Big Data Analytics in Medical Imaging using Deep Learning

Amirhessam Tahmassebi[a,*], Anahid Ehtemami[b], Behshad Mohebali[a], Amir H. Gandomi[c], Katja Pinker[a,d,e], and Anke Meyer-Baese[a]

[a]Department of Scientific Computing, Florida State University, Tallahassee, Florida, USA
[b]Department of Electrical and Computer Engineering, FAMU-FSU College of Engineering, Tallahassee, Florida, USA
[c]School of Business, Stevens Institute of Technology, Hoboken, New Jersey, USA
[d]Department of Radiology, Breast Imaging Service, Memorial Sloan Kettering Cancer Center, New York, USA
[e]Department of Biomedical Imaging and Image-Guided Therapy, Division of Molecular and Gender Imaging, Medical University of Vienna, Vienna, Austria

ABSTRACT

Big data has been one of the hottest topics of scientific discussions in the recent years. In early 2000s, an industry analyst attempted to describe big data as the three Vs: Volume, Velocity, and Variability. With the new technologies such as Hadoop, it is now feasible to store and use extremely large volumes of data that comes in at an unprecedented velocity. The variability of this data can be large as it can come in different formats such as text documents, voice or video, and financial transactions. Big data analytics has been proven to be useful is various fields such as science, sports, advertising, health care, genomic sequence data, and medical imaging. This study presents a brief overview of big data analytics in medical imaging approaches with considering the importance of contemporary machine learning techniques such as deep learning.

Keywords: Big Data, Deep Learning, Medical Imaging, Image processing, Optimization

1. INTRODUCTION

Machine learning particularly pays attention to learning patterns based on the raw data to predict future unseen data. The quality of the data plays an essential role in the performance of the machine learning algorithms. In principle, pre-processing is required to transform the raw data into a good representation of the data with extracting salient features to learn patterns for future unseen data. This is called feature engineering. Feature engineering has a large impact to produce a high-performance machine learning model with a lower complexity. In better words, the more salient features we produce, the simpler machine learning model will be.[1,2]

Big Data as one of the inseparable high-focus of data science has become more important for large companies such as Google, Microsoft, IBM, and Amazon as in an abstract statement of the rise of big data based on Moore's law, "world's data doubling every year". Big Data is associated with the "V-V-V-V" or "4Vs" concept: 1) Volume, 2) Variety, 3) Velocity, and 4) Veracity which brings more data analytics challenges.[3] Big data has been one of the hottest topics of scientific discussions in the recent years. Processing this massive volume of data in order to find out the hidden patterns and correlations is called big data analytics.[4] Big data analytics has been proven to be useful is various fields such as science, sports, and advertising.[5] It has also been utilized in health care practices; aiding care-delivery and disease exploration. Researchers have been using large volumes of images, signals, and genomics data, individually or combining multimodal data from different sources to achieve meaningful results.[6]

The use of technology in medicine has been tremendously increased in the past decades as new techniques are developed. Medical imaging is visualizing the oragans morphology to aid clinical diagnosis, treatments,

* Corresponding Author: Amirhessam Tahmassebi
E-mail: atahmassebi@fsu.edu
URL: http://www.amirhessam.com

and monitoring the treatment response. The analysis of medical imaging is an interdisciplinary research that has revolutionized the abilities of our health-care providers. The large number of health-care organizations and patients has lead us to develop and use more computer-aided medical diagnostics and decision support systems.[7] Medical image processing and analytics improve the interpretability of the contents. In addition to analytical methods, collecting, sharing, and compressing techniques are important in handling medical imaging data.

Most of the challenges of medical imaging analysis are similar to the ones from any kind of big data analytics. However, there are several steps that are added mainly due to the type of the captured data in each individual analysis. Big data are characterized by high dimensionality and often large sample sizes which result in specific challenges that include noise accumulation, spurious correlations, incidental homogeneity, computationally expensive models, and algorithm instability. Many algorithms and methods perform well for moderate sized datasets, but they fail to cope with rapid increase of dimensionality.[8]

2. TYPES OF MEDICAL DATA

Medical data (images) are produced by interaction of different forms of radiation with tissue and it can range from a simple chest X-ray to more complex images produced by functional magnetic resonance imaging (fMRI). Different techniques of medical imaging such as radiology, nuclear medicine, or optical imaging, provide images with different spatial and temporal resolutions.

2.1 Radiography

X-ray is a form of electromagnetic radiation which consists of photons. It was discovered by Wilhelm Konrad Rongtgen while he was studying cathode tubes. He found out that the tube was emitting light as well as a new mysterious kind of radiation which he called X-rays. Soon he discovered that this radiations can travel through different material and also be captured on a photographic plate. Before long, x-rays were being used for medical purposes.[9] The resolution of the images produced by radiographic systems depends on several parameters including the size of the focal spot, thickness of the body part, and the light scattering properties of the fluorescent screen.[10] X-ray photons by nature carry some quantum noise. The noise amplitude is corresponding to the square root of the signal amplitude and the signal-to-noise ratio (SNR) behaves as the square root of the signal amplitude. Therefore, dose reduction is not unpunished in image quality. Some conversions during imaging process also add noise and reduce the SNR.[11]

2.2 X-ray Computed Tomography

Computed tomography (CT) became feasible only after the development of modern computers. It is a tool that reconstructs images from measured data citect1. Tomographic imaging consists of capturing x-ray images of an object from multiple orientations and measuring the decrease in intensity along several linear paths. Then, an algorithm reconstructs the distribution of X-ray attenuation in the volume that is scanned.[12] CT data consists of a sequence (thousands) of images which can be visualized using different 2D or 3D image processing tools. Volume rendering and isosurfacing are the two standard modes of 3D visualization of CT data. CT values are the gray-level numbers in images. Volume rendering involves mapping each CT value to a color and an opacity. Some phases can be rendered transparent, revealing the internal structure. Isosurfacing is defining 3D contour surfaces distinguishing the boundaries between CT numbers, separating the elevation values on a topo map.[13]

2.3 Magnetic Resonance Imaging (MRI)

In magnetic resonance imaging (MRI), a powerful magnet to generate images that cannot be captured using X-rays or CT such as joints, cartilage, ligaments, and tendons.[14] The MRI machine is used to create a static electromagnetic field to align the proton spins of oxygen atoms in blood. A short radio frequency wave reorients the nuclei of the atoms and the atoms absorb this energy. When the interfering wave stops, the protons gradually returns to their aligned spin and release the energy that is stored in them. This produces a radio signal that is measured by the scanner and interpreted into images. Protons in different tissues generate different signals which is used to distinguish various types of tissue.[15] The MRI data is captured at a very high rate (multiple slices per second). The trade-off for this high speed is a low spatial resolution. MRI data also suffers from a variety of distortions because of the effect of magnetic field inhomogeneities. Furthermore, unwanted motions

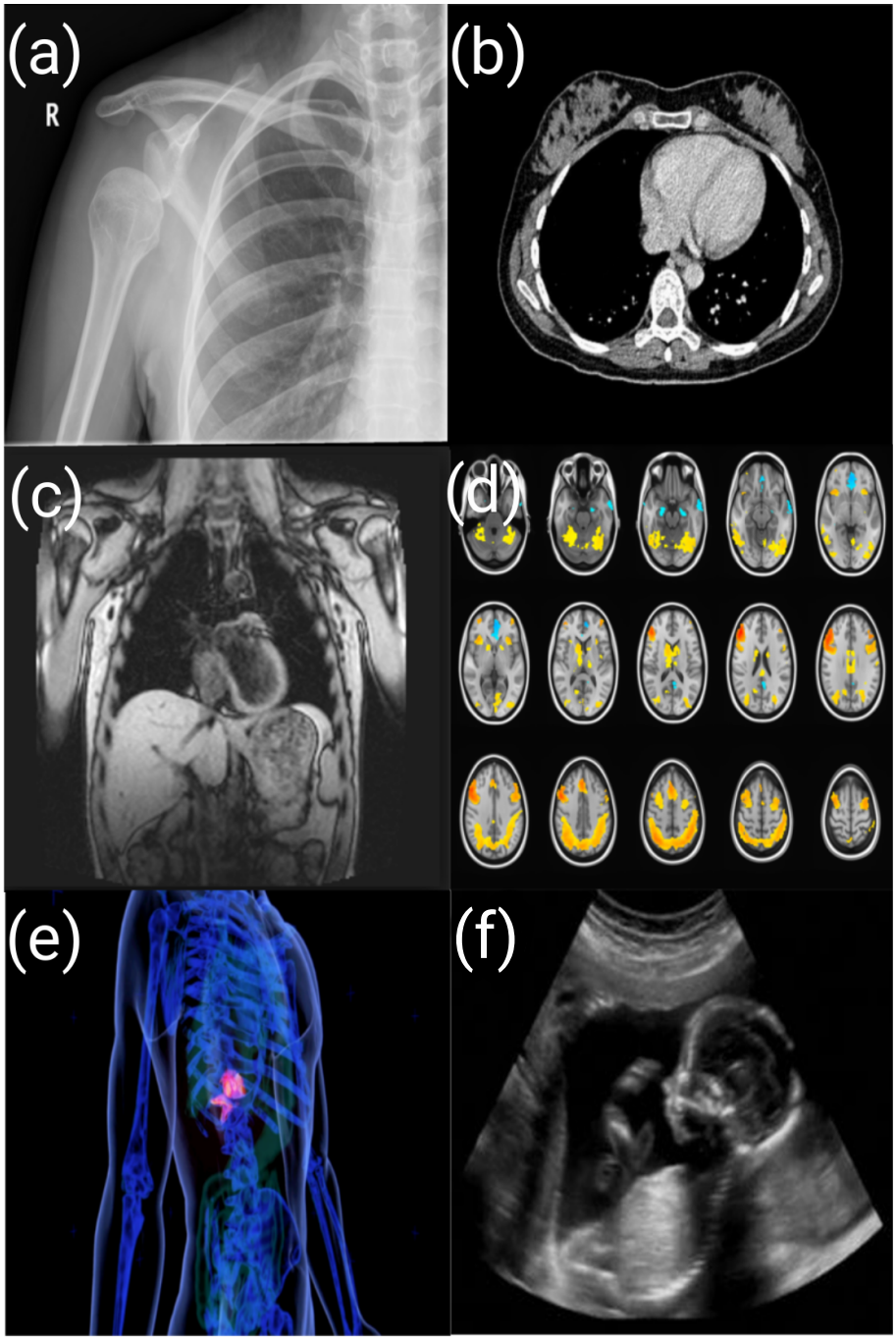

Figure 1. (a) An inferior shoulder dislocation radiology. (b) CT lung cancer screening. (c) Chest MRI. (d) fMRI brain scans of functional connectivity. (e) A sample nuclear medicine image of torso on the human body. (f) Fetal ultrasound.

during scans are the other substantial source of noise in MRI data sets. Even slight movements can result in a big change in the captured signal. For this reason, realignment is often one of the primary step in analyzing this data.[16]

2.4 Functional MR Imaging (fMRI)

Regardless of any external stimuli, live brains show activity unceasingly. The neurons with higher activity consume more oxygen. Functional MRI is a non-invasive method that locates and measures the fluctuations in blood-oxygen-level dependent fluctuations and provides a map of the functional connectivity in brain.[17] The signal that is measured is complex valued. Both real and imaginary components are measured with independent error that is normally distributed. The reconstructed voxel data is also complex valued since Fourier Transform is a linear operation. In most studies, the phase portion is discarded and only magnitude is used since it carries most of the useful information. It is important to know the behavior of the signal and noise presented in fMRI data to be able to property model the components.[18] Neural activity unfolds in time and space. Therefore, spatial and temporal resolution of data can result in some limitations in deriving conclusions. Temporal resolution aids distinguishing brain events in time and spatial resolution, across spatial locations. The nature of fMRI experiments prevent having ideal spatial and temporal resolution at the same time. Therefore, it is important to find the perfect balance between the spatial and temporal resolution requirements for each specific experiment.[19] MRI and fMRI can use either Arterial Spin Labeling (ASL) or Blood Oxygen Level Dependent (BOLD) signals. The method is selected based on the required sensitivity and other specifications of the experiment.[20]

2.5 Nuclear Medicine Imaging

The method of observing the radiation from different parts of the body after a radioactive tracer is injected or orally given to the patient, is nuclear medicine imaging which is used for observing tumors, infections, and thyroid or bone scintigraphy. It is ensured that the radiation exposure to patients is as low as possible. The two common types of nuclear medicine imaging are Single Photon Emission Computed Tomography (SPECT) and Positron Emission Tomography (PET). PET and SPECT both produce three dimensional images and the main difference is in the radiotracer that is used in the process. Comparing to SPECT, PET is more costly but it produces better contrast and spatial resolution. PET data can be captured dynamically or statically. Dynamic acquisition lets us observe the long-term behavior of the tracer in the tissue and is a great way to get quantitative measurements of the target area. Static acquisition provides semi-quantitative information and it works by specifying one time frame over the course of imaging. Static images can also be obtained from dynamic data by finding the average of radioactivity over a set of time frames.[21] In SPECT, the camera moves around the patient and the images are captured from at least 180 degrees. After the scan, reconstruction is done by filtered-back projection methods. The images are viewed in the transverse, sagittal or coronal planes or as three dimensional models. The useful property of SPECT is that the reconstructed images can be viewed in multiple planes and it is possible to separate overlapping structures.[22]

2.6 Ultrasound

Ultrasound is a widely available, safe, and non-invasive method for producing real-time images of the structures inside of the body or the blood flow, by using sound waves. In ultrasound scanning or sonography, high frequency sound waves are transmitted into the body and the transducer collects the reflected signal to create an image. Ultrasound can produce images of thin sections of the body. However, it is possible to create three dimensional images from the acquired data.

3. HIGH PERFORMANCE COMPUTING

As discussed, medical images include a wide spectrum of different image acquisition methodologies for different purposes. They can vary from two, three, to four dimensional images. Some modern techniques can provide very high resolution data. Higher resolution and dimensions result in the volume of data to grow exponentially which requires advanced analytical methods and high performance computing.[6] Integrating images from different modalities can also be done to provide more information and better accuracy in diagnosis.

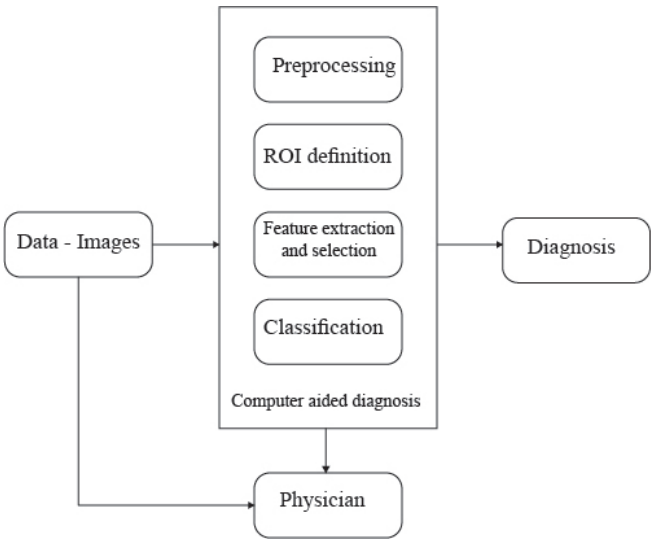

Figure 2. Computer aided diagnosis using ROIs.

3.1 Compression Techniques

For transmission and storage efficiency, it is vital to reduce the data volume to a manageable level by eliminating the redundancies and irrelevant information. It is important that compression does not degrade the quality of the images and reduce the useful information that can be obtained from them. Compression techniques are lossy or lossless. In lossless compression, unlike lossy, the entire image is preserved. Lossy compression techniques are acceptable in some field such as natural photography, where some limited loss of details is acceptable even though it might not be visually noticeable.[23]

Lossless medical image compression techniques include:

- Lossless JPEG (Using Huffman or arithmetic entropy code)
- JPEG-LS (Providing low complexity lossless image compression using entropy coding and LOCO-I algorithm)
- JPEG-2000 (For high quality final result, using wavelet transform image decomposition and arithmetic coding)
- PNG (Using LZ77 and Huffman encoding)
- CALIC (High compression ratio using arithmetic entropy codes)[24]

3.2 Cloud Computing & Parallelization

With big data, comes the need for more efficient storage methods than the traditional database management systems (DBMS). MapReduce started the trend of developing simple and scalable data processing systems that is used for big data and now is the most actively researched big data processing system.[25] Based on MapReduce now there are different systems that offer more flexibility and have other advantages. Hadoop is an open source implementation of Google MapReduce that is very commonly used outside Google. The two main components of Hadoop are: Hadoop distributed file system (HDFS) and MapReduce which is the processing component. Hadoop has Master-Slave architecture. Slave nodes (data nodes) are responsible for processing read and write requests from the file system clients. The master node (name node, job tracker) handles the file system namespace operations such as opening and closing files and renaming directories.[26] Redundancy is important to avoid failures. That is why there are three master nodes.[27] The MapReduce framework works with key-value

Table 1. Characteristics and targets of automatic clouds.

Characteristics	Target
Resources/services variability	Auto Scaling Load balancing Scheduling Adaptive resource provisioning
Contextual behavior	Self healing Self configuration Automatic pricing mrchanism
Easy deployment and reliability management	Reliability Optimal deployment selection Composition Discovery migration

pairs. First, it maps the user functions and by using the pairs as input to distributed machines, generates key-value pairs. After that in the reduce phase, the intermediate pairs are reduced to single results. MapReduce provides fault tolerance by partitioning the data into several splits and storing each split on a different machine. MapReduce simplifies the process of parallelization for users. It can work with different data formats and can be scaled to several thousand processors.[28]

Currently, the common disk storage capacity goes up to several Terabytes but soon it will be Exabytes and Zettabytes and more. Cloud computing systems provide high performance and scalable data storage which are necessary for processing and sharing big data. Clouds are networks of servers with different functions that can store the data and offer higher level services. Cloud storage facilitates big data mining and collection. However, with cloud, there are security and privacy issues that need to be managed. Most distributed frameworks like MapReduce, do not provide security protections. Security checks need to be performed in real time in order to prevent unauthorized mappers and protect the data. This can sometimes be less feasible because of the massive volume of the data being generated constantly. Encryption is necessary for access control methods because data storage devices are vulnerable. Inability to encrypt the data during the tagging on logging of data or distributing it is another vulnerability by some data stores that can cause security threats.

Apache Spark and Apache Arrow, which are fast and general engines for large-scale data processing were chosen to expedite the running process of the proposed pipeline.[29,30] Spark began life in 2009 as a project within the AMPLab at the University of California, Berkeley Spark engines can run programs 100x faster than Hadoop MapReduce in memory and 10x faster on disk. Since the proposed pipeline was written in Python, PySpark was used to implement the Spark context. Figure 3 depicts the data flow in PySpark in details. PySpark to create a Spark context used Py4J which is a BSD licensed Java collection method that enables Java programs to call back Python objects and create a Java Virtual Machine (JVM) which is an abstract computing machine that enables a computer to run a Java program. In addition to this, Spark is compatible with most of the Python libraries such as Pandas. In fact, the codes can be written using Pandas API and converted into Spark format. This process can also be done while loading the data into pipeline by choosing Parquet or Spark formats in schema structure other than common formats such as Comma Separated Values (CSV).[1]

4. ANALYTICS

4.1 Preprocessing

Medical images, similar to most real-world data, suffer from issues that if not treated, can increase the inaccuracy of the results of analyzing the images. Contrast adjusting, noise reduction, physiological artifacts removal, and handling the missing data are some of the reasons for performing preprocessing steps, prior to analysis, in order to validate model assumptions.[18]

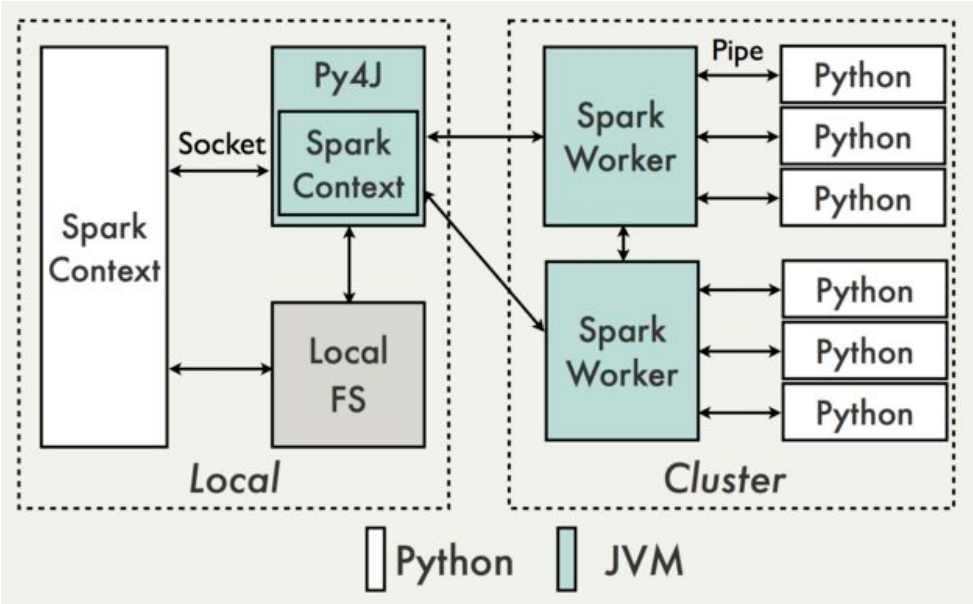

Figure 3. An illustration of PySpark data flow. The Python environments were shaded in white and the Java environments were shaded in blue.

4.2 Segmentation

Segmentation is partitioning the image into sets of regions to extract the areas of interest. Regions can be defined by a particular shape, border, color, or texture. Classical clustering methods that perform segmentation by finding pixels similar in intensity values, RGB values, texture, and more, include Iterative K-means and Isodate clustering. Histogram methods and different variations of it such as Ohlander's Recursive Histogram technique assume that the homogeneous objects in the image can be extracted as clusters on the histogram.[31] Region growing is the other method for segmentation. In this method, the algorithm starts from one point in the image (usually the top left corner) and grows the region until the pixels are too different from the current region and form a set of connected pixels with same population mean and variance.[32] The other type of region-based segmentation algorithm is Threshold Segmentation which directly ddivides the gray scale information based on the value of different targets. For images that include several touching objects, the Watershed Segmentation methods could be applied. The watershed transform seeks catchment basins and watershed ridge lines in an image to distinguish between foreground and background and the region that each pixel belongs to.[33]

4.3 Region of Interest (ROI)

ROI analysis involves extracting the signal from specified regions by selecting clusters of pixels or voxels in the image and it can be used for analysis withing one subject or across multiple ones. Using ROI techniques reduces the type I errors that can occur in analysis, by limiting the number of statistical tests to a few ROIs.[34] In analyzing medical images, loss of data results in loss of vital information. ROI is mainly used for medical images. Each image is divided into two parts, foreground; the areas that carry diagnostically important information, and background; the rest of the image. To preserve the quality of the diagnostic part, lossless compression techniques are favorable. Additionally, the diagnostic part has higher priority in transmissions.[35] Exploratory ROIs are spheres of the same diameter, placed at the local maxima in the statistical map. The locations of the ROIs can be selected based on anatomical templates such as Talairach atlas for brains, or functionally based on the data from images produced by techniques such as fMRI, or based on previous studies.[34] For different types of medical images, there are several tools that help with placing and analyzing ROIs such as SPM by Wellcome Trust Centre for Neuroimaging for brain images or Matlab for different types of medical images. After extracting all the ROI coordinates, a value should be calculated for each point of interest. The simplest way of calculating this value is finding the mean for each point. However, the mean can be easily affected by outliers in which case

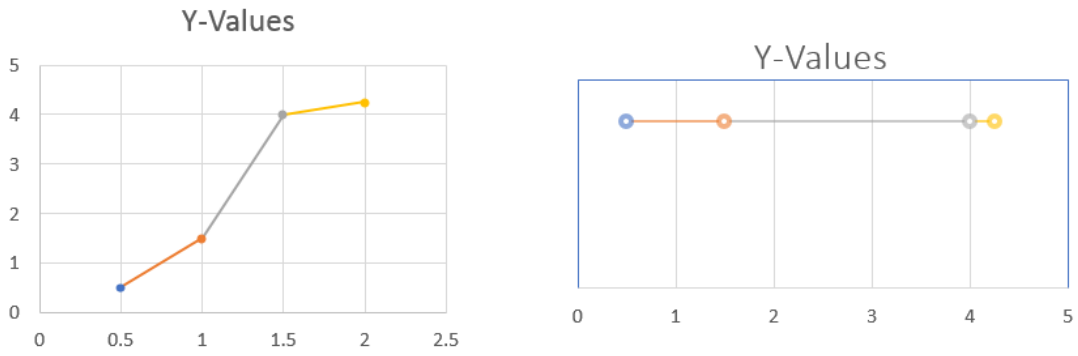

Figure 4. PCA Transformation of a 2D (left) to a 1D line graph (right) - In this dataset, the value of X is barely informative since the distances are regular. By removing the X values and projecting the Y values onto a 1D chart, we can oberve the variances more visibly.

the median is more suitable as a summary function. Weighted mean is another function that can be used in the calculations.

4.4 Feature Extraction

Features are information extracted from ROIs in images that are most representative of the data than can be fed into machine learning models. Using features reduces redundancies as well as decreasing the dimensionality where the computational cost of the model is high.[36] The algorithms that extract features are also known as image descriptors.[37]

In medical image analysis, features can be selected using traditional (statistical) or biology based methods. The method should be selected according for each specific application. Measured features can include: eigenvalues, energy, phase, amplitude, moments. Structural images include peaks, lines, edges, and other parametric models. In general, image descriptor algorithms can be based on intensity, shape, or texture. When there is a high correlation in images (e.g. between voxels in each area) transformation functions can be used to extract the relevant information. Principal component analysis and discrete Fourier transform are some examples of these tranformation functions.

4.4.1 Principal Component Analysis (PCA)

In PCA, the data is linearly transformed to a new coordinate system. These projection of the data onto the lower dimension system is done such that the higher variances are emphasized.[38] The first principal component contains the highest variability in the data and is the most informative component. PCA is also a method of data compression due to dimensionality reduction. The linear combination of these orthogonal components represents the entire data with minimal loss of information.[39–41]

4.4.2 Fourier Transform

The Fourier Transform is suitable for image processing including filtering, compression, and reconstruction, to decompose the image into sine and cosine components which represent the image. The Discrete Fourier Transform (DFT) provides a sample of all frequencies in the image that is large enough to fully represent the geometric characteristics of a spatial domain image. DFT can provide a good representation of signal changes and behavior for discrete time signals.[42] The characteristics that change with time cannot be represented using DFT since it can only be used for slices (windows) of the signals that have a fixed time duration.[43]

5. WHY DEEP LEARNING?

Over the past twenty years, advances in medical imaging with employing novel methods including convolutional neural networks (CNN) to learn directly from input image data have shown numerous success in various medical imaging tasks including microcalcification on digital mammograms.[44] Deep learning algorithms have the ability to sap salient information as high-level complex features via stacking multiple layers. As the network goes deeper, the features will be more complex. However, the combination of big data analytics and deep learning algorithms in the context of medical imaging is barely tapped. Traditionally, deep learning algorithms and neural networks have been used to extract features or in better works to reduce the dimensionality and classical machine learning algorithms such as linear support vector machines were used to finish the detection task. The most essential step towards linking deep learning algorithms into big data analytics of medical data is to remove human knowledge in parsing any substantial information and move towards artificial intelligence. This would require extracting more features and obviously more data is needed which brings up the issue that the number of possible combinations of the features is exponentially related to the number of extracted features.[2] In this section, we address only a few examples of the essential challenges based on the most recent studies in which deep learning algorithms were used to improve the results. In addition to this, the most recent technologies available to implement deep learning algorithms including libraries, software, optimization algorithms, and methods to tune the deep learning hyper-parameters are also presented.

In order to plan the diagnosis and therapy process of prostate cancer, analysis of T_2-weighted MRIs is required. The high variability of prostate morphology along with artifacts and noisy signals around the boundaries makes the segmentation process arduous. Therefore, employing the traditional algorithms for segmenting the prostate MRIs such as manual contouring different views including axial, sagittal, and coronal of prostate slice by slice would not lead to gain high accuracy. In addition to this, this process is time-consuming and labor-intensive. Thus, implementing an automated segmentation of the 3D T_2W MRIs with higher accuracy and robust to noise and artifacts around the boundaries, heterogeneous intensity distribution of the prostate MRI, and the complex morphological structure within and surrounding the prostate would open new avenue in segmentation process of prostate cancer research. Cheng et al[45] have proposed a multi-blocks pipeline for automatic segmentation using deep learning. In this pipeline, as the first block, the AlexNet deep CNN model was employed to refine the prostate boundaries. This model contains five 2D convolutional layers along with three fully connected layers. In order to overcome over-fitting, dropout was employed after each fully connected layer. It should be noted that it is required to have fixed patch sizes (64×64) in order to implement the pipeline using the AlexNet.[46] In addition to patch-based CNN with AlexNet, they implemented a holistically nested network (HNN) to learn the prostate interior image-labeling map for segmentations. HNN helps to efectively learn the edges and the prostate boundaries using both deep and fully connented architures.

In order to incorporate low-dose CT in lung cancer clinical situations, a suitable de-noising approach is needed. Various studies have shown that, the results gained from noisy images of low-dose CT cannot be reliable normally.[47,48] Nishio et al.[47] have employed a convolutional auto-encoder instead of general denoising auto-encoder. Denoising auto-encoder (DAE) is a special type of 3-layered neural network comprising an input layer, one-hidden layer, and an output layer. It adds a Gaussian noise to the original input to build a pair of original and noisy input. Then, DAE transforms the noisy inputs using an activation function into another subspace. Then, the weights that represent the noisy inputs are used to reconstruct the original input. It should be noted that the DAE does not preserve the locality of the 2D images. This obstacle is solved by using convolutional auto-encoder (CAE) which has the same architecture except CAE receives 2D images as its inputs. CAEs are more advantageous with respect to DAEs since they can be implemented on any CT scanner without using raw CT data. Although it requires to use deep layers of network to gain better accuracy (less error) which needs good computational equipment and solid knowledge of parameters optimization.[49]

Deep learning models involve optimization in many contexts. The essential idea of optimization is finding the best weights at each layer of the topology of the deep neural networks with the hope that these weights decrease the cost on the entire training set. Optimization by itself is an arduous and time-consuming problem and that becomes more challenging for difficulties in convex problems such as ill-conditioning, local minima, saddle points, and steep cliff and non-convex problems including deep neural networks. To consider the previously mentioned challenges, various optimization strategies including grid-search and advanced strategy such

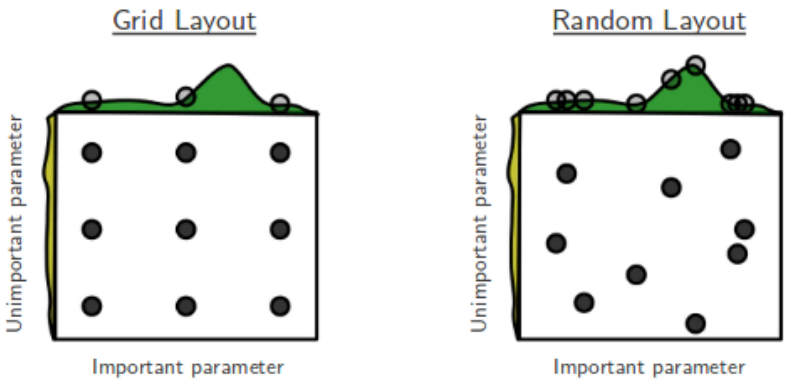

Figure 5. An illustration of grid-search and random-search layouts.

as Bayesian optimization along with several optimization algorithms including basic algorithms and algorithms with adaptive learning rates for the training phase of the deep models are available that can be found in the supplementary algorithms section. However, it is previously shown that the fine-tuned network architectures such as VGGNet16 which is an ImageNet pre-trained model can be used to initialize the deep architecture with better performance.[50]

Deep learning models as discussed require the determination of various hyper-parameters including the number of layers, the number of neurons per layer, activation functions, weight penalty, learning rate, momentum, and etc.[1] The simplest way would be an exhaustive grid-search over specified parameter values for the deep model. In this way, all the possible combinations of the specified hyper-parameters will be checked.[51,52] This process requires a good amount time and responsive computing resources as discussed in cloud computing section.[53] It should be noted that this burden can be reduced by using a random search over the specified region instead of an exhaustive grid-search with getting roughly the best results. Figure 5 presents an illustration of grid-search and random search for hyper-parameters tuning. In principle, in random search not all parameter values would be tried out, but rather a fixed number of parameter settings will be sampled from the specified distributions. This is quite interesting since not all the alternative values in each list for hyper-parameters play an important role in the outcomes. Therefore, by random sampling, the most important hyper-parameters can be determined and the other hyper-parameters settings can be fixed the same as before. In this way, the same results would not be replicated anymore and the learning slope will be positive. To overcome over-fitting depending on the size of the data k-folds cross-validation can be employed.

Employing machine learning to predict what combinations are likely to work well could help to rescue from the huge computational time. It requires to predict the regions of the hyper-parameter space that might give better outcomes. It also requires to predict how well a new combination will do and model the uncertainty of that prediction using Gaussian Process models. Gaussian processes (GPs) provide a principled, practical, and probabilistic approach in machine learning. GPs simply have an essential assumption that similar inputs give similar outputs. This simple and weak prior are actually very sensible for the effects of hyper-parameters. GPs are able to learn for each input dimension what the appropriate scale is for measuring similarity. GPs predict a Gaussian distribution of values rather than just predicting a single value. Bayesian optimization is a constrained global optimization approach built upon Bayesian inference and Gaussian process models to find the maximum value of an unknown function in the most efficient ways (less iterations).[1]

6. SUPPLEMENTARY ALGORITHMS

Algorithm 1: Bayesian Optimization

Input: Training data S
Output: Posterior distribution

for $i \in \{1, 2, \ldots, n\}$ **do**
  Find x_i by optimizing the acquisition function over the $GP : x_i = argmaxu(x|S_{1:i-1})$;
  Sample the objective function: $y = f(x_i) + \epsilon_i$;
  Augment the data $S_{1:i} = \{S_{1:i-1}, (x_i, y_i)\}$;
  Update the GP;
end

Algorithm 2: SGD

Input: Training data S, learning rate η, weights w
Output: Updated weights w

$w \leftarrow w_0$;
while *stopping criterion is not met* **do**
  Randomly shuffle the training data S ;
  Sample a minibatch of size m:$\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\} \in S$;
  for $i \in \{1, \ldots, m\}$ **do**
    $\hat{G} \leftarrow \frac{\partial}{\partial w_i} cost(w, (x^{(i)}, y^{(i)}))$; Gradient calculation
  end
  $w \leftarrow w - \eta \hat{G}$;
end

Algorithm 3: Adagrad

Input: Training data S, learning rate η, weights w, fuzz factor ϵ, learning rate decay over each update r
Output: Updated weights w

$\epsilon \leftarrow \epsilon_0 \approx 10^{-8}$;
$w \leftarrow w_0$;
$r \leftarrow 0$;
while *stopping criterion is not met* **do**
  Randomly shuffle the training data S ;
  Sample a minibatch of size m:$\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\} \in S$;
  for $i \in \{1, \ldots, m\}$ **do**
    $\hat{G} \leftarrow \frac{\partial}{\partial w_i} cost(w, (x^{(i)}, y^{(i)}))$; Gradient calculation
  end
  $r \leftarrow r + \hat{G} \odot \hat{G}$;
  $w \leftarrow w - \frac{\eta}{\epsilon + \sqrt{r}} \odot \hat{G}$;
end

Algorithm 4: Adadelta

Input: Training data S, learning rate η, weights w, decay rate ρ, fuzz factor ϵ
Output: Updated weights w

$\rho \leftarrow \rho_0$;
$\epsilon \leftarrow \epsilon_0 \approx 10^{-8}$;
$w \leftarrow w_0$;
$E[\hat{G}^2]_{t=0} \leftarrow 0$;
$E[\Delta w^2]_{t=0} \leftarrow 0$;
for $t \in \{1, \ldots, T\}$ **do**
    Randomly shuffle the training data S ;
    Sample a minibatch of size m:$\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\} \in S$;
    for $i \in \{1, \ldots, m\}$ **do**
        $\hat{G}_t \leftarrow \frac{\partial}{\partial w_i} cost(w_t, (x^{(i)}, y^{(i)}))$; Gradient calculation
    end
    $E[\hat{G}^2]_t \leftarrow \rho E[\hat{G}^2]_{t-1} + (1-\rho)\hat{G}_t^2$;
    $\Delta w_t \leftarrow -\frac{\sqrt{E[\Delta w^2]_{t-1}+\epsilon}}{\sqrt{E[\hat{G}^2]_t+\epsilon}}\hat{G}_t$;
    $E[\Delta w^2]_t \leftarrow \rho E[\Delta w^2]_{t-1} + (1+\rho)\Delta w^2{}_t$;
    $w_{t+1} \leftarrow w_t + \Delta w_t$;
end

Algorithm 5: RMSprop

Input: Training data S, learning rate η, weights w, decay rate ρ, fuzz factor ϵ, learning rate decay over each update r
Output: Updated weights w

$\rho \leftarrow \rho_0$;
$\epsilon \leftarrow \epsilon_0 \approx 10^{-8}$;
$w \leftarrow w_0$;
$r \leftarrow 0$;
while *stopping criterion is not met* **do**
    Randomly shuffle the training data S ;
    Sample a minibatch of size m:$\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\} \in S$;
    for $i \in \{1, \ldots, m\}$ **do**
        $\hat{G} \leftarrow \frac{\partial}{\partial w_i} cost(w, (x^{(i)}, y^{(i)}))$; Gradient calculation
    end
    $r \leftarrow \rho r + (1-\rho)\hat{G} \odot \hat{G}$;
    $w \leftarrow w - \frac{\eta}{\sqrt{\epsilon+r}} \odot \hat{G}$;
end

Algorithm 6: Adam

Input: Training data S, learning rate η, weights w, fuzz factor ϵ, learning rates decay over each update r_1 and r_2, exponential decay rates β_1 and β_2

Output: Updated weights w

$\epsilon \leftarrow \epsilon_0 \approx 10^{-8}$;
$w \leftarrow w_0$;
$r_1 \leftarrow 0$;
$r_2 \leftarrow 0$;
$t \leftarrow 0$;
while *stopping criterion is not met* **do**
  Randomly shuffle the training data S ;
  Sample a minibatch of size m:$\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\} \in S$;
  for $i \in \{1, \ldots, m\}$ **do**
    $\hat{G} \leftarrow \frac{\partial}{\partial w_i} cost(w, (x^{(i)}, y^{(i)}))$; Gradient calculation
    $t \leftarrow t + 1$;
  end
  $r_1 \leftarrow \beta_1 r_1 + (1 - \beta_1)\hat{G}$;
  $r_2 \leftarrow \beta_2 r_2 + (1 - \beta_2)\hat{G} \odot \hat{G}$;
  $\hat{r_1} \leftarrow \frac{r_1}{1-\beta_1^t}$;
  $\hat{r_2} \leftarrow \frac{r_2}{1-\beta_2^t}$;
  $w \leftarrow w - \eta \frac{\hat{r_1}}{\epsilon + \sqrt{\hat{r_2}}}$;
end

Algorithm 8: Nadam

Input: Training data S, learning rate η, weights w, fuzz factor ϵ, learning rates decay over each update r_1 and r_2, momentum decay rate γ, exponential decay rates β_1 and β_2

Output: Updated weights w

$\epsilon \leftarrow \epsilon_0 \approx 10^{-8}$;
$w \leftarrow w_0$;
$t \leftarrow 0$;
$r_1 \leftarrow 0$;
$r_2 \leftarrow 0$;
while *stopping criterion is not met* **do**
  Randomly shuffle the training data S ;
  Sample a minibatch of size m:$\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\} \in S$;
  for $i \in \{1, \ldots, m\}$ **do**
    $\hat{G}_t \leftarrow \frac{\partial}{\partial w_i} cost(w_t, (x^{(i)}, y^{(i)}))$; Gradient calculation
    $t \leftarrow t + 1$;
  end
  $r_{1t} \leftarrow \beta_1 r_{1t-1} + (1 - \beta_1)\hat{G}_t$;
  $\hat{r_{1t}} \leftarrow \frac{r_{1t}}{1-\beta_1^t}$;
  $w_{t+1} \leftarrow w_t - \frac{\eta}{\epsilon + \sqrt{r_{2t}}}(\beta_1 r_{1t} + \frac{1-\beta_1}{1-\beta_1^t}\hat{G}_t)$;
end

Algorithm 7: Adamax

Input: Training data S, learning rate η, weights w, fuzz factor ϵ, learning rate decay over each update r, exponentially weighted infinity norm u, exponential decay rates β_1 and β_2

Output: Updated weights w

$\epsilon \leftarrow \epsilon_0 \approx 10^{-8}$;
$w \leftarrow w_0$;
$r \leftarrow 0$;
$u \leftarrow 0$;
$t \leftarrow 0$;
while *stopping criterion is not met* **do**
  Randomly shuffle the training data S ;
  Sample a minibatch of size m:$\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\} \in S$;
  for $i \in \{1, \ldots, m\}$ **do**
    $\hat{G}_t \leftarrow \frac{\partial}{\partial w_i} cost(w_t, (x^{(i)}, y^{(i)}))$; Gradient calculation
    $t \leftarrow t + 1$;
  end
  $r_t \leftarrow \beta_1 r_{t-1} + (1 - \beta_1)\hat{G}_t$;
  $u_t \leftarrow \max(\beta_2 u_{t-1}, |\hat{G}_t|)$;
  $w_t \leftarrow w_{t-1} - \frac{\eta r_t}{(1-\beta_1^t)u_t}$;
end

REFERENCES

[1] Tahmassebi, A., "ideeple: Deep learning in a flash," in [*Disruptive Technologies in Information Sciences*], **10652**, 106520S, International Society for Optics and Photonics (2018).

[2] Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E., "Deep learning applications and challenges in big data analytics," *Journal of Big Data* **2**(1), 1 (2015).

[3] Dumbill, E., "What is big data? an introduction to the big data landscape," *O'Reilly* (2012).

[4] Sagiroglu, S. and Sinanc, D., "Big data: A review," in [*Collaboration Technologies and Systems (CTS), 2013 International Conference on*], 42–47, IEEE (2013).

[5] Lohr, S., "Opinion — big data's impact in the world," (Feb 2012).

[6] Belle, A., Thiagarajan, R., Soroushmehr, S., Navidi, F., Beard, D. A., and Najarian, K., "Big data analytics in healthcare," *BioMed research international* **2015** (2015).

[7] Reid, P. P., Compton, W. D., Grossman, J. H., Fanjiang, G., et al., "Information and communications systems: The backbone of the health care delivery system," (2005).

[8] Fan, J., Han, F., and Liu, H., "Challenges of big data analysis," *National science review* **1**(2), 293–314 (2014).

[9] Webb, A. and Kagadis, G. C., "Introduction to biomedical imaging," *Medical Physics* **30**(8), 2267–2267 (2003).

[10] Suetens, P., [*Fundamentals of medical imaging*], Cambridge university press (2017).

[11] Hendee, W. R., Ritenour, E. R., and Hoffmann, K. R., "Medical imaging physics," *Medical Physics* **30**(4), 730–730 (2003).

[12] Momose, A., Takeda, T., Itai, Y., and Hirano, K., "Phase–contrast x–ray computed tomography for observing biological soft tissues," *Nature medicine* **2**(4), 473–475 (1996).

[13] Lee, C. I., Haims, A. H., Monico, E. P., Brink, J. A., and Forman, H. P., "Diagnostic ct scans: assessment of patient, physician, and radiologist awareness of radiation dose and possible risks," *Radiology* **231**(2), 393–398 (2004).

[14] Ogawa, S., Lee, T.-M., Kay, A. R., and Tank, D. W., "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *Proceedings of the National Academy of Sciences* **87**(24), 9868–9872 (1990).

[15] Kalita, J. and Misra, U., "Comparison of ct scan and mri findings in the diagnosis of japanese encephalitis," *Journal of the neurological sciences* **174**(1), 3–8 (2000).

[16] Jezzard, P., Clare, S., et al., "Sources of distortion in functional mri data," *Human brain mapping* **8**(2-3), 80–85 (1999).

[17] Ehtemami, A., *Statistical data analysis of resting state fMRI: A study of nicotine addiction treatment*, PhD thesis, The Florida State University (2016).

[18] Lindquist, M. A. et al., "The statistical analysis of fmri data," *Statistical science* **23**(4), 439–464 (2008).

[19] Huettel, S. A., Song, A. W., McCarthy, G., et al., [*Functional magnetic resonance imaging*], vol. 1, Sinauer Associates Sunderland (2004).

[20] Stewart, S. B., Koller, J. M., Campbell, M. C., and Black, K. J., "Arterial spin labeling versus bold in direct challenge and drug-task interaction pharmacological fmri," *PeerJ* **2**, e687 (2014).

[21] Yoder, K. K., "Basic pet data analysis techniques," in [*Positron Emission Tomography-Recent Developments in Instrumentation, Research and Clinical Oncological Practice*], InTech (2013).

[22] Van Laere, K. J., Warwick, J., Versijpt, J., Goethals, I., Audenaert, K., Van Heerden, B., and Dierckx, R., "Analysis of clinical brain spect data based on anatomic standardization and reference to normal data: an roc-based comparison of visual, semiquantitative, and voxel-based methods," *Journal of Nuclear Medicine* **43**(4), 458–469 (2002).

[23] Sridevi, S., Vijayakumar, V., and Anuja, R., "A survey on medical image compression techniques,"

[24] Ukrit, M. F., Umamageswari, A., and Suresh, G., "A survey on lossless compression for medical images," *International Journal of Computer Applications* **31**(8), 47–50 (2011).

[25] Dean, J. and Ghemawat, S., "Mapreduce: simplified data processing on large clusters," *Communications of the ACM* **51**(1), 107–113 (2008).

[26] Alam, A. and Ahmed, J., "Hadoop architecture and its issues," in [*Computational Science and Computational Intelligence (CSCI), 2014 International Conference on*], **2**, 288–291, IEEE (2014).

[27] Bakshi, K., "Considerations for big data: Architecture and approach," in [*Aerospace Conference, 2012 IEEE*], 1–7, IEEE (2012).

[28] Zhang, Y., Cao, T., Li, S., Tian, X., Yuan, L., Jia, H., and Vasilakos, A. V., "Parallel processing systems for big data: a survey," *Proceedings of the IEEE* **104**(11), 2114–2136 (2016).

[29] Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., et al., "Mllib: Machine learning in apache spark," *The Journal of Machine Learning Research* **17**(1), 1235–1241 (2016).

[30] Shoro, A. G. and Soomro, T. R., "Big data analysis: Apache spark perspective," *Global Journal of Computer Science and Technology* **15**(1) (2015).

[31] Nitzberg, M., Mumford, D., and Shiota, T., [*Filtering, segmentation and depth*], vol. 662, Springer (1993).

[32] Adams, R. and Bischof, L., "Seeded region growing," *IEEE Transactions on pattern analysis and machine intelligence* **16**(6), 641–647 (1994).

[33] Shafarenko, L., Petrou, M., and Kittler, J., "Automatic watershed segmentation of randomly textured color images," *IEEE transactions on Image Processing* **6**(11), 1530–1544 (1997).

[34] Poldrack, R. A., "Region of interest analysis for fmri," *Social cognitive and affective neuroscience* **2**(1), 67–70 (2007).

[35] Kaur, A. and Goyal, M., "Roi based image compression of medical images," *International Journal of Computer Science Trends and Technology* , 2347–8578 (2014).

[36] Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R. G., Granton, P., Zegers, C. M., Gillies, R., Boellard, R., Dekker, A., et al., "Radiomics: extracting more information from medical images using advanced feature analysis," *European journal of cancer* **48**(4), 441–446 (2012).

[37] Dey, N., Ashour, A. S., Shi, F., and Balas, V. E., [*Soft Computing Based Medical Image Analysis*], Academic Press (2018).

[38] Gray, V., [*Principal Component Analysis: Methods, Applications and Technology*], Nova Science Publishers, Incorporated (2017).

[39] Tahmassebi, A., Gandomi, A. H., McCann, I., Schulte, M. H., Schmaal, L., Goudriaan, A. E., and Meyer-Baese, A., "An evolutionary approach for fmri big data classification," in [*2017 IEEE Congress on Evolutionary Computation (CEC)*], 1029–1036, IEEE (2017).

[40] Tahmassebi, A., Gandomi, A. H., Schulte, M. H., Goudriaan, A. E., Foo, S. Y., and Meyer-Baese, A., "Optimized naive-bayes and decision tree approaches for fmri smoking cessation classification," *Complexity* **2018** (2018).

[41] Tahmassebi, A., *Pattern Recognition in Medical Imaging: Supervised Learning of fMRI and MRI Data*, PhD thesis, The Florida State University (2018).

[42] Lathi, B. P., [*Signal processing and linear systems*], Oxford University Press New York (1998).

[43] Ballard, D. H. and Brown, C. M., "Computer vision. englewood cliffs," *J: Prentice Hall* (1982).

[44] Zhang, W., Doi, K., Giger, M. L., Wu, Y., Nishikawa, R. M., and Schmidt, R. A., "Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network," *Medical Physics* **21**(4), 517–524 (1994).

[45] Cheng, R., Roth, H. R., Lay, N. S., Lu, L., Turkbey, B., Gandler, W., McCreedy, E. S., Pohida, T. J., Pinto, P. A., Choyke, P. L., et al., "Automatic magnetic resonance prostate segmentation by deep learning with holistically nested networks," *Journal of Medical Imaging* **4**(4), 041302 (2017).

[46] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in [*Advances in neural information processing systems*], 1097–1105 (2012).

[47] Nishio, M., Nagashima, C., Hirabayashi, S., Ohnishi, A., Sasaki, K., Sagawa, T., Hamada, M., and Yamashita, T., "Convolutional auto-encoder for image denoising of ultra-low-dose ct," *Heliyon* **3**(8), e00393 (2017).

[48] Chen, H., Zhang, Y., Zhang, W., Liao, P., Li, K., Zhou, J., and Wang, G., "Low-dose ct via convolutional neural network," *Biomedical optics express* **8**(2), 679–694 (2017).

[49] Tahmassebi, A., Gandomi, A. H., McCann, I., Schulte, M. H., Goudriaan, A. E., and Meyer-Baese, A., "Deep learning in medical imaging: fmri big data analysis via convolutional neural networks," *Proceedings of the Practice and Experience on Advanced Research Computing. ACM* (2018).

[50] Xie, S. and Tu, Z., "Holistically-nested edge detection," in [*Proceedings of the IEEE international conference on computer vision*], 1395–1403 (2015).

[51] Mohebali, B., Breslend, P., Graber, L., and Steurer, M., "Validation of a scattering parameter based model of a power cable for shipboard grounding studies," in [*ASNE Electric Machines Symposium (EMTS)*], **201**, 28–29 (2014).

[52] Graber, L., Mohebali, B., Bosworth, M., Steurer, M., Card, A., Rahmani, M., and Mazzola, M., "How scattering parameters can benefit the development of all-electric ships," in [*2015 IEEE Electric Ship Technologies Symposium (ESTS)*], 353–357, IEEE (2015).

[53] Tahmassebi, A., Gandomi, A. H., and Meyer-Bäse, A., "High performance gp-based approach for fmri big data classification," in [*Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*], 57, ACM (2017).