IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 21, 2020, accepted November 16, 2020, date of publication November 24, 2020,
date of current version December 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040245

A Two-Dimensional Sparse Matrix Profile DenseNet for COVID-19 Diagnosis Using Chest CT Images

QIAN LIU[1,2], CARSON K. LEUNG[2], (Senior Member, IEEE), AND PINGZHAO HU[1,2,3]
[1]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, MB R3E 0J9, Canada
[2]Department of Computer Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
[3]Research Institute in Oncology and Hematology, CancerCare Manitoba, Winnipeg, MB R3E 0J9, Canada

Corresponding author: Pingzhao Hu (pingzhao.hu@umanitoba.ca)

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

ABSTRACT COVID-19 is a newly identified disease, which is very contagious and has been rapidly spreading across different countries around the world, calling for rapid and accurate diagnosis tools. Chest CT imaging has been widely used in clinical practice for disease diagnosis, but image reading is still a time-consuming work. We aim to integrate an image preprocessing technology for anomaly detection with supervised deep learning for chest CT imaging-based COVID-19 diagnosis. In this study, a matrix profile technique was introduced to CT image anomaly detection in two levels. At one-dimensional level, CT images were simply flatted and transformed to a one-dimensional vector so that the matrix profile algorithm could be implemented for them directly. At two-dimensional level,a matrix profile was calculated in a sliding window way for every segment in the image. An anomaly severity score (CT-SS) was calculated, and the difference of the CT-SS between the COVID-19 CT images and Non-COVID-19 CT images was tested. A sparse anomaly mask was calculated and applied to penalize the pixel values of each image. The anomaly weighted images were then used to train standard DenseNet deep learning models to distinguish the COVID-19 CT from Non-COVID-19 CT images. A VGG19 model was used as a baseline model for comparison. Although extra finetuning needs to be done manually, the one-dimensional matrix profile method could identify the anomalies successfully. Using the two-dimensional matrix profiling method, CT-SS and anomaly weighted image can be successfully generated for each image. The CT-SS significantly differed among the COVID-19 CT images and Non-COVID-19 CT images ($p - value < 0.05$). Furthermore, we identified a potential causal association between the number of underlying diseases of a COVID-19 patient and the severity of the disease through statistical mediation analysis. Compared to the raw images, the anomaly weighted images showed generally better performance in training the DenseNet models with different architectures for diagnosing COVID-19, which was validated using two publicly available COVID-19 lung CT image datasets. The metric Area Under the Curve(AUC) on one dataset were 0.7799(weighted)vs. 0.7391(unweighted), 0.7812(weighted) vs. 0.7410(unweighted), 0.7780(weighted) vs. 0.7399(unweighted), 0.7045(weighted) vs. 0.6910(unweighted) for DenseNet121, DenseNet169, DenseNet201, and the baseline model VGG19, respectively. The same trend was observed using another independent dataset. The significant results revealed the critical value of using this existing state-of-the-art algorithm for image anomaly detection. Furthermore, the end-to-end model structure has the potential to work as a rapid tool for clinical imaging-based diagnosis.

INDEX TERMS Rare pattern mining, matrix profile, COVID-19 CT images, risk score, DenseNet, mediation analysis.

I. INTRODUCTION

Unsupervised anomaly detection using rare pattern mining is one of the most intuitive medical imaging–based disease diagnosis methods [1]. People without medical expertise could find an obvious lesion in a medical image if the lesion is extremely different from other parts in the image. Actually, even radiologists also read the images in that way. As the most obvious lesion is so conspicuous that it can be noticed immediately by radiologists at their first glance of the image.

The associate editor coordinating the review of this manuscript and approving it for publication was Yudong Zhang.

This step is unsupervised, and it depends only on the intrinsic information in the image itself. Then using the normal human sectional anatomy knowledge, radiologists could further tell whether this lesion is critical or not [2].

COVID-19 is a newly identified disease that is very contagious and has been rapidly spreading across different countries around the world [3]. Common symptoms from COVID-19 are fever, dry cough, but in more serious cases, patients can experience difficulty in breathing [4]. At present, the NA-PCR (Nucleic Acid Polymerase Chain Reaction) testing is considered as the most effective, cheap and rapid detection method of COVID-19. However, a bottleneck to use this technique is that there are a short of supplies of NA-PCR in some countries [5]. Several alternative methods have been considered for individuals to test positive for COVID-19, including CT (computed tomography) scans of the lungs. Lung CT scanning is fast and easy to detect COVID-19. As the number of infected patients increases exponentially, it can be hard to provide testing scans for patients because of the limited number of doctors. It is recommended that artificial intelligence (AI) systems can be developed to analyse the lung CT scans of patients to determine COVID-19 status [6].

To build the AI system, we used a two-step strategy: detection and enhancement of CT image anomaly and modelling of the anomaly enhanced CT images. For the first step, we used a naïve two-dimensional sliding window approach to calculate the matrix profile of the image. Summing up this matrix profile could make an image-specific severity score (SS) indicating the severity of the image anomalous. This CT-SS can be computed automatically as compared to the manually calculated CT-SS [7], [8], and it has the potential to rapidly identify COVID-19 patients. At the same time, the matrix profile could be easily used to generate a salience map [9] for each CT image to detect lung anomaly. A saliency map is a topographic map that represents visual saliency of an image [10]. Overlapping the image and its salience map could further give us a weighted CT image to enhance the anomaly; For the second step, the weighted CT images could be input into a deep convolutional neural network for further classification or regression tasks. This naïve two-dimensional method is easy to apply, but the benefits of those ultra-fast Fourier algorithms developed by Keogh *et al.* [11] are lost in this situation. To speed up the calculation, raw images were pooled to a lower resolution and the stride of the sliding window was set to the same as the length of the image segment. In this way, the nearest neighbour, matrix profile, and deep learning technologies were effectively integrated together. The proposed algorithm was tested using two publicly available COVID-19/Non-COVID-19 lung CT image sets [12], [13], respectively. Please be noted that the Non-COVID-19 group contains the images from both healthy controls and other types of lung disease cases.

The main contribution of this study could be divided into two parts. The first one is the innovative application of a classic low-dimensional time-series rare pattern mining technique to unsupervised high-dimensional medical imaging anomaly detection. Another contribution is the application of dense deep learning networks for COVID-19 diagnosis using the anomaly enhanced CT images.

II. RELATED WORK

Unsupervised rare pattern mining in medical imaging has been proved to be beneficial [14], but most of the classic rare pattern mining techniques (Apriori [15], [16] based and FP-Growth [17] based) cannot be directly applied to image data as it is two dimensional with space-related information [18]. There are several methods developed for unsupervised image anomaly detection. According to Ehret *et al.*, these unsupervised methods could be classified as nearest neighbour-based anomaly detection, clustering-based anomaly detection, statistical anomaly detection, spectral anomaly detection, and information theoretic anomaly detection [19]. In fact, if applied in a static image situation, they all belong to the first category to some extent [1], since they all measure certain distances and try to identify the discord distances of data instances. The assumption of the nearest neighbour-based anomaly detection in a static image is that normal pixel segments are always similar to each other. Therefore they have a close distance with their nearest neighbours, while anomalies are dislike their closest neighbours [20].

If the aforementioned assumption of the nearest neighbour-based anomaly detection holds, then the problem of image anomaly detection can be transformed as a problem of scanning of a given segment (or window) of images and the retrieval of the nearest neighbours of the scanned segments. This is exactly the same as the definition of a similarity join problem defined by Yeh *et al.* [11]: given a collection of data objects, retrieve the nearest neighbours for every object. To solve the problem, Keogh *et al.* proposed a new data structure called matrix profile and developed a series of matrix profile based algorithms to solve the similarity join problem for time series data [11], [21]–[24]. A matrix profile consists of two components: a distance profile and a profile index. The distance profile is a vector of minimum Euclidean distances among the subsequences within the time-series. The profile index contains the index of subsequences' first nearest neighbours. In other words, the profile index is the location of a subsequence's most similar subsequence. In order to apply the matrix profile technique to large databases, Fast Fourier transform was introduced to make the matrix profile algorithm ultra-fast, therefore it can be applied to big time-series data without sacrificing the time efficacy [11], [25]. These algorithms were proved to be efficient for one-dimensional time series data. However, matrix profile technique has not been introduced to high dimensional data, such as two-dimensional image data.

Besides the limitation of matrix profile in high-dimensional data, there are also lack of studies exploring its integration with advanced machine learning techniques. Although nearest neighbour is a successful long-standing technique, and the matrix profile provides us a strategy using

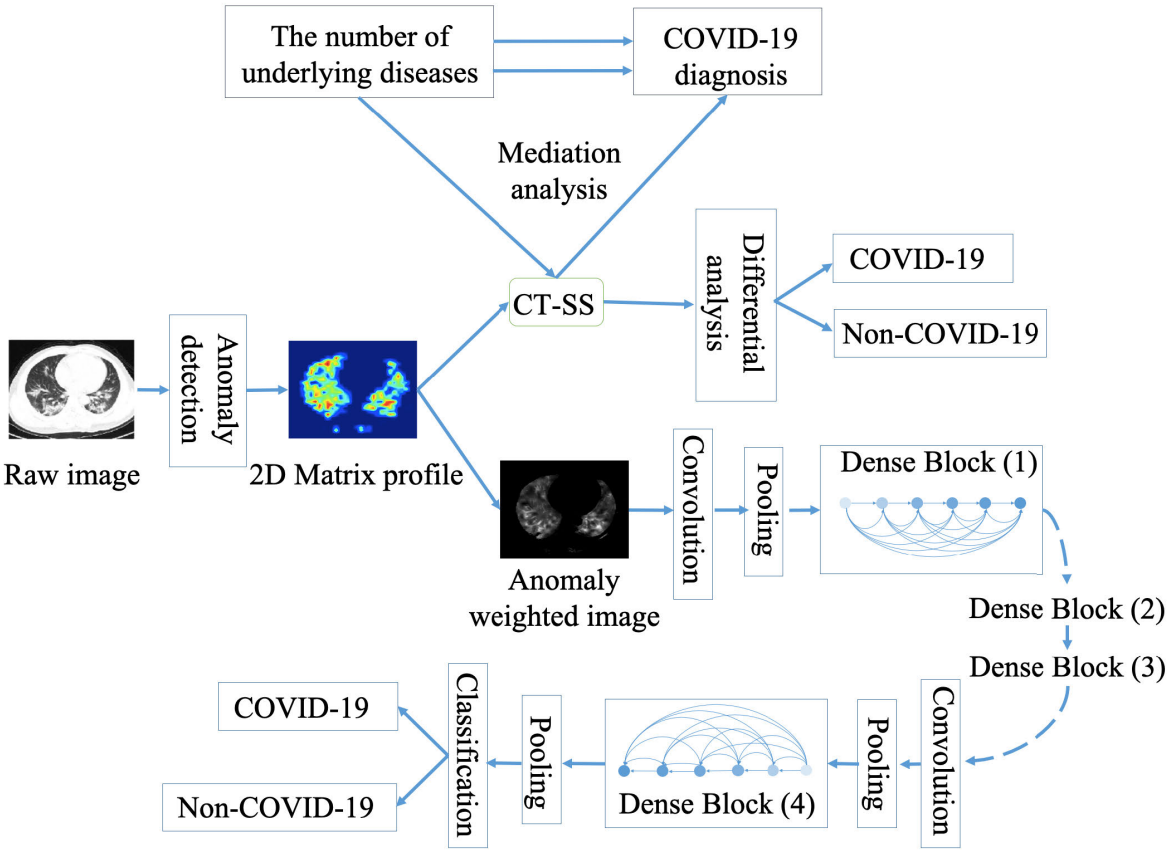


FIGURE 1. Workflow of the proposed study. CT-SS refers to CT-severity score.

nearest neighbour technique for the anomaly detection, it has not been well integrated with current advanced deep learning techniques. A recent work designed a model to use nearest neighbour for identifying anomaly at the image set level [26]. In their model, a set of normal images were input into a deep learning-based feature extractor for building a feature library. Once a new image is arrived, it would undergo the same feature extractor. Then the nearest distances of these features extracted from the new image with those features stored in the feature library will be computed. By verifying if the distance is larger than a predefined threshold, they could determine if the new image is normal or anomalous [26], [27]. However, as we mentioned, this design is at image set level which could not detect the segment level anomalies. Moreover, it is actually a semi-supervised approach as it needs the label information to build the feature library.

CT-severity score(CT-SS) was proposed by a study for COVID-19 rapid diagnosis [7]. In order to obtain the CT-SS, the authors need to manually measure and access the ground-glass opacity, interstitial opacity, and air trapping ratio of the lungs in the CT images. These three features are typical pneumonia symptoms, and different doctors may obtain different values of them even using the same image. Therefore, although the CT-SS has been proved to be significantly associated with the COVID-19 severity, the extra manual measurement burden added to the radiologists and the potential bias of their expertise limits its application in clinic practice. Another consideration is the lack of theoretical analysis of why CT-SS is associated with COVID-19 severity. It was reported that medical image phenotypes, such as CT-SS, could work as the mediators of genetic variations or other basic clinical characteristics' effects on disease outcomes [28], [29]. Therefore, mediation analysis could be used to test the significance of the indirect causal relationship among patient's clinical characteristics, CT-SS, and COVID-19 severity.

Although there are no many studies on deep learning nearest neighbour-based image anomaly detection, supervised deep learning has been widely involved in COVID-19 diagnosis. There were a lot of deep convolutional neural networks (CNN) proposed by different studies on COVID-19 diagnosis [12], [13], [30]–[33]. In this study, we propose a state-of-the-art end-to-end matrix profile - based DenseNet [34] model for COVID-19 diagnosis. We also compare the performance of different DenseNet architectures and the basic convolutional architecture called VGG (Visual geometry Group Network) [35].

III. DATA AND METHODS

The whole workflow is summarized in **Fig. 1**. The proposed anomaly detection algorithm first preprocesses the raw CT images using the matrix profile technique, the CT-SS and the

anomaly weighted images are then calculated. For the CT-SS, differential analysis and mediation analysis are performed to explore its potential application to diagnosis for COVID-19 and its clinical interpretability. The anomaly weighted images are finally applied to build DenseNet-based deep learning models for COVID-19 diagnosis.

A. ANOMALY DETECTION

Assume a chest CT image can be defined as a matrix P of pixel values p_{ij},

$$P = \begin{Bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nm} \end{Bmatrix} \tag{1}$$

where m is the width of P, and n is the height of P. i ranges from 1 to m, and j ranges from 1 to n. In principle, there are two approaches to detect anomalies in the chest CT image. The first method is to flatten the image matrix P into a long vector. The vector could be treated as a time series, thus those well-developed algorithms for the time series analysis could be applied easily. The flattened operation could be done along the row P_{row} as shown in Equation (2) or the column P_{col} as shown in Equation (3) of the P.

$$P_{row} = p_{11}, p_{12}, \cdots, p_{1m}, \cdots, p_{n1}, p_{n2}, \cdots, p_{nm} \tag{2}$$

$$P_{col} = p_{11}, p_{21}, \cdots, p_{n1}, \cdots, p_{1m}, p_{2m}, \cdots, p_{nm} \tag{3}$$

For the flat image P_{flat} (P_{row} or P_{col}), we can apply the ultra-fast Fourier transform algorithms to speed up the calculation of the matrix profile to detect anomalies. The detailed description of the algorithms can be found in the original papers [11], [21]–[24]. After the anomalies are detected, we could trace the anomalies back to the position in the matrix P by joining them across the rows and columns. In this way, the two-dimensional anomaly detection problem is transferred into two one-dimensional anomaly detection problems. We could think of this as scanning the image along two directions in a greedy snake way [36]. Then we can find the overlapped anomalies detected by these two greedy snakes. This is the proposed one-dimensional method to calculate the matrix profile for a image.

The second method is to find the local anomaly regions or two-dimensional segments of the image P directly. We define a segment $P_{ij,wh}$ of P as a matrix, which has a size of $w \times h$ and starts from P_{ij} as shown in Equation (4).

$$P_{ij,wh} = \begin{pmatrix} p_{ij} & \cdots & p_{i(j+w-1)} \\ \vdots & \ddots & \vdots \\ p_{(i+h-1)j} & \cdots & p_{(i+h-1)(j+w-1)} \end{pmatrix} \tag{4}$$

We define a sparse segment set S as shown in Equation (5) of the P as an ordered set of sparsely selected segments of the P obtained by a sliding window of size $w \times h$ and a stride s across P. Where S_{ij} could be used to denote $P_{ij,wh}$.

$$S = \begin{pmatrix} p_{11,wh} & p_{1(1+s),wh} & \cdots & p_{1(m-w+1),wh} \\ p_{(1+s)1,wh} & p_{(1+s)(1+s),wh} & \cdots & p_{(1+s)(m-w+1),wh} \\ \vdots & \vdots & \ddots & \vdots \\ p_{(n-h+1)1,wh} & p_{(n-h+1)(1+s),wh} & \cdots & p_{(n-h+1)(m-w+1),wh} \end{pmatrix} \tag{5}$$

We define a sparse two-dimensional matrix profile ($2DM$) as a matrix of the Euclidean distances between each segment $P_{ij,wh}$ in the sparse segments set S and its nearest neighbours in S. The $2DM$ has the same size as S, but the elements in S are matrices while the elements in $2DM$ are numbers. To calculate $2DM$, the pairwise Euclidean distance between one element in S with every other element in S will be calculated. The minimum of these distances will be stored in the same position of $2DM$ as the element in S. According to the assumption of nearest neighbour-based anomaly detection in a static image, a segment that has the smaller nearest distance will probably be a normal segment,while a segment that has the larger nearest distance will probably be an anomaly. Therefore, the value of $2DM$ could represent the anomaly level of the segments in S. The algorithm of building the $2DM$ is shown in **Algorithm 1**.

Algorithm 1 Calculate $2DM$

Input: an image P, window size $w \times h$,stride s
Output: a matrix profile $2DM$

```
2D_M ← inf
for S_ij in SlidingWindow(P, s, w): do
    for S_i'j' in SlidingWindow(P, s, w): do
        distance ← EuclideanDistance(S_ij, S_i'j')
        if distance < 2DM_ij then
            2DM_ij ← distance
        else
            PASS
        end if
    end for
end for
```

After $2DM$ is calculated, the values in the $2DM$ matrix are summed [8] and scaled to a range of 0 to 100 as CT-SS of the image. The difference of the CT-SS between the patient groups are tested using student t-test. And a statistic mediation analysis [37] is performed to identify the indirect effects of age, gender, and underlying diseases on COVID-19 severity through the two-dimensional matrix profile based CT-SS using the R package "mediation". In the mediation analysis model, the COVID-19 severity and CT-SS are treated as dependent variable and mediator,separately. While the age, gender, and the number of underlying diseases(how many underlying diseases the patient has) are treated as independent variable, separately. There are three steps for conducting the mediation analysis. The first step is three simple regression analyses with the dependent variable of COVID-19 severity and the independent variable of the age,

TABLE 1. VGG and DenseNet architectures * used in this study.

VGG19	DenseNet121	DenseNet169	DenseNet201
$3 \times 3\, conv^* \times 2$	$7 \times 7\, conv$	$7 \times 7\, conv$	$7 \times 7\, conv$
$max\ pool$	$3 \times 3\, max\ pool$	$3 \times 3\, max\ pool$	$3 \times 3\, max\ pool$
$3 \times 3\, conv \times 2$	$\begin{Bmatrix} 1 \times 1\, conv \\ 3 \times 3\, conv \end{Bmatrix} \times 6$	$\begin{Bmatrix} 1 \times 1\, conv \\ 3 \times 3\, conv \end{Bmatrix} \times 6$	$\begin{Bmatrix} 1 \times 1\, conv \\ 3 \times 3\, conv \end{Bmatrix} \times 6$
$max\ pool$	$1 \times 1\, conv$	$1 \times 1\, conv$	$1 \times 1\, conv$
	$2 \times 2\, average\ pool$	$2 \times 2\, average\ pool$	$2 \times 2\, average\ pool$
$3 \times 3\, conv \times 4$	$\begin{Bmatrix} 1 \times 1\, conv \\ 3 \times 3\, conv \end{Bmatrix} \times 12$	$\begin{Bmatrix} 1 \times 1\, conv \\ 3 \times 3\, conv \end{Bmatrix} \times 12$	$\begin{Bmatrix} 1 \times 1\, conv \\ 3 \times 3\, conv \end{Bmatrix} \times 12$
$max\ pool$	$1 \times 1\, conv$	$1 \times 1\, conv$	$1 \times 1\, conv$
	$2 \times 2\, average\ pool, stride\ 2$	$2 \times 2\, average\ pool, stride\ 2$	$2 \times 2\, average\ pool, stride\ 2$
$3 \times 3\, conv \times 4$	$\begin{Bmatrix} 1 \times 1\, conv \\ 3 \times 3\, conv \end{Bmatrix} \times 24$	$\begin{Bmatrix} 1 \times 1\, conv \\ 3 \times 3\, conv \end{Bmatrix} \times 32$	$\begin{Bmatrix} 1 \times 1\, conv \\ 3 \times 3\, conv \end{Bmatrix} \times 48$
$max\ pool$	$1 \times 1\, conv$	$1 \times 1\, conv$	$1 \times 1\, conv$
	$2 \times 2\, average\ pool, stride\ 2$	$2 \times 2\, average\ pool, stride\ 2$	$2 \times 2\, average\ pool, stride\ 2$
$3 \times 3\, conv \times 4$	$\begin{Bmatrix} 1 \times 1\, conv \\ 3 \times 3\, conv \end{Bmatrix} \times 16$	$\begin{Bmatrix} 1 \times 1\, conv \\ 3 \times 3\, conv \end{Bmatrix} \times 32$	$\begin{Bmatrix} 1 \times 1\, conv \\ 3 \times 3\, conv \end{Bmatrix} \times 32$
$max\ pool$	$7 \times 7\, global\ average\ pool$	$7 \times 7\, global\ average\ pool$	$7 \times 7\, global\ average\ pool$
$FC^* \times 3,\ softmax$	$FC,\ softmax$	$FC,\ softmax$	$FC,\ softmax$

$*This\ table\ is\ modified\ from\ the\ table\ in\ DenseNet\ original\ paper.$ [34]

$*conv = A\ sequence\ of\ batch\ normalization, ReLu, Convolution\ layers.$

$*FC = Fully\ connected\ layer$

gender and the underlying diseases, respectively. The second step is also three simple regression analyses predicting the mediator, which is the two-dimensional matrix profile based CT-SS, from the age, gender and the underlying diseases, respectively. The third step is three multiple regression analyses predicting the dependent variable of COVID-19 severity from the CT-SS and age, CT-SS and gender, CT-SS and the underlying diseases, respectively.

An up-sampling step is performed to impute $2DM$ to the same size of the image P. In this way, an anomaly mask is made for P. This map is actually a salience map if we plot it on top of the raw image P. The anomaly map is then used as a weight matrix to be fed into a simple linear model for making a weighted image P_w, which can potentially enhance the anomaly in the raw CT image P.

$$P_w = P + 2DM \cdot P \tag{6}$$

Here, the 2DM is a matrix with the same dimension of the raw image P. We calculate the dot product of these two matrices (2DM and P). Then we add P with the product. P_w is then passed to a classification-based deep learning model for model training and testing.

B. VGG AND DenseNet

We treat the lung CT imaging-based diagnosis of COVID-19 as a binary classification problem (e.g. COVID-19 or Non-COVID-19). VGG and DenseNet model are applied to perform the classification. VGG has 16 convolutional layers and 3 fully connected layers and won the 2014 Large Scale Visual Recognition Challenge. It is a CNN model with a deeper architecture by increasing the number of convolutional layers and reducing the size of convolutional layers [35].

DenseNet is a relatively new framework of convolutional deep learning. The idea of DenseNet is to build a deeper architecture which has connections between each convolutional layer to every other layer within the same dense block in a feed-forward fashion. Unlike the ResNet [38], the connections of DenseNet are in feature-level instead of weight-level. The parameters of each layer will be trained only once, and the resulted feature-maps will be concatenated together as the input of the layer they connect to. In this way, the weights could be more efficient, and the gradients would not be vanished. The performance of DenseNet has been estimated on several benchmark datasets [34]. With different number of convolutional layers in each dense block, DenseNet could have different settings. The three architectures of DenseNet (DenseNet121, DenseNet169, and DenseNet201) used in this study are listed in **Table 1**.

The anomaly weighted images P_w are then used in the training, validation and testing of the above mentioned VGG and DenseNet models.

C. CHEST CT DATASETS

This proposed workflow of the deep learning models were applied to analyse the weighted (P_w) and unweighted (P) raw lung CT images, respectively. The raw data used in this study came from two publicly available datasets. One was downloaded from a GitHub repository published by The University

TABLE 2. Data splits used in this study.

		Non-COVID-19	COVID-19	Total
Dataset 1*	Training	146	183	329
	Validation	15	57	72
	Testing	34	35	69
Dataset 2*	Training	7984	3201	11185
	Validation	997	400	1397
	Testing	998	400	1398

$*Dataset\,1 = The\,University\,of\,California\,San\,Diego\,dataset;$

$*Dataset\,2 = Wuhan\,Huazhong\,University\,of\,Science\,and\,Technology\,dataset.$

of California San Diego [12]. It contains 275 COVID-19 lung CT images and 195 Non-COVID-19 lung CT images. This dataset was built by reading the captions of the published papers about COVID-19. The author of the dataset manually searched for quite a number of COVID-19 CT imaging papers, and copied the CT images contained in those papers as figures. The label information of these CT images, such as whether they were obtained from COVID-19 patients or Non-COVID-19 patients, was collected by reading the captions of the figures in those papers. The split of the training, testing, and validation of this dataset followed the authors' suggestion (**Table 2**).

We did not borrow the data augmentation and transfer learning steps as done by the authors of the dataset and our training strategy is relatively simple in terms of the training epochs and model structure because the goal of this work is to test the effect of the anomaly detection-based image preprocessing. The other dataset was published by Wuhan Huazhong University of Science and Technology [13]. This one has 4,001 COVID-19 lung CT images and 9,979 Non-COVID-19 lung CT images. The quality of this dataset is better than the first one since it was directly obtained from Wuhan's hospitals. The images are all in DICOM (Digital Imaging and Communications in Medicine) format with the similar FOV (Field of View) and resolution (200K). We borrowed the lung parenchyma splitting algorithm from the author of the dataset to split the lung regions from the other body parts [13]. After this, we randomly selected 80% of the images as a train set, 10% as a validation set, and 10% as a test set. The detailed number of images are listed in **Table 2**. We choose this data split strategy to be consistent with the split strategy of the first data set. Images from both datasets are resized to a uniform resolution of 224 × 224. Training parameters are kept the same in both the anomaly detection-based framework and anomaly detection-removed framework.

D. PERFORMANCE EVALUATION

To evaluate our model performance, we used below performance measures: *Accuracy* (a ratio of correctly predicted observations to the total observations), *Precision* (a ratio of correctly predicted positive observations to the total predicted positive observations), *Recall* (also called sensitivity, the ratio of correctly predicted positive observations to the all observations in actual class), *AUC* (Area Under the Curve) and $F1$ (a weighted average of Precision and Recall). All performance metrics of anomaly weighted images were stored in a vector, while all performance metrices of raw images were stored in another vector. Then a t-test was done to test the significance between these two vectors.

E. IMPLEMENTATION OF THE ALGORITHM

We made the data splits with our code publicly available for reproducing our results (https://github.com/qianliu1219/iMP). The raw data used in this study could be downloaded from UCSD (https://github.com/UCSD-AI4H/COVID-CT) and ICTCF (http://ictcf.biocuckoo.cn/index.php). The proposed two-dimensional matrix profile algorithm and the training of the VGG model and the three DenseNet models on the two datasets took around 80 hours for a Nvidia GeForce GTX 1080 GPU machine.

IV. RESULTS

A. THE SPARSE MATRIX PROFILE AND CT-SS

After applying the ultra-fast matrix profile algorithm to the one-dimensional flattened images, we could obtain meaningful patches, which indicate the potential anomaly pixels in the chest CT images. **Fig. 2** showed one of the examples.

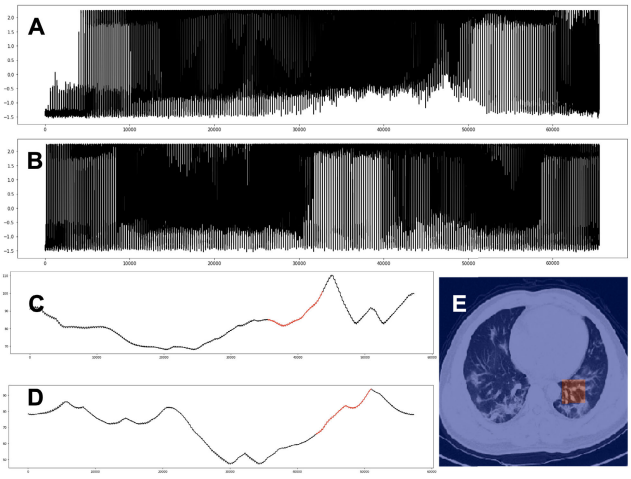


FIGURE 2. Examples of sparse matrix profile. The row flattened image (A) and the column flattened image (B). The one-dimensional matrix profiles of A and B were plotted as black lines (C, D). The meaningful rare patterns were highlighted using the red colours. Their overlap was traced back to the raw image. A meaningful anomaly patch was observed (E).

However, the top discords (top smallest distances) require to be carefully defined in order to find a meaningful patch using the one-dimensional matrix profile algorithm. If we

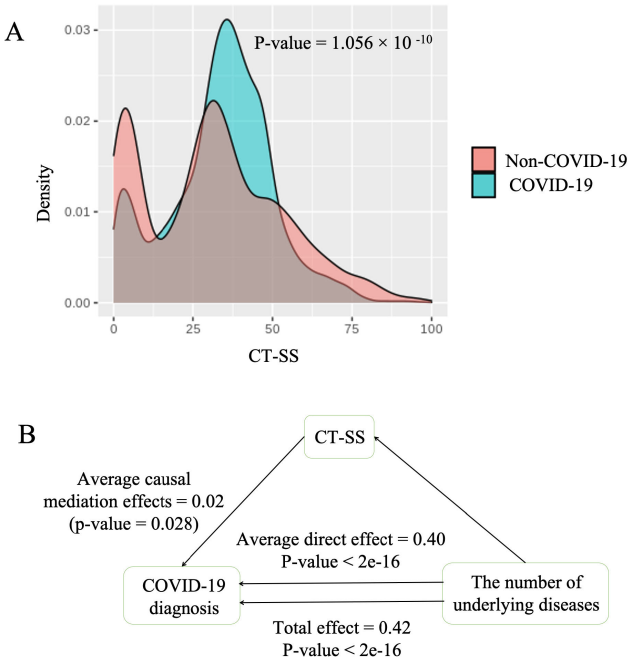


FIGURE 3. **The results of CT-SS differential analysis and mediation analysis. Top panel (A) is the density distributions of the two-dimensional matrix profile algorithm-based CT severity scores (CT-SS) for COVID-19 and Non-COVID-19 groups, respectively. Bottom panel (B) is the causal association analysis among the underlying diseases, CT-SS, and COVID-19 diagnosis. The total effect and its significance of underlying diseases on COVID-19 diagnosis, the direct component of the total effect and indirect effect through the CT-SS are showing on the paths.**

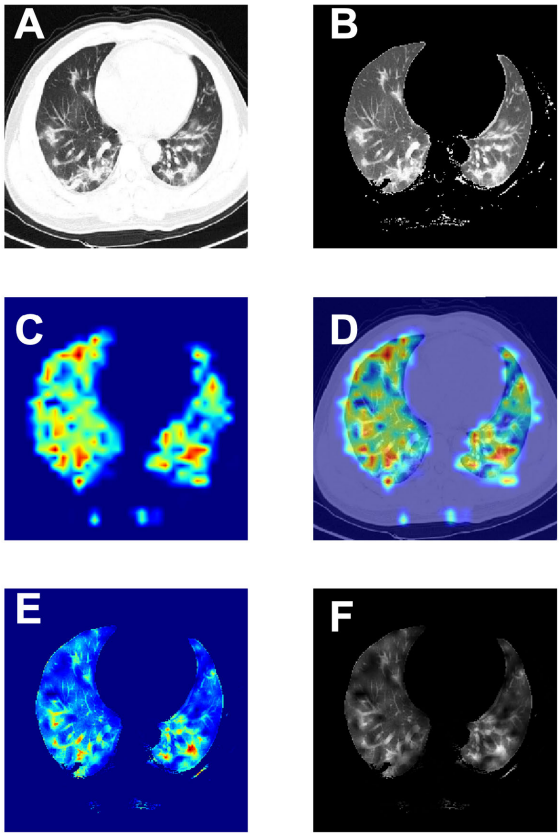


FIGURE 4. **Examples of generation of salience maps. The raw image (A), the lung regions (B), identified two-dimensional matrix profile heatmap (C,D), and anomaly weighted image (E,F). In the weighted image, valuable pixels in the lung regions and lesion regions were highlighted, while meaningless regions were suppressed.**

select the top 1 discord, the highlighted patch sometimes would present in the corner or along the edge of the image or body parts in the image. Those meaningless discords need to be manually filtered out, which is inconvenient. Therefore, the approach is not practical in clinical reality.

Using the two-dimensional matrix profile algorithm, we got an anomaly severity score (CT-SS) for each image. The density distribution of the CT-SS is shown in **Fig. 3A**. These CT-SS were significantly different between the COVID-19 group and Non-COVID-19 group ($p - value < 0.05$).

The mediation analysis identified a significant causal relationship among the number of underlying diseases, CT-SS, and COVID-19 severity as shown in **Fig. 3B**. The number of underlying diseases has a total effect of 0.42 ($p - value < 0.05$) on COVID-19 severity. Five percentage of this effect, which is 0.02 ($0.02 \div 0.42 = 5\%$), could be explained by the two-dimensional matrix profile based CT-SS. The rest 95% is the direct effect that needs to be explained by other mechanisms.

Based on the two-dimensional matrix profile, we could get a salience map pasted on the top of the raw image to suppress the meaningless pixel values without losing information in meaningful regions such as the lung region and lesion region (**Fig. 4**).

B. THE ANOMALY WEIGHTED LUNG CT IMAGES IMPROVED THE COVID-19 DIAGNOSIS

We trained the VGG and DensNet models (**Table 1**) using both the training sets of raw CT images and the anomaly weighted CT images for the Datasets 1 and 2 (**Table 2**), respectively, which were evaluated using their validation sets (**Table 2**). As shown in **Fig. 5 (Top)**, due to the small sample size in Dataset 1, the performance of DenseNet121 on the raw images and the anomaly weighted images is inconsistent for different performance measures. Also due to the poor quality (in terms of both the format (Normal image format, not medical image format) and the various of image resolutions(from 9K to 1.6M)) and small sample size, models trained on Dataset 1 need more training epochs to converge than models trained on Dataset 2. Except accuracy, the anomaly weighted images have better performance than the raw images for all other four performance measures in DenseNet121. For Dataset 2, the anomaly weighted images have shown better performance than the raw images for all the five performance measures in **Fig. 5 (Bottom)**. Other DenseNet architectures as well as the VGG network as shown in **Table 1** also showed the similar trend (results were not shown). The overall performance winner was always the model trained with the anomaly

TABLE 3. Classification performance on testing sets.

		VGG19		DenseNet121		DenseNet169		DenseNet201	
		With anomaly detection	Without anomaly detection	With anomaly detection	Without anomaly detection	With anomaly detection	Without anomaly detection	With anomaly detection	Without anomaly detection
Dataset 1	Accuracy	**0.7700**	0.7446	0.7810	**0.8095**	**0.7931**	0.7702	**0.7807**	0.7792
	Precision	**0.6318**	0.6109	**0.6777**	0.6204	**0.7019**	0.6734	**0.6930**	0.6800
	Recall	**0.6993**	0.6799	**0.7257**	0.7025	0.7111	**0.7271**	**0.7319**	0.7222
	F1	**0.6638**	0.6436	**0.7008**	0.6589	**0.7065**	0.6992	**0.7119**	0.7005
	AUC	**0.7301**	0.7012	**0.7799**	0.7391	**0.7812**	0.7410	**0.7780**	0.7399
Dataset 2	Accuracy	**0.9590**	0.9279	**0.9735**	0.9413	**0.9638**	0.9803	**0.9788**	0.9591
	Precision	**0.9628**	0.9691	**0.9752**	0.9712	**0.9806**	0.9693	**0.9799**	0.9665
	Recall	**0.9787**	0.9497	**0.9880**	0.9458	**0.9837**	0.9351	**0.9914**	0.9559
	F1	**0.9707**	0.9593	**0.9816**	0.9583	**0.9821**	0.9518	**0.9856**	0.9611
	AUC	**0.9809**	0.9721	**0.9957**	0.9844	**0.9996**	0.9801	**0.9991**	0.9849

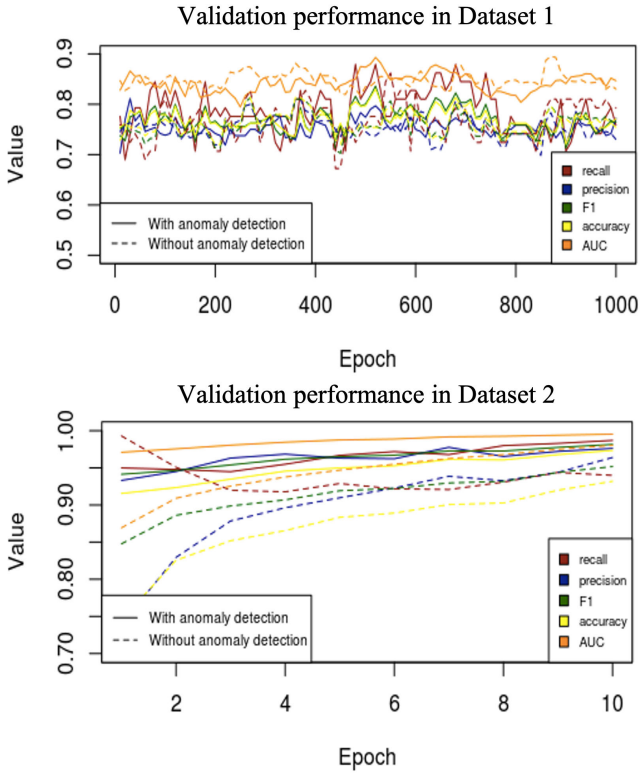


FIGURE 5. The validation performances of the DenseNet121 models on the Datasets 1 and 2.

weighted images instead of the model trained with the raw images ($p - value < 0.05$) as shown in **Table 3**.

The trained DenseNet models and the VGG model were applied to the testing sets of the Datasets 1 and 2, respectively. We showed the model performance on the testing sets in **Table 3**. We also showed the performance of DenseNet121 on the validation set as example **Fig. 5**. The performance on validation set is relatively higher than that of the testing set (**Table 3**) in the Dataset 1, which means there is a potential over-fitting no matter whether we used the raw data or the anomaly weighted data. This is not surprising since the sample size is very small and the image quality in the Dataset 1 is various. We were expecting that the sparsity introduced by the anomaly mask could help avoiding over-fitting [39]. However, it turns out that the anomaly mask has limited effect on preventing over-fitting. This might because the main content of the weighted images is still from the raw images, and only a small proportion comes from the anomaly mask. The sparsity introduced by the mask might be not enough for avoiding over-fitting. For the second dataset, over-fitting was not observed as the performance was stable during validation and testing. As shown in Table 3, we also observed that the DenseNet models under different network architectures (Table 1) have better performance than the VGG model (Table 1). Generally speaking, The DenseNet and the VGG models also showed the improved performance using the anomaly weighted images than raw images using the testing sets in both Datasets 1 and 2 (**Table 3**).

V. DISCUSSION

Matrix profile is a successful technique in unsupervised rare pattern-based time-series anomaly detection [40]. It was developed based on the nearest neighbour algorithm. In this study, matrix profile was introduced to static image anomaly detection at one-dimensional level and two-dimensional level, separately. At one-dimensional level, image matrix was flattened to a long vector which could be considered as a time-series. With this transformation, the set of entire subsequences could be scanned efficiently using fast Fourier algorithms. This method works fine in identifying anomalies within an image. However, extra manual operations need to be done for choosing a suitable discord. And it is unnecessary to visit an image pixel by pixel for anomaly detection. Instead,

it is more reasonable to directly calculate the image matrix profile at the two-dimensional level. In our two-dimensional matrix profile method, a predefined size of sub-segment is scanned and its nearest distance with other sub-segments is calculated. The combination of all these nearest distances is mapped to the same coordinates of the corresponding sub-segments in the original images. The generated anomaly mask has the ability to indicate the meaningful lesion pixels in the original images. We further transformed each of the images into a severity score as a fast tool to indicate the normality of the images. This severity score showed significant difference between COVID-19 group and Non-COVID-19 group, which means it could work as an automatic and easy-calculated clinical tool to support COVID-19 diagnosis. To understand the potential causal effect of the severity score on COVID-19 diagnosis, we performed a statistical mediation analysis to examine the association between COVID-19 diagnosis and the number of underlying diseases through the score. We identified the significant indirect effect of the number of underlying diseases on COVID-19 severity through this severity score.

The anomaly mask can also be used to weight the original image for completing further tasks. In this study, we evaluated the performance of the anomaly weighted images to classify the COVID-19 and Non-COVID-19 lung CT images using a deep learning model. The anomaly weighted images were shown to be better in training the deep classification model than the raw images. This is likely due to the enhanced information introduced by the preprocessing. We made the whole working flow connected so that it could be implemented easily [41]. What's more, unsupervised anomaly detection and supervised deep convolutional neural network could be combined together in an end-to-end manner.

To control the runtime of the algorithm, we downsized the raw images to a smaller resolution. Although to obtain the best classification performance of the deep learning model is not the main task of this study, we realized the degradation in resolution might decrease the performance of the deep convolutional neural network based image classifier used in this study [42]. It might be better to keep the original resolution if the runtime is not a consideration or the computer configuration could be improved. Another potential future direction of this study is to develop ultra-fast algorithms for two-dimensional matrix profile calculation using a two-dimensional fast Fourier transformation [43]. Also, for the one-dimensional method, currently the one-dimensional fast Fourier transformation (FFT) [44] involved in the core algorithm does not consider the sparsity of medical image data. We could introduce sparse Fourier transform (SFT) [45] into the core calculation of one-dimensional matrix profile algorithm. At the application level, this technique is not limited to analysis of COVID-19 CT images, it could be extended to other diseases and other image types. Although this study focuses on the two-dimensional image anomaly detection problem, the matrix profile technique could be potentially further extended to analyze three-dimensional volume rendered CT scans which are more commonly used in medical practice. The potential application of CT-SS could also be explored if more clinical information are provided. For example, if the clinical outcomes (prognosis, treatment response, etc.) of the patients are available, the associations of CT-SS with these clinical outcomes could be further analysed [8]. Although this study is not intended to compete with the most state-of-art work in completing a classification task, we could integrate our sparse matrix profile method for enhancing anomalies in images with other advanced deep learning models [46], [47] and some data augmentation techniques to achieve the best classification performance.

VI. CONCLUSION

Inspired by the success of matrix profile in time-series data anomaly detection, we attempted to extent its application to image anomaly detection. Two possible clinical utilities have been tested, which are the CT-SS and the anomaly weighted images. The CT-SS could significantly distinguish the COVID-19 and Non-COVID-19 patients. This ability might come from the mechanism of its mediation effect on the number of underlying diseases' association with COVID-19 severity. The anomaly weighted images performed better in training different settings of DensNet models than the raw images. These significant results revealed its potential use for the lung CT imaging-based COVID-19 rapid diagnosis. This work has opened a window for raising the one-dimensional rare pattern mining algorithm to solve two-dimensional rare pattern detection problem. Unsupervised anomaly detection and advanced deep convolutional neural network were utilized in an unbroken and connected manner. The proposed algorithm is also dimension-extendible and explainable, in terms of its nearest neighbour theory. Furthermore, the implemented algorithm package might become a clinical tool for COVID-19 rapid diagnosis and assessment.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[2] Z. Alaverdyan, "Unsupervised representation learning for anomaly detection on neuroimaging. Application to epilepsy lesion detection on brain MRI," Ph.D. dissertation, Dept. Comput. Sci., Univ. Lyon, Lyon, France, 2019.

[3] Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, and M. Wang, "Presumed asymptomatic carrier transmission of COVID-19," *J. Amer. Med. Assoc.*, vol. 323, no. 14, pp. 1406–1407, Apr. 2020.

[4] Y. Li and L. Xia, "Coronavirus disease 2019 (COVID-19): Role of chest CT in diagnosis and management," *Amer. J. Roentgenol.*, vol. 214, no. 6, pp. 1–7, Jun. 2020.

[5] B. Udugama, P. Kadhiresan, H. N. Kozlowski, A. Malekjahani, M. Osborne, V. Y. C. Li, H. Chen, S. Mubareka, J. B. Gubbay, and W. C. W. Chan, "Diagnosing COVID-19: The disease and tools for detection," *ACS Nano*, vol. 14, no. 4, pp. 3822–3835, Apr. 2020.

[6] S. A. Harmon, T. H. Sanford, S. Xu, E. B. Turkbey, H. Roth, Z. Xu, D. Yang, A. Myronenko, V. Anderson, A. Amalou, and M. Blain, "Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets," *Nature Commun.*, vol. 11, no. 1, pp. 1–7, Dec. 2020.

[7] R. Yang, X. Li, H. Liu, Y. Zhen, X. Zhang, Q. Xiong, Y. Luo, C. Gao, and W. Zeng, "Chest CT severity score: An imaging tool for assessing severe COVID-19," *Radiol. Cardiothoracic Imag.*, vol. 2, no. 2, Apr. 2020, Art. no. e200047.

[8] Z. Feng *et al.*, "Early prediction of disease progression in COVID-19 pneumonia patients with chest CT and clinical characteristics," *Nature Commun.*, vol. 11, 2020, Art. no. 4968, doi: 10.1038/s41467-020-18786-x.

[9] N. Courty and E. Marchand, "Visual perception based on salient features," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, vol. 1, Oct. 2003, pp. 1024–1029.

[10] E. Niebur, "Saliency map," *Scholarpedia*, vol. 2, no. 8, p. 2675, 2007.

[11] C.-C.-M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, Z. Zimmerman, D. F. Silva, A. Mueen, and E. Keogh, "Time series joins, motifs, discords and shapelets: A unifying view that exploits the matrix profile," *Data Mining Knowl. Discovery*, vol. 32, no. 1, pp. 83–123, Jan. 2018.

[12] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, "COVID-CT-dataset: A CT scan dataset about COVID-19," 2020, *arXiv:2003.13865*. [Online]. Available: http://arxiv.org/abs/2003.13865

[13] W. Ning, S. Lei, J. Yang, Y. Cao, and P. Jiang, "iCTCF: An integrative resource of chest computed tomography images and clinical features of patients with COVID-19 pneumonia," *Res. Square*, 2020, doi: 10.21203/rs.3.rs-21834/v1.

[14] M. L. Antonie, O. R. Zaiane, and A. Coman, "Application of data mining techniques for medical image classification," in *Proc. 2nd Int. Conf. Multimedia Data Mining*, 2001, pp. 94–101.

[15] B. Liu, W. Hsu, and Y. Ma, "Mining association rules with multiple minimum supports," in *Proc. KDD*, 1999, pp. 337–341.

[16] L. Szathmary, A. Napoli, and P. Valtchev, "Towards rare itemset mining," in *Proc. 19th IEEE Int. Conf. Tools Artif. Intell. (ICTAI)*, Oct. 2007, pp. 305–312. [Online]. Available: https://hal.inria.fr/inria-00189424

[17] S. Tsang, Y. S. Koh, and G. Dobbie, "RP-Tree: Rare pattern tree mining," in *Proc. Int. Conf. Data Warehousing Knowl. Discovery* (Lecture Notes in Computer Science), vol. 6862. Berlin, Germany: Springer, 2011, pp. 277–288.

[18] A. Borah and B. Nath, "Rare pattern mining: Challenges and future perspectives," *Complex Intell. Syst.*, vol. 5, no. 1, pp. 1–23, Mar. 2019, doi: 10.1007/s40747-018-0085-9.

[19] T. Ehret, A. Davy, J.-M. Morel, and M. Delbracio, "Image anomalies: A review and synthesis of detection methods," *J. Math. Imag. Vis.*, vol. 61, no. 5, pp. 710–743, Jun. 2019.

[20] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. London, U.K.: Pearson, 2016.

[21] C.-C.-M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh, "Matrix profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 1317–1322.

[22] Y. Zhu, Z. Zimmerman, N. S. Senobari, C.-C.-M. Yeh, G. Funning, A. Mueen, P. Brisk, and E. Keogh, "Matrix profile II: Exploiting a novel algorithm and GPUs to break the one hundred million barrier for time series motifs and joins," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 739–748.

[23] C.-C.-M. Yeh, H. Van Herle, and E. Keogh, "Matrix profile III: The matrix profile allows visualization of salient subsequences in massive time series," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 579–588.

[24] C.-C.-M. Yeh, N. Kavantzas, and E. Keogh, "Matrix profile VI: Meaningful multidimensional motif discovery," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 565–574.

[25] H. Hassanieh, P. Indyk, D. Katabi, and E. Price, "Nearly optimal sparse Fourier transform," in *Proc. Annu. ACM Symp. Theory Comput.* New York, NY, USA: ACM Press, 2012, pp. 563–577. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2213977.2214029

[26] L. Bergman, N. Cohen, and Y. Hoshen, "Deep nearest neighbor anomaly detection," 2020, *arXiv:2002.10445*. [Online]. Available: http://arxiv.org/abs/2002.10445

[27] N. Sarafijanovic-Djukic and J. Davis, "Fast distance-based anomaly detection in images using an inception-like autoencoder," in *Proc. Int. Conf. Discovery Sci.* Cham, Switzerland: Springer, 2019, pp. 493–508.

[28] X. Bi, L. Yang, T. Li, B. Wang, H. Zhu, and H. Zhang, "Genome-wide mediation analysis of psychiatric and cognitive traits through imaging phenotypes," *Hum. Brain Mapping*, vol. 38, no. 8, pp. 4088–4097, Aug. 2017.

[29] N. Luo, J. Sui, J. Chen, F. Zhang, L. Tian, D. Lin, M. Song, V. D. Calhoun, Y. Cui, V. M. Vergara, and F. Zheng, "A schizophrenia-related Genetic-Brain-Cognition pathway revealed in a large chinese population," *EBioMedicine*, vol. 37, pp. 471–482, Nov. 2018.

[30] F. Shan, Y. Gao, J. Wang, W. Shi, N. Shi, M. Han, Z. Xue, D. Shen, and Y. Shi, "Lung infection quantification of COVID-19 in CT images with deep learning," 2020, *arXiv:2003.04655*. [Online]. Available: http://arxiv.org/abs/2003.04655

[31] O. Gozes, M. Frid-Adar, H. Greenspan, P. D. Browning, H. Zhang, W. Ji, A. Bernheim, and E. Siegel, "Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis," 2020, *arXiv:2003.05037*. [Online]. Available: http://arxiv.org/abs/2003.05037

[32] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, and B. Xu, "A deep learning algorithm using CT images to screen for corona virus disease (COVID-19)," *MedRxiv*, Jan. 2020, doi: 10.1101/2020.02.14.20023028.

[33] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, K. Wang, and L. Ye, "Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423–1433, Jun. 2020, doi: 10.1016/j.cell.2020.04.045.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[36] N. M. Khan and K. Raahemifar, "A novel accelerated greedy snake algorithm for active contours," in *Proc. 24th Can. Conf. Electr. Comput. Eng. (CCECE)*, May 2011, pp. 000186–000190.

[37] D. P. MacKinnon, A. J. Fairchild, and M. S. Fritz, "Mediation analysis," *Annu. Rev. Psychol.*, vol. 58, pp. 593–614, Jan. 2007.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[39] X. Sun, X. Ren, S. Ma, and H. Wang, "meProp: Sparsified back propagation for accelerated deep learning with reduced overfitting," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70. Sydney, NSW, Australia: PMLR, International Convention Centre, Aug. 2017, pp. 3299–3308. [Online]. Available: http://proceedings.mlr.press/v70/sun17c.html

[40] L. Feremans, V. Vercruyssen, B. Cule, W. Meert, and B. Goethals, "Pattern-based anomaly detection in mixed-type time series," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2019, pp. 240–256.

[41] D. Wu, K. Kim, B. Dong, G. El Fakhri, and Q. Li, "End-to-end lung nodule detection in computed tomography," 2017, *arXiv:1711.02074*. [Online]. Available: http://arxiv.org/abs/1711.02074

[42] S. P. Kannojia and G. Jaiswal, "Effects of varying resolution on performance of CNN based image classification an experimental study," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 9, pp. 451–456, Sep. 2018.

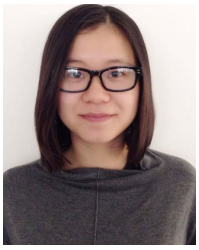
[43] V. S. Tutatchikov, "Two-dimensional fast Fourier transform: Batterfly in analog of Cooley-Tukey algorithm," in *Proc. 11th Int. Forum Strategic Technol. (IFOST)*, Jun. 2016, pp. 495–498.

[44] H. J. Nussbaumer, "The fast Fourier transform," in *Fast Fourier Transform and Convolution Algorithms*. Heidelberg, Germany: Springer, 1981, pp. 80–111.

[45] H. Hassanieh, P. Indyk, D. Katabi, and E. Price, "Simple and practical algorithm for sparse Fourier transform," in *Proc. 23rd Annu. ACM-SIAM Symp. Discrete Algorithms*, Jan. 2012, pp. 1183–1194.

[46] M. Z. Alom, M. M. S. Rahman, M. S. Nasrin, T. M. Taha, and V. K. Asari, "COVID_MTNet: COVID-19 detection with multi-task deep learning approaches," Apr. 2020, *arXiv:2004.03747*. [Online]. Available: http://arxiv.org/abs/2004.03747

[47] D. Davletshina, V. Melnychuk, V. Tran, H. Singla, M. Berrendorf, E. Faerman, M. Fromm, and M. Schubert, "Unsupervised anomaly detection for X-ray images," 2020, *arXiv:2001.10883*. [Online]. Available: http://arxiv.org/abs/2001.10883



QIAN LIU received the B.Sc. degree in medical imaging from Sichuan University, China, in 2011, and the M.Sc. degree in computational biology from the University of Manitoba, Canada, in 2019, where she is currently pursuing the Ph.D. degree in computational biology. Her research interests include machining learning, data mining, medical imaging, and molecular omics data analysis.

CARSON K. LEUNG (Senior Member, IEEE) received the B.Sc. (Hons.), M.Sc., and Ph.D. degrees from The University of British Columbia, Canada. He is currently a Full Professor with the Department of Computer Science, University of Manitoba, Canada. He is the Founder and Director of the Database and Data Mining Laboratory. His research interests include data mining and analysis, machine learning, data science, big data, databases (including image databases), data management, data warehousing, data visualization and visual analytics, bioinformatics, health informatics and electronic health, Web technology and services, as well as social computing and social network analysis.

PINGZHAO HU received the master's degree from Dalhousie University and the Ph.D. degree in computer science from York University. He is currently an Associate Professor with the Department of Biochemistry and Medical Genetics and an Adjunct Professor with the Department of Computer Science, University of Manitoba, and an Assistant Professor (Status) with the Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto. His research interests include statistical genetics and bioinformatics, especially in developing statistical and computational algorithms for integrative analysis of genomic data to understand the causal associations between human genome and phenome.