


MediMLP: Using Grad-CAM to Extract Crucial Variables for Lung Cancer Postoperative Complication Prediction

Tao He, *Student Member, IEEE*, Jixiang Guo, *Member, IEEE*, Nan Chen, Xiuyuan Xu, Zihuai Wang, Kaiyu Fu, Lunxu Liu, and Zhang Yi , *Fellow, IEEE*

Abstract—Lung cancer postoperative complication prediction (PCP) is significant for decreasing the perioperative mortality rate after lung cancer surgery. In this paper we concentrate on two PCP tasks: (1) the binary classification for predicting whether a patient will have postoperative complications; and (2) the three-class multi-label classification for predicting which postoperative complication a patient will experience. Furthermore, an important clinical requirement of PCP is the extraction of crucial variables from electronic medical records. We propose a novel multi-layer perceptron (MLP) model called medical MLP (MediMLP) together with the gradient-weighted class activation mapping (Grad-CAM) algorithm for lung cancer PCP. The proposed MediMLP, which involves one locally connected layer and fully connected layers with a shortcut connection, simultaneously extracts crucial variables and performs PCP tasks. The experimental results indicated that MediMLP outperformed normal MLP on two PCP tasks and had comparable performance with existing feature selection methods. Using MediMLP and further experimental analysis, we found that the variable of “time of indwelling drainage tube” was very relevant to lung cancer postoperative complications.

Index Terms—Feature selection, Lung cancer, MLP, Neural networks, Postoperative complication.

I. INTRODUCTION

LUNG cancer is among the most dangerous cancers. To decrease the mortality rate of lung cancer, many machine learning methods have been used. These methods can be divided into preoperative detection and postoperative prediction. Among them, preoperative detection methods have been widely

developed. Volumetric thoracic computed tomography (CT) is a common diagnostic tool for clinicians. As deep learning and big data have developed, many preoperative detection methods have been proposed based on CT image analysis. Automatic preoperative lung nodule detection in CT images has been widely exploited and helps to locate potential tumors on CT slices. Similar image-based techniques, such as magnetic resonance imaging (MRI) and color doppler (CD) image analysis have also been widely developed. Most of these preoperative detection methods are computer vision methods because the preoperative examination produces many images of thoracic organs or tissues.

Postoperative prediction is very important because the patients might suffer from complications during postoperative period and the lung cancer is easy to relapse after clinical surgery, radiation oncology or chemotherapy. The automatic postoperative prediction would be very helpful for postoperative patients full recovery by preventing them from postoperative complications and the relapse of lung cancer. Among these postoperative prediction tasks, lung cancer postoperative complication prediction (PCP) has large clinical significance because postoperative complication is one of the major causes of perioperative death after surgery. The major clinical requirement of lung cancer PCP is to predict whether a patient will have complications and which complications the patient will experience.

The electronic medical records (EMRs) of postoperative patients record sufficient surgery information about clinical surgeries. We collected the EMRs of 8,459 patients from Department of Thoracic Surgery, West China Hospital and West China School of Medicine.¹ The dataset included 72 columns of variables. The detailed variables are listed in Table I. We attempted to develop an automatic lung cancer PCP machine learning method based on the EMR dataset. However, this was very challenging for the following reasons: (1) There were missing values in the EMRs. For some variables, patients did not know the detailed information. For example, some patients did not know their family history of cancer. Some patients' EMRs had missing values because the corresponding preoperative examination was not applied. For example, for some patients, the heart EDD value was missing because they did not take the cardiac function examination. (2) Patients' EMR variables had diverse data types, including integer, Boolean, float, and multi-integer, where multi-integer indicates multiple values with the integer data type.

Manuscript received June 20, 2019; revised September 30, 2019; accepted October 22, 2019. Date of publication October 25, 2019; date of current version June 5, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61432012 and in part by the Major Scientific and Technological Projects of the New Generation of Artificial Intelligence in Sichuan Province in 2018 under Grant 2018GZDZX0035. (Corresponding author: Zhang Yi.)

T. He, J. Guo, X. Xu, and Z. Yi are with the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: taohe@stu.scu.edu.cn; guojixiang@scu.edu.cn; xuxiuyuan@stu.scu.edu.cn; zhangyi@scu.edu.cn).

N. Chen, Z. Wang, K. Fu, and L. Liu are with the Department of Thoracic Surgery, West China Hospital and West China School of Medicine, Sichuan University, Chengdu 610065, China (e-mail: chennan@stu.scu.edu.cn; wangzihuai24@163.com; fky1525840@163.com; lunxu_liu@aliyun.com).

Digital Object Identifier 10.1109/JBHI.2019.2949601

¹<http://english.cd120.com/>

TABLE I

LIST OF 72 VARIABLES. THE THREE COLUMNS INDICATE NAMES OF VARIABLES, NUMBERS OF MISSING VALUES, AND DATA TYPE, RESPECTIVELY

Variables	No. of Missing Value	Data Type	Variables	No. of Missing Value	Data Type
age	5	Integer	cardiac function FS (%)	6556	Float
gender	0	Boolean	height	904	Float
ethnic groups	205	Integer	weight	471	Float
occupation	0	Integer	lung function FVC (L)	1464	Float
initial symptom	376	Integer	lung function FVCP (%)	1471	Float
if comorbidity	0	Boolean	lung function FEV1 (L)	1460	Float
family history of cancer	2566	Boolean	lung function FEVP (%)	1465	Float
history of surgery	947	Boolean	lung function FEV1/FVC (%)	1466	Float
history of smoking	67	Boolean	lung function PEF (L/S)	1517	Float
if quitting smoking	0	Boolean	lung function PEFP (L/S)	4640	Float
lobe of lesions	293	Multi-integer	lung function MVV (L)	1536	Float
size of lesions	1350	Float	lung function MVVP (%)	1540	Float
location of lesions	620	Multi-integer	lung function CO (ml/mm)	1805	Float
if swollen lymphaden	1544	Integer	lung function COP (%)	1831	Float
preoperative 8th TNM: T	632	Integer	six minute walking (m)	819	Float
preoperative 8th TNM: N	453	Integer	surgeon	50	Integer
preoperative 8th TNM: M	632	Integer	surgery mode	7	Integer
preoperative 8th TNM	326	Integer	No. of thoracoscopic port	2386	Integer
cardiac function EDD (mm)	6585	Float	location of thoracotomy	3094	Integer
cardiac function ESD (mm)	6586	Float	length of thoracotomy	4402	Float
cardiac function EDV (ml)	6460	Float	if breaking rib	3507	Boolean
cardiac function ESV (ml)	6465	Float	No. of breaking rib	4726	Integer
cardiac function SV (ml)	6462	Float	pleural adhesions	361	Integer
cardiac function EF (%)	6386	Float	interloped fissure	726	Integer
if transferred to thoracotomy	2370	Boolean	intraoperative analgesia mode	0	Integer
amount of bleeding (ml)	1958	Float	if pleural dissemination	505	Boolean
if blood transfusion	1386	Boolean	lymphaden station	634	Multi-integer
time of surgery (min)	330	Float	if reconstruction	1	Boolean
time of pulmonary resection (min)	5280	Float	if extensive resection	1	Boolean
time of lymphadenectomy	5541	Float	excision extension	7	Multi-integer
maximum diameter of tumors	5056	Float	if indwelling drainage tube	105	Boolean
if pleural effusion	5023	Boolean	No. of drainage tube	4084	Integer
amount of pleural effusion (ml)	7987	Float	volume of drainage of three days	710	Integer
ASA grading	5057	Integer	time of indwelling drainage tube	403	Float
time of anesthesia (min)	5170	Float	if ICU therapy	144	Boolean
if intraoperative local analgesia	0	Boolean	histology	0	Integer

An open problem in many machine learning tasks is encoding inputs with a suitable code [1]. (3) Implicit irrelative or weak relative variables existed in the EMRs. These variables would make no contribution to automatic lung cancer PCP; however, we could not directly eliminate them because we did not know exactly what they were.

In this paper, we define variables, which are very relevant to the postoperative complication, as the crucial variables. A more important requirement for lung cancer PCP is to extract the crucial variables. This requirement has large clinical importance because it allows clinicians to optimize the diagnosis scheme or use specific remedies for a specific patient according to these crucial variables. Therefore, to satisfy the clinical requirements and solve the aforementioned challenges, we defined lung cancer PCP as binary complication classification and multi-label complication classification tasks, and developed a method to extract crucial variables that caused the postoperative complication.

Recently, neural networks have been widely applied in many domains, such as image recognition [2], object detection [3], semantic segmentation [4], natural language processing [5], speech recognition [6] and speech synthesis [7]. In medical tasks, medical image segmentation and medical image classification have widely exploited neural networks. Since the development of U-Net [8], many encoder-decoder architectures [4], [9] with shortcut connections have been developed in which shortcut connections are well designed to transform context information into higher layers. In [10], an architecture that consisted of two sub-networks was proposed for the classification of retinopathy of prematurity. In [11], the proposed method consisted of convolutional neural networks (CNNs) together with a region enhancing mechanism and cross-training algorithm. This method can recognize solid nodules and identify malignant tumors. These neural network methods are mostly trained on medical images, and EMRs have been less involved.

In recent years, as class activation mapping (CAM) [12] and deconvolution [13] have developed, the visual explanation of CNNs has been well exploited. These methods create a high-resolution class-discriminative visualization, which performs as weakly supervised localization. However, these methods cannot be directly applied in a multi-layer perceptron (MLP). Therefore, in this paper, we propose a novel medical MLP (MediMLP), which involves gradient-weighted CAM (Grad-CAM) [14] (a variant of CAM), to perform PCP tasks and extract crucial variables for lung cancer PCP. The contributions of this paper are as follows:

- The MediMLP involves one local connected layer and fully connected layers with a shortcut connection to adaptively use the Grad-CAM algorithm to extract crucial variables for lung cancer PCP.
- The proposed MediMLP together with the Grad-CAM algorithm performs medical EMR classification and crucial variable extraction tasks in one shared model. Grad-CAM provides the interpretability of MediMLP. To the best of our knowledge, this is the first time that Grad-CAM has been used in MLP in addition to the first time that the use of Grad-CAM has been expanded to feature selection.
- The experimental results indicate that Grad-CAM in MediMLP has comparable performance with existing feature selection methods. With further experimental analysis, we confirm that the variable of “time of indwelling drainage tube” is the top crucial variable for lung cancer postoperative complication.

The remainder of this paper is organized as follows: In Section II, we present related works. In Section III, we first define the problem and then introduce the network architecture and corresponding Grad-CAM algorithm. We present experimental results of binary classification, multi-label classification, and crucial variable extraction of lung cancer PCP in Section IV. In Section V, we conclude our work.

II. RELATED WORKS

A. Neural Networks

Since the development of deep learning, neural networks have become the outstanding machine learning methods. Neural networks can be divided into feed-forward neural networks (FNNs) [2], [15]–[17] and recurrent neural networks (RNNs) [18], [19].

RNNs consist of a series of neural networks with interconnected neurons in every hidden layer. The most widely used RNNs is long short-term memory [20]. It has the ability to learn long-term dependencies on time-related sequence learning tasks. Its variants have been widely developed in natural language processing [5], speech recognition [6], and speech synthesis [7].

FNNs include (MLPs) [21], restricted Boltzmann Machines (RBMs) [22], and CNNs [23]. MLPs consist of multiple layers with fully connected neurons and nonlinear activation functions. They are widely used in research because of their ability to solve the object recognition problem. RBMs are a type of generative stochastic FNN that can learn the probability distribution of data. RBMs can be stacked layer by layer as deep belief networks (DBNs). CNNs are specifically good at image-related tasks,

such as image recognition [24], image segmentation [4], and object tracking [25]. They contain convolutional and pooling layers, where convolutional layers perform convolution mapping over images and pooling layers perform nonlinear mapping by shrinking the size of feature maps. In recent years, many popular CNN architectures have been proposed for image recognition problems, such as GoogLeNet [15], Highway network [17], ResNet [2], and DenseNet [16].

B. Visual Explanation of Convolutional Neural Networks

CNNs have been widely exploited and applied in many domains. However, CNNs still have not been explained fully. It is very important to understand the implicit reason why CNNs are successful in image-related tasks and the development of artificial intelligence. Recently, two views have been developed to explain this: the deconvolution view and CAM view.

Deconvolution view: The intuitive approach to understanding images is to understand the feature representations of CNNs by viewing the feature map activation. In [13], the basic idea was to transform hidden layers’ feature activations back into the input pixel space. Deconvolution [26] can be viewed as the reverse convolution operation, which is similar to the visualizing idea in [13]. Therefore, by deploying deconvolutional and unpooling layers, the trained CNNs can be reconstructed as a visualizing architecture, which is stacked in inverted order. The visualizing results present detailed corners, edges, and color features in low layers and significant pose variation, class-specific variation, and complex invariances in high layers.

CAM view: The CAM view considers that good visualizing techniques should be class-discriminative and high-resolution. Class-discriminative indicates that the locations of specific targets should be delineated and high-resolution indicates that pixel-level details should be well captured. In [12], CAM was proposed for visualization by localizing the discriminative regions of a specific class. However, CAM has to deploy a global average pooling layer to the CNNs, which limits the use of CAM. In [14], Grad-CAM was proposed to visual explanation of CNNs via gradient-based localization. Grad-CAM uses the gradients of targets to highlight important localizations in the image. It exploits the usable range of CAM and outperforms many previous methods on weakly supervised localization tasks.

C. Postoperative Complication Prediction

Postoperative complication is one of the main causes of an increase in patient morbidity and mortality. In [27], the world’s postoperative complications were reviewed and their effects were examined with respect to patient-centered outcomes. The researchers concluded that the crucial variables of postoperative complications were the postoperative discharge and length of hospital stay. In [28], the postoperative complication of postoperative pneumonia after major cancer surgery was analyzed. The researchers found that postoperative complications in patients who underwent cancer surgery could be reduced by preoperative oral care.

PCP with machine learning methods has also been developed. In [29], the random forest (RF) algorithm was used to predict

the probability of postoperative complication with head and neck squamous cell carcinoma. In [30], a group of machine learning methods, including the decision tree (DT), RF, neural networks, extreme boost, logistic regression (LR), and support vector machine (SVM), were used to predict the survival factors for breast cancer. The important variables were selected using variable selection methods in an RF. To the best of our knowledge, the machine learning method has been rarely applied to automatic lung cancer PCP.

III. METHODS

In this section, we first define the problem of lung cancer PCP to be solved in this study. Then, we present the proposed MediMLP and the corresponding Grad-CAM algorithm.

A. Definition of the Lung Cancer Postoperative Complication Prediction Problem

Given a set of EMR data, each EMR record \mathbf{r}_j is a concatenation of $[\mathbf{v}_1, \dots, \mathbf{v}_i, \dots, \mathbf{v}_{n-1}, \mathbf{v}_n]$, where \mathbf{v}_i denotes the i th variable in j th record \mathbf{r}_j and n denotes the number of variables. Three tasks to be solved are defined as follows:

Binary Classification of PCP: Each record \mathbf{r}_j has a target t_j to indicate whether the corresponding patient has the postoperative complication. This PCP learning task learns a function mapping:

$$f(\mathbf{r}_j, \theta_b) \rightarrow t_j, \quad (1)$$

where $t_j \in \{0, 1\}$, and θ_b denotes the parameters of the prediction model for the binary classification of PCP.

Multi-label Classification of PCP: Each record \mathbf{r}_j has a binary coded label $\mathbf{l}_j = [l_j^1, \dots, l_j^i, \dots, l_j^{q-1}, l_j^q]$, $l_j^i \in \{0, 1\}$, where each bit l_j^i indicates whether the corresponding patient has the i th postoperative complication and q is the number of postoperative complications. It is worth mentioning that a patient could experience multiple postoperative complications; therefore, this task is a typical multi-label classification problem, which is formulated as follows:

$$f(\mathbf{r}_j, \theta_m) \rightarrow \mathbf{l}_j, \quad (2)$$

where θ_m denotes the parameters of the prediction model for the multi-label classification of PCP.

Crucial Variables Extraction: This task could be viewed as determining the subset from \mathbf{r}_j and verifying that variables of the subset are crucial variables. This task is to learn that:

$$f(\mathbf{r}_j) \rightarrow [\mathbf{u}_1, \dots, \mathbf{u}_i, \dots, \mathbf{u}_{m-1}, \mathbf{u}_m], \quad (3)$$

where $m < n$, m is the number of extracted crucial variables, and $\mathbf{u}_i \in \mathbf{r}_j$ for each i .

B. Architecture of the Proposed MediMLP

The architecture of the proposed MediMLP is shown in Fig. 1. In the input layer, we first apply n locally connected layers to uniformly transform all variables to n neurons individually. The formulation is

$$z_i = f(\mathbf{w}_i \mathbf{v}_i + \mathbf{b}_i), n \geq i \geq 1, \quad (4)$$

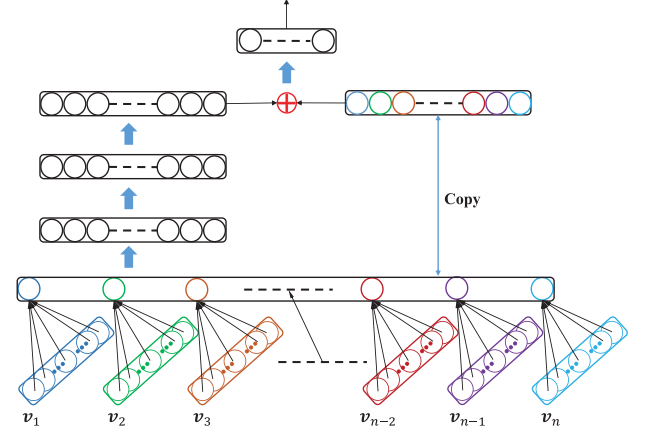


Fig. 1. Architecture of MediMLP. Blue arrows indicate full connection layers.

where \mathbf{w}_i and \mathbf{b}_i are the weights and biases, respectively, of the locally connected layer according to i th variable. Following these locally connected layers, we combine n neurons z_i to form \mathbf{z}_{low} and perform the later connection using the following formulation:

$$\begin{cases} \mathbf{y} &= \mathbf{z}_{high} + \mathbf{z}_{low}, \\ \mathbf{z}_{high} &= Q(\mathbf{z}_{low}, \theta), \end{cases} \quad (5)$$

where $Q(\cdot)$ denotes the fully connected layers of the networks' backbone and final activation \mathbf{y} is the summation of high-level and low-level features. Finally, we apply a linear connection to obtain the network output.

There are three advantages of this design. First, different variables have different encodings, so the locally connected layers produce a uniform presentation of variables. Second, the low-level features and high-level features are summarized to obtain final activations of MediMLP. This design is similar to some popular blocks [2], [16], [17]. The experimental results in Section (IV) indicate that MediMLP is useful for improving the classification performance of lung cancer PCP. Third, by using the Grad-CAM algorithm, final activation \mathbf{y} can be used to indicate the importance of corresponding variables.

C. Grad-CAM of MediMLP

In traditional machine learning methods, complication classification and crucial variable extraction are fully individual tasks; however, for the Grad-CAM algorithm, we can concurrently perform them in one trained MediMLP. We define the importance of variables as:

$$s_i = \frac{\partial y_c}{\partial y_i} \cdot y_i, \quad (6)$$

where s_i is the indicator of the i th variable's importance, y_i is an element of final activation \mathbf{y} , and y_c is the network output of MediMLP for class c . We sort n indicators and select m variables with maximum indicators as crucial variables because these variables contribute a great deal to network training and

producing class c . We verify the performance of this method in Section IV.

IV. EXPERIMENTS

In this section, we will design several experiments for the binary classification of PCP, multi-label classification of PCP, and crucial variable extraction. The dataset was collected from the thoracic surgery department of West China Hospital and contained 8,459 samples. Table I lists detailed information about the variables. The second column presents the number of missing values in the EMRs and the third column presents the data types: integer, Boolean, float, and multi-integer, where multi-integer indicates the multiple values with the integer data type. For instance, lesions can be localized in multiple locations; hence, “location of lesions” is the multi-integer data type. We encoded integer data, Boolean data, float data, and multi-integer data into one-hot coding, binary one-hot coding, float coding, and multi-label one-hot coding, respectively. We use a fixed value of “-1” to denote the missing values. The dataset was split into five subsets for 5-fold cross-validation. Four subsets each contained 1,692 samples and the remaining subset contained 1,691 samples.

A. Binary Classification

In this section, we compare MediMLP with existing machine learning methods for the binary classification task. These methods are as follows:

- **Logistic Regression (LR):** LR has been widely applied as the baseline in many binary classification tasks. It is a nonlinear learning model that consists of a linear model followed by a logistic function.
- **Decision Tree (DT):** DT is a widely used tree-like model in machine learning. It follows the rule that the outcome is the content of its leaf node, and the conditions form a conjunction in the “if” clause. We experientially set the tree depth to 5 to obtain the best performance.
- **Gradient Boosting (GB):** GB is a type of machine learning ensemble method that produces the result in the form of an ensemble of weak prediction models, typically decision trees. We set 100 estimators and the corresponding DT’s maximum depth was 5. The learning loss function was the mean squared error.
- **Random Forest (RF):** RF is also a machine learning ensemble method that constructs a multitude of decision trees. The prediction result is the mode of classification of the individual trees. The main role of RF is to correct the decision trees’ habit of overfitting. We set 100 estimators and the corresponding DT’s maximum depth was 5.
- **Support Vector Machine (SVM):** SVM is a supervised learning method that performs as a nonlinear probabilistic binary linear classifier. We set the kernel using the sigmoid function.
- **MLP:** MLP in this case is a normal MLP. We applied two hidden layers in the MLP and each layer included 320 neurons.

- **MediMLP:** MediMLP used 72 locally connected layers to obtain uniform variable representations, and the backbone was the same as the normal MLP.

All methods were validated using 5-fold cross-validation. GB and RF are ensemble-based methods and the others are single-model methods. DT, GB, RF, and SVM were implemented using the Python scikit-learn toolkit, and LR, MLP, and MediMLP were implemented using Pytorch with one TITAN Xp GPU. All methods, except MediMLP, concatenated 72 variables into a vector of 213 bits for input. LR, MLP, and MediMLP were trained using SGD with a learning rate of 0.01, batch size of 32, and weight decay of 0.01. Training the methods took approximately 2.5 mins, 8 mins and 10 mins, respectively. MLP and MediMLP involved one-dimensional BatchNorm and the LeakyReLU activation function. True positive rate (TPR), F1-measure (F1), area under the curve (AUC), and accuracy (ACC) were the metrics of binary classification.

Because our dataset had a serious label imbalance problem, we used a binary focal loss function [31] to train the networks. The focal loss function is:

$$\begin{cases} y_c^k = \sum_{i=1}^n w_{ci} \cdot y_i^k + b_c, \\ p_c^k = \sum_j^C \frac{y_c^k}{y_j^k}, \\ F_b = -\frac{1}{K} \sum_k^K \sum_{c=1}^C \alpha_c \cdot (1 - p_c^k)^\gamma \cdot t_c^k \cdot \log p_c^k, \end{cases} \quad (7)$$

where p_c^k and t_c^k are the k th sample’s predicted probability and target for class c , respectively, where C is the number of classes, and K is the number of training samples. α and γ are the hyperparameters of the focal loss function. In this experiment, $C = 2$ and we experientially set $\alpha = [0.2, 0.8]$ and $\gamma = 2$.

We gave performance curves of metrics impacted by numbers of neurons and layers of MediMLP in Fig. 2. In Fig. 2(1), We found that the impact of the layers was tiny. Thus, we considered the averaged curve of the metrics, which is indicated by the black dashed line in Fig. 2(1). Finally, we used the two-layer MediMLP in our final experiments. In Fig. 2(2), we found that the metric performance decreased when the number of neurons increased to greater than 384. Considering the averaged curve of those metrics, we used 320 neurons for each hidden layer in our final experiments.

The experimental results compared with other machine learning methods are presented in Table II. GB achieved the best TPR and RF achieved the best F1, AUC, and ACC results. This indicates that ensemble-based methods improved classification performance. Regarding single-model methods, that is, LR, DT, SVM, MLP, and MediMLP, the experimental results confirmed that MediMLP outperformed normal MLP for all metrics. We believe that the performance of MediMLP could also be boosted by the ensemble strategy, but improving classification performance was not the main objective of this paper. We also show the convergence curves of AUC and TPR, comparing MediMLP with normal MLP on the validation set in Fig. 2(3). It can be easily seen that MediMLP converged faster than normal MLP.

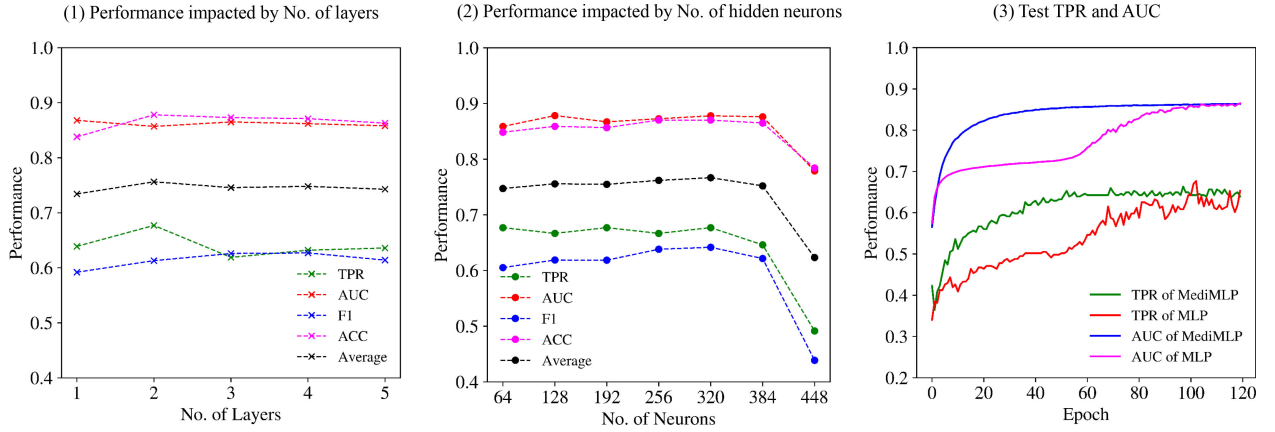


Fig. 2. Performance curves of metrics. The “Average” curve indicates the average values of TPR, AUC, F1, and ACC.

TABLE II
EXPERIMENTAL RESULTS OF BINARY CLASSIFICATION WITH 5-FOLD
CROSS-VALIDATION. THE BEST RESULTS ARE IN BOLD

Models	TPR	F1	AUC	ACC
LR	0.674	0.649	0.831	0.875
DT	0.674	0.638	0.791	0.869
SVM	0.656	0.657	0.793	0.882
MLP	0.684	0.648	0.877	0.873
MediMLP	0.684	0.65	0.88	0.875
GB	0.689	0.647	0.856	0.872
RF	0.651	0.688	0.882	0.899

B. Multi-Label Classification

In this section, we present the experimental results of multi-label classification. Using the complication classes, we define three complication labels: lung complication, other organ complication, and systemic complication. We present their details in Table III, including the number of samples, number of subtypes, and subtype examples. Lung complication indicates that a complication symptom appeared in the lung, such as pneumonia, pleural effusion, or emphysema. Other organ complication indicates that a complication symptom appeared in organs apart from the lung, such as heart failure, venous thrombus, or diarrhea. Systemic complication indicates that a complication symptom appeared as a systemic infection or for an unknown reason, such as nausea and vomiting, coma or shock. The three complication labels were determined by the severity of complications and the correlation with clinical surgery.

The network structures of MLP and MediMLP were the same as that of the binary classification. We applied a multi-label focal loss function to train them using the following formulation:

$$\begin{cases} p_c^k = \text{Sigmoid} \left(\sum_{i=1}^n (w_{ci} \cdot y_i^k + b_c) \right), \\ F_m = -\frac{1}{K} \sum_k^K \sum_{c=1}^C \alpha_c \cdot (1 - p_c^k)^\gamma \cdot l_c^k \cdot \log p_c^k \\ \quad + (1 - \alpha_c) \cdot (p_c^k)^\gamma \cdot (1 - l_c^k) \cdot \log(1 - p_c^k). \end{cases} \quad (8)$$

In this experiment, $C = 3$, and we set $\alpha = [0.15, 0.05, 0.05]$ and $\gamma = 2$.

We used micro TPR, micro F1, and Hamming loss as the metrics of multi-label classification and compared MediMLP with LR, MLP, Perceptron (single hidden layer Perceptron), and Radial Basis Function (RBF) network. Perceptron and RBF have 320 hidden neurons for comparison. Table IV presents the experimental results of multi-label classification with 5-fold cross-validation. The experimental results indicated that MediMLP achieved best performance with respect to micro TPR and F1. The performance of the Hamming Loss is comparable to baseline methods.

C. Crucial Variable Extraction

MediMLP demonstrated good performance for the binary classification of PCP and the multi-label classification of PCP. In this section, we present the experimental results of crucial variable extraction with Grad-CAM of MediMLP. In machine learning, crucial variable extraction can be viewed as a feature selection task. We compared Grad-CAM of MediMLP with the following widely used feature selection methods:

- **Variance Threshold (VT):** Variance is the expectation of the squared deviation of a variable from its mean. In machine learning, variance indicates the potential impact on the learning task. We sorted the variance of the EMR variables to indicate their importance in descending order.
- **Chi-Squared Test (CST):** CST is a statistical method that is applied to evaluate the likelihood that any observed difference between data arose by probability. We used scores of the EMR variables to indicate importance of them.
- **Mutual Information (MI):** MI is a measure of the mutual dependence between two random variables. We counted the dependence scores between each EMR variable and label, and used them as the importance of the EMR variables.
- **Pearson Correlation Coefficient (PCC):** PCC is a widely used measure of the linear correlation between two variables. We used PCC to define the correlation between each

TABLE III
DETAILS OF COMPLICATION TYPES

Complication Types	No Complication	Lung Complication	Other Organs Complication	Systemic Complication
No. of Samples	7013	1239	203	174
No. of Subtypes	None	63	67	49
Subtype Examples	None	Pneumonia, pleural effusion, emphysema, lung abscess.	Heart failure, venous thrombus, infection of incisional wound, diarrhea, urinary tract infection.	Nausea and vomiting, coma, shock, rash, anaphylactic reaction.

TABLE IV
EXPERIMENTAL RESULTS OF MULTI-LABEL CLASSIFICATION
WITH 5-FOLD CROSS-VALIDATION

Models	Micro TPR	Micro F1	Hamming Loss
LR	0.601	0.584	0.055
RBF	0.609	0.598	0.061
Perceptron	0.613	0.569	0.059
MLP	0.607	0.591	0.051
MediMLP	0.618	0.602	0.053

EMR variable and label, and selected crucial variables according to the PCC values.

- **Recursive Feature Elimination (RFE):** RFE selects features by recursively analyzing increasingly smaller sets of features. We used LR as the estimator and selected three features for each iterator. The final scores of RFE were used as the importance of the EMR variables.
- **Gini Importance (GI):** GI performs relevant feature identification based on the RF algorithm. It provides multivariate feature importance scores for each EMR variable. The corresponding RF has 100 DTs with a maximum depth of 5.

We implemented these methods using the Python scikit-learn package. To quantitatively compare Grad-CAM of MediMLP with these methods we first asked clinicians in the thoracic surgery department at West China Hospital to empirically label the importance of the EMR variables with score $s \in (0, 1)$. Then, we reversely sorted these scores. We present the results of the top 10 variables in Table V. We transformed the numerical sequence of the variables into a vector to indicate the uniform importance of the variables. For instance, a vector $v = [1, 2, \dots, m]$ was used to indicate the importance of the clinicians' manual labels. The Euclidean distance (ED) and PCC² were calculated as the correlation metrics between manual labels and the methods, and were used to validate the performance of the methods.

We calculated the metrics of corresponding vectors for the top $m = 10$, $m = 24$ and $m = 36$ bits. The experimental results are listed in Table VI. The best results are shown in

²PCC here has a different use to that used as a feature selection method. PCC used for crucial variable extraction calculates the correlation between EMR variables and complication labels, but this PCC is used as the metric calculated correlation between manual labels and the importance of feature selection methods.

bold and the second-best results are underlined. For the ED metric, Grad-CAM, CST, and PCC presented the best results. For the PCC metric, Grad-CAM and GI presented the best results. However, from the results in Tables V and VI, we can deduce that empirical manual labels were inaccurate for some variables and inconsistent with these feature selection methods.

Another approach to evaluate these crucial variable extraction methods is to present the classification performance by inputting the top m crucial variables. Thus, we trained LR and MediMLP with $m = 72, 66, 60, 54, 48$. LR was used as the classification model for all feature selection methods. We plotted the performance curves of the binary classification in Fig. 3. We found that the performance of VT decreased quickly when $m = 54$, which means that VT was not suitable as the feature selection method for the binary classification of PCP. For AUC, the methods had comparable results, except CST had an outstanding AUC of 0.864 when $m = 66$. For ACC, RFE and PCC had consistently good results for most variables. MI had the lowest results when $m = 66$ and $m = 60$, but performed well for most variables because the performance curves continued to increase when m decreased. GI and Grad-CAM performed well at first, but they decreased when $m < 66$. For TPR, GI, Grad-CAM, and CST had constant good results for most variables. MI had the best result when $m = 66$, but its performance decreased quickly when m decreased. RFE and PCC did not perform well at each point of the performance curve. For F1, RFE, PCC, and GI had constant good results for most variables. MI and CST had good results at first, but their performance decreased when m decreased. MI had the lowest results when $m = 66$ and $m = 60$, but its performance increased when m decreased. To summarize, all methods except VT had comparable performance for binary classification when m decreased.

We also plotted performance curves for multi-label classification in Fig. 4. We found that VT was also not suitable as the feature selection method for the multi-label classification of PCP. For the Hamming loss, MI had constant low losses. Grad-CAM, GI, and RFE had oscillating but low losses. CST and PCC had oscillating but high losses when $m = 66$, $m = 60$, and $m = 54$. For TPR, CST, PCC, Grad-CAM, and GI had almost constant good performance. MI had constant but low results. RFE had low performance at first but its performance increased when $m < 60$. For F1, only MI had constant good results. Grad-CAM and GI had a low result at first, but they maintained good performance when $m < 66$. RFE and PCC had good performance at first, but their performance decreased when

TABLE V
TOP 10 MANUALLY LABELED CRUCIAL VARIABLES

Variables	Manual	Grad-CAM	VT	CST	MI	PCC	RFE	GI
if comorbidity	1	31	1	14	54	66	36	37
age	2	6	43	25	21	5	6	5
if ICU therapy	3	4	31	20	8	9	33	4
time of indwelling drainage tube	4	1	54	1	1	1	2	1
preoperative 8th TNM: N	5	28	25	30	59	21	37	48
amount of bleeding (ml)	6	9	59	17	3	8	56	21
volume of drainage of three days	7	50	72	22	2	23	27	2
preoperative 8th TNM: M	8	22	42	44	58	60	14	49
cardiac function EF (%)	9	19	8	33	18	30	44	56
if blood transfusion	10	51	32	6	61	12	21	22

TABLE VI
ED AND PCC PERFORMANCE OF COMPARED METHODS. THE BEST RESULTS ARE IN BOLD AND SECOND-BEST RESULTS ARE UNDERLINED

Metrics	No. of Variables	Grad-CAM	VT	CST	MI	PCC	RFE	GI
ED	10	<u>77.42</u>	124.53	61.79	110.89	90.33	96.33	91.96
	24	118.69	166.68	142.08	144.4	<u>126.73</u>	160.10	153.88
	36	162.68	184.84	<u>162.04</u>	174.06	141.32	174.78	171.55
PCC	10	0.4964	0.1158	0.0001	0.1671	-0.0024	0.2349	<u>0.4440</u>
	24	<u>0.3789</u>	-0.1415	0.1827	-0.0216	-0.0431	0.2707	0.4579
	36	0.6322	0.0857	<u>0.3787</u>	0.1842	0.2748	0.1822	0.3506

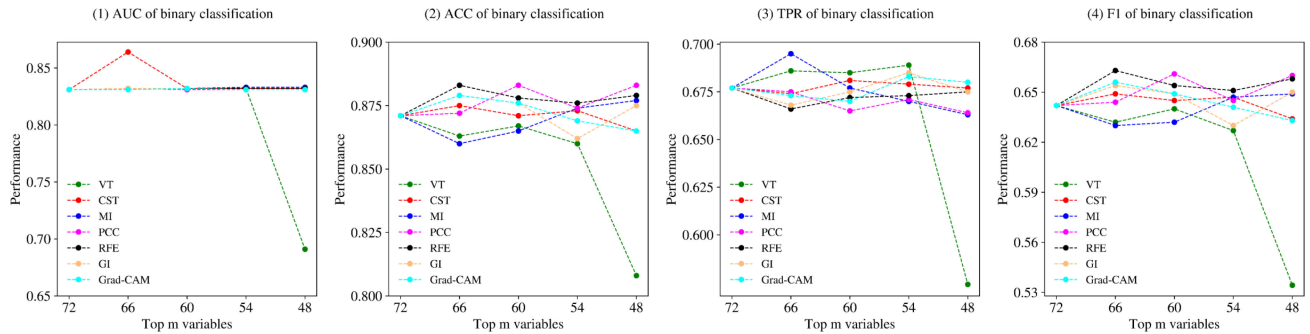


Fig. 3. Performance curves with the top m variables of binary classification.

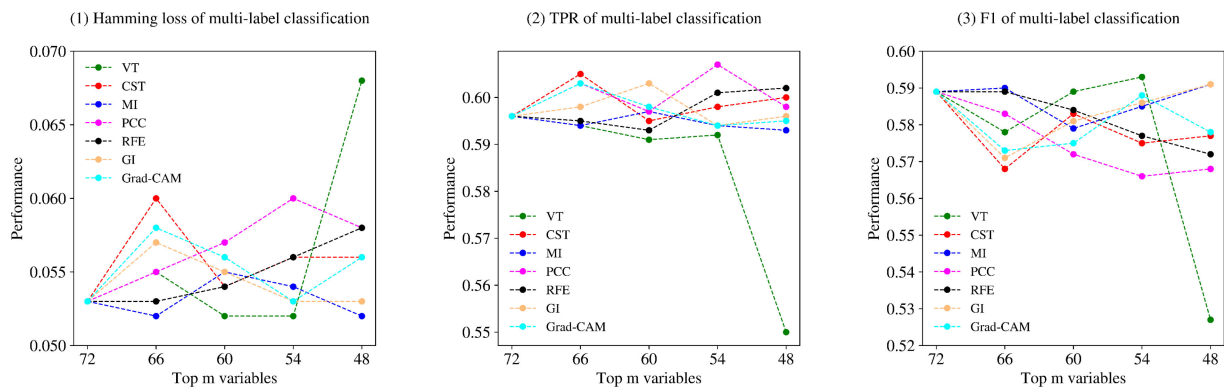


Fig. 4. Performance curves for the top m variables of multi-label classification.

TABLE VII

TOP 10 CRUCIAL VARIABLES BY AVERAGING THE RESULTS OF GRAD-CAM, CST, PCC, AND GI. THE NUMBER IN PARENTHESES INDICATES THE RANKING NUMBER OF MANUAL LABELS. THE WORST RESULTS ARE IN BOLD

Sequence (Manual)	Omitted Crucial Variables	Binary Classification				Multi-label Classification		
		TPR	F1	AUC	ACC	Micro TPR	Micro F1	Hamming Loss
1(4)	time of indwelling drainage tube	0.161	0.229	0.691	0.818	0.468	0.327	0.123
2(3)	if ICU therapy	0.68	0.639	0.831	0.868	0.603	0.574	0.057
3(2)	age	0.684	0.633	0.831	0.864	0.594	0.586	0.053
4(10)	if transferred to thoracotomy	0.670	0.654	0.832	0.879	0.596	0.596	0.052
5(5)	amount of bleeding (ml)	0.665	0.661	0.832	0.883	0.596	0.580	0.055
6(68)	location of thoracotomy	0.669	0.656	0.832	0.880	0.598	0.577	0.056
7(66)	histology	0.681	0.639	0.832	0.868	0.598	0.583	0.055
8(22)	if reconstruction	0.677	0.643	0.832	0.871	0.594	0.584	0.054
9(21)	if extensive resection	0.662	0.664	0.833	0.885	0.597	0.569	0.058
10(50)	height	0.679	0.643	0.833	0.870	0.591	0.586	0.053

$m < 66$. CST had oscillating but low results. To summarize, all methods except VT had comparable performance when m decreased for multi-label classification.

D. Experimental Analysis of the Top 10 Crucial Variables

To obtain most crucial variables in the EMR dataset, we averaged the ranking number of Grad-CAM, CST, PCC, and GI. In Table VII, we present the top 10 crucial variables according to the average result. A variable is a crucial variable if it heavily impacts the performance of lung cancer PCP tasks. Thus, we conducted 10 binary classification and multi-label classification experiments by inputting 71 variables, but without the variable in the corresponding row. From the results in Table VII, we found that the variable of “time of indwelling drainage tube” was very relevant to PCP because the performance of binary classification and multi-label classification decreased a great deal for all metrics when it was omitted. The impact was negligible for the remaining 9 crucial variables.

Although other crucial variables had a negligible impact on postoperative complication classification tasks, we could not assess whether other variables are not crucial for lung cancer PCP for the following reasons. First, the EMR dataset was not sufficiently large. Of the 8,459 samples, only 1,445 samples were labeled as complications. Second, there were many missing values in the EMR records.

To explain the impact of the crucial variable “time of indwelling drainage tube” and missing values on lung cancer PCP, we provide the statistics of missing values of “time of indwelling drainage tube” for false predicted samples of MediMLP in Table VIII. For each “ x/y ,” x and y indicate the number of missing values and FN, FP, or F, where FN, FP, and F indicate false negative samples, false positive samples, and false predicted samples, respectively. Compared with the global missing value rate of “time of indwelling drainage tube,” 4.8% (403/8,459), we obtained the following: 18.1% > 4.8% for FN, 2.8% < 4.8% for FP, and 9.5% > 4.8% for F, respectively. Therefore, we found that MediMLP predicted some false negative samples because of the missing values of “time of indwelling drainage tube.”

TABLE VIII

MISSING VALUE RATE ON EACH TEST SUBSET. FN, FP, AND F INDICATE FALSE NEGATIVE SAMPLES, FALSE POSITIVE SAMPLES, AND FALSE PREDICTED SAMPLES, RESPECTIVELY

Test Subset	FN	FP	F
1	21.4% (22/103)	3.1% (4/129)	11.2% (26/232)
2	12.8% (11/86)	0.0% (0/89)	6.3% (11/175)
3	14.0% (15/107)	1.8% (2/110)	7.8% (17/217)
4	18.4% (16/87)	5.9% (8/135)	10.8% (24/222)
5	24.4% (21/86)	2.0% (3/147)	10.3% (24/233)
Sum	18.1% (85/469)	2.8% (17/610)	9.5% (102/1,079)

V. CONCLUSION

In this paper, we proposed the MediMLP together with the Grad-CAM algorithm for lung cancer PCP. The trained MediMLP was used simultaneously for postoperative complication classification and crucial variable extraction. To the best of our knowledge, this is the first time that Grad-CAM has been used in MLP for feature selection. With our experimental analysis of the top 10 crucial variables, we found that the variable “time of indwelling drainage tube” was the top crucial variable for lung cancer PCP.

In the future, we will focus on improving the experimental results of postoperative complication classification tasks. The most challenging problem is to overcome the problem of missing values. In real clinical applications postoperative complications are typically predicted using X-ray images. We will attempt to extend MediMLP to CNNs for lung cancer PCP on X-ray images in future work.

REFERENCES

- [1] B. Wu, J. Jia, T. He, J. Du, X. Yi, and Y. Ning, “Inferring users’ emotions for human-mobile voice dialogue applications,” in *Proc. IEEE Int. Conf. Multimedia Expo*, 2016, pp. 1–6.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

- [3] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
- [4] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
- [5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3156–3164.
- [6] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.
- [7] A. van den Oord *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3915–3923.
- [8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Medical Image Comput. Comput. Assisted Intervention*, 2015, pp. 234–241.
- [9] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
- [10] J. Hu, Y. Chen, J. Zhong, R. Ju, and Z. Yi, "Automated analysis for retinopathy of prematurity by deep neural networks," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 269–279, Jan. 2019.
- [11] X. Qi *et al.*, "Automated diagnosis of breast ultrasonography images using deep neural networks," *Medical Image Anal.*, vol. 52, pp. 185–198, 2019.
- [12] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2921–2929.
- [13] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 818–833.
- [14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 618–626.
- [15] S. Christian *et al.*, "Going deeper with convolutions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 1–9.
- [16] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2261–2269.
- [17] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 2377–2385.
- [18] L. Zhang, Z. Yi, and S.-I. Amari, "Theoretical study of oscillator neurons in recurrent neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5242–5248, Nov. 2018.
- [19] Z. Yi, "Foundations of implementing the competitive layer model by lotka-Volterra recurrent neural networks," *IEEE Trans. Neural Netw.*, vol. 21, no. 3, pp. 494–507, Mar. 2010.
- [20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [21] D. E. Rumelhart and J. L. McClelland, "Learning internal representations by error propagation. in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. 1. Foundations*. Cambridge, MA, USA: MIT Press, 1987.
- [22] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [23] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Into Imag.*, vol. 9, no. 4, pp. 611–629, 2018.
- [24] T. He, H. Mao, and Z. Yi, "Moving object recognition using multi-view three-dimensional convolutional neural networks," *Neural Comput. Appl.*, vol. 28, pp. 3827–3835, 2016.
- [25] T. He, H. Mao, J. Guo, and Z. Yi, "Cell tracking using deep neural networks with multi-task learning," *Image Vision Comput.*, vol. 60, pp. 142–153, 2016.
- [26] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. Int. Conf. Comput. Vision*, 2011, pp. 2018–2025.
- [27] S. E. Tevis and G. D. Kennedy, "Postoperative complications and implications on patient-centered outcomes," *J. Surgical Res.*, vol. 181, no. 1, pp. 106–113, 2013.
- [28] M. Ishimaru, H. Matsui, S. Ono, Y. Hagiwara, K. Morita, and H. Yasunaga, "Preoperative oral care and effect on postoperative complications after major cancer surgery," *Brit. J. Surgery*, vol. 105, no. 12, pp. 1688–1696, 2018.
- [29] Y. Chen, W. P. Cao, X. Gao, H. Ong, and T. Ji, "Predicting postoperative complications of head and neck squamous cell carcinoma in elderly patients using random forest algorithm model," *BMC Med. Inf. Decision Making*, vol. 15, 2015, Art. no. 44.
- [30] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, "Predicting factors for survival of breast cancer patients using machine learning techniques," *BMC Med. Inform. Decision Making*, vol. 19, no. 1, 2019, Art. no. 48.
- [31] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2999–3007.