Received: 27 February 2024 | Revised: 8 May 2024 | Accepted: 20 May 2024

DOI: 10.1002/sta4.705

ORIGINAL ARTICLE

WILEY

Tensor factor adjustment for image classification with pervasive noises

Xiaochuan Li[1] | Bingnan Li[2] | Wenzhan Song[3] | Yuan Ke[2]

[1]Wells Fargo, Sunnyvale, California, USA

[2]Department of Statistics, University of Georgia, Athens, Georgia, USA

[3]School of Electrical and Computer Engineering, University of Georgia, Athens, Georgia, USA

Correspondence
Yuan Ke, 404 Brooks Hall, University of Georgia, Athens, GA 30024, USA.
Email: yuan.ke@uga.edu

Present address
404 Brooks Hall, University of Georgia, Athens, GA 30024, USA

Funding information
NSF, Grant/Award Numbers: NSF- 2312974, NSF-1940864, NSF-2019311, NSF-2210468, NSF-2243044, NSF-2324389; NIH, Grant/Award Number: NIH-1R01HL172291-01

Abstract

This paper studies a tensor factor model that augments samples from multiple classes. The nuisance common patterns shared across classes are characterised by pervasive noises, and the patterns that distinguish different classes are represented by class-specific components. Additionally, the pervasive component is modelled by the production of a low-rank tensor latent factor and several factor loading matrices. This augmented tensor factor model can be expanded to a series of matrix variate tensor factor models and estimated using principal component analysis. The ranks of latent factors are estimated using a modified eigen-ratio method. The proposed estimators have fast convergence rates and enjoy the blessing of dimensionality. The proposed factor model is applied to address the challenge of overlapping issues in image classification through a factor adjustment procedure. The procedure is shown to be powerful through synthetic experiments and an application to COVID-19 pneumonia diagnosis from frontal chest X-ray images.

KEYWORDS
chest X-ray image, COVID-19 diagnosis, medical image classification, tensor factor model

1 | INTRODUCTION

With the advancement of modern data science, data structured as tensors or multidimensional arrays are becoming increasingly prevalent in diverse fields. For example, images or video clips are naturally multiway tensors. Electrical health records, as another example, collect multiple features of many patients over a period of time and can be treated as tensor time series Zhang et al. (2021). Other examples of tensor data can be found in sensor networks He et al. (2016), magnetic resonance imaging Hasan et al. (2011), spatial-temporal analysis Ran et al. (2016), climate research Zhang et al. (2009), microbiome studies Mor et al. (2022) and quantitative finance Huang et al. (2018). As tensor data are typically high-dimensional and large-scale, learning its information through parsimonious yet flexible models is desirable. To that end, classical latent factor models (Bai, 2003; Bai & Li, 2012; Fan et al., 2008, 2011; Lam & Yao, 2012) have been revitalised by the statistics and machine learning communities to understand their computational, empirical and theoretical properties for tensor data applications (Chen et al., 2019, 2022; Chen & Fan, 2021; He et al., 2022; Wang et al., 2019; Zhang & Han, 2019).

Computer-aided diagnosis for medical images has attracted significant attention in the computer vision domain over the past decade. During the COVID-19 pandemic, deep neural network-based image classification methods have been widely used to detect pneumonia caused by the virus from chest X-ray images (Jain et al., 2021; Ismael & Şengür, 2021; Li et al., 2022). Early detection of lung infection has been proven critical for COVID-19 patients with a high risk of developing severe symptoms, as timely treatment can reduce their mortality rates (Goyal et al., 2020; Sun et al., 2020). A well-observed challenge in medical image classification is that the images from positive and negative classes can be highly overlapping, in the sense that only a small area of the image contains the specific information to distinguish the two classes. In comparison, the rest of the image contains pervasive noises between the two classes, such as background, body shape, skeleton and tissues. In Figure 1, we

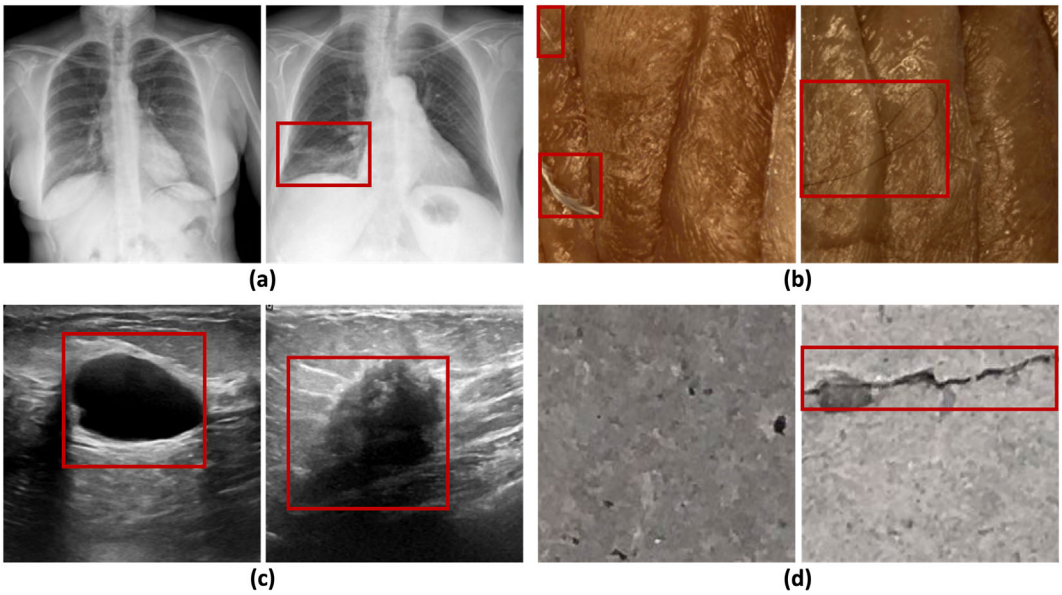

FIGURE 1 Examples of overlapping issues in image classification. The red boxes indicate the key areas used to distinguish positive and negative classes. (a) COVID-19 diagnosis via chest X-ray: normal (left) versus pneumonia (right); (b) food contamination detection: feather (left) versus hair (right); (c) breast cancer detection via ultrasound: benign (left) versus malignant (right); (d) surface crack detection: no crack (left) versus crack (right).

provide several examples of this overlapping phenomenon in image classification applications. As a result, even cutting-edge deep learning-based classifiers may suffer from limited correct classification rates since the weak specific signals are obscured by large pervasive noise.

In this paper, we introduce a novel augmented tensor factor model to address the aforementioned challenge in image classification, which may provide insight into more general heterogeneous tensor data analysis problems. Our model combines samples from multiple classes and decomposes the tensor data into a pervasive component that shares the same pattern across classes and a specific component that varies from class to class. The pervasive component is modelled by the production of a low-rank latent tensor factor and a few factor loading matrices. We propose to matricise the augmented tensor factor model into a series of matrix variate factor models. Then, factor loading matrices and latent tensor factors are estimated using principal component analysis (PCA), which leverages the wisdom of classical latent factor analysis. The ranks of latent tensor factors are estimated using a modified eigen-ratio method. Additionally, we prove the theoretical properties of our estimators. Our estimators for factor loading matrices and latent tensor factors benefit from large sample sizes and high dimensionality, leading to fast convergence rates. The proposed rank estimation method can also consistently recover the ranks of latent tensor factors. Thanks to these desirable and novel theoretical results, we developed a factor adjustment procedure for overlapping image classification. We use the training sample to estimate ranks and factor loading matrices; then, we regress the testing set on the estimated factor loading matrices. Finally, we apply the classifier to the regression residuals, which are consistent estimates of specific components. The factor adjustment procedure improves the signal-to-noise ratio by removing pervasive noise. The effectiveness of this procedure is demonstrated through synthetic experiments. We also show that our method improves the correct COVID-19 pneumonia diagnostic rate from chest X-ray images by 10.5%.

1.1 | Notations and definitions

A tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \ldots \times p_D}$ is a multidimensional or D-way array. The value D specifies the tensor order or the number of tensor modes. We denote scalars by lowercase letters, for example, x, and consider them as 0th-order tensors. We denote vectors by bold lowercase letters, for example, $\mathbf{x}$, and consider them as first-order tensors. We denote matrices by bold capital letters, for example, $\mathbf{X}$, and consider them as second-order tensors. Tensors of dimension three or more are denoted by boldface Euler script letters, for example, $\mathcal{X}$. For a D-way tensor, we define the following.

Fibres: Fibres are created by fixing all but one index of a tensor. For a matrix, rows and columns are mode-1 and mode-2 fibres, respectively. Fibres are always assumed to be oriented as column vectors.

Matricisation: The mode-d matricisation (unfolding) of $\mathcal{X}$ is denoted as $\mathcal{X}^{(d)} \in \mathbb{R}^{p_d \times \prod_{m \neq d} p_m}$ and arranges the mode-d fibres of $\mathcal{X}$ into the columns of $\mathcal{X}^{(d)}$.

Kronecker product: For two matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{p \times q}$, the Kronecker product (denoted by $\otimes$) is given by

$$\mathbf{A}\bigotimes\mathbf{B}=\begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}\in\mathbb{R}^{pm\times qn}.$$

Mode-d Product: The mode-d product between a tensor $\mathcal{X}\in\mathbb{R}^{p_1\times\dots\times p_D}$ and a matrix $\mathbf{A}\in\mathbb{R}^{r_d\times p_d}$ is denoted by $\mathcal{Y}=\mathcal{X}\times_n\mathbf{A}\in\mathbb{R}^{p_1\times\dots\times p_{dn-1}\times r_d\times p_{d+1}\dots\times p_D}$. Element-wise, this operation can be expressed as follows:

$$y_{p_1\dots p_{d-1}r_dp_{d+1}\dots p_D}=\sum_{p_d}x_{p_1\dots p_D}a_{p_dr_d}.$$

2 | AUGMENTED TENSOR FACTOR MODEL

2.1 | Model setup

Suppose we observe two D-way tensor valued random samples $\{\mathcal{X}_i^{(+)}\in\mathbb{R}^{p_1\times\dots\times p_D}\}_{i=1}^{n_1}$ and $\{\mathcal{X}_i^{(-)}\in\mathbb{R}^{p_1\times\dots\times p_D}\}_{i=1}^{n_2}$ from positive and negative classes, respectively. We assume the two samples share a common pervasive component $\mathcal{P}$ but different specific components $\mathcal{S}^{(+)}$ and $\mathcal{S}^{(-)}$. To be specific,

$$\begin{aligned}\mathcal{X}_i^{(+)} &=\mathcal{P}_i+\mathcal{S}_i^{(+)},\ \text{for}i=1,\dots,n_1,\\ \mathcal{X}_i^{(-)} &=\mathcal{P}_i+\mathcal{S}_i^{(-)},\ \text{for}i=1,\dots,n_2.\end{aligned}$$

As the pervasive component does not contain useful information to distinguish the positive and negative classes, we aim to adjust it and implement the binary classification on the specific components instead of the original observations.

Let $\{\mathcal{X}_i\in\mathbb{R}^{p_1\times\dots\times p_D}\}_{i=1}^{n}=\{\mathcal{X}_i^{(+)}\}_{i=1}^{n_1}\cup\{\mathcal{X}_i^{(-)}\}_{i=1}^{n_2}$ be the augmented sample with $n=n_1+n_2$, $\mathcal{X}_i=\mathcal{X}_i^{(+)}$ if $i\le n_1$ and $\mathcal{X}_i=\mathcal{X}_{i-n_1}^{(-)}$ if $i>n_1$. Without loss of generality, we can decompose $\mathcal{X}_i$ into two parts as follows:

$$\mathcal{X}_i=\mathcal{P}_i+\mathcal{S}_i,\ \text{for}i=1,\dots,n, \tag{1}$$

where $\mathcal{P}_i$ is a pervasive component and $\mathcal{S}_i$ is the augmented specific component that satisfies $\mathcal{S}_i=\mathcal{S}_i^{(+)}$ if $i\le n_1$ and $\mathcal{S}_i=\mathcal{S}_{i-n_1}^{(-)}$ if $i>n_1$. Further, we assume that $\mathcal{P}_i$ admits a Tucker tensor decomposition (Tucker, 1966) as follows:

$$\mathcal{P}_i=\mathcal{G}_i\times_1\mathbf{U}_1\times_2\dots\times_D\mathbf{U}_D,\ \text{for}i=1,\dots,n, \tag{2}$$

where $\mathcal{G}_i\in\mathbb{R}^{r_1\times\dots\times r_D}$ is a latent tensor factor with $r_d\ll p_d$ for $d=1\dots,D$ and $\mathbf{U}_d\in\mathbb{R}^{p_d\times r_d}$ is a factor loading matrix for the d-th mode of $\mathcal{X}_i$.

The tensor factor model (2) is flexible in the sense that both the latent tensor factor $\mathcal{G}_i$ and factor loading matrices $\{\mathbf{U}_1,\dots,\mathbf{U}_d\}$ are unobservable. Thus, like classical factor models, model (2) has identifiability issues. To be specific, let $\{\mathbf{H}_d\in\mathbb{R}^{r_d\times r_d}\}_{d=1}^{D}$ be a set of orthogonal matrices. Then, model (2) also holds for the latent tensor factor $\mathcal{G}_i\times_1\mathbf{H}_1^{-1}\times_2\dots\times_d\mathbf{H}_d^{-1}$ and factor loading matrices $\{\mathbf{U}_1\mathbf{H}_1,\dots,\mathbf{U}_d\mathbf{H}_d\}$. To facilitate the discussions in estimation and theoretical analysis, we impose the identification conditions as $p_d^{-1}\mathbf{U}_d^{\mathsf{T}}\mathbf{U}_d=\mathbf{I}_{r_d}$ for $d=1,\dots,D$. Please note that the identification conditions will not create any problems for our estimation method as we are mainly interested in estimating the linear space spanned by the columns of $\mathbf{U}_d$, which is the same as the one spanned by $\mathbf{U}_d\mathbf{H}_d$. Additionally, the proposed augmented tensor factor model can easily be generalised to multiple samples. To keep our presentation focused, we will not discuss this further in this paper.

2.2 | Matricisation and PCA estimation

Motivated by the PCA estimation of the classical factor models, we introduce a PCA type tensor factor model estimation method based on tensor matricisation. In this subsection, we assume that the ranks $r_1,\dots,r_D$ are known for the ease of presentation. The estimation of ranks will be

discussed in Section 2.3. Recall that we denote the mode-d matricisation of $\mathcal{X}_i$ as $\mathcal{X}_i^{(d)} \in \mathbb{R}^{p_d \times \prod_{m \neq d} p_m}$. Then, we can rewrite (1) and (2) as D matrix variate factor models (Chen et al., 2019; Chen & Fan, 2021; Wang et al., 2019), that is,

$$\begin{cases} \mathcal{X}_i^{(1)} &= \boldsymbol{U}_1 \mathcal{G}_i^{(1)} \boldsymbol{U}_{-1}^{\mathsf{T}} + \mathcal{S}_i^{(1)}, \\ &\vdots \\ \mathcal{X}_i^{(d)} &= \boldsymbol{U}_d \mathcal{G}_i^{(d)} \boldsymbol{U}_{-d}^{\mathsf{T}} + \mathcal{S}_i^{(d)}, \text{ for} i = 1, \ldots, n, \\ &\vdots \\ \mathcal{X}_i^{(D)} &= \boldsymbol{U}_d \mathcal{G}_i^{(D)} \boldsymbol{U}_{-D}^{\mathsf{T}} + \mathcal{S}_i^{(D)}, \end{cases} \tag{3}$$

where $\mathcal{G}_i^{(d)}$ and $\mathcal{S}_i^{(d)}$ are mode-d matricisations of $\mathcal{G}_i$ and $\mathcal{S}_i$ and $\boldsymbol{U}_{-d} = \bigotimes_{m \neq d} \boldsymbol{U}_m$.

To estimate the factor loading matrix $\boldsymbol{U}_d$, we first compute the model-d sample variance covariance matrix $\boldsymbol{\Sigma}_d \in \mathbb{R}^{p_d \times p_d}$ as follows:

$$\boldsymbol{\Sigma}_d = \frac{1}{n p_\pi} \sum_{i=1}^{n} \mathcal{X}_i^{(d)} \mathcal{X}_i^{(d)\mathsf{T}}, \text{ for} d = 1, \ldots, D,$$

where $p_\pi = \prod_{d=1}^{D} p_d$ is the product of all D dimensions.

Next, we apply an eigen-decomposition to each $\boldsymbol{\Sigma}_d$. Let $\lambda_1^{(d)} \geq \ldots \geq \lambda_{p_d}^{(d)}$ be the eigenvalues of $\boldsymbol{\Sigma}_d$ sorted in the descending order and $\boldsymbol{v}_1^{(d)}, \ldots, \boldsymbol{v}_{p_d}^{(d)}$ be the corresponding eigenvectors. For a given rank r_d, we propose to estimate $\boldsymbol{U}_d$ by

$$\begin{aligned} \hat{\boldsymbol{U}}_d :=& \ \hat{\boldsymbol{U}}_d(r_d) \\ =& \ \sqrt{p_d} \left(\boldsymbol{v}_1^{(d)}, \ldots, \boldsymbol{v}_{r_d}^{(d)} \right), \text{ for} d = 1, \ldots, D. \end{aligned}$$

By collecting all estimates of factor loading matrices and utilising identification conditions, we estimate the tensor latent factors by

$$\hat{\mathcal{G}}_i = \frac{1}{p_\pi} \mathcal{X}_i \times_1 \hat{\boldsymbol{U}}_1^{\mathsf{T}} \times_2 \ldots \times_D \hat{\boldsymbol{U}}_D^{\mathsf{T}}, \text{ for} i = 1, \ldots, n.$$

Then, naturally, we estimate the pervasive and specific components by

$$\begin{aligned} &\widehat{\mathcal{P}}_i = \hat{\mathcal{G}}_i \times_1 \hat{\boldsymbol{U}}_1 \times_2 \ldots \times_D \hat{\boldsymbol{U}}_d \\ \text{and} \ \ &\widehat{\mathcal{S}}_i = \mathcal{X}_i - \widehat{\mathcal{P}}_i, \text{ for} i = 1, \ldots, n. \end{aligned} \tag{4}$$

2.3 | Estimation of ranks

Estimating the ranks of tensor factor models is a challenging problem due to the unsupervised nature of the task. The matricisation decomposes the mode-D tensor factor model into D matrix variate factor models as in (3). This motivates us to borrow the wisdom from classical factor model inference literature (Ahn & Horenstein, 2013; Bai & Ng, 2002; Chamberlain & Rothschild, 1982; Chang et al., 2015; Lam & Yao, 2012; Stock & Watson, 2002).

For every mode d, we use the modified eigen-ratio method Chang et al. (2015) to estimate the fixed but unknown rank r_d. Let $\lambda_k(\boldsymbol{\Sigma}_d)$ the k-th largest eigenvalue of the model-d sample variance–covariance matrix, $K_{\max}$, be a prescribed upper bound and C be a small positive constant. We propose to estimate r_d by

$$\hat{r}_d = \arg \min_{1 \leq k \leq K_{\max}} \frac{\lambda_k(\boldsymbol{\Sigma}_d) + C}{\lambda_{k+1}(\boldsymbol{\Sigma}_d) + C}, \text{ for} d = 1, \ldots, D. \tag{5}$$

The threshold $K_{\max}$ controls the computational cost. It represents the belief of the largest possible value of r_d. To avoid random spikes when $\lambda_{k+1}(\boldsymbol{\Sigma}_d)$ is close to 0, a small positive constant C is used.

2.4 | Theoretical results

In this subsection, we investigate the theoretical properties of the proposed estimators. Due to space limitations, we provide all assumptions, technical lemmas and proofs in appendices found in a separate supporting information file.

Theorem 1 Convergence of factor loading matrices. Suppose Assumptions 1–3 in Appendix A in the supporting information hold. We assume that the ranks $\{r_d\}_{d=1}^D$ are fixed but allow sample size n and dimensions $\{p_d\}_{d=1}^D$ to diverge. Then, we have

$$\|\hat{\boldsymbol{U}}_d - \boldsymbol{U}_d \boldsymbol{H}_d\|_F^2 = O_p\left(\frac{1}{p_d} + \frac{p_d^2}{np_\pi}\right), \text{ for} d = 1, ..., D,$$

where $\{\boldsymbol{H}_d \in \mathbb{R}^{r_d \times r_d}\}_{d=1}^D$ is a set of orthogonal matrices.

Theorem 1 shows that our model-d factor loading matrix estimator can consistently estimate the truth up to an r_d by r_d orthogonal matrix. Further, detailed analysis in the proof of Theorem 1 shows that $\boldsymbol{H}_d = \frac{1}{np_d}\sum_{i=1}^n \mathcal{G}_i^{(d)} \mathcal{G}_i^{(d)\mathrm{T}} \boldsymbol{U}_d^{\mathrm{T}} \hat{\boldsymbol{U}}_d \boldsymbol{\Lambda}_d^{-1}$ with $\boldsymbol{\Lambda}_d = diag(\lambda_1 ..., \lambda_{r_d})$ being a diagonal matrix.

Theorem 2 Convergence of latent tensor factors. Suppose Assumptions 1–3 in Appendix A in the supporting information hold. We assume that the ranks $\{r_d\}_{d=1}^D$ are fixed but allow sample size n and dimensions $\{p_d\}_{d=1}^D$ to diverge. With the same $\{\boldsymbol{H}_d\}_{d=1}^D$ as in Theorem 1, we have, for $i = 1, ..., n$,

$$\|\hat{\mathcal{G}}_i - \mathcal{G}_i \times_1 \boldsymbol{H}_1^{-1} \times_2 ... \times_D \boldsymbol{H}_d^{-1}\|_F^2 = O_p\left(\frac{1}{n} + \sum_{d=1}^D \frac{1}{p_d}\right).$$

Theorem 2 provides an estimation error upper bound for the latent tensor factors. The convergence rate depends on both sample size, n, and dimension $\{p_1, ..., p_D\}$. This 'blessing of dimensionality' phenomenon is in line with classical factor model analysis, as noted in references such as Bai (2003) and Fan et al. (2008). Theorems 1 and 2 allow for consistent estimation of both the pervasive and specific components. When $\sum_{i=1}^D p_d^2 < p_\pi$, some simple algebra shows that, for $i = 1, ..., n$, we have

$$\begin{aligned} \left\|\hat{\mathcal{P}}_i - \mathcal{P}_i\right\|_F^2 &= O_p\left(\frac{1}{n} + \sum_{d=1}^D \frac{1}{p_d}\right) \\ \text{and } \left\|\hat{\mathcal{S}}_i - \mathcal{S}_i\right\|_F^2 &= O_p\left(\frac{1}{n} + \sum_{d=1}^D \frac{1}{p_d}\right). \end{aligned}$$

Theorem 3 Convergence of ranks. Suppose Assumptions 1–3 in Appendix A in the supporting information hold. We assume that the true ranks $\{r_d\}_{d=1}^D$ are fixed but allow sample size n and dimensions $\{p_d\}_{d=1}^D$ to diverge. Then, we have

$$P(\hat{r}_d = r_d) \to 1, \text{ for} d = 1, ..., D.$$

Theorem 3 demonstrates that the modified eigen-ratio method can correctly estimate all ranks with high probability. Theorem 3 offers a theoretical guarantee for the challenging rank estimation problem and is of independent interest.

3 | FACTOR ADJUSTMENT FOR IMAGE CLASSIFICATION

In this section, we focus on applying the augmented tensor factor model proposed in Section 2 to address challenges in overlapping image classification. To illustrate our ideas, we consider a binary classification between two sets of grey-scale images. Let $\{\mathcal{X}_i^{(+)} \in \mathbb{R}^{p_1 \times p_2}\}_{i=1}^{n_1}$ and $\{\mathcal{X}_i^{(-)} \in \mathbb{R}^{p_1 \times p_2}\}_{i=1}^{n_2}$ be two observable sets of images with positive and negative class labels, respectively. Without loss of generality, we assume

the images have been preprocessed to have a fixed size of p_1 by p_2 pixels. Denote $\{\mathcal{X}_i\}_{i=1}^{n} = \{\mathcal{X}_i^{(+)}\}_{i=1}^{n_1} \cup \{\mathcal{X}_i^{(-)}\}_{i=1}^{n_2}$ be the augmented sample with $n = n_1 + n_2$. Our goal is to train a classifier on this sample and predict the class label of images in a testing sample $\{\tilde{\mathcal{X}}_i \in \mathbb{R}^{p_1 \times p_2}\}_{i=1}^{m}$.

To avoid the distraction of the pervasive component and improve the classification accuracy, we propose a tensor factor adjustment image classification procedure which can be introduced by the following two phases.

Training phase

1. We treat the training sample as a three-way tensor $\{\mathcal{X}_i\}_{i=1}^{n}$ and model it with the augmented tensor factor model

$$\mathcal{X}_i = \mathcal{G}_i \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 + \mathcal{S}_i, \text{ for} i = 1, ..., n.$$

2. Obtain the matricisation of the above tensor factor model by the method introduced in Section 2.2.
3. Estimate the ranks by the modified eigen-ratio method introduced in Section 2.3.
4. Estimate factor loading matrices and latent tensor factors by the method introduced in Section 2.2. Then, by (4), we estimate the specific components as $\{\hat{\mathcal{S}}_i\}_{i=1}^{n}$.
5. We train a classifier on $\{\hat{\mathcal{S}}_i\}_{i=1}^{n}$ with the true class labels.

Testing phase

1. We regress the testing sample $\{\tilde{\mathcal{X}}_i\}_{i=1}^{m}$ onto the estimated factor loading matrices.
2. Estimate the specific components of the testing sample as the regression residuals, denoted as $\{\tilde{\mathcal{S}}_i\}_{i=1}^{m}$.
3. We apply the classifier trained in the training process to classify $\{\tilde{\mathcal{S}}_i\}_{i=1}^{m}$. The classification results will be used to predict the class labels of $\{\tilde{\mathcal{X}}_i\}_{i=1}^{m}$.

The flowcharts for the training and testing phases are summarised in Figures 2 and 3, respectively.

4 | SYNTHETIC EXPERIMENTS

In this section, we carefully evaluate the performance of the proposed tensor factor adjustment image classification procedure using various synthetic experiments. In Section 4.1, we present the experiment settings and implementation details. The results of the experiments are reported and analysed in Section 4.2.

4.1 | Experiment settings

For each experiment, we generate a positive sample $\{\mathcal{X}_i^{(+)}\}_{i=1}^{n_1}$ and a negative sample $\{\mathcal{X}_i^{(-)}\}_{i=1}^{n_2}$ from the following tensor factor model:

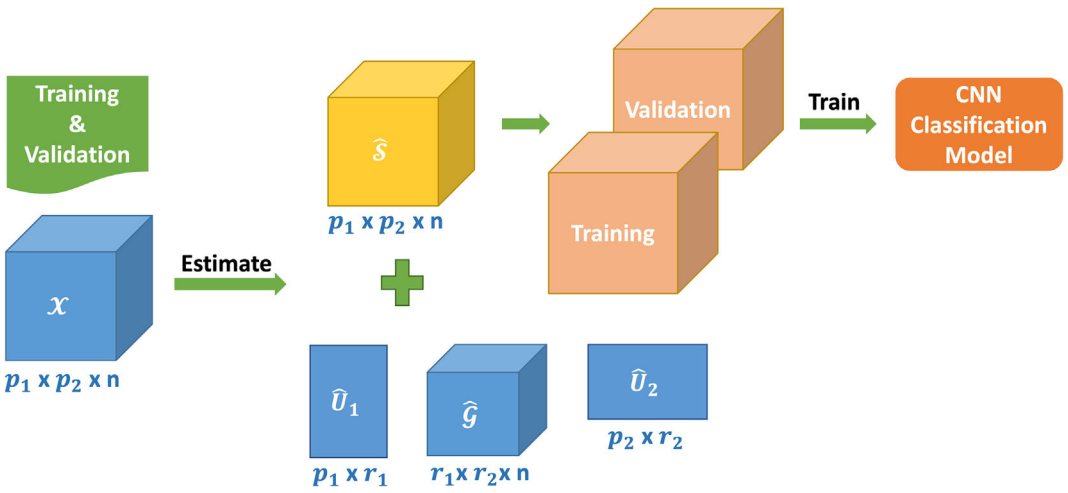

FIGURE 2 A flowchart for the training phase of the factor adjustment method.

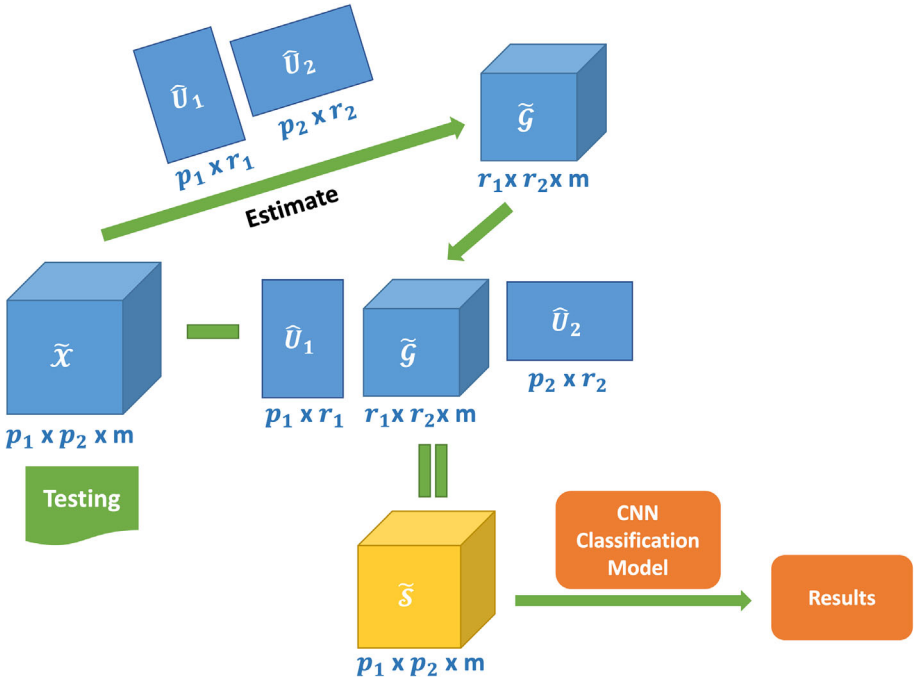

FIGURE 3 A flowchart for the testing phase of the factor adjustment method.

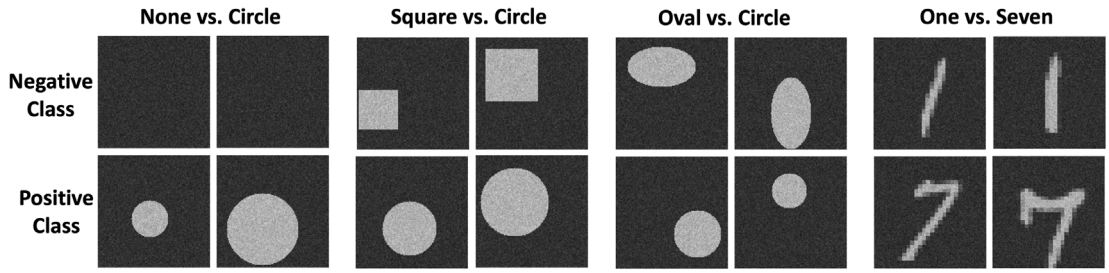

FIGURE 4 Synthetic specific component examples for each setting.

$$\begin{cases} \mathcal{X}_i^{(+)} = \beta \mathcal{G}_i \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 + \mathcal{S}_i^{(+)} + \mathcal{W}_i, \\ \mathcal{X}_i^{(-)} = \beta \mathcal{G}_i \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 + \mathcal{S}_i^{(-)} + \mathcal{W}_i, \end{cases}$$

where β is a nonnegative scalar parameter to control the overlapping amount between the positive and negative samples and $\mathcal{W}_i$ is a white noise term to add some randomness to the specific components.

The data-generating process for each component in the above model is summarised as follows.

1. The elements in latent tensor factors $\{\mathcal{G}_i\}_{i=1}^{n_1+n_2}$ are drawn from i.i.d. uniform distribution between $-a$ and a, where a is a positive parameter to partially control the signal to noise ratio.
2. The factor loading matrices $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are designed to be orthonormal matrices to satisfy the identification conditions.
3. The elements in the white noise term $\mathcal{W}_i$ are drawn from i.i.d. standard Gaussian distribution.
4. The specific components $\mathcal{S}_i^{(+)}$ and $\mathcal{S}_i^{(-)}$ are generated from some class-specific image patterns. To be specific, we consider the following four pairs of image patterns: (a) none versus random filled circle; (b) random filled square versus random filled circles; (c) random filled oval versus random filled circles and (d) random handwritten 1 versus random handwritten 7 drawn from the MNIST dataset. We illustrate these four settings in Figure 4.

The ranks of the pervasive component will have an impact on the signal-to-noise ratio of the generated images. So, to keep the comparison of all settings fair, we will adjust the parameter a as $a(r_1,r_2)$ to keep all settings have approximately the same signal-to-noise ratio. The computational details are omitted here but can be recovered from the replication codes.

Further, we use the parameter β to control the overlapping ratio between the two classes. When we increase β, the pervasive components weigh more in both classes; hence, the true signals in the specific components are harder to recognise. Figure 5 provides a list of synthetic examples with β increased from 1 to 9. When the pervasive component dominates the data-generating process, as expected, the images in positive and negative classes are very hard to distinguish.

4.2 | Implementation and results

As outlined in Section 4.1, we generate synthetic images using the four signal settings demonstrated in Figure 4. For each setting, we set the image sizes to be $p_1 \times p_2 = 128 \times 128$ and the sample sizes $n_1 = n_2 = 2{,}400$. The ranks (r_1, r_2) are set as $(8,8)$, $(16,16)$ and $(32,32)$ to cover low to moderate rank settings. The generated samples are then divided into training, validation and testing sets as specified in Table 1. Throughout this paper, we use a modified VGG16 model Li and Ke (2022) as our binary image classifier. We use the training set to train our classifier and the validation set to tune hyperparameters. The trained classifier will be used to predict the class labels in the test set. The prediction accuracy is measured by the correct classification rate.

For each signal setting, we consider and compare the following three scenarios.

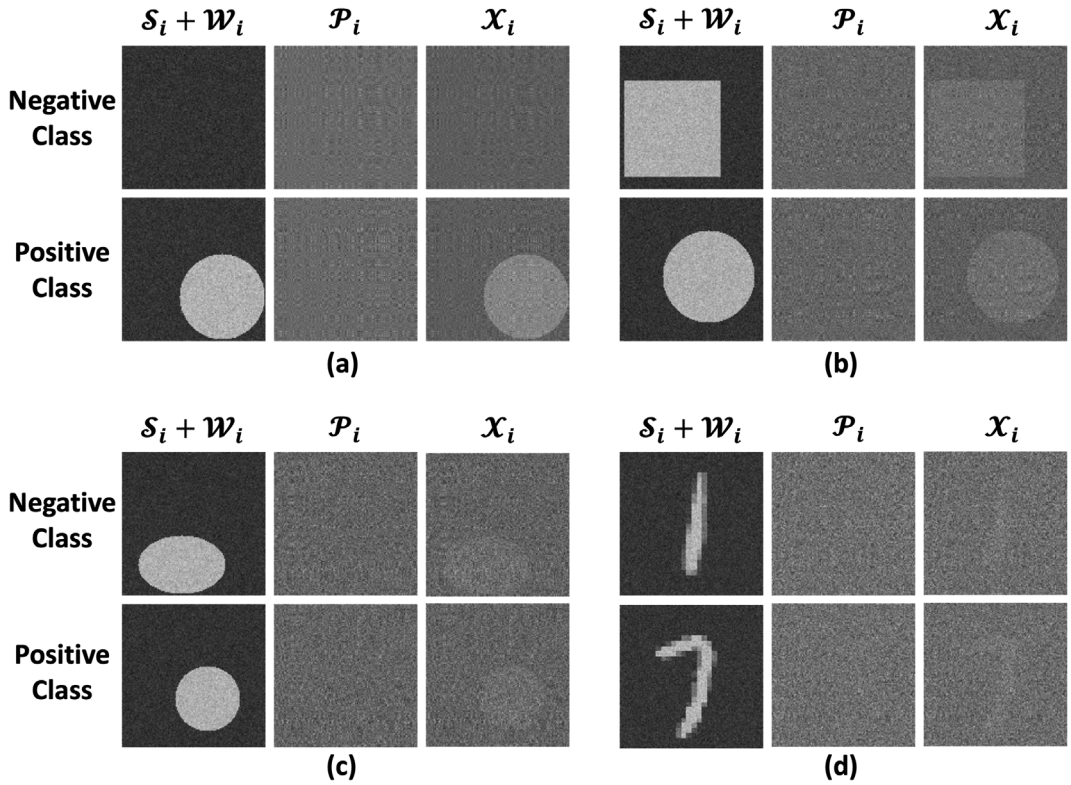

FIGURE 5 Visualisation of synthetic examples. (a) None versus circle with ranks $(8,8)$ and $\beta = 1$; (b) square versus circle with ranks $(16,16)$ and $\beta = 3$; (c) oval versus circle with ranks $(32,32)$ and $\beta = 6$; (d) one versus seven with ranks $(64,64)$ and $\beta = 9$. For each panel, the three columns from left to right represent specific components (signal with white noise), pervasive components (overlapping noise) and observed images.

TABLE 1 Synthetic and real data splitting sample sizes.

	Synthetic images		COVID-19 chest X-ray images	
	Negative class	Positive class	Negative class	Positive class
Training	2000	2000	1223	2325
Validation	200	200	200	200
Testing	200	200	200	200

Nonoverlapping: We generated a pair of nonoverlapping positive and negative samples $\{\mathcal{X}_i^{(+)}\}_{i=1}^{n_1}$ and $\{\mathcal{X}_i^{(-)}\}_{i=1}^{n_2}$ by setting the overlapping parameter $\beta = 0$. Then, it is equivalent to directly train and test the classifier on specific components plus white noises, that is, $\{\mathcal{S}_i^{(+)} + \mathcal{W}_i\}_{i=1}^{n_1}$ and $\{\mathcal{S}_j^{(-)} + \mathcal{W}_j\}_{j=1}^{n_2}$.

Overlapping: We set the overlapping parameter β to be large enough such that the modified VGG16 classifier behaves as badly as random guesses, that is, the correct classification rate on the testing set is about 0.5. In this scenario, we simulate the images that are highly overlapping between two classes, and most existing classifiers are not tailored to handle such cases.

Factor adjustment: We use the same highly overlapping data as generated in the overlapping scenario but apply the factor adjustment procedure as described in Section 3. Then, the classifier is trained and tested on the estimated specific components $\{\hat{\mathcal{S}}_i\}_{i=1}^{n_1+n_2}$ instead of the overlapping raw samples.

Table 2 presents the detailed experiment results. In all scenarios, the modified VGG16 classifier can perfectly predict the true class labels in the testing set for nonoverlapping cases but performs as poorly as random guesses for overlapping cases. This demonstrates that the presence of pervasive components can greatly hinder even simple image classification tasks, as the two classes are highly overlapping or correlated with each other. However, as expected, the proposed factor adjustment method effectively recovers the true signals dominated by the pervasive component and decorrelates the two overlapping classes. As a result, the factor adjustment method achieves near-perfect correct classification rates for all scenarios. In general, the factor adjustment method works as if we were classifying with the true signals.

The performances of these combinations over the validation set are visualised in Figure 6. Each point in the figure represents a prediction accuracy value of a rank option. Blue indicates a higher accuracy and red indicates lower accuracy. The darker the colour, the more extreme the accuracy value. The darkest blue specifies an accuracy of 1, and the darkest red implies an accuracy of 0.5. As observed, all the points in the $R_3 = 3$ and $R_3 = 8$ subfigures are red. For the rest subplots, we have blue dots, but the blue dots are only located in the top-right corner. This phenomenon signifies only if all the dimensions of the chosen rank are equal to or larger than the corresponding true rank; the proposed procedure is likely to have an ideal performance over the validation set. Therefore, if the grid search results provided multiple options as they achieved similar prediction accuracy, we consider choosing the rank with the smallest values, that is, conducting decomposition using the smallest core shape. This strategy also allows us to keep more information on the observed images.

5 | FRONTAL CHEST X-RAY IMAGE CLASSIFICATION FOR COVID-19 DIAGNOSTIC

The repeated waves of COVID-19 have infected over 660 million people and resulted in over 6.5 million deaths worldwide since its emergence in early 2020. Recently, a new wave of infections hit China as it relaxed its 'Zero-COVID' policy. While nucleic acid amplification tests and antigen tests are widely accepted for detecting positive cases, frontal chest X-ray (chest X-ray) image analysis remains a prominent method for diagnosing lung infections and pneumonia, which can be critical indicators for potential severe symptoms of COVID-19 (Jain et al., 2021; Ismael & Şengür, 2021; Li et al., 2022). However, the typical appearance of a COVID-19 lung infection can be complex for nonexperts to recognise, as the differences between negative and positive images are subtle and largely obscured by pervasive noises such as skeletons, organs and shadows. In that sense, it is essential to develop an accurate and efficient classification method to adjust for pervasive noises and identify positive chest X-ray images from negative ones. In this section, we apply the proposed augmented tensor factor model and factor adjustment method to tackle this task.

TABLE 2 Correct classification rate on testing set for synthetic experiments.

		Ranks (r_1, r_2)		
	Dataset	**(8, 8)**	**(16, 16)**	**(32, 32)**
None	Nonoverlapping		0.995	
versus	Overlapping	0.5	0.5	0.5
circle	Factor adjusted	0.995	0.995	0.9975
Square	Nonoverlapping		1.0	
versus	Overlapping	0.5	0.5	0.5
circle	Factor adjusted	0.995	0.995	0.9975
Oval	Nonoverlapping		1.0	
versus	Overlapping	0.5	0.5	0.5
circle	Factor adjusted	1.0	0.9925	0.975
One	Nonoverlapping		0.9925	
versus	Overlapping	0.5	0.5	0.5
seven	Factor adjusted	0.9875	0.9875	0.9825

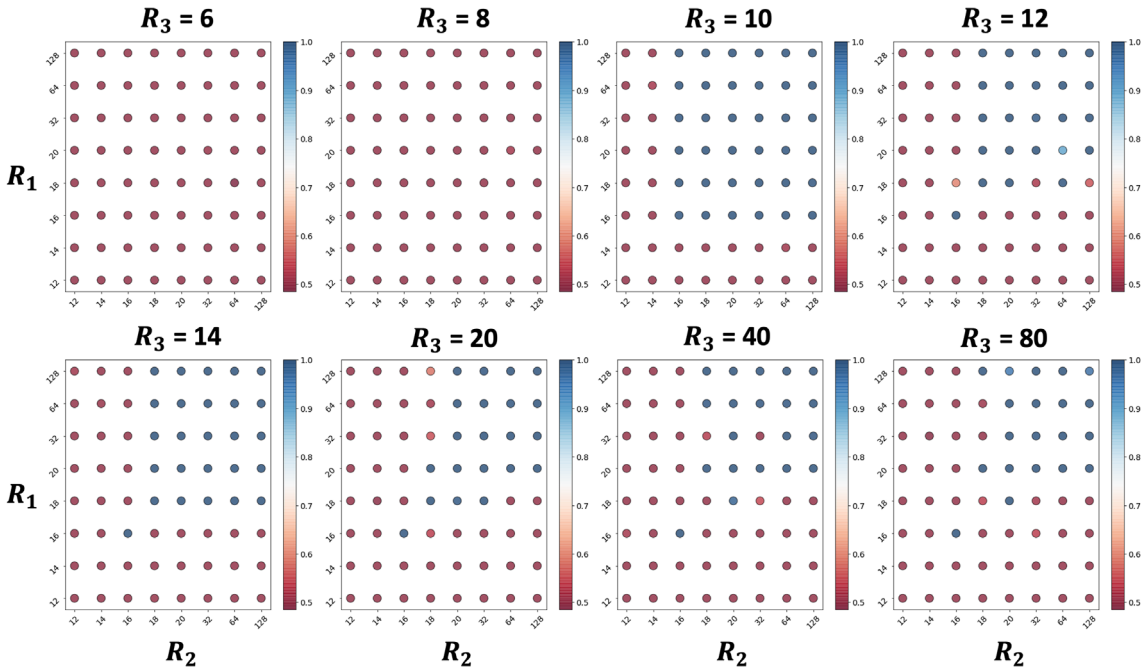

FIGURE 6 Prediction accuracy over the validation set during the grid search process.

5.1 | Data description

The dataset we study is collected and provided by the Society for Imaging Informatics in Medicine (SIIM).[1] The dataset combines high-quality chest X-ray images from the existing public BIMCV Vayá et al. (2020) and MIDRC-RICORD Tsai et al. (2021) COVID-19 datasets and annotates these chest radiographs by 22 radiologists. Following the guidelines listed in Lakhani et al. (2021), radiologists have annotated chest X-ray images into four mutually exclusive categories, including 'Negative for Pneumonia', 'Typical Appearance', 'Indeterminate Appearance' and 'Atypical Appearance'. Such a labelling strategy is based on the prior knowledge of radiographic manifestations of COVID-19 and can contribute to the consistency in radiology reportingLitmanovich et al. (2020).

In this section, we consider a binary classification problem by focusing on images from the'Negative for Pneumonia' and 'Typical Appearance for COVID-19' classes. To keep the presentation simple, we denote 'Negative for Pneumonia' as the negative class and 'Typical Appearance for COVID-19' as the positive class. The negative and positive classes contain 1623 and 2725 images, respectively. All images have been preprocessed to have a fixed size of 128×128 grey-scale pixels. Figure 7 lists several sample images from both positive and negative classes. From a nonprofessional perspective, the images from the two classes are highly overlapping as they are dominated by skeletons and organs with no valuable information to diagnose COVID-19. The useful information resides in the lung area, but the signals are weak and largely obscured by pervasive noises. In our experiment, we randomly divide both positive and negative classes into training, validation and testing sets, as detailed in Table 1.

5.2 | Analysis results

We begin by performing image classification on positive and negative classes using a modified VGG16 classifier. The classifier is trained on the training set, and the hyperparameters are fine-tuned utilising the validation set. Subsequently, the trained classifier is employed to predict class labels in the test set. Given that the two classes are imbalanced, we use a threshold of 0.66 instead of the standard 0.5 to account for the sample size ratio in the training set. The resulting classification accuracy is 71%, and we treat it as the baseline.

Next, we follow the factor adjustment procedure outlined in Section 3. We estimate and adjust the pervasive components when we train and predict class labels. The ranks are selected by the modified eigen-ratio method described in Section 2.3. To be specific, after setting the prescribed upper bound as one-fourth of the tensor size, that is, $K_{max} = 32$, the eigen-ratio method chooses $(r_1, r_2) = (28, 30)$. Similar to Figure 6, we visualise the performance of various rank combinations over the validation set in Figure 8. Figure 9 illustrates several pairs of positive and negative chest X-ray images in the testing set, along with their estimated pervasive and specific components. The pervasive components effectively identify the regions of the human skeleton as expected. In comparison with the raw images, the estimated specific components are visually more focused on

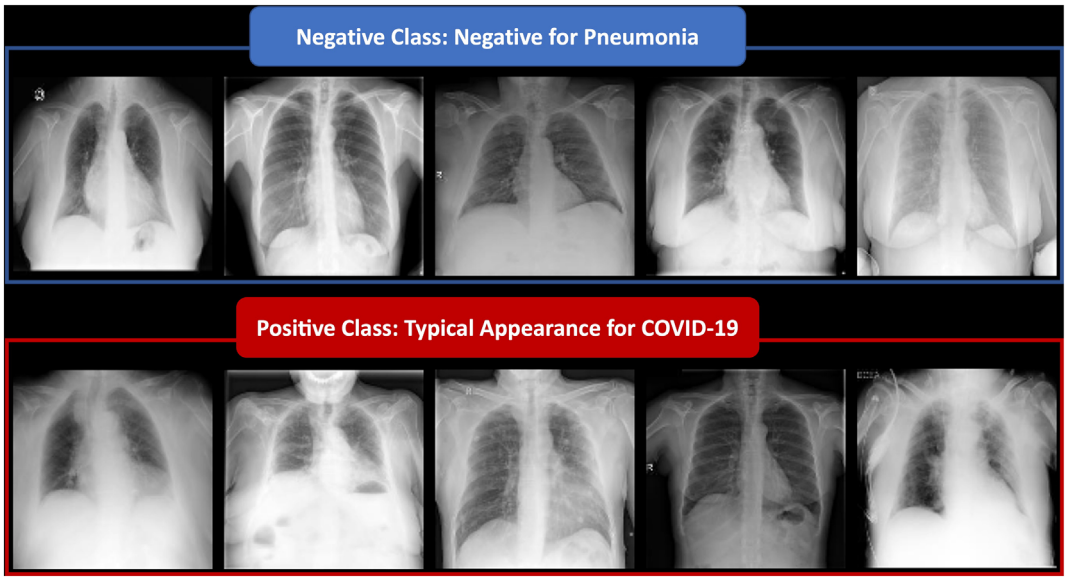

FIGURE 7 Sample images in the COVID-19 chest X-ray dataset.

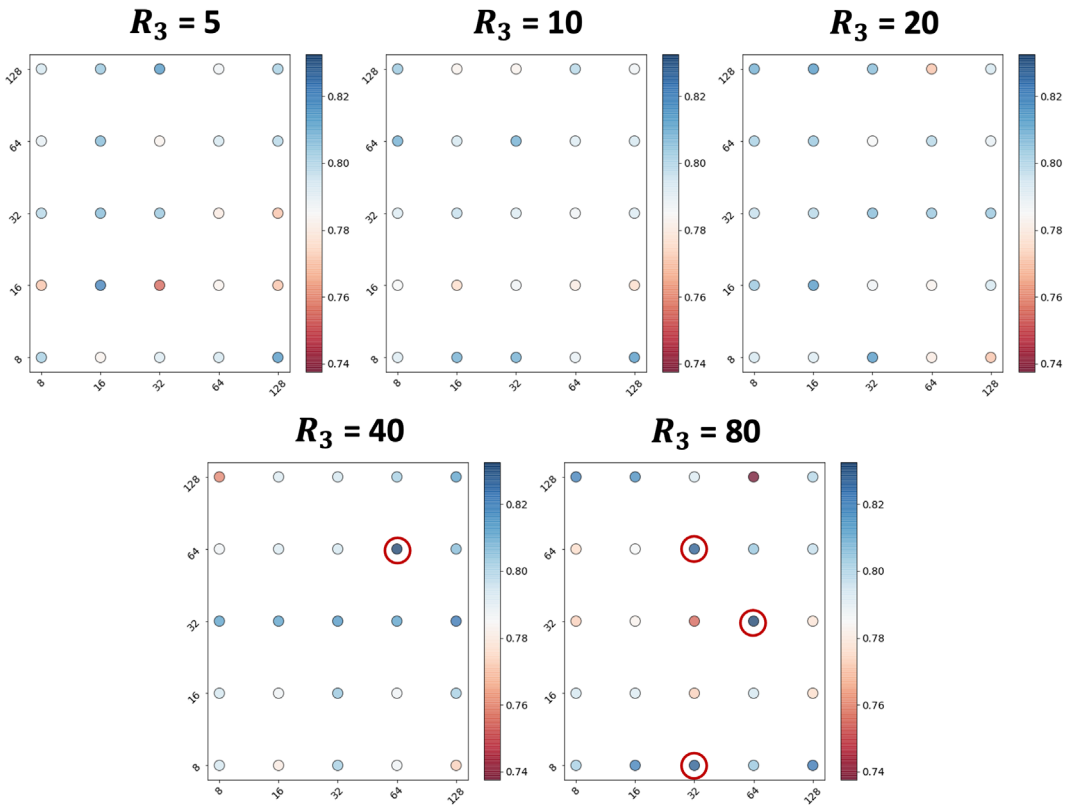

FIGURE 8 Real data grid search results over the validation set.

the lung area. Figure 10 visualises the classification probabilities of test images predicted by the modified VGG16 model trained on raw overlapping samples (left panel) and the factor adjusted samples (right panel). In the raw samples, the positive and negative classes are heavily overlapping with each other, making it difficult to establish a threshold that can clearly separate the two classes. In contrast, with the factor adjusted method, the two classes are less overlapped. Additionally, the negative class prediction distribution is more skewed towards 0, while that of the positive class is skewed towards 1, indicating that the estimated specific components are more distinguishable compared with the raw images. The resulting correct classification rate for the factor adjusted method is 81.5%, which is a 10.5% increase from the baseline.

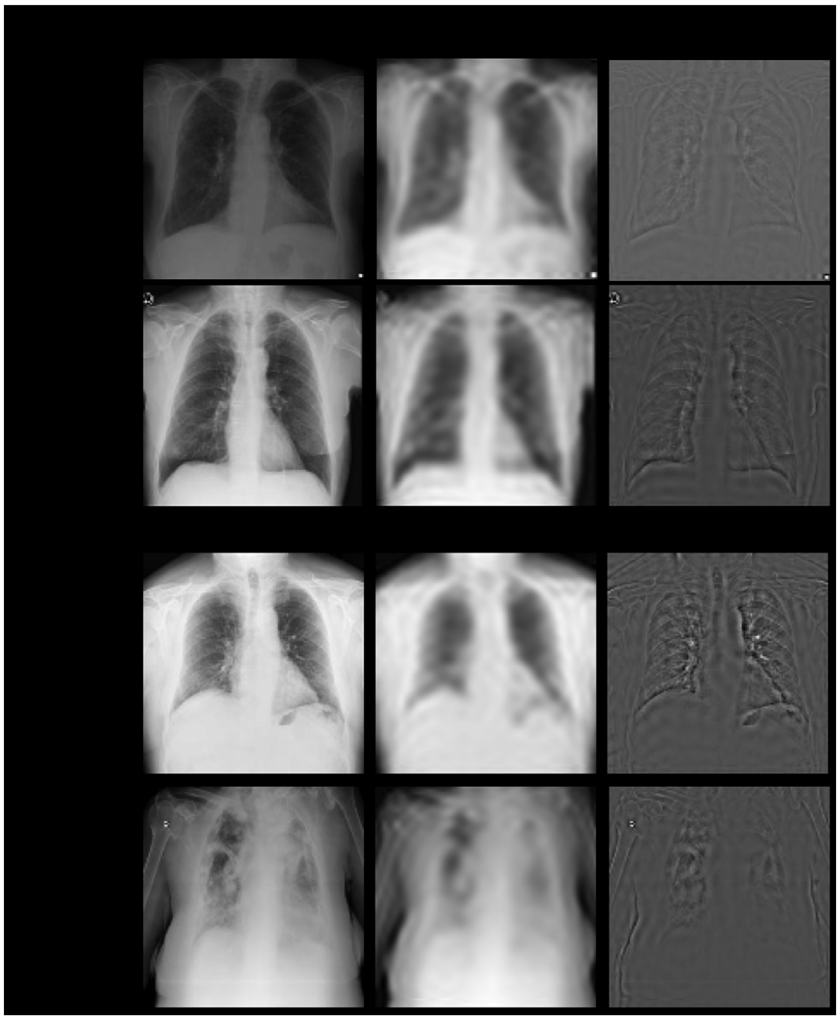

FIGURE 9 Examples of the raw images, the estimated pervasive components, and the estimated specific components from the test chest X-ray images.

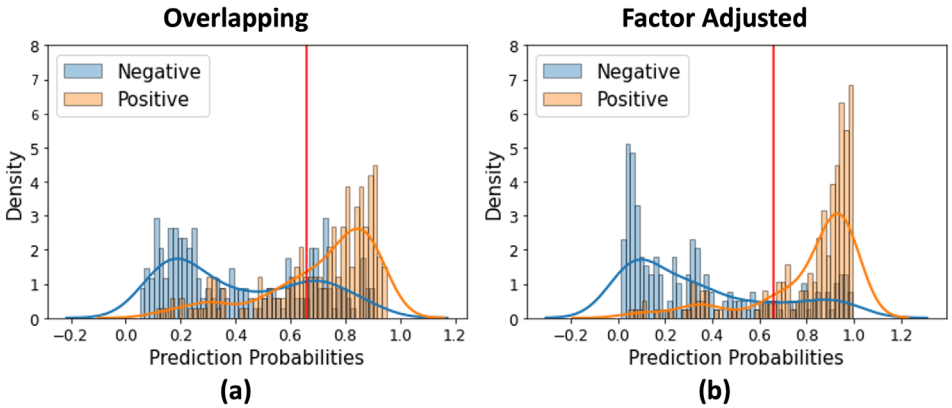

FIGURE 10 Histograms and density plots of classification probabilities for test images: (a) the modified VGG16 model trained on overlapping and (b) factor adjustment method. In both panels, the red vertical line represents the classification threshold.

6 | CONCLUSION AND DISCUSSION

This paper is motivated by a COVID-19 chest X-ray image classification problem, where the positive and negative classes are largely overlapping as the images are dominated by common noises such as skeletons, organs and shadows. The specific signals that can separate the two classes are

weak and obscured by pervasive noises. We introduced a tensor factor model that decomposes the augmented samples into pervasive and specific components. The pervasive component is characterised by the production of latent tensor factors and factor loading matrices. We proposed a matricisation plus PCA strategy to estimate this tensor factor model and investigated the theoretical properties of our estimators. Then, we developed a factor adjustment procedure to remove the pervasive component and apply the classifier directly to the specific components. The intuition is to adjust the noises present in the pervasive component and improve the signal-to-noise ratio in classification. The empirical performance of this procedure is well justified by various synthetic experiments. In the application of frontal chest X-ray image-based COVID-19 diagnosis, our method improves the baseline by 10.5% in terms of classification accuracy.

In this discussion, we address a significant suggestion from the reviewer regarding the potential integration of the tensor factor model directly into the deep neural network classifier's training pipeline. This would entail an end-to-end system where preprocessing, via the tensor factor model, and classification are seamlessly connected, possibly by conceptualising the tensor factor model as an autoencoder. Such a framework promises not only to streamline processing but also to enhance learning efficacy by jointly optimising the preprocessing and classification tasks. While intriguing, this approach extends beyond the scope of our current research. However, it represents a compelling direction for future studies, promising to bridge the gap between preprocessing and classification in deep learning workflows.

AUTHOR CONTRIBUTIONS

Xiaochuan Li contributed to the methodology development, algorithm validation, numerical experiments, and writing for this paper. Bingnan Li contributed to numerical experiments and writing for the paper. Wenzhan Song contributed to methodology validation and application for the paper. Yuan Ke contributed to problem formulation, methodology development, theoretical analysis, design of experiments, and writing for the paper.

ACKNOWLEDGEMENTS

The authors would like to thank the Chief Editor, Associate Editor and anonymous reviewers for their comments which significantly improved the paper. Song's research is partially supported by NSF-1940864, NSF-2019311, NSF-2324389, NSF-2312974, Georgia Research Alliance and NIH-1R01HL172291-01. Ke's research is partially supported by NSF-2210468, NSF-2243044, NSF-2324389 and NIH-1R01HL172291-01.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in BIMCV-COVID-19 (https://github.com/BIMCV-CSUSP/BIMCV-COVID-19.git).

ORCID

Bingnan Li https://orcid.org/0000-0002-7131-9428

ENDNOTE

[1] Avaliable at https://www.kaggle.com/c/siim-covid19-detection

REFERENCES

Ahn, S. C., & Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, *81*(3), 1203–1227.
Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*(1), 135–171.
Bai, J., & Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, *40*(1), 436–465.
Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221.
Chamberlain, G., & Rothschild, M. (1982). *Arbitrage, factor structure, and mean-variance analysis on large asset markets*: National Bureau of Economic Research Cambridge, Mass., USA.
Chang, J., Guo, B., & Yao, Q. (2015). High dimensional stochastic regression with latent factors, endogeneity and nonlinearity. *Journal of Econometrics*, *189*(2), 297–312.
Chen, E. Y., & Fan, J. (2021). Statistical inference for high-dimensional matrix-variate factor models. *Journal of the American Statistical Association*, *2021*, 1–18.
Chen, E. Y., Tsay, R. S., & Chen, R. (2019). Constrained factor models for high-dimensional matrix-variate time series. *Journal of the American Statistical Association*.
Chen, R., Yang, D., & Zhang, C.-H. (2022). Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, *117*(537), 94–116.
Fan, J., Fan, Y., & Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, *147*(1), 186–197.
Fan, J., Liao, Y., & Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics*, *39*(6), 3320.
Goyal, D. K., Mansab, F., Iqbal, A., & Bhatti, S. (2020). Early intervention likely improves mortality in COVID-19 infection. *Clinical Medicine*, *20*(3), 248.
Hasan, K. M., Walimuni, I. S., Abid, H., & Hahn, K. R. (2011). A review of diffusion tensor magnetic resonance imaging computational methods and software tools. *Computers in Biology and Medicine*, *41*(12), 1062–1072.
He, J., Sun, G., Zhang, Y., & Geng, T. (2016). Data recovery in heterogeneous wireless sensor networks based on low-rank tensors. In *2016 IEEE Symposium on Computers and Communication (ISCC)*, IEEE, pp. 616–620.

He, Y., Li, L., & Trapani, L. (2022). Statistical inference for large-dimensional tensor factor model by weighted/unweighted projection. arXiv preprint arXiv: 2206.09800.
Huang, J., Zhang, Y., Zhang, J., & Zhang, X. (2018). A tensor-based sub-mode coordinate algorithm for stock prediction. In *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, IEEE, pp. 716–721.
Ismael, A. M., & Şengür, A. (2021). Deep learning approaches for COVID-19 detection based on chest X-ray images. *Expert Systems with Applications*, *164*, 114054.
Jain, R., Gupta, M., Taneja, S., & Hemanth, D. J. (2021). Deep learning based detection and analysis of COVID-19 on chest X-ray images. *Applied Intelligence*, *51*(3), 1690–1700.
Lakhani, P., Mongan, J., Singhal, C., Zhou, Q., Andriole, K. P., Auffermann, W. F., Prasanna, P., Pham, T., Peterson, M., & Bergquist, P. J. (2021). The 2021 SIIM-FISABIO-RSNA machine learning COVID-19 challenge: Annotation and standard exam classification of COVID-19 chest radiographs.
Lam, C., & Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics*, *2012*, 694–726.
Li, H., Zeng, N., Wu, P., & Clawson, K. (2022). Cov-Net: A computer-aided diagnosis method for recognizing COVID-19 from chest x-ray images via machine vision. *Expert Systems with Applications*, *207*, 118029.
Li, X., & Ke, Y. (2022). Privacy preserving and communication efficient information enhancement for imbalanced medical image classification. In *Annual Conference on Medical Image Understanding and Analysis*, Springer, pp. 663–679.
Litmanovich, D. E., Chung, M., Kirkbride, R. R., Kicska, G., & Kanne, J. P. (2020). Review of chest radiograph findings of COVID-19 pneumonia and suggested reporting language. *Journal of thoracic imaging*, *35*(6), 354–360.
Mor, U., Cohen, Y., Valdés-Mas, R., Kviatcovsky, D., Elinav, E., & Avron, H. (2022). Dimensionality reduction of longitudinal OMICS data using modern tensor factorizations. *PLOS Computational Biology*, *18*(7), e1010212.
Ran, B., Tan, H., Wu, Y., & Jin, P. J. (2016). Tensor based missing traffic data completion with spatial–temporal correlation. *Physica A: Statistical Mechanics and its Applications*, *446*, 54–63.
Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, *97*(460), 1167–1179.
Sun, Q., Qiu, H., Huang, M., & Yang, Y. (2020). Lower mortality of COVID-19 by early recognition and intervention: Experience from Jiangsu province. *Annals of intensive care*, *10*(1), 1–4.
Tsai, E. B., Simpson, S., Lungren, M. P., Hershman, M., Roshkovan, L., Colak, E., Erickson, B. J., Shih, G., Stein, A., & Kalpathy-Cramer, J. (2021). The RSNA international COVID-19 open radiology database (RICORD). *Radiology*, *299*(1), E204–E213.
Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, *31*(3), 279–311.
Vayá, M. I., Saborit, J. M., Montell, J. A., Pertusa, A., Bustos, A., Cazorla, M., Galant, J., Barber, X., Orozco-Beltrán, D., & García-García, F. (2020). BIMCV COVID-19+: A large annotated dataset of RX and CT images from COVID-19 patients. *arXiv preprint arXiv:2006.01174*.
Wang, D., Liu, X., & Chen, R. (2019). Factor models for matrix-valued high-dimensional time series. *Journal of econometrics*, *208*(1), 231–248.
Zhang, A., & Han, R. (2019). Optimal sparse singular value decomposition for high-dimensional high-order data. *Journal of the American Statistical Association*, *114*(528), 1708–1725.
Zhang, C., Fanaee-T, H., & Thoresen, M. (2021). Feature extraction from unequal length heterogeneous EHR time series via dynamic time warping and tensor decomposition. *Data Mining and Knowledge Discovery*, *35*(4), 1760–1784.
Zhang, Q., Berry, M. W., Lamb, B. T., & Samuel, T. (2009). A parallel nonnegative tensor factorization algorithm for mining global climate data. In *International conference on computational science*, Springer, pp. 405–415.

AUTHOR BIOGRAPHIES

Xiaochuan Li is a Quantitative Analytics Specialist at Wells Fargo. She recently received her Ph.D. in Statistics degree from the University of Georgia.

Bingnan Li is a Ph.D. student in the Department of Statistics at the University of Georgia. He contributed to numerical experiments and writing for the paper.

Wenzhan Song is the Georgia Power Mickey A. Brown Professor in the School of Electrical and Computer Engineering at the University of Georgia.

Yuan Ke is an Associate Professor in the Department of Statistics, at the University of Georgia.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Li, X., Li, B., Song, W., & Ke, Y. (2024). Tensor factor adjustment for image classification with pervasive noises. *Stat*, *13*(3), e705. https://doi.org/10.1002/sta4.705