. Liu, H. Kuang, and J. Wang. A fully automated multimodal mri-based multi-task learning for glioma segmentation and idh genotyping. *IEEE Transactions on Medical Imaging*, 41(6):1520–1532, June 2022. 3

[4] X. He, Y. Wang, S. Zhao, and X. Chen. Co-attention fusion network for multimodal skin cancer diagnosis. *Pattern Recognition*, 133:108990, 2023. 3

[5] S. C. Huang, L. Shen, M. P. Lungren, and S. Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021. 3

[6] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 3

[7] Msoud Nickparvar. Brain tumor MRI dataset. Data set, 2021. Accessed on 3rd March. 1, 3, 4

[8] Maria E Plissiti, Panagiotis Dimitrakopoulos, Giorgos Sfikas, Christophoros Nikou, Orestis Krikoni, and Avraam Charchanti. Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3144–3148. IEEE, October 2018. 1, 2, 3, 4

[9] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 1, 2, 3, 4