Machine vision-assisted identification of the lung adenocarcinoma category and high-risk tumor area based on CT images

Highlights

- We study machine vision-assisted lung adenocarcinoma classification using CT images
- We design a holistic machine vision framework, improving classification performance
- Our method outperforms famous deep CNNs and medical imaging classification methods
- Our method better explains relations between CT patterns and pathological diagnoses

Authors

Liuyin Chen, Haoyang Qi, Di Lu, ..., Long Wang, Guoyuan Liang, Zijun Zhang

Correspondence

zijzhang@cityu.edu.hk

In brief

In this study, the authors developed a holistic machine vision framework for developing a data-driven model to identify the lung adenocarcinoma category based on CT images only and improved its generalization in datasets from different resources through a knowledge distillation procedure. The authors demonstrate that the CT image features could be adopted to infer the pathological classification of lung adenocarcinoma and further discussed the relationship between CT features and pathological characteristics.
Article

Machine vision-assisted identification of the lung adenocarcinoma category and high-risk tumor area based on CT images

Liuyin Chen,1,6 Haoyang Qi,1,5 Di Lu,2,5 Jianxue Zhai,2 Kaican Cai,2 Long Wang,3 Guoyuan Liang,4 and Zijun Zhang1,6,*

1School of Data Science, City University of Hong Kong, Hong Kong SAR, China
2Department of Thoracic Surgery, Nanfang Hospital, Southern Medical University, Guangzhou, China
3Department of Computer Science and Technology, University of Science and Technology Beijing, Beijing, China
4Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
5These authors contributed equally
6Lead contact
*Correspondence: zijzhang@cityu.edu.hk
https://doi.org/10.1016/j.patter.2022.100464

SUMMARY

Computed tomography (CT) is a widely used medical imaging technique. It is important to determine the relationship between CT images and pathological examination results of lung adenocarcinoma to better support its diagnosis. In this study, a bilateral-branch network with a knowledge distillation procedure (KDBBN) was developed for the auxiliary diagnosis of lung adenocarcinoma. KDBBN can automatically identify adenocarcinoma categories and detect the lesion area that most likely contributes to the identification of specific types of adenocarcinoma based on lung CT images. In addition, a knowledge distillation process was established for the proposed framework to ensure that the developed models can be applied to different datasets. The results of our comprehensive computational study confirmed that our method provides a reliable basis for adenocarcinoma diagnosis supplementary to the pathological examination. Meanwhile, the high-risk area labeled by KDBBN highly coincides with the related lesion area labeled by doctors in clinical diagnosis.

INTRODUCTION

According to the WHO 2015 report,1 approximately 8.8 million deaths were caused by cancer, of which lung cancer constituted 20%. Lung adenocarcinoma is the most common type of lung cancer, whose early diagnosis and proper treatment are important. According to the classification standard of lung tumors described by the International Association for the Study of Lung Cancer, American Thoracic Society, and European Respiratory Society classification in 2011 as well as the WHO in 2015, lepidic-predominant adenocarcinomas ≤3 cm in size can be classified into (1) adenocarcinoma in situ (AIS), which shows the entire lepidic growth, (2) minimally invasive adenocarcinoma (MIA) with invasion of no more than 5 mm, and (3) invasive adenocarcinoma (IAC), based on the degree of infiltration.2 It is believed that this classification standard of lung
adenocarcinoma in pathophysiology helps improve the predictive ability of clinical outcomes and therapeutic benefits, which are important in the diagnosis.\(^3\)

In real-world practice, lung adenocarcinoma is usually classified based on the results of pathological examination, which evaluates the degree of infiltration, such as determining the foci of stromal, vascular, and pleural invasion as well as measuring the largest single focus of the invasion and central scans.\(^4\) When it comes to computer-vision-based methods, histopathological images are chosen in most datasets.\(^5\) However, pathological examination is not performed as a routine evaluation to diagnose lung diseases, which may lead to misdiagnosis, as this examination might not be conducted especially at the early stage of lung cancer. In clinical practice, computed tomography (CT) is a commonly adopted auxiliary lung cancer diagnosis technique owing to its value in accurately inspecting chronic changes in the lung parenchyma.\(^6\) In fact, research results have shown that AIS and MIA are tumors most likely to be detected on imaging procedures.\(^7\) Therefore, it is meaningful and urgent to develop a non-pathological method to help identify adenocarcinoma types and to compensate for the limitations of pathological examination. A CT-image-based method turns out to be the best alternative.

Recent studies\(^8\)–\(^10\) have been conducted by radiologists and pathologists to discover and further confirm that degrees of tumor invasiveness can lead to different symptoms on CT images, indicating the feasibility of adopting CT imaging as an auxiliary diagnostic technique for classifying lung adenocarcinoma. Yanagawa et al.\(^9\) reported that the irregular margin, the air bronchogram with disruption and/or irregular dilatation, and pleural indentation might distinguish IAC from AIS and that the solid portion size on CT could be significantly different between IAC and MIA (Figure S1). Other studies\(^9\)–\(^10\) have reported similar results. The findings of existing studies\(^9\) have indicated that the percentage of solid volume and the proportion of solid mass in the entire nodule increased as the adenocarcinoma became more invasive histopathologically. Previous studies\(^8\)–\(^10\) have provided solid evidence on the feasibility of further studying the computer vision model for automating the adenocarcinoma classification by examining lung CT images; however, experimental verification in a larger number of patients is still warranted. Our study fills this gap in the literature.

This paper develops a holistic modeling framework based on convolutional neural networks (CNN) to facilitate lung adenocarcinoma diagnosis. The developed model allows an automated identification of adenocarcinoma categories and detection of the tumor area on CT images that most likely contribute to the identification of the specific type of adenocarcinoma. The proposed framework consists of three major data-analytical stages: preprocessing, feature engineering, and final classification. In the preprocessing stage, the segmentation and rebalancing units are developed to exclude the redundant background of CT images and rebalance the long-tailed data distribution, respectively. The preprocessed datasets are then fed into the representation branch and the rebalance branch in the feature-engineering stage to generate the weighted features of the images. Finally, in the classification stage, these features are passed on to the fully connected layer as a classifier to identify the adenocarcinoma categories. The high-risk area on images can be generated simultaneously by utilizing a class activation map (CAM),\(^11\) which highlights the regions related to the classification process. A knowledge distillation procedure is developed for the proposed framework to facilitate the model generalization to different datasets. To further explore the potential of deep-learning-based methods for the identification of adenocarcinoma categories using CT images, the features of different images and the overall imbalanced distribution of the dataset are considered in entirety, thus obtaining ideal results. Meanwhile, the developed framework has been verified to achieve a state-of-the-art performance based on additional datasets collected from multiple sources and by comparing our method with a set of solid benchmarking methods.

RESULTS

In this study, we developed a CNN-based bilateral-branch-network diagnosis framework with a knowledge distillation procedure (KDBBN) that can identify the lung adenocarcinoma category based on CT images rather than the results of pathological examinations or histopathological images (Figure 1). To rebalance the extreme long-tailed distribution, we innovatively use two kinds of data: representation data and rebalanced data. To process representation and rebalanced data respectively in latent feature engineering, two CNN-based branches, the representation branch and the rebalance branch, are developed in the framework. The final latent features are obtained by aggregating the features generated by these two branches. The performance of the model was evaluated based on overall...
accuracy, precision, F1-score, and area under the receiver-operating characteristic (ROC) curve (AUC). Figure 2 illustrates the details of the overall architecture of the proposed framework. 

Processing the special long-tailed dataset with representation data and rebalanced data

The lung adenocarcinoma dataset has long-tailed distribution. Figure 3 demonstrates the data distribution of long-tailed datasets, in which most samples belong to several head categories, whereas the other tail categories have few samples. This distribution indicates that the data are extremely imbalanced.

Meanwhile, it is challenging to obtain sufficient evidence to infer the real distribution because of the high rate of missed diagnosis of MIA/AIS. Achieving a high overall accuracy is obviously insufficient. Traditional imbalanced learning techniques, such as the synthetic minority oversampling technique (SMOTE), also do not work ideally, because lung patterns should remain clear and analyzable clinically to explain the relationship between CT imaging features and results of pathological examination.

Therefore, two types of data were considered: representation data and rebalanced data. Representation data retain the observed distribution of lung adenocarcinoma categories and serve as a basis consistent with the traditional medical image-processing system. Rebalanced data transform the long-tailed distribution to a more balanced distribution, whose main purpose is to elevate the identification performance of the data-driven model on tail categories. The process of generating these two types of data is regarded as the preprocessing stage.

Preprocessing stage 1: Segmenting vital area

The segmentation unit aims to extract the region of interest (ROI) from the raw data by removing the abounding background in these CT images because the diagnosis-related information only lies in the central lung section and the background of CT images can be misleading. Three image-segmentation methods were applied in the proposed framework: the crop-background extractor, simple ROI (SROI) extractor, and conditional random field (CRF) extractor. Of the three extractors, both the SROI and CRF extractors segment the complete lung regions along the edge of the section, where the CRF extractor has a better performance on the precise division of boundaries. However, the computing complexity of the CRF extractor is high and dispensable when the quantity rather than the quality of the result is of greater concern. The SROI extractor is designed as a cost-efficient supplement to the CRF images in the rebalanced data, which enhances the data diversity and reduces the data duplication in the tail categories. The combination of SROI and CRF extractors potentially strengthens the robustness and reduces overfitting of the framework, whereas common oversampling methods, such as SMOTE, could transform the features of the samplers, causing the lung area to appear unclear or deformed. Moreover, special attention should be paid to the specific solid portion area, and for this reason the crop-background extractor is introduced. To fine-tune the crop-background extractor, the proportion of the solid portion in the entire image should be controlled. The rebalanced data composed of all three kinds of images presents an ideal distribution of the adenocarcinoma samples, taking the clinical diagnostic preference into consideration.

Generally, the result from the CRF extractor has a more precise boundary, reducing unnecessary information loss, although the result from the SROI extractor is also acceptable. The boundary of the lung section in the ROI image with a black background offers a higher contrast than that with a white background.
Crop-background extractor

The crop-background extractor is a semi-automated segmentation algorithm, focusing on roughly highlighting the cancer-related area, given

$$\rho = \frac{\sum \text{pixel}}{\text{number of total pixels}}$$  \hspace{1cm} (Equation 2)

$$\mu = \frac{\sum \text{row:col pixel}}{\text{number of col/row}}$$  \hspace{1cm} (Equation 3)

where pixel denotes the pixel value of a CT image and $\mu$ represents the average pixel value of one row or column. One characteristic of the lung CT image is that the solid portion of the nodule in the lung parenchyma is close to the white pixel while the redundant background and lung parenchyma are close to the black pixel. Thus, a larger $\mu$ indicates that this row/column is more likely to contain nodule information. The role of $\rho$ is similar to that of $\mu$, except that $\rho$ represents the entire image rather than the row or column in $\mu$, serving as a supplement to $\mu$. Based on setting different thresholds on $\mu$ and $\rho$, images containing different densities of information are generated as shown in Figure S3. Intuitively, the larger the $\mu$ and $\rho$, the smaller the crop-background image, whereas a larger solid nodule accounts for the entire image. These images provide nodule areas with less redundant information and higher information density, highlighting the central areas in training. However, unlike the CRF or SROI extractor, which excludes the background outside the lung section, the crop-background extractor might also remove some lung areas whose information density is lower than the threshold. Thus, by fine-tuning $\mu$ and $\rho$, the importance of the marginal area in the lung section for identification of lung adenocarcinoma category can be determined.

Preprocessing stage 2: Rebalancing

Figure 4 briefs the main idea of the rebalancing unit. The first random sampler $S_1$ is applied to identically distributed long-tailed datasets, from Data 1 to Data $N$, individually with preservation of the original long-tailed distribution. The parameter $\beta$ denotes the target proportion of the total number of samples from Data 1 to Data $N$. The second random sampler $S_2$ acts on the integrated dataset generated by $S_1$ and transforms its distribution from the long-tailed to a new target distribution $\gamma$, which in our case is roughly a uniform distribution.

To integrate $N$ different datasets $D_1, D_2, \ldots, D_N$ with the same data distribution $\psi$ into one dataset $D$ with the specified data distribution $\psi'$, the following equation describes the relationship:

$$D = S_2(\gamma; S_1(\beta; D_1, D_2, \ldots, D_N)),$$  \hspace{1cm} (Equation 4)

where $S_1(\cdot; \cdot)$ and $S_2(\cdot; \cdot; \cdot)$ represent two random samplers. In $S_1(\cdot; \cdot)$ and $S_2(\cdot; \cdot; \cdot)$, the first parameter is the data proportion of each category after resampling, while the second represents the input dataset. $\beta = (\beta_1, \beta_2, \ldots, \beta_N)$, $\beta_1 + \beta_2 + \ldots + \beta_N = 1$, in which $\beta_i (i = 1, \ldots, N)$ denotes the proportion of the corresponding dataset $D_i$ in each category of the result dataset $D$. $\gamma = (\gamma_A, \gamma_M, \gamma_I)$, $\gamma_A + \gamma_M + \gamma_I = 1$, where $\gamma_i (i = A, M, I)$ denotes the proportion of corresponding categories AIS, MIA, and IAC in the result dataset $D$. The detailed algorithm of the data rebalancing unit is described in algorithm 2 (Table S6).
The rebalancing unit aims to rebalance the dataset distribution and increase the diversity of the training images without changing the regular pattern of the lung in the CT images. Maintaining the regular pattern also avoids the global position information of the lesion area in the CT images from being affected. Moreover, with the parameter $b$, the proportion of datasets generated by different extractors can be controlled. Among them, $b_c$, controlling crop-background images, is the most vital. It enhances the data variety and controls the degree of attention of the specialized area in the rebalance branch.

**Two-branch network architecture**

In the feature-engineering stage, the feature extraction network is naturally divided into two branches. The representation branch processes the representation data and performs representation learning, whereas the rebalance branch processes the rebalanced data and improves the identification performance of the network in the tail categories. Finally, the features generated by the two branches are aggregated via weighted average to acquire the output feature.

The feature extraction task in the two branches is handled by CNN backbones, which can be adjusted and chosen accordingly. Features generated by appropriate CNN models aim to facilitate representation learning and mitigate the impact of the data imbalance on the final identification results. In this study, we expect the CNN backbone in the representation branch to automatically obtain the most representative latent features for the original dataset while the CNN backbone in the rebalance branch should pay more attention to the tail category. Two well-known CNN structures widely applied in medical imaging, DenseNet169 and ResNet50, are considered as candidates for each feature-engineering branch.

The feature vectors $f_r$ and $f_e$ are generated by the global average pooling layers from DenseNet backbone in the representation branch and the ResNet backbone in the rebalance branch, with the same dimension. Subscripts $r$ and $e$ denote the features or parameters in the representation and rebalance branches, respectively. The features are further incorporated to one vector $f$ by weighting $f_r$ and $f_e$ with a parameter $\alpha$, which can be formulated as follows:

$$f = \alpha f_r + (1 - \alpha)f_e.$$  \hspace{1cm} (Equation 5)

The output logits are formulated as follows:

$$z = W^T f.$$ \hspace{1cm} (Equation 6)

where $W$ denotes the weight matrix of the final fully connected (FC) layer. The softmax function layer calculates the probability distribution for the adenocarcinoma categories via

$$\tilde{\rho} = \frac{\exp(z)}{\sum \exp(z)}.$$ \hspace{1cm} (Equation 7)

If we denote $E$ as the cross-entropy loss function, the weighted cross-entropy loss for the identification process combining two branches can be illustrated as

$$L = \alpha E(\tilde{\rho}, y_r) + (1 - \alpha)E(\tilde{\rho}, y_e)$$ \hspace{1cm} (Equation 8)

and the whole identification network is end-to-end trainable.

**DISCUSSION**

**Comparison between the proposed framework and other state-of-the-art methods**

Popular image-classification models are considered as benchmarking methods in computational experiments; these models are used to evaluate the performance of the proposed framework. The benchmarking methods are divided into four groups according to the different feature extractors and classifiers. The first group of benchmarking methods is type I, which adopts LBP\textsuperscript{21} or GLCM\textsuperscript{22} as the feature extractor, and k-nearest neighbor (KNN),\textsuperscript{23} or support vector machine (SVM)\textsuperscript{24} as the classifier, denoted as, for example, “LBP + KNN” or “LBP + SVM.” The type I method represents the performance of the classical machine-learning classification model with a traditional feature-engineering method.
Table 1. Comparison of the performance of benchmarking methods and the proposed framework (percentage, mean ± SD)

| Method          | Class Precision | F1         | Accuracy | AUC   |
|-----------------|-----------------|------------|----------|-------|
| Type I          |                 |            |          |       |
| LBP + KNN       | IAC 91.8 ± 0.4  | 94.8 ± 0.2 | 90.3 ±   | 89.1 ±|
|                 | MIA 85.1 ± 6.3  | 58.2 ± 6.4 | 0.4      | 1.4   |
|                 | AIS 55.2 ± 6.0  | 39.0 ± 7.8 |          |       |
| LBP + SVM       | IAC 92.8 ± 0.4  | 94.8 ± 0.2 | 87.3 ±   | 87.7 ±|
|                 | MIA 84.1 ± 5.3  | 56.7 ± 5.4 | 0.5      | 1.4   |
|                 | AIS 54.2 ± 4.1  | 39.6 ± 7.1 |          |       |
| GLCM + KNN      | IAC 91.9 ± 0.4  | 94.8 ± 0.2 | 86.5 ±   | 87.1 ±|
|                 | MIA 83.1 ± 3.3  | 58.8 ± 6.4 | 0.6      | 0.4   |
|                 | AIS 54.0 ± 6.4  | 34.0 ± 4.8 |          |       |
| GLCM + SVM      | IAC 92.5 ± 0.4  | 94.8 ± 0.2 | 87.2 ±   | 88.0 ±|
|                 | MIA 85.1 ± 2.3  | 58.2 ± 6.4 | 0.4      | 1.2   |
|                 | AIS 55.0 ± 6.0  | 38.1 ± 6.5 |          |       |
| Type II         |                 |            |          |       |
| DB + KNN        | IAC 91.8 ± 0.6  | 95.4 ± 0.3 | 91.5 ±   | 89.1 ±|
|                 | MIA 88.4 ± 4.1  | 61.2 ± 7.6 | 0.7      | 1.9   |
|                 | AIS 81.5 ± 12.0 | 39.0 ± 13.1|          |       |
| RB152 + KNN     | IAC 92.1 ± 0.3  | 95.4 ± 0.2 | 91.1 ±   | 88.0 ±|
|                 | MIA 79.6 ± 12.4 | 58.6 ± 7.2 | 0.6      | 2.2   |
|                 | AIS 76.5 ± 9.9  | 41.4 ± 5.4 |          |       |
| RB50 + KNN      | IAC 91.8 ± 0.6  | 95.2 ± 0.6 | 90.8 ±   | 87.4 ±|
|                 | MIA 77.8 ± 13.6 | 57.7 ± 13.5| 1.0      | 2.2   |
|                 | AIS 72.0 ± 5.4  | 36.8 ± 7.0 |          |       |
| DB + SVM        | IAC 91.0 ± 0.6  | 95.4 ± 0.3 | 91.2 ±   | 89.0 ±|
|                 | MIA 88.8 ± 3.6  | 61.8 ± 6.8 | 0.4      | 1.5   |
|                 | AIS 80.5 ± 12.0 | 40.0 ± 12.6|          |       |
| RB152 + SVM     | IAC 92.1 ± 0.3  | 95.1 ± 0.5 | 91.3 ±   | 88.3 ±|
|                 | MIA 79.9 ± 11.4 | 58.7 ± 6.2 | 0.5      | 1.9   |
|                 | AIS 77.1 ± 8.8  | 41.8 ± 5.4 |          |       |
| RB50 + KNN      | IAC 91.6 ± 0.7  | 95.2 ± 0.6 | 90.4 ±   | 87.2 ±|
|                 | MIA 78.1 ± 12.6 | 58.2 ± 12.5| 0.8      | 2.0   |
|                 | AIS 71.7 ± 5.5  | 38.8 ± 7.3 |          |       |
| Type III        |                 |            |          |       |
| D169            | IAC 96.4 ± 0.5  | 98.0 ± 0.3 | 95.9 ±   | 94.8 ±|
|                 | MIA 93.8 ± 1.9  | 85.1 ± 5.7 | 0.4      | 2.6   |
|                 | AIS 96.6 ± 3.4  | 80.1 ± 3.7 |          |       |
| R152            | IAC 92.9 ± 0.8  | 96.2 ± 0.5 | 92.9 ±   | 90.9 ±|
|                 | MIA 95.8 ± 3.1  | 65.7 ± 5.2 | 0.8      | 1.8   |
|                 | AIS 91.6 ± 6.0  | 62.0 ± 4.1 |          |       |
| Inception-v4    | IAC 91.0 ± 1.2  | 93.6 ± 1.1 | 87.6 ±   | 72.9 ±|
|                 | MIA 53.6 ± 16.7 | 49.1 ± 8.8 | 2.0      | 3.5   |
|                 | AIS 10.3 ± 17.8 | 64.1 ± 11.0|          |       |
| Type IV         |                 |            |          |       |
| Guan et al.     | IAC 97.4 ± 0.8  | 98.6 ± 0.5 | 96.9 ±   | 95.9 ±|
|                 | MIA 94.9 ± 2.1  | 84.1 ± 4.2 | 0.8      | 1.8   |
|                 | AIS 90.2 ± 4.0  | 84.1 ± 4.1 |          |       |
| Jin et al.      | IAC 97.8 ± 0.8  | 98.8 ± 0.5 | 97.1 ±   | 96.6 ±|
|                 | MIA 92.5 ± 3.1  | 86.0 ± 3.7 | 0.9      | 1.6   |
|                 | AIS 90.2 ± 6.6  | 84.1 ± 4.9 |          |       |
| Proposed framework | IAC 98.9 ± 0.5  | 99.3 ± 0.2 | 97.8 ±   | 96.8 ±|
|                 | MIA 90.7 ± 0.9  | 87.6 ± 4.6 | 0.4      | 1.9   |
|                 | AIS 88.6 ± 1.1  | 86.7 ± 3.2 |          |       |

The second group of benchmarking methods, type II, follows a framework that combines the features extracted by CNNs and the classical machine-learning classifier. The Inception network,25 ResNet,26 and DenseNet,27 which represent three state-of-the-art deep CNNs, are considered. Their structures are denoted as the DenseNet169 backbone (DB), ResNet152 backbone (RB152), and ResNet50 backbone (RB50), respectively. These combinations are denoted as, for example, “DB + KNN” or “DB + SVM.”

The third group of benchmarking methods, type III, utilizes deep-learning-based transfer learning methods, Inception-v4, R152, and D169, with pretrained weights on ImageNet while all hyperparameters are the same as those in the proposed framework.

The fourth group of benchmarking methods, type IV, represents effective medical imaging classification or segmentation methods that have been applied in other tasks. Guan et al. proposed an attention-guided CNN framework for the thorax disease classification task and achieved state-of-the-art performance on the ChestX-ray1428 dataset.29 Jin et al. compared several state-of-the-art medical imaging segmentation algorithms, and finally chose to combine U-net++30 with CNN to rapidly identify Covid-19 from other lung diseases.31 Two methods are adapted to solve the lung adenocarcinoma classification problem in this study, and their results serve as the benchmark.

As shown in Table 1, the computational results of the four groups of benchmarking methods and our proposed framework are listed according to different evaluation metrics. The proposed framework outperforms all the other methods in terms of overall accuracy and class precision. Moreover, it shows higher robustness based on different folds of data split.

In Figure 5, the speed of convergence for the proposed framework lies between the representation and rebalance branches. This demonstrates that the proposed framework combines the performances of the two branches and prevents overfitting (Figure 6).

Relationship between the two branches

The performances of the representation branch, rebalance branch, and proposed framework are summarized in Table 2. We report two results for the rebalance branch owing to the different executing processes. The rebalance branch with a simple rebalancing method (SR branch) means that the rebalanced data in the rebalance branch are generated by the traditional oversampling method. The rebalance branch with the rebalancing unit (RU branch) indicates that the rebalance data are generated according to the preprocessing stage in our framework. It is observable that the RU branch performs better in terms of overall accuracy and precision in each individual category; this finding demonstrates that the proposed rebalancing unit effectively improves the identification performance in the tail category while avoiding overfitting. The proposed framework outperforms either a single representation branch or a single rebalance branch, as reported in Table 2. These results illustrate that the framework can integrate the advantage of the representation branch in IAC with that of the rebalance branch in MIA and IAS.

The value of \( \alpha \) varies from 0.3 to 0.7, or it can also be dynamically set (Figure S5). This illustrates that the performance of the

6 Patterns 3, 100464, April 8, 2022
extractors and U-net as well as the raw dataset as the input. This study, the classification performances of D169 are compared, based on the input datasets generated by different modules. The best performance appears when both branches are trained.

As shown in Table 3, SROI in black and CRF outperform the raw data. This validates the effectiveness of utilizing the SROI lung extractor and the CRF lung extractor to extract more useful lung regions. However, SROI in white performs slightly worse than the raw data. The probable reason for this is that the white lung section region cannot be distinguished from the white background, which causes difficulties in the identification when the number of samples remains the same as other datasets in Table 3. The performance of the U-net is close to that of the CRF. However, manual labeling masks are requested for U-net, while the CRF extractor can automatically segment the ROI. As shown in Table 4, SROI has a significant advantage in terms of calculation time.

Three datasets generated by the crop-background extractor present three different types of requirements. The first one, \( m = 180 \) and \( r > 100 \), only removes the redundant background. The second one, \( m = 200 \) and \( r > 100 \), removes all of the background together with some white lung regions. The apparent worse performance indicates that the loss of information has a negative impact on identification. The third one, \( m = 230 \) and \( r > 160 \), only retains the central black lung region with the solid portion nodule; its performance is even worse than that of the second one. However, according to the values of the precision and F1 score for each category in the three datasets, the datasets show uniform variations in different categories. The overall performance is maintained at an acceptable level. This phenomenon indicates that the overall skeleton of the information required for identification is retained. Some minor information is lost as \( m \) and \( r \) grow. Through the second and third crop-background datasets, we conjecture that the central black region, especially the solid portion nodule, holds the most vital information for the identification of lung adenocarcinoma categories. The relative size of the solid portion and the white lung region might provide minor information for identification.

**Comparison between the identification results of the CT images and those of pathological examinations**

Figure 7 shows examples of correctly (Figure 7A) and inaccurately (Figure 7B) identified cases, together with the corresponding raw input images, output heatmaps, detected high-risk areas, and probability score for each class. As shown in Figure 7A, the proposed framework can correctly identify the categories of lung CT images with high confidence scores. The detection results also align with the lesion areas shown in the raw input images marked in a red circle by skilled doctors. In Figure 7B, some inaccurate cases occur between IAC and correct categories (AIS/MIA). There is considerable overlap between these cases in the detection and category prediction results. The detected risky areas are not veracious and frequently deviate from the lesion. We observe an interesting phenomenon.
in that their heatmaps appear bicircular, like the lung area. The highest probability scores are both obtained in the incorrect IAC category; however, the correct categories rank second, and the scores are more competitive compared with the scores of the incorrect categories in Figure 7A. The second-highest scores are still in the same order of magnitude as the top-highest ones. From the phenomenon in heatmaps, the framework might have been confused with the images and tried to set the whole lung region as an interesting area but failed; thus, the identification also failed. Another inaccurate case occurs between AIS and the correct category of IAC. The risky area is also quite inaccurate, and the probability scores this time are even closer to each other, showing that the inaccurate risky area will mislead the framework. One possible reason for the misdiagnosis is that the lesion areas in those images are not typical and insufficiently evident because of the excessively low contrast in the MIA and AIS images.

In summary, the proposed framework ensures more accurate identification of lung adenocarcinoma categories compared with other methods and can provide references to experienced radiologists to distinguish lung diseases more effectively.

### Framework validation of other datasets directly or through knowledge distillation

Datasets from different sources over different years are utilized to prove the generalizability of the framework. A detailed description of all datasets can be found in overview of the datasets in experimental procedures. The dataset utilized in the previous discussions is denoted as dataset 1. Dataset 1 is divided randomly into three parts—train, test, and validation (val)—with a split ratio of 60%, 30%, and 10%, respectively. The trained framework is then validated by implementing it in two smaller labeled datasets (denoted as datasets 2 and 3). The experimental design aims to validate the feasibility of the framework on unfamiliar data.

As shown in Table 5, the results are undoubtedly close between the test and val parts of dataset 1, although the val part does not participate in the learning process. Considering these results as the baseline, the performance of the proposed framework degrades slightly on datasets 2 and 3, the most likely reason being that the framework has not been fine-tuned based on each dataset because of the small data volume. However, the high specificity and sensitivity still indicate that the misdiagnosis and missed diagnosis rates are considerably low.

A semi-supervised knowledge distillation (KD) procedure is designed to further demonstrate the performance and transferability of the framework (Figure 8). Dataset 1 is also divided randomly into three parts. The proposed framework trained by the train/test parts serves as the first teacher model M1. The pseudo labels of the unlabeled data (dataset 4) will be generated by M1. M1 is then fine-tuned to obtain the second teacher model M2 by minimizing the labeled loss function $L_{\text{labeled}}$ on labeled data. The student model has the same architecture as the teacher models. Its parameters come from fine-tuning M2 by minimizing the unlabeled loss function $L_{\text{unlabeled}}$ on unlabeled data. Here, $s$ denotes the ground-truth label for the labeled data, $s'$ denotes the prediction of the teacher model for labeled or unlabeled data, and $s''$ denotes the prediction of the student model for labeled or unlabeled data.
Table 3. Identification performance on different ROI datasets (percentage, mean ± SD)

| Dataset     | Class | Precision  | F1  | Accuracy | AUC  |
|-------------|-------|------------|-----|----------|------|
| μ = 180, p > 100 | IAC   | 95.9 ± 0.5 | 97.7 ± 0.1 | 96.0 ± | 94.1 ± |
|             | MIA   | 91.4 ± 5.3 | 79.3 ± 8.2 | 0.3    | 1.5   |
|             | AIS   | 92.8 ± 4.2 | 77.2 ± 4.2 |        |       |
| μ = 200, p > 100 | IAC   | 95.3 ± 0.6 | 97.1 ± 0.3 | 94.5 ± | 93.6 ± |
|             | MIA   | 91.4 ± 3.7 | 78.6 ± 6.4 | 0.5    | 2.1   |
|             | AIS   | 87.4 ± 6.7 | 71.8 ± 2.5 |        |       |
| μ = 230, p > 160 | IAC   | 95.9 ± 0.5 | 97.7 ± 0.1 | 94.0 ± | 92.9 ± |
|             | MIA   | 95.2 ± 0.5 | 95.7 ± 0.1 | 0.5    | 1.1   |
|             | AIS   | 91.4 ± 5.3 | 79.3 ± 8.2 |        |       |
| Raw data    | IAC   | 96.4 ± 0.5 | 98.0 ± 0.3 | 95.9 ± | 94.8 ± |
|             | MIA   | 93.8 ± 1.9 | 85.1 ± 5.7 | 0.4    | 2.6   |
|             | AIS   | 96.6 ± 3.4 | 80.1 ± 3.7 |        |       |
| SROIinwhite | IAC   | 92.8 ± 4.2 | 77.2 ± 4.2 | 95.7 ± | 94.2 ± |
|             | MIA   | 92.8 ± 4.2 | 77.2 ± 4.2 | 0.5    | 1.0   |
|             | AIS   | 89.3 ± 5.9 | 76.0 ± 1.6 |        |       |
| SROIinblack | IAC   | 96.1 ± 0.4 | 97.7 ± 0.1 | 96.3 ± | 94.5 ± |
|             | MIA   | 95.7 ± 3.0 | 83.6 ± 3.9 | 0.4    | 1.0   |
|             | AIS   | 88.2 ± 8.4 | 75.3 ± 2.0 |        |       |
| CRF         | IAC   | 97.2 ± 0.5 | 98.4 ± 0.3 | 96.8 ± | 94.5 ± |
|             | MIA   | 94.9 ± 1.9 | 87.1 ± 4.9 | 0.6    | 3.6   |
|             | AIS   | 92.3 ± 3.4 | 83.7 ± 3.3 |        |       |
| U-net       | IAC   | 96.8 ± 0.5 | 98.3 ± 0.3 | 96.9 ± | 94.4 ± |
|             | MIA   | 97.3 ± 1.5 | 86.7 ± 4.6 | 1.2    | 1.9   |
|             | AIS   | 97.1 ± 2.4 | 82.9 ± 3.2 |        |       |

The computational results of the comparative experiments demonstrated that the proposed framework outperformed three groups of benchmarking methods in terms of overall classification accuracy and precision for each category, especially in the tail categories. The high-risk area heatmap results showed that there was considerable overlap between the heatmap and the solid lesion area detected by skilled doctors, which provided additional evidence to support the theory that the solid portion area strongly contributes to the different categories of lung adenocarcinoma. Furthermore, the results of the evaluation of different image-segmentation algorithms suggest that the relative size of the solid portion and the white lung region are related to adenocarcinoma identification.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

The lead contact for this study is Zijun Zhang: zijzhang@cityu.edu.hk.

Materials availability

This study did not generate new unique materials.

Data and code availability

Code can be accessed at https://github.com/lynnvahh/KDBBN. The latest DOI is https://zenodo.org/badge/latestdoi/454431366. The accession number for the data from the Nanfang Hospital and the explanatory file of data from open access reported in this paper is OSFHOME: https://osf.io/5aqe4/. Data from restricted access cannot be disclosed in our paper in accordance with the Creative Commons license.
Overview of the datasets

Four datasets are utilized in the study. Dataset 1 and dataset 2 were collected from the hospital, while dataset 3 and dataset 4 are composed of samples from open-access online repositories.

Dataset 1

The effectiveness of our proposed framework is mainly evaluated based on a clinical CT image dataset (see Table S1) provided by Nanfang Hospital in China. In total, 2,571 lung section CT images from 520 patients across China were collected before 2019 with the corresponding clinical diagnostic records including the lung adenocarcinoma category, which is diagnosed by experienced doctors. The patients together with the corresponding CT images can be categorized into three categories, IAC, AIS, and MIA, according to the diagnostic results, which are regarded as the ground truth. The CT images were acquired from the CT scans with a resolution of 512 × 512 or 484 × 484 as DICOM format. Since several images cannot provide complete diagnostic information or do not match with the diagnostic records, we finally selected 2,425 labeled images from 488 patients for our experiments. Among these, 2,118 images belong to the IAC category, 153 images belong to MIA, and 154 images belong to AIS.

Because different CT instruments are used, the view of images can be round or square (Figure S4). Inter-class variances of symptoms appearing in lungs are apparent because of different CT appliances, creating extra challenges for the machine vision-based diagnosis.

Dataset 2

To further validate the model, this dataset (see Table S2) was collected by Nanfang Hospital in 2021 through a similar procedure. After excluding substandard data, 670 labeled images from 98 patients were selected. Among these, 542 images belong to the IAC category, 32 images belong to MIA, and 96 images belong to AIS.

Dataset 3

To prove the generalizability of the model, this dataset (seedata and code availability) was collected from different sources. One portion of images are obtained from the data collection Lung Fused-CT-Pathology in The Cancer Imaging Archive (TCIA). Other images were obtained from online repositories and papers. In total there were 267 IAC images, 30 AIS images, and 19 MIA images.

Dataset 4

To increase training data diversity, this dataset was obtained from the TCIA data collection NSCLC-Radiomics-Genomics, CPTAC-LUAD, NSCLC-Radiomics, and APOLLO-5-LUAD. Lepidic-predominant adenocarcinoma samples could be selected based on the clinical records, and the corresponding middle slice CT images were collected as unlabeled data for the later training procedure.

Human research ethics statement

This study involves archival data without disclosing the personal identity or private information, and has obtained the ethics approval from the Human Subjects Ethics Committee at City University of Hong Kong. The reference number of this ethics approval is 8-2021-47-E.

Experiment settings

Before training starts, we perform the standardization and normalization of each image in the original dataset and resize the images to 224 × 224 resolution. Such data preprocessing aims to save computing resources and time so that, in the application, it may run on personal computers rather than high-performance servers. Computational experiments have been conducted to prove that this change of resolution does not impact much on the performance of the proposed framework (Table S3). During the training process, data

| Image Category | IAC | MIA | AIS |
|----------------|-----|-----|-----|
| Original Input Images | ![image](image1.png) | ![image](image2.png) | ![image](image3.png) |
| Heatmaps | ![heatmap1.png] | ![heatmap2.png] | ![heatmap3.png] |
| High Risk Area Detection Results | ![detection1.png] | ![detection2.png] | ![detection3.png] |
| Category Prediction Results | AIS: 4.023e-06, MIA: 7.229e-05, IAC: 9.999e-01 | AIS: 3.133e-07, MIA: 9.999e-01, IAC: 3.842e-07 | AIS: 9.999e-01, MIA: 1.913e-06, IAC: 2.476e-06 |

Table 5. Adenocarcinoma category identification performance of the proposed framework on different test or validation datasets

| Dataset | Class   | Sensitivity | Specificity | Accuracy | AUC  |
|---------|---------|-------------|-------------|----------|-------|
| Dataset 1, test | IAC: 99.9 | 90.3 | 97.9 | 96.9 |
|          | MIA: 84.8 | 99.4 |     |       |
|          | AIS: 93.0 | 99.7 |     |       |
| Dataset 1, val | IAC: 99.7 | 90.4 | 97.9 | 96.9 |
|          | MIA: 85.0 | 99.2 |     |       |
|          | AIS: 93.1 | 99.5 |     |       |
| Dataset 2 | IAC: 99.6 | 91.4 | 97.2 | 96.0 |
|          | MIA: 81.2 | 99.1 |     |       |
|          | AIS: 88.5 | 99.7 |     |       |
| Dataset 3 | IAC: 99.6 | 87.8 | 96.8 | 95.9 |
|          | MIA: 80.0 | 99.3 |     |       |
|          | AIS: 84.2 | 99.3 |     |       |

Figure 7. Examples of the prediction results from the proposed framework

(A and B) The original labels of the CT images diagnosed by skilled doctors through pathological examinations and the corresponding probability scores for three categories. The detected high-risk area by the framework is also shown by the heatmaps and detection results. (A) Correctly identified cases. (B) Inaccurately identified cases.
augmentation strategies are performed on the images by random horizontally flipping and rotating. Computational experiments are conducted to determine the more suitable CNN backbone in each of the two feature-engineering branches as reported in Table S4. DenseNet169 (D169) and ResNet50 (R50) are chosen as the CNN backbones in the representation branch and rebalance branch, respectively. In the model development, the CNN backbones with pre-trained weights on the ImageNet are further optimized by the Adam optimizer with a batch of 16 images per step, while the total training epochs for each backbone is set at 30. Values of the momentum, the learning rate, and the weight decay are set to 0.9, 0.001, and $1 \times 10^{-6}$, respectively. The $z$ is fine-tuned from 0.3 to 0.7, or set dynamically (Figure S5). Finally, it is set to 0.6 based on the best results. $\gamma$ is set to 0.334, 0.333, and 0.333 for IAC, AIS, and MIA. $\beta = (0.2, 0.2, 0.15, 0.15, 0.1)$ respectively correspond to the CRF images, black and white SROI images, and three sets of crop-background images, when (1) $\mu = 180$, $r > 100$, (2) $\mu = 200$, $r > 100$, (3) $\mu = 230$, $r > 160$. The framework is built on Keras 2.1.2 and TensorFlow 1.8.0 and is implemented under the Ubuntu operating system with GPU NVIDIA GeForce RTX 2080Ti.

To confirm the robustness of the proposed framework, we utilize 4-fold cross-validations in the experiment.

**CNN models adopted in the framework**
The architectures of the CNN models adopted in the framework are the same as those in He et al.26 and Huang et al.27

**Implementation**
Algorithms 3 and 4 (Tables S7 and S8) detail the proposed framework in the training and inference phase, respectively. In the training phase, for each training epoch the training dataset is first fed into different extractors in the segmentation unit to generate corresponding segmented image datasets $D_{crf}$, $D_{sroi}$, $D_{crop}$. These datasets are then passed to the uniform sampler or the rebalanced sampler accordingly to produce the representation data $f(x_{ri}, y_{ri})_{n_i=1}^g$ and the rebalanced data $f(x_{ei}, y_{ei})_{n_i=1}^g$ for training, in which suffix $r$ and $c$ respectively represent the representation branch and rebalance branch, while $i$ means the $i$th sample and $n$ represents the total number of samples. After feeding training samples into the representation branch and rebalance branch, respectively, corresponding features $f_r$, $f_c$ and the final weighted feature $f$ are obtained, while the output logits $z$ and the probability distribution for categories $\hat{p}$ can be calculated. At the end of this training epoch, the classification loss function is computed and the weights of the CNN backbones are updated by optimizing the loss function. Finally, after reaching the preset number of epochs in training, the CNN backbones and the classifier together with their weights obtained are then applied in the inference phase.

In the inference phase, the testing dataset is simply fed into the CRF extractor for preprocessing. Processed images are then passed to two branches to obtain the representation feature and rebalance feature, respectively. Through the weighted aggregation of two features, we obtain the final weighted feature $f$. 

![Figure 8. Knowledge distillation procedure](https://example.com/figure8)

The main purpose of designing a knowledge distillation procedure is to better transfer the original framework to other smaller datasets (datasets 2 and 3). The teacher model and student model are both set to be the proposed framework. Besides, to compare different train test partitioning portions without decreasing training data diversity, the procedure is semi-supervised, introducing unlabeled data (dataset 4).

![Figure 9. Area under the ROC curve (AUC) scores on different datasets based on different data-partitioning proportions](https://example.com/figure9)
feature maps for the classification, and the probability distributions are calculated to determine the predicted categories having the largest probabilities.

**Evaluation metrics**

Considering the imbalanced distribution of the dataset, evaluation metrics that can reflect the performance on each category are particularly needed. Therefore, we employ two metrics, Precision and F1-score, to evaluate the performance of identifying each category.

\[
\text{precision}_j = \frac{TP_j}{TP_j + FP_j} \quad \text{(Equation 11)}
\]

\[
F1 - \text{score}_j = 2 \times \frac{TP_j}{2 \times TP_j + FN_j + FP_j} \quad \text{(Equation 12)}
\]

where suffix \(j\) represents the index of the categories (IAC, MIA, or AIS). \(TP_j\), \(FP_j\), \(TN_j\), and \(FN_j\) respectively denote the number of true positive, false positive, true negative, and false negative validation samples in the corresponding index \(j\).

Similarly, we also introduce the sensitivity and specificity to evaluate the misdiagnosis or missed diagnosis rate.

\[
\text{sensitivity}_j = \frac{TP_j}{TP_j + FN_j} \quad \text{(Equation 13)}
\]

\[
\text{specificity}_j = \frac{TN_j}{TP_j + FP_j} \quad \text{(Equation 14)}
\]

We adopt the overall accuracy as the overall performance evaluation. Furthermore, the AUC value, which represents the area under the ROC curve, is also reported to evaluate the property of the classifiers, and the curves are plotted with weighting of each class.

**SUPPLEMENTAL INFORMATION**

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100464.

**ACKNOWLEDGMENTS**

This work was supported in part by the Hong Kong Research Grants Council General Research Fund Project, no. 11204419; in part by the National Natural Science Foundation of China Youth Scientist Fund, no. 52007160; in part by the project of Fundamental Research Funds for the Central Universities and the Youth Teacher International Exchange & Growth Program, no. QNXM20210037; in part by the Joint Fund of National Natural Science Foundation of China with Shenzhen City, no. U1813209; and in part by the Labora-
tory for Al-Powered Financial Technologies Limited.

**AUTHOR CONTRIBUTIONS**

Conceptualization, Z.Z.; methodology, L.C., H.Q., and Z.Z.; software, L.C. and H.Q.; validation, L.C., H.Q., D.L., and Z.Z.; formal analysis, L.C. and H.Q.; investigation, L.C., D.L., J.Z., and K.C.; resources, D.L., J.Z., and K.C.; data curation, L.C. and D.L.; writing—original draft, L.C. and H.Q.; writing—review and editing, Z.Z., L.W., and G.L.; supervision, Z.Z.; project administration, Z.Z. and G.L.; funding acquisition, Z.Z. and L.W.

**DECLARATION OF INTERESTS**

The authors declare no competing interests.

Received: November 17, 2021
Revised: December 15, 2021
Accepted: February 8, 2022
Published: March 3, 2022

**REFERENCES**

1. Siegel, R.L., Miller, K.D., and Jemal, A. (2015). Cancer statistics, 2015. CA: A Cancer J. Clin. 65, 5–29. https://doi.org/10.3322/caac.21254.

2. Behera, M., Owonikoko, T.K., Gal, A.A., Steuer, C.E., Kim, S., Pillai, R.N., Khuri, F.R., Ramalingam, S.S., and Sica, G.L. (2016). Lung adenocarcinoma staging using the 2011 IASLC/ATS/ERS Classification: a pooled analysis of adenocarcinoma in situ and minimally invasive adenocarcinoma. Clin. Lung Cancer 17, e57–e64. https://doi.org/10.1016/j.clc.2016.03.009.

3. Eguchi, T., Kadota, K., Park, B.J., Travis, W.D., Jones, D.R., and Adusumilli, P.S. (2014). The new IASLC-ATS-ERS lung adenocarcinoma classification: what the surgeon should know. Semin. Thorac. Cardiovasc. Surg. 26, 210–222. https://doi.org/10.1053/j.smctvs.2014.09.002.

4. Kawakami, T., Nabeshima, K., Makimoto, Y., Hamasaki, M., Iwasaki, A., Shirakusa, T., and Iwasaki, H. (2007). Micropapillary pattern and grade of stromal invasion in PT1 adenocarcinoma of the lung: usefulness as prognostic factors. Mod. Pathol. 20, 514–521. https://doi.org/10.1038/modpathol.3800765.

5. Wang, X., Chen, H., Gan, C., Lin, H., Dou, Q., Tsougenis, E., Huang, Q., Cai, M., and Heng, P.-A. (2020). Weakly supervised deep learning for whole slide lung cancer image analysis. IEEE Trans. Cybern. 50, 3950–3962. https://doi.org/10.1109/tcyb.2019.2935141.

6. Buzug, T.M. (2011). Computed Tomography (Springer Handbook of Medical Technology), pp. 311–342. https://doi.org/10.1007/978-3-540-74658-4_16.

7. Boland, J.M., Froemming, A.T., Wampfler, J.A., Maldonado, F., Peikert, T., Hyland, C., de Andrade, M., Aubry, M.C., Yang, P., and Yi, E.S. (2016). Adenocarcinoma in situ, minimally invasive adenocarcinoma, and invasive pulmonary adenocarcinoma—analysis of Interobserver Agreement, survival, radiographic characteristics, and gross pathology in 296 nodules. Hum. Pathol. 51, 41–50. https://doi.org/10.1016/j.humpath.2015.12.010.

8. Yanagawa, M., Johkoh, T., Noguchi, M., Mori, E., Shintani, Y., Okumura, M., Hata, A., Fujisawa, M., Honda, O., and Tomyama, N. (2017). Radiological prediction of tumor invasiveness of lung adenocarcinoma on thin-section CT. Medicine 96, e6331. https://doi.org/10.1097/md.000000000006331.

9. Ko, J.P., Suh, J., Ibdapo, O., Escalon, J.G., Li, J., Pass, H., Naichid, D.P., Crawford, B., Tsai, E.B., Koo, C.W., et al. (2016). Lung adenocarcinoma: correlation of quantitative CT findings with pathologic findings. Radiology 280, 931–939. https://doi.org/10.1148/radiol.2016142975.

10. Lim, H.-J., Ahn, S., Lee, K.S., Han, J., Shin, Y.M., Woo, S., Kim, J.-H., Yie, M., Lee, H.Y., and Yi, C.A. (2013). Persistent pure ground-glass opacity lung nodules \( \geq 10 \) mm in diameter at CT Scan. Chest 144, 1291–1299. https://doi.org/10.1378/chest.12-2987.

11. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.319.

12. Buda, M., Maki, A., and Mazurowski, M.A. (2018). A systematic study of the class imbalance problem in Convolutional Neural Networks. Neural Network. 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011.

13. Huang, C., Li, Y., Loy, C.C., and Tang, X. (2016). Learning deep representation for imbalanced classification. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.580.

14. Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). Class-balanced loss based on effective number of samples. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2019.00949.

15. Fernandez, A., Garcia, S., Herrera, F., and Chawla, N.V. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J. Artif. Intell. Res. 67, 863–905. https://doi.org/10.1613/jair.1.11192.
