Building An SVM Classifier for Automated Selection of Big Data

Junhua Ding, Jiabin Wang
Department of Computer Science
East Carolina University
Greenville, NC, USA
dingj@ecu.edu, wangj15@students.ecu.edu

Xiaojun Kang
School of Computer Science
China University of Geoscience
Wuhan, Hubei, China
xj_kang@126.com

Xin-Hua Hu
Department of Physics
East Carolina University
Greenville, NC, USA
hux@ecu.edu

***Abstract*—The quality of big data could great impact the value extracted from the data. Automated filtering of noisy data from big data is an ideal approach for improving the quality of big data. However, due to large volume and variety of big data, automated filtering of noisy data from big data is a grand challenging task. In this paper, we propose a support vector machine based approach for automated classification of big data so that the noisy data are classified as separated categories from the regular data. In order to improve the classification accuracy and training performance, we design an experiment for improving the classification model through finding the optimized learning feature set and an approach for iteratively improving the quality of the training data set. We conducted a thorough experimental study of automated classification of massive image data of biology cells to explain the approach of automated selection of big data and demonstrate its effectiveness. Finally, we compare the performance of the SVM based classification and a deep learning based classification of the same data set. The proposed approach and experience collected from the experimental study can help big data researchers and practitioners to design strategies for improving the quality of big data, designing high performance classifier, and building tools for automated selection of big data.**

***Keywords*-machine learning, support vector machine, diffraction image, GLCM, feature selection, deep learning, big data**

I. Introduction

Big data has four characteristics defined on the volume of data, velocity of growing, variety of data types, and value to be extracted [1]. The volume and velocity of big data refer to the unprecedented amount of data and the speed of its generation, and the variety means big data are complex and heterogeneous. However, the most important characteristic of big data is its non-determined big values. Special tools and techniques such as new algorithms, scalable and high performance data processing infrastructure and analytics tools are needed to extract value from big data. However, the quality of the data great impacts the value extraction. Big data quality attributes include availability, usability, reliability, and relevance. Each attribute includes detail quality attributes such as credibility, integrity, and completeness [2] [1]. Existing research results have shown abnormal data can significantly decrease the accuracy of data analytics, which can damage the credibility of the values extracted from the data [3] [4]. In order to address the problem, one can improve the machine learning algorithm using for value extraction to handle poor data or to filter out the poor quality data to reduce their impact [5].

Due to the massive scale of big data, automated filtering of noisy data in big data using machine learning algorithms is a prefer choice. However, similar work is rare. In this paper, we introduce a machine learning based approach for automated separation of noisy data from massive scale biomedical image data. The noisy data include invalid data items and valid data items but they were incorrectly labelled known as class label noise. The approach is developed based on an support vector machine (SVM) [6] classifier for automatically classifying big data into several categories, where noisy data and regular data are classified into different categories. In order to improve the performance of the SVM classification, we design an approach for improving the SVM model and the quality of the data set. Improving the quality of the machine learning model or the training data set is the two fundamental approaches for improving the performance of machine learning techniques. In this paper, the performance of the SVM model is tuned using an optimized feature set, which is found by an extensive experimental study. The quality of the training data set is improved through multiple rounds of selection using the SVM classifier.

We introduce the proposed approach and demonstrate its effectiveness through classifying diffraction images of biology cells into three categories, which include two categories of noisy data and one category of regular data. Diffraction images of cells are acquired using a polarization diffraction imaging flow cytometer (p-DIFC), which was invented and developed by co-author Hu for quantifying and profiling 3D morphology of single cells [7]. The 3D morphological features of a cell captured in the diffraction image can be used for accurately classifying cell types, which is central to many branches of biology and life science research. Co-authors Ding and Hu have been studying cell morphology assay and classification for over a decade [8] [4] [9]. p-DIFC can take the diffraction images of near 100 cells each second, and we have collected large amount of diffraction images for different types of cells. Although different approaches for the classification of diffraction images were proposed, they are either too complex or low accurate. The problems are partially due to the low quality of the collected diffraction images. Cell samples for p-DIFC

978-1-5386-1996-4/17 $31.00 © 2017 IEEE
DOI 10.1109/BigDataCongress.2017.12

imaging include non-cell particles such as ghost cell bodies or aggregated spherical particles and cell debris. The diffraction images taken from non-cell particles are also collected together with the images taken from cells. When we train a machine learning classifier, the diffraction images taken from non-cell particles, which are noisy data, may impact the classification accuracy. It is necessary to remove the images of non-cell particles from the training data set.

Manually separating the noisy images from a large amount of diffraction image data is a heavy labor work. Therefore, we developed an SVM classifier for automated selection of diffraction images. In this approach, diffraction images are separated into three categories: diffraction images of viable cells of intact structures (or simply called as *cells*), diffraction images of ghost cell bodies or aggregated spherical particles (or simply called *fractured cells*), and diffraction images of cell debris or small particles (or simply called *debris*). In order to train the SVM classifier, each diffraction image is converted into a group of textual features calculated from its Gray-Level Co-Occurrence Matrix (GLCM) [10]. The SVM classifier is trained with the GLCM matrix of the training data set. Each training vector includes 32 GLCM features. Some of these features probably don't contribute to the classification accuracy or even decrease classification accuracy and performance. Feature selection is an important task in building machine learning models. Its approaches can be grouped into wrappers and filters. Wrappers use machine learning models to evaluate feature sets, and filters evaluate each feature with some criteria. We conducted an experimental study to find the optimized GLCM feature set for the SVM classifier. The training data is iteratively improved via multiple rounds of selection using the classifier to filter out the noisy images. The experimental result shows the classification accuracy of the SVM classifier for diffraction images is not high enough, but replacing SVM with a different machine learning algorithm such as deep learning could result in a high effective classifier. A new direction for automated selection of big data based on deep learning [11] is discussed and the preliminary result demonstrated the advantages of deep learning for classifying diffraction images.

The contributions of this paper can be summarized as follows: 1. Proposed a machine learning approach for automated selection of big data. noisy data in big data can be classified into separated categories by the machine learning classifier. The strategy for improving the performance of the classifier through improving the machine learning model and the data set is essential to the design of any machine learning based classifier. The quality of big data can be significantly improved through multiple rounds of filtering. The approach can be easily adapted for improving data quality in other domain specific applications of big data. 2. Conducted a comprehensive experimental study to find an optimized feature set for building an optimized SVM classifier. The experimental result demonstrates the correlation between a feature set and the classification accuracy. It offers an evidence to show how to improve the performance of a machine learning classifier through finding the optimized feature set. The experimental approach can be used for finding the optimized feature set for many other machine learning algorithms. 3. Compared the SVM classifier and a deep learning classifier and pointed out the new direction for automated selection of big data. The comparison study provides useful information for other big data researchers to select machine learning algorithms and feature representations for building an effective classifier.

The rest of this paper is organized as follows: Section 2 introduces the background of this research including cell imaging and SVM based automated classification of big data. Section 3 describes the design of an SVM classifier for automated classification of diffraction images and compares its performance to a deep learning classifier. Section 4 discusses the related work and Section 5 concludes the paper.

II. Classification of Diffraction Images

In this section, we first describe morphology based cell imaging and classification, and then discuss the background of automated classification of diffraction images using machine learning.

A. Morphology Based Cell Classification

Cells are basic elements of life and possess highly varied and convoluted 3-dimensional (3D) structures by intracellular organelles to sustain their phenotypic variations and functions. Cell classification are central to many branches of biology and life science research. Morphology based cell classification at the single-cell level attracts intense research efforts recently for their direct relations to cellular functions. p-DIFC is used to acquire cross-polarized Diffraction Image (p-DI) pairs such as s-polarization and p-polarization image pair from single cells [7]. The s-polarization image and p-polarization image are images acquired by only the s-polarization or the p-polarization of the scattered light, respectively. In this paper, each diffraction image pair includes an s-polarization image and a p-polarization image. Three sample diffraction image pairs are shown in Fig. 1. Different from images acquired by non-coherent light, the p-DI pairs present characteristic patterns due to the coherent light scatter emitted by the intracellular molecular dipoles induced by an incident laser beam. The p-DI data thus provide a big data source to probe the 3D morphology of the illuminated cells that requires powerful machine learning tools for extracting morphological and molecular information. During past decade, co-authors Hu and Ding *et al.* have developed different machine learning approaches including SVM and deep learning for rapid and accurate cell morphology analysis based on diffraction images of cells [8] [4] [12] [13]. But a systematic investigation of improving the machine learning performance hasn't been conducted.

B. GLCM

GLCM defines the textual pattern of an image with the statistics of the spatial relationship of pixels. It defines how often different combinations of gray level pixels occur in

an image for a given displacement/distance d in a particular angle θ. The distance d and angle θ refer to the distance and direction between the pixel under observation and its neighbor. The definitions of GLCM features for diffraction images include 14 original GLCM features and 3 extended features for diffraction images, which can be found at Ding *et al.* previous publications [8]. For example, GLCM feature *contrast* measures the local variations in an image, *correlation* means the linear dependence of gray levels between the pixels of neighboring gray tones, *homogeneity* measures the homogeneity of an image, and *entropy* measures the heterogeneity of an image. The 17 GLCM features quantitatively characterize the textual pattern in a diffraction image. Each diffraction image is converted into a group of feature values for SVM training and testing. We developed a parallel program using NVIDIA's CUDA on GPUs for calculating GLCM and the 17 features to achieve computational speedup. The size of the co-occurrence matrix scales quadratically with the number of gray levels in the image. The diffraction image in our study is normalized to an 8-bit gray-level range from the originally captured 14-bit image. Based on previous experimental results, we set the distance d to 1, and GLCM is calculated by average of GLCM calculated from 4 different angles, where θ is set to $\{0^o, 45^o, 90^o, 135^o\}$.

C. SVM based Image Classification

SVMs are supervised learning models associated with learning algorithms that build a set of hyperplanes in a high-dimensional space through analyzing data for classification or regression analysis [6]. An SVM performs binary classification in general. Given a training data set, each data item in the training data set is labelled by the category it belongs to or the other of two categories, and then an SVM training algorithm constructs a model to classify test items to one category or the other. However, several SVM classifiers can be combined to implement a multiclass classifier by comparing *'one against the rest'* or *'one against one'*. LIBSVM [14], an open source toolkit for SVM is used for conducting SVM classification in our projects. LIBSVM has wide range of nice features such as allowing the user to set different parameters, experiment different learning kernels, display useful statistics, and efficient multiclass classification.

SVM has been used for classifying the cell types based on diffraction images [13] [9] [4] [8]. The procedure of building an SVM classifier for diffraction images can be summarized as follows: 1. Calculate the GLCM features for each diffraction image in the training data set and the test data set. 2. Each diffraction image is labelled for its category such as *cells* or *debris*. A feature vector of a diffraction image is consisted of its GLCM feature values and its label. The feature vectors of all diffraction images in the training data set form a feature matrix. 3. Train the SVM classifier using the feature matrix. 4. Test the classifier with diffraction images in the test data set, and validate the classification accuracy using N-fold Cross Validation (NFCV) and confusion matrix. In order to improve the classification accuracy and performance, optimized features need to be identified.

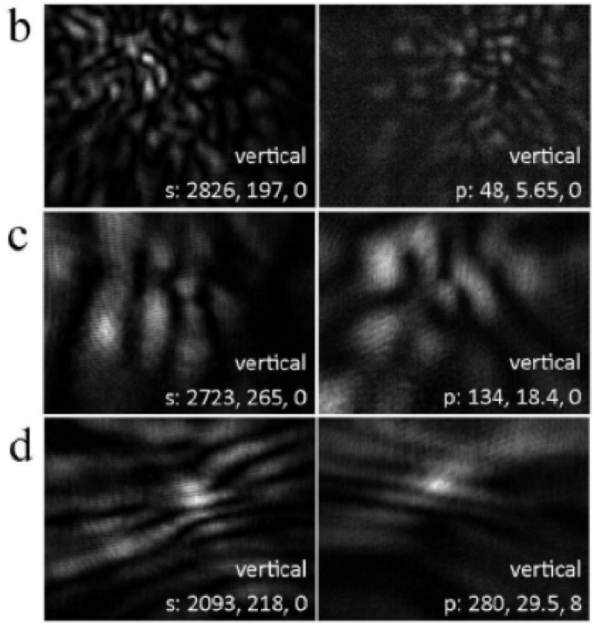

Fig. 1. Sample p-DIFC acquired diffraction image pair of (b) a viable cell of intact structures, (c) the cell debris or small particles, and (d) a ghost cell body or aggregated spherical particles.

III. Building An SVM Classifier for Diffraction Images

In this section, we describe how to build an SVM classifier for classifying the three categories of diffraction images: the images of viable cells of intact structures, the images of ghost cell bodies or aggregated spherical particles, and the images of cell debris or small particles. We also discuss how to find the optimized feature set for the SVM classifier using an experimental study, and finally we compare the SVM classification and a deep learning classification of the diffraction images.

A. Data Set

The textual patterns of the three categories of diffraction images are different. The diffraction image of a cell contains lots of normal speckle patterns, the diffraction image of a fractured cell consists of significant strip patterns, and the diffraction image of the debris generally includes small number of large diffuse speckle patterns. We label the three categories of diffraction images as *cells* for viable cells of intact structures, *strips* for ghost cell bodies or aggregated spherical particles, and *debris* for cell debris or small particles. The sample of each category of diffraction image is shown in Fig. 1.

We collected 3000 p-DIFC acquired diffraction image pairs, and each diffraction image pair includes an s-polarization diffraction image and its paired p-polarization image. Both the s-polarization and its paired p-polarization image are taken from the same particle using two different cameras. Therefore,

the data set includes 3000 s-polarization diffraction images and their paired 3000 p-polarization images. Within the 3000 diffraction image pairs, 1000 diffraction images pairs are manually labeled as *cells*, 1000 diffraction images pairs are labeled as *debris*, and 1000 diffraction image pairs are labeled as *strips*. Because SVM requires the training and test instances represented by the vector of numerical data, we assigned the the three-categories {cell, debris, strips} as {0,0,1}, {0,1,0} and {1,0,0}. Each diffraction image has 17 GLCM feature values. Therefore, a diffraction image pair includes 34 GLCM feature values. The same GLCM feature of the s-polarization or the p-polarization image is distinguished with "s-" or "p-" attached to its original name. Since the value of feature *Minimal Probability (MIP)*, which measures texture uniformity solely on the basis of the lowest probability of the pixel combinations, is 0 for all diffraction images in the data set, feature MIP is not counted in the classification of diffraction images. Then the number of combined features of a diffraction image pair is 32. The SVM classifier is built on SVM library LIBSVM, which supports multi-classification using SVM [14].

B. Feature Selection

Each diffraction image pair is converted into a feature vector consisting of 32 GLCM feature values and the label of the image category. The feature matrix to be used for training is the collection of the feature vectors of all images in the training data set. The test data set also includes a set of feature vectors of the test images. 10 fold cross validation (10FCV) and confusion matrix are used for validate the classification result. In 10FCV, the data is equally split into 10 groups where each group is held out for test in turn and the classifier is trained on the remaining nine-tenths; then its error rate is calculated on the holdout test set. This test procedure is repeated for a total of 10 times so that in the end, every image has been used exactly once for testing. Finally, the classification accuracy is averaged on the classification accuracy of the 10 tests. A confusion matrix presents the distribution of the test images in each category so that to understand how the test images are correctly classified for its real category and incorrectly classified for other categories.

We first train the SVM classifier using all 32 GLCM features. The average classification accuracy of 10FCV for category *cells, debris* and *strips* is 74.50%, 81.50% and 62.00% for the data set of diffraction images, respectively. The confusion matrix is shown in Table I. Although the classification accuracy is not bad, it is a concern that 24% diffraction images of the ghost cell body or aggregated spherical particles (labelled as strips) were incorrectly classified as cell debris or small particles, and 14% were incorrectly classified as cell viable cells of intact structures (labelled as cells). Over 1/3 diffraction images of category strips were incorrectly classified. Therefore, we need find a way to improve the performance of the classifier. The first attempt is to find an optimized feature set for the classifier. We conduct an experimental study to find the optimized feature set.

The process of the experiment is summarized as follows:

TABLE I
A CONFUSION MATRIX OF CLASSIFICATION WITH ALL FEATURES

	Cells	**Debris**	**Strips**
Cells	74.50%	16.00%	9.50%
Debris	6.50%	81.50%	12.00%
Strips	14.00%	24.00%	62.00%

1) Train and test the SVM classifier using only one GLCM feature each time and repeat 32 times so that every feature has a chance to be used for training and test. List the average classification accuracy for each feature. For simplicity reason, we measure the classification accuracy using the average classification accuracy for all three categories instead of one classification accuracy for each category.
2) Choose one GLCM feature that has the highest average classification accuracy from step 1, then select one more feature from the remaining 31 features and use the two features to train and test the SVM classifier and check the average classification accuracy. Repeat the two-feature training and testing step for each of the 31 remaining features. Then record the average classification accuracy of the two-feature training and testing that have the highest average classification accuracy.
3) Add one feature to the two-features set that was selected in step 2, and repeat the procedure in step 2 to experiment the training and testing with the third one. The third one is the one that is combined with the two-features set producing the highest average classification accuracy among the two features combining with each of the other 29 features.
4) Repeat the same procedure of step 3 for selecting 4 features, 5 features until all 32 features are checked.

The experimental result for classifying the three categories of diffraction images using different number of features is shown in Table II, where n step represents the number of features that are using for the training and testing and the feature column lists n^{th} feature that is added to the last feature set and produces the highest classification accuracy. For example, step 5 uses 5 features, which are *p-MAP, s-DIS, p-COR, s-SAV*, and *p-SVA*, and *p-SVA* is the feature that produces the highest classification accuracy with the feature set in step 4. The experimental result shows the average classification accuracy has the highest average classification accuracy with only 17 features instead of 32 full features. However, the difference between the optimized feature set and full feature set is only about 2%, a non-significant difference. It is also interesting to see that the 12-features set produced even slightly better average classification accuracy than the full-features set, which is consistent to Ding *et al.* previous results [8]. However, it is infeasible to experiment full combinations of all features even we tested many combinations. We also checked whether a different combination would produce better average classification accuracy.

In the second approach, we start the experiment with all

TABLE II
AVERAGE CLASSIFICATION ACCURACY OF DIFFERENT FEATURE SETS BUILT BY ADDING FEATURES

Step	Feature	Accuracy	Step	Feature	Accuracy
1	p-MAP	49.33%	17	p-ASM	74.83%
2	s-DIS	56.67%	18	s-ASM	74.50%
3	p-COR	59.33%	19	p-DVA	74.50%
4	s-SAV	63.67%	20	p-CLS	74.17%
5	p-SVA	65.50%	21	s-DVA	73.83%
6	s-VAR	66.83%	22	p-SAV	74.17%
7	p-IDM	67.83%	23	p-MEA	74.17%
8	p-SEN	70.33%	24	p-VAR	74.50%
9	s-CON	71.00%	25	p-CON	74.17%
10	p-ENT	71.33%	26	s-CLP	74.33%
11	s-MAP	72.17%	27	s-ENT	73.50%
12	p-DEN	73.17%	28	s-MEA	73.50%
13	p-CLP	73.33%	29	s-CLS	73.33%
14	s-DEN	73.83%	30	s-SEN	73.33%
15	s-IDM	74.00%	31	s-SVA	73.17%
16	s-COR	74.50%	32	p-DIS	72.67%

TABLE III
AVERAGE CLASSIFICATION ACCURACY OF DIFFERENT FEATURE SETS BUILT BY REMOVING FEATURES

Step	Feature	Accuracy	Step	Feature	Accuracy
1	p-DIS	73.17%	17	p-DEN	74.50%
2	s-SVA	73.33%	18	p-MEA	74.67%
3	p-DVA	73.50%	19	p-CON	73.67%
4	s-CLS	73.33%	20	s-COR	73.33%
5	p-MAP	73.83%	21	s-ASM	72.83%
6	s-SEN	74.33%	22	p-CLS	72.00%
7	s-DVA	74.50%	23	p-SAV	72.50%
8	p-VAR	74.67%	24	s-DEN	71.67%
9	p-ENT	74.83%	25	s-SAV	70.33%
10	p-CLP	74.67%	26	s-CON	69.33%
11	p-ASM	74.33%	27	s-VAR	66.50%
12	s-ENT	74.33%	28	p-COR	64.50%
13	p-SVA	74.50%	29	s-MAP	60.67%
14	s-DIS	74.67%	30	p-IDM	53.83%
15	s-MEA	74.17%	31	s-IDM	47.83%
16	s-CLP	74.83%	32	p-SEN	n/a

32 GLCM features first, and then remove one feature each time to check the average classification accuracy using the new feature set. Each time, the feature that decreases the most of the classification accuracy when it is deleted from the feature set is removed from the future feature set. Therefore, the experiment repeats 32 times to find the 31 features would produce the highest classification accuracy in the first step, and then repeats 30 times in the second step and so on until the feature set only has one feature. The experimental result for classifying the three categories of diffraction images using different number of features is shown in Table III, where n step represents the number of features are deleted from the feature set and the feature column lists n^{th} feature that is removed from the feature set. From the experimental results, it is easy to see that two sets of feature sets produced the highest classification accuracy, and one feature set includes 23 features, the other includes 16 features. The three features sets including the one found in last section that produced the highest average classification accuracy (all of them are 74.83%) are different. In addition, we see that as soon as the feature set includes 12 or more features, the average classification accuracy among the feature sets is only slightly different so that the difference could be ignored. The observation was further confirmed by other experiment with different feature combinations. Therefore, we conduct additional experiments to improve the classification accuracy of the SVM classifier through improving the quality of the data set.

C. Data Quality Improvement

There are two basic ways to improve the performance of the machine learning classifier: find a better machine learning model or use better training data. One way to improve the quality of data is to filter noisy data from training data set. The noisy data include invalid data and class label noisy data. In this section, we discuss different attempts to improve the quality of training data.

1) Iterative Selection of Data: Each p-DIFC acquired diffraction image is labelled for its category by biologists. As we see in Fig. 1, the difference of the textual pattern among the three categories of diffraction images could be very small such as a strips pattern could be very similar to a normal speckle pattern and vice versa. It is very difficult to know whether these images were correctly labelled or not since the original samples don't exist anymore. It is easy to understand some of these diffraction images could be incorrectly labelled due to the confusion of the images. Some diffraction images could be accidentally incorrectly labelled. In order to improve the quality of the training data, we need separate the vague diffraction images and class label noisy images and then remove the vague images and correct the label of the class label noisy images. We first conduct a 10FCV of all diffraction images, and mark the images that are incorrectly classified by the SVM classifier. Then we remove the marked images from the data set since we believe these images are low quality. We conducted one more 10FCV of the new data set using the SVM classifier. However, our experimental result showed the new "better" data set results in a lower classification accuracy. Table IV is the confusion matrix of the 10FCV result of the classification with the original data set, and Table V is the confusion matrix of the classification with the "better" data set, which doesn't include the data that were incorrectly classified in the first round of 10FCV. It is easy to see that the classification accuracy of debris and strips with the reduced data set has been significantly decreased. We suspect that majority of the removed images actually were high quality images, which means they have significant features to be correctly classified and they are correctly labelled in the data set. Removing large portion of the data such as over 1/3 of strips images from the original data set could great impact the performance of the classification especially considering only 1000 diffraction image pairs for each category in this study. Therefore, it is necessary to manually check the incorrectly classified images and removed only the real noisy images. Since around 1500 diffraction images (i.e., 750 DI pairs) including p-polarization and s-polarization images were incorrectly classified in the first round of experiment, we only manually inspected 600

TABLE IV
A CONFUSION MATRIX OF CLASSIFICATION WITH ORIGINAL DATA

	Cells	Debris	Strips
Cells	77.50%	11.00%	11.50%
Debris	6.00%	81.50%	12.50%
Strips	14.00%	20.50%	65.50%

TABLE V
A CONFUSION MATRIX OF CLASSIFICATION WITH REDUCED DATA

	Cells	Debris	Strips
Cells	77.50%	14.00%	8.50%
Debris	11.00%	76.00%	13.00%
Strips	19.00%	27.00%	54.00%

images and removed the images that were extremely difficult to be classified by experts and relabeled the images that were incorrectly labelled. We used the new data set to conduct a 10FCV, and average classification accuracy is improved from 74.83% to 80.33%. The classification accuracy of cells is improved from 77.50% to 88.76%, and debris from 81.50% to 88.75%. Then if we check the new classification result again and manually remove the noisy images one more time, it is possible the classification accuracy could be improved more.

2) Pre-selection of Data: Hu *et al.* experimented a different approach for improving the quality of the data set [4]. In the approach, the data are pre-selected by image processing and clustering to remove low quality data [4]. The process of the pre-selection of diffraction images is summarized as follows:

1) Find a set of borderline length parameters for differentiating the strips pattern from the speckle pattern. It can be completed through converting the image into binary based on the threshold of the average of the pixel intensity and then applying Sober operators to find the boardlines and measure the length.
2) Measure the speckle size using the frequency histogram calculated using 2D fast Fourier transform (FFT) for each diffraction image.
3) K-means clustering algorithm is applied to calibrate image data and separate diffraction images into strips and speckles, which are further separated as normal speckles and large diffuse speckles.
4) Classify the calibrated diffraction images into large diffuse speckles ($i.e.$, debris) and normal speckles ($i.e.$, normal cells) using SVM [4].

The experimental result showed the classification accuracy of the three categories of diffraction images closes to 100%. Therefore, pre-selection of training data using a clustering algorithm to remove noisy data is an effective approach for improving the quality of big data. It further confirmed that data quality is important to the machine learning performance. However, the pre-selection for the data quality improvement is very complex and it is difficult to be adapted for other domain specific big data applications.

3) Single Polarization Image Data: p-DIFC takes one s-polarization image and one p-polarization image for each particle each time. The s-polarization image and p-polarization

TABLE VI
A CONFUSION MATRIX OF CLASSIFICATION WITH P-POLARIZATION DIFFRACTION IMAGES

	Cells	Debris	Strips
Cells	77.50%	11.50%	11.00%
Debris	26.00%	66.50%	7.50%
Strips	26.50%	31.00%	42.50%

TABLE VII
A CONFUSION MATRIX OF CLASSIFICATION WITH S-POLARIZATION DIFFRACTION IMAGES

	Cells	Debris	Strips
Cells	47.50%	40.00%	12.50%
Debris	14.00%	69.50%	16.50%
Strips	16.00%	30.50%	53.50%

TABLE VIII
A CONFUSION MATRIX OF CLASSIFICATION WITH SINGLE-POLARIZATION DIFFRACTION IMAGES

	Cells	Debris	Strips
Cells	82.50%	7.25%	10.25%
Debris	48.50%	26.50%	25.50%
Strips	38.50%	12.75%	48.75%

image would capture different morphology information from the particle. It is interesting to check how the paired images increase the classification accuracy and how the low quality data impact the classification accuracy. We experimented different data sets that include only single polarization such as only s-polarization or p-polarization images, or individual images without considering polarization. We experimented the classification with the three different data sets of individual diffraction images: the first one only contains s-polarization images (i.e., the data set has only the 3000 s-polarization images), the second one only includes p-polarization images (i.e., the data set has only the 3000 p-polarization images), and the third one separates p-polarization images and s-polarization images without pairing them (i.e., the data set includes 1500 s-polarization images and 1500 p-polarization images). Since the feature optimization was built based on diffraction image pairs, they are not appropriate for the single diffraction images. The experiment on the single polarization images used all 16 GLCM features. The confusion matrices of 10FCV classification results are shown in Tables VI, VII and VIII. Comparing the results to the classification result with all features shown in Table I, the data sets that only have single polarization diffraction images resulted in a much lower classification accuracy. The classification with mixed single polarization of diffraction images has a very poor classification accuracy. It further confirms a high quality data set is important to the performance of the SVM classification. In next section, we will see a better classification model (i.e. the deep CNN model) would result in a high classification accuracy even with a low quality data set.

D. A Deep Learning Classifier

A deep learning neural network has multiple hidden layers. One of the promises of deep learning is replacing handcrafted

features with efficient algorithms for unsupervised or semi-supervised feature learning and hierarchical feature extraction [15]. Various deep learning architectures such as convolutional deep neural networks, deep belief networks and recurrent neural networks have been applied to fields like computer vision, automatic speech recognition, natural language processing, audio recognition and bioinformatics where they have been shown to produce state-of-the-art results on various tasks. Convolutional Neural Network (CNN) is a widely used deep learning network for image classification. CNN AlexNet has been widely cited for its success in 2012 Large Scale Visual Recognition Challenge (ILSVR2012). Since then, many sophisticated and deeper CNNs have been proposed for image classification in ILSVR such as VGG [16] and GoogLeNet [17].

Comparing to other images, diffraction images are relatively simple due to its low resolution and no background noisy. Therefore, we select the model of AlexNet implemented in Caffe to build the deep learning classifier for the classification of diffraction images thanks to its relatively simple architecture. AlexNet includes 5 convolutional layers and multiple max-polling layers in addition to 3 fully connected layers, whose output of the last layer is fed to a 1000-way softmax to produce a distribution over the 1000 classes [18]. Each convolutional layer filters every channel of the input image with multiple kernels. Due to the large number of features used in a deep learning, the volume of the training data set required for a deep learning is also large. The original AlexNet was trained with 1.2 million images. We prepared 100,000 diffraction images for each category of the three categories of diffraction images based on several thousand of original diffraction images that include the majority of the 3000 diffraction images that were used for building the SVM classifier. Each original diffraction image is downsampled into many small diffraction images in size 227 * 227 pixels, which is the image size accepted by AlexNet. Both cropping and pooling techniques were used for downsampling the original images and produce a large amount of training images. A small image is labelled as the original image where the small image is cropped or pooled from. Table IX shows a confusion matrix of 8FCV of the classification of the three categories of the diffraction images using the deep learning classifier. From the example, it is easy to see that the classification accuracy of deep learning classifier is much higher than the SVM classifier. However, deep learning is not able to be applied to the original diffraction images directly, and a large amount of artificial diffraction images have to be produced from the original images to build a large enough training data set. In addition, training the deep learning classier is much slower than training the SVM classifier.

TABLE IX
A CONFUSION MATRIX OF DEEP LEARNING CLASSIFICATION

	Cells	**Debris**	**Strips**
Cells	94.20%	3.90%	1.90%
Debris	1.60%	97.50%	0.90%
Strips	4.30%	5.40%	90.30%

E. Discussion

In this section, we described an SVM classifier for classifying three categories of diffraction images. In order to improve the accuracy of the classification, extensive experiments were conducted to find an optimized feature set for the SVM classifier. The experimental study showed more than one optimized feature set may exist, but the difference between an optimized feature set and the full features set is very small in term of the classification accuracy. The experimental study showed full features are not necessary to result in a higher classification accuracy than a subset of full features. High quality data set is also important to ensure the quality of machine learning. The quality of a data set can be improved via multiple rounds of classification through removing low quality data and correct class label noisy data. We conducted an experimental study to compare the classification accuracy of the SVM classifier applying to diffraction image pairs and single polarization diffraction images. The experimental result showed the quality of the data set is important to the performance of the machine learning classification. Finally, we compared the classification accuracy of the SVM classifier and a deep learning classifier for the diffraction images. The classification accuracy of the deep learning classifier was much higher than the SVM classifier. However, the incompatible size of the diffraction image, the limited number of available diffraction images, and highly demand of deep learning computing hardware could cause difficulties for building an effective deep learning classifier.

IV. RELATED WORK

Quality of big data would greatly impact the value extraction from big data. Poor quality data could cause serious problems such as wrong prediction or low accuracy of the classification. The quality attributes of big data such as availability, usability and reliability have been well defined in some publications [2] [1]. Gao *et al.* have given an overview of the issues, challenges and tools of validation and quality assurance of big data [19], where they defined big data quality assurance as the study and application of quality assurance techniques and tools to ensure the quality attributes of big data. Although general techniques and tools were developed for quality assurance of big data, much more work are on the quality assurance of domain specific big data such as health care management data, social media data and finance data. For example, there are much work on the evaluation of the veracity of web sources [20] [21]. Finding the duplicated information from different data sources is an important task of quality assurance of big data. Machine learning algorithms such as Gradient Boosted Decision Tree (GBDT) were used for detecting the duplication [22]. Data filtering is an approach for quality assurance of big data through removing bad data from data sources. In this paper, we proposed an SVM approach for automated classification of large scale of diffraction images data to select needed data from a large amount of diffraction images that may include lots of noisy data. The impact of the class label

noisy data and invalid data can be iteratively reduced through multiple rounds of selection.

Machine learning researchers have to make the trade-off between using better learning models and using better training data when they look for a machine learning based classification solution [23]. Feature selection approaches can be classified into two groups: wrapper approaches and filter approaches. The two approaches are used together in many cases. Thati *et al.* [8] reported their work on feature selection for classifying diffraction images, where an experimental approach called Extensive Feature Correction Study (EFCS) was used to select optimized GLCM features for classifying diffraction images. The selected feature set was cross checked with an algorithm based feature selection approach called Correlation based Feature Selection Algorithm (CFS). Feature selection guided by combinatorial method has been reported. Dreisetitl *et al.* [24] proposed and experimented a feature selection based on the classification performance of pairwise feature sets. In this paper, we checked how a machine learning model can be improved via feature optimization and demonstrate how the data quality can impact the performance of machine learning. The feature subsets were evaluated in forward and in backward through adding or removing one feature each time for the highest classification accuracy.

V. Summary and Future Work

The large volume and variety of big data require an automated approach for the selection of different categories of data. In this paper, we introduced an SVM approach for automated big data selection. The approach includes an experimental study for feature selection and iterative data improvement through multiple rounds of data selection to ensure the quality of the classification. We demonstrated how the quality of the data set could impact the classification accuracy of the SVM classifier, and how the SVM classifier could be improved through feature optimization and how different machine learning algorithms could affect the classification accuracy through comparing the SVM classifier and a deep learning classifier. The experimental study explained the strategy for improving the quality of machine learning based classification through machine learning model improvement and data quality improvement. The machine leaning based approach for automated selection of big data approach can be easily adapted for other domain specific big data applications.

Acknowledgment

This research is supported in part by grants #1262933 and #1560037 from the National Science Foundation. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

[1] V. Gudivada, R. Raeza-Yates, and V. Raghavan, "Big data: Promises and problems," *IEEE Computer*, vol. 48, no. 3, pp. 20–23, 2015.

[2] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Science Journal*, vol. 14:2, pp. 1–10, 2015.

[3] E. Giannoulatou, S.-H. Park, D. Humphreys, and J. Ho, "Verification and validation of bioinformatics software without a gold standard: a case study of bwa and bowtie," *BMC Bioinformatics*, vol. 15(Suppl 16):S15, 2014.

[4] J. Zhang, Y. Feng, M. S. Moran, J. Lu, L. Yang *et al.*, "Analysis of cellular objects through diffraction images acquired by flow cytometry," *Opt. Express*, vol. 21, no. 21, pp. 24 819–24 828, 2013.

[5] J. A. Saez, B. Krawczyk, and M. Wozniak, "On the influence of class noise in medical data classification: Treatment using noise filtering methods," *Applied Artificial Intelligence*, vol. 30, no. 6, pp. 590–609, Jul. 2016.

[6] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, Jan. 1998.

[7] K. Jacobs, J. Lu, and X. Hu, "Development of a diffraction imaging flow cytometer," *Opt. Lett.*, vol. 34, no. 19, p. 29852987, 2009.

[8] S. K. Thati, J. Ding, D. Zhang, and X. Hu, "Feature selection and analysis of diffraction images," in *4th IEEE Intl. Workshop on Information Assurance*, Vancouver, Canada, August 2015.

[9] Y. Feng, N. Zhang, K. Jacobs, W. Jiang, L. Yang *et al.*, "Polarization imaging and classification of jurkat t and ramos b cells using a flow cytometer," *Cytometry A*, vol. 85, no. 11, pp. 817–826, 2014.

[10] R. Haralick, "On a texture-context feature extraction algorithm for remotely sensed imagery," in *Proceedings of the IEEE Computer Society Conference on Decision and Control*, Gainesville, FL, Dec. 1971, pp. 650–657.

[11] Y. Bengio, "Learning deep architectures for ai," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[12] J. Ding, D. Zhang, and X. Hu, "An application of metamorphic testing for testing scientific software," in *1st Intl. workshop on metamorphic testing with ICSE*, Austin, TX, May 2016.

[13] K. Dong, Y. Feng, K. Jacobs, J. Lu, R. Brock *et al.*, "Label-free classification of cultured cells through diffraction imaging," *Biomed. Opt. Express*, vol. 2, no. 6, p. 17171726, 2011.

[14] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[15] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug 2013.

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

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015. [Online]. Available: http://arxiv.org/abs/1512.00567

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012, pp. 1097–1105.

[19] J. Gao, C. Xie, and C. Tao, "Big data validation and quality assurance – issuses, challenges, and needs," in *2016 IEEE Symposium on Service-Oriented System Engineering (SOSE)*, March 2016, pp. 433–441.

[20] X. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang, "Knowledge-based trust: Estimating the trustworthiness of web sources," *Proc. VLDB Endow.*, vol. 8, no. 9, pp. 938–949, May 2015.

[21] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," *IEEE Trans. on Knowl. and Data Eng.*, vol. 20, no. 6, pp. 796–808, Jun. 2008.

[22] C. H. Wu and Y. Song, "Robust and distributed web-scale near-dup document conflation in microsoft academic service," in *2015 IEEE International Conference on Big Data (Big Data)*, Oct. 2015, pp. 2606–2611.

[23] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[24] S. Dreiseitl and M. Osl, "Feature selection based on pairwise classification performance," in *Computer Aided Systems Theory - EUROCAST 2009*, 2009, pp. 769–776.