Breast Cancer Biopsy Predictions Based on Mammographic Diagnosis Using Support Vector Machine Learning

Julia H. Miao, Kathleen H. Miao, and George J. Miao, *Senior Member, IEEE*

***Abstract*—Globally, breast cancer is one of the major causes of cancer death in females. Of the available diagnostic methods for breast cancer, mammography is commonly used as a non-invasive method for distinguishing malignant tumors from benign ones. However, its diagnosis accuracy varies widely; surgical biopsy, an expensive and invasive surgery, is typically needed to confirm a tumor's state of malignancy. In this research, a mammographic diagnostic method is presented for breast cancer biopsy outcome predictions using support vector machine (SVM) learning classification. The developed SVM learning classification is a nonlinear classifier based on a Gaussian radial basis function (RBF) kernel, which allows more flexibility in dealing with any non-separable mammographic mass data. The developed SVM learning classification can provide a not only higher but also more reliable percentage of accuracy in diagnosing malignant breast cancer and benign disease for breast biopsy outcome predictions. The testing results showed that the developed SVM learning classification had a sensitivity (or recall) of 94.54% in diagnosing malignant breast cancer, a specificity of 93.44% in diagnosing benign disease, a precision of 93.15%, a *F*-score of 0.94, and an overall accuracy of 93.98% in diagnosing both malignant breast cancers and benign disease. Furthermore, an estimated area of the receiver operating characteristic (ROC) curve analysis and its associated standard error was 0.9630±0.0516 for breast biopsy outcome predictions, which outperformed the diagnostic accuracies of previously reported methods. Therefore, the developed SVM learning classification with mammography can provide highly accurate and consistent diagnoses in distinguishing malignant and benign cases for breast cancer biopsy outcome predictions, thus reducing the number of unnecessary biopsies for patients.**

***Index Terms*—breast cancer, benign disease, biopsy, Gaussian radial basis function (RBF) kernel, malignant breast cancer, mammography, precision, receiver operating characteristic (ROC) curve, sensitivity, specificity, support vector machine (SVM)**

Manuscript received September 9, 2015; revised October 12, 2015.

Julia H. Miao is with Cornell University, Ithaca, NY 14853, USA (e-mail: jhm344@cornell.edu).

Kathleen H. Miao is with Cornell University, Ithaca, NY 14853, USA (e-mail: khm37@cornell.edu).

George J. Miao is with Flezi, LLC, San Jose, CA 95134, USA (e-mail: g.j.miao@ieee.org).

I. Introduction

EVERY year, 14 million people are diagnosed with cancer, and 8 million people worldwide die from cancer according to the Center for Disease Control and Prevention [1]. In the United States, cancer is the second leading cause of death [2]. Specifically, among all cancer cases for females, breast cancer is ranked as the second leading causes of cancer death and new cancer cases [3]-[5].

Breast cancer usually forms lumps or masses referred to as tumors [3]. Most breast cancer tumors at early stages are benign, which are considered diseases that are not yet malignant and life-threatening [6]. Thus, before symptoms develop, early detection of breast cancer masses is one of the most important factors influencing patients' chances of long-term survival.

Despite the sensitivity of mammography in detecting breast cancer, the positive predictive value of breast biopsy outcomes is low. Likewise, mammography lacks high diagnostic accuracy in distinguishing malignant breast cancer and benign disease. Its diagnostic accuracy is reported to range anywhere from 68% to 79% [7]. False-negative results (FNR) and false-positive results (FPR) are also problems [8]. FNR occur when mammograms appear normal even though breast cancer is present, leading to delays in treatment for affected breast cancer patients. On the other hand, FPR occur when radiologists decide mammograms are abnormal, but no breast cancer is actually present.

Consequently, when using mammography to detect a breast cancer tumor, surgical biopsy, which has a reported accuracy of breast cancer diagnosis close to 100% [7], is usually needed to confirm the state of its malignancy [9]; as a result, the low positive predictions of mammogram explanations lead to a high number of unnecessary biopsies for benign outcomes [10]. In fact, several hundreds of thousands of unnecessary biopsies are performed on benign rather than malignant cases each year [11], [12]. However, surgical biopsy is expensive and invasive, which can often not be suitable for the patient.

To increase mammographic accuracy, various computer aided diagnosis (CAD) systems were developed to distinguish malignant breast cancer and benign disease to predict biopsy outcomes. A breast image reporting and database system (BIRADS), established by the American College of Radiology, is the most common way for radiologists to

describe mammogram findings and to make an assessment to support the decision of a physician to perform a breast biopsy [13]. Using BIRADS with various characteristics, such as mass shape, obtained from a mammogram, several CAD approaches were reported to predict breast cancer biopsy outcomes based on an intelligible decision process [10] and a case-based reasoning classifier using different similarity measures based on Euclidean and Hamming distances [12], [14]-[16]. Other CAD methods included an artificial neural network approach based on BIRADS descriptions [17], [18], a classification based on a decision tree approach [19], and a prediction model based on a distributed genetic programming approach [20]. Recently, one article reported mammographic diagnosis for breast cancer biopsy predictions using a neural network classification model [21]. These CAD methods were proposed to predict breast cancer biopsy outcomes and/or to classify malignant and benign lumps using mammogram data.

Recently, another type of CAD method using a support vector machine [22] had been applied in short-term prognosis evaluation of breast cancer patients in terms of survival or recurrence outcomes after a given follow-up period. Another variation method, a relevance vector machine [23], was used in cancer classification, which is made according to the site of origin of the malignant cells.

For this research, an enhanced statistical learning approach is developed for mammographic diagnosis of breast cancer biopsy outcome predictions utilizing support vector machine (SVM) learning classification. The developed SVM learning classification contained two separate models: a learning classification (or training) model and a diagnostic (or prediction) model. The learning classification model was a SVM nonlinear classifier based on a Gaussian radial basis function (RBF) kernel, which computed the inner product in feature space between two vector arguments. The diagnostic model distinguished and classified malignant breast cancers and benign diseases for breast cancer biopsy outcome predictions using new patient data. The probabilities of misclassification error and prediction accuracy as well as the performance of the developed SVM learning classification were evaluated using the model sensitivity, specificity, precision, *F*-score, and receiver operating characteristic (ROC) curve analysis.

The test results of the developed SVM learning classification along with mammography can provide highly accurate and consistent diagnoses in distinguishing malignant and benign cases for breast cancer biopsy outcome predictions. Therefore, the developed SVM learning classification model with mammography can reduce the number of unnecessary breast biopsies for patients with benign outcomes.

II. Materials and Methods

In this section, we first present the mammographic dataset of breast cancer tumors. Then, prediction methods of the developed SVM learning classification along with its sensitivity, specificity, precision, *F*-score, and ROC curve analysis are introduced in detail, including a nonlinear SVM learning classification using a Gaussian RBF kernel, a diagnostic (or prediction) model to distinguish and classify malignant breast cancers and benign diseases for breast biopsy outcome predictions, and their corresponding algorithms, approaches, and implementation architecture.

A. Mammographic Mass Dataset

The mammographic mass dataset was obtained from the Mammographic Mass Database, which is available in the UCI Machine Learning Repository [24]. There are a total of 961 clinical instances including 516 benign and 445 malignant cases in the mammographic mass dataset. Among all of the clinical instances, 131 clinical instances have missing attribute values, which were removed from the dataset in this research. This resulted in a dataset of 830 clinical instances, in which 427 clinical instances are benign disease cases and 403 clinical instances are malignant breast cancer cases. Each clinical instance in the mammographic mass dataset contains five input attributes, including BIRADS, age, mass shape, mass margin, and mass density, as well as one class attribute of severity. The class attribute of severity is a binary value of 0 or 1, which indicates benign disease or malignant cancer diagnoses, respectively.

B. Nonlinear SVM Learning Methods

In this section, the mammographic diagnostic method to classify malignant breast cancers and benign diseases for breast biopsy outcome predictions is established using a nonlinear SVM learning classification. The nonlinear SVM learning classification and diagnostic (or prediction) models, associated with their algorithms, methods, and implementation architecture are presented in detail. The performances of the nonlinear SVM learning classification and diagnostic model were evaluated by using the model sensitivity, specificity, precision, *F*-score, and the model ROC curve analysis.

The SVM Learning Classification and Diagnostic Models

SVM is a kernel-based machine learning method derived from statistical learning theory [25], [26]. Since the publication of its algorithm and theory, SVM has become one of the most important tools for classification, prediction, and regression [27]. In particular, SVM has shown promise in a variety of medical and biological classifications on tumor types, such as brain and lung cancer [28], [29], leukemia [30], and lymphoma [31], as well as gene expression [32].

In binary classification problems, N training data $\{(\boldsymbol{x}_1, t_1),\ldots, (\boldsymbol{x}_N, t_N)\}$ are given, where $\boldsymbol{x}_i$ is a vector corresponding to an input sample data, including n input attributes, and t_i is a binary class label with ± 1. In this research, the n input attributes were 5 input attributes for the Mammographic Mass Dataset. In most cases, the N training data were often not linearly separable. To deal with this situation, a SVM learning classification model is used to find the function, $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$, and ξ_i by solving the following optimization problem [33]-[36]:

$$\min_{w,b,\xi} \left\{\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_i \xi_i\right\}, \text{ Subject to } t_i(\boldsymbol{w}\boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad (1)$$

where $\xi_i \geq 0$ for *i = 1,2,..., N*. The formulation in Equation (1) has a trade-off of two objectives: (1) finding a hyperplane with a large margin between two groups of data by minimizing the first term of $\frac{1}{2}\|\boldsymbol{w}\|^2$; and (2) finding a hyperplane that can separate the training data well by minimizing the second term of $C\sum_i \xi_i$. The parameter *C* is used to control the trade-off. In other words, the second term of $C\sum_i \xi_i$ is used to reduce the number of training errors in the case of data that are nonlinearly separable. Equation (1) is also referred to as the *soft margin* SVM. Thus, the basic idea behind SVM is to search for a balance point between the regularization of the first term of $\frac{1}{2}\|\boldsymbol{w}\|^2$ and the regularization of the training errors. Through this balance, the SVM is thereby able to achieve a high accuracy of classification in an optimal sense.

Using Lagrange multipliers of $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$, the previous constrained problem in Equation (1) can be expressed as follows:

$$L(\boldsymbol{w},\boldsymbol{b},\boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left[t_i(\boldsymbol{w}\boldsymbol{x}_i + b) - 1 + \xi_i\right] - \sum_{i=1}^{N}\mu_i\xi_i \tag{2}$$

where $\alpha_i \geq 0$ and $\mu_i \geq 0$ for *i = 1,2,..., N*.

To establish a nonlinear separating surface, SVM performs a nonlinear mapping of transferred data from a lower dimensional space to a high dimensional space given by:

$$\boldsymbol{x} \rightarrow \emptyset(\boldsymbol{x}), \tag{3}$$

where $\emptyset(\boldsymbol{x})$ is a general transformation function. Then, SVM is applied based on a linear separating hyperplane in the feature space, which corresponds to a nonlinear surface in the original feature space.

To optimize $\boldsymbol{w}$, $\boldsymbol{b}$, and ξ_i, derivatives are applied on Equation (2) by solving: $\frac{\partial L}{\partial \boldsymbol{w}} = 0$, $\frac{\partial L}{\partial \boldsymbol{b}} = 0$, and $\frac{\partial L}{\partial \xi_i} = 0$. Then, the results are obtained:

$$\boldsymbol{w} = \sum_{i=1}^{N}\alpha_i\, t_i \emptyset(x_i), \tag{4}$$

$$\sum_{i=1}^{N}\alpha_i\, t_i = 0, \tag{5}$$

$$\alpha_i = C - \mu_i. \tag{6}$$

Using Equations (4), (5), and (6), a dual Lagrange is obtained in the following form:

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\, t_i t_j K(x_i, x_j), \tag{7}$$

where $K(x_i, x_j) = \emptyset(x_i)\emptyset(x_j)$ is referred to as a kernel function.

Substituting Equation (4) into $f(\boldsymbol{x})$, we obtain the function:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N}\alpha_i\, t_i K(\boldsymbol{x}, x_i) + b. \tag{8}$$

Equation (8) is a statistical classification model, which can be used to predict new data points for each class. In this research, Equation (8) was called as a kernel-based diagnostic (or prediction) model for the mammographic mass data. If a subset of the data points has $\alpha_i = 0$, the data points do not contribute to the statistical classification model in this case. Thus, the remaining data points are called *support vectors*. These data points should have $\alpha_i > 0$, thereby leading to

$$t_i f(\boldsymbol{x}_i) = 1 - \xi_i. \tag{9}$$

If $\alpha_i < C$, then Equation (6) indicates that $\mu_i > 0$, which requires $\xi_i = 0$ since $\mu_i\xi_i = 0$ based on a constrained optimization of the Karush-Kuhn-Tucker (KKT) conditions [35]. Hence, these data points lie on the margin. If $\xi_i \leq 1$, the data point with $\alpha_i = C$, which is inside the margin, will be classified correctly; if $\xi_i > 1$, the data point with $\alpha_i = C$ will be classified incorrectly.

For a diagnostic model of classifying malignant breast cancer and benign diseases for breast biopsy outcome predictions, Equation (8) was used to generate prediction results for new patient mammographic data. If the equation $f(\boldsymbol{x}) > 0$, the results belonged to the category of benign disease. If the equation $f(\boldsymbol{x}) < 0$, the results belonged to the category of malignant breast cancer.

To determine the parameter *b* in Equation (8), those support vectors, with $0 < \alpha_i < C$, have $\xi_i = 0$. Thus, Equation (9) becomes $t_i f(\boldsymbol{x}_i) = 1$ that will satisfy

$$t_i\left[\sum_{j\in S}\alpha_j t_j K(x_i, x_j) + b\right] = 1. \tag{10}$$

Therefore, by averaging, a numerically stable solution for the parameter *b* is given

$$b = \frac{1}{NM}\sum_{i\in M}\left[t_i - \sum_{j\in S}\alpha_j t_j K(x_i, x_j)\right], \tag{11}$$

where *M* denotes the indices of a subset of the data points in which $0 < \alpha_i < C$.

The Kernel Functions Used in SVM Models

A SVM learning classification model can be considered as either a linear classifier or a nonlinear classifier, depending on the application of the type of kernel functions. However, in many cases, data are nonlinearly separable. Thus, nonlinear kernel functions are often used for the SVM learning classification model. The most widely used nonlinear kernel function is the Gaussian RBF kernel given by

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2\right), \tag{12}$$

where the bandwidth parameter σ creates a large and flexible class model. By turning the bandwidth parameter σ well, the Gaussian RBF kernel is able to capture the underlying functions behind a wide variety of training data sets.

The Methods of SVM Model Evaluation

In order to evaluate the performances of the SVM learning classification model, one of the best methods is to analyze the model's accuracy, sensitivity, specificity, precision, and *F*-score as well as its ROC curve analysis. In this research, these analyses depended on the number of false positive and false negative instances of the mammographic mass data according to the reference [6], [21]. Table 1 shows the diagnostic results in terms of positive or negative for distinguishing malignant breast cancer and benign disease by using the developed SVM learning classification model.

The *sensitivity* is defined as the probability of correctly identifying malignant breast cancers given by [21],

TABLE 1
THE DEVELOPED SVM LEARNING CLASSIFICATION MODEL'S DIAGNOSTIC RESULTS FOR DISTINGUISHING MALIGNANT BREAST CANCER AND BENIGN DISEASE

	Actual Malignant	Actual Benign	Total Number
Predicted Malignant	True Positive (TP)	False Positive (FP)	TP + FP
Predicted Benign	False Negative (FN)	True Negative (TN)	FN + TN
Total Number	TP + FN	FP + TN	TP + FP + FN + TN

$$Sensitivity = \frac{TP}{TP+FN}. \quad (13)$$

The sensitivity is also referred to as the *true positive rate*, *recall* or *capture rate* in the area of machine learning.

The *specificity* is defined as the probability of correctly identifying benign diseases given by,

$$Specificity = \frac{TN}{FP+TN}. \quad (14)$$

The specificity is sometimes called the *true negative rate*. The difference of (1 – specificity) is known as the *false positive rate*.

The *precision* or the *positive predictive value* is defined as

$$Precision = \frac{TP}{TP+FP}. \quad (15)$$

Notice that the recall in Equation (13) is a measure of quantity while the precision in Equation (15) is a measure of quality. Both the precision and recall are in a mutual relationship based on the understanding and measure of relevance.

Thus, for the probability of the misclassification error (PME), it is obtained by

$$PME = \frac{FN+FP}{TP+FN+FP+TN}, \quad (16)$$

and for the model's accuracy, it is defined by

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN}, \quad (17)$$

where the model's accuracy = (1 – *PME*).

Additionally, based on the harmonic mean of precision and recall, the *F*-score is defined as

$$F\_score = 2\left(\frac{Precision \times Recall}{Precision+Recall}\right), \quad (18)$$

where the *F*-score can be used as a single measure of the performance of the test or a single measure of a model's accuracy of the test. The *F*-score can also be interpreted as a weighted average of the precision and recall. A *F*-score equal to 1 would signify the best score of a accuracy. A *F*-score of 0 would be the worst score.

The SVM-based ROC Curve Analysis

A ROC curve analysis of the developed SVM learning classification model was based on a graph plot, which was generated by changing a set of trade-off points between the sensitivity and the difference of (1 – specificity) for cases classified as malignant breast cancer. A corresponding estimated area under the ROC curve analysis was considered as an effective measure of inherent validity of a diagnostic test [37] and an evaluation metric for the performance of classification and prediction models [20]. The estimated area under the ROC curve analysis of the developed SVM learning classification model was determined by using a trapezoidal approximation formula [21], [38]:

$$\int_0^1 S(x)dx \cong \sum_{i=0}^{N}\left(\frac{y_i+y_{i+1}}{2}\right)(x_{i+1}-x_i), \quad (19)$$

where *S(x)* denoted the function of the ROC curve analysis, y_i and x_i represented the sensitivity and (1-specificity) at the *i*th ($i = 0, 1, 2, \ldots, M$) point, respectively. Additionally, a standard error (SE) of the area of the ROC curve analysis for the developed SVM learning classification model was obtained by [39]

$$SE = \sqrt{\frac{S(1-S)+(N_1-1)(Q_1-S^2)+(N_2-1)(2-S^2)}{N_1N_2}}, \quad (20)$$

where *S* would be an estimated area of the ROC curve analysis for the SVM learning classification model, ranging from 0 to 1; N_1 and N_2 denoted the number of clinical instances of malignant (positive) and benign (negative) cases in the mammographic mass dataset, respectively; $Q_1 = S/(2-S)$ and $Q_2 = 2S^2/(1+S)$. Assuming that the future breast cancer clinical instances are drawn from the same distribution, the estimated area of the ROC curve analysis and its standard error shows how well and accurately the developed SVM learning classification model will perform in diagnosing new clinical instances of mammographic mass data within a predictive confidence interval.

The estimated area of the ROC curve analysis is statistically interpreted as the probability of the classification model to correctly classify malignant breast cancer and benign disease. Thus, the estimated area of the ROC curve analysis can be used to evaluate and rank the quality of the developed SVM learning classification models. When the estimated area of the ROC curve analysis is equal to 1, the learning classification model is a perfect modeling in terms of diagnostic accuracy in distinguishing malignant breast cancer from benign disease. Therefore, the higher the estimated area of the ROC curve analysis is, the better the learning classification model performs [21]. As a result, this subsequently leads to the least probability of misclassification error of distinguishing malignant and benign diagnoses for breast cancer biopsy outcome predictions.

III. RESULTS

In this research, a total of 830 clinical instances of the mammographic mass dataset was used: 427 (51.45%) benign diseases and 403 (48.55%) malignant breast cancers. The developed SVM learning classification model was trained and tested using all of the available clinical instances of the mammographic mass dataset.

The SVM learning classification (or training) model is shown in Figure 1 part (a), which includes a scale function, a *k*-fold cross validation, a subset of data selection, a SVM learning classification, and a training error. All of the input attributes for the clinical instances were scaled to zero mean and unit variance by using the scale function before these

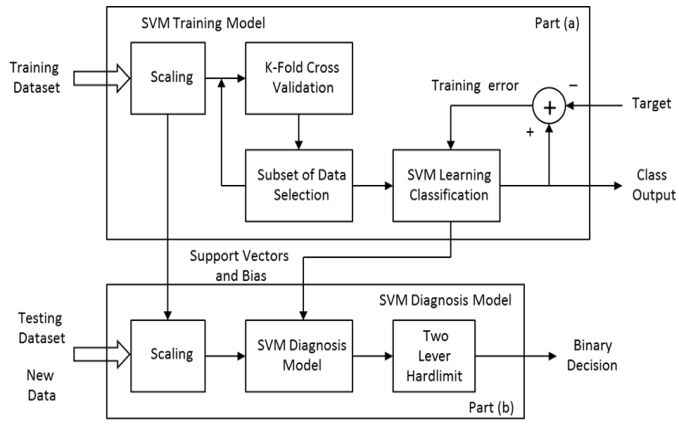

Fig. 1: The developed SVM learning classification contained two separate models: a learning classification (or training) model in part (a) and a diagnostic (or prediction) model in part (b). The learning classification model was a SVM nonlinear classifier based on a Gaussian RBF kernel. The 10-fold cross-validation was used to adjust hyper-parameters for the developed SVM learning classification models during the training periods. The training results, such as support vectors and bias, were fed into the SVM diagnosis model. The diagnostic model distinguished and classified malignant breast cancers and benign diseases for breast cancer biopsy outcome predictions when new patient data were used.

clinical instances were fed into the developed SVM learning classification.

To evaluate the performances of the developed SVM learning classification model, the probability of accuracy and the area of the ROC curve analysis were estimated using a nonparametric approach based on a rotation method [40], also referred to as a *k*-fold cross validation [41]. The *k*-fold cross validation had less bias toward the model training and test results for calculating the probability of accuracy and the area of the ROC curve analysis.

A. The SVM Learning Classification Model Results

Training the SVM learning classification (or training) model depended on the selection of the regularization parameter *C*, which was used to control the trade-off in Equation (1), and the kernel parameters. In this research, a Gaussian RBF kernel as shown in Equation (12) was used for the developed SVM learning classification model. The parameter σ, a bandwidth parameter for the Gaussian RBF kernel, was the only kernel parameter to be determined for the developed SVM learning classification model.

The 10-fold cross-validation, a method for adjusting hyper-parameters (such as the regularization parameter *C* and the bandwidth parameter σ) for the developed SVM learning classification models during the training periods, was also used. The 830 clinical instances of the mammographic mass dataset were first partitioned into 10 subsets of the data that were equally sized. Each data point from the 830 clinical instances was randomly assigned to one of the subsets of data. Then an individual SVM learning classification model was trained by applying SVM algorithms to 9 of the subsets of data (training data). This model was then evaluated using the one remaining subset of data (testing data). Furthermore, a cross-validation error was computed by using an average of the 10 outcomes of the developed SVM learning classification model evaluations, which were used to predict the performance of the developed SVM learning classification model algorithms when applied to the entire set of clinical instances.

In order to choose the regularization parameter *C* and the Gaussian RBF kernel bandwidth parameter σ using the 10-fold cross-validation, the cross-validation error was computed for the developed SVM learning classification models based on different values for the parameters *C* and σ. The regularization parameters *C* and the bandwidth parameter σ were finally determined, based on the lowest cross-validation error, and were finally used to train the developed SVM learning classification model on the entire 830 clinical instances dataset.

For the training results, the final parameter values for the developed SVM learning classification model were obtained: the regularization parameter $C = 630$, the bandwidth parameter for the Gaussian RBF kernel $\sigma = 39$, the model intercept value $b = -0.01427455$, and the number of support vectors was 541. The training error rate for the developed SVM learning classification model was 6.0241% when setting the tolerance of termination criterion to 0.0005.

In dealing with multidimensional features, visualization and understanding are often aided by representing the observations in a lower-dimensional space. In particular, two-dimensional scatter plots based on principal component analysis [40] are helpful in exploring relationships between the malignant breast cancers and benign diseases groups, in assessing the group-conditional distributions, and in identifying a typical feature observation. Thus, Figure 2 shows a scatter plot of the 830 clinical instances of the mammographic mass dataset, with the first principal component as the *x*-axis and the second principal component as the *y*-axis. In this figure, a red symbol of the "+" represented malignant breast cancer and a blue

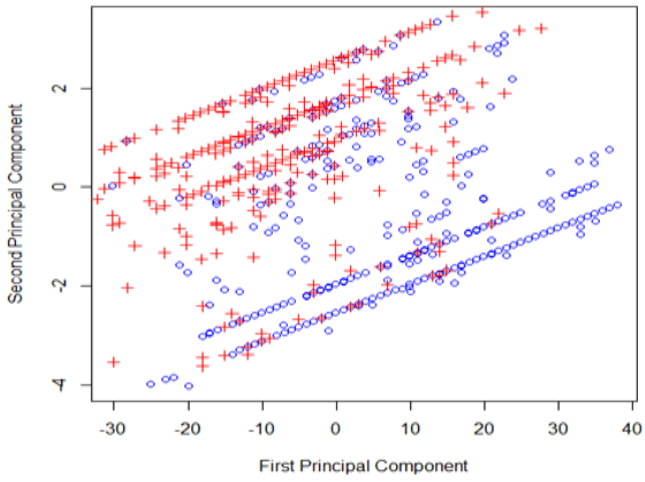

Fig. 2: A scatter plot of the 830 clinical instances of the mammographic mass dataset based on the first principal component as the *x*-axis and the second principal component as the *y*-axis, where a "+" (red) represented malignant breast cancer and a "o" (blue) indicated benign disease. As shown, there was a large overlaid area of intersection between the malignant breast cancers and benign diseases, clearly illustrating that the clinical instances of the mammographic mass dataset were not separable.

symbol of the "o" indicated benign disease. As shown, there was a large overlaid area of intersection between the malignant breast cancers and benign diseases, clearly illustrating that the clinical instances of the mammographic mass dataset were not separable. This observation thereby led to the use of a nonlinear Gaussian RBF kernel in this developed SVM learning classification model.

B. The SVM Diagnostic Model Results

In this section, we present the testing results of the kernel-based SVM diagnostic (or prediction) model, as shown in part (b) of Figure 1, to estimate its probability of accuracy on distinguishing malignant breast cancer and benign disease for breast biopsy outcome predictions.

After completing the training of the developed SVM learning classification model, the final support vectors and hyper-parameters obtained from the learning classification model were loaded into the kernel-based SVM diagnostic model. Using the same 830 clinical instances of the mammographic mass dataset, the kernel-based SVM diagnostic model was tested to estimate its probability of accuracy in distinguishing malignant breast cancer and benign disease for breast biopsy outcome predictions.

For the kernel-based SVM diagnostic model, the testing accuracy in diagnosing and classifying malignant breast cancer and benign disease was 93.98%, with details shown in Table 2. Accordingly, using Equations (13), (14), (15) and (18), the sensitivity (recall), specificity, precision, and *F*-score results were 94.54%, 93.44%, 93.15%, and 0.94, respectively.

TABLE 2
THE TEST RESULT OF THE KERNEL-BASED SVM DIAGNOSTIC MODEL IN DIAGNOSING MALIGNANT BREAST CANCER AND BENIGN DISEASE

	Actual Malignant	Actual Benign	Total
Predicted Malignant	381	28	409
Predicted Benign	22	399	421
Total	403	427	830
Probability of Misclassification Error	5.45%	6.56%	6.02%

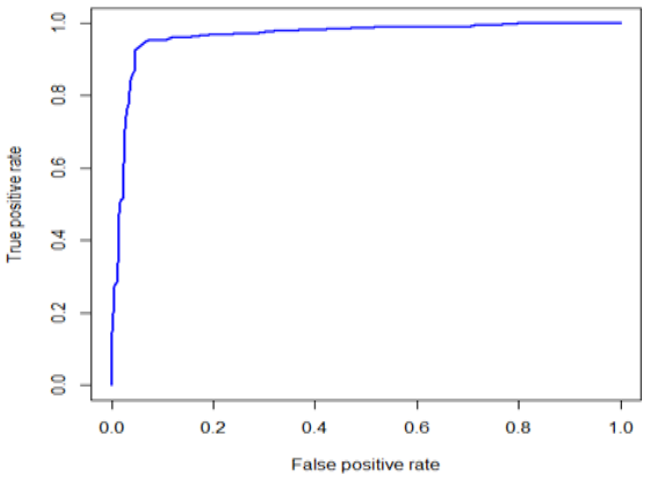

Fig. 3: An estimated area of the ROC curve analysis of the developed SVM learning classification model for cases classified as malignant breast cancer and benign disease, where the true positive rate is sensitivity on the *y*-axis and the false positive rate is the difference (1 – specificity) on the *x*-axis. As shown, the estimated area under the ROC curve analysis was 0.9630.

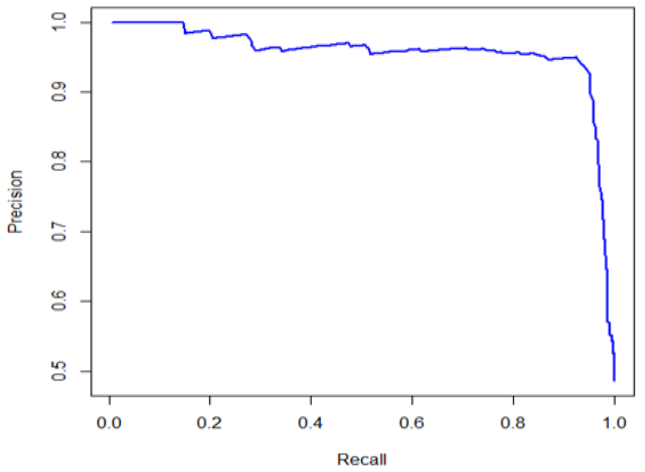

Fig. 4: A curve plot of mutual relationship between the precision and recall, where the precision is on the *y*-axis and the recall is on the *x*-axis. As shown, this curve can be used to determine an optimal cutoff point for the recall and precision, which is the curve's upper-right corner with minimized distance to the point of (1,1).

C. The Area Under the ROC Curve Analysis Results

The ROC curve analysis results of the developed SVM learning classification model were produced by varying a set of trade-off points between the model sensitivity on the y-axis and the difference value (1 – specificity) on the x-axis as shown in Figure 3. The estimated area under the ROC curve analysis was 0.9630. Correspondingly, the associated SE of the area under the ROC curve analysis obtained by using Equation (20) was 0.0516. Thus, the estimated area of the ROC curve analysis results implied that the proposed SVM learning classification model can provide a consistently high accuracy in diagnosing and classifying malignant breast cancer and benign disease for breast biopsy outcome predictions.

In addition, Figure 4 shows a curve plot of the relationship between the precision and recall. The precision could be considered as a measure of exactness (or quality) and the recall as a measure of completeness (or quantity). Both the precision and recall in the curve plot were in a mutual relationship based on the understanding and measure of relevance. Thus, this curve plot allowed us to find an optimal solution for the model accuracy by determining a set of trade-off points based on the precision on the y-axis and the recall on the x-axis.

IV. DISCUSSION

The developed SVM learning classification model was utilized to diagnose and classify malignant breast cancer and benign disease for breast biopsy outcome predictions. It was trained to determine the model hyper-parameters using the

10-fold cross-validation and tested using the same mammographic mass dataset. The test accuracy of the developed SVM learning classification model was 93.98%. Furthermore, the model sensitivity (or recall) was 94.54%, the model specificity was 93.44%, the model precision was 93.15%, and the model *F*-score was 0.94. The estimated area under the ROC curve analysis for the developed SVM learning classification model was 0.9630, and its corresponding SE was 0.0516. Thus, based on these results, the diagnostic accuracy of the developed SVM learning classification model would be 93.98% accurate in distinguishing between benign disease and malignant breast cancer for a new patient with mammographic mass data. Additionally, due to the high estimated area (0.9630±0.0516) of the ROC curve analysis along with the high recall and precision probabilities, the developed SVM learning classification model was able to achieve a consistently high accuracy in diagnosing malignant breast cancer and benign disease for breast biopsy outcome predictions.

In comparison to related papers, there were several different methods developed using the same mammographic mass dataset including: an artificial neural network (ANN) classifier [18], [19], a case-based reasoning classifier (CBRC) [7], [12], [14]-[16], a distributed genetic programming approach-based prediction model (DGPA) [20], a decision tree approach (DTA) [19], and a neural network classification model (NNCM) [21]. Table 3 shows the testing results of the performances of the previously reported methods and the developed SVM learning classification model (SVMLCM) based on analysis of the estimated areas under the ROC curves.

In this assessment, as seen in Table 3, the estimated area under the ROC curve analysis of the developed SVM learning classification model is comparably much higher than most of those of the previously published methods. Moreover, the developed SVM learning classification model used the Gaussian RBF kernel, which is a nonlinear kernel function with a large selection for the bandwidth parameter σ. Thus, the developed SVM learning classification model had more flexibility, regardless of whether or not there were overlapping data (or clusters) between the malignant breast cancer and benign disease cases; moreover, the SVM model provided a more reliable and greater percentage of accuracy in distinguishing malignant breast cancer and benign disease for breast biopsy outcome predictions. Therefore, the proposed SVM learning classification model with mammography has significant potential in reducing the number of unnecessary, expensive, and invasive surgical breast biopsies in clinical practices.

TABLE 3
COMPARISON OF THE PERFORMANCES OF THE TEST RESULTS FROM PREVIOUSLY REPORTED METHODS AND THE DEVELOPED SVM LEARNING CLASSIFICATION MODEL (SVMLCM) BASED ON THE ESTIMATED AREAS OF THE ROC CURVES ANALYSIS

Methods	Area and Standard Error Under ROC Curve
SVMLCM	0.9630±0.0516
ANN	0.847±0.017 ~ 0.880±0.01
CBRC	0.857±0.016 ~ 0.890±0.01
DTA	0.838±0.017 ~ 0.870±0.01
DGPA	0.859±0.032 ~ 0.860±0.03
NNCM	0.9626±0.0069

V. CONCLUSION AND FUTURE WORK

In this paper, we developed a SVM learning classification model and evaluated its recall, precision, specificity, *F*-score, and ROC curve analysis to diagnose and classify malignant breast cancer and benign disease for breast biopsy outcome predictions. The developed SVM learning classification model was trained using the 10-fold cross-validation technique and tested using all of the available 830 clinical instances of the mammographic mass dataset. The developed SVM learning classification model, based on the Gaussian RBF kernel, was a nonlinear classifier that had high flexibility in adjusting the model hyper-parameters for any non-separable mammographic mass data.

The testing results showed that the developed SVM learning classification model had a sensitivity (or recall) of 94.54% in diagnosing malignant breast cancer, specificity of 93.44% in diagnosing benign disease, precision of 93.15%, model *F*-score of 0.94, and overall accuracy of 93.98% in diagnosing both malignant breast cancer and benign disease. An estimated area of the ROC curve analysis and its associated SE for the proposed SVM learning classification model was 0.9630±0.0516. Therefore, the developed SVM learning classification model along with mammography can provide highly accurate and consistent diagnoses for breast biopsy outcome predictions, allowing future patients to bypass unnecessary surgical biopsies.

In future research, we would propose a set of enhanced classification and prediction models, including random forest approaches along with other types of classification models to form an ensemble learning classification and prediction (ELCP) model, which combines prediction results from different individual models using a weighting function. The ELCP model typically has better prediction results than those of the individual models, especially in dealing with non-linear separable clusters in data sets. Thus, the ELCP model is capable of further enhancing the diagnosis accuracy of breast cancer biopsy predictions.

ACKNOWLEDGMENT

The authors gratefully acknowledge that the mammographic mass dataset of clinical breast cancer cases was obtained from the Mammographic Mass Database available in the UCI Machine Learning Repository. This dataset, which contains mammographic information of breast cancer clinical instances, was contributed by Dr. Rüdiger Schulz-Wendtland from the Institute of Radiology, Gynaecological Radiology, University Erlangen-Nuremberg in Germany.

REFERENCES

[1] Department of Health and Human Services Centers for Disease Control and Prevention, World Cancer Day, February 3, 2015, Available http://www.cdc.gov/cancer/dcpc/resources/features/worldcancerday/.

[2] Department of Health and Human Services Centers for Disease Control and Prevention, United States Cancer Statistics, Technical Notes 2007, Available http://www.cdc.gov/cancer/npcr/uscs/2007/technical_notes/.

[3] American Cancer Society, *Cancer Facts & Figures 2012*, Atlanta, Georgia, American Cancer Society, pp. 1–63, 2012.

[4] American Cancer Society, *Breast Cancer Facts and Figures 2011-2012*, Atlanta, Georgia, American Cancer Society, pp. 1-32, 2011.

[5] National Cancer Institute, *Cancer Trend Progress Report – 2011/2012 Update*, U.S. Depart. of Health & Human Services, National Institutes of Health, Available http://progressreport.cancer.gov/introduction.asp.

[6] G. J. Miao, K. H. Miao, and J. H. Miao, "Neural pattern recognition model for breast cancer diagnosis," *Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Bioinformatics (JBIO)*, August Edition, pp. 1–8, September 2012.

[7] S. W. Fletcher, W. Black, R. Harris, B. K. Rimer, and S. Shapiro, "Report of the international workshop on screening for breast cancer," *Journal of the National Cancer Institute*, vol. 85, pp. 1644–1656, 1993.

[8] National Cancer Institute at the National Institute of Health, *Mammograms*, U.S. Depart. of Health & Human Services, Available http://www.cancer.gov/cancertopics/factsheet/detection/mammograms.

[9] M. Elter, R. Schulz-Wendtland, and T. Wittenberg, "The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process," *Medical Physics*, vol. 34, no. 11, pp. 4164–4172, 2007.

[10] M. S. Hung, M. Shanker, and M. Y. Hu, "Estimating breast cancer risks using neural netwroks," *Journal of Operational Research Society*, vol. 52, pp. 1–10, 2001.

[11] D. B. Kopans, "The positive predictive value of mammography," *American Journal of Roentgenology*, vol. 158, pp. 521–526, 1992.

[12] C. E. Floyd, J. Y. Lo, and G. D. Tourassi, "Case-based reasoning computer algorithm that uses mammographic findings for breast biopsy decisions," *American Journal of Roentgenology*, vol. 175, pp. 1347–1352, November 2000.

[13] American College of Radiology, BI-RADS Atlas, Available http://www.acr.org/Quality-Safety/Resources/BIRADS

[14] A. O. Bilska-Wolak and C. E. Floyd, "Investigating different similarity measures for a case-based reasoning classifier to predict breast cancer", *Proceedings of SPIE*, vol. 4322, pp. 1862–1866, 2001.

[15] A. O. Bilska-Wolak and C. E. Floyd, "Development and evaluation of a case-based reasoning classifier for prediction of breast biopsy outcome with BI-RADS lexicon", *Medical Physics*, vol. 29, pp. 2090–2100, 2002.

[16] A. O. Bilska-Wolak, C. E. Floyd, J. Y. Lo, and J. A. Baker, "Computer aid for decision to biopsy breast masses on mammography: validation on new cases," *Academic Radiology*, vol. 12, pp. 671–680, 2005.

[17] J. A. Baker, P. J. Kornguth, J. Y. Lo, M. E. Williford, and C. E. Floyd, "Breast cancer: Prediction with artificial neural network based on BI-RADS standardized lexicon," *Radiology*, vol. 196, pp. 817–822, 1995.

[18] M. K. Markey, J. Y. Lo, R. Vargas-Voracek, G. D. Tourassi, and C. E. Floyd "Perception error surface analysis: A case study in breast cancer diagnosis," *Computers in Biology and Medicine*, vol. 32, pp. 99–109, 2002.

[19] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[20] S. A. Ludwig, "Prediction of breast cancer biopsy outcomes using a distributed genetic programming Approach," *Proceedings of the 1st ACM International Health Informatics Symposium*, ACM, pp. 694–699, New York, 2010.

[21] K. H. Miao and G. J. Miao, "Mammographic diagnosis for breast cancer biopsy predictions using neural network classification model and receiver operating characteristic (ROC) curve evaluation," *Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Bioinformatics (JBIO)*, September Edition, Vol. 3, Issue 9, pp. 1–10, October 2013.

[22] D. Conforti, D. Costanzo, and R. Guido, "Cancer prognostic evaluation via support vector machines," *International Scientific Journal of Computing*, Vol. 3, Issue 3, pp. 29–34, 2004.

[23] A. Bharathi and K. Anandakumar, "Cancer classification using relevance vector machine learning approach," *Journal of Medical Imaging and Health Informatics*, Vol. 5, No. 3, pp. 630–634, June 2015.

[24] UCI Machine Learning Repository, Mammographic Mass Data Set, Available http://archive.ics.uci.edu/ml/datasets/Mammographic+Mass.

[25] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[26] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transaction on Neural Network*, Vol. 10, pp. 988–999, 1999.

[27] A. Karatzoglou, Alex Smola, and Kurt Hornik, "Kernel-based machine learning lab," *R Repository: CRAN Package Kernlab*, Version 0.9–19, pp. 1–108, November, 2013.

[28] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, and M. Angelo, et al., "Prediction of central nervous system embryonaltumour outcome based on gene expression," *Nature*, Vol. 415, No. 24, pp. 436–442, 2002.

[29] A. Bhattacharjee, W. G Richards, J. Staunton, C. Li, and S. Monti, et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proc. Natl. Acad. Sci*. USA, Vol. 98, pp. 13790–13795, 2001.

[30] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, and M. Gaasenbeek, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, Vol. 286, No. 5439, pp. 531–537, 1999.

[31] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, and J. L. Kutok, et al., "Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning," *Nat Med*, Vol. 8, No. 1, pp. 68–74, 2002.

[32] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, and C. Sugnet, et al., "Knowledge-based analysis of microarray gene expression data using support vector machines," *Proc. Natl. Acad. Sci.* USA, Vol. 97, No. 1, pp. 262–267, 2000.

[33] S. Mukherjee, "Classifying microarray data using support vector machines," *Whitehead Institute for Genome Research and Center for Biological and Computational Learning at MIT*, Chapter 9, pp. 1–20, August 2002.

[34] C. W. Hsu, and C. J. Lim, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, Vol. 13, No. 2, pp. 415–425, March 2002.

[35] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science & Business Media, LLC, 2006.

[36] B. Clarke, E. Fokoue, and H. H. Zhang, *Principles and Theory for Data Mining and Machine Learning*, Springer Science & Business Media, LLC, 2009.

[37] R. Kumar and A. Indrayan, "Receiver operating characteristic (ROC) curve for medical researchers," *Indian Pediatrics*, Vol. 48, pp. 277–287, April 2011.

[38] R. L. Finney, F. D. Demana, B. K. Waits, and D. Kennedy, *Calculus: A Complete Course*, Second Edition, Addison Wesley Longman, Inc., 2000.

[39] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, Vol. 143, No. 1, pp. 29–36, April 1982.

[40] G. J. Miao and M. A. Clements, *Digital Signal Processing and Statistical Classification*, Artech House, Inc., 2002.

[41] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer Science & Business Media, LLC, 2009.

Julia H. Miao is an undergraduate student at Cornell University, Ithaca, New York.

She is a National AP Scholar, a National Elks Foundation Scholar, and a National Siemens Competition Semifinalist. She is the co-author of the peer-reviewed journal paper "*Neural Pattern Recognition Model for Breast Cancer Diagnosis*." She has received a number of academic, science, and technology awards, including the AP Scholar with Distinction Award, the Synopsys Silicon Valley Science & Technology Championship Second-Place and Third-Place Awards, IEEE Award for Best Electro-Technology, National Mathematics Honor Society Mu Alpha Theta Winner Award, and Society of Women Engineers Santa Clara Valley Section Winner Award. She has also received the USA National Presidential Gold Level Volunteer Service Award.

Julia received national, state, and local scholarships from the Elks National Foundation, including Most Valuable Student Achievement Scholarships from the National Lodge, California-Hawaii Elks Association Level Lodge, and Sunnyvale Lodge, respectively. She is the Co-Founder of The National Wishing Star Organization for the American Cancer Society, dedicated to

raising awareness of cancer and other illnesses in the global community. Her current research interests include biological sciences, cancer research, and medicine.

Kathleen H. Miao is an undergraduate student at Cornell University, named to Dean's List of the College of Arts and Sciences, and received a Cornell Tradition Fellowship from Cornell University, Ithaca, New York.

She is a National AP Scholar and a USA Biology Olympiad Semifinalist. She is the co-author of several peer-reviewed international journal paper publications and is an undergraduate student researcher at Cornell University. Presently, she is an Associate Editor of Biological and Biomedical Sciences for the international research journal *Journal of Young Investigators* and a Managing Editor for *The Research Paper* at Cornell University.

She received a number of academic, science, and technology awards including the AP Scholar with Distinction Award, the Elks National Foundation Most Valuable Student Achievement Scholarship awards, the Intel Science Talent Search Research Report Badge Award, Synopsys Silicon Valley Science & Technology Championship Honorable Mention (Third place Award), IEEE Award for Best Electro-Technology, National Mathematics Honor Society Mu Alpha Theta Winner Award, and Society of Women Engineers Santa Clara Valley Section Winner Award. She also received Stanford University Medical Center Auxiliary Volunteer Honors for valuable contributions in community service.

Kathleen is the Co-Founder of The National Wishing Star Organization for the American Cancer Society and a Vice President of Community Service for the National Society of Collegiate Scholars at Cornell University. Her current research interests include cancer research and the medical sciences as well as multidisciplinary topics involving computational modeling of biological processes and medical diagnoses of cancer.

George J. Miao received a B.Eng. joint degree from Shanghai University of Science and Technology (now Shanghai University) and Shanghai Second Medical University (now Shanghai Jiao Tong University School of Medicine), China; a M.S. in Statistics from Columbia University, New York, New York; and a Ph.D. in Electrical Engineering from the Georgia Institute of Technology, Atlanta, Georgia.

He is a Vice President and a Chief Scientist at Flezi, LLC. He worked and consulted for a number of U.S. Fortune 500 companies, universities, research institutes, investment and asset management firms. He is the co-author of the textbook *Digital Signal Processing and Statistical Classification* (Artech House, 2002) and the author of the textbook *Signal Processing in Digital Communications* (Artech House, 2007). He holds 16 granted U.S. patents in the area of digital signal processing.

Dr. Miao is a Senior Member of the IEEE and was a Chairman of the IEEE New Jersey Coast Chapter of Signal Processing (2003-2006). He has been a manager of the California Hedge Fund Association since 2012. He received a number of awards, including the IEEE Region-1 Technology Award, the IEEE Section Technical Achievement Award, the IEEE Chapter Distinguished Service Award, and the IEEE Signal Processing Society Certificate of Appreciation. His current research interests are in statistical learning theory, data mining and machine learning of sophisticated algorithm-based quantitative models, classification and prediction modeling, in conjunction with advanced and dynamic digital signal processing for trading strategies, trading signal detection, equity long and short strategies, equity pricing prediction and volatility forecasting, statistical and volatility arbitrage on alpha generation, cross-asset correlation trading as well as quantitative portfolio optimization and management in big data analytics.