Hindawi
Journal of Healthcare Engineering
Volume 2021, Article ID 9998819, 11 pages
https://doi.org/10.1155/2021/9998819

Review Article

Biomedical Image Classification in a Big Data Architecture Using Machine Learning Algorithms

Christian Tchito Tchapga,[1] Thomas Attia Mih,[1] Aurelle Tchagna Kouanou ,[1,2] Theophile Fozin Fonzin ,[2,3] Platini Kuetche Fogang,[4] Brice Anicet Mezatio,[2] and Daniel Tchiotsop[5]

[1]*College of Technology, University of Buea, Buea, Cameroon*
[2]*Department of Research, Development,Innovation and Training, InchTech's, Yaoundé, Cameroon*
[3]*Department of Electrical and Electronic Engineering, Faculty of Engineering and Technology (FET), University of Buea, P.O. Box 63, Buea, Cameroon*
[4]*Research Unity of Condensed Matter, Electronics and Signal Processing, Department of Physics, Faculty of Science, University of Dschang, P.O. Box 67, Dschang, Cameroon*
[5]*Research Unity of 'Automatic and Applied Informatic,IUT-FV of Bandjoun, University of Dschang-Cameroun, B.P. 134 Bandjoun, Dschang, Cameroon*

Correspondence should be addressed to Aurelle Tchagna Kouanou; tkaurelle@gmail.com

Received 6 March 2021; Revised 9 May 2021; Accepted 25 May 2021; Published 31 May 2021

Academic Editor: Aiping Liu

Copyright © 2021 Christian Tchito Tchapga et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In modern-day medicine, medical imaging has undergone immense advancements and can capture several biomedical images from patients. In the wake of this, to assist medical specialists, these images can be used and trained in an intelligent system in order to aid the determination of the different diseases that can be identified from analyzing these images. Classification plays an important role in this regard; it enhances the grouping of these images into categories of diseases and optimizes the next step of a computer-aided diagnosis system. The concept of classification in machine learning deals with the problem of identifying to which set of categories a new population belongs. When category membership is known, the classification is done on the basis of a training set of data containing observations. The goal of this paper is to perform a survey of classification algorithms for biomedical images. The paper then describes how these algorithms can be applied to a big data architecture by using the Spark framework. This paper further proposes the classification workflow based on the observed optimal algorithms, Support Vector Machine and Deep Learning as drawn from the literature. The algorithm for the feature extraction step during the classification process is presented and can be customized in all other steps of the proposed classification workflow.

1. Introduction

The healthcare field has experienced rapid growth in medical data in recent years. In 2018, the USA generated a zettabyte of healthcare data [1]. In the wake of this agglomeration of medical data, especially images, the use of new methods based on big data technologies, machine learning (ML), and artificial intelligence (AI) has therefore become necessary. Big data is generally identified by five major characteristics called the "5V": volume (amount of data generated), variety (data from different categories), velocity (speed of data generation), variability (inconsistency of data), and veracity (quality of captured data) [1–8]. The application of information technologies to the healthcare field raises opportunities for the development of new diagnostics and treatments, making it a critical area of investigation. The new ideas, concepts, and technologies based on big data, ML, and AI are proposed to improve the healthcare field. Nowadays, many works are performed to use big data to manage and analyze healthcare systems. El aboudi and Behlima proposed a big data management approach for healthcare systems [9]; Tchagna et al. proposed a complete big data workflow for biomedical image analysis [7]; Belle et al. showed the impact of big data analysis in healthcare [10]; Luo et al. in [2] performed a literature review of big data application in biomedical research and healthcare; Viceconti et al., as far as they are concerned, examined the possibility of using big data for personalized healthcare [11]; Archenaa and Anita in 2015 showed the need for big data analytics in healthcare to improve the quality of healthcare as follows: providing patient-centric services, detecting spreading diseases earlier, monitoring the hospital's quality, and improving the treatment methods [12]. Thus, when big data technologies are incorporated into a framework or applications, better data handling and higher performance can be achieved [13]. Based on those works, it was noticed that the biomedical system is converging to a big data platform that presents us with an opportunity to efficiently manage and analyze this huge and growing amount of biomedical data.

A vast quantity of data in healthcare constitutes data images captured from medical imaging (Computed Tomography Scan, Echography, Mammography, MRI, etc.). To achieve complete management and analysis of biomedical images, we have to automate all steps proposed in [7]. One of the most necessary steps is classification. Classification in ML concerns a problem of identifying to which set of categories a new population belongs [7]. A good classification performed essentially leads to a good automatic diagnosis of diseases on an image. This is in order that the diagnostic algorithms can adapt accordingly to the image groups resulting from the classification. So, classification is an important step in a biomedical automatic system.

In a new concept for biomedical images analysis using big data architecture proposed in 2018 by Tchagna et al. in [7], the authors present a workflow performing the steps of acquisition of biomedical image data, analysis, storage, processing, querying, classification, and automatic diagnosis of biomedical images. The workflow was performed with unstructured and structured image data based on a NoSQL database. The authors proposed a Spark architecture that allows developing appropriate and efficient methods to leverage a large number of images for classification. However, in their work, they did not explain very well the algorithm used for biomedical image classification. Based on this gap, the paradigm in this paper is to present and discuss methods and algorithms used to perform a good classification for the biomedical image in big data architecture.

This paper specifically focuses on biomedical imaging with big data technologies along with ML for classification. It presents a set of algorithms that can be used to accomplish the classification step in big data architecture. It further describes the importance of applying the classification of biomedical images in big data architecture. Based on the Spark framework, this work proposes an algorithm to perform the steps of the proposed classification workflow. ML plays an important role in biomedical image classification, and when combined with big data technologies, the processing is done with less time and can handle a lot of images at the same time. The rest of this paper is organized as follows. Section 2 reviews published methods in the field. In Section 3, these methods are explored theoretically throughout our work. Section 4 presents the Spark algorithm. A conclusion and future work are provided in Section 5.

2. A Survey of Biomedical Image Classification Methods

The healthcare field is distinctively different from other fields. Healthcare is generally delivered by health professionals. Pharmacy, dentistry, nursing, midwifery, medicine, audiology, optometry, occupational therapy, psychology, physical therapy, and other health professions are all part of healthcare. Healthcare is a high-priority field and people expect the highest level of care and services. It is most difficult for specialists to identify complex disease patterns from large amounts of images. Hence, each specialist will be limited to visualizing only the biomedical images essentially related to his field of competence, which is somewhat restrictive. In contrast, ML, deep learning (DL), and AI excel at automatic pattern recognition from large amounts of biomedical image data. In particular, machine learning and deep learning algorithms (e.g., support vector machine, neural network, and convolutional neural network) have achieved impressive results in biomedical image classification [14–23]. Classification helps to organize biomedical image databases into image categories before diagnostics [24–30]. Many investigations have been performed by researchers to improve classification for biomedical images [6, 7, 31–36]. In 2016, Miranda et al. surveyed medical image classification techniques. They reviewed the state-of-the-art image classification techniques to diagnose human body disease and covered identification of medical image classification techniques, image modalities used, the dataset, and tradeoff for each technique [31]. They concluded that artificial neural network (ANNs) classifier and SVM are the most used technique for image classification because these techniques give high accuracy, high sensitivity, high specificity, and high classification performance results [31]. In the same logic, Jiang et al. in 2017 made an investigation on ML algorithms for healthcare [32]. They grouped algorithms by the category of ML (Supervised, Unsupervised, and Semisupervised) and provided a graphical representation. Supervised learning algorithms are used for classification. They showed that SVM and ANNs are two famous algorithms used to classify biomedical image data. In medical imaging, SVM and ANN take up to 42% and 31%, respectively, of the most used algorithms [32]. Similarly, Wang et al. in [6] confirmed that the SVMs and ANNs are good classifiers.

In 2007, Jiang et al. used the Rough Set Theory (RTS) to improve SVM for classifying digital mammography images [33]. They reported 96.56% accuracy. However, their work is only limited to mammography images, and they used structured data. But in reality, the vast majority of images' data come from many sources that are unstructured. Jeved et al. proposed in [34] a technique to classify brain images from Magnetic Resonance Imaging (MRI) using perceptual texture features, fuzzy weighting, and support vector machine. Their proposed technique classifies normal and different classes of abnormal images, and they used fuzzy logic to assign weights to different feature values based on its discrimination capability. Lu and Wang in 2012 applied the SVM to breast multispectral magnetic resonance images to classify the tissues of the breast. They compared their method with the commonly used C-means for performance evaluation and proved that the SVM is a promising and effective spectral technique for MR image classification [35]. Although the two methods majorly mentioned above are efficient, their applications in image classification are limited only to a few kinds of medical images. Deep learning comes with many hidden networks to improve the efficiency of classification performance when the datasets are very large. For this reason, khan et al. in [36] proposed a modified convolutional neural network (CNN) architecture for automatically classifying anatomies in medical images by learning features at multiple levels of abstractions from the data obtained. They also provided an insight into the deep features that have been learned through training, which will help in analyzing various abstraction of features ranging from low level to high level and their role in the final classification, and obtained a test accuracy of 81% [36]. Li et al. in [37] proposed a customized CNN network for lung image patch classification and designed a fully automatic neural-based machine learning framework to extract discriminative features from training samples and perform classification at the same time. They showed that the same CNN architecture can be generalized to perform other medical image or texture classification tasks. In 2018, Ker et al. discussed the DL applications in medical image classification, localization, detection, segmentation, and registration [38]. They focused on CNN and explained all methods to perform each task. Concerning classification, they gave examples of disease classification tasks by using CNN. In 2020, Zhang et al. looked for how to accelerate the processes of learning time with large-scale multilabel image classification using the CNN method for learning and building the classifier with an unknown novel group that came in a stream during the training stage [39]. However, their classifier/model is essentially on the ability of the novel-class detector that can give the worse result when multiple novel classes may exist. Vieira et al. provided a review of the studies of applying DL to neuroimaging data to investigate neurological disorders and psychiatric. They compare the different ML algorithms with DL algorithm in neuroimaging and show that DL gives good results compared to ML such as SVM when the dataset is important [40]. Nalepa and Kawulok in 2019 performed an extensive survey on existing techniques and methods to select SVM training data from large datasets and concluded that the DL will be more efficient than SVM for large datasets [41]. Badar et al. show how to apply DL in Retina image classification and identification to detect diseases such as diabetic retinopathy, macular bunker, age-related macular degeneration, retinal detachment, retinoblastoma, and retinitis pigmentosa [42]. Yan et al. in 2019 proposed a novel hybrid CNN and RNN for breast cancer image classification by using the richer multilevel feature representation of the histopathological biomedical image patches [43]. Zareapoor et al. used a combination of DL and a nonlinear-SVM to deal with extremely large datasets to improve the learning time of their model. However, the learning time remains and DL is today a problem with their model [44]. Fang et al. proposed a CNN architecture method for breast cancer classification by constructing a multi-SVM-based biomedical image kernel using quality scores got to achieve the classification [45]. Kotsiantis in [46] compares the features of learning techniques for classification. Table 1 shows the summary of this comparison. As Jiang et al. in [32], they concluded that the SVM and ANNs are the best algorithms used for classification problem in biomedical image. However, they established this conclusion when the amount of data is not large. So today, we can replace ANNs with CNN when we work on a large dataset.

Based on the previously cited literature in this section, it was observed that the classifier algorithms depend on the amount of data of images in the input of the classification system. For example, for a medium dataset, SVM outperforms another classification algorithm like DL. Indeed, the SVM classifier often outperforms many commonly used ML algorithms, even though it may not be an ideal choice to handle large datasets [47–51]. For a large dataset, DL outperforms another classification algorithm [23, 39, 52–55]. Despite the notable advantages of DL and SVM, challenges in applying them to the biomedical domain still remain. It was noticed that none of these works have made their classification with big data tools. Indeed, classification can be performed with big data technologies for these reasons:

(1) In order to swiftly work with both unstructured and structured biomedical images (inferring knowledge from complex heterogeneous patient data/leveraging the patient data image correlations in longitudinal records)

(2) Rapid queries and access to biomedical images database

(3) Prospect of a database based on NoSQL technologies

(4) Personalized classification algorithm to the patient

(5) Opportunity to efficiently handle massive amounts of biomedical image data

(6) Easy to analyze data images using machine learning and artificial intelligence

(7) Implementation of the MapReduce programming (parallel programming) in those frameworks (Hadoop, Spark)

Furthermore, in the previously cited works in this section, the authors did not show what is the impact of big data in their works, if any. This drawback is one of the main interests of this paper. This paper, therefore, presents a designed workflow for biomedical image classification based on SVM and DL (CNN), which could be implanted in big data architecture.

Table 1: Comparison of classification methods in biomedical image based on the literature [32, 46].

	Decision trees	Neural networks	Naïve bayes	KNN	SVM	Rule-learning
Accuracy	**	***	*	**	****	**
Speed of classification	****	****	****	*	****	****
Tolerance to redundant attributes	**	**	*	**	***	**
Speed of learning	***	*	****	****	*	**
Tolerance to missing values	***	*	****	*	**	**
Tolerance to highly interdependent attributes	**	***	*	*	***	**
Dealing with discrete/binary/ continues attributes	****	*** (not discrete)	*** (not continuous)	*** (not directly discrete)	** (Not discrete)	*** (not directly discrete)
Tolerance to noise	**	**	***	*	**	*
Dealing with a danger of overfitting	**	*	***	***	**	**
Attempts for incremental learning	**	***	****	****	**	*

****Very good. ***Good. **Fairly Good. *Bad.

3. System Classification Workflow for Biomedical Images

The application of ML technology with SVM, especially DL with CNN, to biomedical image classification field research has become more and more popular recently. The main objective of medical image classification is to identify which parts of the human body are infected by the disease and not only to reach high accuracy. In the proposed workflow, according to the previous section, there are two algorithms to perform classification with good accuracy: one for a medium dataset and the other for a large-scale dataset. SVM and DL are then used, respectively, in this regard. The classification processes are as depicted in Figure 1.

In Figure 1, the system workflow to perform a biomedical image classification is presented. As shown in the workflow, the classification process is performed in two basic steps. In the first step, a classifier model is built based on the labeled biomedical image using ML (SVM or CNN) algorithms. When the classifier model has been derived, any unlabeled biomedical images can be presented to the model in order to make predictions about the group to which such images belong. Figure 2 presents a DL along with CNN architecture for image classification. The following section presents how we deal with training and testing datasets in classification.

3.1. The Training Phase. The first part of Figure 1 is the training phase. The training phase in classification concerns the phase where you present your data from the training dataset (labeled biomedical images in this case), extract features, and train your model, by mapping the input with the expected output. Here, the network can learn by using a Gradient Descent Algorithm (GDA). The purpose of GDA is to find the different values of the network weights that best minimize the error between the true and estimated outputs [56–59]. Backpropagation is the name of this propagation procedure and permits the network to predict how much the weights from the lower layers network have to be changed by the GDA. The training phase has traditionally three main steps: labeled biomedical image dataset retrieval, feature extraction, and machine learning algorithm (SVM or CNN).

3.1.1. Labelled Biomedical Image. In general, labeled images (training dataset) are used to perform the machine learning of the class (group) description which in turn is used for unknown (unlabeled) images [60]. Since the supervised learning paradigm is adopted in this workflow, the labeled biomedical images dataset is the most suitable for the learning phase in this workflow.

3.1.2. Feature Extraction. An image is represented by a set of descriptors that structure the feature vectors and is formed by pixels, which may or may not represent features. A feature is defined as an interesting part of an image and is used as a starting point for computer vision algorithms [61]. When the features are extracted from a labeled biomedical image dataset, classification is then done using a classification method such as SVM or DL. When the classification is performed by using DL, the features are called deep features.

3.1.3. Machine Learning Algorithm (SVM or CNN). The support vector machine (SVM) is a supervised learning method that generates input-output mapping functions from a set of labeled training data [62]. Originally, SVM is a binary classifier that works by identifying the optimal hyperplane and correctly divides the data points into two classes [63]. There will be an infinite number of hyperplanes and SVM will select the hyperplane with maximum margin. The margin indicates the distance between the classifier and the training points (support vector) [63, 64]. SVM is mainly used to deal with classification and regression problems. There are three steps of the SVM algorithm, Identification of Hyperplane, classification of classes, and finding hyperplane to separate classes [64, 65]. The training principle behind SVM is to find the optimal linear hyperplane so that the expected classification error for unseen test samples should be minimized [34, 60]. SVM is a margin-based classifier that achieves superior classification performance compared to

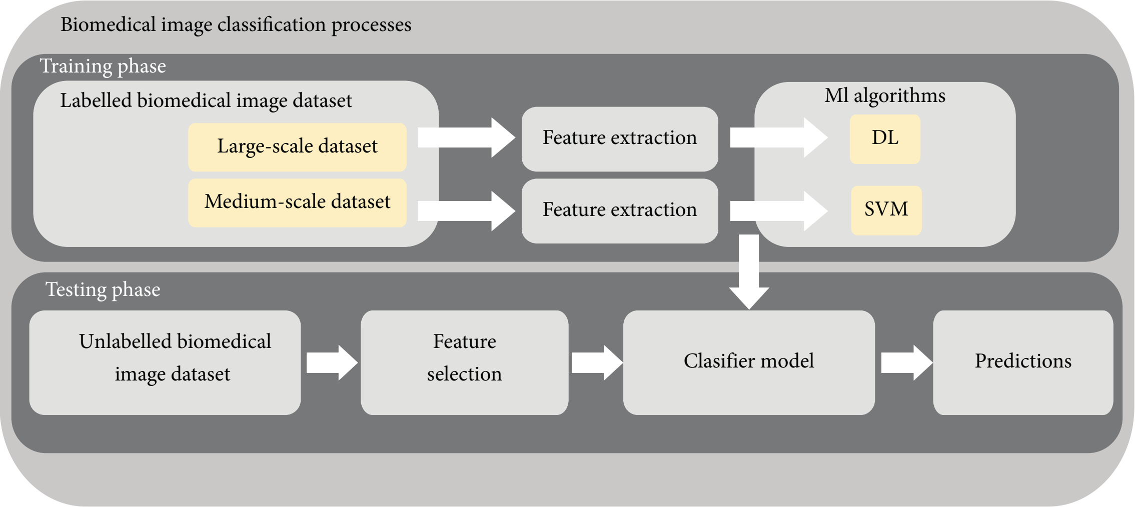

Figure 1: Classification system workflow for training and testing processes.

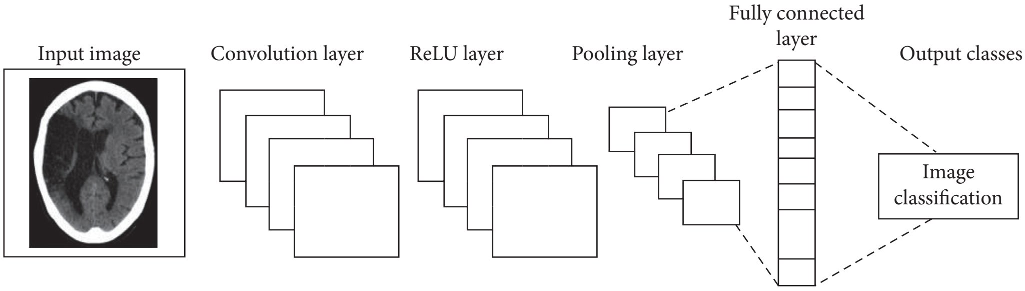

Figure 2: Convolutional neural network (CNN) architecture for biomedical image classification.

other algorithms when the amount of dataset training is medium [34, 51, 60].

DL techniques are conquering the prevailing traditional approaches of the neural network; when it comes to the huge amount of dataset, applications requiring complex functions demanding increase accuracy with lower time complexities [22, 66, 67]. DL particularly CNN has shown an intrinsic ability to automatically extract the high-level representations from big data [36]. CNN is an artificial neural network with many hidden layers of units between the input and output layers and millions or billions of parameters [21, 68–71]. General, DL architecture is composed of one or more convolutional layers with many hidden networks, one or more max pooling operations, and a full connection layer. This feeds into a final Fully Connected Layer which assigns class scores or probabilities, thus classifying the input into the class with the highest probability [38]. To apply a CNN on an image, we have this image in the input to the network. The network has an input layer that takes this image as the input, an output layer from where we obtain the trained output, and the intermediate layers called the hidden layers. The network has a series of subsampling and convolutional layers. The layers produce together an approximation of input image data. DL is very good at learning the local and global structures from image data [37]. However, the CNN exploits spatially local correlation by enforcing a local connectivity pattern between neurons of adjacent layers [72]. Deep learning enables the extraction of multiple feature levels from data directly without explicit definition. It provides a higher level of feature abstraction, thus potentially providing better prediction performance [73, 74]. Research works based on CNN significantly improved the best performance for many image databases [37, 75]. So, to apply DL, the dataset of the image has to contain many images.

3.2. Testing Phase. In the testing phase, the feature vectors of the unlabeled biomedical image dataset serve as input. A classifier decided on the basis of the classifier model, with its own classification rules, to which class/group that feature vector belongs. The testing phase has four main steps: unlabeled biomedical image capturing, feature extraction, classifier model, and prediction. The feature extraction step in the testing phase is performed as in the training phase.

3.2.1. Unlabeled Biomedical Image. The unlabeled biomedical images dataset is used to provide an unbiased evaluation of a final model fit on the labeled biomedical images dataset.

3.2.2. Classifier. A classifier is trained on the extracted features. The goal of a classifier is to distinguish images of the known class from images of alien classes. Thus, a classifier is required to learn so that it can identify out-of-class (alien) images. The SVM and DL classifiers are used to perform verification for the next stage of prediction.

3.2.3. Prediction. Based on SVM or DL classifier, the prediction stage in the workflow allows predicting automatically into which class an image belongs. Here, we can evaluate the prediction average accuracy for both SVM and DL. However, as drawn from the literature, it is established that for a large dataset, the accuracy of the DL classifier is generally better than the SVM classifier. And for a medium dataset, the classifier of SVM is better than the DL classifier.

One of the characteristics of big data is the volume (amount of data generated). To apply the classification workflow of Figure 1 in big data architecture, we have to verify this rule for the dataset that is presented to the workflow's training phase. However, taking into consideration the size of the dataset we can perform classification on, we can use SVM or DL as explained in the previous subsection. Here, the performance of the network can be evaluated by several performance parameters such as sensitivity, accuracy, specificity, and F-score. Sensitivity refers to the proportion of true positives correctly identified, specificity refers to true negatives correctly identified, and the accuracy of a classifier/model represents the overall proportion of correct classifications [58, 59]. Spark framework is one of the best frameworks used to perform big data processing. In the next section, an algorithm to perform some stages of Figure 1 is presented, based on the Spark framework of image classification in [7].

4. Spark Algorithm for Biomedical Image Classification

Apache Spark is a distributed computing platform used in the big data scenario that has become one of the most powerful frameworks. Spark offers a unified and complete framework to manage the different requirements for big data processing with a variety of datasets (graph data, image/video, text data, etc.) from different sources (batch, real-time streaming) [7]. Spark framework has been created to overcome the problems of the Hadoop framework according to its creators. Indeed, the Spark framework has proved to perform faster than Hadoop in many situations (more than 100 times in memory). With capabilities like in-memory data storage and near real-time processing, the performance can be several times faster than other big data technologies. Spark framework is able to make data suitable for iteration, query it repeatedly, and load it into memory. In the Spark framework, the main program (driver) controls multiple slaves (workers) and collects results from them, whereas slaves' nodes read data partitions (blocks) from a distributed file system execute some computations and save the result to disk. Spark as Hadoop is based on parallel processing MapReduce that aims at automatically processing data in an easy and transparent way through a cluster of computers. In addition to Map and Reduce operations, Spark also supports SQL queries, streaming data, machine learning, and graph processing data [7]. In Spark, sometimes, we can program and execute our algorithm on many clusters at the same time. For instance, Figure 3 shows the links between four nodes to perform data processing.

Figure 3 shows the possibility of the processing of data in four nodes, where the master node and the slave nodes are defined. The master manages and distributes the job to the slave. According to the volume of the dataset, you can choose more or less than three slaves. The number of slaves leads to the gaining of processing time.

In Spark DataFrame, the importation and representation of images follow the pipeline as shown in Figure 4.

This pipeline consists typically of the image import, preprocessing, model training, and inferencing stages.

In this section, we propose a Spark algorithm to perform some stages of Figure 1. Image feature is an image pattern, based on which we can describe the image with what we see. The main role of features in image biomedical classification is to transform visual information into vector space. Thus, we can perform mathematical operations on them and find similar vectors. To perform feature selection, the first issue is to detect features on a biomedical image. The number of them can be different depending on the image, so we add some clauses to make our feature vector always have the same size. Then, we build vector descriptors based on our features; each descriptor has the same size. It should be noted that there are different approaches to write this Spark algorithm for each step of Figure 1. Algorithm 1 presents a method to perform feature extraction by using Spark framework with its MapReduce programming. In this algorithm, it should be noted that the feature extraction from the unlabeled or labeled image is performed with many images in the big data context with respect to the different V of big data (volume, velocity, variety, variability, and veracity). However, the performances of the Spark framework can be decreased in some situations: especially during the feature extraction, in a situation where there are some small images in the dataset (unlabeled biomedical images/labeled biomedical images). Another instance is if the size of the image considered is too different from one another, it will cause an unbalanced loading in the Spark. To solve these problems, in [76], the authors introduced two methods: sequence in feature extraction and feature extraction by segmentation. The implementation of one of these two methods can resolve the problem of unbalanced loading and the running time of each job can also be the same.

The process of classification is a function that is started when new unlabeled data comes to the system. Algorithm 2 predicts the image's class by identifying to which set of categories this image belongs. In the algorithm, both prediction and query are performed in the same **MapReduce** phase.

By adopting the Spark framework, there comes an advantage to work in a big data environment and use its embedded libraries like MLlib (Machine Learning libraries).

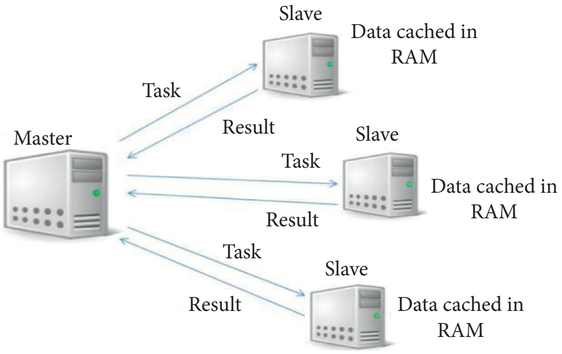

Figure 3: Job execution Apache Spark in four clusters: one master and three slaves.

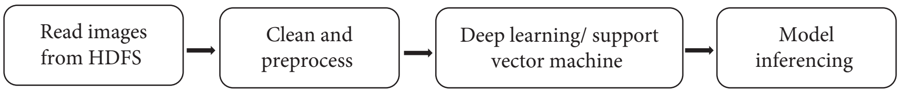

Figure 4: Importing images in Spark DataFrame.

```
INPUT: DataI
//DataI is the unlabeled or labeled biomedical image to be processed in order to extract features
MapReduce $e$ dataI
Find the feature to $e$ in the unlabeled or labeled biomedical image and the outputs of a tuple with the feature ID (key) and $e$ (value) (MAP)
The tuple is sent to the correspondent node according to its key (SHUFFLE)
Features = the standard elements for each biomedical image used to build a classifier model based on SVM or DL. The output will consist of a tuple with $e$ (key) and the features model (value) (REDUCE)
End MapReduce
```

Algorithm 1: Feature extraction process.

```
INPUT: query, cl
//query is the data image to be queried
//cl represents the number of class for prediction
MapReduce $e$ data
Find the feature node to $e$ in the class model and the outputs of a tuple with the class ID (key) and $e$ (value) (MAP)
The tuple is sent to the correspondent node according to its key (SHUFFLE)
Features = the standard elements for each image in order to retrieve the classifier model of $e$. The output will consist of a tuple with $e$ (key) and the classifier model (value) (REDUCE)
For each tuple in features, the model returns the most-voted class from the classifier model. This value will be the class image predicted for the given image
End MapReduce
```

Algorithm 2: Prediction process.

5. Comparison of Machine Learning Methods and Applications' Use in the Literature

To perform this comparison, we are based on some works done in the literature. Table 2 gives us an overview of different works done in the literature in ML with their application.

Table 2 helps us to see that ML algorithms are very used in biomedical application and today in a lot of applications also. Many researches as Wang et al., Tchagna Kouanou et al., or Chowdharya et al. performed a good job and published a lot of papers in this exciting domain.

We can also conclude that, nowadays, as the size of the training data set grows, ML algorithms become more effective. Therefore, when combining big data technologies with ML, we will benefit twice: these algorithms can help us keep up with the influx of data, and the amount

Table 2: Some ML methods and application comparison.

Authors	Deep learning methods	Machine learning method	Big data technologies	Applications
Luo et al. [2]	No	No	Yes (Hadoop)	Healthcare
Tchagna Kouanou et al. [7]	No	No	Yes (Spark and Hadoop)	Biomedical images
Manogaran and Lopez [8]	No	Yes	Yes	Healthcare
Thrall et al. [17]	No	Yes	No	Radiology
Fujiyoshi et al. [24]	Yes	No	No	Image recognition
Tchagna Kouanou et al. [77]	No	Yes (K-Means- unsupervised learning)	No	Biomedical image compression
Tchagna Kouanou et al. [78]	No	Yes (K-Means- unsupervised learning)	Yes (Hadoop)	Biomedical image compression
Tchagna Kouanou et al. [79]	No	Yes (K-Means- unsupervised learning)	No	Image compression
Alla Takam et al. [80]	Yes (CNN)	No	Yes (Spark)	Biomedical image
Chowdhary and Acharjya [81]	No	Yes (fuzzy C-means)	No	Feature extraction and segmentation
Bhattacharya et al. [82]	Yes	No	No	Biomedical image
Chowdhary et al. [83]	Yes	No	No	Biomedical images (breast cancer classification)
Wang et al. [84]	Yes (CNN, hierarchical loss)	No	No	Biomedical images (breast cancer classification)

and variety of the same data can help and grow the algorithms. ML and big data are the current blue chips in the IT industry. The storage of big data technologies analyzes and extracts information from a large amount of data. On the other hand, ML is the ability to learn and improve automatically from experience without explicit programming [7].

6. Conclusion

In this paper, we have focused on the concept of big data for biomedical image classification tasks and, in particular, on exploring machine learning algorithms (SVM and DL) for biomedical classification following the Spark programming model. We have proposed a workflow with essential steps for biomedical image classification. Based on the literature surveyed, the SVM and DL were found to be the two possible candidate algorithms that can be used to perform biomedical image classification. In the survey, it was established from the literature that SVM gives a good performance when the size of the dataset is medium while the DL is established to have good performance when the dataset is of large scale. Therefore, we can choose which machine learning algorithm to use for classification based on the size of the dataset at hand. Spark is the framework that we proposed for the implementation of the proposed workflow. We have given a Spark algorithm to perform feature extraction in our proposed workflow. It should be noted that this algorithm can be customized and applies to another step. As future work, we propose to make a real-world implementation of our Spark algorithm and calculate all performance parameters as in [77–79], where the authors implemented the algorithm for image compression that we can use in the workflow proposed in [7].

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors acknowledge and thank Dr. Romanic Kengne and the InchTech's team (http://www.inchtechs.com), for their support and assistance during the conception of this work.

References

[1] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, and S. S. Iyengar, "Computational health informatics in the big data Age," *ACM Computing Surveys*, vol. 49, no. 1, pp. 1–36, 2016.

[2] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, "Big data application in biomedical research and health care: a literature review," *Biomedical Informatics Insights*, vol. 8, pp. 1–10, 2016.

[3] A. Yang, M. Troup, and J. W. K. Ho, "Scalability and validation of big data bioinformatics software," *Computational and Structural Biotechnology Journal*, vol. 8. , 2017 In press.

[4] S. Istephan and M.-R. Siadat, "Unstructured medical image query using big data - an epilepsy case study," *Journal of Biomedical Informatics*, vol. 59, pp. 218–226, 2016.

[5] A. Oussous, F.-Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big Data technologies: a survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, 2018.

[6] L. Wang, Y. Wang, and Q. Chang, "Feature selection methods for big data bioinformatics: a survey from the search perspective," *Methods*, vol. 111, pp. 21–31, 2016.

[7] A. Tchagna Kouanou, D. Tchiotsop, R. Kengne, D. T. Zephirin, N. M. Adele Armele, and R. Tchinda, "An optimal big data workflow for biomedical image analysis," *Informatics in Medicine Unlocked*, vol. 11, pp. 68–74, 2018.

[8] G. Manogaran and D. Lopez, “A survey of big data architectures and machine learning algorithms in healthcare,” *International Journal of Biomedical Engineering and Technology*, vol. 25, no. 2/3/4, pp. 182–211, 2017.

[9] N. El aboudi and L. Benhlima, “Big data management for healthcare systems: architecture, requirements, and implementation,” *Advances in Bioinformatics*, vol. 2018, Article ID 4059018, 10 pages, 2018.

[10] A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, “Big data analytics in healthcare,” *BioMed Research International*, vol. 2015, Article ID 370194, 16 pages, 2015.

[11] M. Viceconti, P. Hunter, and R. Hose, “Big data, big knowledge: big data for personalized healthcare,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1209–1215, 2015.

[12] J. Archenaa and E. A. M. Anita, “A survey of big data analytics in healthcare and government,” *Procedia Computer Science*, vol. 50, pp. 408–413, 2015.

[13] M. A. Amanullah, R. A. A. Habeeb, and F. H. Nasaruddin, “Deep learning and big data technologies for IoT security,” *Computer Communications*, vol. 151, 2020.

[14] N. Pitropakis, E. Panaousis, T. Giannetsos, E. Anastasiadis, and G. Loukas, “A taxonomy and survey of attacks against machine learning,” *Computer Science Review*, vol. 34, Article ID 100199, 2019.

[15] M. L. Giger, *Machine Learning in Medical Imaging*, American College of Radiology, Reston, VA, USA, 2017.

[16] H. T. Nguyen and L. T. Nguyen, “Fingerprints classification through image analysis and machine learning method,” *Algorithms*, vol. 12, p. 241, 2019.

[17] J. H. Thrall, X. Li, Q. Li et al., *Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success*, American College of Radiology, Reston, VA, USA, 2017.

[18] M. Torrisi, G. Pollastri, and Q. Le, “Deep learning methods in protein structure prediction,” *Computational and Structural Biotechnology Journal*, vol. 1, no. 9, 2020.

[19] X. Wu, D. Sahoo, C. Steven, and H. Hoi, “Recent advances in deep learning for object detection,” *Neurocomputing*, vol. 396, 2020.

[20] P. Hähnel, J. Mare, J. Monteil, and A. O’Donnch, “Using deep learning to extend the range of air pollution monitoring and forecasting,” *Journal of Computational Physics*, vol. 408, Article ID 109278, 2020.

[21] J. Katz, I. Pappas, S. Avraamidou, and E. N. Pistikopoulos, “Integrating deep learning models and multiparametric programming,” *Computers and Chemical Engineering*, vol. 136, 2020.

[22] S. Hayakawa and T. Suzuki, “On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces,” *Neural Networks*, vol. 123, pp. 343–361, 2020.

[23] M. T. Young, J. D. Hinkle, R. Kannan, and A. Ramanathan, “Distributed Bayesian optimization of deep reinforcement learning algorithms,” *Journal of Parallel and Distributed Computing*, vol. 139, pp. 43–52, 2020.

[24] H. Fujiyoshi, T. Hirakawa, and T. Yamashita, “Deep learning-based image recognition for autonomous driving,” *IATSS Research*, vol. 43, pp. 244–252, 2019.

[25] Y. Sun, L. Li, and L. Zheng, “Image classification base on PCA of multi-view deep representation,” *Journal of Visual Communication and Image Representation*, vol. 62, pp. 253–258, 2019.

[26] I. Rizwan I Haque and J. Neubert, “Deep learning approaches to biomedical image segmentation,” *Informatics in Medicine Unlocked*, vol. 18, 2020.

[27] B. Ma, X. Li, Y. Xia, and Y. Zhang, “Autonomous deep learning: a genetic DCNN designer for image classification,” *Neurocomputing*, vol. 379, 2019.

[28] C. Zhu, F. Song, Y. Wang et al., “Breast cancer histopathology image classification through assembling multiple compact CNNs,” *BMC Medical Informatics and Decision Making*, vol. 19, p. 198, 2019.

[29] L. C. C. Bergamasco and L. S. Nunes Fatima, “Intelligent retrieval and classification in three-dimensional biomedical images—a systematic mapping,” *Computer Science Review*, vol. 31, pp. 19–38, 2019.

[30] H. Cevikalp, B. Benligiray, and O. N. Gerek, “Semi-supervised robust deep neural networks for multi-label image classification,” *Pattern Recognition*, vol. 100, 2019.

[31] E. Miranda, M. Aryuni, and E. Irwansyah, “A survey of medical image classification techniques,” in *Proceedings of the International Conference on Information Management and Technology (ICIMTech)*, pp. 56–61, Bandung, Indonesia, November 2016.

[32] F. Jiang, Y. Jiang, H. Zhi et al., “Artificial intelligence in healthcare: past, present and future,” *BMJ Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 1–14, 2017.

[33] Y. Jiang, Z. Li, L. Zhang, and P. Sun, “An improved SVM classifier for medical image classification,” in *RSEISP*, et al. pp. 764–773, Springer-Verlag, Berlin, Germany, 2007.

[34] U. Javed, M. M. Riaz, A. Ghafoor, and T. A. Cheema, “MRI Brain Classification Using Texture Features, Fuzzy Weighting and Support Vector Machine,” *Progress in Electromagnetics Research B*, vol. 53, pp. 73–88, 2013.

[35] C. S. Lo and C. M. Wang, “Support vecto machine for breast MR image classification,” *Computers & Mathematics with Applications*, vol. 64, no. 5, pp. 1153–1162, 2012.

[36] S. Khan and S. P. Yong, “A deep learning architecture for classifying medical images of anatomy object,” in *Proceedings of APSIPA Annual Summit and Conference*, APSIPA, Kuala Lumpur, Malaysia, December 2017.

[37] Q. Li, W. Cai, Z. Wang, Y. Zhou, D. G. Feng, and M. Chen, “Medical image classification with convolutional neural network,” in *Proceedings of the 13th International Conference on Control, Automation, Robotics & Vision Marina Bay Sands (ICARCV)*, pp. 844–848, IEEE, Singapore, December 2014.

[38] J. Ker, L. Wang, J. Rao, and T. Lim, “Deep learning applications in medical image analysis,” *IEEE Special Section on Soft Computing Techniques for Image Analysis in the Medical Industry Current Trends, Challenges and Solutions*, vol. 6, pp. 9375–9389, 2018.

[39] C. Zhang, P. Yue, D. Tapete, B. Shangguan, M. Wang, and Z. Wu, “A multi-level context-guided classification method with object-based convolutional neural network for land cover classification using very high resolution remote sensing images,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 88, Article ID 102086, 2020.

[40] S. Vieira, W. H. L. Pinaya, and A. Mechelli, “Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications,” *Neuroscience and Biobehavioral Reviews*, vol. 74, pp. 58–75, 2017.

[41] J. Nalepa and M. Kawulok, “Selecting training sets for support vector machines: a review,” *Artificial Intelligence Review*, vol. 52, pp. 857–900, 2019.

[42] M. Badar, M. Haris, and A. Fatima, "Application of deep learning for retinal image analysis: a review," *Computer Science Review*, vol. 35, Article ID 100203, 2020.

[43] R. Yan, F. Ren, Z. Wang et al., "Breast cancer histopathological image classification using a hybrid deep neural network," *Methods*, vol. 173, 2019.

[44] M. Zareapoor, P. Shamsolmoali, D. K. Jain, H. Wang, and J. Yang, "Kernelized support vector machine with deep learning: an efficient approach for extreme multiclass dataset," *Pattern Recognition Letters*, vol. 115, 2017.

[45] Y. Fang, J. Zhao, L. Hu, X. Ying, Y. Pan, and X. Wang, "Image classification toward breast cancer using deeply-learned quality features," *Journal of Visual Communication and Image Representation*, vol. 64, Article ID 102609, 2019.

[46] S. B. Kotsiantis, "Supervised machine learning: a review of classification techniques," *Informatica*, vol. 31, pp. 249–268, 2007.

[47] G. Chen, R. Xu, and Z. Yang, "Deep ranking structural support vector machine for image tagging," *Pattern Recognition Letters*, vol. 105, pp. 1–9, 2017.

[48] V. F. Murilo, M. V. F. Menezes, L. C. B. Torres, and A. P. Braga, "Width optimization of RBF kernels for binary classification of support vector machines: a density estimation-based approach," *Pattern Recognition Letters*, vol. 128, pp. 1–7, 2019.

[49] O. Okwuashi and C. E. Ndehedehe, "Deep support vector machine for hyperspectral image classification," *Pattern Recognition*, vol. 103, 2020.

[50] Y. An, S. Ding, S. Shi, and J. Li, "Discrete space reinforcement learning algorithm based on support vector machine Classification," *Pattern Recognition Letters*, vol. 111, 2018.

[51] H. Almeida, M. J. Meurs, L. Kosseim, G. Butler, and A. Tsang, "Machine learning for biomedical literature triage," *PLoS One*, vol. 9, no. 12, Article ID e115892, 2014.

[52] A. Aldweesh, A. Derhab, and Z. A. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: a survey, taxonomy, and open issues," *Knowledge-Based Systems*, vol. 189, 2019.

[53] X. Yang, L. Wu, W. Ye, K. Zhao et al., "Deep learning signature based on staging CT for preoperative prediction of sentinel lymph node metastasis in breast cancer," *Academic Radiology*, vol. 27, no. 9, 2019.

[54] Y.-W. Chen and L. C. Jain, "Medical image classification using deep learning, deep learning in healthcare," *Intelligent Systems Reference Library*, vol. 171, 2020.

[55] J. Chen, S. Zhou, Z. Kang, and Q. Wen, "Locality-constrained group lasso coding for microvessel image classification," *Pattern Recognition Letters*, vol. 130, no. 5, 2020.

[56] C. Cao, F. Liu, H. Tan et al., "Deep learning and its applications in biomedicine," *Genomics Proteomics Bioinformatics*, vol. 16, pp. 17–32, 2018.

[57] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, pp. 102–127, 2019.

[58] "The impact of deep learning," *Zeitschrift für Medizinische Physik*, vol. 29, pp. 83-84, 2019.

[59] A. Maier, C. Syben, T. Lasser, and C. Riess, "A gentle introduction to deep learning in medical image processing," *Zeitschrift für Medizinische Physik*, vol. 29, pp. 86–101, 2019.

[60] S. A. Lashari and R. Ibrahim, "A framework for medical images classification using soft set," *Procedia Technology*, vol. 11, pp. 548–556, 2013.

[61] P. M. Ferreira, M. A. T. Figueiredo, and P. M. Q. Aguiar, "Content-based image classification: a non-parametric approach," 2018.

[62] L. Wang, Ed., *Support Vector Machines: Theory and Applications*, p. 503, Springer, Berlin, Germany, 2005.

[63] N. I. S. Bahari, A. Ahmad, and B. M. Aboobaider, "Application of support vector machine for classification of multispectral data," *IOP Conf. Series: Earth and Environmental Science*, vol. 20, pp. 1–8, 2014.

[64] I. Qabajeh, F. Thabtah, and F. Chiclana, "A recent review of conventional vs. automated cybersecurity anti-phishing techniques," *Computer Science Review*, vol. 29, pp. 44–55, 2018.

[65] A. Goyal, V. Gupta, and M. Kumar, "Recent named entity recognition and classification techniques: a systematic review," *Computer Science Review*, vol. 29, pp. 21–43, 2018.

[66] S. Dutta, B. C. S. Manideep, S. Rai, and V. Vijayarajan, "A comparative study of deep learning models for medical image classification," *IOP Conference Series: Materials Science and Engineering*, vol. 263, Article ID 042097, 2017.

[67] T. Huang, S. Wang, and A. Sharma, "Highway crash detection and risk estimation using deep learning," *Accident Analysis and Prevention*, vol. 135, Article ID 105392, 2020.

[68] M. A. Ferrag, L. Maglaras, and J. H. Moschoyiannis, "Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study Image classification," *Journal of Information Security and Applications*, vol. 50, Article ID 102419, 2020.

[69] Z. Chen, X. J. Wu, and J. Kittler, "Low-rank discriminative least squares regression for image classification," *Signal Processing*, vol. 173, 2020.

[70] X. Zhu, Z. Li, X. Li, S. Li, and F. Dai, "Attention aware perceptual enhancement nets for low-resolution image classification," *Information Sciences*, vol. 515, 2019.

[71] A. Di Ciaccio and G. M. Giorgi, "Deep learning for supervised classification," *Rivista Italiana di Economia Demografia e Statistica*, vol. LXVIV, no. 2, pp. 1–10, 2015.

[72] D. Jaswal, V. Sowmya, and K. P. Soman, "Image classification using convolutional neural networks," *International Journal of Advancements in Research & Technology*, vol. 3, no. 6, pp. 1661–1668, 2014.

[73] J. Kim, J. Hong, and H. Park, "Prospects of deep learning for medical imaging," *Precision and Future Medicine*, vol. 2, pp. 37–52, 2018.

[74] C. Parmar, J. D. Barry, A. Hosny, J. Quackenbush, and H. J. W. L. Aerts, "Data analysis strategies in medical imaging," *Clinical Cancer Research*, vol. 24, no. 15, pp. 3492–3499, 2018.

[75] J. Ahn, J. Park, D. Park, J. Paek, and J. Ko, "Convolutional neural network-based classification system design with compressed wireless sensor network images," *PLoS One*, vol. 13, no. 5, Article ID e0196251, 2018.

[76] X. Zhang, Y. Yang, and L. Shen, "Spark-SIFT: a spark-based large-scale image feature extract system," in *Proceedings of the 13th International Conference on Semantics, Knowledge and Grids*, pp. 69–76, IEEE, Beijing, China, August 2017.

[77] A. Tchagna Kouanou, D. Tchiotsop, R. Tchinda, C. Tchito Tchapga, A. N. Kengnou Telem, and R. Kengne, "A machine learning algorithm for biomedical images compression using orthogonal transforms," *Int. J. of Image, Graphics and Signal Processing (IJIGSP)*.vol. 10, no. 11, pp. 38–53, 2018.

[78] A. Tchagna Kouanou, D. Tchiotsop, Z. Djoufack Tansa'a, and R. Tchinda, "A machine learning algorithm for image compression with application to big data architecture: a comparative study," *iMedPub Journals, British Biomedical Bulletin*.vol. 7, no. 1, p. 316, 2019.

[79] A. Tchagna Kouanou, D. Tchiotsop, T. Fozin Fonzin, M. Bayangmbe, and R. Tchinda, "Real-time image

compression system using an embedded board," *Science Journal of Circuits, Systems and Signal Processing*, vol. 7, no. 4, pp. 81–86, 2018.

[80] C. Alla Takam, O. Samba, A. Tchagna Kouanou, and D. Tchiotsop, "Spark Architecture for deep learning-based dose optimization in medical imaging," *Elsevier Informatics*, vol. 29, pp. 1–13, 2020.

[81] C. L. Chowdhary and D. P. Acharjya, "Segmentation and feature extraction in medical imaging: a systematic review," *Procedia Computer Science*, vol. 167, pp. 26–36, 2020.

[82] S. Bhattacharya, P. K. R. Maddikunta, Q. V. Pham et al., "Deep learning and medical image processing for coronavirus (COVID-19) pandemic: a survey," *Sustainable Cities and Society*, vol. 65, p. 102589, 2021.

[83] C. L. Chowdhary, P. G. Shynu, and V. K. Gurani, "Exploring breast cancer classification of histopathology images from computer vision and image processing algorithms to deep learning," *International Journal of Advanced Science and Technology*, vol. 29, no. 3, pp. 43–48, 2020.

[84] Z. Wang, N. Dong, W. Dai, S. D. Rosario, and E. P. Xing, "Classification of breast cancer histopathological images using convolutional neural networks with hierarchical loss and global pooling," , 2018.