Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Advances in data preprocessing for bio-medical data fusion: An overview of the methods, challenges, and prospects

Shuihua Wang [a], M. Emre Celebi [b], Yu-Dong Zhang [c], Xiang Yu [c], Siyuan Lu [c], Xujing Yao [c], Qinghua Zhou [c], Martínez-García Miguel [d], Yingli Tian [e,*], Juan M Gorriz [f,g,*], Ivan Tyukin [a,*]

[a] *School of Mathematics and Actuarial Science, University of Leicester, Leicester, LE1 7RH, UK*
[b] *Department of Computer Science, University of Central Arkansas, AR, 72035, USA*
[c] *School of Informatics, University of Leicester, Leicester, LE1 7RH, UK*
[d] *Department of Aeronautical and Automotive Engineering, Loughborough University, Loughborough, LE11 3TU, UK*
[e] *Department of Electrical Engineering, The City College of New York, New York, 10031, USA*
[f] *Department of Psychiatry, University of Cambridge, Cambridge, CB2 1TN, UK*
[g] *Data Science and Computational Intelligence Institute, Granada, 52005, Spain*

ARTICLE INFO

Keywords:
Data fusion
Noise
Data scarcity
High dimensionality
Missing data
Small dataset

ABSTRACT

Due to the proliferation of biomedical imaging modalities, such as Photo-acoustic Tomography, Computed Tomography (CT), Optical Microscopy and Tomography, etc., massive amounts of data are generated on a daily basis. While massive biomedical data sets yield more information about pathologies, they also present new challenges of how to fully explore the data. Data fusion methods are a step forward towards a better understanding of data by bringing multiple data observations together to increase the consistency of the information. However, data generation is merely the first step, and there are many other factors involved in the fusion process like noise, missing data, data scarcity, and high dimensionality. In this paper, an overview of the advances in data preprocessing in biomedical data fusion is provided, along with insights stemming from new developments in the field.

1. Introduction

Due to the proliferation of biomedical imaging modalities [1] such as Photo-acoustic Tomography (PAT) [2], Computed Tomography (CT) [3, 4], Optical Microscopy and Tomography (OMT) [2], Single Photon Emission Computed Tomography (SPECT) [5], Magnetic Resonance [6] (MR) Imaging, Ultrasound, Positron Emission Tomography (PET) [7,8], Magnetic Particle Imaging (MPI) [9], Electroencephalogram (EEG) [10]/ Magnet-encephalography (MEG) [11], Electron Tomography (ET) [12], and Atomic Force Microscopy (AFM) [13], massive amounts of biomedical and health informatics data are being generated on a daily basis. It is commonly known that it is difficult to gain full understanding of the data through a single analysis modality. Take, for example, a malignant tumor, which is difficult to diagnose through a single modality for many reasons, like the low positive predictive values, low specificity, etc. Therefore, it is necessary to exploit the information provided by multiple modalities simultaneously for better diagnosis. The acquisition of multimodal data is an important initial step of the process. In many instances, however, the real crux of the problem is how to fully explore all sources of information available. Data fusion provides a step forward towards a complete understanding of a given pathology.

Data fusion is inspired by how humans and animals process sensory signals by merging multiple inputs from different internal and external sensors to reliably collect information about their environment for survival purposes. Data fusion has been widely used in many fields, such as geographic information systems [14,15], wireless sensor networks [16–18], chem-informatics [19], and bioinformatics [20,21]. Data fusion involves the integration of data from different resources to interact and inform each other to enhance a variety of data analysis tasks such as detection, estimation, segmentation, and classification. Data fusion can be carried out at different levels [22–24], including raw data-level, feature-level, and decision-level. However, performing a data

* Corresponding authors.
E-mail addresses: shuihuawang@ieee.org (S. Wang), ecelebi@uca.edu (M.E. Celebi), yudongzhang@ieee.org (Y.-D. Zhang), xy144@le.ac.uk (X. Yu), sl672@le.ac.uk (S. Lu), xy147@le.ac.uk (X. Yao), qz105@le.ac.uk (Q. Zhou), m.martinez-garcia@lboro.ac.uk (M.-G. Miguel), ytian@ccny.cuny.edu (Y. Tian), gorriz@ugr.es (J.M. Gorriz), it37@le.ac.uk (I. Tyukin).

https://doi.org/10.1016/j.inffus.2021.07.001
Received 24 May 2021; Received in revised form 4 July 2021; Accepted 5 July 2021
Available online 10 July 2021

fusion task in a particular application field can be extremely challenging. For example, in the field of biomedical data analysis, a number of problems may occur at different fusion levels, including the following aspects:

- Data noise [25]: Data noise is one of the major limitations in imaging and is an important issue for biomedical image preprocessing. Noise can be defined as unwanted information in images [26].
- Missing values [27]: Missing values refer to the absence of data items for a subject. Missing values are pervasive in real-world data sets, and they present a significant challenge at the different levels for multimodal data fusion as they may be of importance to the data analysis task at hand.
- Alignment and registration [28,29]: Alignment and registration aim at reducing spatial or temporal in-homogeneities between samples, including differences in acquisition frequencies, sampling devices, and sample physiology. In biomedical data, registration is a standard prerequisite for the analysis and fusion of multimodal data [30].
- Small datasets [31–33]: Data scarcity can be a problem in domains where data collection is difficult, time-consuming, and/or expensive. A prime example is the scarcity of medical data. Analyzing small data sets and building statistical models using them are both challenging tasks.
- High dimensionality [34]: With the proliferation of diverse modalities, it has been made possible to acquire large amounts of biomedical data with high dimensionality. Processing the high-dimensional data incurs a high computational cost and is inherently inefficient since many of the values that describe a data object are redundant due to noise and linear or nonlinear dependencies. Other potential issues include instabilities [35], distance concentrations, and insufficient data volumes to make use of high-dimensional data [36]. Consequently, the dimensionality, i.e., the number of values that are used to describe a data object, needs to be reduced prior to any subsequent processing of the data.

This paper provides a review of methods to deal with those challenges and their prospects towards the fusion of medical imaging data. The rest of this paper is organized as follows: Section 2 introduces some common biomedical image acquisition methods, including CT, MRI, X-ray, etc. Section 3 describes the challenge of noise and the available solutions. Section 4 provides a review of the missing value problem and imputation methods that deal with it. Then, Section 5 illustrates the alignment and registration methods. Section 6 provides a review of the small datasets issue. Section 7 describes the high dimensionality issue and its solutions. In Section 8, conclusions are drawn and novel trends are discussed. Table 11 provides all the abbreviations and their representations in this paper.

2. Data acquisition methods

Medical imaging [37–39] refers to a range of technologies for visualizing specific parts of the body for clinical diagnosis and medical treatment. Medical imaging can also visualize the function of tissues or organs. With medical imaging technologies, clinicians can investigate the internal structure of the skin and bones as well as diagnose and provide treatment. Medical imaging also assists in building datasets of physiology and normal anatomy to make it possible for researchers to conduct further analyses [40].

Medical imaging is a part of the broader domain of biological imaging, which includes many different types of imaging technologies, such as X-ray [41,42], ultrasound [43,44], magnetic resonance imaging (MRI) [45], nuclear medicine functional imaging techniques -e.g., positron emission tomography (PET) [8,46] and single-photon emission computed tomography (SPECT), etc. The details of some common medical imaging technology are next described.

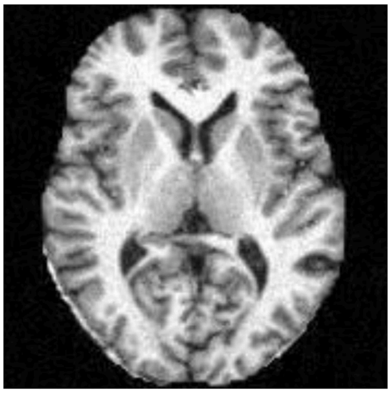


Fig. 1. MRI brain image

2.1. *Magnetic resonance imaging*

Magnetic resonance imaging (MRI) [47], which is a noninvasive medical imaging technique to produce three-dimensional detailed anatomical images, utilizes strong magnetic fields, magnetic field gradients, and radio waves to produce pictures of the anatomy and the physiological processes of the body. The patient to be scanned should be positioned within an MRI scanner that forms a strong magnetic field around the specific area of interest. The specific region is then defined by the X and Y gradient coils with energy caused by an oscillating magnetic field temporarily applied at the appropriate resonance frequency. The receiving coil then measures the radio frequency (RF) signal level emitted by the excited atoms. The RF signal can be used to infer the position information as the RF level and phase change due to the changing the local magnetic field by gradient coils. The contrast between various tissues is determined by the rate of excited atoms returning to the equilibrium state. Patients might be given contrast agents, like gadolinium, to make the image more clear [48].

The advantage of the MRI [49] is that it does not rely on ionizing radiation or X-rays, which are harmful and may cause direct tissue damage or cancer. MRI is an outstanding imaging technology in regard to image details, though the scanning process takes long time and produces loud noises. MRI has been widely used to image joints [50,51], brain [52], wrists [53], ankles [54,55], breasts [56], heart [6,57] and blood vessels [58]. However, MRI is usually expensive and may not be able to offer the resolution and enough information to detect all types of cancers, such as breast cancer which is indicated by micro-calcifications; currently, it cannot differentiate the benign disease and malignant tumors. Moreover, some patients might be allergic to the contrast agents or have chronic kidney disease, which prevents them from ingesting these agents [59]. In addition, it may be unsafe for a patient to go through the MRI scanner if the patient has medical implants or other non-removable metal inside their body [60]. Fig. 1 shows an MRI brain image.

2.2. *Computed tomography*

Computed tomography (CT) [3,61,62] uses rotating X-ray machines and computers to create cross-sectional images to visualize different body parts, including the head, shoulders, spine, and heart. CT provides a non-invasive way to visualize the inside of the body. CT can show the details of damage to bones, injuries of an internal organ, problems with blood flow, stroke, and cancer. For instance, CT can provide information about size, location, and shape of a tumor prior to radiotherapy or to guide needle biopsies.

Since its introduction in the 1970s, CT has become a significant technique to supplement X-rays and ultrasonography in medical imaging [63]. CT has the following advantages: (1) CT can provide high image resolution, therefore, better details. (2) CT can exclude the

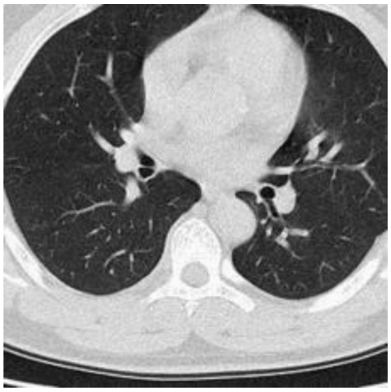


Fig. 2. CT brain image

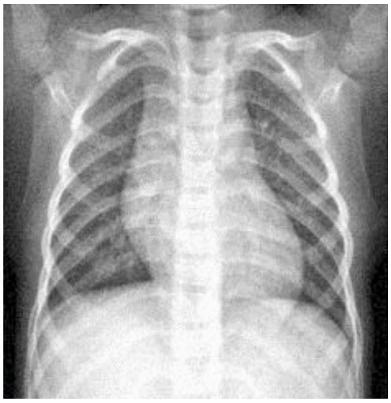


Fig. 3. X-ray image from a healthy subject

superimposition of images of structures outside of the region of interest. (3) CT imaging data can be viewed in multi-planar transverse, coronal, or sagittal plane depending on the corresponding diagnostic task.

However, one in 80 people could be at a risk of developing cancer due to being subjected to CT scans [4]. It is estimated by one study that 0.4% of the cancers in the United States were, in fact, caused by CT scanning procedures. Though a study by Tubiana [61] disputed the estimate as there is no consensus that the low level of radiation used in CT scans causes damage. CT is therefore not usually recommended unless the patient exhibits certain symptoms. Fig. 2 shows a CT brain image.

2.3. X-ray

X-ray waves, found in 1895 by Röntgen, are one type of high-energy electromagnetic radiation. X-ray waves have been widely used for medical imaging since their introduction as they can pass through the body to create images of different parts of your body by variable shades of black and white [42,64,65]. For patients to be scanned, they need to be positioned so that the body part to be imaged is located between an X-ray source and an X-ray detector. When X-rays pass through the body, they can be absorbed at different rates due to the different densities of different body parts. Then, an image can be generated as a detector on the other side of the body picks up the X-rays after they pass through the body. When the X-rays pass through high-density body parts, such as bones, they will be shown as clear white areas on the image. In contrast, low-density parts, such as lungs and hearts, it will be shown as darker areas on the image.

X-ray imaging is widely used for the examination of bone fractures and breaks [62], tooth problems such as root infection and loose teeth [66], and scoliosis [67]. X-rays are also commonly used as an imaging method to produce mammograms for detecting breast cancers. X-ray imaging [68] is painless, fast, and non-invasive. However, X-rays expose the patients to radiation. Therefore, it should be used judiciously. Fig. 3 shows an X-ray image from a healthy subject.

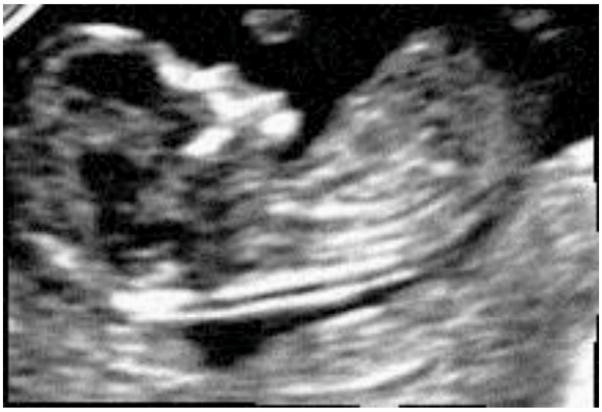


Fig. 4. Fetal Ultrasound

2.4. Ultrasound

The principle of Ultrasound is that high-frequency sound can travel through soft tissues and fluids, and then it bounces back or echoes off denser surfaces to generate images [68]. The echoes determine the ultrasound image features in shades of gray, which reflect different densities as more ultrasound bounces back when hitting a denser object. Different ultrasound frequencies can generate images with different qualities. For example, high frequencies can provide high-quality images, but they are more readily absorbed by the skin and other tissue, and thus, they cannot penetrate as deeply as lower frequencies. As Ultrasound uses radio waves instead of radiation to form images, it is much safer compared to X-ray and CT. Ultrasound is suitable for use during pregnancy to monitor the baby's development. 3D ultrasound can provide a static 3D image of the baby, while 4D can provide a moving video. Besides monitoring the fatal development, ultrasound can be used for the diagnosis of internal organs, such as the liver, kidneys, and thyroid nodules. Fig. 4 shows a fetal ultrasound image.

2.5. Positron emission tomography

Positron emission tomography (PET) is a nuclear medicine imaging, which is based on the radioactive substances known as the radiotracers to provide clear visualization of the changes in metabolic processes, blood flow, regional chemical composition, absorption, etc. [69]. With

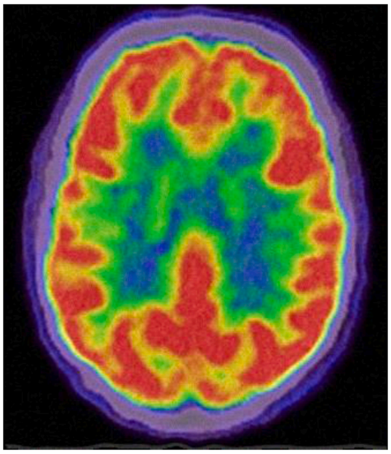


Fig. 5. PET normal brain image

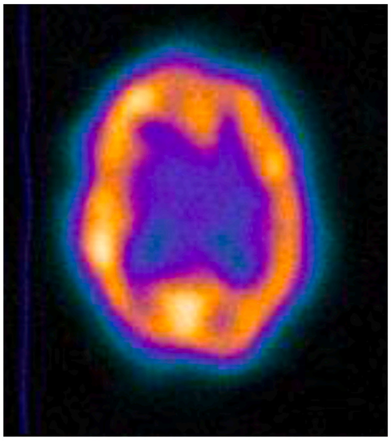


Fig. 6. SPECT brain image

the injection a small amount of liquid radioactive material as the tracer into the body, gamma rays by tracers are then emitted and detected by the gamma cameras to generate a 3D image, which is similar to an X-ray image. Variable tracers may be used depending on the purpose of the scan.

Due to the high cost and complexity of the support infrastructure, like cyclotrons, PET scanners, etc., PET was mainly used for researches in the past. However, in recent years, due to the advanced technology and the proliferations of PET scanners, PET is also employed in clinical applications to help with disease diagnosis, which can help improve the understanding of disease pathogenesis. PET can be applied for the diagnosis of movement disorders, epilepsy, brain tumors, stroke and neuronal plasticity, neuropharmacology, dementia, and some possible future applications with different types of tracers. Usually, PET can be used together with CT or MRI to help doctors to get a more detailed view of the illness, and therefore to get a better assessment of the patient's condition. Fig. 5 shows an example of PET normal brain image.

2.6. Single-photon emission computed tomography

Single-photon emission computed tomography (SPECT) is another type of nuclear medicine tomographic imaging technique that is also based on gamma rays [70]. SPECT can provide true 3D information that is traditionally shown as cross-sectional slices through the patients and is free to be reformatted and manipulated according to the application requirements.

For the SPECT imaging, patients need to take an injection of the gamma-emitting radioisotope into the bloodstream. Usually, the radioisotope is a simple soluble dissolved ion, like an isotope of gallium (III). In most cases, a marker radioisotope is used to create radioligand when it is attached to a specific ligand. The properties of the radioligand bind it to specific types of tissues. Then, the coupled combination of ligand and radiopharmaceutical can be carried to bound to the region of interest in the body, followed by the gamma camera can see the ligand concentration.

Different from traditionally taking a picture of the anatomical structure, SPECT allows monitoring of the biological activity at each place in the 3-D region analyzed. The amounts of blood flow are indicated by the emission from the radionuclide in the capillaries of the imaged regions. The images obtained from SPECT imaging by using a gamma camera are multiple 2-D images from different angles. Afterwards, a tomographic reconstruction algorithm is applied to the multiple projections, yielding a 3-D data set. The imaging principle of SPECT is similar to PET as they both use radioactive tracers and the detection of gamma rays. Differently, SPECT emits gamma radiation, which is measured directly, while PET tracers emit positron annihilate with electrons up to a few millimeters away, making two gamma photons to be emitted in opposite directions. As PET imaging can immediately find these emissions coincident as to provide more radiation event localization information, it can provide higher resolution than SPECT. However, SPECT is significantly cheaper than PET as they are able to use longer-lived and more easily obtained radioisotopes.

Table 1
Summary of common biomedical image acquisition methods

Modalities	Imaging method	Advantage	Disadvantage	Application
MRI	Magnetic fields and radio waves	• Less radiation compared to CT and X-ray	• Expensive • Noise • Radiofrequency energy	joints, brain [52], wrists, ankles[54, 55], breasts, heart, blood vessels, and etc.
CT	Ionizing radiation	• Cheaper than MRI • High image resolution • Accurate, fast, and painless	• Potential allergy to the contrast agent • Harmful to the unborn baby	damage to bones, injuries of an internal organ, problems with blood flow, stroke, cancer and etc.
X-ray	Ionizing radiation	• Painless • Fast	• No 3D information • Radiation	bone fractures, tooth problems, scoliosis, lung problems, etc.
Ultrasound	Sound waves	• Safe, quick, and easy • Do not use radiation	• Fewer details as X rays • cannot be applied in areas that contain gas (such as lungs) • doesn't pass through bones	diagnosis of internal organs fetal development
PET	Radiotracers	• Painless, noninvasive	• Cause a major allergic reaction • Harmful to babies if pregnant	Detect cancer at an earlier stage
SPECT	gamma rays	• Less time compared to PET • Cheaper than PET	• Long scan times • Low-resolution images	Monitor brain disorders, heart problems and bone disorders.

SPECT can be utilized as a complement of any gamma imaging study as it can provide a true 3D representation, such as tumor imaging, infection (leukocyte) imaging, thyroid imaging, or bone scintigraphy. As SPECT provides accurate localization in 3D space, it can be used to provide information about localized function in internal organs, like functional cardiac or brain imaging. Fig. 6 shows a SPECT image from a patient with uncontrolled complex partial seizures.

Table 1 shows a summary of the common biomedical image methods, including MRI, CT, X-ray etc., with their corresponding imaging method, advantages, disadvantages and their applications.

3. Preprocessing of noisy data

3.1. Background

Data noise is one of the major factors affecting the quality of imaging outputs, and addressing the negative impact of noise is an important step for biomedical image processing. Noise can be defined as any unwanted information in images. Consider, for example, an image where input sensory information is presented as a grey-level matrix or tensor, then, an element in an image can be expressed as a pair

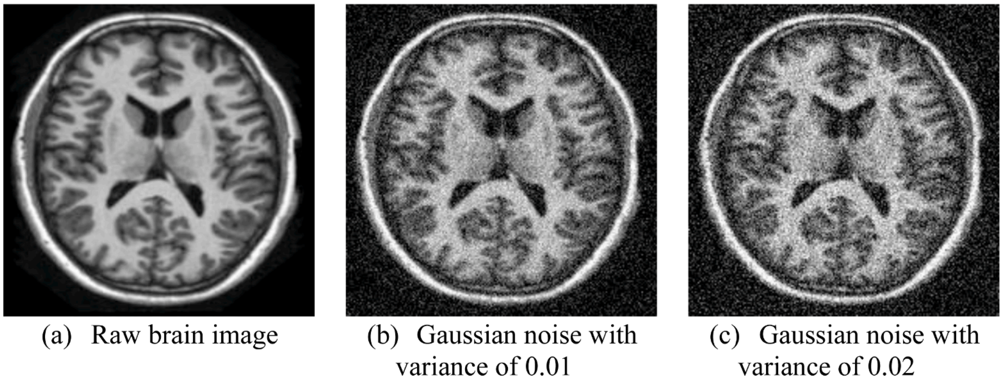

(a) Raw brain image (b) Gaussian noise with variance of 0.01 (c) Gaussian noise with variance of 0.02

Fig. 7. Gaussian noise

$$(i, v(i)) \tag{1}$$

where $i \in I$, denotes the coordinate and $v(i)$ represents the corresponding grey level. Note that the number of variables defining the coordinate (sometimes referred to as a dimension of) i can vary depending on specific tasks and processes. $v(i)$ is a real value in grey level images while in color images $v(i)$ is a triplet for red, green, and blue channels, respectively. The value of every pixel $v(i)$ is obtained by the measurement of light intensity, which can be implemented by a charged-coupled device (CCD) matrix. The capacitors in the CCD device count the number of photons in a period of time to generate the intensities. According to the central limit theorem, the numbers of photons received by the captors fluctuate around their mean values if the subject of the image is in constant light. However, in real-life applications, the capacitors can receive bogus heat photons if they are not cooled down appropriately. In such circumstances, noise appears in the final image, which can be expressed as

$$v(i) = u(i) + n(i) \tag{2}$$

where $v(i)$ is the value obtained by observation, $u(i)$ is the original true value and $n(i)$ stands for the noise value. Various factors can contribute to the noise $n(i)$, such as calibration error and quantization degradation, which are unavoidable in measurement. Specifically, for biomedical imaging like CT, there can be random noise, electronic noise, statistical noise, and round-off noise.

The quality of CT images is related to several factors. For example, inappropriate protocol parameter values and the movement of patients can blur the reconstructed images. The movement is sometimes unavoidable in practical applications due to the breathing and heart beating. Field of view is also a significant factor for CT imaging. The reconstructed images can be degraded if the field of view is too small or too big. Artifact is another major factor for CT, which is defined as the difference between the desired CT numbers and the obtained CT numbers [71]. The quality of MRIs is also related to a bunch of factors, such as movement of the objects and scanning times [72].

In the remainder of this section, we focus on image noise and denoising methods. The common noise models are discussed, and the state-of-the-art denoising techniques are presented, including wavelet-based methods, Markov random field-based algorithms, anisotropic diffusion filtering, non-local methods, bilateral and trilateral methods, and deep learning-based denoising.

3.2. Noise models

The prior knowledge of noise models is beneficial for denoising processing. Since the image noise often appears randomly, it is suitable to describe it by random variables and probability density functions (PDFs).

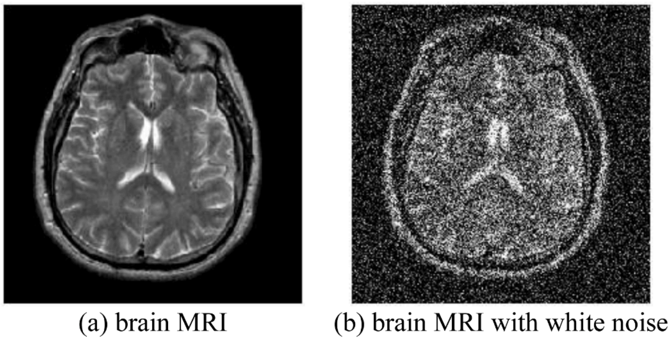

(a) brain MRI (b) brain MRI with white noise

Fig. 8. White noise

3.2.1. Gaussian noise

The Gaussian noise model is often used to simulate thermal noise. For the univariate Gaussian noise n, its PDF is written as

$$p_n(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty \tag{3}$$

where μ denotes the mean value and σ^2 represents the variance. In digital images, x stands for the gray-level, so it is non-negative and often defined as integer $x \in [0,255]$. Fig. 7(a) shows a raw brain image extracted from Open Access Series of Imaging Studies (OASIS) brain dataset [73]. Fig. 7(b-c) present the Gaussian noise injection results with variance of 0.01 and 0.02, respectively.

3.2.2. White noise

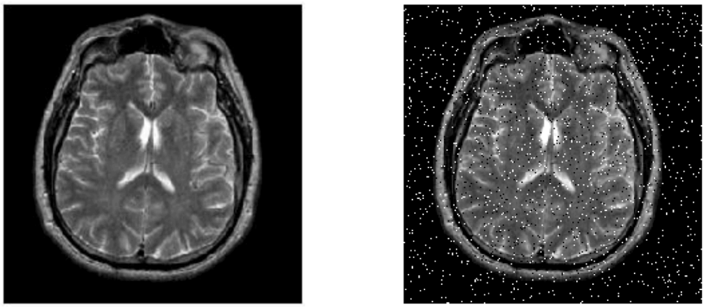
Gaussian noise is defined by its PDF, but white noise is based on the noise power. From the view of the spectrum, white noise power is a constant value. In an image with white noise, the intensity value of each pixel is different from its neighboring values. Fig. 8 presents an example of a brain MRI image with white noise.

3.2.3. Impulse valued noise

Impulse valued noise, also known as salt and pepper noise, is another type of image noise that is commonly seen during transmission. The definition can be expressed as

$$q_{SP}(i,j) = \begin{cases} \gamma & x(i,j) = N_s \vee N_p \\ 1-\gamma & \text{otherwise} \end{cases} \tag{4}$$

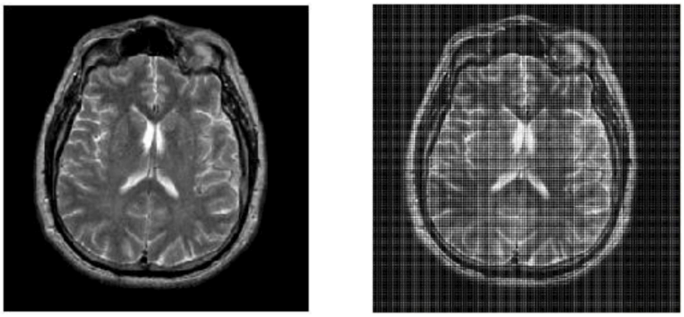
where $N_s = 255$ denotes sault noise, and $N_p = 0$ denoting pepper noise. $x(i,j)$ represents the pixel value at position (i,j) after salt-and-pepper noise is added to the original image. γ means the noisy density, which is a factor meaning how many percentages of all pixels will add salt-and-pepper noise. The salt and pepper noise does not corrupt the whole image but changes parts of the pixel values. Because in data trans-



(a) brain MRI

(b) brain MRI with impulse valued noise (25% of the pixels corrupted)

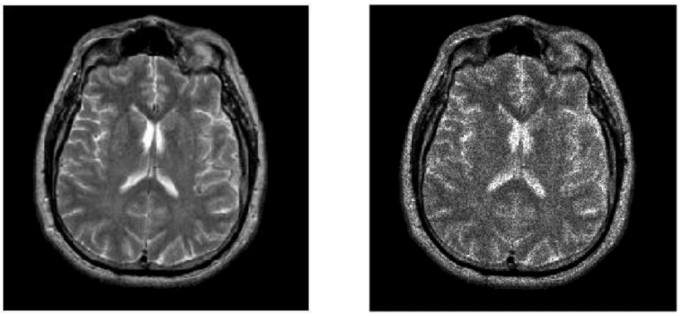
Fig. 9. Impulse valued noise



(a) brain MRI

(b) brain MRI with periodic noise

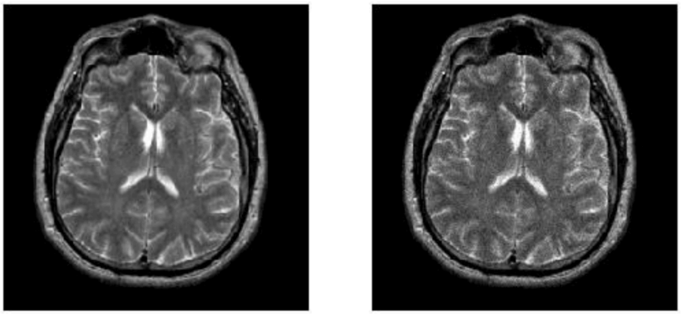
Fig. 10. Periodic noise



(a) brain MRI

(b) brain MRI with speckle noise

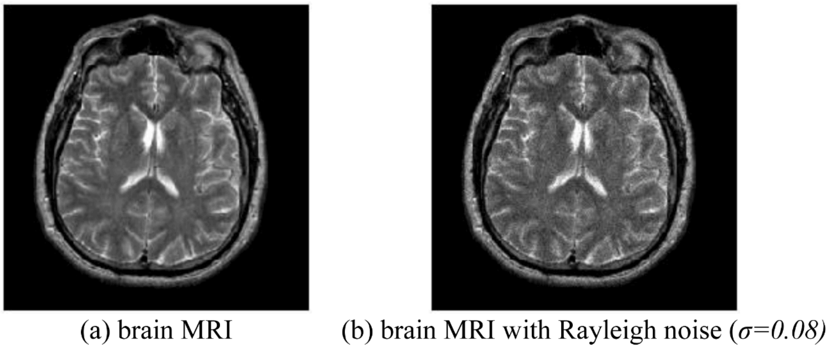
Fig. 11. Speckle noise



(a) brain MRI

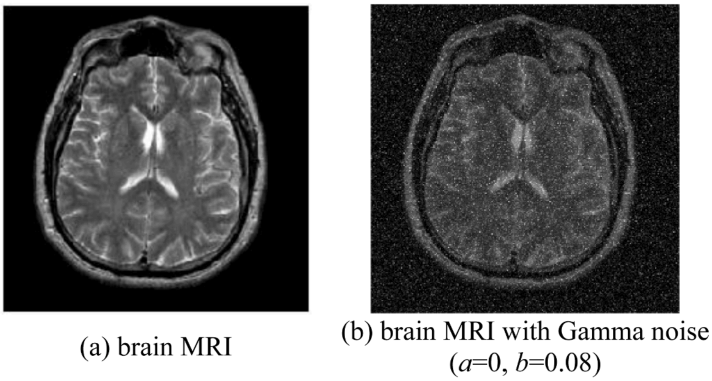
(b) brain MRI with Poisson noise

Fig.12. Poisson noise


(a) brain MRI

(b) brain MRI with Rayleigh noise (σ=0.08)

Fig. 13. Rayleigh noise


(a) brain MRI

(b) brain MRI with Gamma noise (a=0, b=0.08)

Fig. 14. Gamma noise

mission, some values of pixels can be corrupted and substituted by either pure black or pure white values. Fig. 9 illustrates an example of an MRI brain image and its contamination with impulse valued noise.

3.2.4. *Periodic noise*

Periodic noise is caused by electric interference during the image capture process. Periodic noise has a certain pattern that repeats independently in the spatial domain. Fortunately, periodic noise can be easily removed by filtering in the frequency domain. An instance of sine periodic noise is provided in Fig. 10.

3.2.5. *Speckle noise*

Speckle noise is multiplicative noise, which is usually caused by bad information channels. As this noise is multiplicative with the original signals, it appears with the signals and disappears when the pixel values are zero. The speckle noise is modelled as multiplicative noise, defined as

$$I_o = I_f + N_m I_f + N_a \tag{5}$$

where I_f means noise-free image, and I_o the observed image. N_m and N_a denotes the multiplicative noise and additive noise, respectively. An example of speckle noise with 0.05 variance is given in Fig. 11.

3.2.6. *Poisson noise*

Poisson noise is so named because it obeys the Poisson distribution. Poisson noise is caused by the quantum characteristic of light. The number of the quanta that arrives on the surface of the photoelectric detection device follows statistical fluctuation so that image is granular. As a result, the contrast of the image diminishes, and the detailed information is covered. An instance is presented in Fig. 12.

Table 2
summary of different noise types

Type of noise	Description
Gaussian noise	Gaussian noise obeys Gaussian distribution, which can be defined by mean and variance.
White noise	White noise is defined based on the noise power, which is a constant value.
Impulse valued noise	Impulse valued noise is also known as salt and pepper noise, which is often seen during transmission.
Periodic noise	Periodic noise has certain pattern that repeats independently in spatial domain, which can be easily removed by filtering in frequency domain.
Speckle noise	Speckle noise is multiplicative noise, which is usually caused by bad information channels.
Poisson noise	Poisson noise obeys the Poisson distribution, which is caused by the quantum characteristic of light.
Rayleigh noise	Rayleigh noise is often seen in radar images.
Gamma noise	Gamma noise often occurs in laser images.

3.2.7. *Rayleigh noise*

Rayleigh noise is often seen in radar images, and its PDF is given by

$$p_n(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, \ x \geq 0 \tag{6}$$

where the σ^2 represents the variance. Fig. 13 presents an example of Rayleigh noise.

3.2.8. *Gamma noise*

Gamma noise often occurs in laser images with the PDF given by

$$p_n(x) = \frac{a^b x^{b-1}}{(b-1)!} e^{-ax}, \ x \geq 0 \tag{7}$$

where $a > 0$ and b is a positive integer. Fig. 14 illustrates an instance of

Gamma noise.

Considering the length of this review, only the above eight common noise models are discussed. Table 2 offers a summary of these different types of noise.

3.3. Denoising Methods

Extensive research has been done for image denoising. In this section, we present the well-known denoising algorithms, and these methods sometimes are combined to get better denoising results.

3.3.1. Wavelet-based methods for denoising

Wavelet transform is the most widely used method for signal analysis, which offers multi-resolution analysis by scale-space domain transform. Two factors determine the wavelet transform result: wavelet basis function and decomposition level. The wavelet basis function is responsible for generating components of different frequencies. The decomposition level controls the threshold for the wavelet transform. Generally, wavelet-based denoising methods contain the following steps: wavelet decomposition, thresholding, and wavelet reconstruction. Denoising by wavelet algorithms is computationally efficient, and it requires less manual intervention as there is no parameter tuning. Edge preservation ability of wavelet methods is also outstanding. Hence, wavelet transform has been used for image denoising, enhancement, and feature extraction.

Kazubek [74] suggested using Wiener filtering to analyze the coefficients of the wavelet transform. Portilla, et al. [75] proposed a scale mixture denoising model in the wavelet domain. Firstly, the noisy image was decomposed into wavelet coefficients, and the covariance in the local neighborhood was estimated by a Gaussian vector and a hidden positive scalar multiplier. Then, a Bayesian based estimation algorithm was leveraged for removing Gaussian noise. Finally, the denoised image was obtained by wavelet reconstruction. Ghazel, et al. [76] conducted a detailed study on fractal wavelet coding for image denoising and restoration. Gruber, et al. [77] proposed to locally decompose the embedded noisy signals into lower dimension space and use local independent component analysis to remove the noise. Afterwards, the noise-free signals can be obtained by reconstruction. They also put forward a delayed algorithm for multiple unknown signals extraction for denoising. In the experiment of denoising in nuclear magnetic resonance spectra, the performance of their denoising algorithms was compared with kernel principle component analysis. Luisier, et al. [78] suggested employing Stein's unbiased risk estimate to obtain the weights in their model and proposed a wavelet thresholding algorithm. Khmag, et al. [79] designed a cluster-based denoising method in the wavelet domain. The coefficients from the second level of wavelet decomposition were used to generate sparse multi-resolution features from the noisy images. The clustered coefficients were linked based on the sparsity as well as self-similarity information. Bao, et al. [80] combined the wavelet transform with a deep learning algorithm. Firstly, the monogenic wavelet transform was selected as the feature extractor to generate amplitude and phase representations from the noisy images. Then, these representations served as the input to a deep deconvolutional neural network model for denoising. Finally, the denoised coefficients formed the clean image by inverse transform. Chen, et al. [81] proposed a new indicator called weight sum variance of digital number probability (WSVODP), which is only related to the difference of the sensors. The proposed WSVODP was capable of determine the optimal wavelet filter coefficients for denoising. Gökdağ, et al. [82] used wavelet transform to remove white Gaussian noise from confocal laser scanning microscopy images. They developed a systematic algorithm to get the best parameters for wavelet thresholding and utilized the analysis of variance to monitor the interactions between these parameters. Wavelet transform can also be combined with swarm optimization algorithm for denoising [83].

3.3.2. Markov random field-based methods

Markov random field (MRF) is a popular graphical model for status prediction. In an MRF, the status of a certain position is only dependent on the status of its neighboring positions but independent of any other units. Naturally, an image can be seen as an MRF, with the intensity values being the status and the coordinates as the positions. Therefore, MRF can be leveraged for image denoising.

Malfait and Roose [84] proposed a denoising method with multivariate probability functions. However, it's difficult to determine those probability functions in practice. Therefore, MRF was utilized to obtain the probabilities in an indirect way. For noise suppression, three probability functions were modeled: a posteriori, a priori, and a conditional probability function. In implementation of the MRF, 3×3 neighborhood was chosen. Hua, et al. [85] used MRF as a regularization method in image denoising. As they employed a two-state Gaussian mixture model, and the dependent relationship in the spatial domain between the wavelet coefficients was specified by the MRF model. To determine the hyper-parameters and configurations, expectation-maximization and iterated conditional modes were leveraged. Experiment results suggested that their denoising algorithm achieved a better signal-to-noise ratio than traditional wavelet transform. Barbu [86] developed a real-time denoising system by a novel active random field training. The active random field was proposed based on MRF with conditional random field. To train this active random field, an optimization algorithm based on supervised learning was proposed. The proposed novel random field technique yielded state-of-the-art performance with thousands of times faster speed so that it can be applied in real-time applications. Cao, et al. [87] designed a three-layer MRF to suppress the image noise. Each layer was aimed at a specific task. The texture regions were embedded in layer-1. In layer-2, the training target was stored, which was the images without noise. The layer-3 is composed of the noisy images. Maximum a posteriori estimation between the layer-1 and layer-2 was implemented by iterated conditional modes. Simulation results revealed that their multi-layer MRF could suppress the noise while maintaining the details in the images. Xu and Shi [88] proposed a denoising algorithm for parallel MRI (pMRI). Fields of experts is a type of high-order MRF, which was used for priors learning in the statistics in pMRIs. A loss-specific training algorithm was also proposed to optimize the parameters in the fields of experts. The experiment was carried out on real data, and the denoising performance of the proposed approach was robust. Lekadir, et al. [89] developed a denoising and fiber reconstruction method for multi-slice cardiac diffusion tensor images (DTIs) based on MRF. The MRF was combined with a statistical constraint for missing fiber and a consistency term to enable the obtained meshes continuous. Their method was evaluated on both synthetic and real data and produced satisfactory results.

Generally speaking, MRF can preserve the texture structures by spatial correlation information effectively. However, the optimization of the MRF models is usually based on iterated conditional modes, which is computationally expensive.

3.3.3. Anisotropic diffusion filtering for denoising

Anisotropic diffusion is used for image smoothing. Unlike the Gaussian blur, anisotropic diffusion is capable of denoising while maintaining the details in the images like edges and corners. The concept of anisotropic diffusion was originally invented in thermal theory. The idea of anisotropic diffusion in image denoising is that the pixels can be regarded as heat flows. If the pixel value is similar to its neighbors, that flow will diffuse to the neighbors. Otherwise, if the difference between the pixel and some of its neighbors is obvious, there can be some edges in the neighbors, so the flow will not diffuse to those directions. Therefore, the edges can be preserved. In essence, anisotropic diffusion filtering is edge-preserving filtering. For an image I, the iteration expression in four directions is

$$I_{t+1} = I_t + \lambda\left(cN_{x,y}\nabla_N(I_t) + cS_{x,y}\nabla_S(I_t) + cE_{x,y}\nabla_E(I_t) + cW_{x,y}\nabla_W(I_t)\right) \tag{8}$$

where t is the iteration time, and the four derivatives and thermal coefficients are defined as

$$\begin{cases} \nabla_N\left(I_{x,y}\right) = I_{x,y-1} - I_{x,y} \\ \nabla_S\left(I_{x,y}\right) = I_{x,y+1} - I_{x,y} \\ \nabla_E\left(I_{x,y}\right) = I_{x-1,y} - I_{x,y} \\ \nabla_W\left(I_{x,y}\right) = I_{x+1,y} - I_{x,y} \end{cases} \tag{9}$$

$$\begin{cases} cN_{x,y} = \exp\left(- \frac{\nabla_N\left(I_{x,y}\right)^2}{k^2} \right) \\ cS_{x,y} = \exp\left(- \frac{\nabla_S\left(I_{x,y}\right)^2}{k^2} \right) \\ cE_{x,y} = \exp\left(- \frac{\nabla_E\left(I_{x,y}\right)^2}{k^2} \right) \\ cW_{x,y} = \exp\left(- \frac{\nabla_W\left(I_{x,y}\right)^2}{k^2} \right) \end{cases} \tag{10}$$

where the λ and k are hyper-parameters. Significant efforts have been made to further improve the performance of basic anisotropic diffusion methods in applications.

Ben Abdallah, et al. [90] found that anisotropic diffusion filtering works when an image is contaminated with speckle noise but fails on other types of noise. The reason is that the noise model was estimated wrong. Based on this finding, they put forward an adaptive anisotropic diffusion filtering (AADF). In AADF, the noise estimation was done at every iteration so that the color noise can be effectively removed. The image quality was improved in their experiment in comparison with conventional anisotropic diffusion filtering as well as fast non-local mean filtering algorithms. Kim, et al. [91] proposed to use region adaptive smoothing strength to improve the quality of the restored images by anisotropic diffusion. In each iteration, an adaptive classifier was trained to obtain a promising estimation on the smoothing strength with respect to the changing noise. The training samples for the classifier were also carefully selected in order to ensure good results. In their implementation, decision tree was selected as the classification algorithm. They also proposed a region analysis approach to reduce the computational complexity. The proposed method yielded better peak signal to noise ratio than several anisotropic diffusion variant techniques. Xia, et al. [12] proposed a denoising technique for phase images based on anisotropic diffusion. They introduced a synthetic noise estimation technique to the anisotropic diffusion to accurately classify the noise pixels from the desired signal pixels so that the diffusion process can be iterated with the corresponding coefficients effectively. The proposed method was evaluated on artificial and real mouse artery images and achieved good denoising performance while preserving detailed information. Beitone, et al. [92] proposed a gradient anisotropic diffusion to reconstruct heat sources from noisy temperature fields. The gradient anisotropic diffusion was optimized to generate the possible heat source in an aluminum plate. Ben Abdallah, et al. [93] developed a segmentation technique for blood vessel images based on anisotropic diffusion. To remove the noise in the RGB fundus images, an adaptive anisotropic diffusion filter was used, with the combination of noise level functions. Then, the images were converted to gray-scale images for blood vessel segmentation. The noise level function was defined as the local variance in the images, which can be calculated intensities of the pixels. In their improved version of speckle noise-reducing anisotropic diffusion, the noise level function values of homogeneous regions were computed for noise estimation. Chen, et al. [94] put forward a denoising algorithm for seismic data analysis with anisotropic diffusion and isotropic diffusion. They found that the conventional Chambolle–Lions anisotropic diffusion (CLAD) method fails in separating noise from features when the characteristic of them is in multiple scales because it is difficult to find an appropriate threshold. Hence, an energy-based dynamic CLAD was developed which can distinguish noise from real information and employ different diffusion strategies for different regions dynamically. The threshold was defined as the mean of gradient magnitude, and it was updated during the iterations of diffusion. Hadj Fredj and Malek [95] studied the oriented speckle reducing anisotropic diffusion (OSRAD) and tried to improve its computational efficiency so that the OSRAD can be applied in real-time denoising. They implemented a CUDA-based OSRAD, which runs on GPU. Compared with traditional OSRAD running on CPU, this CUDA-based OSRAD ran thirty times faster with the same denoising effect, which was better than other denoising algorithms like wavelet and bilateral based methods in their experiment. The removal of speckle noise in ultrasound images poses a major challenge in medical image analysis. Jubairahmed, et al. [96] discovered that conventional anisotropic diffusion could cause the loss of contour information in ultrasound images. They suggested employing contourlet transform to decompose the ultrasound images into coefficients and leverage thresholding for denoising. Then, the denoised coefficients formed the image by reconstruction. Finally, to remove speckle noise, the adaptive nonlinear anisotropic diffusion was performed on the reconstructed image. Kamalaveni, et al. [97] proposed to improve the anisotropic diffusion by dynamic diffusion rate for different regions of images with the aim to maintain more details like lines and corners. Firstly, the structure tensor of each pixel was computed and decomposed to get the eigenvalues and eigenvectors. Then, the maximum and minimum gradient variations for every pixel were generated by its eigenvalues and eigenvectors. Afterwards, the edge functions and the corresponding derivatives along the gradient directions can be obtained. Finally, the self-snake diffusion filter was used to remove speckle noise, and an edge stopping term was added for sharpness improvement. Bai and Feng [98] put forward a generalized anisotropic diffusion, inspired by the fractional order anisotropic diffusion. A novel derivative named G-derivative was presented, and the generalized anisotropic diffusion can be given based on Euler–Lagrange equations. Detailed analyses of stability and simulation results were also presented. Elsharif, et al. [99] developed a hybrid denoising system for ultrasound images. They performed two level discrete wavelet transform (DWT), and employed anisotropic diffusion was used for speckle removal. The nonlinear filtering was also performed on the DWT coefficients, and the total variation was utilized to obtain better quality. Their method outperformed several traditional denoising algorithms in terms of image quality measurements like signal-to-noise ratio, etc. Guo, et al. [100] introduced weighted Euclidean distance to detect edges in synthetic aperture radar images so that the coefficients of anisotropic diffusion can be updated adaptively. The comparison of Gaussian weighting and nonlinear weighting mechanism was discussed as well. Mishra, et al. [101] proposed to harness the edge density probability function and the local information of pixels to better adjust the diffusion directions in speckle reducing anisotropic diffusion. The false contours can be removed by edge density information, and the phenomenon of over smoothing can be alleviated by the relativity of pixels. Experimental results revealed that their method produced better sharpness of the lines in ultrasound images. Mei, et al. [102] suggested to used phase asymmetry to recognize lines and edges in ultrasound images. Based on this phase asymmetry idea, they proposed a new fractional total variation method. The coefficients of fractional order anisotropic diffusion were developed based on phase asymmetry. The entire denoising model was optimized by gradient descent. The edge preserving and denoising performance was improved, and the staircase phenomenon was alleviated as well.

3.3.4. Non-local methods for denoising

Non-local methods take a different perspective to remove the noise in images. Instead of denoising based on local information, such as linear filtering and median filtering, non-local methods employ the redundancy information in images for noise removal. The entire image is divided into several blocks. To remove noise in certain block, non-local methods will search the other blocks with similar structures and compute a weighted average estimation as the denoised pixel value.

Suppose a pixel i with observed gray-level $v(i)$ in a noisy image I, the estimation by non-local means (NLM) filtering is defined as

$$NLM(v(i)) = \sum_{j\in I} w(i,j)v(j) \tag{11}$$

where $w(i,j)$ serves as the weighting factor, which satisfies

$$\begin{cases} 0 \leq w(i,j) \leq 1, \forall j \in I \\ \sum_{j\in I} w(i,j) = 1 \end{cases} \tag{12}$$

The similarity in the pixel pair (i,j) is measured by the gray-level vectors of their blocks. With this weighting strategy, non-local means not only considers the single-pixel values but takes the neighborhood structure information into account, so that it can produce robust denoising results. Non-local methods have been widely used for various types of image denoising applications, and variations have been proposed as well.

Yang, et al. [103] proposed a hybrid speckle removal system based on NLM for ultrasound images. The local structure information was used to generate speckle noise statistics in local blocks. To reduce mixed noise, Chen, et al. [104] proposed a robust bi-sparsity model to generate the similarity based on the prior information. The coefficients and non-local means were utilized to recognize the similar structures, which were regularized by L_0 norm. With the aim of handling outliers and improving robustness, a weighting mechanism was also added to their system. The experimental results demonstrated the superiority of their method. Yu, et al. [105] developed a probabilistic NLM method to remove speckle noise from optical coherence tomography images. Rank-ordered absolute difference (ROAD) was employed to distinguish the noisy pixels in the local blocks of the image, and the uncorrupted probability can be calculated. Consequently, better similarity estimation can be performed between the image blocks. Finally, the noisy pixels can be restored using weighted means. The improved NLM yielded promising speckle removal performance and preserved the structure information at the same time. In [106], a discontinuity indicator was proposed to classify the edges and noise, and an adaptive bandwidth parameter was used to replace the fixed one to obtain better denoising performance. Mandal, et al. [107] put forward a super-resolution algorithm with only a single noisy image. The noise strength was evaluated by the gradients of local blocks. To implement sparse representation, an adaptive thresholding algorithm was proposed. An additional term was included to reserve edges and contours in the reconstructed image. Qian, et al. [108] suggested using principal component analysis to estimate the noise level to improve the efficacy of NLM. Tang, et al. [109] found that there is a correlation between the image blocks without noise, which can be revealed by low-rank representation. Hence, a corrupted probability term was employed for regularization. Multiple estimations of the local blocks were harnessed to get the aggregated denoised image. Bindilatti, et al. [110] combined Wiener filter with non-local weighting to remove signal-dependent noise like Poisson noise. They proposed to estimate parameters by non-local block information based on stochastic distance. Georgiev, et al. [111] developed a 3D image analysis algorithm based on non-local denoising in complex domains. Panigrahi, et al. [112] firstly transformed the images into the curvelet domain, and the approximation and detailed coefficients were obtained. Then, a multi-scale NLM was proposed to remove noise and maintain edges. To distinguish noise and signal, hard thresholding was used. The experiment was implemented with both gray-level and color images. Shahdoosti and Rahemi [113] used a log-likelihood to get denoised pixels in images, and the non-local information of the blocks was generated based on Pearson distance for filtering. Hou, et al. [114] proposed pixel-level non-local self-similarity which performed better than block-level similarity, because it is easier to obtain similar pixels than similar blocks. Mei, et al. [115] proposed an optimized Bayesian non-local means to estimate the noise-free ultrasound images. Redundancy index of every block was computed to locate the low-redundancy areas of the image. Zeng, et al. [116] integrated non-local filtering and low-rank regularization to remove noise from hyperspectral images. The image was first divided into overlapping blocks. To separate the clean blocks from noisy blocks and maintain structure information simultaneously, a local rank-constrained low-rank technique was proposed. Finally, an NLM algorithm was used to remove noise.

3.3.5. Bilateral and Trilateral filtering

Bilateral and trilateral filters are nonlinear filters that are capable of removing noise as well as preserving details, such as edges and corners in images. Good denoising results by bilateral and trilateral methods are contributed by the weighted sum of intensity values in the neighborhood of the pixels. The weights are carefully chosen, which are related to not only the pixel spatial distances but also the intensity distances. Given an image I, the bilateral filtering is expressed as

$$\mathrm{Bil}(I(\mathbf{x})) = \frac{1}{C}\sum_{\mathbf{y}\in N(\mathbf{x})} e^{-\frac{\mathbf{y}-\mathbf{x}^2}{2\sigma_d^2}} e^{-\frac{|I(\mathbf{y})-I(\mathbf{x})|^2}{2\sigma_r^2}} I(\mathbf{y}) \tag{13}$$

where the σ_d and σ_r are hyper-parameters that control the tradeoff between spatial distance and intensity distance, $N(\mathbf{x})$ is the neighboring field of pixel x and C is the constant value obtained by

$$C = \sum_{\mathbf{y}\in N(\mathbf{x})} e^{-\frac{\mathbf{y}-\mathbf{x}^2}{2\sigma_d^2}} e^{-\frac{|I(\mathbf{y})-I(\mathbf{x})|^2}{2\sigma_r^2}} \tag{14}$$

The values of the two hyper-parameters σ_d and σ_r are crucial for the final denoising results. Unfortunately, there is little theoretical research on how to determine optimal hyper-parameters. In practice, they are usually determined by trial and error. Zhang and Gunturk [117] tried to model the two as noise variance functions by empirical analysis. They revealed that σ_r is the more significant factor of the two, which is linear to the noise's standard deviation. An improved multi-resolution bilateral filtering was proposed by integrating bilateral filtering with a wavelet thresholding method. Akdemir Akar [118] suggested harnessing the genetic algorithm to optimize the parameters in bilateral filter model for Rician noise elimination in MRIs. Balocco, et al. [119] proposed a speckle reducing bilateral filter (SRBF) for ultrasound denoising. To preserve the image details, the statistical characteristics of noise were embedded into the conventional bilateral filter. Lin, et al. [120] developed an automated system to remove impulse noise and Gaussian noise based on a switching mechanism. Firstly, the features for textures and boundaries were generated for every pixel in the image. Then, each pixel was classified as impulse noise, Gaussian noise or real signal based on a sorted quadrant median vector approach. Finally, the switching bilateral filter was employed to remove these two different types of noise based on the classification labels. Their method can work efficiently without weighting parameters. Zhang, et al. [121] also suggested removing impulse and Gaussian noise within one framework. Firstly, they found out all the impulse noise pixels by a detector and edge component value. Then, they suggested connecting the edges to get refined regions. Finally, an adaptive bilateral filter was proposed to remove the two types of noise with different strategies automatically based on the label information from the detector. Wei, et al. [122] put forward a two-stage denoising algorithm for 3D optical and laser scanning. In the first stage, a joint bilateral filter method was developed to remove most noise in the 3D mesh while preserving texture features. In the second stage, they proposed to add the boundaries and lines as constraints to the traditional Laplacian smoothing because the feature lines are easy to obtain after the denoising in the first stage. Phophalia and Mitra [123] integrated the bilateral filter with rough set theory (RST) to improve denoising efficacy as well as preserving more details. The RST was employed for generating edge mask and labels at pixel level, which can be used to guide the

Table 3
Summary of image denoising methods

No.	Authors	Type of noise	Methods	Datasets	Results
1	Kazubek [74]	White Gaussian noise	A thresholding for pre-processing and Wiener filtering for denoising	Standard test images (Barbara and Lena)	The proposed method achieved state-of-the-art PSNRs with less computational complexity.
2	Portilla, et al. [75]	White Gaussian noise	Wavelet decomposition, Gaussian vector, and Bayesian estimation	Standard test images (Lena, Barbara, Boats, House, and Peppers)	Their algorithm achieved substantially better PSNR and mean squared error than some previous methods.
3	Ghazel, et al. [76]	White Gaussian noise	Fractal wavelet coding	Standard test image (Lena)	The fractal wavelet coding produced better denoising performance in terms of PSNR and root mean squared error.
4	Clauset, et al. [151]	White Gaussian noise	Local ICA and kernel PCA	Standard test image (Lena) and nuclear magnetic resonance spectra	The kernel PCA achieved better denoising effects on nuclear magnetic resonance spectra images.
5	Pedersen, et al. [152]	White Gaussian noise	Stein's unbiased risk estimate and wavelet thresholding	Standard test images (Al, Bridge, Crowd, Goldhill, Barbara, Boats, House, and Peppers)	Near-optimal denoising results were achieved with less computation requirement.
6	Fielding, et al. [153]	White Gaussian noise	Dictionary learning, cluster, and wavelet decomposition	8 benchmark images (girl, baboon, couple, bark, etc.)	The proposed method achieved state-of-the-art denoising performances within less execution time.
7	Pedersen, et al. [152]	White Gaussian noise	Convolutional neural network and monogenic transform	Berkeley segmentation dataset	The combined framework achieved state-of-the-art performance in terms of both visualization and PSNR.
8	Molenberghs, et al. [154]	Stripe noise	Weight sum variance of digital number probability	Remote sensing images	Their denoising approach yielded better PSNR, which consequently helped the cloud segmentation.
9	Gökdağ, et al. [82]	White Gaussian noise	Wavelet thresholding and analysis of variance	Confocal laser scanning microscopy images	The proposed method achieved satisfactory denoising results.
10	Golilarz, et al. [83]	White Gaussian noise	Multi-population differential evolution-assisted Harris hawks optimization algorithm and thresholding neural network	Satellite images	Utilization of particle intelligent algorithms for parameter optimization enhanced the denoising performance.
11	Malfait and Roose [84]	White Gaussian noise	Wavelet decomposition, Markov random field, and probability functions	Standard test images (House, Peppers, and aerial photographs)	The denoising effect of the proposed method was better than other wavelet-based methods.
12	Hua, et al. [85]	Speckle noise	Wavelet decomposition, Markov random field, Gaussian mixture model, and expectation maximization	Synthetic aperture radar images	The proposed method outperformed conventional wavelet methods for denoising.
13	Kim and Curry [155]	Gaussian noise	Active random field	Standard test images (Lena, Barbara, Boats, House, and Peppers)	Active random field performed better than conventional Markov random field as well as thousands of times speedup.
14	Dong and Peng [156]	White Gaussian noise	Hierarchical Markov random field and iterated conditional modes	Standard test images (Lena, Bark, Straw, Tile roof, Baboon, Barche, Brodatz, and Elaine)	The proposed method can preserve more texture information as well as efficiently denoise.
15	Cismondi, et al. [157]	Non-central Chi distributed noise	Markov random field, sliding window scheme, and Gaussian mixture model	Parallel magnetic resonance images	Their method was effective and robust in comparison with state-of-the-art approaches.
16	Do, et al. [158]	Fiber noise	Markov random field and a consistency term	Multi-slice cardiac diffusion tensor images	Their method improved the performance of denoising and reconstruction on 3D images.
17	Roland, et al. [159]	Gaussian noise, multiplicative noise, and mixed color signal-dependent noise	Adaptive anisotropic diffusion filtering	Berkeley segmentation dataset and retinal images	The image quality was improved in their experiment in comparison with conventional anisotropic diffusion filtering as well as fast non-local mean filtering algorithms.
18	Mirkes, et al. [160]	White Gaussian noise	Anisotropic diffusion and region adaptive smoothing strength	Kodak dataset	The proposed method yielded better peak signal-to-noise ratio than several anisotropic diffusion variant techniques.
19	Idri, et al. [161]	Random noise and speckle noise	Anisotropic diffusion and synthetic noise estimation technique	Phantom and mouse artery images	The proposed method achieved good denoising performance while maintaining detailed information.
20	Myrtveit, et al. [162]	White Gaussian noise	Gradient anisotropic diffusion	Infrared thermography	Their method can accurately generate heat sources in noisy field images.
21	Wang and Rao [163]	Gaussian noise, Speckle noise, and Poisson noise	Adaptive anisotropic diffusion and noise level function	STARE Project database and DRIVE database	The proposed denoising algorithm is beneficial to image segmentation.
22	Stamatakis and Alachiotis [164]	Gaussian noise	Anisotropic diffusion and energy-based dynamic CLAD	Seismic data	Their method was effective in removing noise from seismic data.
23	Little [165]	Speckle noise	CUDA based oriented speckle reducing anisotropic diffusion	Synthetic data and real ultrasound video images	This CUDA-based OSRAD ran thirty times faster with the same denoising effect, which was better than other denoising algorithms like wavelet and bilateral-based methods in their experiment.
24	Jubairahmed, et al. [96]	Speckle noise	Contourlet transform and adaptive nonlinear anisotropic diffusion	The US image database	The despeckling performance of this approach was better than several state-of-the-art methods.
25		Speckle noise			

(continued on next page)

Table 3 (*continued*)

No.	Authors	Type of noise	Methods	Datasets	Results
	Lefort, et al. [166]		Structure tensor, maximum and minimum gradient variations, and self-snake diffusion filter	Standard test images (Lena, Fruits, Camera, Ship, Lift, Wheel, Cat, Trui1, Barbara, and House)	The proposed method outperformed conventional diffusion algorithms.
26	Schafer and Graham [167]	White Gaussian noise	Generalized anisotropic diffusion	Standard test images (Lena, Barbara, Boat, and Peppers)	The generalized anisotropic diffusion was effective in denoising.
27	Elsharif, et al. [99]	Speckle noise	Wavelet decomposition and nonlinear filtering	Ultrasound images	Their method outperformed several traditional denoising algorithms in terms of image quality measurements.
28	Guo, et al. [100]	Speckle noise	Weighted Euclidean distance and nonlinear filtering	Synthetic aperture radar images	The proposed method can remove speckle noise and better maintain the edge information at the same time.
29	Ratitch, et al. [168]	Speckle noise	Edge density probability function and nonlinear filtering	Ultrasound images	Experiments revealed that their method produced better sharpness of the lines in ultrasound images.
30	Mei, et al. [102]	Speckle noise	Fractional total variation, gradient descent and fractional order anisotropic diffusion	Ultrasound images	The edge-preserving and denoising performance was improved, and the staircase phenomenon was alleviated as well.
31	Tsiatis and Davidian [169]	Speckle noise	Non-local means filtering with local structure information	Ultrasound images	The denoising performance was better compared with original non-local means.
32	Gottfredson, et al. [170]	Gaussian noise, salt-and-pepper noise and random valued impulse noise	Non-local means filtering and robust bi-sparsity model	Standard test images (F16, Lena, Peppers, House, Barbara, Boat, Bridge, Pentagon, and Couple)	The denoising performance of their approach was better than several state-of-the-art methods.
33	Gad and Darwish [171]	Speckle noise	Probabilistic non-local means and rank-ordered absolute difference	Optical coherence tomography images	The improved NLM yielded promising speckle removal performance and preserved the structure information at the same time.
34	Roy [172]	White Gaussian noise	Non-local means filtering, discontinuity indicator and adaptive bandwidth	USC-SIPI image database	The proposed method yielded better PSNR than some mainstream algorithms.
35	Laird [173]	White Gaussian noise, Rayleigh noise and uniform noise	Non-local means filtering and adaptive thresholding algorithm	Standard optical images, Middlebury database and BSD100 dataset	Their method was robust and effective in the experiments on different datasets as well as under different noise conditions.
36	Rotnitzky and Wypij [174]	White Gaussian noise	Non-local means filtering and principal component analysis	Brillouin optical time domain analyzer signals	The proposed system can denoise without distortion.
37	Tang, et al. [109]	Speckle noise	Non-local means filtering and low-rank representation	Optical coherence tomography images	The experiments on real world images suggested that their method outperformed several state-of-the-art approaches in denoising.
38	Robins, et al. [175]	Poisson noise	Wiener filter with non-local weighting	Standard test images (Cameraman, Peppers, Barbara, Boat and Head CT)	The proposed algorithm was effective in denoising and it can preserve more edge information.
39	Georgiev, et al. [111]	Gaussian noise	Complex-domain non-local denoising	3D time-of-flight data	Their method showed superiority in complex domain denoising.
40	Vansteelandt, et al. [176]	White Gaussian noise	Curvelet transform and multi-scale non-local means	Standard test images (Barbara, Boat, Building, Cameraman, Couple, Goldhill, House, Lake, Lena and Peppers) and color images	The proposed algorithm achieved state-of-the-art performance in terms of PSNR.
41	Stekhoven and Bühlmann [177]	Speckle noise	Log-likelihood and Pearson distance	Ultrasound images	The denoising performance of their method exceeded some state-of-the-art approaches in terms of PSNR.
42	Donders, et al. [178]	White Gaussian noise and real-world noise	Pixel-level non-local self-similarity	BSD68 dataset, Cross-Channel dataset, and Darmstadt Noise Dataset	Their method achieved competitive denoising results.
43	Waljee, et al. [179]	Speckle noise	Optimized Bayesian non-local means	Ultrasound images	Their method improved the denoising performance and preserved more edge information compared with original non-local means.
44	Brick and Kalton [180]	Gaussian noise, white Gaussian noise, stripe noise and impulse noise	Non-local means filtering and low-rank regularization	Hyperspectral images	The proposed scheme achieved state-of-the-art denoising performance.
45	Zhang and Gunturk [117]	White Gaussian noise and real-world noise	Bilateral filtering and wavelet thresholding	Standard test images (Barbara, Boat, Goldhill, House, Lena and Peppers) and color images	The evaluation on both synthetic and real-world images revealed the effectiveness of their method.
46	Zhang [181]	Rician noise	Bilateral filtering and genetic algorithm	Brain magnetic resonance images	The performance of bilateral filtering is dependent on the parameter selection, and genetic algorithm improved the denoising performance.
47	Horton and Lipsitz [182]	Speckle noise	Speckle reducing bilateral filter	Ultrasound images	Their algorithm was applicable in various speckle noise situations.
48	Lin, et al. [120]	Impulse noise and Gaussian noise	Switching bilateral filtering and sorted quadrant median vector approach	Standard test images (Boat, Goldhill, Airplane, Lena, and Bridge)	Their method can work efficiently without weighting parameters.
49	Fuller and Kim [183]	Impulse noise and Gaussian noise	Impulse noise detector, adaptive bilateral filtering and improved artificial bee colony	Standard test images (Airplane, Boats, Bridge, Goldhill, House, Lena, Monarch, Pepper, etc.)	The denoising results of their algorithm was better than some state-of-the-art filters.
50				3D optical and laser scanning	

(*continued on next page*)

Table 3 (*continued*)

No.	Authors	Type of noise	Methods	Datasets	Results
	Yenduri and Iyengar [184]	Gaussian noise and real-world noise	Joint bilateral filter and improved Laplacian smoothing		The proposed scheme was effective and feasible in 3D mesh denoising.
51	Biessmann, et al. [185]	Gaussian noise and Rician noise	Bilateral filter with rough set theory	Open Access Series of Imaging Studies and Brain Tumor Segmentation challenge data	The proposed denoising method achieved better denoising performance on two benchmark datasets.
52	Nelwamondo, et al. [186]	Impulse noise and speckle noise	Adaptive wavelet shrinkage algorithm and trilateral filtering	Ultrasound images	The proposed method can denoise while improve the sharpness of the edges.
53	Do and Batzoglou [187]	Real-world noise	Trilateral smoothing algorithm	Indian Pines, Salinas, and the University of Pavia	Their method achieved better performance than traditional bilateral filtering.
54	Nelwamondo, et al. [186]	Gaussian noise, salt and pepper noise, uniform impulse noise, and speckle noise	Switching bilateral filter and domain weight pattern	Standard test images (Lena, Baboon, Girl, Pentagon, House, Airplane, Sailboat, Aerial, Stream-and-bridge, etc.)	The proposed approach outperformed several bilateral filter-based methods in eliminating noise.
55	Cui, et al. [127]	Speckle noise	Guided trilateral filter scheme and maximum likelihood estimation	Ultrasound images	Their method was effective in image denoising and less sensitive to parameter settings.
56	Zhang, et al. [188]	Interpolation noise	Motion estimation and trilateral filtering	Standard image sequences (Football, Tennis, Garden, Mobile, Paris, and Container)	Trilateral filtering can help improve the video quality in frame rate up-conversion.
57	Schuler, et al. [139]	White Gaussian noise	Multilayer perceptron	Berkeley segmentation dataset	The proposed method can be applied in real-world image de-blurring.
58	Pampaka, et al. [189]	Gaussian noise and rain streaks	Self-learning image decomposition framework	Real-world images	Their method outperformed state-of-the-art approaches in removing rain streaks and Gaussian noise.
59	Reiter and Raghunathan [190]	White Gaussian noise	Stacked denoising autoencoder	CIFAR-bw dataset	The proposed autoencoder was evaluated on a large dataset and achieved state-of-the-art denoising performance.
60	Van Buuren [191]	White Gaussian noise	Residual blocks and multi-scale feature selection	Standard test images (Boats, Lena, Pepper, etc.), and Berkeley Segmentation Dataset	The experiment suggested that their model achieved better denoising results than eight state-of-the-art approaches.
61	Allison [192]	Stripe noise	Residual blocks and wider CNN structure	Meteorological satellite infrared cloud images	Their CNN model was better than several state-of-the-art methods in denoising.
62	Zhang [193]	Real-world noise	Spectral difference mapping algorithm and principal component analysis	Hyperspectral images and airborne data	The proposed method reduced computational complexity and preserved more spectra details while denoising.
63	Allison [192]	Gaussian noise and Poisson noise	3-D atrous denoising convolution neural network	Hyperspectral images	The proposed architecture outperformed several state-of-the-art methods.
64	Sinharay, et al. [194]	Gaussian noise	Residual learning and batch normalization	BSD68	Their model worked effectively for denoising with less computational time.
65	Zheng, et al. [147]	Gaussian noise	Privacy-preserving deep neural network	ChestX-ray8	Their method can be applied in a cloud computing environment for denoising.
66	Little [165]	Rain streaks	Recurrent network, residual mapping, and bilateral LSTM	Rain100H	The proposed model achieved satisfactory denoising results on real-world images.
67	Little [165]	Gaussian noise and real-world noise	Attention-guided denoising convolutional neural network	Berkeley Segmentation Dataset and Waterloo Exploration Database	The denoising performance of the proposed network was comparable to state-of-the-art models.
68	Wu and Bailey [195]	Gaussian noise and real-world noise	Batch-renormalization denoising network	Berkeley Segmentation Dataset and Waterloo Exploration Database	The proposed model yielded better performance than state-of-the-art denoising algorithms.

bilateral filter. Zhang, et al. [124] proposed an adaptive wavelet shrinkage algorithm and combined it with trilateral filter to eliminate impulse and speckle noise in ultrasound images.

The trilateral filter is an improved form of bilateral filter, which is proposed to deal with impulse noise elimination. The trilateral filter introduces a ROAD function to determine whether a pixel is on an edge or it is impulse noise. Chen, et al. [125] found that conventional bilateral filter fails when the centroid of a neighborhood is labeled as noise pixel in hyperspectral image. Hence, they proposed a trilateral smoothing algorithm to solve this challenge. Langampol, et al. [126] suggested improving the performance of switching bilateral filters by introducing a domain weight pattern. The domain weight pattern was proposed to describe the intensity distribution of the center pixel and its neighborhood. With this novel pattern, the mixed noise and the strength can be obtained so that the bilateral filter achieved better results. Cui, et al. [127] proposed a guided trilateral filter scheme and applied it for denoising in ultrasound images, which was generated by the maximum likelihood estimation over the residual of noisy images and target images. Bilateral and trilateral filters can also be used to improve the quality of videos, like frame rate up-conversion and signal-to-noise-ratio improvement [128].

3.3.6. *Deep learning for denoising*

Deep learning is one of the most active research topics in computer science nowadays, which has been applied in various practical problems, such as image recognition [129], semantic segmentation [130], and restoration [131]. In fact, the CNN model was proposed as early as 1989 by LeCun, et al. [132] for recognition of handwritten zip codes, but restricted by ineffective training algorithms and limited computational resources-thus, the mainstream scientific community and practitioners did not pay much attention to CNNs. Deep learning was popularized by the AlexNet [133] in 2012, with its exciting performance on ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). Since then, various CNN models have been invented such as VGG [134], ResNet [135], DenseNet [136], SqueezeNet [137], MobileNet [138], etc. These popular CNN models are designed for image classification, but they can also be used for image denoising.

Schuler, et al. [139] proposed to use multilayer perceptron for image deconvolution, which sharpens a blurry image. The input to the multilayer perceptron was the noisy images, while the output was the denoised clean images. The mapping by multilayer perceptron worked without feature selection, and it can be used to remove different types of noise as well as mixed noise. Huang, et al. [140] proposed a self-learning image decomposition framework that can be applied to denoising. With sparse representation and clustering, their method does require training images, so it can be used for single image denoising. Li [141] designed a denoising autoencoder and stacked these autoencoders together for noise removal. Sun, et al. [142] first employed residual learning to produce a denoised reference image for the input noisy image. Then, a multi-scale feature selection structure based on residual blocks was proposed to restore the details with both the input noisy image and the denoised reference. Xiao, et al. [143] also use residual learning with the aim of reducing mapping size. Then, a wider CNN architecture with more convolutional layers was employed for denoising, and the representations from different CNN layers were harnessed to recover the details and texture information. They also extend their research for single image denoising. Xie, et al. [144] put forward a spectral difference mapping algorithm for hyperspectral image denoising. The denoised key band was proposed to implement efficient computing, which was obtained by principal component analysis. Liu and Lee [145] presented a 3-D atrous denoising convolution neural network for denoising. Both the spatial and spectral domains were leveraged to extract features. In order to prevent overfitting as well as to preserve more detail, multi-scale and multi-branch analysis was conducted. Tian, et al. [146] combined residual learning with batch normalization to accelerate the training process for image denoising. Zheng, et al. [147] proposed a denoising deep neural network and applied it for privacy preservation in the cloud. Ren, et al. [148] first trained two single recurrent networks and coupled them to extract both rain streaks and noise-free backgrounds. Then, bilateral LSTM was designed to integrate the two models to propagate the rain streaks and background. Tian, et al. [149] introduced an attention mechanism into deep CNN to remove image noise, and they also proposed to use batch re-normalization to fuse two deep CNNs for denoising in [150].

To summarize, in this section, we provided a detailed account of various image denoising methods, including wavelet-based methods, Markov random fields, anisotropic diffusion filtering, non-local methods, bilateral and trilateral methods, and deep learning-based denoising. A brief summary of these denoising methods is listed in Table 3. In many practical applications, however, a combination of denoising methods is likely to produce more satisfactory results. Image denoising is still a challenge in image pre-processing, and the quest for better denoising performance is likely to continue for years to come.

4. Missing value challenge

At the data-level for multimodal data fusion, we often meet a significant unavoidable challenge—missing values. This ‘missing value challenge’ pervasively exists in the majority of real-world data sets, and four commonly seen scenarios are listed below. The first is: when dealing with locally missing samples in a single dataset, a clear and complete data entry will not be easy to obtain due to many reasons [196]. For instance, the selected detector is inappropriate, the detector is partially occluded or malfunctioned during the working process, or the data is omitted during the collection process. All these factors may lead to the data-missing challenge. The next point is when multiple modalities are involved in a system, obtaining the data from only one modality cannot present complete and accurate information of the system. For instance, MEG and EEG are always recorded at the same time to compensate information for each other [197]. Thirdly, when taking samples at different modalities, if the utilized sampling points are incomparable, the obtained data will possibly be seen as structurally missing. For this scenario, each modality is appropriately sampled on its own. However, the points on the common sampling grid would be seen as missing data if they miss the data from all modalities [198]. The fourth circumstance is connection prediction, which often appears in social network analysis and recommender systems. For instance, the challenge in the analysis of social networks is how to well predict social connections according to an existing database of connections [199], where known entries are far from enough.

Table 4
Univariate missingness pattern

	x_1	x_2	x_3	x_4	x_5	x_6
C_1	76	109	56	83	17	207
C_2	123	82	111	100	106	?
C_3	67	73	89	8	29	?
C_4	25	106	45	34	10	?
C_5	213	55	38	145	89	?
C_6	89	45	90	17	96	?

Table 5
Multivariate missingness pattern: monotone pattern

	x_1	x_2	x_3	x_4	x_5	x_6
C_1	76	109	56	83	17	207
C_2	123	82	111	100	106	?
C_3	67	73	89	8	?	?
C_4	25	106	45	?	?	?
C_5	213	55	?	?	?	?
C_6	89	?	?	?	?	?

Table 6
Multivariate missingness pattern: arbitrary pattern

	x_1	x_2	x_3	x_4	x_5	x_6
C_1	76	109	56	83	17	?
C_2	?	82	111	100	106	80
C_3	67	73	?	8	29	96
C_4	25	?	45	34	?	109
C_5	213	55	38	145	89	310
C_6	89	45	90	?	96	95

4.1. Missingness mechanisms

Missingness mechanisms can be defined as the nature and categories of missing values. If considered from the perspective of missing distribution, three unique categories can be listed: Missing Completely at Random (MCAR), Missing at Random (MAR) [185], and Missing not at Random (MNAR) [187].

In the case of MCAR, the missing data distribution is random and uncorrelated with the values of any variables. In the case of MAR, the missing distribution is not completely random; that is, the missing of such data is not associated with the missing values themselves but possibly has some relations with the observed values. In the case of MNAR, the absence of data is dependent on missing values themselves [193,200,201].

4.2. Missing data patterns

In general, missing data patterns can be categorized into two: univariate and multivariate. Under the circumstance of a univariate missingness pattern, missing values can only exist in one variable. A typical example can be seen in Table 4, where x stands for the variable and C stands for the row. It is explicit that variable x_6 is the only variable with missing data [196].

Under the circumstance of multivariate missingness pattern, missing values will exist in no less than two variables. Moreover, this pattern could be categorized into a monotone pattern and arbitrary patterns. In the situation of monotone, if the data for column x_i is missing, then all subsequent data will be missing, as shown in Table 5 [196].

For an arbitrary pattern, as illustrated in Table 6, missing values can appear anywhere, and no matter how one arranges variables, no special structure would appear [196].

All in all, if not handling well, missing data would greatly affect the quality of multimodal data fusion in various ways. One obvious impact is efficient information, and statistical power may be reduced [202,203], and thus some useful data analysis approaches will become difficult to employ [204]. In addition, some bias may be introduced into estimations derived from the statistical model [159,160,200]. Therefore, to improve the quality of the knowledge obtained from data fusion and other intelligent data analysis approaches, the first issue we must handle with is missing values.

The remainder of this section introduces a solution called Missing Data Imputation Techniques. The meaning for each abbreviation used is stated out in Table 11.

4.3. Missing Data Imputation Techniques

Missing Data Imputation Techniques (MDITs) are commonly utilized to deal with the missing values [205]. Rather than delete or tolerate the cases associated with missing values, it can well handle the missing values by imputing appropriate new values and at the same time retaining the originally known values in the dataset. There exist various methods in the field of MDITs. They can be divided into two categories: non-ignorable (NI) missing data imputation and ignorable missing data imputation.

4.3.1. Non-ignorable (NI) missing data imputation methods

4.3.1.1. Likelihood-based methods. One main category of NI imputation methods is the likelihood-based method. To well utilized this method, the first thing that needs to be determined is the mechanism of missingness. This is because, under the case of MNAR, having the specified information of primary data and missing data mechanism becomes a necessity, as they must be jointly modeled to prevent bias from being introduced into estimations. A typical way is to integrate a parametric model for NI and the complete data log-likelihood [206–208].

There exist three commonly used alternative likelihood-based methods: Selection Models (SMs), Pattern Mixture Models (PMMs), and Shared Parameter Models (SPMs), proposed by Mahapatra, et al. [209]. PMMs and SMs can be considered as two decomposing possibilities of the joint distribution. For SMs, a specification of the distribution for complete primary data and the probability distribution for the missing data patterns is needed [196] [210,211]. While for PMMs, which supposes that there exists a mixture of patterns in the missingness, need to take the circumstances of model parameters for each pattern into consideration and operate the computation separately. However, PMMs cannot directly provide marginal estimates [64,212–214]. Instead of incorporating common parameters into models, SPMs are usually applied when the missingness is possibly related to the true underlying response for a subject when the data settings are clustered and longitudinal [215–219].

4.3.1.2. Non-likelihood-based methods. The non-likelihood-based methods require the joint distribution of the complete data to be like a non-parametric (or semi-parametric) model. In contrast, the mechanism of missingness to be like a parametric model [220–222].

A well-known non-likelihood-based method is sensitivity analysis, which has been utilized in many types of research [223, 224]. However, the sensitivity analysis is known to have a few defects. The first one lies in practice; its presentation of results is not simplified and concise enough. Secondly, sensitivity analysis has the limitation that it is usually confined to a relatively small number of parameters. Last but not least, if various sensitivity analysis could be predicted, contradictory conclusions would possibly be generated [225].

4.3.1.3. Comparison and summary. The non-likelihood-based methods are more widely utilized in comparison to the likelihood-based methods because it is difficult to seek a non-response model that is perfectly specified as a function of reported values in most real-world cases. Nevertheless, the NI missing data imputation methods are not simple and flexible enough, as they need not only the model for the complete data but also the specified information of missingness distribution. In the next part, we will introduce some Ignorable missing data imputation approaches.

4.3.2. Ignorable missing data imputation methods

Ignorable missing data imputation methods could be categorized as single and multiple imputation methods.

4.3.2.4. Single imputation methods. Single imputation means substituting each missing value with a single value. After filling in all the missingness and achieving a new complete dataset, more standard data analysis approaches will be able to come into use. Moreover, it merely handles one time of missing values, implying that a consequent consistency of results could be achieved from the same analysis [226], which signifies that the single imputation approaches are suitable for utilizing in the field of machine learning. The following introduced are various traditional utilized main basic single imputation techniques and two modern single imputation techniques: deep learning approach and Expectation Maximization (EM) approach.

Mean imputation. Mean imputation is one of the commonly utilized imputation approaches. It supposes the average value of a variable is the best estimation of all the circumstances in which information about the variable is missing [179,227]. Therefore, when the data missingness case is MCAR, an average value will be assigned to the known values of the same variable [180]. Suppose x_4 in Table 5 is a continuous variable, the blanks of missing values which are marked with '?' will be filled in by the mean values of the three observed values of the variable x_4, according to:

$$x_4 = \frac{1}{3}\sum_{i=1}^{3} C_i(x_4) \tag{15}$$

However, imputing the sub-group average of all the sub-group missing data may not be the best choice. Cohen [228] proposed an improved method to split the missing values into two parts and impute according to the following equations:

$$\overline{X}_{obs} \pm \sqrt{\frac{n + n_{obs} - 1}{n_{obs} - 1}}\sigma_{obs} \tag{16}$$

$$\sigma^2{}_{obs} = \frac{1}{n_{obs}}\sum_{i=1}^{n_{obs}}\left(X_i - \overline{X}_{obs}\right)^2 \tag{17}$$

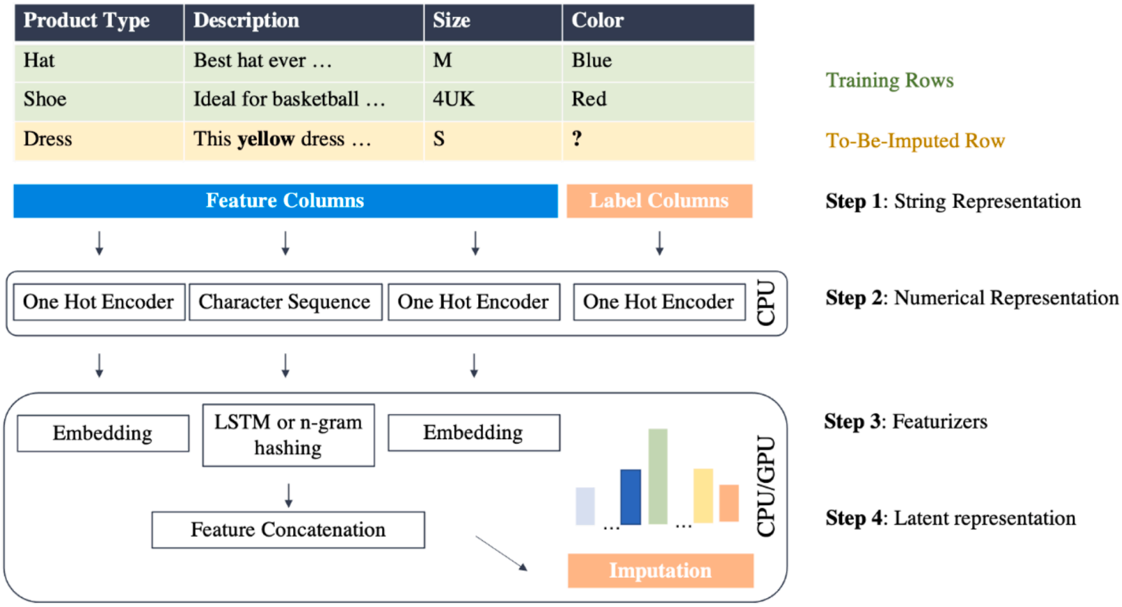
where $\overline{X}_{obs}$ represents the mean of observed values, and n_{obs} represents the number of observed values.

To conclude, mean imputation is a good choice for the MCAR circumstance. It is rapid and easy to put into practice. Nevertheless, one defect of this method is it may result in underestimation of the population variance, and thus a small standard error and a possibly Type I error.

Regression imputation. Regression imputation aims at substituting each missing data blank with a newly predicted value on the basis of a regression model in the case of MAR [181].

Generally, the process of regression estimation is divided into two stages. In the first phase, a regression model is established utilizing all the existing complete observed values, and then the value for missingness blank will be computed according to the established regression model.

Regression imputation preserves the size information of the sample by retaining the absence of values, which is superior to Multiple Imputation (MI). However, since the imputation data is computed by the regression model that needs specifying, there are exaggerations of correlation and covariance. A larger sample size becomes a need to give out



Product Type	Description	Size	Color
Hat	Best hat ever …	M	Blue
Shoe	Ideal for basketball …	4UK	Red
Dress	This **yellow** dress …	S	?

Fig. 15. An example imputation flow path for Datawig on non-numerical data

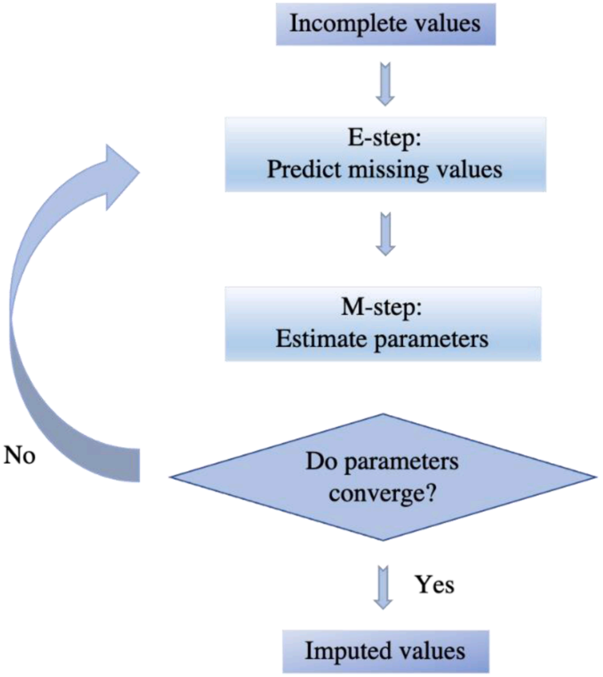


Fig. 16. The procedure of EM

stable estimations [229].

Hot-deck imputation. The general process of this approach is: first, stratifying the data set based on some auxiliary variables; then saving the complete cases in the classes of the active file; finally, imputing each missing blank of the variable for a non-respondent with the observed response from the most 'similar' respondent [230].

Random and deterministic are two typical hot-deck imputation methods. The random method randomly selects the respondent from a range of potential respondents. If the corresponding class has no observations, it will be combined with other classes and the imputation would be performed according to the merged class [196]. While for the deterministic hot-deck, there are many instances. Similar Response Pattern Imputation (SRPI) determines the most similar case with no missingness and copies the values, in this case, to substitute the blank in those cases with missingness. The K-NN imputation approach starts with searching the missing value of the K's nearest neighbor and then substitutes the blank with the mean value of the variable value corresponding to the K's nearest neighbor. [226,231].

The hot-deck preserves the associations and distribution of the available information by replacing different missingness with different observed values and holds the appropriate measured level of variables. The results are usually superior to those from the mean imputation and the regression imputation [196].

Deep Learning (Datawig). To deal with the missing value challenge in large-scale datasets containing millions of rows or in tables with heterogeneous data types, including unstructured text, a deep learning imputation method called 'Datawig' was introduced. It is a robust, scalable approach for missing value imputation that combines deep learning feature extractors with automatic hyperparameter tuning and could offer more flexible modelling options as well as achieve relatively accurate results when compared to other imputation methods. An example imputation flow path for Datawig on non-numerical data is shown in Fig. 15 [185].

Expectation Maximization (EM). The EM method in missing data handling is an approach of seeking maximum likelihood estimation of parameters of an underlying distribution in the data set with missingness issue [232].

As Figure 16 shows, it starts with predicting the missingness according to assumed values for the parameters. Next, it utilizes the predictions for the updating of parameters. Then repeats these two steps until the sequence of parameters converges to maximum likelihood estimations [233].

The EM method is favored for its statistical properties. In most cases, it outperforms popular incomplete data handling approaches (e.g., mean imputation) because it supposes the missingness circumstance as MAR. This method guarantees the convergence to the local maximum value of the likelihood function. If the degree of missingness is high, then the speed of convergence will be slow. Otherwise, the speed will be fast. However, one limitation of the EM method is it adds little uncertainty component to the estimation, which neglects the estimation variability. Moreover, EM does not guarantee the convergence to a global maximum likelihood solution [234,235].

Even if single imputation may sometimes be considered as a potential approach to address the missing data problem, it has little

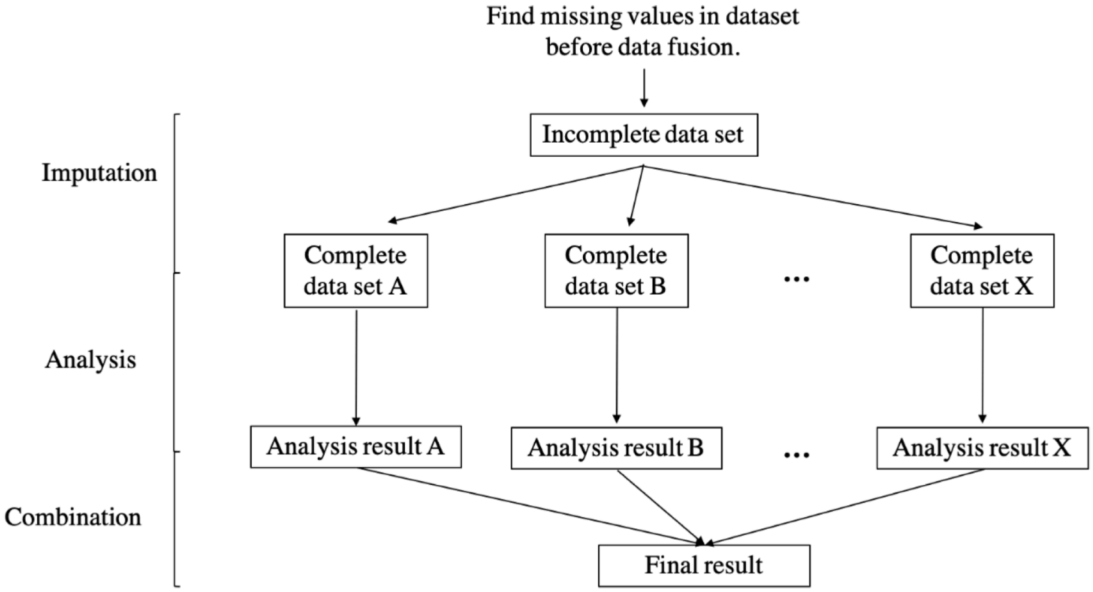


Fig. 17. General procedure of MI

uncertainty in missing data estimations. As a result, bias may be introduced into the available sample size and the standard deviation. Apart from that, confidence intervals for parameter estimates could become too narrow, and a severe Type I error will exist. Thus, to focus on introducing the uncertainty into the model, the information of multiple imputation methods will be discussed in the next section [188].

4.3.3. Multiple imputation methods

In order to effectively deal with missingness in the circumstance under MAR and multivariate normality assumptions, Multiple Imputation (MI) is generally introduced. MI compensates for the obvious shortcomings of single imputation while retaining most of its main benefits [221]. The main idea of MI can be utilized to introduce statistical uncertainty into the model by multiple imputations of missing data. This uncertainty is utilized to simulate the sample variability of a complete dataset. MI is very effective even when dealing with a dataset with a small number of samples. However, each operation of MI may generate imputed results that are slightly different from each other, so the results are not reproducible. MI is also computationally intensive and will become time-consuming when dealing with the workload of imputing multiple (usually more than 5) data sets. In addition, different categories of imputation models require different result integration approaches, giving restrictions in the selection of models [236] [237].

4.3.3.5. General procedure. MI has several desirable features. First of all, it can introduce suitable random error into the imputation process, which can enable an approximately unbiased estimation of all parameters. Other deterministic imputation methods are difficult to achieve this under general settings [238]. Besides, MI can deal with all categories of data and analysis with no need for specific software. Better

Table 7
A summary table for the missing data imputation methods

S. n.	Reference	Method	Missing Data Category	Imputation Category	Characteristic
1	Little [165]	SM	Non-ignorable	Likelihood-based	Under the MNAR assumption Require the specific distribution of the missing mechanism
2	Little [165]	PMM	Non-ignorable	Likelihood-based	Under the MNAR assumption Do not directly provide marginal estimates
3	Wu and Bailey [195]	SPM	Non-ignorable	Likelihood-based	Under the MNAR assumption The calculation is complicated.
4	Robins, et al. [175]	Sensitivity analysis	Non-ignorable	Non-likelihood-based	Presentation of results is not simplified and concise enough. Usually confined to a relatively small number of parameters. Contradictory conclusions would possibly be generated.
5	Waljee, et al. [179]	Mean imputation	Ignorable	Single imputation (Traditional)	Good choice for the MCAR circumstance Simple and rapid. May result in underestimation of the population variance.
6	Zhang [241]	Regression imputation	Ignorable	Single imputation (Traditional)	Good choice for the MAR circumstance Preserves the size information Need a larger sample size to give out stable estimations
7	Andridge and Little [242]	Hot-deck imputation	Ignorable	Single imputation (Traditional)	Good choice for the MAR circumstance Preserve the associations and distribution of the available information
8	Biessmann, et al. [185]	Datawig (Deep learning)	Ignorable	Single imputation (Modern)	Robust, scalable, flexible, and accurate.
9	Do and Batzoglou [187]	EM	Ignorable	Single imputation (Modern)	More effective than other single imputation methods in most cases. Add a little uncertainty component to the estimation. May fall into the local extreme value. The convergence rate is not very fast, and the calculation is very complicated.
10	Zhang [193]	MCMC	Ignorable	Multiple imputation	A long enough Markov chain is constructed for the distribution of the elements to stabilize to a stationary distribution.

standard error estimates can also be obtained by utilizing repeated imputations [239]. Last but not least, even when the number of imputing times is limited, MI can still complete the task very well. In some applications, just 3–5 imputations are sufficient to obtain excellent results. One famous MI method that can be well used to deal with non-monotone missing pattern circumstances is Markov Chain Monte Carlo (MCMC). It is a Monte Carlo integration method utilizing Markov chains. In each iteration of this method, the imputations are drawn from the target probability distribution, and then the unknown parameter values of the predictive distribution are simulated according to the draws from the completed data posterior [193]. The basic procedure of MI mainly includes the following three stages: Imputation stage, Analysis stage, and Combination stage, as illustrated in Fig. 17 [240].

4.3.3.6. Selection of Multiple imputation model. The imputation model built for MIs will focus on two factors. The first thing it needs to concentrate on is: the selected imputation methods have to be proper, which means they should be compatible with the analysis methods [240]. The choice is often made according to the category about the missingness patterns, the mechanisms of missing values, and the distribution of data. The second factor that needs to be considered is the variables. The variables that are used by the analysis model should absolutely be included. While those that are not used for analysis can also be included if they are highly related to the missing values. However, when dealing with complicated circumstances, it is sometimes difficult to find a perfect imputation as there exists a bias in the estimator of MI variance for domains that are not part of the imputation model [196].

4.3.4. Comparison and summary

To conclude, single imputation tries substituting each missing value with a single value. In contrast, MI handles the missingness based on repeated simulation, with a good reflection of sampling variability for the values in the real world. Both of them have the potential of preparing the input dataset for data fusion. A summary table for the missing data imputation methods listed above is displayed in Table 7.

All in all, there is no perfect imputation strategy that can deal with all categories of missing value challenges in the dataset. Each imputation strategy may perform well on some datasets and missing data types but may perform poorly on others. Unless a specific strategy is determined to be used for a particular type of missing value due to obvious setting rules, it is best to experiment and evaluate which model works best for your own dataset.

5. Alignment and Registration

Alignment and registration aim to reduce spatial or temporal inhomogeneities between samples, including differences in acquisition frequencies, sampling devices, and sample physiology. In biomedical data, registration is a standard prerequisite for the analysis and fusion of multimodal data. Registration is prevalent in neuroimaging due to the human brain’s relative in-elasticity [243], while studies involving registration of other anatomical regions have also been conducted [244]. Image alignment and registration are commonly required in the clinical analysis and biomedical research of imaging data [240]. Registration provides the benefits of correlation between individual samples and independent subjects. In modern clinical treatment, a reliable diagnosis is often based on multiple clinical measurements that provide complementary information, e.g., X-ray and MRI provides adequate visualization of bone and tissue structure, respectively [240]. While clinicians are trained to utilize a variety of measurements to achieve a diagnosis, integrating imaging modalities through alignment and registration can provide a more efficient diagnosis while also providing a basis for procedures like image-guided radiotherapy [245, 246] and techniques like video microscopy [247,248]. From the perspective of modern research, a significant challenge is the inhomogeneities between individual samples. The effect of this inhomogeneity is especially significant in the case of neuroimaging, where a lack of geometric comparability between subject brains impedes the identification of specific characteristics [249]. Registration is a fundamental prerequisite for neuroimaging research, providing the basis for subsequent procedures like volumetric feature extraction, atlas construction, and 3D brain reconstruction. In the data fusion either in research or application, homogeneity amongst single modalities and structural homogeneity amongst multiple modalities yields better performance. Application of alignment and registration can be the basis for the fusion of information from multiple modalities.

5.1. Transformations and Interpolations

Transformations for registration adjust the sample towards the desired target space that reduces inhomogeneities. Transformations can be categorized into linear and nonlinear transformations. Linear transformation involves the calculation of rotational and translational vectors, mitigating global positional changes. A typical type of linear transformation is rigid transformations, which involve rotations, translations, scaling, and shearing [250]. The transformations can be encoded in a matrix *M*, for which the transformed data is the product of the matrix with the original data, i.e., $x' = M{\cdot}x$. Linear transformations are suitable for data with minor distortion or deformations. A prime example is in neuroimaging: where patients are mainly stationary during the imaging process, and the skull provides a structural containment for the brain [251,252]. However, linear transformations are not suitable for organs involved in constant moderate-scale motions, e.g., heart and lungs. These motions present local deformations, which can be adjusted by nonlinear transformations. There are two main types of nonlinear transformation: (1) physical-model-based (2) basis function-based. Physical-model-based transformations like the linear elasticity transformation predominately model the deformation of objects based on stress and strain theory, where internal forces of the current state and external forces of deformation interact towards equilibrium. Other physical models, like the fluid flow or medical *a priori* based on the human anatomical structure, can also be used for nonlinear transformations [246]. Apart from transformations, another essential subject in registration is interpolation. Interpolation is used to approximate values of points outside of set grid positions, a common scenario for registration between samples of varying sizes and resolutions. The most common method is linear interpolations, where the interpolated value of a point is dependent on the distances to the neighboring points. Computational complexity increases with the use of more neighboring points and more complex interpolation methods. Other interpolation methods include the nearest neighbor, windowed sinc, and stochastic interpolation. Interpolations can cause fluctuations in registration measures and create artifacts in registered images [253]. Transformations and interpolations form the basis for any registration method set used with or prior to the data fusion process and are generally chosen based on the type of data and modalities involved in fusion.

5.2. Intensity-based registration

Intensity-based registration relies on the information of individual image voxels to derive registration measures, which are often iteratively optimized for better transformation from source data to a template or reference data. For images, these methods are also known as voxel-based registration. Standard measures include mutual information, cross-correlation, and the sum of squared differences (SSD). Optimization methods, detailed in the following Section 5.5, estimate the best parameters for the transformation model based on these measures. Here we provide some basic examples of these commonly used measures. Sum of square differences is one of the fundamental measures of registration, with a transformation function of *f*,

$$\mathrm{SSD} = \sum_i \sum_j \left(R_{i,j} - f\left(S_{i,j}\right)\right)^2 \tag{18}$$

where *R* is the reference, and *S* is the source. A modified mass-preserving SSD has been applied for the registration of lung CT [65]. A similar alternative to SSD is the sum of absolute differences (SAD), where we take the sum of scaled voxel-wise differences between the registered and reference images.

$$\mathrm{SAD} = \frac{1}{n}\sum_i \sum_j ||R_{i,j} - f\left(S_{i,j}\right) \| \tag{19}$$

SAD has been widely applied for intensity-based registration, including its role as a measure of registration quality measure for a generative adversarial network (GAN) which registers cardiac and retinal MRI images [254]. Another class of intensity-based registration methods is based on mutual information, which captures non-parametric statistical dependencies with no *a priori* requirements. Mutual information (MI) between two images can be calculated with measures of entropy, whose basic formulation is

$$MI = E(R) + E(f(S)) - E(R, f(S)) \tag{20}$$

where *E* represents entropy estimation methods, e.g., Tsallis entropy, Renyi entropy, and the original Shannon entropy. Maximization of mutual information between the registered image and reference leads to robust and reliable registration. Mutual information and normalized MI are some of the most commonly applied method classes. Previous applications of mutual information include focusing on increasing local MI with regional mutual information [255], adaption to multimodal data [256], and combination with gradient information [33,257]. With a lack of dependency on initialization and preprocessing of the source, MI has also been applied in unsupervised registration with no or limited references [258]. MI has also been used as a constraint in the loss function of a cyclic-GAN model, optimizing an overlay of MRI on CT images for better image-guided thermal ablation of liver tumors [259]. These techniques can be easily applied or extended for more data-fusion-focused applications. As the basis of intensity-based registration methods, the above-mentioned measures are calculated based on the global intensity values of the data. Therefore, intensity-based registration does not capture spatial or temporal dependencies for data in higher than one dimension. This drawback is often mitigated with feature-based registration in the following subsection.

5.3. Feature-based registration

Feature-based registration derives registration measures from homogeneous features between samples. The lower dimensionality of the homogeneous features requires less computational power compared to voxel-based registration. Features can be categorized into two types: (1) artificial identification points or extrinsic landmarks introduced into the data, (2) anatomical and geometric identifications, or intrinsic landmarks contained within the data. Extrinsic landmarks are also known as artificial identification points (AIPs). These AIPs can be foreign markers that can be artificially implanted or injected onto the subject, e.g., molds, contrast, radioactive tracers. With recent developments in the medical apparatus, AIP is usually non-invasive or minimally invasive. However, for certain types of biomedical data, AIP can be highly invasive, e.g., radioactive tracers for nuclear medicine. Extrinsic landmarks are limited by the physical placement or injection of the AIPs, which may not be optimal. AIP provides a basis for the registration of highly deformable or elastic anatomy like skin and soft tissue [260]. However, compared to AIP implanted on rigid anatomical structures like bones, the movement of AIP in soft tissue also poses a problem in providing a robust positioning basis. Registration measures can be efficiently calculated by comparing the AIP and fiducial markers, specially designed identification points [261]. Intrinsic landmarks are anatomical and geometric identifications that are important within samples, providing local and uniform information over the entire sample. These landmarks include morphological features of anatomical components and geometric landmarks of image features like corners, intersections, local minima, and local maxima. These intrinsic landmarks can be identified either through manual segmentation or algorithmic pipelines. By computing measures between these identifications between samples, we can provide measures of registration similarity. Commonly used distances include Euclidean distance, Mahalanobis distance, and Manhattan distance. Registration methods that use intrinsic landmarks can be categorized by their use of different morphological features into three types: point-based methods, curve-based methods, and surface-based methods. The following subsections introduce these methods.

5.3.1. Point-based methods

Point-based methods usually identify clear anatomical structures in an image and position feature points on these structures as the basis for scale-space registration. These methods heavily depend on the quality of information on the targeted anatomical structures but are often universally applicable to different modalities. Computer vision algorithms like the Harris detection algorithm have been applied in neuroimaging to detect and select corners as feature points. Scale-invariant feature transform (SIFT) is also a point-based method that compares keypoint features invariant to translation, rotation, and scaling. The Iterative Closest Point (ICP) algorithm is often applied to optimize feature point selection and register feature points between samples [262]. The ICP iterates the procedure of finding a set of close reference points, calculating the distance measure, and performing the transformation until convergence to an optimal registration. ICP is often applied for multimodal registration and fusion [263]. Point-based methods are often used for or with external landmarks, while feature points can be also be selected based on maximal information content in the context of anatomical geometry [264]. Point-based methods have been applied for landmark registration prior to the fusion of MRI and PET images [265].

5.3.2. Curve-based methods

Point-based methods usually identify characteristic curves or lines, which contain features like edges, object contours, gradient minima, maxima, and crest lines. These features can form representations of anatomical structures and their boundaries. These methods include standard edge detection methods like the Canny edge detector with its subsequent improvements, e.g., the use of curvelets as a replacement of Gaussian filters [266,267]. Second-order approaches based on Laplacian of Gaussian and recent fuzzy logic approaches are also used for feature detection. An alternate approach is the use of contours. The classic 'snake' of active contours are energy-minimizing splines, driven by internal forces and external constraints to approach the lines and edges of anatomical structures [268]. Similar elastic contour approaches have been applied to various image registration applications [269]. Improvements in the classic 'snake' include balloon-based models to reduce dependency on contour initialization [270] and united snakes to combine B-spline functions, FEM functions, etc., [271]. More recent applications include the use of active contours, which tolerate discontinuities by replacing smoothness constraints with masked regularization [272], and the use of curve-based registration for time-series of intensity change in dynamic contrast-enhanced MRI [273].

5.3.3. Surface-based methods

Surfaces or regions are characterized by homogeneous local surface shapes or distinct boundaries. Surfaces inherently provide more redundancy than curves and point landmarks, crucial for non-rigid transformations. Surface-based methods are inherently similar to curve or line-based methods. For some of the prior mentioned point-based or curve-based methods, we can directly extend their formulation to surfaces, e.g., the extension of non-rigid ICP to a point cloud [274]; the extension of 'snake' methods of active contours to 'level sets' involves contour initialization to a surface [275]. While these 'level sets' provide a basis for segmentation, surface or region-based methods can also involve the use of segmentation to isolate surfaces of interest for registration. Brain registration, especially cortical surface registration, involves transforming cortical features to 2D, ellipsoid, or spherical planes or other surfaces for subsequent rigid transformation or deformable warping [276, 277]. Warping is a method of measuring a deformation field between processed and source images. Warping does not necessarily involve registered image and source image but can also include warping based on prior knowledge, e.g., the gradient nonlinearity of MRI magnetic fields. Recent studies combined biomechanical prior with geometric shape prior for surface registration of MRI to transrectal ultrasound [278,279]. Other recent registration studies with segmentation include the 3D active contour segmentation of the liver from abdominal MRI for registration [280].

5.4. Hybrid registration methods

Hybrid registration methods combine intensity-based and feature-based registration methods for higher quality registration. Standard hybrid registration includes a feature-based step and an intensity-based step, where each step is designed to register global or local information [281]. The combination of surface and intensity registration methods is commonly used in neuroimaging to obtain specific brain structures [282], while a recent study has successfully registered 3D curves with 3D surfaces [283]. Another type of hybrid registration uses a hierarchical approach, where sample data are converted to a hierarchy of resolutions, where registration is performed at each level and combined for final registration. The use of hierarchical registration avoids local minima with the global information provided by low-resolution levels, while multiple registrations at various medium to high-resolution levels reduce the requirement for bootstrapping optimizations, resulting in higher computational efficiency [284]. For intensity-based methods, this process can involve a single atlas or reference resized to various hierarchies or multiple atlases in each level [285], while hierarchical registration can also be applied with feature-based registration at each or selected levels [286]. The combinations of multiple registration levels or various registration methods improve registration quality [287]. Hierarchical registration is often applied to registrations with differing methods on partial data or transformed data common in multimodal data fusion [288].

5.5. Optimization for registration

The majority of registration and alignment procedures can be formulated as optimization problems. Therefore, the method of optimization is of vital importance. In this section, we will introduce some of the fundamental approaches to optimization, including gradient descent, Newton's method, and Powell's method. We will also mention recent advances in global optimization, including evolutionary algorithms and deep learning. Gradient descent (GD) is a major category of optimization methods used for registration. GD searches for local minima in a step-wise fashion, moving towards negative gradient regions. A simple representation is,

$$x_{t+1} = x_t - \gamma \nabla f(x_t) \tag{21}$$

where the x represents variables in registered data, t represents a timestep in optimization, and ∇f is the gradient of the objective function f. GD methods are constrained by defining a convergence criterion, where the optimization process is stopped if the criterion is satisfied. A variety of optimization methods have been derived from fundamental GD [289]. Examples include steepest gradient descent, which applies a simplified first-order Taylor, and conjugate gradient descent, which applies the Gram-Schmidt procedure to orthogonalize gradient vectors in each step of the descent.

Another category of optimization method is based on the classical root find method – Newton's method. A simple representation of this second-order derivative-based method is,

$$x_{t+1} = x_t - \gamma \nabla f(x_t) H_f(x_t)^{-1} \tag{22}$$

where H_f is a matrix of second-order partial derivatives, or Hessian, of the objective function f. Quasi-Newton methods, like the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method, are often applied to avoid the calculation of the complete Hessian. A simplified BFGS, with unit step size, can be used to compute the Hessian using,

$$H_f(x_t) = \frac{\nabla f(x_t) - \nabla f(x_{t-1})}{x_t - x_{t-1}} \tag{23}$$

This is also known as the Secant method. Compared to GD, optimization algorithms based on Newton's method often converges more efficiently due to second-order and iterative information. The Levenberg-Marquardt method combines GD with Newton's method for even higher efficiency.

As an alternative to these gradient-based methods, Powell's method is a gradient-free alternative for registration optimization. It uses iterative line search minimizations to find optimal values of individual variables i and determine the next step with a scalar variable α_i.

$$x_{t+1} = x_t - \sum_i \alpha_i s_i \tag{24}$$

Termination criterions similar to GD convergence criteria are applied to stop the optimization process [290]. Without the need for derivations, Powell's method is significantly more efficient than gradient-based methods like GD and Quasi-Newton. It is often used for the optimization of image registration problems [291]. Powell's method is inherently limited in its degrees of freedom; to mitigate this limitation, alternative gradient-free methods have been developed, e.g., the Nelder-Mead method.

The gradient-based and gradient-free methods discussed above are the most classical and generic algorithms available for optimization—more modern approaches, like evolutionary algorithms, model biological processes. Genetic algorithms (GA) are a popular branch of evolutionary algorithms inspired by natural selection. In GA, solutions are modeled as individuals within a population. Each solution contains a number of parameters modeled as genes, which are used to evaluate the individual's fitness using a fitness function. Using the concept of 'survival of the fittest,' pairs of individuals combine their genes to produce new individuals or solutions, known as offsprings. Other concepts like mutation and termination are also present in most GA. Compared to GD-derived methods and Powell's method, GA is a global optimization method.

Another new approach is the use of universal approximators, like neural networks. Depending on the depth of the neural network used, this approach is also known as deep learning [292]. By reformulating the registration problem as supervised transformation or unsupervised transformation estimation based on similarity metrics, we can incorporate the iterative optimization procedure into the training and optimization of neural networks, opening a range of new methodologies for registration and alignment of biomedical data [293]. Studies have shown that deep learning methods can outperform standard registration and alignment in multiple fields. By formulating neural network outputs into registration metrics, studies have performed deformable registration of neonatal brain MRI and showed better performance compared to mutual information-based methods [293]. Similarly, deep learning has also been applied for unsupervised registration without ground truth references, where a twin translation and transformation network also outperformed a range of standard registration methods [294]. Apart from standard applications of neural networks in a reformulation of the similarity optimization problem, deep learning also includes reinforcement learning and generative models. Reinforcement learning involves a trained agent, typically composed of a policy and value network, which explores the space of transformations for registration. Generative models are neural networks that can generate new data from the provided source; these models include the previously mentioned GAN and cyclic-GAN. These models can either be used as an improved similarity measure optimization framework, unsupervised registration method, or as a means to convert multimodal registration to unimodal registration. More detailed descriptions of these methods, along with examples, can be found in the survey by Blessy and Sulochana [295].

5.6. Quality Assessment

Quality assessment is an essential part of the registration and alignment, especially for data fusion, where the quality of the registered and aligned images directly impacts fusion quality. Quality assessment can be done on two bases: with ground truth reference and without a ground truth reference. The ground truth reference, or optimal solution, can be used to directly calculate registration accuracy and robustness. In recent studies of applying deep learning for registration, classical algorithmic registration results were used as the optimal solution, while the neural networks were used to increase efficiency significantly [296]. However, in most cases, the ground truth reference is not available. Conventional methods include the use of fiducial markers [296], i.e., extrinsic landmarks detailed in Section 5.4, and visual inspection based on morphology. Phantom studies have also been used for quality assessment in intra-modality registrations, especially in thorax imaging [297]. Apart from the previously introduced use of mutual information for unsupervised registration, alternative quality assessment metrics have also been applied, including MSE, peak signal-noise ratio (PSNR), gradient smoothness, and redundant information estimation. Another alternative quality assessment method for registration without ground truth references is consistency analysis, which involves registering images in reverse order from the registered data to source data [298,299]. This process directly compares the reconstructed sample and the ground truth original, which allows for computations and optimizations based on consistency measures. The quality assessment procedure in the registration and alignment of biomedical data provides a basis for evaluating fusion quality and is therefore essential to the fusion process.

5.7. Practical Applications

5.7.1. Neuroimaging – MRI-PET Registration

A typical application of registration among modalities is in the fusion of MRI and PET data in neuroimaging. The fusion of MRI and PET data usually requires the registration of MRI data and the co-registration of PET data. MRI registration usually requires B_1-field and gradient nonlinearity correction, which corrects magnetic field inhomogeneities in the imaging apparatus [300]. This correction is usually followed by intensity normalization with histogram peak sharpening and removing the skull and cerebellum components from the brain images with bootstrapping threshold approximations. MRI images are then registered to a brain template for spatial normalization, usually through linear intensity-based registration. Brain templates are the spatial standards for the human brain, generated from neuroimaging or autopsies of a single individual, e.g., Talairach or Collin-27, or a group of subjects, e.g., MNI-152. In many studies, PET images are taken with MRI images. The PET images are then aligned to the corresponding registered MRI images through rigid alignment [301,302]. Through the process of co-registration, spatial alignment and normalization are inherited by the co-registered PET images. The two modalities' fusion can then be performed from the feature-level to voxel-level [303]. Many studies also performed brain segmentation into specific anatomical regions, where ROI-specific features like volume or cortical thickness can be obtained. With co-registered PET images, ROI-specific features of MRI and PET can be combined to expand feature space.

Table 8
Summary of references in Section 5 Alignment and Registration

Reference	Year	Task/Summary	Type	Application
Cohen and Cohen [224]	1993	Active contour models with balloon models.	Curve-based	Various
Maurer, et al. [215]	1997	Head volume registration with fiducial markers	Point-based	Neuroimaging
Studholme, et al. [310]	1999	Entropy measure for regional mutual information (MI)	Intensity-based	Neuroimaging
Maksimovic, et al. [222]	2000	Active contour models for 3D reconstruction and segmentation	Curve-based	Head trauma CT
Christensen and Johnson [247]	2001	Consistence registration through both forward and reverse transformations	Surface-based	Neuroimaging
Jenkinson, et al. [311]	2002	Brain image linear registration and motion correction	Intensity-based	Neuroimaging
Vemuri, et al. [312]	2003	Level-sets of contours for image registration	Surface-based	Neuroimaging
Hellier and Barillot [313]	2003	Hybrid of photometric and landmark-based registration	Hybrid	Neuroimaging
Houhou, et al. [234]	2005	Hierarchical atlas for image registration	Hybrid	Neck CT
Greve and Fischl [205]	2009	Brain image alignment	Surface-based	Neuroimaging
Loizou, et al. [225]	2007	Active contour segmentation of intima-media (carotid artery)	Curve-based	Cardiac ultrasound
Almhdie, et al. [217]	2007	ICP algorithm with lookup matrix	Point-based	Lung and heart data
Xiao-chun, et al. [237]	2007	Lucas-Kanade algorithm based on gradient descent	Intensity-based	Neuroimaging
Postelnicu, et al. [314]	2008	Combination of volumetric and surface registration	Hybrid	Neuroimaging
Gebäck and Koumoutsakos [220]	2009	Edge detection with curvelets	Curve-based	Microscopy images
Danilchenko and Fitzpatrick [245]	2010	Quality assessment with fiducial markers	Point-based	Neuroimaging
Dietzel, et al. [65]	2011	Fusion of DCE-MRI and X-ray mammograms	Hybrid	Breast DCE-MRI and X-ray
De Nigris, et al. [236]	2010	Hierarchical model with adaptive local mutual information	Hybrid	Neuroimaging
Freiman, et al. [255]	2011	Abdominal CT registration with local-affine diffeomorphic demons	Hybrid	Abdominal CT
Gorbunova, et al. [208]	2012	Mass preserving registration for lung CT	Intensity-based	Lung CT
Hu, et al. [221]	2012	Hierarchical image registration based on multi-scale and contour line	Curve-based	Neuroimaging
Lazar, et al. [256]	2013	Batch-effect removal for gene expression data	Intensity-based	Gene expression data
Kim and Tai [235]	2014	Hierarchical model with feature-based registration	Hybrid	Neuroimaging

(continued on next page)

Table 8 (*continued*)

Reference	Year	Task/Summary	Type	Application
Suk, et al. [250]	2014	Hierarchical registration and fusion for deep learning classification of AD/MCI	Hybrid	Neuroimaging
Khallaghi, et al. [230]	2015	Surface registration with biomechanical prior for image fusion	Surface-based	Prostate MRI and transrectal ultrasound
Khallaghi, et al. [230]	2015	Surface registration with statistical biomechanical prior for image fusion	Surface-based	Prostate MRI and transrectal ultrasound
Simonovsky, et al. [243]	2016	Similarity measure modelling as neural network classification task for image registration	Intensity-based	Neuroimaging
Zhang, et al. [248]	2016	Quality assessment based on backward registration	Feature-based	Various
Xu, et al. [254]	2016	Registration methods for Abdominal CT	Various	Abdominal CT
Che, et al. [201]	2017	Ultrasound-to-ultrasound registration	Feature-based	Ultrasound images
Liu, et al. [252]	2017	Multi-level fusion of features for classification of Alzheimer's disease	Hybrid	Neuroimaging
Mahapatra, et al. [209]	2018	Deformable registration with generative adversarial networks (cyclic GAN)	Intensity-based	Retinal images & Cardiac MRI
Li, et al. [226]	2018	Active contour motion segmentation that preserves discontinuities	Curve-based Surface-based	Liver MRI
Raposo and Barreto [232]	2018	Registration of 3D curves with 3D surfaces	Hybrid	Orthopedic models
Liu, et al. [315]	2018	Multi-modal registration for deep learning classification of Alzheimer's disease	Hybrid	Neuroimaging
Mohammadian, et al. [204]	2019	Microscopy image registration with fiducial markers	Point-based	Correlative Microscopy
Xu, et al. [210]	2019	Multi-modal registration with mutual information (MI)	Intensity-based	Various
Alfano, et al. [316]	2019	Breast tumour localisation with pose registration based on breast surface point cloud	Surface-based	Breast CT
Wei, et al. [214]	2019	MRI-CT intra-procedural registration with cycle-GAN for tumour thermal ablation	Intensity-based	Liver MRI and CT
de Vos, et al. [317]	2020	Mutual information with unsupervised deep learning	Intensity-based	Breast MRI and Cardiac MRI
Bhavana [318]	2020	Landmark registration for medical image registration and fusion	Point-based	CT and MRI images
Sun and Feng [227]	2020	Registration for intensity changes in dynamic contrast enhanced (DCE) MRI	Curve-based	Liver DCE-MRI
He and Razlighi [229]	2020	Volumetric registration of brain cortical regions via landmarks and deformation diffeomorphisms	Surface-based	Neuroimaging
Haskins, et al. [240]	2020	Application of deep learning in medical image registration	Various	Various
Arar, et al. [244]	2020	Unsupervised multi-modal image registration with task-specific three neural networks	Intensity-based	General

5.7.2. *Chest and Abdominal Imaging*

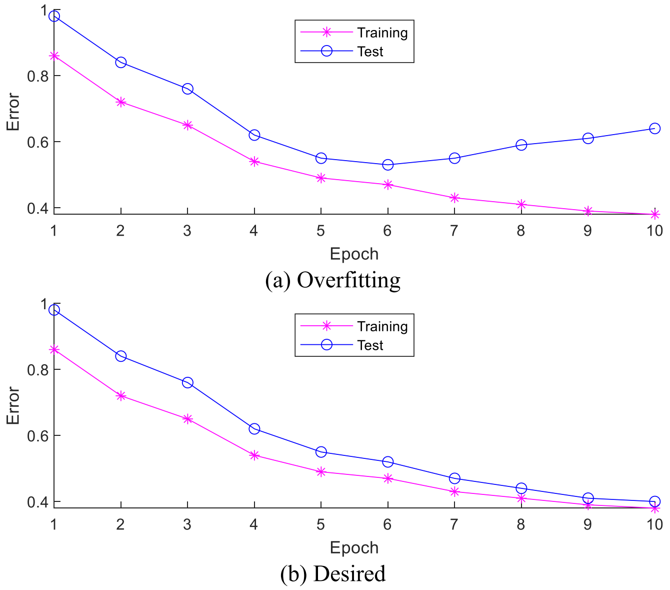
For various conditions and diseases involving the chest and abdominal region, combinations of X-ray, ultrasound, CT, PET, and MRI are often used for diagnosis. Spatial or temporal registration of samples or multiple modalities can provide better bases for fusion. Unlike the registration of relatively non-elastic neuroimaging data, chest and abdominal imaging cover multiple easily deformable organs and therefore require more sophisticated nonlinear registration methods. Feature-based registration with extrinsic landmarks of skin markers has been studied for thorax CT and SPECT registration [304]. The registration of MRI with real-time ultrasound was applied for better biopsy procedures in potential breast cancer cases [305]. Talas, et al. [306] combined 2D X-ray mammograms with 3D MR mammograms with a nonlinear deformation model. Goodfellow, et al. [307] adapted multiple image registration pipelines, e.g., FSL, ANTs, designed for neuroimaging data for abdominal CT, while non-rigid techniques based on local affine assumptions have been applied to CT and DTI [308]. In most cases, rigid transformations cannot adapt to the abdominal and chest regions' elastic and deformable physiology.

5.7.3. *Genetic data*

Although registration and alignment are primarily targeted at neuroimaging data, similar methods are also applied in 1D or sequential data. A prime example is gene expression data, where data from multiple studies are often combined to increase sample size, or perform the meta-analysis. There are often multiple identifiers for a single gene. Therefore, gene identifiers from a single source or multiple sources often contain platform-specific identifiers like Illumina Gene ID or Affymetrix Gene ID, which must be aligned into a single framework, e.g., Entrez Gene, Ensemble Gene Identifiers. Due to the use of multiple studies, it is common for a study to contain gene expression data from multiple platforms, which have inherently different methods for the measurement of gene expression. This combination will result in batch effects and differences in scale within the genetic data, which need to be removed to align different studies [309]. The summary of the recent Alignment and registration research is shown in Table 8.

6. Preprocessing for small size dataset: Data Augmentation

Medical data are normally of small size [319]. The successes of deep learning algorithms fuel the interest in applying deep neural network models to medical image analysis, classification, segmentation, data fusion, etc. However, a small-size dataset will impair the generalization ability of deep neural network models. This generalization means the performance gap of a model evaluated on the test set and training set. This section will give a brief survey on data augmentation, which is an efficient image-domain solution to overfitting.



(a) Overfitting

(b) Desired

Fig. 18. Training and test performance: (a) Overfitting; (b) Desired

6.1. Background

Fig. 18(a) shows the overfitting curves, where the test error increases after epoch 6 as training error continues to decrease. Fig. 18(b) presents a pair of desired curves, where both training and test error decrease until convergence.

Traditional solutions to small-size dataset problems consist of data generation (DG), regularization, and ensemble approaches (EA). DG creates data from a sampled data source. The synthetic minority oversampling technique (SMOTE) [320] is a typical algorithm for DG. Regularization is mainly for the weights of models. Large weights will make the models unstable because minor variations on the inputs will yield large differences in the output for large weights. Smaller weights are regarded to be more regular (i.e., less specialized). Hence, this type of technique is called weight regularization. EA methods use multiple models to obtain better predictive performance than any model alone [320].

Data augmentation (DA) is an approach that solves overfitting by addressing the root of the problem, the training set. The augmented data represent a more comprehensive set of training data, thus minimizing the distance between the training set and test set. Fig. 19(a) shows the distance between the training and test set, where each dot means a sample image. It shows training set cannot cover the characteristics of the test set, so the trained model may overfit. Fig. 19(b) shows the training set zone is enlarged and covers the test set zone; hence, now the distance between the augmented training set and the test set is minimized.

It should be noted that data augmentation is mainly used for image recognition, particularly medical image classification. This is because medical image collection is quite expensive and labor-intensive. Medical images are usually generated by positron emission tomography (PET), computer tomography (CT), ultrasound (US), single photon emission computed tomography (SPECT), magnetic resonance imaging (MRI), functional MRI (fMRI), Magnetic resonance spectroscopy imaging (MRSI) scanning, etc. Other factors also complicate medical image collection, such as expensive and laborious imaging scanning, patient privacy concerns, disease rarity, and the requirement of radiologists' delineation. However, data augmentation can also be used in object detection carried out by R-CNN [321], fast FCNN [322], and faster RCNN [323], YOLO [323], YOLO9000 [324], YOLOv3 [324], etc. Semantic segmentation is a rising application field of data augmentation.

The safety of a type of data augmentation is another important factor. Suppose an image I, and its corresponding label is C. A safe data augmentation D is defined as

$$C\left[D_{\text{safe}}(I)\right] = C(I) \tag{25}$$

Namely, the data augmentation is label-preserving. In some cases, the unsafe data augmentation method will change the labels as

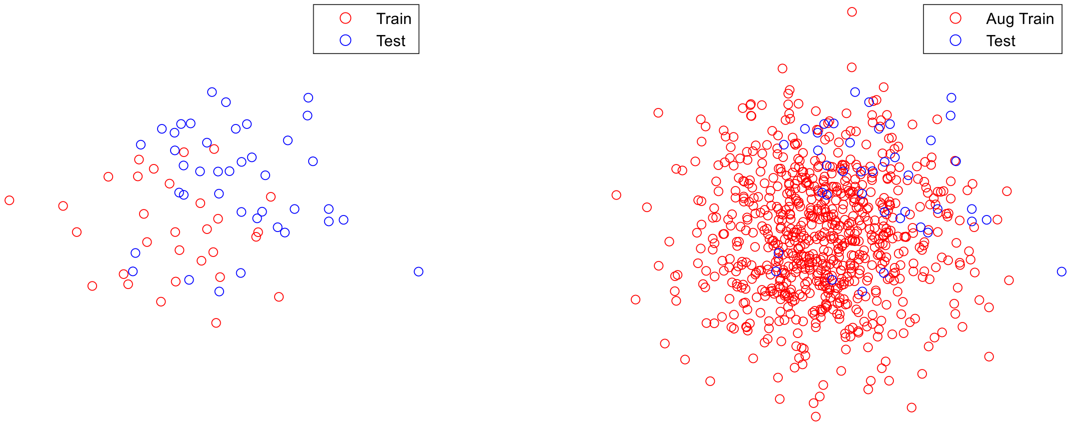
$$C\left[D_{\text{not}-\text{safe}}(I)\right] \neq C(I) \tag{26}$$

Note that "safety" is domain-dependent [325], and its certification needs expert knowledge. For example, rotation is safe for vehicle classification (See Fig. 20(a)) but not safe for digit recognition since 9 will be rotated to 6 (See Fig. 20(b)). The injection of a small amount of noise is safe for lung disease recognition (See Fig. 20(c)), but adding a large amount of noise unsafe for the same task (See Fig. 20(d)).

6.2. Data Augmentation versus other methods

In the context of deep learning, particularly convolutional neural network (CNN) models, there are some special methods to solve small-size dataset problems, such as batch normalization, dropout, pretraining and transfer learning (which will be discussed in the following sections), zero-shot learning, and one-shot learning.

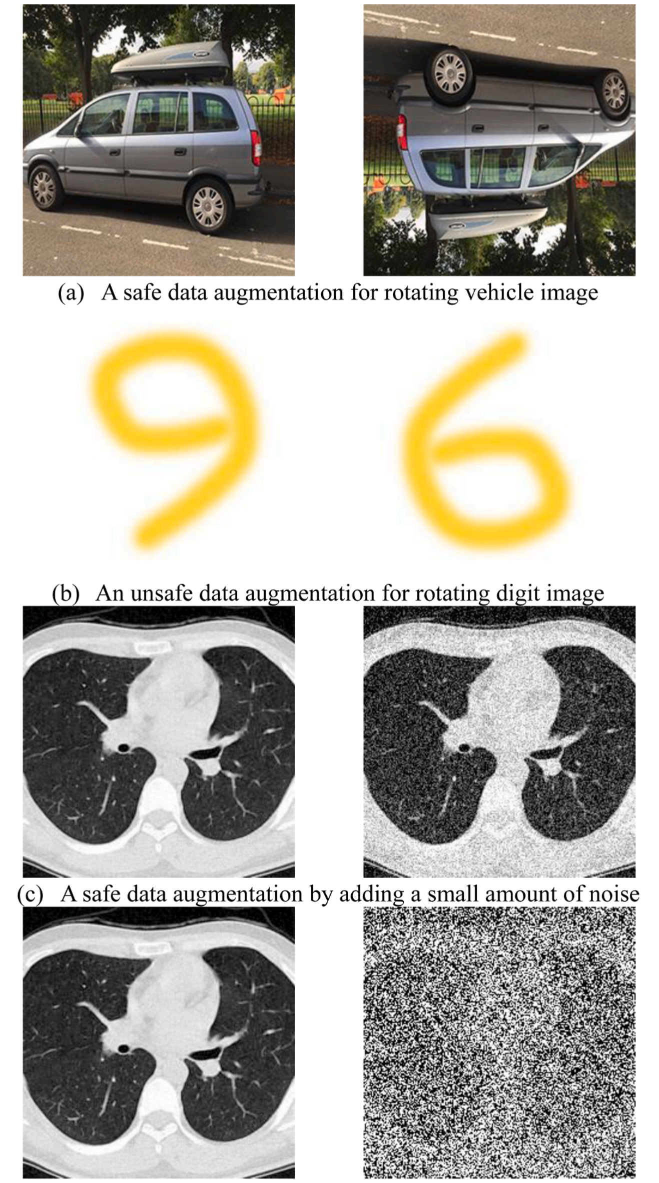
The motivation of batch normalization (BN) is to solve the "internal



(a) Without DA

(b) With DA

Fig. 19. DA help reduce the distance between training and test set



(a) A safe data augmentation for rotating vehicle image

(b) An unsafe data augmentation for rotating digit image

(c) A safe data augmentation by adding a small amount of noise

(d) An unsafe data augmentation by adding a large amount of noise

Fig. 20. Realistic samples of safe and non-safe data augmentations

covariant shift (ICS)", which means the effect of the randomness of the distribution of inputs to internal CNN layers during training. The existence of ICS will worsen the CNN's performance [326]. Suppose that we have N minibatch samples, BN normalizes the internal layer's inputs $\mathrm{B} = \{\beta_i\}$ over every mini-batch, in order to make sure the batch normalized output $V = \{v_i\}$ has a uniform distribution. Mathematically, BN involves the learning of a function of the form

$$\Bigg\{\underbrace{\beta_i, i = 1, 2.\cdots, N}_{\mathrm{B}}\Bigg\} \mapsfrom \Bigg\{\underbrace{v_i, i = 1, 2, \cdots, N}_{V}\Bigg\} \tag{27}$$

During training, the empirical mean a_m and empirical variance a_v can be computed as

$$\begin{cases} a_m = \dfrac{1}{N}\left(\sum_{i=1}^{N}\beta_i\right) \\ a_v = \dfrac{1}{N}\sum_{i=1}^{N}(\beta_i - a_m)^2 \end{cases} \tag{28}$$

The input $\beta_i \in \mathrm{B}$ was first normalized to $\hat{\beta}_i$

$$\hat{\beta}_i = \frac{\beta_i - a_m}{\sqrt{(a_v + a_s)}} \tag{29}$$

where a_s in denominator in Eq. (29) is stability factor, used to enhance the numerical stability. Now the $\hat{\beta}_i$ have zero-mean and unit-variance characteristics. In order to have a more expressive deep neural network [327] (here expressive means the network's expressive power, i.e., the ability to express functions), a transformation is usually carried out as

$$v_i = b_1 \times \hat{\beta}_i + b_2, i = 1, 2, \cdots, N \tag{30}$$

where the parameters b_1 and b_2 are two learnable parameters during training. The transformed output $v_i \in V$ is then passed to the next layer and the normalized $\hat{\beta}_i$ remains internal to the current layer.

Fan, et al. [328] proposed the concept of dropout neurons (DNs) by randomly dropping neurons and setting their neighboring weights to zero during training. The selections of DNs are random with a retention probability (γ_p). Suppose we have a neuron $N(i,j)$ and its corresponding original weights are $t(i,j)$, and the collection of DNs is δ.

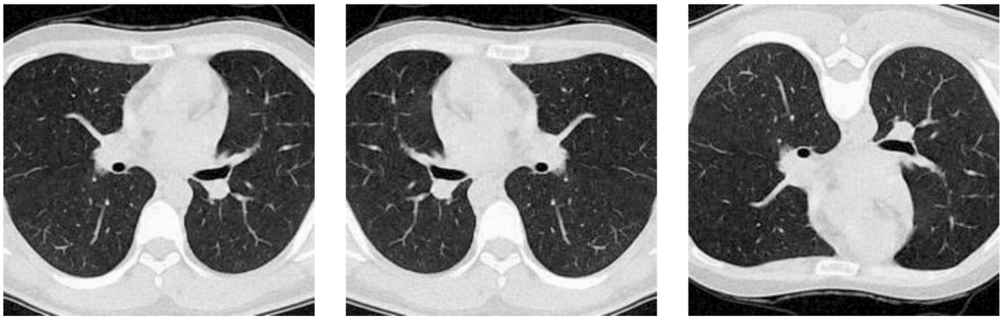
$$t_t(i,j) = \begin{cases} t(i,j) & N(i,j) \in \delta \\ 0 & N(i,j) \notin \delta \end{cases} \tag{31}$$

where $t_t(i,j)$ means the weights of neuron $N(i,j)$ during training. γ_p has a default value of 0.5, viz., $\gamma_p = 0.5$. During inference, we run the entire CNN without dropout, but the weights of FCLs using DNs are downscaled (viz., multiplied) by γ_p:

$$t_i(i,j) = \gamma_p \times t(i,j) \tag{32}$$

where $t_i(i,j)$ denotes the weight of neuron $N(i,j)$ during inference.

One-shot learning or few-shot learning is to learn information about



(a) Original Image

(b) Horizontal flipping

(c) Vertical flipping

Fig. 21. Horizontal flipping versus vertical flipping

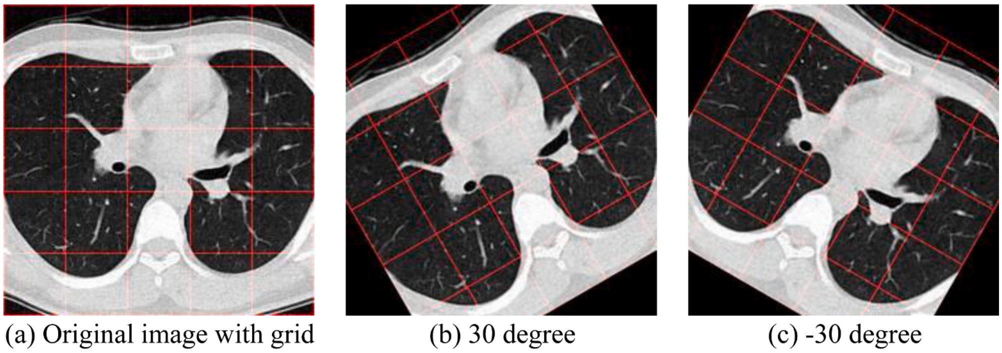

(a) Original image with grid (b) 30 degree (c) -30 degree

Fig. 22. Rotation results

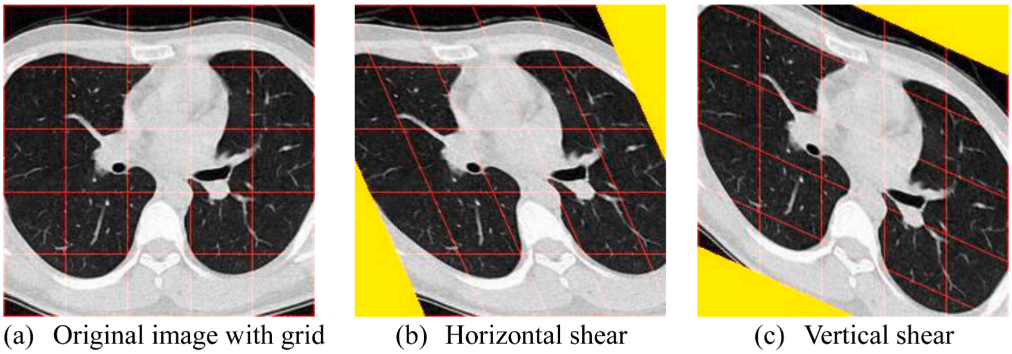

(a) Original image with grid (b) Horizontal shear (c) Vertical shear

Fig. 23. Shear transform results

object classes from one or only a few training samples, respectively. Their motivation is given by humans' ability to learn object classes from few examples. One-shot learning is now successfully applied in medical image registration [329], hand gesture recognition [330], expert-aided systems [331], etc. Two commonly-used methods are Siamese networks (which learn a distance function) [332] and memory-augmented networks [273]. Zero-shot learning [333] is an extreme paradigm where at test time, the trained classifier needs to predict samples from classes that were not observed during training.

6.3. Geometric Transforms

Flipping. Flipping in geometry means the image is reflected along a line, leading to a mirror image of the original one. Vertical flipping is less common than horizontal flipping. The flipping is one of the simplest and most straightforward data augmentation methods [275]. Experiments on ImageNet, CIFAR-10, and other biomedical datasets prove the effectiveness of flipping. Note that on datasets such as SVHN or MNIST, which involve texts and digits, flipping is unsafe. Fig. 21(a) shows an original lung window image, and Fig. 21(b and c) present the corresponding horizontal and vertical flipping results.

Rotation. Rotation is a motion of an image around a point. Usually, a clockwise rotation is a negative magnitude, while a counterclockwise rotation is a positive magnitude. In a data augmentation situation, the image is rotated around the central point [334]. Slight rotation such as within $[-15^{\circ}, 15^{\circ}]$ are usually safe for digit recognition and text recognition, but a wide rotation such as within $[-90^{\circ}, 90^{\circ}]$ may be unsafe, i.e., the label is no longer preserved. Fig. 22(a) shows the original lung image with grid lines colored in red. Fig. 22(b and c) present the rotation results with rotation angles of 30 degrees and -30 degrees, respectively.

Shear. Shear mapping displaces each point in a fixed direction by an amount that is proportional to its signed distance from the line passing through the origin and parallel to that direction [277]. If we suppose the original pair of coordinates is $[x_1, y_1]$, and the pair of coordinates after shear transform is $[x_2, y_2]$, then horizontal shear is defined as

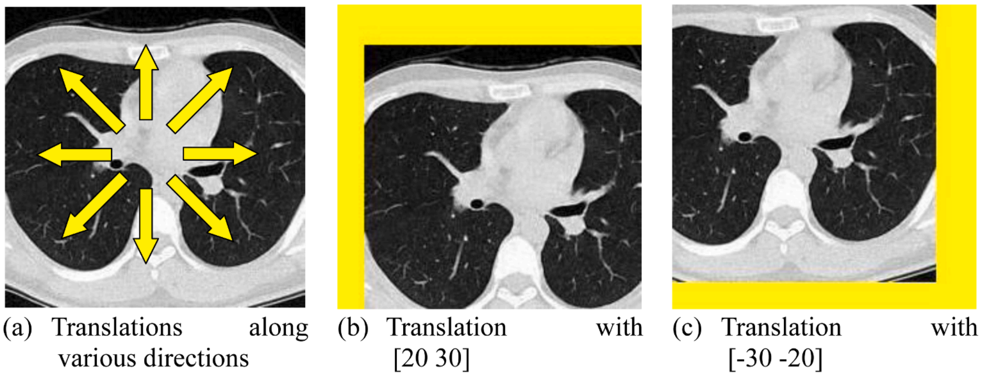

(a) Translations along various directions (b) Translation with [20 30] (c) Translation with [-30 -20]

Fig. 24. Translation schematic and results

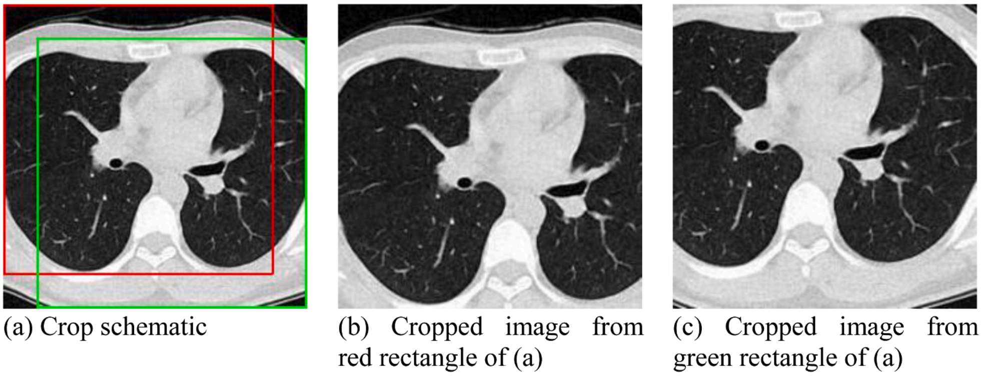

(a) Crop schematic (b) Cropped image from red rectangle of (a) (c) Cropped image from green rectangle of (a)

Fig. 25. Crop results

$$[x_2, y_2, 1] = [x_1, y_1, 1] \times \begin{bmatrix} 1 & 0 & 0 \\ a_{hs} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{33}$$

where a_{hs} is the horizontal shear factor. Similarly, for vertical shear we can define as

$$[x_2, y_2, 1] = [x_1, y_1, 1] \times \begin{bmatrix} 1 & a_{vs} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{34}$$

where a_{vs} is the vertical shear factor. Fig. 23(a) presents the original lung image with red grids, and Fig. 23(b and c) present the corresponding horizontal and vertical shear results, respectively.

Translation. Translation in geometry is to move every pixel in the image by the same distance along the same direction. The translation is commonly used in face recognition. Face images are typically collected in almost perfectly centered positions, which will necessitate the AI classifier to be tested on similarly centered images. Using the translation data augmentation method, the dataset will be filled with other translated images (face not in the center), so the classifier will become more robust and can work efficiently on images where faces are not centered. There will be “missing” values when images are translated outwards of the original image size, so we need to fill in those missing values with either a constant such as 0 (black) or 255 (white) or random noise [335]. Fig. 24(a) presents a schematic showing the translation can move the image along the same direction. Fig. 24(b-c) provides two translation results of [20,30], and [-30, -20], respectively.

Cropping. In traditional image processing and computer vision tasks, cropping is an efficient tool to extract patches from a large image or a mixed-size image set [336,337]. Then algorithms are run on the patches instead of the images themselves. In the data augmentation domain, cropping cuts a patch with a predefined size out of the original image. The difference between cropping and translation is that cropping reduces the spatial size while translation preserves the spatial size. For example, if the original size is $[W_0, H_0]$, then the size after cropping is $[W_c, H_c]$ and the size after translation is $[W_t, H_t]$, we have

$$\begin{cases} W_t = W_0, H_t = H_0 \\ W_c < W_0, H_c < H_0 \end{cases} \tag{35}$$

Fig. 25(a) shows the crop schematic where two rectangles (red and green) delineating the regions to be cropped. Fig. 25(b and c) show the cropped images from red and green rectangles, respectively.

Geometric transformations are popular data augmentation solutions to increase the amount of training data [337]. The advantage of geometric transforms is that they are easy to carry out. The disadvantage is additional computation cost and storage memory, and extra training time. The geometric transformation must be observed carefully since some of them may alter the image labels.

6.4. Noise Injection

Gaussian Noise. Noise injection means adding noise to the inputs of a deep neural network model during training. The noise is usually set as a Gaussian noise, which is statistical noise having a probability density function (PDF) equal to normal distribution. The description of Gaussian noise is illustrated n section 3.2.1. Noise injection has proved successful in robot speech commands [338], fruit classification [339], plant leaf disease recognition [340], etc.

Salt-and-pepper Noise. Salt-and-pepper noise, as described in section 3.2.3 is another common noise to be added to input images. Calderoni, et al. [341] used salt and pepper noise for the identification of early esophageal cancer.

Speckle Noise. Speckle noise mentioned in section 3.2.5 is a granular interference that inherently exists in medical ultrasound (US) images, active radar, synthetic aperture radar, etc. Data augmentation with speckle noise has been proved efficient in radar images [342] and neonatal hip US images [343].

Noise added in other layers. All the previous methods inject noises at

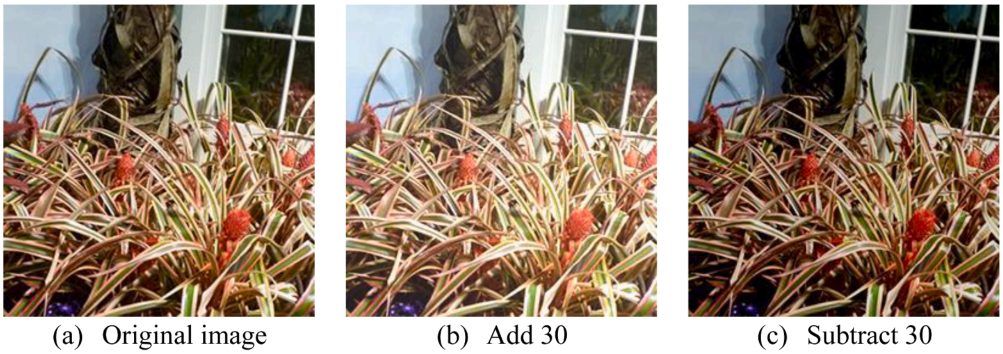
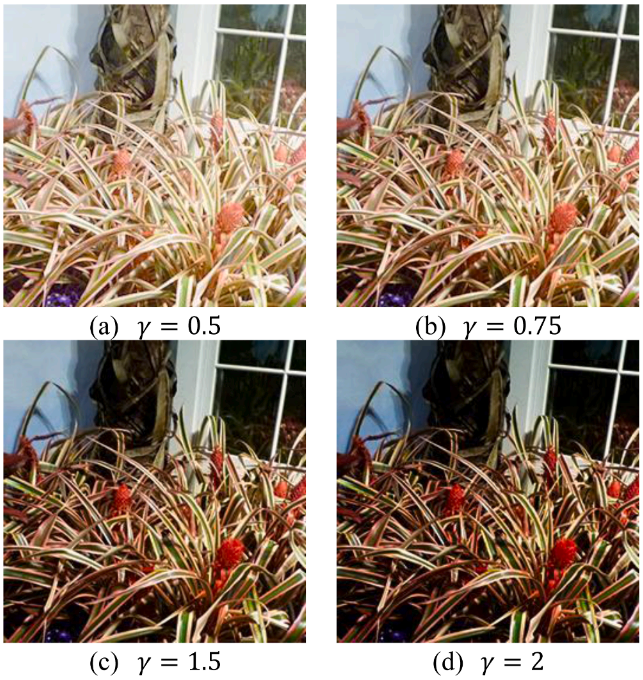

(a) Original image (b) Add 30 (c) Subtract 30

Fig. 26. Simple photometric transform by adding and subtracting a constant value



(a) $\gamma = 0.5$ (b) $\gamma = 0.75$

(c) $\gamma = 1.5$ (d) $\gamma = 2$

Fig. 27. Gamma correction

the input layer; however, noises can be added at other layers. For example, Davatzikos, et al. [344] added noise, interpolated and extrapolated in learned feature spaces. Gothelf, et al. [345] added noise to the loss layer and presented a novel method, “DisturbLabel”, which randomly replaced a part of labels as an incorrect categorical value in each iteration. Their experiments demonstrated that DistrubLabel could prevent the network training from overfitting.

6.5. *Photometric Transforms*

Photometric transforms, also known as color space transform, is to manipulate the gray values of a grayscale image or to manipulate RGB color values of a color image [346]. A simple method is to add or subtract a constant value to increase or decrease the gray values of the image, making it brighter or darker. Fig. 26 gives a simplistic example, where (a) shows the raw image and (b-c) present the result by adding 30 to and subtracting 30 from the raw image, respectively.

Gamma Correction. Gamma correction is a nonlinear operation to adjust the luminance values of the images. It is defined by the power-law expression:

$$f_o = A \times f_i^{\gamma} \tag{36}$$

where f_i and f_o denote the input and output gray values, and their values are normalized into the range of $[0,\ 1]$ so $A = 1$ will preserve the gray scale range. Two important ideas exist: (i) gamma compression associated with $\gamma < 1$; and (ii) gamma expansion associated with $\gamma > 1$ [338]. The top row in Fig. 27 presents two samples of gamma compression, i.e., $\gamma = (0, 5, 0.75)$ respectively. The bottom row in Fig. 27 presents two other samples of gamma expansion with γ equivalent to 1.5, and 2, respectively.

Color Jittering. Color jittering (CJ) [347] shifts the color values in original images by adding or subtracting a random value. The benefit of CJ is that it can help bring in randomness change to the color channels, so it can aid the production of fake color images. Fig. 28 shows six color jittering examples on the raw image in Fig. 26(a).

PatchShuffle. Kang, et al. [293] presented a new PatchShuffle method. In each minibatch, images are split into nonoverlapped patches, and each patch undergoes a transformation such that pixels within that patch are shuffled. They conducted experiments with different filter sizes n and different swapping probabilities ε. Suppose that the original image is X with size of $N \times N$, and X is partitioned into a block matrix with non-overlapping patches $X = \{a_{ij}\}$, a_{ij} means the patch at i-th row and j-th column within the patch matrix.

$$X = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1,N/n} \\ a_{21} & a_{22} & \cdots & a_{2,N/n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N/n,1} & a_{N/n,2} & \cdots & a_{N/n,N/n} \end{bmatrix} \tag{37}$$

The PatchShuffle transformation acts on each patch by

$$\widetilde{a}_{ij} = p^r_{ij} \times a_{ij} \times p^c_{ij} \tag{38}$$

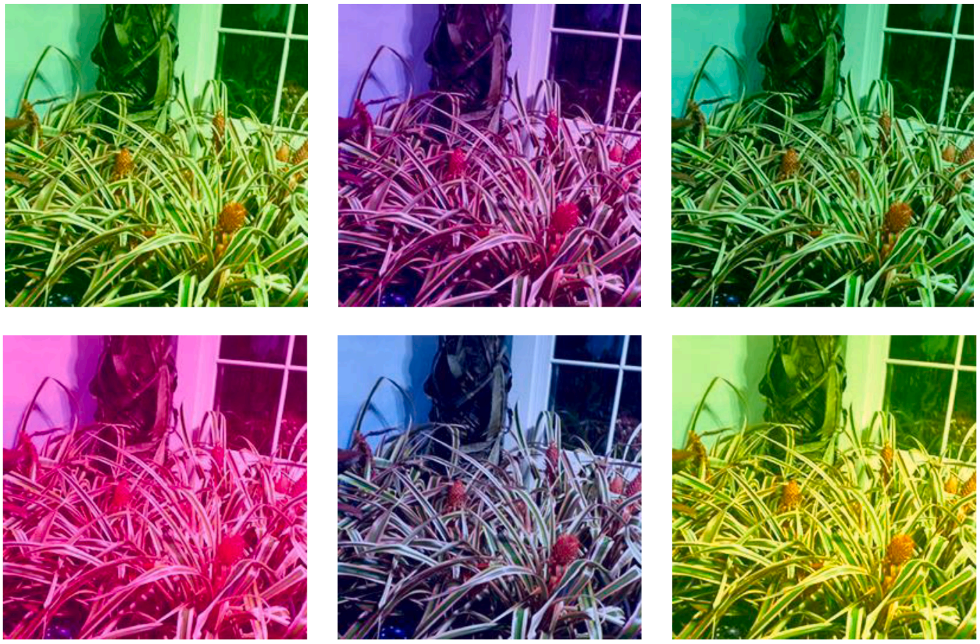
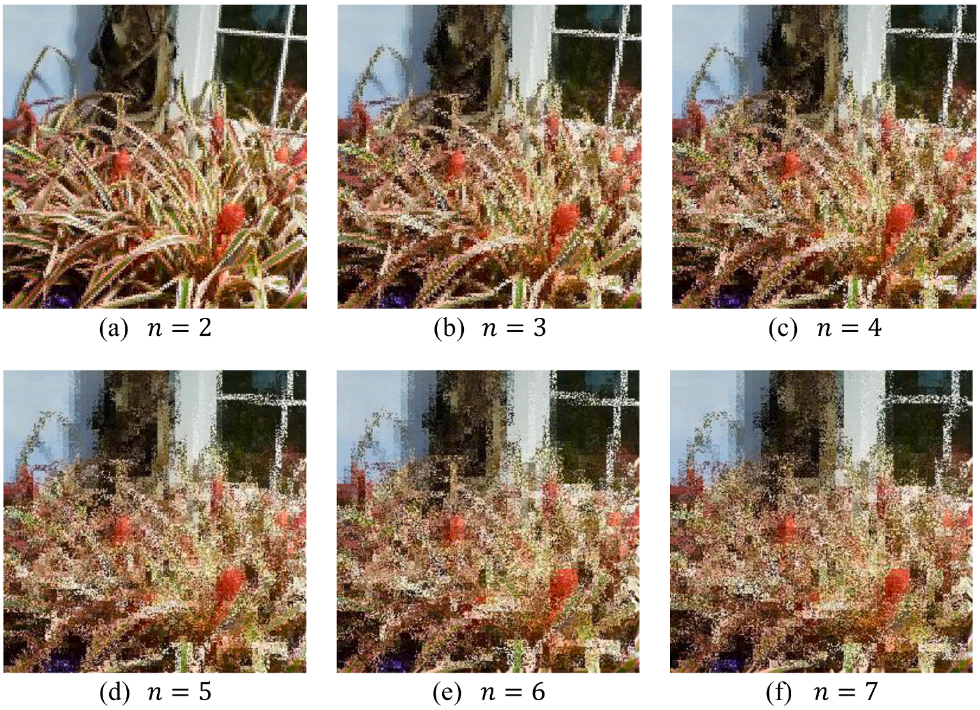


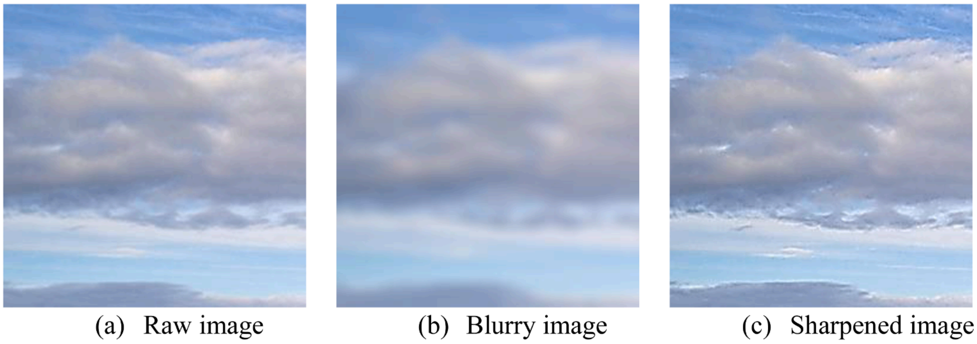
Fig. 28. Color jittering examples



(a) $n = 2$ (b) $n = 3$ (c) $n = 4$

(d) $n = 5$ (e) $n = 6$ (f) $n = 7$

Fig. 29. PatchShuffle results



(a) Raw image (b) Blurry image (c) Sharpened image

Fig. 30. Blurring and sharpening results

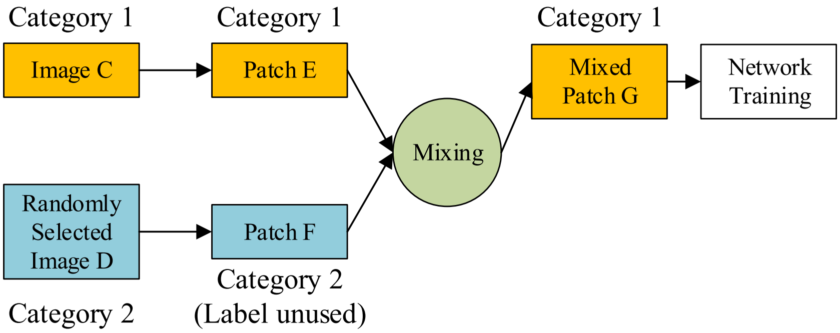
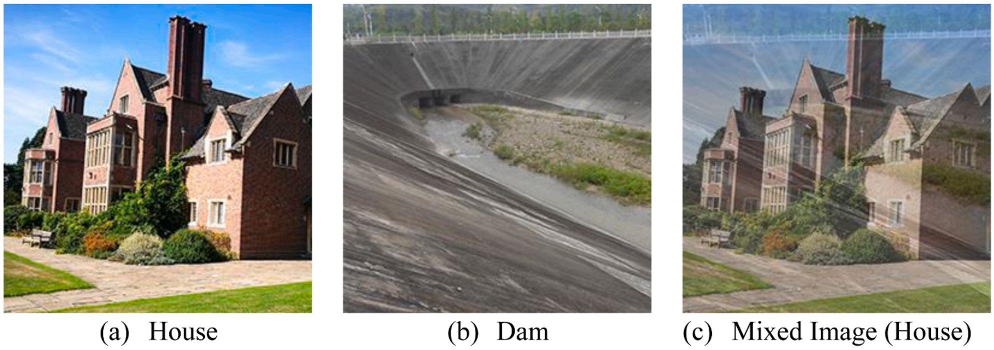


Fig. 31. Schematic of SamplePairing

where $\widetilde{a}_{ij}$ denotes the transformed patch, p^r_{ij} and p^c_{ij} denote the row and column permutation matrixes, respectively. Their experiments showed the optimal hyperparameter is $n = 2$ and $\varepsilon = 0.05$. Fig. 29 shows the PatchShuffle results with $n = 2, 3, \cdots, 7$. In their paper, Tibshirani [348] reported PatchShuffle could be applied not only on images but also on feature maps.

Sharpening and blurring. Kernel filters can be used to sharpen and blur images. The kernel filters slide an $n \times n$ kernel across the image with either a Gaussian blur filter [294] or an unsharp masking [349]. The former yields a blurry image, while the latter yields a sharpened image. Fig. 30(a) shows a raw cloud image, while Fig. 30(b and c) show the blurry and sharpened images, respectively.



(a) House (b) Dam (c) Mixed Image (House)

Fig. 32. A SamplePairing example of missing house and dam images

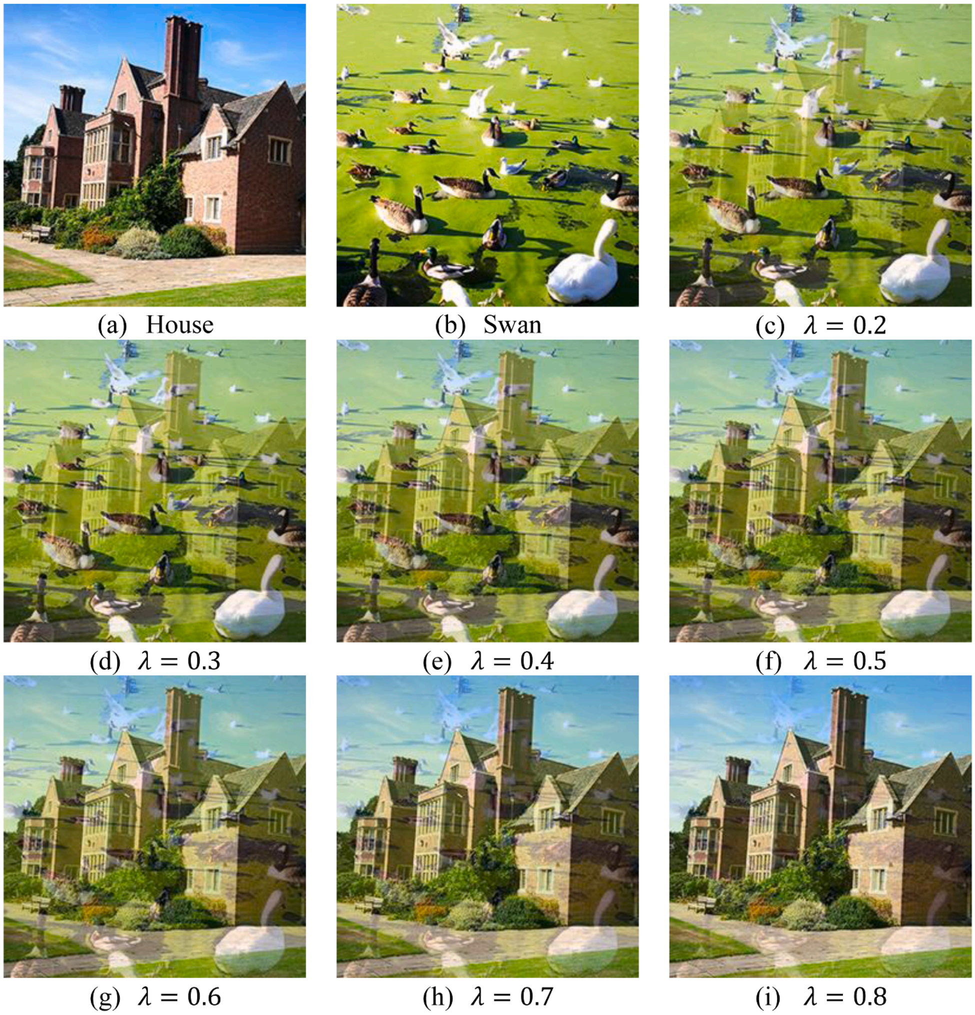
Fig. 31.

Intuitively, blurring images can help classifiers better resist the blur (Gaussian, motion, average, etc.) during the test, and sharpened images bring about more edge and contrast details for object category classifications. Both sharpening and blurring operations are quite common in data augmentation.

6.6. *Image Mixing*

Sample Pairing. The afore-mentioned transformation methods are single-image augmentation methods. Now we will discuss more novel methods working on two or more images. McIntosh and Lobaugh [350] proposed SamplePairing technique, which synthesizes a new training sample from one image by overlaying another image randomly chosen from the training data. That is, to take an average of two images in a pixel-wise way.

Suppose there is an image C of category 1, and another randomly selected image D of category 2. SamplePairing method first generates two patches E and F from the image C and D, respectively, by random cropping method and random horizontal flipping. The category 2 label is discarded. Then the two patches E and F are mixed to generate the mixed patch G by averaging intensities of two patches pixelwise.

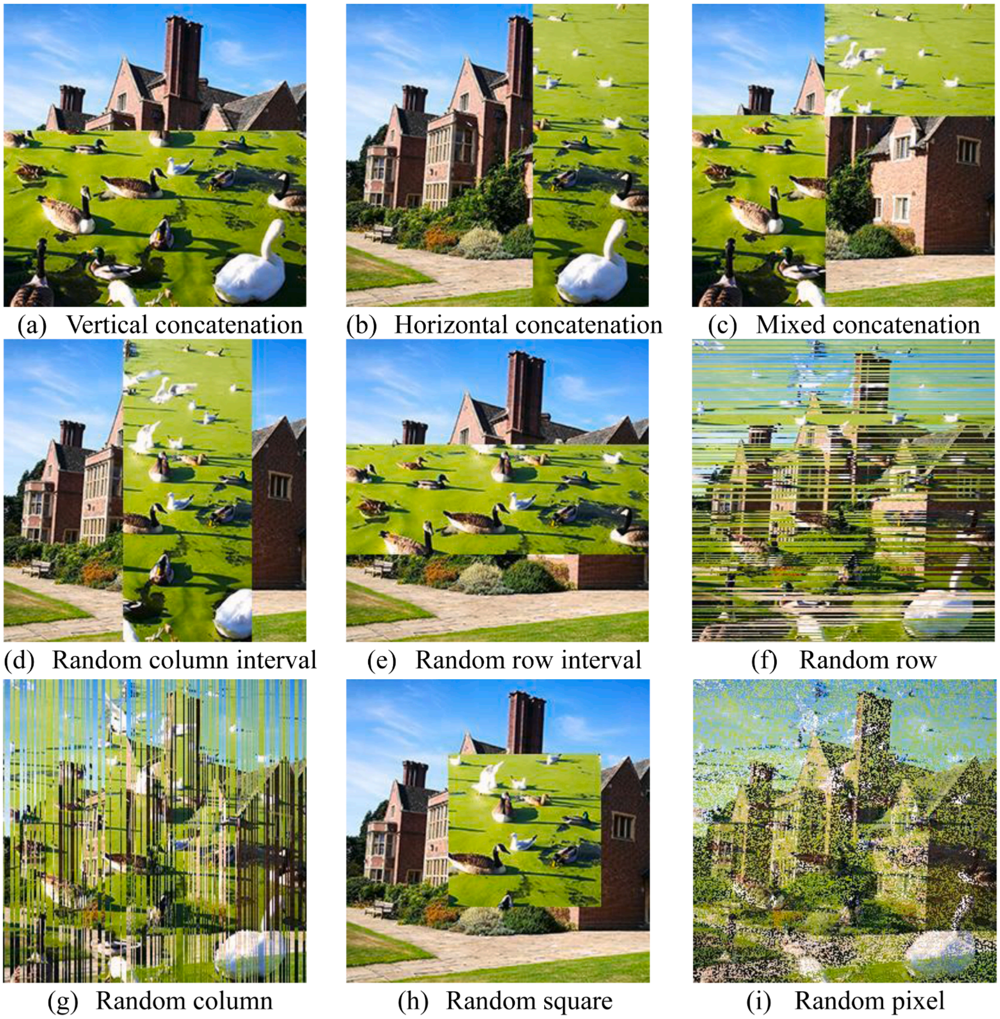


(a) House (b) Swan (c) $\lambda = 0.2$

(d) $\lambda = 0.3$ (e) $\lambda = 0.4$ (f) $\lambda = 0.5$

(g) $\lambda = 0.6$ (h) $\lambda = 0.7$ (i) $\lambda = 0.8$

Fig. 33. Mixup results


(a) Vertical concatenation (b) Horizontal concatenation (c) Mixed concatenation

(d) Random column interval (e) Random row interval (f) Random row

(g) Random column (h) Random square (i) Random pixel

Fig. 34. Nonlinear mixing method

$$\underbrace{G(i,j)}_{\text{C1}} = \left[\underbrace{E(i,j)}_{\text{C1}} + \underbrace{F(i,j)}_{\text{C2}}\right] \Big/ 2 \tag{39}$$

where C1 and C2 mean the category labels. Then the mixed patch G is used for network training. The authors claimed using their Sample-Pairing method can generate N^2 new samples from N training sample dataset [351]. Fig. 32(a-b) presents the house and dam images, respectively. Fig. 32(c) shows the mixed image with a label of “house”.

Mixup. A data-agnostic augmentation routine, mix-up, was proposed by Yan, et al. [352]. In their paper, a hyperparameter $\lambda \in [0,1]$ was introduced, and one-hot label encoding was used to use the information of categories of both images. Suppose (E,F) mean the two randomly selected samples, and t the label of corresponding categories, we can get the mixup sample and labels G and t_G as

$$\begin{cases} G = \lambda \times E + (1-\lambda) \times F \\ t_G = \lambda \times t_E + (1-\lambda) \times t_F \end{cases} \tag{40}$$

where (t_E, t_F) are labels of two samples randomly selected from the training set. Briefly, mixup extends the training dataset by linearly interpolating two randomly selected images. Fig. 33(a-b) gives two randomly selected images: House and Swan, photographed from Leicester botanic garden and Abbey park, respectively. Fig. 33(c-i) presents the mixup results with $\lambda = 0.2, 0.3, \cdots, 0.8$, respectively.

Nonlinear mixing. Vounou, et al. [353] expanded linear combination to nonlinear mixing methods. The authors proposed novel nonlinear mixing methods. Suppose $\lambda \in [0,1]$ is a random variable, the vertical concatenation (VC) combines the top λ fraction of image E and the bottom $(1-\lambda)$ fraction of image F, instead of pixelwise average. Suppose (W,H) are the width and height of the input image, and (w,h) are the width and height index, we have

$$G^{VC}(w,h) = \begin{cases} E(w,h) & h \le \lambda H \\ F(w,h) & \text{otherwise} \end{cases} \tag{41}$$

And the horizontal concatenation (HC) is described as

$$G^{HC}(w,h) = \begin{cases} E(w,h) & w \le \lambda W \\ F(w,h) & \text{otherwise} \end{cases} \tag{42}$$

Mixed concatenation (MC) is an application of horizontal concatenation to the vertical concatenation of two input images. Namely, suppose we have $0 \le \lambda_1, \lambda_2 \le 1$ two random variables, then

$$G^{MC}(w,h) = \begin{cases} E(w,h) & h \le \lambda_1 H \wedge w \le \lambda_2 W \\ F(w,h) & h \le \lambda_1 H \wedge w > \lambda_2 W \\ F(w,h) & h > \lambda_1 H \wedge w \le \lambda_2 W \\ E(w,h) & h > \lambda_1 H \wedge w > \lambda_2 W \end{cases} \tag{43}$$

Random column interval selects a random column interval, and that interval part of the image E is replaced with image F. Random row interval does the same thing on row direction. The random row method selects each row at random either from image E or F. The random row method can be regarded as a higher frequency of vertical concatenation. Similarly, we can deduce the random column method. Random square is

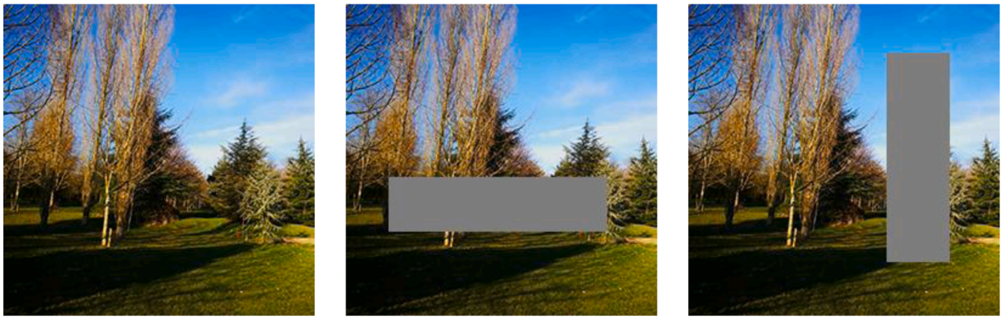

(a) Arboretum (b) 1st sample of RE (c) 2nd sample of RE

Fig. 35. Example on random erasing

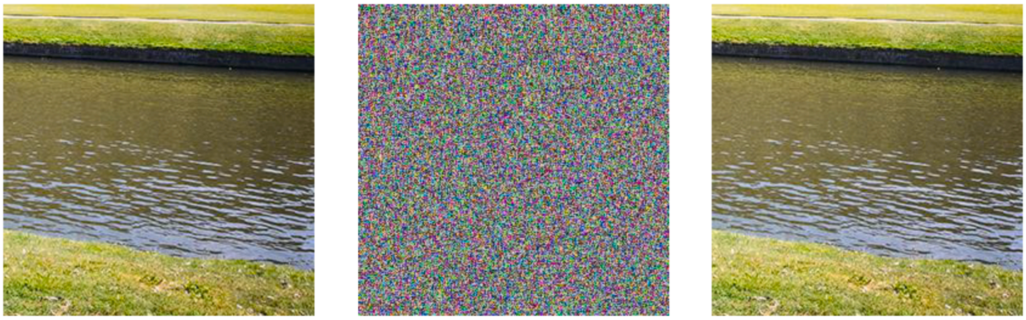

(a) “river”, 77.8% confidence (b) Perturbation by DeepFool (c) “House”, 99.8% confidence

Fig. 36. Adversarial training examples

to cut out a random square in image *E* with that corresponding region in image *F*. Random pixel samples each pixel separately from both images. Fig. 34 shows nine nonlinear mixing methods.

Random Erasing. Kohannim, et al. [354] introduced a new random erasing (RE) method, which randomly selects a rectangle region and erases its pixels with random values. This RE method is useful to combat image recognition tasks on account of occlusion, which means some parts of the object are blocked. RE forces the model to learn more global features from other unblocked parts.

In practice, RE randomly selects an $n \times m$ patch of an image and masks it with either 0s (black), 255s (white), mean pixel values, or random values. The best patch fill method was proven to be random values. Two hyperparameters in RE are the fill method and the size of the masks [300]. Fig. 35(a) shows one arboretum picture photographed in Shady Lane Arboretum, Leicester. Fig. 35(b-c) shows the two random erasing samples, which we can still observe this is an arboretum.

It should note that RE is not always “safe”. In digit recognition tasks, if the top bar was erased, then “7” may look like “1”. In other fine-grained tasks [355], such as tumor grade classification, the random erasing method may block the tumor itself. Therefore, some intervention strategies should be performed to guarantee the “safety” of the augmented dataset. Also, identifying the makes of vehicles may be impaired since RE may block the brands of vehicles.

6.7. *Deep learning-based Methods*

Adversarial Training. Originally, adversarial machine learning attempts to fool models with deceptive inputs. The adversarial attacking consists of a rival network that learns deceptive augmentation of images that cause misclassification in its rival classification network. Suppose we have an image *A* of category C1, and now we add a small amount of noise to it εN, in which the noise is designed strategically. The summation *B* will fall into another category C2.

$$\underbrace{A}_{\text{C1}} + \varepsilon \times N \underbrace{B}_{\text{C2}} \tag{44}$$

where ε is a small value, usually $\varepsilon < 0.01$. Fig. 36(a) shows an image labelled as “river” with 77.8% confidence. After adding the perturbation by DeepFool [356], this image will be labelled by AI models as “House” with 99.8% confidence.

The adversarial training can be used as an effective data augmentation method to fix weak spots in the traditional AI model. Hence, those trained models will be more robust and resistant to attackers. Adversarial training may not increase the test performance, but it will improve the performance of adversarial examples, i.e., improving the security and robustness of trained AI models. It is noteworthy to add, though, that high-dimensional deep learning models may become inherently unstable to perturbations with high probability as works [303,304] demonstrate.

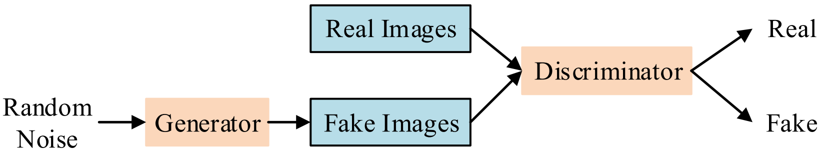


Fig. 37. A schematic of GAN

Table 9
Summary of various DA types and operations

Type	Operation	Description
Geometric Transform	Flip	The image is reflected along a line, leading to a mirror image of the original one. Horizontal flipping is more popular than vertical flipping [275].
	Rotation	Rotation [276] is a motion of an image around a point. The image is rotated around the central point
	Shear	Shear mapping [277] displaces each point in a fixed direction by an amount that is proportional to its signed distance from the line passing through the origin and parallel to that direction.
	Translation	Translation [278] is to move every pixel in the image by the same distance along the same direction
	Cropping	Cropping [279, 280] extracts patches from a large image or a mixed-size image set
Noise Injection	Gaussian Noise (Input layer)	Gaussian noise [282-284] is statistical noise having a PDF equal to normal distribution
	Salt-and-pepper Noise (Input Layer)	It is a type of image noise commonly seen during transmission [285].
	Speckle Noise (Input Layer)	It is multiplicative noise [286, 287], which is usually caused by bad information channels.
	Noise added in other layers	Noises can be added at other layers, such as learnt feature spaces [288] and loss layers [289].
Photometric Transform	Gamma Correction	Gamma correction [291] is a nonlinear operation to adjust the luminance values of the images.
	Color Jittering	CJ [292] shifts the color values in original images by adding or subtracting a random value.
	Patch Shuffle	Images are split into nonoverlapped patches, and each patch undergoes a transformation such that pixels within that patch are shuffled [293].
	Sharpening	Kernel filters of unsharp masking [295]
	Blurring	Gaussian blur filters [294]
Image Mixing	SampleParing	SampleParing [296] synthesizes a new training sample from one image by overlaying another image randomly chosen from the training data.
	Mixup	Mixup [297] extends the training dataset by linearly interpolating two randomly selected images.
	Nonlinear mixing	Nonlinear mixing [298] includes vertical concatenation, horizontal concatenation, mixed concatenation, random column interval, random row interval, random row, random column, random square, random pixel, etc.
	Random Erasing	RE [299] randomly selects a rectangle region and erases its pixels with random values.
DL-based Methods	Adversarial Training	Adversarial training is used as an effective DA method to fix weak spots in the traditional AI model. Those trained models will be more robust and resistant to attackers [303, 304].
	GAN	GAN [305] consists of two neural networks contesting with each other in a zero-sum game, where one network's gain is the other network's loss.

Generative adversarial network. The Generative adversarial network (GAN) consists of two neural networks contesting with each other in a zero-sum game [357], where one network's gain is the other network's loss. There are many generative models that currently exist, but GAN is leading the performances in computation speed and quality. An intuitive anecdote for GAN is a competition between police (Discriminator) and a counterfeiter (Generator), or a predator and prey [358]. Both sides are improving their techniques, so finally, the counterfeiter can make tickets that are hard to recognize as real or fake by the police, see Fig. 37.

The success of the generator makes it powerful for generative modeling. GANs have been proved to be effective in data augmentation. Rao, et al. [359] proposed the first GAN based on multilayer perceptron to handle MNIST handwritten digit image, the size of which is only $28 \times 28 \times 1 = 784$ pixels. Nowadays, the images in recent biomedical datasets are finer resolution and more complicated than MNIST images. Hence some important variants of GANs were commonly used in data augmentation in the biomedical field.

For example, Wan, et al. [360] proposed a new attribute-preserving GAN (APGAN), that provides both attribute-preserving and good visual qualities after style transfer. Marquand, et al. [361] presented a new modified generator GAN (MG-GAN). The difference between MG_GAN and the basic GAN is that the generator in MG-GAN is fed with original data and multivariate noise to produce data with Gaussian distribution. The authors reported MG-GAN improved accuracy by 18.8% and 11.9% compared to KNN and basic GAN, respectively. Krishnan, et al. [362] compared deep convolutional GAN (DCGAN) with auxiliary classifier GAN (ACGAN) for liver lesion classification. The authors found DCGAN provided better results and showed that the GAN-generated CT images could serve as synthetic data augmentation, thus improving the performance of CNN. Using classic data augmentation, the classifier yielded 78.6% sensitivity and 88.4% specificity. While adding synthetic data augmentation, the classifier improved to 85.7% sensitivity and 92.4% specificity.

In summary, we have discussed five types of data augmentation methods: geometric transforms, noise injection, photometric transforms, image mixing, and deep learning-based methods. In practical AI model designing and training, the AI users will try to test one or several different data augmentation methods and combine them together to attempt to achieve better performance. The problem of choosing appropriate data augmentation is still an active research topic. Due to the page limit, the above DA types and operations are itemized in Table 9.

7. Preprocessing for high dimensionality

Dimensionality reduction (DR), or feature reduction, the process to remove noisy and redundant data, is a crucial pre-processing step in data fusion to improve the accuracy of the subsequent modules. If proper methods are applied, the overfitting issue can be avoided while the accuracy and generalization can be greatly improved by the fused data. Dimensionality reductions techniques are implemented through feature selection and feature extraction, where feature selection aims at selecting features from the original features, while feature extraction focuses more on creating new features based on the original features. Broadly, DR techniques can be divided into *supervised* and *unsupervised* techniques, respectively. Common *supervised* techniques include filter techniques, wrapper techniques, and embedded techniques. *Unsupervised* techniques include data-driven-based techniques such as Principal Component Analysis (PCA) and domain knowledge-driven techniques. These dimensionality reduction methods can also be integrated with deep learning models to improve the performance of those models.

The advent of deep learning introduces new solutions to traditional computer vision tasks such as image classification and detection. Given the advantages such as high robustness and high performance with the help of Graphical Processing Unit (GPU), deep learning has been the main focus in some areas of computer science and can avoid trivial image preprocessing procedures. However, data reduction, as an important preprocessing step, can be integrated into machine learning models, especially for high-dimensional data analysis. According to the information of features, features can be divided into three classes, including suitable, unnecessary, and repeated. Therefore, what data

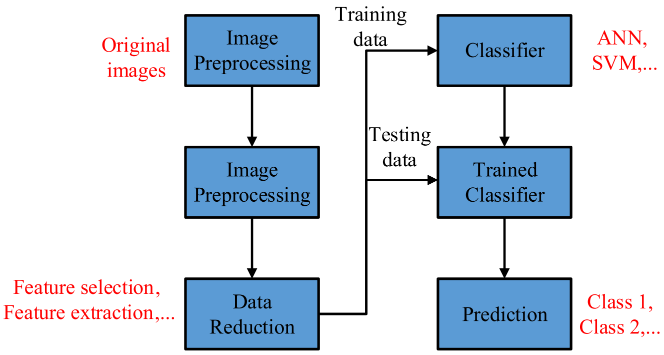


Fig. 38. Classification that introduced data reduction

reduction aims at is to refine the features by utilizing all the available information and contribute to the improvements of models' performance. Data reduction can be implemented in two ways through feature selection and feature extraction, respectively. In feature selection, only essential features are selected from the input data set. On the other hand, feature extraction creates new features from the original features. Both feature extraction and selection methods can be isolated or combined for the performance improvement of machine learning models. A classification example that introduces data reduction is shown in Fig. 38.

For feature selection, there are usually three key steps, including subset generation, subset evaluation, and termination, as shown in Fig. 39.

Subset generation aims at specifying a candidate subset for evaluation in each state. This process is determined by two key elements, including the search starting point and search strategy. To begin with the process, the search starting point, which indirectly determines the search direction, must be predefined. The search point could be an empty set where features are successively added to the set until the desired output is found. Inversely, the search point could also start from a full feature set where features can be successively removed from the set to produce the final feature output. Also, the search points could start with both ends and then add and remove features simultaneously until the desired output is generated. The second key element is the search strategy. Given a data set with N features, 2^N candidate subsets can be chosen from the data set. The search space makes it a challenging task to implement an exhaustive search even when N is moderate. Different search strategies, including *sequential search, random search,* and *complete search*, therefore, have been explored. *Sequential search* methods add or remove features once at a time to find the subset. However, completeness is therefore abandoned, and no optimal subsets can be guaranteed. To facilitate the searching process, p features can be added in one step while q features are removed in the next step ($p>q$) [363]. *The random search* starts with a randomly selected subset where the search can proceed in two different ways. One is to introduce randomness into the classical sequential approaches such as *simulated annealing* and *random-start-hill-climbing* [362]. The other is known as the *Las Vegas* algorithm that produces the next subset in a random manner. Nevertheless, randomness in these methods helps avoid local optima in the search space, although the optimality of the subset selection is resource-dependent.

Subset evaluation is a procedure to evaluate the newly generated feature subset by specific evaluation criteria, which can be broadly classified into two groups, independent criteria, and dependent criteria, regarding the dependency on the mining algorithms. Independent criteria are commonly used for the evaluation of feature subsets generated by the filter models. Popular independent criteria are *information measures, dependency measures, distance measures,* and consistency *measures* [364]. *Information measures* are used to measure the information gain from a feature. The definition of the information gain from a feature is the difference between the prior uncertainty and the expected posterior uncertainty. For two given features A and B, we prefer A if the information gain from A is greater than that from B. *Dependency* measures, also known as correlation measures, measure the capability of predicting the value of one variable from the value of another. These measures depict the association between a feature and the class. In a classification problem, feature A turns out to be more preferable if the association between A and class Z is higher than the association between feature B and class Z. *Distance measures* are also known as discrimination measures. For a two-class classification problem, if feature A produces a larger difference between two-class conditional probabilities than B, then A has a higher priority than B. *Consistency measures* aim at finding a minimized number of features that can separate classes consistently, just like the full set of features can. Inconsistency is to describe the phenomenon of two instances with the same feature values but having different class labels. In the wrapper models, which can be interpreted as a black box for feature selection by classification, predetermined mining algorithms are required for feature selection. Then the dependent criteria measure the performance of the mining algorithms applied on the selected subset and therefore determine which features to be selected. The drawback of these measures is that the computational cost is expensive as predetermined mining algorithms are introduced.

Stopping criteria determine when the feature selection process should stop. There are usually four popular stopping criteria. The first one is the completed search. It is quite straightforward that the search should stop when the search space has been completely explored. The second criterion is when some given bound is reached. Here the bound could be a specified number of features such as the minimum number of features or the maximum number of features. The third criterion is that feature selection should stop when the addition of any feature does not lead to a better subset. The last one is that an acceptable subset is selected in terms of the acceptable performance of a subsequent classifier.

Result validation could be directly implemented by using prior knowledge about the data. If the relevant features are known to us beforehand, we can then compare the known set of features with the selected features. The irrelevant and redundant features can also help remove unwanted features. However, we don't have such prior knowledge in practice and have to rely on some indirect methods instead. When considering a feature selection for a classification problem, the indirect method for validation of the selected features is to compare the performance of the models trained by the subset features and the full set features.

Feature extraction uses some transformation to map the original features to more significant features with possibly lower dimensionality,

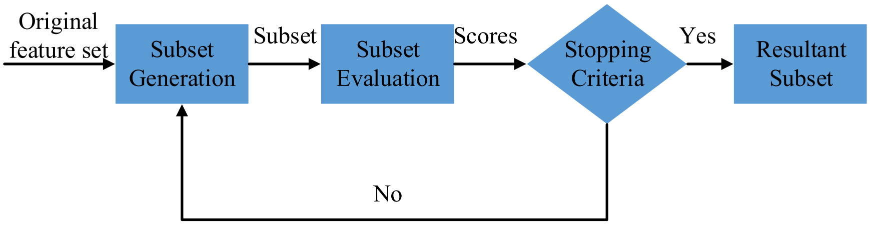


Fig. 39. Three key modules of feature selection

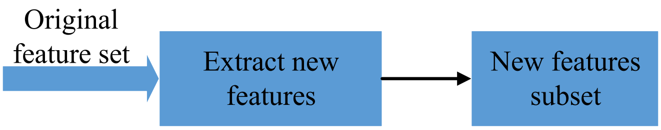


Fig. 40. Feature extraction

as is shown in Fig. 40. Finding a suitable representation of multivariate data is crucial for artificial neural networks and other classifiers [365]. Feature extraction can be used to reduce the complexity of data by representing each variable in feature space by linear combinations of original variables. Principal component analysis (PCA), a simple nonparametric method that extracts the most relevant information from redundant and noisy data, has been used as the most popular approach in feature extraction. Hence, many variants of PCA have also been proposed in the field. The choice of feature selection and feature extraction should be careful, but feature extraction shows advantages on computational cost [366]. In [367], the authors compared data reduction methods implemented in feature subset selection and feature extraction on the classification of two different types of datasets, including email data and drug discovery data. Information gain (IG) and wrapper methods were used to select features when implementing feature selection. However, it was found the wrapper shows better performance than IG in terms of classification accuracy. Compared to feature extraction methods, wrapper methods tend to produce the smallest features subsets while the classification accuracy is quite competitive to that of the feature extraction methods. Admittedly, the computational cost of wrapper methods is much more expensive compared to feature extraction methods. Also, some works integrate feature selection and feature extraction [368]. In the work of Ref. [369], features are firstly selected in the first level of dimensionality reduction based on mutual correlation. In the second level, PCA is used to extract features in the first level. Experiments on several standard datasets showed that the proposed method is more advantageous than single-level dimensionality reduction techniques.

Also, data reduction, which is referred to as feature reduction henceforth, can be implemented through supervised and unsupervised techniques depending on the learning patterns of these methods. In the following sections, we will introduce feature reduction techniques that are supervised and unsupervised in a sequence.

7.1. Supervised feature reduction techniques

Supervised feature reduction techniques require high-dimensional data input and output labels for the selection of relevant features while removing redundant features and noise. These techniques can be subdivided into three categories, including filter, wrapper, and embedded methods. There are three main differences between these three categories. Firstly, for filter techniques such as t-tests and Pearson correlation coefficient, simple statistical measures are used to measure the relevance of features when detecting group-level differences. Features are then ranked based on relevance. Secondly, an objective function from a machine learning model is used in wrapper techniques to rank features regarding their relevance to the model. Finally, embedded methods yield a small subset of relevant features by enforcing penalties on a machine learning model for feature selection. In the following sections, we will introduce filter techniques, wrapper techniques, and embedded techniques one by one.

7.1.1. Filter techniques

Pearson correlation coefficient (PCC) is one of the representative filter feature reduction techniques. PCC calculates the linear correlation between individual features and labels and ranks the features regarding the linear correlations [370]. If we assume a group classification problem with predictors variables X and class labels Y. Then X_i denotes the m dimensional vector of the ith variable for the training examples while y is the m dimensional vector containing all the target values. The Pearson correlation coefficient between predictor variables and the labels can be expressed as:

$$P_i = \frac{\sum_{k=1}^{m}\left(x_{k,i} - \bar{x}_i\right)(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}\left(x_{k,i} - \bar{x}_i\right)^2 \sum_{k=1}^{m}(y_k - \bar{y})^2}} \tag{45}$$

where the bar notion denotes the average over the index k. The higher values of the correlation coefficient P_i, the greater relevance of the feature in discriminating between the classes. Users have to manually predefine a threshold to select relevant features for the following machine learning analysis. Therefore, cross-validation procedures and a varied range of thresholds have to be carried out for the exploration of the optimal threshold that gives the best generalization of the method. The advantage is that PCC filters can be applied to situations when there are multi-group tasks but only linear dependencies between features and targets can be found, which becomes the major drawback of PCC especially when high-dimensional data with multivariate relationships must be considered. Numerous studies have used PCC filters for relevant feature selection. In the work [371], the authors calculated PCC between genes. Highly correlated genes that are considered to be dependent or coregulated form a cluster. The signal-to-noise ratio (SNR) method is then used to rank the correlated genes. Genes with the highest SNR are used as the representatives of each group. Besides, PCC filters have been widely used in gender classification and Alzheimer's disease (AD) classification [372,373].

T-test, as one of the typical statistical hypothesis testing techniques, has been widely used in feature reduction as well. Let $\bar{x}_1$ and $\bar{x}_2$ be the mean values of the two groups of the observed samples, s_1 and s_2 denote the corresponding standard deviations. Then the t-score of a feature is calculated as:

$$t = \frac{\left|\bar{x}_1 \; - \bar{x}_2 \right|}{\sqrt{\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2-2} \cdot \left[\frac{1}{N_1} + \frac{1}{N_2}\right]}} \tag{46}$$

where N_1 and N_2 are the numbers of subjects in each group. After calculation, a user-defined threshold of significance, e.g., p-value, that statistically shows whether or not the probability is greater in magnitude than t under the null hypothesis is introduced. Similar to PCC, the selection of the optimal threshold can be achieved by the cross-validation process [374]. Application of *t-test* in feature extraction allows fast computation and scalable to high-dimensional data. However, there are still several limitations of *t-test* based feature reduction techniques. One is that these methods are univariate in that no interactions between multiple features and spatial patterns are considered. Another is that t-tests are only to explore the difference between two groups, although this can be compensated by the equivalent analysis of variance (ANOVA) technique. Nevertheless, several studies have used t-test to select relevant features for machine learning [375,376]. An improved version of the t-test called ANOVA technique is usually used to select features in multiple groups. There is also extensive utilization of ANOVA technique in the field of feature reduction and selection [377–379]. Notably, ANOVAs provide the same benefit as t-tests, while the process of choosing the optimal threshold is the same. Another multivariate extension of ANOVA, which is named MANCOVA, has also been widely used in numerous feature selection tasks [379,380].

7.1.2. Wrapper techniques

Wrapper techniques select feature subset through the classification technique. The selected feature subset is evaluated by the objective function through search algorithms. Wrapper techniques can be classified into two categories, including Sequential Selection Algorithms

(SSA) and Heuristic Search Algorithms (HSA). SSA starts with a proper subset and includes or removes one feature at a time. The selection ends when the subset meets the output requirement according to predefined criteria. SSA is simple to implement and is fast to generate results as the size of the search space is usually low. Therefore, SSA gives up the compactness to find the best subsets [346]. HSA, which uses heuristic information to guide the search, can neither ensure the best subset to be found but usually finds an acceptable subset in a reasonable time [365, 381]. HSA can also be subdivided into two types: specific heuristics and general-purpose metaheuristics. The former is designed to solve a certain problem, while the latter aims at solving more general problems. According to the search direction, wrapper approaches can be further divided into the forward selection and backward elimination. In forward selection, the search starts with an empty feature set while features are added into the feature set step by step until the optimal subset with the optimal number of features found. By contrast, backward elimination starts with full features and iteratively removes a few features at each step until the optimal feature subset found. In this section, we will introduce recursive feature elimination (RFE) method, which is a popular backward elimination technique.

Given a two-class classification task, we have a set of features x_i and corresponding target labels y_i. And the training data is subdivided into two subsets, including 'Training' and 'Evaluation'. The observation weights a_i is then obtained from a machine learning algorithm. Feature relevance weights are then calculated through:

$$W = \sum_{x_i \in nzo} a_i y_i x_i \tag{47}$$

where *nzo* stands for *objects* with *non-zero* weights. The absolute values of the weights *W* are then ranked based on their importance, where the lowest-ranked features at a predefined percentage are removed. In the following step, the model is trained with features that have excluded the most irrelevant features, and the accuracy on the evaluation set is reported by the newly trained model. This process iterates until a stopping criterion is met or until the feature set is empty. Finally, the subset of features that results in the highest accuracy is chosen for the training of the final machine learning model while the rest of the features are discarded.

RFE requires two predefined parameters, which could be troublesome. One is the stopping criteria. When keeping removing low ranking features iteratively, it's likely the empty subset will be generated. However, when this happens, the iteration that gives the highest accuracy on the evaluation set is selected. Another scenario is that the performance of the model in the current iteration is not significantly better than that of the previous iteration. Then the procedure should be terminated as explored by De Martino et al. [382]. Another parameter that needs to be predefined is the percentage of the removal of the features in each step. There are studies used varied parameter such as 2%, 8%,10% [383–385]. However, the impact between the choice of the parameter and the overall performance of the model remains to be an open research question. Besides, the computational cost increases significantly if a very small percentage of features is removed at each iteration, while relevant features could be removed when the percentage is chosen to be a relatively large one.

RFE has two main benefits. The first is that RFE considers multivariate interactions between spatial patterns in the data. The second is REF might lead to better generalization ability as it uses a predictive model to remove redundant features. However, the drawback of RFE is also obvious that high computational cost is usually as it performs a completely heuristic search of the feature input space [386]. Nevertheless, there are still popular usage of RFE in different areas including ASD [387–389], AD [390], psychosis [391], schizophrenia [392], MDD [331], MCI [387], mood disorder [393].

7.1.3. Embedded techniques

The least absolute shrinkage and selection operator (LASSO) [348, 351], the Elastic Net [349], and the partial least square (PLS) method [350] are the three most popular embedded methods. In LASSO techniques and Elastic Net, both machine learning and feature reduction procedures are integrated into a regularization framework that produces a selected subset. However, PLS selects features by analyzing associations between the variables, either independent or dependent. We next describe these feature reduction methods.

7.1.3.7. LASSO. Assume that we have a binary classification task with a set of features x_i^j and corresponding target labels y_i, where $i = 1,2, \ldots, N$ and $j = 1,2,\ldots, M$. N and M stand for the number of observations and the dimensionality of features, respectively. Furthermore, each feature is assumed to be normalized by subtracting its mean and dividing by its standard deviations. Then, the coefficients $\hat{\gamma}$ are computed by minimizing the function [351]:

$$\sum_{i=1}^{N} \left(y_i - \sum_j x_i^j \gamma^j \right)^2 + \alpha \sum_{j=1}^{M} |\gamma^j| \tag{48}$$

where α is a predefined parameter that controls the balance between sparsity and high predictive accuracy. When α approaches 1, the model becomes sparser, which means few relevant features. On the other hand, the model is less sparse when α approaches 0, which means more relevant features. The selection of the most optimal α involves cross-validation procedures that test a range of α. Then the one that contributes to the highest model accuracy is selected. To solve the LASSO function, usual optimization procedures such as the coordinate descent algorithm can be used.

There are two main benefits of this method in the feature reduction process. One is that the majority features are discarded as the majority of the coefficient $\hat{\gamma}$ are set to zero. The second one is that LASSO can handle the situation where the number of observations is fewer than the number of predictor variables. There are numerous successful applications of feature selection using LASSO including AD classification [352–354], gender classification [394], autism spectrum disorder (ASD) classification [355] and so on [356,395].

7.1.3.8. Elastic Net. Elastic net is quite similar to LASSO but with an additional quadratic term [349]. If we consider the two-class classification task like the one in LASSO, then Elastic net computes model coefficients $\hat{\gamma}$ by minimizing the objective function [357,358]:

$$\sum_{i=1}^{N} \left(y_i - \sum_j x_i^j \gamma^j \right)^2 + \alpha_1 \sum_{j=1}^{M} |\gamma^j| + \alpha_2 \sum_{j=1}^{M} |\gamma^j|^2 \tag{49}$$

where α_1 and α_2 are two user-defined parameters that control the degree of penalty. The penalty $\sum_{j=1}^{M} |\gamma^j|^2$ leverage the sparsity by resulting in few features with non-zero weights. These two parameters are usually selected via an objective parameter grid-search process which determines the best parameters from a range of parameters in the two dimensions domain. However, grid-search can be computationally expensive. Previous applications of feature reduction implemented through the Elastic Net include AD classification [359,360], and treatment response prediction in ADHD [361].

7.1.3.9. Partial Least Squares. Partial least squares correction (PLSC, [362]) and partial least squares regression (PLSR, [363]) are two main categories of the partial least squares feature reduction method. Compared to PLSR, PLSC is usually more popular in the medical imaging area. Therefore, we will discuss PLSC in this section.

Let consider the previous two-class classification example with the normalized features x_i^j with the corresponding target y_i. Then PLSC starts with the computation of cross product of the features and the target vectors as follows:

$$P = Y^T X \tag{50}$$

The resulting matrix P is then decomposed by singular value decomposition (SVD) [362], which is:

$$P = USV^T \tag{51}$$

P, therefore, can be decomposed into two singular vectors (U and V) and a diagonal matrix S containing the 'singular values' in the diagonal. Weights in U identifies the variables in X that contribute the most in explaining the relationship between features and the targets. Finally, latent variables of X and Y are reconstructed based on the following two equations:

$$\begin{cases} P_x = XV \\ P_y = YU \end{cases} \tag{52}$$

P_x and P_y stand for the reduced latent variables of original features in X and the latent variables for the target variables, respectively. By doing so, the original features of high-dimensionality are now represented by low-dimensional latent variables. The applications of feature reduction using the partial least squares method include Age classification (young vs. old) [364], multimodal feature reduction tasks [365], and so on so forth [366].

7.2. Unsupervised feature reduction techniques

Unsupervised feature reduction techniques, also known as feature extraction techniques, extract relevant features through linear or nonlinear combinations of the original features. Principal component analysis (PCA) and independent component analysis (ICA) are the two most popular unsupervised dimensionality reduction techniques. We start this section with PCA and ICA and end with Coordinate-Based Mate-Analysis (CBMA) techniques, a technique that relies on existing 'domain knowledge' for feature reduction.

7.2.1. Principal component analysis

PCA linearly transforms the correlated variables into unrelated variables with reduced dimensionality [367]. In essence, these principal components are the linear combinations of the original features while keeping most of the variance in the features. The first step to construct principal components from high-dimensional features is to normalize the original features by subtracting the sample mean, and the resultant features are then divided by the standard deviation. Secondly, eigendecomposition is performed based on the covariance matrix, which is calculated from the standardized features. The eigenvalues are sorted in a decreasing order that indicates the decreasing variance of the features. By multiplying the original normalized features with the most significant eigenvectors, the features are then mapped into a lower-dimensional space. The number of eigenvectors is predefined by the user to meet certain requirements.

PCA has been extensively used in reducing relevant features in medical data classification tasks such as schizophrenia [368,369], AD [370], face recognition [371], and psychosis [372]. Notably, there are also regression studies that involved PCA such as age prediction [373] and AD clinical scores prediction[374].

PCA contributes two major benefits to dimensionality reduction in medical data analysis. The first one is the easy implementation and computational efficiency. The second is that this technique is unsupervised so that the categorical labels or annotations are not required for the extraction of relevant features. However, PCA also has some shortcomings. First, users are required to predefine the number of principal components, which leads to repetitive experiments before the best number can be found. Though there are some attempts at simplifying the procedures [375], it remains a big challenge of PCA. Second, the interpretability is poor since principal components are linear combinations of the original features. Lastly, classical PCA may not adequately explore more complex nonlinear feature interactions as principal components are built through a linear transformation [376]. Having said this, various nonlinear generalizations of the PCA have now been proposed to alleviate some of these issues [377].

7.2.2. Independent Component Analysis

ICA, a multivariate data-driven technique, falls into the category of *blind-source separation* methods, which separate features into underlying *independent* information components. ICA separates the mixed signals or features into independent and relevant features. ICA assumes that source signals are independent and unknown but linearly mixed [378].

Let the feature matrix be $X \in \mathbb{R}^{m \times n}$, where m and n stand for the number of observations and number of attributes (dimensionality), respectively. The source matrix is denoted by $S \in \mathbb{R}^{m' \times n}$, where m' is the expected number of independent components. Another matrix $A \in \mathbb{R}^{m \times m'}$ is defined as the mixing matrix whose columns contain the associated n components. Based on the above variables, X can be expressed as [379]:

$$X = AS \tag{53}$$

Additionally, we can have:

$$Y = WX \tag{54}$$

Therefore, ICA focuses on estimating the unmixing matrix $W \in \mathbb{R}^{n \times m}$, which renders Y to be a good approximation of the true signal sources S.

In fMRI, most of the ICA dimensionality reduction studies extracted relevant independent components from the spatial dimension. But there is another category of ICA that subdivides the methods into *individual-level* ICA and *group-level* ICA. Briefly, each subject's features are input into individual ICA analysis while sets of components for the groups are estimated and reconstructed to obtain individual-subject independent components in *group-level* ICA.

The advantages of ICA mainly come from two aspects. Firstly, unlike univariate methods, no regressors of interest need to be specified in ICA as the specification of regressor may require prior knowledge and assumptions. The second advantage of ICA mainly comes from brain signals processing that has been proved to be successful in disentangling the brain signals such as separating motion, scanner related, and physiological components [379]. However, ICA also has some drawbacks. One is the expensive computational cost of ICA algorithms [380]. Another is that ICA remains to be improved as ICA algorithms may not be able to adequately separate default mode networks and respiration signals in fMRI.

Nevertheless, there are numerous ICA studies in the medical image feature reduction field [346,365,381–385]. Note that there is a significant difference between PCA and ICA. In PCA, features that could be correlated are mapped into sets of uncorrelated features. In ICA, however, original features are statistically transformed into a set of independent features. The common between PCA and ICA is that both of these two techniques are unsupervised and require no labeled data.

7.2.3. Coordinate-based meta-analysis

CBMA techniques are different from the other unsupervised methods in that CBMA techniques rely on existing domain knowledge for feature reduction while the other unsupervised methods are mainly data-driven. Meta-analysis techniques have been involved in the studies of modeling, analyzing, and reporting brain activations [386]. Representative meta-analysis techniques include multi-level kernel density estimation [387], kernel-density estimation [388], and activation likelihood estimation (ALE) [389].

CBMA has been widely used in feature reduction in medical imaging. For example, a CBMA technique is applied in [390] to select features for the classification of working memory, emotion, and pain using fMRI.

Table 10
Summary of feature reduction method in recent research

Reference	Task	Mode	Technique	FET
L. Goh et al. [326]	Classification of gene expression data	Supervised	Filter (PCC based)	PCC between genes calculated for the classification task.
Z. Dai et al. [327]	Analysis of early Alzheimer's disease	Supervised	Filter (PCC based)	PCC was calculated for the measurement of the functional connectivity among regions.
Y. Fan et al. [328]	Gender classification	Supervised	Filter (PCC based)	PCC is used to measure the relevance of each feature to the classification
B. Mwangi et al. [329]	Diagnostic classification of depressive disorder	Supervised	Filter (T-test based)	The optimal threshold is obtained by a cross-validation process for T-test
R. Chaves et al. [330]	Diagnosis of Alzheimer's disease	Supervised	Filter (T-test based)	T-test feature selection for classification by SVM
C. Chu et al. [331]	Effectiveness of feature selection	Supervised	Filter (T-test based)	Common feature selection methods are compared
S. G. Costafreda et al. [332]	Exploration of diagnostic specificity	Supervised	Filter (ANOVA)	ANOVA for modeling of diagnostic group effect
S. G. Costafreda et al. [396]	Analysis of the structural neuroanatomy of depression	Supervised	Filter (ANOVA)	ANOVA for selection of areas of maximum group differences between observations and
J. H. Yoon et al. [333]	Deficits in distributed representations in schizophrenia	Supervised	Filter (ANOVA)	Voxelwise ANOVA applied in the study
E. A. Allen et al. [397]	Multivariate comparison of resting-state networks	Supervised	Filter (MANCOVA)	Applied MANCOVA for interpretability of variability in the multivariate response
S. Calderoni et al. [341]	ASD analysis	Supervised	Wrapper (RFE)	RFE and SVM are combined to identify the most discriminating voxels in gray matter segments (SVM-RFE).
C. Ecker et al. [342]	Investigation of the predictive value of whole-brain structural MR scans in autism	Supervised	Wrapper (RFE)	SVM-RFE for detection of subtle differences in brain networks between ASD patients and healthy subjects.
M. Ingalhalikar et al. [343]	Constructing abnormality markers of pathology based on diffusion	Supervised	Wrapper (RFE)	Features are ranked and then selected.
C. Davatzikos et al. [344]	Detection of prodromal Alzheimer's disease via pattern classification	Supervised	Wrapper (RFE)	RFE is used to find the minimal set of features to be fed into the classifier.
D. Gothelf et al. [345]	Developmental changes in multivariate neuroanatomical patterns	Supervised	Wrapper (RFE)	30% of worst-discriminating voxels are removed at a time until the performance started deteriorating
E. Castro et al. [346]	Characterization of groups using composite kernels and multi-source fMRI analysis data	Supervised	Wrapper (RFE)	The RFE algorithm is based on the calculation of discriminative weights
K. Nho et al. [347]	Automatic prediction of conversion from mild cognitive impairment to probable AD	Supervised	Wrapper (RFE)	SVM-RFE algorithm, which returns a ranking of all the features and then selects features accordingly.
J. Mourão-Miranda [398]	Risk assessment of mood disorders from low-risk adolescent	Supervised	Wrapper (RFE)	RFE is used to determine the optimal subset of brain voxels that results in the best discrimination accuracy. Also, RFE helps to accurately localize the most discriminative brain voxels.
J. Yan et al. [352]	Multimodal neuroimaging predictors based on structured sparse learning	Supervised	Embedded (LASSO)	Modeled the interrelated structure within the predictor variables by incorporating LASSO
M. Vounou et al. [353]	Sparse reduced-rank regression detects genetic associations	Supervised	Embedded (LASSO)	Proposed the application of a penalized multivariate model, sparse reduced-rank regression (sRRR).
O. Kohannim et al. [354]	Discovery and replication of gene influences on brain structure	Supervised	Embedded (LASSO)	The gene effects in genome-wide association studies (GWAS) of brain images are evaluated by LASSO.
R. Casanova et al. [394]	Gender classification	Supervised	Embedded (LASSO)	Random Forest and LASSO are combined for classification.
E. Duchesnay et al. [355]	ASD classification	Supervised	Embedded (LASSO)	Feature selection is used to predict the clinical status of a highly imbalanced dataset.
I. Rish et al. [356]	Predicting temporal lobe volume	Supervised	Embedded (LASSO)	The proposed feature selection method helped to predict a tensor-based morphometry-derived measure of temporal lobe volume.
A. Rao et al. [359]	Classification of AD	Supervised	Embedded (Elastic Net)	A sparsity penalty is introduced into the log-likelihood and served feature selection algorithm.
J. Wan et al [360]	Hippocampal Surface Mapping	Supervised	Embedded (Elastic Net)	The association between single nucleotide polymorphisms (SNPs) and quantitative traits (QTs) is examined by Elastic Net.
A. F. Marquand et al. [361]	Treatment response prediction in ADHD	Supervised	Embedded (Elastic Net)	Sparse multinomial logistic regression (SMLR) with an elastic net penalty is proposed.
K. Chen et al. [364]	Age classification	Supervised	Embedded (PLSC)	A partial least square (PLS) algorithm is used to form a covariance-maximized combined latent variable.
J. Sui, T. et al. [365]	Analysis of multimodal feature reduction tasks	Supervised	Embedded (PLSC)	Numerous multivariate methods have been reviewed and analysed.
L. Menzies et al. [366]	Analysis of obsessive-compulsive disorder	Supervised	Embedded (PLSC)	PLSC is used to measure the correlation between the grey matter systems and stop-signal reaction time (SSRT).
P. Alvarado-Alanis et al. [368]	Abnormal white matter integrity in psychosis	Unsupervised	PCA	The white manner tracts are grouped into four factors by PCA.
P.-R. Loh et al. [369]	Fast variance-components analysis of schizophrenia	Unsupervised	PCA	Features are obtained by PCA for bivariate analyses.
L. Khedher et al. [370]	Early diagnosis of AD	Unsupervised	PCA	Multivariate approaches for feature selection including PCA
L. C. Paul et al. [371]	Face recognition	Unsupervised	PCA	PCA method performed worse than PLS feature extraction and linear SVM classifier.
A. B. Bendixen et al. [372]	Psychosis	Unsupervised	PCA	PCA is conducted on Geriatric Anxiety Inventory (GAI) for disorders differentiation.
K. Franke et al. [373]	Age prediction	Unsupervised	PCA	Training a relevance vector machine based on PCA-reduced features.

(continued on next page)

Table 10 (*continued*)

Reference	Task	Mode	Technique	FET
Y. Wang et al. [374]	AD clinical scores prediction	Unsupervised	PCA	The Relevance Vector Machine (RVM) is built for regression based on PCA-reduced features.
P. K. Douglas et al. [381]	fMRI decoding	Unsupervised	ICA	Six different machine learning algorithms are evaluated on the ICA-reduced features.
J. R. Sato et al. [382]	ADHD prediction	Unsupervised	ICA	Evaluation of three different feature extraction methods while the classifiers showed almost the same performance.
E. P. Duff et al. [383]	Prediction using fMRI	Unsupervised	ICA	A task-specific Independent Component Analysis (ICA) procedure is proposed.
A. Hyvarinen et al. [384]	Feature extraction	Unsupervised	ICA	Novel time-contrastive learning model combined with linear ICA.
C. Zhao et al. [385]	Anomaly detection in hyperspectral imagery	Unsupervised	ICA	Improved ICA for feature extraction.
T. Yarkoni et al. [390]	Synthesis of human functional neuroimaging data	Unsupervised	CBMA	No heavy reliance on the automatically extracted information.
T. M. Mitchell et al. [391]	A tool for the automated synthesis of fMRI data	Unsupervised	CBMA	A CBMA technique in multicenter studies with a good generalization performance.

Another CBMA feature reduction framework is known as *Neurosynth,* which is a tool for the automated synthesis of fMRI data [391]. To classify AD subjects, Dukart et al. applied a CBMA technique in multicenter studies with a good generalization performance reported [392]. In the work [331], Chu et al. reported that ROIs selected via a prior domain knowledge lead to better generalization ability compared to features selected through data-driven approaches such as RFE and t-test. The benefit of CBMA techniques is a posteriori certainty can be improved and makes neuroimaging studies less sensitive to type II errors [387]. But CBMA techniques may suffer from information loss as well because features are represented with a high degree of sparseness [393].

7.2.4. *Summary and Discussion*

This section introduced popular data reduction techniques that can be divided into two categories, namely supervised and unsupervised. In supervised techniques, we further introduced three subsets of methods, including filter, wrapper, and embedded. Filter techniques discard redundant features according to statistical feature ranking, as shown in Table 10. There are two common drawbacks of these methods. One is that no interactions between multiple features are considered as they are not multivariate. The second is the difficulty in predefining a proper feature threshold value. By contrast, wrapper techniques are multivariate but computationally expensive. The performance of embedded feature reduction methods highly relies on penalization parameters that are generally chosen by cross-validation. The major difference between supervised methods and unsupervised methods is obviously the information using that supervised methods select relevant features in aim at group-level differentiation while unsupervised methods consider features independent of the final interest.

The performance of feature reduction techniques, as mentioned before, is determined by several factors. One is the annoying optimal threshold values, either for the number of features to be chosen or the number of parameters to be determined in the process. The second one is the randomness in the process of training, and testing models as the reduced features may vary from fold to fold.

In summary, feature reduction techniques have been widely used in the medical imaging area to improve predictive accuracy in spite of curse-of-dimensionality or small sample problems. While numerous studies compared different feature reduction methods, no method emerged as optimal in all medical imaging machine learning tasks.

8. Conclusion

Advances in sensor technologies have made it possible to leverage modern machine learning and AI methods, with the aim of harnessing a multitude of data sources for biomedical information analysis. The diverse nature of such data makes it unrealistic to ignore the interdependencies between the different data sources. It is commonly known that integrating a multitude of data from different imaging modalities can produce more consistent, accurate, robust to equipment and measurement induced noise and functional information than that generated by a single data source. However, the fusion of multi-sourced data may bring various challenges, such as higher complexity in denoising the data, missing data values, data scarcity, larger costs in sensor hardware and data processing, and high dimensionality. This paper has reviewed these challenges and discussed state-of-the-art methodologies to effectively tackle them.

Although AI broadens the already existing large spectrum of sensor fusion methodologies, a number of research frontiers and caveats still persist. Sensor fusion methods, especially when incorporating AI, have lower interpretability than classical approaches and may suffer from generalization problems when the data is scant or not fully representative of the problem at hand – thus, human intervention and monitoring are still necessary. This is even more true for the case of biomedical applications, where the cost of algorithmic errors can be prohibitive. Hence, research efforts must be focused on increasing the interpretability of multi-source data pipelines processing biomedical data and on strengthening the level of integration with medical personnel. Although the level of automation in biomedical decision-making is expected to massively increase in the short term, little research is being directed towards establishing how these intelligent systems will interact with human experts. A new research frontier is that of establishing the efficacy of these algorithms when they act in a symbiotic manner with the medical personnel. It is conceivable that the algorithms that yield maximum performance in autonomous decision-making tend to induce human error in actual biomedical operations.

Another exciting research frontier is that of finding new ways to tackle model drift effects such as data and concept drift. For example, is it possible to make these intelligent systems adaptive to situations such as data drift produced by wear in the data logging equipment, concept drift resulting from environmental factors – which may make some conditions more likely than others, essentially changing the baseline priors – while also adaptive to the different biases of the medical personnel involved? The ultimate goal is to make these systems less expensive to maintain as to reduce the cost of the medical treatments.

Despite these constraints, this is an ever-expanding subject that shows great promise beyond the already existing numerous applications, and many of the surveyed techniques are already applicable if care is taken according to the above considerations.

Table 11
Abbreviations

Abbreviation	Representation
AADF	Adaptive anisotropic diffusion filtering
AD	Alzheimer's disease
ADHD	Attention deficit/hyperactivity disorder
AFM	Atomic force microscopy
AIP	Artificial identification points
ALE	Activation likelihood estimation
ANOVA	Analysis of variance
ANTs	Advanced normalization tools
ASD	Autism spectrum disorder
CBMA	Coordinate-based mate-analysis
CCD	Charged-coupled device
CLAD	Chambolle–lions anisotropic diffusion
CT	Computed tomography
DR	Dimensionality reduction
DTI	Diffusion tensor image
DWT	Discrete wavelet transform
EEG	Electroencephalogram
EM	Expectation maximization
ET	Electron tomography
FEM	Finite element method
FSL	FMRIB software library
GA	Genetic algorithms
GAN	Generative adversarial network
GD	Gradient descent
GIP	General iterative principal
GPU	Graphical processing unit
HSA	Heuristic search algorithms
ICA	Independent component analysis
ICP	Iterative closest point
IG	Information gain
ILSVRC	ImageNet large-scale visual recognition challenge
KDD	Knowledge Discovery from Databases
LASSO	Least Absolute Shrinkage and Selection Operator
MANCOVA	Multivariate analysis of covariance
MAR	Missing at random
MCAR	Missing completely at random
MCI	Mild cognitive impairment
MDD	Major depressive disorder
MDITs	Missing data imputation techniques
MEG	Magnetoencephalography
MI	Multiple imputation
MNAR	Missing not at random
MPI	Magnetic particle imaging
MR	Magnetic resonance
MRF	Markov random field
MRI	Magnetic resonance imaging
MSE	Mean squared error
NI	Non-ignorable
NLM	Non-local means
OASIS	Open access series of imaging studies
OMT	Optical microscopy and tomograph
OSRAD	Oriented speckle reducing anisotropic diffusion
PAT	Photoacoustic tomography
PCA	Principal component analysis
PCC	Pearson correlation coefficient
PDF	Probability density function
PET	Positron emission tomography
PLS	Partial least square
PLSC	Partial least squares correction
PLSR	Partial least squares regression
PMMs	Pattern mixture models
pMRI	Parallel MRI
RF	Radio frequency
RFE	Recursive feature elimination
RML	Raw maximum likelihood
ROAD	Rank-ordered absolute difference
ROIs	Regions of interest
RST	Rough set theory
SAD	Sum of absolute differences
SMs	Selection models
SNR	Signal-to-noise ratio
SPECT	Single photon emission computed tomography
SPMs	Shared parameter models
SRBF	Speckle reducing bilateral filter
SRPI	Similar response pattern imputation
SSA	Sequential selection algorithms
SSD	Sum of squared differences
WSVODP	Weight sum variance of digital number probability

Declaration of Competing Interest

There is no conflict of interest.

Acknowledgment

This study was supported by the Royal Society International Exchanges Cost Share Award, UK (RP202G0230); Medical Research Council Confidence in Concept Award, UK (MC_PC_17171); Hope Foundation for Cancer Research, UK (RM60G0680); Sino-UK Industrial Fund, UK (RP202G0289); Global Challenges Research Fund (GCRF), UK (P202PF11) and the United States National Science Foundation (1946391).

References

[1] R. Acharya, R. Wasserman, J. Stevens, C. Hinojosa, Biomedical imaging modalities: a tutorial, Comput. Med. Imaging Graph. 19 (1995) 3–25.

[2] J. Yao, L.V. Wang, Photoacoustic tomography: fundamentals, advances and prospects, Contr. Media Mol. Imaging 6 (2011) 332–345.

[3] C.J. Garvey, R. Hanlon, Computed tomography in clinical practice, BMJ 324 (2002) 1077–1080.

[4] R. Smith-Bindman, J. Lipson, R. Marcus, K.-P. Kim, M. Mahesh, R. Gould, Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer, Arch. Intern. Med. 169 (2009) 2078–2086.

[5] A. Villalobos, B. Cheng, W. Wagstaff, I. Sethi, Z. Bercu, D.M. Schuster, Tumor-to-normal ratio relationship between planning Technetium-99 Macroaggregated Albumin and Posttherapy Yttrium-90 Bremsstrahlung SPECT/CT, J. Vasc. Interv. Radiol. (2021).

[6] A. Ammar, O. Bouattane, M. Youssfi, Automatic cardiac cine MRI segmentation and heart disease classification, Comput. Med. Imaging Graph. 88 (2021), 101864.

[7] V.P. Sudarshan, G.F. Egan, Z. Chen, S.P. Awate, Joint PET-MRI image reconstruction using a patch-based joint-dictionary prior, Med. Image Anal. 62 (2020), 101669.

[8] Z. Liu, Y. Song, V.S. Sheng, C. Xu, C. Maere, K. Xue, MRI and PET image fusion using the nonparametric density model and the theory of variable-weight, Comput. Methods Programs Biomed. 175 (2019) 73–82.

[9] L.C. Wu, Y. Zhang, G. Steinberg, H. Qu, S. Huang, M. Cheng, A review of magnetic particle imaging and perspectives on neuroimaging, AJNR. Am. J. Neuroradiol. 40 (2019) 206–212.

[10] S.T. Brinker, C. Crake, J.R. Ives, E.J. Bubrick, N.J. McDannold, Scalp sensor for simultaneous acoustic emission detection and electroencephalography during transcranial ultrasound, Phys. Med. Biol. 63 (2018), 155017–155017.

[11] S.P. Singh, Magnetoencephalography: basic principles, Ann. Indian Acad. Neurol. 17 (2014) S107–S112.

[12] S. Xia, Y. Huang, S. Peng, Y. Wu, X. Tan, Adaptive anisotropic diffusion for noise reduction of phase images in Fourier domain Doppler optical coherence tomography, Biomed. Optics Exp. 7 (2016) 2912–2926.

[13] D. Alsteens, H.E. Gaub, R. Newton, M. Pfreundschuh, C. Gerber, D.J. Müller, Atomic force microscopy-based characterization and design of biointerfaces, Nat. Rev. Mater. 2 (2017) 17008.

[14] L. Xu, Z. Du, R. Mao, F. Zhang, R. Liu, GSAM: A deep neural network model for extracting computational representations of Chinese addresses fused with geospatial feature, Comput., Environ. Urban Syst. 81 (2020), 101473.

[15] Y. Zhang, Q. Li, W. Tu, K. Mai, Y. Yao, Y. Chen, Functional urban land use recognition integrating multi-source geospatial data and cross-correlations, Comput., Environm. Urban Syst. 78 (2019), 101374.

[16] R. Shafran-Nathan, Y. Etzion, D.M. Broday, Fusion of land use regression modeling output and wireless distributed sensor network measurements into a high spatiotemporally-resolved NO2 product, Environ. Pollut. 271 (2021), 116334.

[17] X. Liu, R. Zhu, A. Anjum, J. Wang, H. Zhang, M. Ma, Intelligent data fusion algorithm based on hybrid delay-aware adaptive clustering in wireless sensor networks, Future Gener. Comput. Syst. 104 (2020) 1–14.

[18] K. Xiao, R. Wang, H. Deng, L. Zhang, C. Yang, Energy-aware scheduling for information fusion in wireless sensor network surveillance, Inform. Fus. 48 (2019) 95–106.

[19] Y. Wang, Y. Yang, H. Sun, J. Dai, M. Zhao, C. Teng, Application of a data fusion strategy combined with multivariate statistical analysis for quantification of puerarin in Radix puerariae, Vib. Spectrosc. 108 (2020), 103057.

[20] S. Huang, S.L. Sing, G. de Looze, R. Wilson, W.Y. Yeong, Laser powder bed fusion of titanium-tantalum alloys: Compositions and designs for biomedical applications, J. Mech. Behav. Biomed. Mater. 108 (2020), 103775.

[21] J. Du, M. Fang, Y. Yu, G. Lu, An adaptive two-scale biomedical image fusion method with statistical comparisons, Comput. Methods Programs Biomed. 196 (2020), 105603.

[22] S. Salcedo-Sanz, P. Ghamisi, M. Piles, M. Werner, L. Cuadra, A. Moreno-Martínez, Machine learning information fusion in Earth observation: a comprehensive review of methods, applications and data sources, Inform. Fus. 63 (2020) 256–272.

[23] P. Zhang, T. Li, G. Wang, C. Luo, H. Chen, J. Zhang, Multi-source information fusion based on rough set theory: a review, Inform. Fus. 68 (2021) 85–117.

[24] B. He, X. Cao, Y. Hua, Data fusion-based sustainable digital twin system of intelligent detection robotics, J. Clean. Prod. 280 (2021), 124181.

[25] L. Bonomi, L. Fan, X. Jiang, Noise-tolerant similarity search in temporal medical data, J. Biomed. Inform. 113 (2021), 103667.

[26] F. Chen, Z. Cao, E.M. Grais, F. Zhao, Contributions and limitations of using machine learning to predict noise-induced hearing loss, Int. Arch. Occup. Environ. Health 94 (2021) 1097–1111.

[27] A. Purwar, S.K. Singh, Hybrid prediction model with missing value imputation for medical data, Expert Syst. Appl. 42 (2015) 5621–5631.

[28] S. Suthaharan, E.A. Rossi, V. Snyder, J. Chhablani, R. Lejoyeux, J.-A. Sahel, Laplacian feature detection and feature alignment for multimodal ophthalmic image registration using phase correlation and Hessian affine feature space, Signal Process. 177 (2020), 107733.

[29] R.R. Sood, W. Shao, C. Kunder, N.C. Teslovich, J.B. Wang, S.J.C. Soerensen, 3D Registration of pre-surgical prostate MRI and histopathology images via super-resolution volume reconstruction, Med. Image Anal. 69 (2021), 101957.

[30] W. Wimmer, L. Anschuetz, S. Weder, F. Wagner, H. Delingette, M. Caversaccio, Human bony labyrinth dataset: Co-registered CT and micro-CT images, surface models and anatomical landmarks, Data in Brief 27 (2019), 104782.

[31] I. Izonin, R. Tkachenko, M. Gregus ml, K. Zub, P. Tkachenko, A GRNN-based approach towards prediction from small datasets in medical application, Procedia Comput. Sci. 184 (2021) 242–249.

[32] T. Shaikhina, N.A. Khovanova, Handling limited datasets with neural networks in medical applications: a small-data approach, Artif. Intell. Med. 75 (2017) 51–63.

[33] V.J. Kadam, S.M. Jadhav, Performance analysis of hyperparameter optimization methods for ensemble learning with small and medium sized medical datasets, J. Discrete Math. Sci. Cryptogr. 23 (2020) 115–123.

[34] S. Faisal, G. Tutz, Imputation methods for high-dimensional mixed-type datasets by nearest neighbors, Comput. Biol. Med. (2021), 104577.

[35] A.N. Gorban, I.Y. Tyukin, D.V. Prokhorov, K.I. Sofeikov, Approximation with random bases: Pro et Contra, Inform. Sci. 364-365 (2016) 129–145.

[36] E.M. Mirkes, J. Allohibi, A. Gorban, Fractional norms and quasinorms do not help to overcome the curse of dimensionality, Entropy 22 (2020) 1105.

[37] C. Wang, G. Yang, G. Papanastasiou, S.A. Tsaftaris, D.E. Newby, C. Gray, DiCyc: GAN-based deformation invariant cross-domain information fusion for medical image synthesis, Inform. Fus. 67 (2021) 147–160.

[38] S. Singh, R.S. Anand, Ripplet domain fusion approach for CT and MR medical image information, Biomed. Signal Process. Control 46 (2018) 281–292.

[39] W. Kong, Y. Chen, Y. Lei, Medical image fusion using guided filter random walks and spatial frequency in framelet domain, Signal Process. 181 (2021), 107921.

[40] A.S. Lundervold, A. Lundervold, An overview of deep learning in medical imaging focusing on MRI, Zeitschrift für Medizinische Physik 29 (2019) 102–127.

[41] B. Kanrar, K. Sanyal, S. Dhara, Quantification and distribution of trace elements in fusion bead and pressed pellet specimens using a table top micro-X-ray fluorescence spectrometer, Spectrochim. Acta Part B 177 (2021), 106063.

[42] C. Assis, E.M. Gama, C.C. Nascentes, L.S. de Oliveira, M.J. Anzanello, M.M. Sena, A data fusion model merging information from near infrared spectroscopy and X-ray fluorescence. Searching for atomic-molecular correlations to predict and characterize the composition of coffee blends, Food Chem. 325 (2020), 126953.

[43] G. Muhammad, M. Shamim Hossain, COVID-19 and Non-COVID-19 Classification using Multi-layers Fusion From Lung Ultrasound Images, Inform. Fus. 72 (2021) 80–88.

[44] M. Singh, S. Singh, S. Gupta, An information fusion based method for liver classification using texture analysis of ultrasound images, Inform. Fus. 19 (2014) 91–96.

[45] M. Klingebiel, C. Arsov, T. Ullrich, M. Quentin, R. Al-Monajjed, D. Mally, Reasons for missing clinically significant prostate cancer by targeted magnetic resonance imaging/ultrasound fusion-guided biopsy, Eur. J. Radiol. 137 (2021), 109587.

[46] Y. Liu, C. Zhang, C. Li, J. Cheng, Y. Zhang, H. Xu, A practical PET/CT data visualization method with dual-threshold PET colorization and image fusion, Comput. Biol. Med. 126 (2020), 104050.

[47] C. Sun, X. Liu, C. Bao, F. Wei, Y. Gong, Y. Li, Advanced non-invasive MRI of neuroplasticity in ischemic stroke: techniques and applications, Life Sci. 261 (2020), 118365.

[48] S.H. Choi, S.Y. Kim, Y.-S. Lim, Selection of MRI contrast agent and diagnostic criteria for HCC to maximize the advantages of contrast agents, J. Hepatol. 73 (2020) 714–715.

[49] O. Tanaka, Y. Nishigaki, H. Hayashi, T. Iida, T. Yokoyama, E. Takenaka, The advantage of iron-containing fiducial markers placed with a thin needle for radiotherapy of liver cancer in terms of visualization on MRI: an initial experience of Gold Anchor, Radiol. Case Rep. 12 (2017) 416–421.

[50] A.R. El-Najjar, A.M. Abu-Elsoaud, H.T. Mohammed, K.M. Shawky, Diagnostic potential of magnetic resonance imaging (MRI) of the first carpometacarpal joint in hand osteoarthritis, Egypt. Rheumatol. 43 (2021) 59–64.

[51] G. Zeng, F. Schmaranzer, C. Degonda, N. Gerber, K. Gerber, M. Tannast, MRI-based 3D models of the hip joint enables radiation-free computer-assisted planning of periacetabular osteotomy for treatment of hip dysplasia using deep learning for automatic segmentation, Eur. J. Radiol. Open 8 (2021), 100303.

[52] H. Shahamat, M. Saniee Abadeh, Brain MRI analysis using a deep learning based evolutionary approach, Neural Netw. 126 (2020) 218–234.

[53] B.S. Cherian, A.K. Bhat, K.V. Rajagopal, S.B. Maddukuri, D. Paul, N.J. Mathai, Comparison of MRI & direct MR arthrography with arthroscopy in diagnosing ligament injuries of wrist, J. Orthop. 19 (2020) 203–207.

[54] A. Teramoto, Y. Akatsuka, H. Takashima, H. Shoji, Y. Sakakibara, K. Watanabe, 3D MRI evaluation of morphological characteristics of lateral ankle ligaments in injured patients and uninjured controls, J. Orthop. Sci. 25 (2020) 183–187.

[55] R. Heiss, A. Guermazi, M. Jarraya, L. Engebretsen, T. Hotfiel, P. Parva, Prevalence of MRI-Detected Ankle Injuries in Athletes in the Rio de Janeiro 2016 Summer Olympics, Acad. Radiol. 26 (2019) 1605–1617.

[56] L.W. Turnbull, P. Ballard, R. Tetlow, S.J. Bowsley, D.J. Manton, S.J. Burton, Early changes induced by tamoxifen on the endometrium of postmenopausal women with breat cancer: preliminary TVS, MRI and pathological findings, Eur. J. Ultrasound 6 (1997) S14–S15.

[57] J.K. Steinweg, G.T.Y. Hui, M. Pietsch, A. Ho, M.P.M. van Poppel, D. Lloyd, T2* placental MRI in pregnancies complicated with fetal congenital heart disease, Placenta 108 (2021) 23–31.

[58] J. Liu, G. Gambarota, H. Shu, L. Jiang, B. Leporq, O. Beuf, On the identification of the blood vessel confounding effect in intravoxel incoherent motion (IVIM) Diffusion-Weighted (DW)-MRI in liver: An efficient sparsity based algorithm, Med. Image Anal. 61 (2020), 101637.

[59] M.O. Baerlocher, M. Asch, A. Myers, Allergic-type reactions to radiographic contrast media, CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne 182 (2010), 1328–1328.

[60] C.A. Mathew, S. Maller, Maheshwaran, Interactions between magnetic resonance imaging and dental material, J. Pharm. Bioall. Sci. 5 (2013) S113–S116.

[61] M. Tubiana, Computed tomography and radiation exposure, N. Engl. J. Med. 358 (2008) 850, author reply 852-3.

[62] J. Palle, N.K. Wittig, A. Kubec, S. Niese, M. Rosenthal, M. Burghammer, Nanobeam X-ray fluorescence and diffraction computed tomography on human bone with a resolution better than 120 nm, J. Struct. Biol. 212 (2020), 107631.

[63] P.M. Shikhaliev, Large-scale MV CT for cargo imaging: a feasibility study, Nucl. Instrum. Methods Phys. Res. Sect. A 904 (2018) 35–43.

[64] Ż. Górecka, J. Teichmann, M. Nitschke, A. Chlanda, E. Choińska, C. Werner, Biodegradable fiducial markers for X-ray imaging–soft tissue integration and biocompatibility, J. Mater. Chem. B 4 (2016) 5700–5712.

[65] M. Dietzel, T. Hopp, N. Ruiter, R. Zoubi, I.B. Runnebaum, W.A. Kaiser, Fusion of dynamic contrast-enhanced magnetic resonance mammography at 3.0 T with X-ray mammograms: Pilot study evaluation using dedicated semi-automatic registration software, Eur. J. Radiol. 79 (2011) e98–e102.

[66] Y. Zhao, P. Li, C. Gao, Y. Liu, Q. Chen, F. Yang, TSASNet: tooth segmentation on dental panoramic X-ray images by two-stage attention segmentation network, Knowl.e-Based Syst. 206 (2020), 106338.

[67] H.W.D. Hey, C.-K. Kim, W.-G. Lee, H.-S. Juh, K.-T. Kim, Supra-acetabular line is better than supra-iliac line for coronal balance referencing—a study of perioperative whole spine X-rays in degenerative lumbar scoliosis and ankylosing spondylitis patients, The Spine J. 17 (2017) 1837–1845.

[68] N. Hättenschwiler, Y. Sterchi, M. Mendes, A. Schwaninger, Automation in airport security X-ray screening of cabin baggage: Examining benefits and possible implementations of automated explosives detection, Appl. Ergon. 72 (2018) 58–68.

[69] J.R. Garcia, A. Compte, C. Galan, M. Cozar, M. Buxeda, S. Mourelo, 18F-choline PET/MR in the initial staging of prostate cancer. Impact on the therapeutic approach, Revista Española de Medicina Nuclear e Imagen Molecular (English Edition) 40 (2021) 72–81.

[70] A. Villalobos, B. Cheng, W. Wagstaff, I. Sethi, Z. Bercu, D.M. Schuster, Tumor-to-Normal Ratio Relationship between Planning Technetium-99 Macroaggregated Albumin and Posttherapy Yttrium-90 Bremsstrahlung SPECT/CT, J. Vasc. Interv. Radiol. 32 (2021) 752–760.

[71] M. Diwakar, M. Kumar, A review on CT image noise and its denoising, Biomed. Signal Process. Control 42 (2018) 73–88.

[72] S. Aja-Fernandez, A. Tristan-Vega, A review on statistical noise models for Magnetic Resonance Imaging, in: presented at the LPI, ETSI Telecomunicacion, Spain, 2013.

[73] D.S. Marcus, A.F. Fotenos, J.G. Csernansky, J.C. Morris, R.L. Buckner, Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults, J. Cogn. Neurosci. 22 (2010) 2677–2684.

[74] M. Kazubek, Wavelet domain image denoising by thresholding and Wiener filtering, IEEE Signal Process Lett. 10 (2003) 324–326.

[75] J. Portilla, V. Strela, M.J. Wainwright, E.P. Simoncelli, Image denoising using scale mixtures of Gaussians in the wavelet domain, IEEE Trans. Image Process. 12 (2003) 1338–1351.

[76] M. Ghazel, G.H. Freeman, E.R. Vrscay, Fractal-wavelet image denoising revisited, IEEE Trans. Image Process. 15 (2006) 2669–2675.

[77] P. Gruber, K. Stadlthanner, M. Böhm, F.J. Theis, E.W. Lang, A.M. Tomé, Denoising using local projective subspace methods, Neurocomputing 69 (2006) 1485–1501.

[78] F. Luisier, T. Blu, M. Unser, A new SURE approach to image denoising: interscale orthonormal wavelet thresholding, IEEE Trans. Image Process. 16 (2007) 593–606.

[79] A. Khmag, A.R. Ramli, N. Kamarudin, Clustering-based natural image denoising using dictionary learning approach in wavelet domain, Soft. Comput. 23 (2018) 8013–8027.

[80] Z. Bao, G. Zhang, B. Xiong, S. Gai, New image denoising algorithm using monogenic wavelet transform and improved deep convolutional neural network, Multimedia Tools App. 79 (2019) 7401–7412.

[81] B. Chen, X. Feng, R. Wu, Q. Guo, X. Wang, S. Ge, Adaptive wavelet filter with edge compensation for remote sensing image denoising, IEEE Access. 7 (2019) 91966–91979.

[82] Y.E. Gökdağ, F. Şansal, Y.D. Gökdel, Image denoising using 2-D wavelet algorithm for Gaussian-corrupted confocal microscopy images, Biomed. Signal Process. Control 54 (2019), 101594.

[83] N.A. Golilarz, M. Mirmozaffari, T.A. Gashteroodkhani, L. Ali, H.A. Dolatsara, A. Boskabadi, Optimized wavelet-based satellite image de-noising with multi-population differential evolution-assisted harris hawks optimization algorithm, IEEE Access. 8 (2020) 133076–133085.

[84] M. Malfait, D. Roose, Wavelet-based image denoising using a markov random field a Priori model, IEEE Trans. Image Process. 6 (1997) 549–565.

[85] X. Hua, L.E. Pierce, F.T. Ulaby, SAR speckle reduction using wavelet denoising and Markov random field modeling, IEEE Trans. Geosci. Remote Sens. 40 (2002) 2196–2212.

[86] A. Barbu, Training an active random field for real-time image denoising, IEEE Trans. Image Process. 18 (2009) 2451–2462.

[87] Y. Cao, Y. Luo, S. Yang, Image denoising based on hierarchical Markov random field, Pattern Recognit. Lett. 32 (2011) 368–374.

[88] Z. Xu, Q. Shi, Denoising model for parallel magnetic resonance imaging images using higher-order Markov random fields, IET Image Proc. 10 (2016) 962–970.

[89] K. Lekadir, M. Lange, V.A. Zimmer, C. Hoogendoorn, A.F. Frangi, Statistically-driven 3D fiber reconstruction and denoising from multi-slice cardiac DTI using a Markov random field model, Med. Image Anal. 27 (2016) 105–116.

[90] M. Ben Abdallah, J. Malek, A.T. Azar, H. Belmabrouk, J.Esclarín Monreal, K. Krissian, Adaptive noise-reducing anisotropic diffusion filter, Neural. Comput. App. 27 (2015) 1273–1300.

[91] S. Kim, S.-J. Kang, Y.H. Kim, Anisotropic diffusion noise filtering using region adaptive smoothing strength, J. Visual Commun. Image Represent. 40 (2016) 384–391.

[92] C. Beitone, X. Balandraud, D. Delpueyo, M. Grédiac, Heat source reconstruction from noisy temperature fields using a gradient anisotropic diffusion filter, Infrared Phys. Technol. 80 (2017) 27–37.

[93] M. Ben Abdallah, A.T. Azar, H. Guedri, J. Malek, H. Belmabrouk, Noise-estimation-based anisotropic diffusion approach for retinal blood vessel segmentation, Neural Comput. App. 29 (2017) 159–180.

[94] H. Chen, J. Feng, B. Zhou, Y. Hu, K. Guo, An anisotropic diffusion-based dynamic combined energy model for seismic denoising, IEEE Geosci. Remote Sens. Lett. 14 (2017) 1061–1065.

[95] A. Hadj Fredj, J. Malek, GPU-based anisotropic diffusion algorithm for video image denoising, Microprocess. Microsyst. 53 (2017) 190–201.

[96] L. Jubairahmed, S. Satheeskumaran, C. Venkatesan, Contourlet transform based adaptive nonlinear diffusion filtering for speckle noise removal in ultrasound images, Cluster Comput. 22 (2017) 11237–11246.

[97] V. Kamalaveni, S. Veni, K.A. Narayanankutty, Improved self-snake based anisotropic diffusion model for edge preserving image denoising using structure tensor, Multimedia Tools App. 76 (2017) 18815–18846.

[98] J. Bai, X.-C. Feng, Image Denoising using generalized anisotropic diffusion, J. Math. Imaging Vis. 60 (2018), 994–007.

[99] R.I. Elsharif, B.A. Ibraheem, Z.A. Mustafa, S.K. Abass, M.M. Fadl allah, Wavelet decomposition–based speckle reduction method for ultrasound images by using speckle-reducing anisotropic diffusion and hybrid median, J. Clin. Eng. 43 (2018) 163–170.

[100] F. Guo, G. Zhang, Q. Zhang, R. Zhao, M. Deng, K. Xu, Speckle suppression by weighted euclidean distance anisotropic diffusion, Remote Sens. 10 (2018) 722.

[101] D. Mishra, S. Chaudhury, M. Sarkar, A.S. Soin, V. Sharma, Edge probability and pixel relativity-based speckle reducing anisotropic diffusion, IEEE Trans. Image Process. 27 (2018) 649–664.

[102] K. Mei, B. Hu, B. Fei, B. Qin, Phase Asymmetry Ultrasound Despeckling with fractional anisotropic diffusion and total variation, IEEE Trans. Image Process. 29 (2020) 2845–2859.

[103] J. Yang, J. Fan, D. Ai, X. Wang, Y. Zheng, S. Tang, Local statistics and non-local mean filter for speckle noise reduction in medical ultrasound image, Neurocomputing 195 (2016) 88–95.

[104] L. Chen, L. Liu, C.L. Philip Chen, A robust bi-sparsity model with non-local regularization for mixed noise reduction, Inform. Sci. 354 (2016) 101–111.

[105] H. Yu, J. Gao, A. Li, Probability-based non-local means filter for speckle noise suppression in optical coherence tomography images, Opt. Lett. 41 (2016) 994–997.

[106] W. Zeng, Y. Du, C. Hu, Noise suppression by discontinuity indicator controlled non-local means method, Multimed. Tools App. 76 (2016) 13239–13253.

[107] S. Mandal, A. Bhavsar, A.K. Sao, Noise adaptive super-resolution from single image via non-local mean and sparse representation, Signal Process. 132 (2017) 134–149.

[108] X. Qian, X. Jia, Z. Wang, B. Zhang, N. Xue, W. Sun, Noise level estimation of BOTDA for optimal non-local means denoising, Appl. Opt. 56 (2017) 4727–4734.

[109] C. Tang, L. Cao, J. Chen, X. Zheng, Speckle noise reduction for optical coherence tomography images via non-local weighted group low-rank representation, Laser Phys. Lett. 14 (2017), 056002.

[110] A.A. Bindilatti, M.A.C. Vieira, N.D.A. Mascarenhas, Poisson Wiener filtering with non-local weighted parameter estimation using stochastic distances, Signal Process. 144 (2018) 68–76.

[111] M. Georgiev, R. Bregović, A. Gotchev, Time-of-flight range measurement in low-sensing environment: noise analysis and complex-domain Non-Local Denoising, IEEE Trans. Image Process. 27 (2018) 2911–2926.

[112] S.K. Panigrahi, S. Gupta, P.K. Sahu, Curvelet-based multi-scale denoising using non-local means & guided image filter, IET Image Proc. 12 (2018) 909–918.

[113] H.R. Shahdoosti, Z. Rahemi, A maximum likelihood filter using non-local information for despeckling of ultrasound images, Mach. Vis. App. 29 (2018) 689–702.

[114] Y. Hou, J. Xu, M. Liu, G. Liu, L. Liu, F. Zhu, NLH: A blind pixel-level non-local method for real-world image denoising, IEEE Trans. Image Process. 29 (2020) 5121–5135.

[115] F. Mei, D. Zhang, Y. Yang, Improved non-local self-similarity measures for effective speckle noise reduction in ultrasound images, Comput. Methods Progr. Biomed. 196 (2020), 105670.

[116] H. Zeng, X. Xie, W. Kong, S. Cui, J. Ning, Hyperspectral image denoising via combined non-local self-similarity and local low-rank regularization, IEEE Access. 8 (2020) 50190–50208.

[117] M. Zhang, B.K. Gunturk, Multiresolution bilateral filtering for image denoising, IEEE Trans. Image Process.: Publ. IEEE Signal Process. Soc. 17 (2008) 2324–2333.

[118] S. Akdemir Akar, Determination of optimal parameters for bilateral filter in brain MR image denoising, Appl. Soft Comput. 43 (2016) 87–96.

[119] S. Balocco, C. Gatta, O. Pujol, J. Mauri, P. Radeva, SRBF: Speckle reducing bilateral filtering, Ultrasound Med. Biol. 36 (2010) 1353–1363.

[120] C. Lin, J. Tsai, C. Chiu, Switching bilateral filter with a texture/noise detector for universal noise removal, IEEE Trans. Image Process. 19 (2010) 2307–2320.

[121] Y. Zhang, X. Tian, P. Ren, An adaptive bilateral filter based framework for image denoising, Neurocomputing 140 (2014) 299–316.

[122] M. Wei, W. Shen, J. Qin, J. Wu, T.-T. Wong, P.-A. Heng, Feature-preserving optimization for noisy mesh using joint bilateral filter and constrained Laplacian smoothing, Opt. Lasers Eng. 51 (2013) 1223–1234.

[123] A. Phophalia, S.K. Mitra, Rough set based bilateral filter design for denoising brain MR images, Appl. Soft Comput. 33 (2015) 1–14.

[124] J. Zhang, L. Wu, G. Lin, Y. Cheng, An integrated de-speckling approach for medical ultrasound images based on wavelet and trilateral filter, Circuits, Syst., Signal Process. 36 (2016) 297–314.

[125] Z. Chen, J. Jiang, C. Zhou, X. Jiang, S. Fu, C. Zhihua, Trilateral smooth filtering for hyperspectral image feature extraction, IEEE Geosci. Remote Sens. Lett. 16 (2019) 781–785.

[126] K. Langampol, K. Srisomboon, V. Patanavijit, W. Lee, Smart switching bilateral filter with estimated noise characterization for mixed noise removal, Math. Probl. Eng. 2019 (2019) 1–23.

[127] W. Cui, M. Li, G. Gong, K. Lu, S. Sun, F. Dong, Guided trilateral filter and its application to ultrasound image despeckling, Biomed. Signal Process. Control 55 (2020), 101625.

[128] W. Ci, Z. Lei, H. Yuwen, T. Yap-Peng, Frame Rate Up-Conversion Using Trilateral Filtering, IEEE Trans. Circuits Syst. Video Technol. 20 (2010) 886–893.

[129] Y. Tan, P. Tang, Y. Zhou, W. Luo, Y. Kang, G. Li, Photograph aesthetical evaluation and classification with deep convolutional neural networks, Neurocomputing 228 (2017) 165–175.

[130] H. Choi, K.H. Jin, Fast and robust segmentation of the striatum using deep convolutional neural networks, J. Neurosci. Methods 274 (2016) 146–153.

[131] Y. Li, W. Xie, H. Li, Hyperspectral image reconstruction by deep convolutional neural network for classification, Pattern Recognit. 63 (2017) 371–383.

[132] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (1989) 541–551.

[133] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neur. Inform. Process. Syst. 25 (2012) 1097–1105.

[134] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,*2014.

[135] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[136] G. Huang, Z. Liu, V. Laurens, K. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2261–2269.

[137] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," *arXiv:1602.07360,*2016.

[138] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks, The IEEE Confer. Comput. Vis. Pattern Recogn. (CVPR) (2018) 4510–4520.

[139] C.J. Schuler, H.C. Burger, S. Harmeling, B. Schölkopf, A machine learning approach for non-blind image deconvolution, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1067–1074.

[140] D.-A. Huang, L.-W. Kang, Y.-C.F. Wang, C.-W. Lin, Self-learning based image decomposition with applications to single image denoising, IEEE Trans. Multimedia 16 (2014) 83–93.

[141] H.M. Li, Deep learning for image denoising, Int. J. Signal Process., Image Process. Pattern Recogn. 7 (2014) 171–180.

[142] X. Sun, N.K. Kottayil, S. Mukherjee, I. Cheng, Adversarial training for dual-stage image denoising enhanced with feature matching. Smart Multimedia, 2018, pp. 357–366.

[143] P. Xiao, Y. Guo, P. Zhuang, Removing stripe noise from infrared cloud images via deep convolutional networks, IEEE Photonics J. 10 (2018) 1–14.

[144] W. Xie, Y. Li, X. Jia, Deep convolutional networks with residual learning for accurate spectral-spatial denoising, Neurocomputing 312 (2018) 372–381.

[145] W. Liu, J. Lee, A 3-D Atrous convolution neural network for hyperspectral image denoising, IEEE Trans. Geosci. Remote Sens. 57 (2019) 5701–5715.

[146] C. Tian, Y. Xu, L. Fei, J. Wang, J. Wen, N. Luo, Enhanced CNN for image denoising, CAAI Trans. Intell. Technol. 4 (2019) 17–23.

[147] Y. Zheng, H. Duan, X. Tang, C. Wang, J. Zhou, Denoising in the dark: privacy-preserving deep neural network based image denoising, IEEE Trans. Dependable Secure Comput. (2019), 1–1.

[148] D. Ren, W. Shang, P. Zhu, Q. Hu, D. Meng, W. Zuo, Single image deraining using bilateral recurrent network, IEEE Trans. Image Process. 29 (2020) 6852–6863.

[149] C. Tian, Y. Xu, Z. Li, W. Zuo, L. Fei, H. Liu, Attention-guided CNN for image denoising, Neural Netw. 124 (2020) 117–129.

[150] C. Tian, Y. Xu, W. Zuo, Image denoising using deep CNN with batch renormalization, Neural Netw. 121 (2020) 461–473.

[151] A. Clauset, C. Moore, M.E. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (2008) 98–101.

[152] A.B. Pedersen, E.M. Mikkelsen, D. Cronin-Fenton, N.R. Kristensen, T.M. Pham, L. Pedersen, Missing data and multiple imputation in clinical epidemiological research, Clin. Epidemiol. 9 (2017) 157.

[153] S. Fielding, P.M. Fayers, A. McDonald, G. McPherson, M.K. Campbell, Simple imputation methods were inadequate for Missing Not At Random (MNAR) quality of life data, Health Qual. Life Outcomes 6 (2008) 1–9.

[154] G. Molenberghs, C. Beunckens, C. Sotto, M.G. Kenward, Every missingness not at random model has a missingness at random counterpart with equal fit, J. R. Stat. Soc.: Ser. B. (Stat. Methodol.) 70 (2008) 371–388.

[155] J.-O. Kim, J. Curry, The treatment of missing data in multivariate analysis, Sociol. Methods Res. 6 (1977) 215–240.

[156] Y. Dong, C.-Y.J. Peng, Principled missing data methods for researchers, Springer Plus 2 (2013) 1–17.

[157] F. Cismondi, A.S. Fialho, S.M. Vieira, S.R. Reti, J.M. Sousa, S.N. Finkelstein, Missing data in medical databases: Impute, delete or classify? Artif. Intell. Med. 58 (2013) 63–72.

[158] K.T. Do, S. Wahl, J. Raffler, S. Molnos, M. Laimighofer, J. Adamski, Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies, Metabolomics 14 (2018) 1–18.

[159] D. Roland, N. Suzen, T.J. Coats, J. Levesley, A.N. Gorban, E.M. Mirkes, What can the randomness of missing values tell you about clinical practice in large data sets of children's vital signs? Pediatr. Res. 89 (2021) 16–21.

[160] E.M. Mirkes, T.J. Coats, J. Levesley, A.N. Gorban, Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes, Comput. Biol. Med. 75 (2016) 203–216.

[161] A. Idri, I. Abnane, A. Abran, Missing data techniques in analogy-based software development effort estimation, J. Syst. Softw. 117 (2016) 595–611.

[162] I. Myrtveit, E. Stensrud, U.H. Olsson, Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods, IEEE Trans. Softw. Eng. 27 (2001) 999–1013.

[163] Q. Wang, J. Rao, Empirical likelihood-based inference in linear models with missing data, Scand. J. Stat. 29 (2002) 563–576.

[164] A. Stamatakis, N. Alachiotis, Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data, Bioinformatics 26 (2010) i132–i139.

[165] R.J. Little, Modeling the drop-out mechanism in repeated-measures studies, J. Am. Statist. Assoc. 90 (1995) 1112–1121.

[166] V. Lefort, J.-E. Longueville, O. Gascuel, SMS: smart model selection in PhyML, Mol. Biol. Evol. 34 (2017) 2422–2424.

[167] J.L. Schafer, J.W. Graham, Missing data: our view of the state of the art, Psychol. Methods 7 (2002) 147.

[168] B. Ratitch, M. O'Kelly, R. Tosiello, Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models, Pharm. Stat. 12 (2013) 337–347.

[169] A.A. Tsiatis, M. Davidian, Joint modeling of longitudinal and time-to-event data: an overview, Stat. Sin. (2004) 809–834.

[170] N.C. Gottfredson, D.J. Bauer, S.A. Baldwin, Modeling change in the presence of nonrandomly missing data: Evaluating a shared parameter mixture model, Struct. Equat. Model.: A Multidiscip. J. 21 (2014) 196–209.

[171] A.M. Gad, N.M. Darwish, A shared parameter model for longitudinal data with missing values, Am. J. Appl. Math. Stat. 1 (2013) 30–35.

[172] J. Roy, Modeling longitudinal data with nonignorable dropouts using a latent dropout class model, Biometrics 59 (2003) 829–836.

[173] N.M. Laird, Missing data in longitudinal studies, Stat. Med. 7 (1988) 305–315.

[174] A. Rotnitzky, D. Wypij, A note on the bias of estimators with missing data, Biometrics (1994) 1163–1170.

[175] J.M. Robins, A. Rotnitzky, D.O. Scharfstein, Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. Statistical Models in Epidemiology, the Environment, and Clinical Trials, Springer, 2000, pp. 1–94.

[176] S. Vansteelandt, E. Goetghebeur, M.G. Kenward, G. Molenberghs, Ignorance and uncertainty regions as inferential tools in a sensitivity analysis, Stat. Sin. (2006) 953–979.

[177] D.J. Stekhoven, P. Bühlmann, MissForest—non-parametric missing value imputation for mixed-type data, Bioinformatics 28 (2012) 112–118.

[178] A.R.T. Donders, G.J. Van Der Heijden, T. Stijnen, K.G. Moons, A gentle introduction to imputation of missing values, J. Clin. Epidemiol. 59 (2006) 1087–1091.

[179] A.K. Waljee, A. Mukherjee, A.G. Singal, Y. Zhang, J. Warren, U. Balis, Comparison of imputation methods for missing laboratory data in medicine, BMJ Open 3 (2013).

[180] J. Brick, G. Kalton, Handling missing data in survey research, Stat. Methods Med. Res. 5 (1996) 215–238.

[181] Z. Zhang, Missing data imputation: focusing on single imputation, Ann. Transl. Med. 4 (2016), 9–9.

[182] N.J. Horton, S.R. Lipsitz, Multiple imputation in practice: comparison of software packages for regression models with missing variables, Am. Stat. 55 (2001) 244–254.

[183] W.A. Fuller, J.K. Kim, Hot deck imputation for the response model, Surv. Methodol. 31 (2005) 139.

[184] S. Yenduri, S.S. Iyengar, Performance evaluation of imputation methods for incomplete datasets, Int. J. Softw. Eng. Knowledge Eng. 17 (2007) 127–152.

[185] F. Biessmann, T. Rukat, P. Schmidt, P. Naidu, S. Schelter, A. Taptunov, DataWig: missing value imputation for tables, J. Mach. Learn. Res. 20 (2019) 1–6.

[186] F.V. Nelwamondo, S. Mohamed, T. Marwala, Missing data: a comparison of neural network and expectation maximization techniques, Curr. Sci. (2007) 1514–1521.

[187] C.B. Do, S. Batzoglou, What is the expectation maximization algorithm? Nat. Biotechnol. 26 (2008) 897–899.

[188] K. Zhang, R. Gonzalez, B. Huang, G. Ji, Expectation–maximization approach to fault diagnosis with missing data, IEEE Trans. Indust. Electron. 62 (2014) 1231–1240.

[189] M. Pampaka, G. Hutcheson, J. Williams, Handling missing data: analysis of a challenging data set using multiple imputation, Int. J. Res. Method Edu. 39 (2016) 19–37.

[190] J.P. Reiter, T.E. Raghunathan, The multiple adaptations of multiple imputation, J. Am. Statist. Assoc. 102 (2007) 1462–1471.

[191] S. Van Buuren, Flexible imputation of missing data, CRC press, 2018.

[192] P.D. Allison, Multiple imputation for missing data: a cautionary tale, Sociol. Methods & Res. 28 (2000) 301–309.

[193] P. Zhang, Multiple imputation: theory and method, Int. Stat. Rev. 71 (2003) 581–592.

[194] S. Sinharay, H.S. Stern, D. Russell, The use of multiple imputation for the analysis of missing data, Psychol. Methods 6 (2001) 317.

[195] M.C. Wu, K.R. Bailey, Estimation and comparison of changes in the presence of informative right censoring: conditional linear model, Biometrics (1989) 939–955.

[196] Q. Song, M. Shepperd, Missing data imputation techniques, Int. J. Bus. Intell. Data Mining 2 (2007) 261–291.

[197] R.A. Chowdhury, Y. Zerouali, T. Hedrich, M. Heers, E. Kobayashi, J.M. Lina, MEG-EEG information fusion and electromagnetic source imaging: from theory to clinical application in epilepsy, Brain Topogr. 28 (2015) 785–812.

[198] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: an overview of methods, challenges, and prospects, Proc. IEEE 103 (2015) 1449–1477.

[199] A. Clauset, C. Moore, M.E.J. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (2008) 98–101.

[200] A.W. Toga, P.M. Thompson, The role of image registration in brain mapping, Image Vision Comput. 19 (2001) 3–24.

[201] C. Che, T.S. Mathai, J. Galeotti, Ultrasound registration: a review, Methods 115 (2017) 128–143.

[202] P. Czajkowski, T. Piotrowski, Registration methods in radiotherapy, Rep. Pract. Oncol. Radiother. 24 (2019) 28–34.

[203] Y. Sakr, M.-J. Dubois, D. De Backer, J. Creteur, J.-L. Vincent, Persistent microcirculatory alterations are associated with organ failure and death in patients with septic shock, Crit. Care Med. 32 (2004) 1825–1831.

[204] S. Mohammadian, J. Fokkema, A.V. Agronskaia, N. Liv, C. de Heus, E. van Donselaar, High accuracy, fiducial marker-based image registration of correlative microscopy images, Sci. Rep. 9 (2019) 1–10.

[205] D.N. Greve, B. Fischl, Accurate and robust brain image alignment using boundary-based registration, Neuroimage 48 (2009) 63–72.

[206] A. Sotiras, C. Davatzikos, N. Paragios, Deformable medical image registration: a survey, IEEE Trans. Med. Imaging 32 (2013) 1153–1190.

[207] J. Tsao, Interpolation artifacts in multimodality image registration based on maximization of mutual information, IEEE Trans. Med. Imaging 22 (2003) 854–864.

[208] V. Gorbunova, J. Sporring, P. Lo, M. Loeve, H.A. Tiddens, M. Nielsen, Mass preserving image registration for lung CT, Med. Image Anal. 16 (2012) 786–795.

[209] D. Mahapatra, B. Antony, S. Sedai, R. Garnavi, Deformable medical image registration using generative adversarial networks, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 1449–1453.

[210] Y. Xu, L. Peng, G.-Y. Li, Multi modal registration of structural features and mutual information of medical image, Future Gener. Comput. Syst. 93 (2019) 499–505.

[211] J.P. Pluim, J.A. Maintz, M.A. Viergever, Mutual-information-based registration of medical images: a survey, IEEE Trans. Med. Imaging 22 (2003) 986–1004.

[212] B. Cuer, C. Mollevi, A. Anota, E. Charton, B. Juzyna, T. Conroy, Handling informative dropout in longitudinal analysis of health-related quality of life: application of three approaches to data from the esophageal cancer clinical trial PRODIGE 5/ACCORD 17, BMC Med. Res. Methodol. 20 (2020) 223.

[213] Y. Tang, An efficient multiple imputation algorithm for control-based and delta-adjusted pattern mixture models using SAS, Stat. Biopharm. Res. 9 (2017) 116–125.

[214] D. Wei, S. Ahmad, J. Huo, W. Peng, Y. Ge, Z. Xue, Synthesis and inpainting-based mr-ct registration for image-guided thermal ablation of liver tumors, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 512–520.

[215] C.R. Maurer, J.M. Fitzpatrick, M.Y. Wang, R.L. Galloway, R.J. Maciunas, G. S. Allen, Registration of head volume images using implantable fiducial markers, IEEE Trans. Med. Imaging 16 (1997) 447–462.

[216] G.C. Sharp, S.W. Lee, D.K. Wehe, ICP registration using invariant features, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 90–102.

[217] A. Almhdie, C. Léger, M. Deriche, R. Lédée, 3D registration using a new implementation of the ICP algorithm based on a comprehensive lookup matrix: application to medical imaging, Pattern Recognit. Lett. 28 (2007) 1523–1533.

[218] A.D. Savva, T.L. Economopoulos, G.K. Matsopoulos, Geometry-based vs. intensity-based medical image registration: a comparative study on 3D CT data, Comput. Biol. Med. 69 (2016) 120–133.

[219] A.A. Tsiatis, M.A. Davidian, Joint modeling of longitudinal and time-to-event data: an overview, Statistica Sinica 14 (2004) 793–818.

[220] T. Gebäck, P. Koumoutsakos, Edge detection in microscopy images using curvelets, BMC Bioinformatics 10 (2009) 75.

[221] J. Hu, Y. Yang, Z. Su, A novel hierarchical medical image registration method based on multiscale and contour line, in: 2012 International Conference on Systems and Informatics (ICSAI2012), 2012, pp. 1834–1837.

[222] R. Maksimovic, S. Stankovic, D. Milovanovic, Computed tomography image analyzer: 3D reconstruction and segmentation applying active contour models—'snakes', Int. J. Med. Inf. 58 (2000) 29–37.

[223] H. Li, B. Manjunath, S.K. Mitra, A contour-based approach to multisensor image registration, IEEE Trans. Image Process. 4 (1995) 320–334.

[224] L.D. Cohen, I. Cohen, Finite-element methods for active contour models and balloons for 2-D and 3-D images, IEEE Trans. Pattern Anal. Mach. Intell. 15 (1993) 1131–1147.

[225] C.P. Loizou, C.S. Pattichis, M. Pantziaris, T. Tyllis, A. Nicolaides, Snakes based segmentation of the common carotid artery intima media, Med. Biol. Eng. Comput. 45 (2007) 35–49.

[226] D. Li, W. Zhong, K.M. Deh, T.D. Nguyen, M.R. Prince, Y. Wang, Discontinuity preserving liver MR registration with three-dimensional active contour motion segmentation, IEEE Trans. Biomed. Eng. 66 (2018) 1884–1897.

[227] Y. Sun, Q. Feng, Liver DCE-MRI registration based on sparse recovery de-enhanced curves, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 705–708.

[228] M. Cohen, A new approach to imputation, Am. Stat. Assoc. Proced. Section Surv. Res. Methods 13 (1996) 293–298.

[229] H. He, Q. Razlighi, Volumetric Registration of Brain Cortical Regions by Automatic Landmark Matching and Large Deformation Diffeomorphisms, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1412–1417.

[230] S. Khallaghi, C.A. Sánchez, A. Rasoulian, Y. Sun, F. Imani, A. Khojaste, Biomechanically constrained surface registration: application to MR-TRUS fusion for prostate interventions, IEEE Trans. Med. Imaging 34 (2015) 2404–2414.

[231] S. Khallaghi, C.A. Sánchez, A. Rasoulian, S. Nouranian, C. Romagnoli, H. Abdi, Statistical biomechanical surface registration: application to MR-TRUS fusion for prostate interventions, IEEE Trans. Med. Imaging 34 (2015) 2535–2549.

[232] C. Raposo, J.P. Barreto, 3D registration of curves and surfaces using local differential information, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9300–9308.

[233] H. Lester, S.R. Arridge, A survey of hierarchical non-linear medical image registration, Pattern Recognit. 32 (1999) 129–149.

[234] N. Houhou, V. Duay, A.S. Allal, J.-P. Thiran, Medical images registration with a hierarchical atlas, in: 2005 13th European Signal Processing Conference, 2005, pp. 1–4.

[235] S. Kim, Y.-W. Tai, Hierarchical non-rigid model for 3D medical image registration, in: 2014 IEEE International Conference on Image Processing (ICIP), 2014, pp. 3562–3566.

[236] D. De Nigris, L. Mercier, R. Del Maestro, D.L. Collins, T. Arbel, Hierarchical multimodal image registration based on adaptive local mutual information, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2010, pp. 643–651.

[237] Z. Xiao-chun, Z. Xin-bo, F. Yan, An efficient medical image registration algorithm based on gradient descent, in: 2007 IEEE/ICME International Conference on Complex Medical Engineering, 2007, pp. 636–639.

[238] M.J. Powell, An efficient method for finding the minimum of a function of several variables without calculating derivatives, Comput. J. 7 (1964) 155–162.

[239] M.P. Wachowiak, T.M. Peters, High-performance medical image registration using new optimization techniques, IEEE Trans. Inf. Technol. Biomed. 10 (2006) 344–353.

[240] G. Haskins, U. Kruger, P. Yan, Deep learning in medical image registration: a survey, Mach. Vis. App. 31 (2020) 8.

[241] Z. Zhang, Missing data imputation: focusing on single imputation, Ann. Transl. Med. 4 (2016).

[242] R.R. Andridge, R.J. Little, A review of hot deck imputation for survey non-response, Int. Stat. Rev. 78 (2010) 40–64.

[243] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, N. Komodakis, A deep metric for multimodal registration, in: International conference on medical image computing and computer-assisted intervention, 2016, pp. 10–18.

[244] M. Arar, Y. Ginger, D. Danon, A.H. Bermano, D. Cohen-Or, Unsupervised multi-modal image registration via geometry preserving image-to-image translation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13410–13419.

[245] A. Danilchenko, J.M. Fitzpatrick, General approach to first-order error prediction in rigid point registration, IEEE Trans. Med. Imaging 30 (2010) 679–693.

[246] T. Makela, P. Clarysse, O. Sipila, N. Pauna, Q.C. Pham, T. Katila, A review of cardiac image registration methods, IEEE Trans. Med. Imaging 21 (2002) 1011–1021.

[247] G.E. Christensen, H.J. Johnson, Consistent image registration, IEEE Trans. Med. Imaging 20 (2001) 568–582.

[248] Y. Zhang, Y. Fang, W. Lin, X. Zhang, L. Li, Backward registration-based aspect ratio similarity for image retargeting quality assessment, IEEE Trans. Image Process. 25 (2016) 4286–4297.

[249] J.D. Gispert, S. Reig, J. Pascau, J.J. Vaquero, P. García-Barreno, M. Desco, Method for bias field correction of brain T1-weighted magnetic resonance images minimizing segmentation error, Hum. Brain Mapp. 22 (2004) 133–144.

[250] H.-I. Suk, S.-W. Lee, D. Shen, A.s.D.N. Initiative, Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis, Neuroimage 101 (2014) 569–582.

[251] M. Liu, D. Cheng, K. Wang, Y. Wang, Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis, Neuroinformatics 16 (2018) 295–308.

[252] J. Liu, J. Wang, Z. Tang, B. Hu, F.-X. Wu, Y. Pan, Improving Alzheimer's disease classification by combining multiple measures, IEEE/ACM Trans. Comput. Biol. Bioinf. 15 (2017) 1649–1659.

[253] C. Piron, P. Causer, R. Jong, R. Shumak, D.B. Plewes, A hybrid breast biopsy system combining ultrasound and MRI, IEEE Trans. Med. Imaging 22 (2003) 1100–1110.

[254] Z. Xu, C.P. Lee, M.P. Heinrich, M. Modat, D. Rueckert, S. Ourselin, Evaluation of six registration methods for the human abdomen on clinically acquired CT, IEEE Trans. Biomed. Eng. 63 (2016) 1563–1572.

[255] M. Freiman, S.D. Voss, S.K. Warfield, Abdominal images non-rigid registration using local-affine diffeomorphic demons, in: International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging, 2011, pp. 116–124.

[256] C. Lazar, S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, Batch effect removal methods for microarray gene expression data integration: a survey, Brief. Bioinform. 14 (2013) 469–490.

[257] A.S. Tarawneh, A.B.A. Hassanat, K. Almohammadi, D. Chetverikov, C. Bellinger, SMOTEFUNA: Synthetic Minority Over-Sampling Technique Based on Furthest Neighbour Algorithm, IEEE Access 8 (2020) 59069–59082.

[258] Z. Jan, B. Verma, Multiple strong and balanced cluster-based ensemble of deep learners, Pattern Recognit. 107 (2020) 11.

[259] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, IEEE, 2014, pp. 580–587.

[260] R. Girshick, Fast R-CNN, in: International Conference on Computer Vision (ICCV), Santiago, CHILE, IEEE, 2015, pp. 1440–1448.

[261] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.

[262] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, IEEE, 2016, pp. 779–788.

[263] J. Redmon, A. Farhadi, YOLO9000: Better, Faster, Stronger, in: 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, IEEE, 2017, pp. 6517–6525.

[264] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," *arXiv: 1804.02767,*2018.

[265] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (2019) 60.

[266] J. Amin, M. Sharif, M.A. Anjum, M. Raza, S.A.C. Bukhari, Convolutional neural network with batch normalization for glioma and stroke lesion detection using MRI, Cogn. Syst. Res. 59 (2020) 304–311.

[267] J. Kileel, M. Trager, J. Bruna, On the expressive power of deep polynomial neural networks, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, La Jolla 32, Neural Information Processing Systems (NIPS), 2019, pp. 1–10.

[268] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from over-fitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.

[269] T. Fechter, D. Baltas, One-shot learning for deformable medical image registration and periodic motion tracking, IEEE Trans. Med. Imaging 39 (2020) 2506–2517.

[270] C.Y. Ma, S.S. Zhang, A.N. Wang, Y.Y. Qi, G. Chen, Skeleton-based dynamic hand gesture recognition using an enhanced network with one-shot learning, Appl. Sci.-Basel 10 (2020) 16.

[271] A. Puzanov, S.Y. Zhang, K. Cohen, Deep reinforcement one-shot learning for artificially intelligent classification in expert aided systems, Eng. Appl. Artif. Intell. 91 (2020) 12.

[272] M. Adiban, H. Sameti, S. Shehnepoor, Replay spoofing countermeasure using autoencoder and siamese networks on ASVspoof 2019 challenge, Comput. Speech Lang. 64 (2020) 13.

[273] N. Challapalle, S. Rampalli, N. Jao, A. Ramanathan, J. Sampson, V. Narayanan, FARM: a flexible accelerator for recurrent and memory augmented neural networks, J. Signal Process. Syst. 92 (2020) 1247–1261.

[274] I. Deznabi, B. Arabaci, M. Koyuturk, O. Tastan, DeepKinZero: zero-shot learning for predicting kinase-phosphosite associations involving understudied kinases, Bioinformatics 36 (2020) 3652–3661.

[275] N.W.D. Cunha, S.A. Birajdhar, K. Manikantan, S. Ramachandran, Face recognition using Homomorphic Filtering as a pre-processing technique, in: 2013 International Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications (C2SPCA), 2013, pp. 1–6.

[276] M. Morita, Y. Fujii, T. Sato, The width under-estimation of 3D objects with image rotation, I-Perception 10 (2019), 43–43.

[277] M. George, B.R. Jose, J. Mathew, Abnormal activity detection using shear transformed Spatio-temporal regions at the surveillance network edge, Multimedia Tools App 79 (2020) 27511–27532.

[278] E. Park, Y.J. Moon, D. Lim, H. Lee, De-noising SDO/HMI Solar Magnetograms by image translation method based on deep learning, Astrophys. J. Lett. 891 (2020) 9.

[279] J. Gawedzinski, K.M. Schmeler, A. Milbourne, P. Ramalingam, P.A. Moghaddam, R. Richards-Kortum, Toward development of a large field-of-view cancer screening patch (CASP) to detect cervical intraepithelial neoplasia, Biomed. Optics Express 10 (2019) 6145–6159.

[280] S.R. Leyh-Bannurah, U. Wolffgang, J. Schmitz, V. Ouellet, F. Azzi, Z. Tian, state-of-the-art weakly supervised automated classification of prostate cancer tissue microarrays via deep learning: can sufficient accuracy be achieved without manual patch level annotation? J. Urol. 203 (2020). E306–E306.

[281] S. Banerjee, R. Chipman, Y. Otani, Simultaneous balancing of geometric transformation and linear polarizations using six-fold-mirror geometry over the visible region, Opt. Lett. 45 (2020) 2510–2513.

[282] Y. Tada, Y. Hagiwara, H. Tanaka, T. Taniguchi, Robust understanding of robot-directed speech commands using sequence to sequence with noise injection, Front. Robot. Ai 6 (2020) 12.

[283] K. Muhammad, Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation, Multimedia Tools App. 78 (2019) 3613–3632.

[284] J.A. Pandian, G. Geetharamani, B. Annette, M A M Coll Engn & Technol, Data augmentation on plant leaf disease image dataset using image manipulation and deep learning techniques, in: 9th International Conference on Advanced Computing, Tiruchirapalli, INDIA, IEEE, 2019, pp. 199–204.

[285] X. Li, Y. Chai, W. Chen, F. Ao, Identification of early esophageal cancer based on data augmentation, in: 39th Chinese Control Conference (CCC), Shenyang, China, China, 2020, pp. 6307–6312.

[286] M. Sheeny, A. Wallace, RADIO: Parameterized generative radar data augmentation for small datasets, Appl. Sci.-Basel 10 (2020) 13.

[287] A. Sezer, H.B. Sezer, Deep convolutional neural network-based automatic classification of neonatal hip ultrasound images: A novel data augmentation approach with speckle noise reduction, Ultrasound Med. Biol. 46 (2020) 735–749.

[288] T. DeVries, G.W. Taylor, Dataset augmentation in feature space, in: 5th International conference on learning representations (ICLR), Toulon, France, 2017, pp. 1–12.

[289] L. Xie, J. Wang, Z. Wei, M. Wang, Q. Tian, Disturblabel: regularizing cnn on the loss layer, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, 2016, pp. 4753–4762.

[290] S. Sebastian, M.A.P. Manimekalai, Color image compression using JPEG2000 with adaptive color space transform, in: International Conference on Electronics and Communication Systems, Coimbatore, INDIA, 2014, pp. 261–267.

[291] S.M. Yavari, H. Amiri, Effect of shadow removal by gamma correction in SMQT algorithm in environmental application, Environ. Dev. Sustain. 22 (2020) 7057–7074.

[292] C. Puttaruksa, P. Taeprasartsit, Color data augmentation through learning color-mapping parameters between cameras, in: 15th International Joint Conference on Computer Science and Software Engineering, Mahidol Univ, Fac ICT, THAILAND, 2018, pp. 6–11.

[293] G. Kang, X. Dong, L. Zheng, and Y. Yang, "Patchshuffle regularization," *arXiv preprint arXiv:1707.07103,*2017.

[294] P. Singhal, A. Verma, A. Garg, A study in finding effectiveness of Gaussian blur filter over bilateral filter in natural scenes for graph based image segmentation, in: 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), 2017, pp. 1–6.

[295] S. Blessy, C.H. Sulochana, Enhanced Homomorphic Unsharp Masking method for intensity inhomogeneity correction in brain MR images, Comput. Methods Biomech. Biomed. Eng.-Imaging Vis. 8 (2020) 40–48.

[296] H. Inoue, "Data augmentation by pairing samples for images classification," *arXiv preprint arXiv:1801.02929,*2018.

[297] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, in: 6th International Conference on Learning Representations (ICLR), Vancouver CANADA, 2018, pp. 1–13.

[298] C. Summers, M.J. Dinneen, Improved mixed-example data augmentation, in: IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, USA, 2019, pp. 1262–1270.

[299] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, USA, 2020, pp. 13001–13008.

[300] J.I. Forcén, M. Pagola, E. Barrenechea, H. Bustince, Learning ordered pooling weights in image classification, Neurocomputing 411 (2020) 45–53.

[301] H. Goeau, A. Mora-Fallas, J. Champ, N.L.R. Love, S.J. Mazer, E. Mata-Montero, A new fine-grained method for automated visual analysis of herbarium specimens: a case study for phenological data extraction, App. Plant Sci. 8 (2020) 11.

[302] S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: a simple and accurate method to fool deep neural networks, in: IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, Nevada, USA, 2016, pp. 2574–2582.

[303] I. Tyukin, D. Higham, A. Gorban, On adversarial examples and stealth attacks in artificial intelligence systems, in: 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–6.

[304] V. Antun, F. Renna, C. Poon, B. Adcock, A.C. Hansen, On instabilities of deep learning in image reconstruction and the potential costs of AI, Proc. Natl. Acad. Sci. 117 (2020) 30088–30095.

[305] S. van Steenkiste, K. Kurach, J. Schmidhuber, S. Gelly, Investigating object compositionality in generative adversarial networks, Neural Netw. 130 (2020) 309–325.

[306] L. Talas, J.G. Fennell, K. Kjernsmo, I.C. Cuthill, N.E. Scott-Samuel, R.J. Baddeley, CamoGAN: Evolving optimum camouflage with Generative Adversarial Networks, Methods Ecol. Evol. 11 (2020) 240–247.

[307] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, Generative Adversarial Nets. Advances in neural information processing systems, 2014, pp. 2672–2680.

[308] S.Y. Liu, H.Y. Guo, J.G. Hu, X. Zhao, C.Y. Zhao, T. Wang, A novel data augmentation scheme for pedestrian detection with attribute preserving GAN, Neurocomputing 401 (2020) 123–132.

[309] P. Chaudhari, H. Agrawal, K. Kotecha, Data augmentation using MG-GAN for improved cancer classification on gene expression data, Soft Comput. 24 (2020) 11381–11391.

[310] C. Studholme, D.L. Hill, D.J. Hawkes, An overlap invariant entropy measure of 3D medical image alignment, Pattern Recognit. 32 (1999) 71–86.

[311] M. Jenkinson, P. Bannister, M. Brady, S. Smith, Improved optimization for the robust and accurate linear registration and motion correction of brain images, Neuroimage 17 (2002) 825–841.

[312] B.C. Vemuri, J. Ye, Y. Chen, C.M. Leonard, Image registration via level-set motion: applications to atlas-based segmentation, Med. Image Anal. 7 (2003) 1–20.

[313] P. Hellier, C. Barillot, Coupling dense and landmark-based approaches for nonrigid registration, IEEE Trans. Med. Imaging 22 (2003) 217–227.

[314] G. Postelnicu, L. Zollei, B. Fischl, Combined volumetric and surface registration, IEEE Trans. Med. Imaging 28 (2008) 508–522.

[315] M. Liu, D. Cheng, K. Wang, Y. Wang, A.s.D.N. Initiative, Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis, Neuroinformatics 16 (2018) 295–308.

[316] F. Alfano, J.O. Fisac, M. García-Sevilla, M.H. Conde, O.B. Zamora, S. Lizarraga, Prone to supine surface based registration workflow for breast tumor localization in surgical planning, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), 2019, pp. 1150–1153.

[317] B.D. de Vos, B.H. van der Velden, J. Sander, K.G. Gilhuijs, M. Staring, I. Išgum, Mutual information for unsupervised deep learning image registration. Medical Imaging 2020: Image Processing, 2020, 113130R.

[318] V. Bhavana, Medical image registration using landmark registration technique and fusion, Computational Vision and Bio-Inspired Computing: ICCVBIC 2019 1108 (2020) 402.

[319] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification, Neurocomputing 321 (2018) 321–331.

[320] J. Doak, "An evaluation of feature selection methods and their application to computer security," *Techninal Report CSE-92-18,*1992.

[321] M. Ben-Bassat, Pattern recognition and reduction of dimensionality, Handb. Stat. 2 (1982) 773–910.

[322] S. Khalid, T. Khalil, S. Nasreen, A survey of feature selection and feature extraction techniques in machine learning, in: 2014 Science and Information Conference, London Heathrow, UK, 2014, pp. 372–378.

[323] A. Janecek, W. Gansterer, M. Demel, G. Ecker, On the relationship between feature selection and classification accuracy. New Challenges for Feature Selection in Data Mining and Knowledge Discovery, 2008, pp. 90–105.

[324] L. Rangarajan, Bi-level dimensionality reduction methods using feature selection and feature extraction, Int. J. Comput. App. 4 (2010) 33–38.

[325] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[326] L. Goh, Q. Song, N. Kasabov, A novel feature selection method to improve classification of gene expression data, Proceed. Second Confer. Asia-Pacif. Bioinform. 29 (2004) 161–166.

[327] Z. Dai, C. Yan, Z. Wang, J. Wang, M. Xia, K. Li, Discriminative analysis of early Alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (M3), Neuroimage 59 (2012) 2187–2195.

[328] Y. Fan, D. Shen, R.C. Gur, R.E. Gur, C. Davatzikos, COMPARE: classification of morphological patterns using adaptive regional elements, IEEE Trans. Med. Imaging 26 (2006) 93–105.

[329] B. Mwangi, K.P. Ebmeier, K. Matthews, J.Douglas Steele, Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder, Brain 135 (2012) 1508–1521.

[330] R. Chaves, J. Ramírez, J. Górriz, M. López, D. Salas-Gonzalez, I. Alvarez, SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting, Neurosci. Lett. 461 (2009) 293–297.

[331] C. Chu, A.-L. Hsu, K.-H. Chou, P. Bandettini, C. Lin, A.s.D.N. Initiative, Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images, Neuroimage 60 (2012) 59–70.

[332] S.G. Costafreda, C.H. Fu, M. Picchioni, T. Toulopoulou, C. McDonald, E. Kravariti, Pattern of neural responses to verbal fluency shows diagnostic specificity for schizophrenia and bipolar disorder, BMC psychiatry 11 (2011) 18.

[333] J.H. Yoon, D. Tamir, M.J. Minzenberg, J.D. Ragland, S. Ursu, C.S. Carter, Multivariate pattern analysis of functional magnetic resonance imaging data reveals deficits in distributed representations in schizophrenia, Biol. Psychiatry 64 (2008) 1035–1041.

[334] V.D. Calhoun, J. Sui, K. Kiehl, J.A. Turner, E.A. Allen, G. Pearlson, Exploring the psychosis functional connectome: aberrant intrinsic networks in schizophrenia and bipolar disorder, Front. Psychiatry 2 (2012) 75.

[335] E.-G. Talbi, Metaheuristics: from Design to Implementation, 74, John Wiley & Sons, 2009.

[336] M. Mafarja, S. Mirjalili, Whale optimization approaches for wrapper feature selection, Appl. Soft Comput. 62 (2018) 441–453.

[337] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, E. Formisano, Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns, Neuroimage 43 (2008) 44–58.

[338] R.C. Craddock, P.E. Holtzheimer III, X.P. Hu, H.S. Mayberg, Disease state prediction from resting state functional connectivity, Magn. Reson. Med.: An Off. J. Int. Soc. Magn. Reson. Med. 62 (2009) 1619–1628.

[339] B. Mwangi, K.M. Hasan, J.C. Soares, Prediction of individual subject's age across the human lifespan using diffusion tensor imaging: a machine learning approach, Neuroimage 75 (2013) 58–67.

[340] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507–2517.

[341] S. Calderoni, A. Retico, L. Biagi, R. Tancredi, F. Muratori, M. Tosetti, Female children with autism spectrum disorder: an insight from mass-univariate and pattern classification analyses, Neuroimage 59 (2012) 1013–1022.

[342] C. Ecker, V. Rocha-Rego, P. Johnston, J. Mourao-Miranda, A. Marquand, E. M. Daly, Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach, Neuroimage 49 (2010) 44–56.

[343] M. Ingalhalikar, D. Parker, L. Bloy, T.P. Roberts, R. Verma, Diffusion based abnormality markers of pathology: toward learned diagnostic prediction of ASD, Neuroimage 57 (2011) 918–927.

[344] C. Davatzikos, Y. Fan, X. Wu, D. Shen, S.M. Resnick, Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging, Neurobiol. Aging 29 (2008) 514–523.

[345] D. Gothelf, F. Hoeft, T. Ueno, L. Sugiura, A.D. Lee, P. Thompson, Developmental changes in multivariate neuroanatomical patterns that predict risk for psychosis in 22q11. 2 deletion syndrome, J. Psychiatr. Res. 45 (2011) 322–331.

[346] E. Castro, M. Martínez-Ramón, G. Pearlson, J. Sui, V.D. Calhoun, Characterization of groups using composite kernels and multi-source fMRI analysis data: application to schizophrenia, Neuroimage 58 (2011) 526–536.

[347] K. Nho, L. Shen, S. Kim, S.L. Risacher, J.D. West, T. Foroud, Automatic prediction of conversion from mild cognitive impairment to probable Alzheimer's disease using structural magnetic resonance imaging, in: AMIA Annual Symposium Proceedings, 2010, p. 542.

[348] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc.: Ser. B (Methodological) 58 (1996) 267–288.

[349] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 67 (2005) 301–320.

[350] A.R. McIntosh, N.J. Lobaugh, Partial least squares analysis of neuroimaging data: applications and advances, Neuroimage 23 (2004) S250–S263.

[351] R. Tibshirani, Regression shrinkage and selection via the lasso: a retrospective, J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 73 (2011) 273–282.

[352] J. Yan, S.L. Risacher, S. Kim, J.C. Simon, T. Li, J. Wan, Multimodal neuroimaging predictors for cognitive performance using structured sparse learning. International Workshop on Multimodal Brain Image Analysis, 2012, pp. 1–17.

[353] M. Vounou, E. Janousova, R. Wolz, J.L. Stein, P.M. Thompson, D. Rueckert, Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease, Neuroimage 60 (2012) 700–716.

[354] O. Kohannim, D.P. Hibar, J.L. Stein, N. Jahanshad, X. Hua, P. Rajagopalan, Discovery and replication of gene influences on brain structure using LASSO regression, Front. Neurosci. 6 (2012) 115.

[355] E. Duchesnay, A. Cachia, N. Boddaert, N. Chabane, J.-F. Mangin, J.-L. Martinot, Feature selection and classification of imbalanced datasets: application to PET images of children with autistic spectrum disorders, Neuroimage 57 (2011) 1003–1014.

[356] O. Kohannim, D.P. Hibar, N. Jahanshad, J.L. Stein, X. Hua, A.W. Toga, Predicting temporal lobe volume on MRI from genotypes using L 1-L 2 regularized regression, in: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), 2012, pp. 1160–1163.

[357] F. Bunea, Y. She, H. Ombao, A. Gongvatana, K. Devlin, R. Cohen, Penalized least squares regression methods and applications to neuroimaging, Neuroimage 55 (2011) 1519–1527.

[358] J.O. Ogutu, T. Schulz-Streeck, H.-P. Piepho, Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions, in: BMC proceedings, 2012, p. S10.

[359] A. Rao, Y. Lee, A. Gass, A. Monsch, Classification of Alzheimer's Disease from structural MRI using sparse logistic regression with optional spatial regularization, in: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2011, pp. 4499–4502.

[360] J. Wan, S. Kim, M. Inlow, K. Nho, S. Swaminathan, S.L. Risacher, Hippocampal surface mapping of genetic risk factors in AD via sparse learning models, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2011, pp. 376–383.

[361] A.F. Marquand, O.G. O'Daly, S. De Simoni, D.C. Alsop, R.P. Maguire, S. C. Williams, Dissociable effects of methylphenidate, atomoxetine and placebo on regional cerebral blood flow in healthy volunteers at rest: a multi-class pattern recognition approach, Neuroimage 60 (2012) 1015–1024.

[362] A. Krishnan, L.J. Williams, A.R. McIntosh, H. Abdi, Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review, Neuroimage 56 (2011) 455–475.

[363] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.

[364] K. Chen, E.M. Reiman, Z. Huan, R.J. Caselli, D. Bandy, N. Ayutyanont, Linking functional and structural brain images with multivariate network analyses: a novel application of the partial least square method, Neuroimage 47 (2009) 602–610.

[365] J. Sui, T. Adali, Q. Yu, J. Chen, V.D. Calhoun, A review of multivariate methods for multimodal fusion of brain imaging data, J. Neurosci. Methods 204 (2012) 68–81.

[366] L. Menzies, S. Achard, S.R. Chamberlain, N. Fineberg, C.-H. Chen, N. Del Campo, Neurocognitive endophenotypes of obsessive-compulsive disorder, Brain 130 (2007) 3223–3236.

[367] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint arXiv: 1404.1100,*2014.

[368] P. Alvarado-Alanis, P. León-Ortiz, F. Reyes-Madrigal, R. Favila, O. Rodríguez-Mayoral, H. Nicolini, Abnormal white matter integrity in antipsychotic-naive first-episode psychosis patients assessed by a DTI principal component analysis, Schizophr. Res. 162 (2015) 14–21.

[369] P.-R. Loh, G. Bhatia, A. Gusev, H.K. Finucane, B.K. Bulik-Sullivan, S.J. Pollack, Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis, Nat. Genet. 47 (2015) 1385.

[370] L. Khedher, J. Ramírez, J.M. Górriz, A. Brahim, F. Segovia, A.s.D.N. Initiative, Early diagnosis of Alzheimer׳ s disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images, Neurocomputing 151 (2015) 139–150.

[371] L.C. Paul, A. Al Sumam, Face recognition using principal component analysis method, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 1 (2012) 135–139.

[372] A.B. Bendixen, C.B. Hartberg, G. Selbæk, K. Engedal, Symptoms of anxiety in older adults with depression, dementia, or psychosis: a principal component analysis of the geriatric anxiety inventory, Dement. Geriatr. Cogn. Disord. 42 (2016) 310–322.

[373] K. Franke, G. Ziegler, S. Klöppel, C. Gaser, A.s.D.N. Initiative, Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters, Neuroimage 50 (2010) 883–892.

[374] Y. Wang, Y. Fan, P. Bhatt, C. Davatzikos, High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables, Neuroimage 50 (2010) 1519–1535.

[375] L.K. Hansen, J. Larsen, F.Å. Nielsen, S.C. Strother, E. Rostrup, R. Savoy, Generalizable patterns in neuroimaging: How many principal components? Neuroimage 9 (1999) 534–544.

[376] B. Mwangi, T.S. Tian, J.C. Soares, A review of feature reduction techniques in neuroimaging, Neuroinformatics 12 (2014) 229–244.

[377] A.N. Gorban, B. Kégl, D.C. Wunsch, A.Y. Zinovyev, Principal Manifolds for Data Visualization and Dimension Reduction, 58, Springer, 2008.

[378] V.D. Calhoun, J. Liu, T. Adalı, A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data, Neuroimage 45 (2009) S163–S172.

[379] V.D. Calhoun, T. Adali, Unmixing fMRI with independent component analysis, IEEE Eng. Med. Biol. Mag. 25 (2006) 79–90.

[380] N. Correa, T. Adalı, V.D. Calhoun, Performance of blind source separation algorithms for fMRI analysis using a group ICA method, Magn. Reson. Imaging 25 (2007) 684–694.

[381] P.K. Douglas, S. Harris, A. Yuille, M.S. Cohen, Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief, Neuroimage 56 (2011) 544–553.

[382] J.R. Sato, M.Q. Hoexter, A. Fujita, L.A. Rohde, Evaluation of pattern recognition and feature extraction methods in ADHD prediction, Front. Syst. Neurosci. 6 (2012) 68.

[383] E.P. Duff, A.J. Trachtenberg, C.E. Mackay, M.A. Howard, F. Wilson, S.M. Smith, Task-driven ICA feature generation for accurate and interpretable prediction using fMRI, Neuroimage 60 (2012) 189–203.

[384] A. Hyvarinen, H. Morioka, Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. Advances in Neural Information Processing Systems, 2016, pp. 3765–3773.

[385] C. Zhao, Y. Wang, F. Mei, Kernel ICA feature extraction for anomaly detection in hyperspectral imagery, Chin. J. Electron. 21 (2012) 265–269.

[386] S.B. Eickhoff, A.R. Laird, C. Grefkes, L.E. Wang, K. Zilles, P.T. Fox, Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty, Hum. Brain Mapp. 30 (2009) 2907–2926.

[387] T.D. Wager, M. Lindquist, L. Kaplan, Meta-analysis of functional neuroimaging data: current and future directions, Soc. Cogn. Affect. Neurosci. 2 (2007) 150–158.

[388] D.W. Scott, Kernel Density Estimators. Multivariate Density Estimation, 2015, pp. 137–216.

[389] A.R. Laird, K.M. McMillan, J.L. Lancaster, P. Kochunov, P.E. Turkeltaub, J. V. Pardo, A comparison of label-based review and ALE meta-analysis in the Stroop task, Hum. Brain Mapp. 25 (2005) 6–21.

[390] T. Yarkoni, R.A. Poldrack, T.E. Nichols, D.C. Van Essen, T.D. Wager, Large-scale automated synthesis of human functional neuroimaging data, Nat. Methods 8 (2011) 665–670.

[391] T.M. Mitchell, From journal articles to computational models: a new automated tool, Nat. Methods 8 (2011) 627–628.

[392] J. Dukart, K. Mueller, H. Barthel, A. Villringer, O. Sabri, M.L. Schroeter, Meta-analysis based SVM classification enables accurate detection of Alzheimer's disease across different clinical centers using FDG-PET and MRI, Psychiatry Res.: Neuroimaging 212 (2013) 230–236.

[393] G. Salimi-Khorshidi, S.M. Smith, J.R. Keltner, T.D. Wager, T.E. Nichols, Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies, Neuroimage 45 (2009) 810–823.

[394] R. Casanova, C. Whitlow, B. Wagner, M. Espeland, J. Maldjian, Combining graph and machine learning methods to analyze differences in functional connectivity across sex, The Open Neuroimaging J. 6 (2012) 1.

[395] I. Rish, G.A. Cecchi, M.N. Baliki, A.V. Apkarian, Sparse regression models of pain perception, in: International Conference on Brain Informatics, 2010, pp. 212–223.

[396] S.G. Costafreda, C. Chu, J. Ashburner, C.H. Fu, Prognostic and diagnostic potential of the structural neuroanatomy of depression, PLoS One 4 (2009) e6353.

[397] E.A. Allen, E.B. Erhardt, E. Damaraju, W. Gruner, J.M. Segall, R.F. Silva, A baseline for the multivariate comparison of resting-state networks, Front. Syst. Neurosci. 5 (2) (2011).

[398] J. Mourão-Miranda, L. Oliveira, C.D. Ladouceur, A. Marquand, M. Brammer, B. Birmaher, Pattern recognition and functional neuroimaging help to discriminate healthy adolescents at risk for mood disorders from low risk adolescents, PLoS One 7 (2012) e29482.