Protein crystallization image classification with Elastic Net

Jeffrey Hung[a], John Collins[a], Mehari Weldetsion[a], Oliver Newland[a], Eric Chiang[a], Steve Guerrero[b], Kazunori Okada[a]

[a]Computer Science Dept., San Francisco State Univ., 1600 Holloway Ave., San Francisco, CA, USA 94132-4163; [b]Genentech, Inc., 1 DNA Way, South San Francisco, CA, USA 94080

ABSTRACT

Protein crystallization plays a crucial role in pharmaceutical research by supporting the investigation of a protein's molecular structure through X-ray diffraction of its crystal. Due to the rare occurrence of crystals, images must be manually inspected, a laborious process. We develop a solution incorporating a regularized, logistic regression model for automatically evaluating these images. Standard image features, such as shape context, Gabor filters and Fourier transforms, are first extracted to represent the heterogeneous appearance of our images. Then the proposed solution utilizes Elastic Net to select relevant features. Its L^1-regularization mitigates the effects of our large dataset, and its L^2-regularization ensures proper operation when the feature number exceeds the sample number. A two-tier cascade classifier based on naïve Bayes and random forest algorithms categorized the images. In order to validate the proposed method, we experimentally compare it with naïve Bayes, linear discriminant analysis, random forest, and their two-tier cascade classifiers, by 10-fold cross validation. Our experimental results demonstrate a 3-category accuracy of 74%, outperforming other models. In addition, Elastic Net better reduces the false negatives responsible for a high, domain specific risk. To the best of our knowledge, this is the first attempt to apply Elastic Net to classifying protein crystallization images. Performance measured on a large pharmaceutical dataset also fared well in comparison with those presented in the previous studies, while the reduction of the high-risk false negatives is promising.

Keywords: elastic net, image classification, high throughput protein crystallization, random forest, logistic regression, cascade classifiers, feature selection, machine learning

1 INTRODUCTION

Structural analysis of protein crystals plays a crucial role in pharmaceutical research.[12] The development of a new medication that mitigates a symptom while avoiding undesired side-effects often requires binding to specific sites on a protein molecule. Such binding site locations are determined through mapping its molecular structure by diffracting X-rays through a protein crystal. Crystals can be formed by supersaturating a solution with the protein molecules. However, protein crystallization is a difficult process that generates extremely low yields of uncontaminated, well-ordered crystals. Therefore many trials must be run in a high-throughput setting during process optimization, forcing crystallographers to manually examine a great many images looking for rare occurrences of crystallization.[12] This tedious inspection task benefits from reliable automation.

A number of researchers have worked on systems to automatically classify the photographic results of protein crystallization experiments. Wilson's work in 2002[25] used features such as consistent line width frequencies, ratio of minimum box to detected object area, intensity variation of a detected object area, and curvature measures of that detected object area, which were found effective in later studies. Bern's work[4] segmented the solution drop first, and used boundary and edge detection to extract curvature and gradient-based features for crystal detection. Yang et al.[26] designed a cascade classifier based on a hand-constructed decision tree separating "clear" from "non-clear", followed by a Fisher linear discriminant classifier (LDA) to distinguish "precipitate" and "crystal". Kawabata et al.[17] developed a classifier based on LDA for separating between crystal and non-crystal classes, which they cascaded with another LDA classifier to distinguish between three crystal sub-classes. Most relevant to our study are the works by Cumbaa and Jurisica[9],[10] in which they applied to large dataset a wide variety of features which included: energy (intensity change over area), Euler numbers (number of objects minus number of holes), Radon-Laplacian, gray-level co-occurrence matrix (GLCM), edge features (Laplacian and Sobel filter based features).

Experimental results from these prior studies are summarized in Table 1. In many studies, researchers relied on overall accuracy as their performance metric, which is indiscriminate of class label and of distinction between type I and type II errors. In the application context of high-throughput protein structure assays, missing potential crystals (false negatives) comes with much higher operational risk than falsely identifying crystals (false positives). This domain specific bias towards reducing false negatives is common in many applications, such as cyber-security[24] and medical informatics[18], and amounts to prioritizing recall measure over precision and accuracy. The best classification model can be designed to achieve the highest overall recall measure with reasonable precision. In addition, only Cumbaa and Jurisica in their 2005 and 2010 studies appear to have large enough datasets to account for the amount of image variance and skew between crystal images and non-crystal images. For example, Kawabata et al.[17] reported very high overall accuracy for crystal identification, however they used a small data set with an unrealistic class distribution between crystal and non-crystal cases.

This study developed an automated process for classifying protein crystallization images by extracting image features, selecting a subset of important features and applying this subset to machine learning-based image classifiers.[7],[19] We designed our study to employ a large dataset that includes only a few crystal cases, corresponding to the low crystal yields of real protein assay experiments, so that our experimental results can reflect the true performance of the proposed system in realistic application scenarios. As one of our main contributions, this study introduces Elastic Net (EN)[28] for both feature selection and as a model of protein crystal image classification. To the best of our knowledge, this is the first attempt to apply EN to this application problem. The proposed EN classification model is experimentally evaluated in comparison with a number of leading supervised machine learning classifiers, including two-tier cascades of them. This study also investigates the effect of spatial variance in feature space design by comparing the system's performance with a few hand-selected spatially-invariant features against the performance with a feature subset automatically selected by a feature selection algorithm from a complete set that includes Gabor filter-based spatially-variant features. The fact that the latter models with a spatially-variant feature set outperformed the former models with the invariant set suggests that there may be a tendency for crystals to form in certain locations within a well. We systematically evaluate the system's performance through overall accuracy and recall measures in both 3-category and 2-category classification scenarios. Careful analysis of the results yields our recommended solution of a regularized logistic regression model that utilizes the Elastic Net to select the features and a two-tier cascade classifier based on naïve Bayes[21] followed by the random forest[5] algorithm, resulting in performance competitive to those reported by the previous studies.

The rest of this paper is organized as follows. In section 2, we describe the images in the dataset. In section 3, we describe the types of features we used for the feature space. In section 4, we describe the general flow for processing the data and describe the feature selectors and classifiers used in the study. In section 5, we list the specific settings and results for each experiment. Finally in section 6, we propose some explanations and some suggestions for future work.

Table 1. Summary of prior works on automatic protein crystal classification

	Number	Number	Number		Recall	
	of	of Crystal	of			
Study (Classifier Used)	Images	Images	Features	Accuracy	Precipitate	Crystal
Wilson '02 (naive Bayes) [25]	~ 600	~ 200	13	N/A	N/A	86.0%
Bern et al. '04 (hand-built DT) [4]	1057	196	7	77.2%	51.2%	81.6%
Bern et al. '04 (C5.0 + AdaBoost) [4]	706	286	7	75.7%	68.5%	78.7%
Cumbaa and Jurisica '05 (LDA) [9]	190572	5600	59	83.0%	N/A	68.0%
Cumbaa and Jurisica '05 (kNN) [9]	190572	5600	59	85.0%	N/A	76.0%
Yang et al. '06 (hand-built DT + LDA) [26]	110	39	16	80.9%	58.3%	84.8%
Kawabata et al. '08 (LDA-based DT) [17]	439	229	14	92.7%	90.1%	93.5%
Cumbaa and Jurisica '10 (RF, crystal vs. clear) [10]	13830	1879	14908	61.6%	N/A	80.2%
Cumbaa and Jurisica '10 (RF, precipitate vs. clear) [10]	9656	2897	14908	86.4%	88.7%	N/A

2 DATA DESCRIPTION

Data used in this study consisted of 11648 24-bit color images that were 1024x1024 pixels in size. The images were converted to grayscale prior to extracting features. Example images in Figure 1 show large variations in illumination, as well as artifacts such as well-joints and poor focus. The visual appearance of both crystals (Figure 1c) and precipitates (Figure 1b) is also highly heterogeneous.

Expert crystallographers assigned labels to the images corresponding to one of three classes: clear, precipitate or crystal. The distribution amongst the classes was 3740 (32.1%) were clear, 7507 (64.5%) had precipitate, and 401 (3.4%) had crystals. To minimize mislabeling, only cases with more than 80% agreement across multiple ground-truth labelers were kept.

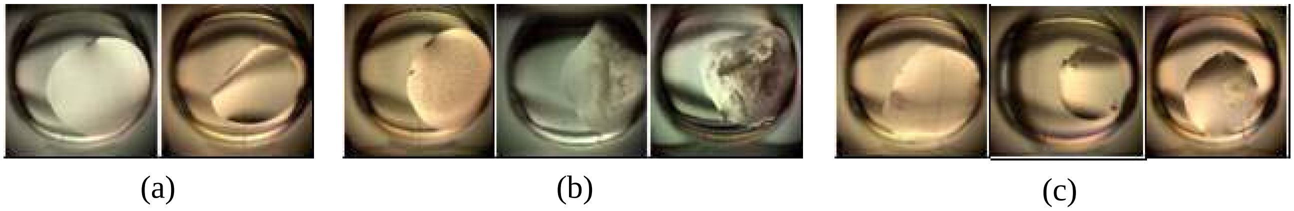

Figure 1. Typical images: (a) Class-1, clear/no hit; (b) Class-2, precipitate/possible hit; (c) Class-3, crystals/hit.

3 FEATURE SPACE

Designing a feature space for our task was difficult, because proteins could crystallize in arbitrary orientations at any location in the experimental well. In addition, there were visually prominent objects which carried no significant information about crystallization. Thus, in theory, the preferred features should be invariant to translation, rotation and scale, and they should not depend on a single prominent object.

The bag-of-words model[15] is applied in this study, because almost every feature of each image had some value. Since the circular well in each image displayed a uniformity of size and position, a static region of interest (ROI) was defined with a center (*W/2*, *W/2*) and a diameter of $7W/10$, where *W* is the image width, in order to ignore pixels outside the well. Within the ROI for each image, a subset of the feature types summarized in Table 2 is extracted.

Table 2. Feature types and their number of coefficients included. Acronyms for the types are given in parenthesis for each feature and will be used throughout this paper.

Type	Count		Type	Count
Gabor filters (GF)	63240		Discrete Fourier transform (FFT)	16384
Gabor marginal scale (GMS)	12648		Gray-level co-occurrence matrix (GLCM)	22
Gabor marginal orientation (GMO)	7905		Shape context descriptors (SC)	20
Gabor average scale (GAS)	5		Corner metrics (CM)	3
Gabor average orientation (GAO)	8			

3.1 Gabor filters

Gabor filters have seen extensive use in both 1-D signal processing and 2-D image processing[20]. In applications such as face recognition[27] and optical character recognition (OCR)[16], they have been applied with particular success. The study problem deals with images of a fixed size which are centered on a well, suggesting the images exhibit a context similarity. Drawing an analogy to face recognition, variation in facial features corresponds to variation in well features.

Gabor filters are bandpass filters. In a one dimensional sense, a Gabor filter is the response of a given input signal to another specifically designed signal, and this has a natural extension to images in two dimensions. The response to the Gabor function is calculated by first changing the representation of the image into its frequency representation using the Fourier transform to produce the coefficients of the different signals. Next, different Gabor filters are applied in order to accentuate or eliminate different signals. The magnitude of the response is measured as the absolute value of the result after inverse-transformation and filtering.

This study utilized forty complex-valued filters comprising eight orientation dimensions and five scale dimensions. A set of magnitude responses for an example image are shown in Figure 2. The responses were then sampled at every 16 pixels inside the region of interest. A total of 40 responses for each of the 1581 sample points gave a total feature set contribution of 63240 dimensions.

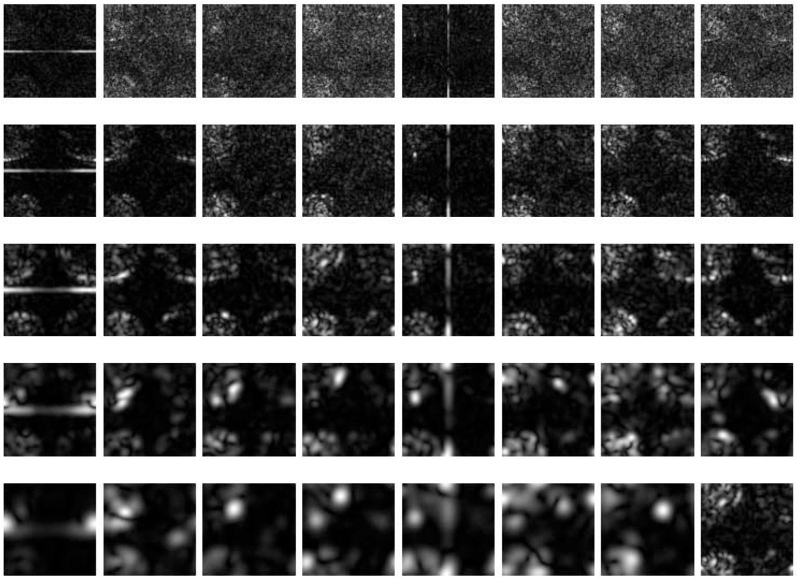

Figure 2. Images showing magnitudes of Gabor filter responses. Each row shows responses for a specific frequency scale while each column shows responses for a specific edge orientation.

Crystals come in different shapes, sizes and orientations. The importance of scale or rotation invariance was not clear, so it was prudent to look at marginals in both of the orientation and scale dimensions. The marginals were defined by taking the average over respective dimensions in the array of responses in Figure 2. Since scale was varied on the vertical axis, there existed eight images representing marginals-of-scale. Similarly, there were five images representing marginals-of-orientation for each image in the dataset. Sampling each image at 1581 points on a grid gave feature space contributions of 12648 marginals-of-scale and 7905 marginals-of-orientation. Furthermore, it was possible to create spatially invariant versions of the marginals by averaging the data for all of the points at each scale and each orientation, resulting in 5 spatially invariant marginals for scales and 8 spatially invariant marginals for orientations.

3.2 Discrete Fourier transform

Fourier transforms are a method for decomposing a signal from an arbitrary domain into its corresponding frequency domain. For frequency-based features, the spectral plot of 2-dimensional discrete Fourier transforms was computed for each image using the fast Fourier transform (FFT). Of note is the relationship between Fourier transforms and Gabor filters. Gabor filters can be considered as a Fourier-type transform that is localized by a Gaussian window in the frequency domain.

In this study, the transforms were downscaled from the original image size to 128x128, resulting in 16,384 flat features. An example of the 2-D FFT performed on three full size images with fairly different appearance characteristics is shown in Figure 3. In comparison, Figure 4 shows the 2-D FFT of the downscaled images. The size of 128x128 was the lowest resolution of rescaling that captured the information of the original spectral plots based on visual inspection.

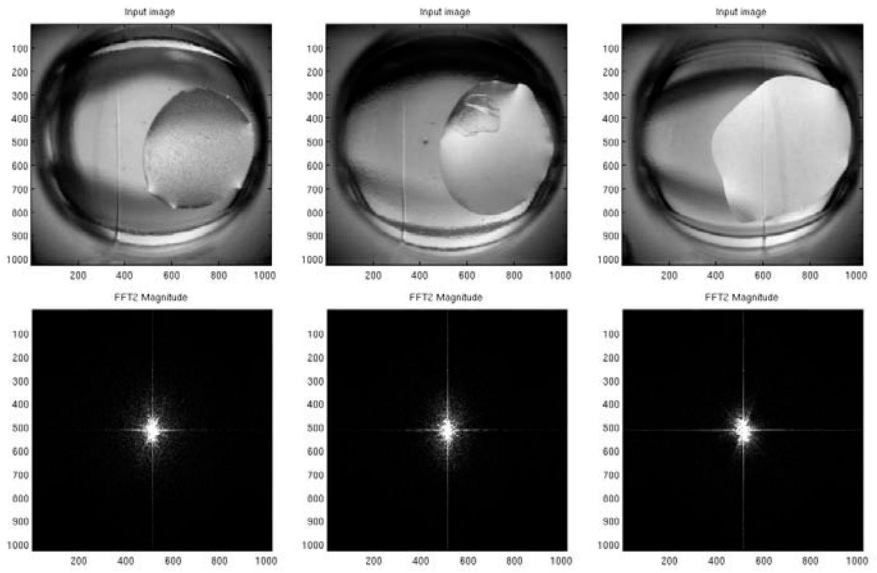

Figure 3. Examples of discrete Fourier transform. Top row: original images with the full 1024 by 1024 resolution. Bottom row: corresponding 2D DFT spectrum image.

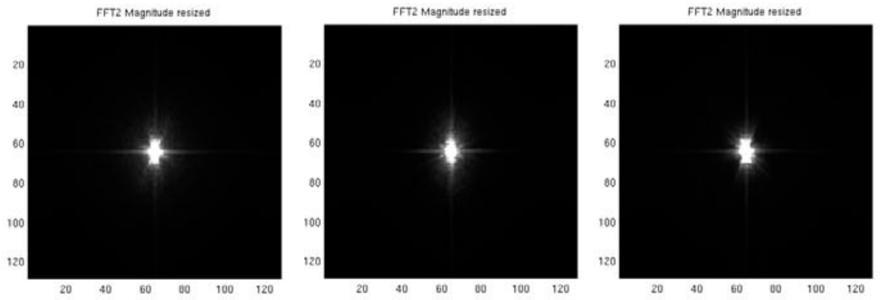

Figure 4. Examples of discrete Fourier transform with down-sampled input images (128 x128). The same input images in Figure 3 are used.

3.3 Gray-level co-occurrence matrix

The gray-level co-occurrence matrix (GLCM)[13] is a technique used in image processing and recognition to summarize an image's texture based on patterns in the intensity spectrum. A GLCM is calculated from a grayscale image. If a grayscale image *I* has *N* scales, then the GLCM is an *N* x *N* matrix *M* such that $M_{i,j} = \sum_{x=1}^{N} \sum_{y=1}^{N} 1$, if $I(x,y) = i$ and $I(x,y+1) = j$ and 0 otherwise. Many different features can be extracted from this matrix, all of which reflect the global texture of a given image in some way. A list of the GLCM texture features implemented is given in Table 3. All measures are calculated from a GLCM matrix *M*, and the following notation is used in the table. M_x and M_y represent the marginals along the vertical (x) and the horizontal (y) dimension. The means and standard deviations are represented by μ_x, μ_y and σ_x, σ_y. | *M* | denotes $\sum_i \sum_j M_{i,j}$. $M_{x+y}(k) = \sum_i \sum_j M_{i,j}$ | $i + j = k$ for k = 2, 3, …, 2*N* and $M_{x-y}(k) = \sum_i \sum_j M_{i,j}$ | $|i - j| = k$ for k = 0, 1, …, *N*-1. Finally, marginal entropies are defined as follows: $H^1_{x,y} = -\sum_i \sum_j M_{i,j} \log M_x(i) M_y(j)$, $H^2_{x,y} = -\sum_i \sum_j M_x i M_y(j) \log M_x(i) M_y(j)$, and $H_{x,y} = -\sum_i \sum_j M_{i,j} \log M_{i,j}$.

Table 3. Description of GLCM texture features

Feature	Equation	Feature	Equation
Autocorrelation[23]	$\sum_i \sum_j (ij) M_{i,j}$	Maximum probability [23]	$\max_{i,j} M_{i,j}$
Contrast [23]	$\sum_{n=0}^{N-1} \{ \sum_i \sum_j M_{i,j} \,\|\, i-j \,\|= n \}$	Sum of squares: variance [13]	$\sum_i \sum_j (i-\mu_x) M_{i,j}$
Correlation$_1$ [6]	$\sum_i \sum_j \frac{(i-\mu_x)(j-\mu_y) M_{i,j}}{\sigma_x \sigma_y}$	Sum average [13]	$\sum_{i=2}^{2N} i M_{x+y}(i)$
Correlation$_2$ [13]	$\sum_i \sum_j \frac{(ij) M_{i,j} - \mu_x \mu_y}{\sigma_x \sigma_y}$	Sum variance [13]	$\sum_{i=2}^{2N} (i + \sum_{i=2}^{2N} M_{x+y}(i) \log M_{x+y}(i))^2 M_{x+y}(i)$
Cluster prominence [23]	$\sum_i \sum_j (i+j-\mu_i-\mu_j)^4 M_{i,j}$	Sum entropy [13]	$-\sum_{i=2}^{2N} M_{x+y}(i) \log M_{x+y}(i)$
Cluster shade [23]	$\sum_i \sum_j (i+j-\mu_i-\mu_j)^3 M_{i,j}$	Difference variance [13]	$\sum_i \sum_j i^2 M_{x-y}$
Dissimilarity [23]	$\sum_i \sum_j \|i-j\| M_{i,j}$	Difference entropy [13]	$\sum_i \sum_j M_{x-y}(i) \log M_{x-y}(i)$
Energy [23]	$\sum_i \sum_j M_{i,j}^2$	Information of correlation$_1$ [13]	$\frac{H_{x,y} - H_{x,y}^1}{\max\{H_x, H_y\}}$
Entropy [13]	$-\sum_i \sum_j M_{i,j} \log M_{i,j}$	Information of correlation$_2$ [13]	$\sqrt{1 - exp\{-2(H_{x,y}^2 - H_{x,y})\}}$
Homogeneity$_1$ [3]	$\sum_i \sum_j \frac{M_{i,j}}{1+\|i-j\|}$	Inverse difference [6]	$\sum_i \sum_j \frac{M_{i,j}/\|M\|}{1+\|i-j\|}$
Homogeneity$_2$ [23]	$\sum_i \sum_j \frac{M_{i,j}}{1+(i-j)^2}$	Inverse difference moment [6]	$\sum_i \sum_j \frac{M_{i,j}/\|M\|}{1+(i-j)^2}$

3.4 Shape context descriptors

Shape context descriptors[1] were originally proposed for the machine vision problem of OCR. They are principally utilized to test if image objects match prototype shapes in a manner that is resistant to shear, rotation, and scaling. This method assumes that the background and foreground are easily separable, and the objects are consistent with pre-defined prototype shapes.

The core of the shape context algorithm is to consider each point as a "focus point" from the set of boundary points, and capture the relationship of all other sampled points to it. From the "focus point", its magnitude and angle to all other sampled points is computed. Every angle-log (magnitude) pair is then binned into a 2D histogram, and this 2D histogram is created for each focus point. The binning gives the perspective tolerance for a moderate amount of change, while not allowing a radical change in the overall shape.

For this study, the standard shape context features were modified to handle our problem due to the lack of consistency in appearance between images. Since the foreground is not easily defined semantically and crystals can vary radically in structure, binarization and edge detection targeting crystal like objects is very difficult. Two modifications were made in our implementation. First, basic edge detection rather than a form of gradient boundary detection is used to determine the sample points. Second, rather than try to create a prototypical shape to match other shape contexts against, a composite shape context descriptor was used as a feature for each image. This feature is based on a 5x5 histogram, scaled locally to the image, and is the arithmetic mean of all the histograms of approximately 100 sampled edge points. This feature is still resistant to rotation, scaling, and shear, but it is less descriptive than a prototypical shape context.

It is important to note that there were only 20 useful features to come out of this 5x5 matrix. The remaining features were consistently zero for our dataset; consequently we removed them from the feature space.

3.5 Corner metrics

Crystalline objects visually appear to have more corners than other well artifacts, suggesting corner metrics may be a useful feature. Three different corner metrics are implemented in this study. A Harris operator corner detector[14] was built with a tolerance of 0.1. Another Harris corner detector was built with an identical tolerance of 0.1, but was limited to returning a maximum of 2000 corner points. Lastly, a Shi-Tomasi corner detector[22] was built with a quality of 0.5.

4 METHOD

The general method for the experiments consisted of the following steps. A set of features was first extracted from each image. Feature selection was performed next to reduce the input dimensions, thereby avoiding overfitting. Finally, classification models were built from the reduced feature subset and tested by 10-fold cross-validation to determine the best performing model.

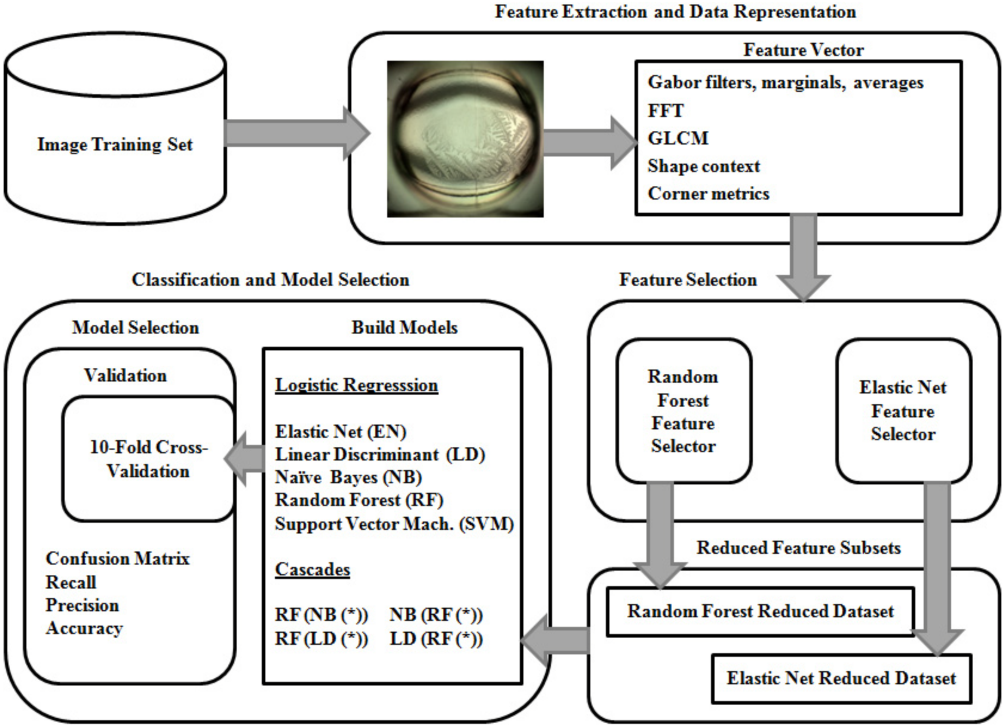

Figure 5. Schematic diagram of our automatic protein crystal classification system and experimental validation of its performance

4.1 Feature selectors

The number of starting raw features is quite large. It is beneficial to perform feature selection to obtain a subset comprising the well-performing features to achieve the maximal classification accuracy, while easing the joint burdens of time and space complexity. Two data driven selection methods are considered in this study.

The random forest (RF) selector comes with a built-in variable importance measure that is based on ranking either average Gini or information gain increases of decision trees[5]. This ranking can then be used to select the most important features without rerunning the RF training. By selecting the top *N* number of features, it is possible to achieve accuracy arbitrarily close to the accuracy of using the entire feature set. More details regarding RF are given in the next section.

The elastic net (EN) selector is built into the EN classifier[28]. The EN classifier creates a natural form of feature selection via L^1 regularization penalties. A combination of ridge and lasso penalties tend to set the values of non-predictive variables to zero and reduce highly correlated groups of features to a single feature. More details regarding EN are also given in the next section.

4.2 Supervised machine learning classifiers

Several supervised classification models were considered in our application context: naive Bayes (NB), linear discriminant analysis (LD), random forest (RF), elastic net (EN), and support vector machines (SVM). In addition, some classifiers were combined in a two-tier cascade. All classifiers were trained from our dataset for our three-class problem and subjected to 10-fold cross validation for model selection.

The NB classifier[21] is a standard probabilistic classifier with the naive assumption that features are conditionally independent. We learned the feature-wise likelihood and class prior distributions from our annotated dataset. A NB classifier then outputted the class label that maximizes its posterior probability given a test datum.

The LD classifier[11] is another standard classifier based on a feature space transformation that maximizes the between-class scatter, while concurrently minimizing the within-class scatter. For our 3-class problem, LD yielded a 2-D subspace onto which test data can be projected for easier classification. An LD classifier then compared distances between the test to the mean for each class in the subspace and outputted the label of the mean closest to the test.

The RF classifier[5] is a popular ensemble classifier based on decision trees (DTs). Bootstrapped samples were first created by randomly sampling a training set with replacement. A RF was built by training a decision tree with each bootstrapped sample, using only a randomly sampled subset (with replacement) of features at each node. Given a test set, each DT generated a classification; a plurality voting of these decisions yielded the RF's output. For our dataset, 100 trees and 2000 features were used for constructing our RF models. The number of features to split on at each node is set to 2000, which approximates the square root of the total number of features. In our pilot study, we observed that increasing the number of trees beyond this default value did not improve performance. Precision reached a maximum using the 2,000 most highly ranked features as determined by RF's innate variable importance.

The EN classifier[28] is a logistic regression method that fits the linear regression model, $y = \beta * x$, with an output range of 0 to 1, where y is the class likelihood for the input x, and β are the regression coefficients. For a multi-class problem like ours, the model is fit for each class and the classifier determines which class maps to the input by picking the class with the highest likelihood for those features. EN is a generalized linear regression technique, used here in the logistic sense for classification. It efficiently chooses an optimal model which is robust in combining the L^2 regularization of the ridge regression and L^1 regularization of the lasso regression in a weighted combination: $(1 - \alpha)$ x L^2-penalty + α x L^1-penalty with $\alpha \in [0, 1]$. The resulting model with an appropriate normalization inherits the natural built-in feature selection of the lasso method to produce a sparse model while avoiding its shortcomings. For example, at most n variables can be selected when $p >> n$, where p and n denote the numbers of features and data cases, respectively. To choose the optimal parameters for EN, a grid search was run on α and λ, where α was varied between 0.1 and 1.0 in increments of 0.1 and a full 100 point regularization path was built for λ according to the elastic net. In choosing the best (α, λ) combination, we optimized for classification accuracy by adjusting the sensitivity and the cost of incorporating new variables into the regression.

The SVM classifier[8] is a flexible and popular classification method that inherently handles high-dimensional data. It supports kernels for transforming what are normally linear hyperplanes into manifolds of arbitrary dimensions and shapes. This allows for more simple linear training algorithms to train models that are more complex than linear relationships. In this study, we utilized a linear kernel to avoid overfitting.

Cascade classifiers have been used in previous studies in this area. A classifier of the form X-RF is one in which classifier X is used to detect positives and only afterwards is RF used on those data deemed negative by X. A classifier of the form RF-X is one in which RF is first used to detect negatives, and only afterwards is X used to detect positives from the remainder. For our experiments, we evaluated four cascade combinations that took the form X-RF or RF-X, where X was either the NB or the LD classifier.

5 RESULTS

The experimental results are covered in this section. Since techniques that have different natural validation methodologies are compared, for example RF's out-of-bag error, we imposed standardization by insisting on using 10-fold cross-validation for all cases. For this application, a good 3-category classifier is one which is weighted against predicting crystals to be negatives. Thus, we closely examined the recall of the crystal class, along with the corresponding precision and accuracy. The precipitate and clear classes were not included, because the goal was to positively identify crystals.

The recall, precision and accuracy are defined in the following manner. Recall is the number of true positives divided by the number of actual positives. Recall is calculated from the confusion matrix by taking the lower right term and dividing it by the sum of the bottom row, where a confusion matrix is organized with columns indicating predicted labels and rows indicating true labels in the order of clear, precipitate, and crystal. Precision is the number of true positives divided by the number of positives. Precision is calculated by taking the lower right term and dividing it by the sum of the right-most column. Accuracy is simply the number of correct classifications divided by the total number of predictions. Correct classifications are found on the main diagonal of the confusion matrix.

Due to the skew in the dataset caused by the difficulty in positively identifying crystals and the importance of not missing any potential crystal, experimental results were also calculated for the 2-category classification task. Two categories were created from the original three categories by combining the precipitate and crystal classes together. We reported only the accuracy, precision and recall of the combined precipitate/crystal class in the tables.

5.1 Spatially-invariant features with single stage classifier

A spatially-invariant feature set composed of Gabor average scale (GAS) and orientation (GAO), GLCM, shape context descriptors (SC) and corner metrics (CM) was chosen for this experiment. The size of the set totaled 58 features. Due to the small number of features, feature selection was not performed for this experiment. For the EN classifier, α was set to 0.2 and λ was set to 0.0005365409 based on the grid parameter search. For the SVM classifier, a linear kernel was used with unweighted classes and a regularization cost C of 1. The experimental results are shown in Table 4.

The 3-category recall is poor for all classifiers, indicating no model is adept at positively identifying the crystal class. This result is undesirable to crystallographers, given that so few images yield crystals. Even precipitate formations can give hints to which cocktails could be successful.

Loss in this study is defined as the classification of a crystal or precipitate image as clear. Precipitate loss is serious, and crystal loss is clearly unacceptable beyond the absolute minimum. In addition, the subjective nature of expert labeling can result in the initial misclassification of some crystal images as precipitate images. In order to avoid these issues, the 3-category problem can be reduced to a two categories by combining the precipitate and crystal classes together. The 2-category recall of the combined precipitate/crystal class captures the information most significant to the crystallographer.

The 2-category recall for the EN and RF classifiers shows these models are capable of finding an image of interest approximately 88% of the time. However, the 2-category recall for the SVM performed significantly worse, only recalling precipitate/crystal images at a rate of 68%.

Table 4. Classification performance of the spatially-invariant features. (GAS, GAO, GLCM, SC, CM)

		2-Category			3-Category		
Selector	Classifier	Accuracy	Precision	Recall	Accuracy	Precision	Recall
none	EN	81.54%	84.39%	87.88%	78.22%	63.64%	1.72%
none	RF	82.88%	85.42%	88.88%	79.49%	75.00%	0.74%
none	SVM	61.43%	80.02%	68.41%	58.85%	0.00%	0.00%

An assessment of the effectiveness of FFT features was made by adding FFT to the feature set of the previous experiment. The size of this set totaled 16439 features. Feature selection was not performed to ensure that all the FFT features were evaluated. For the EN classifier, α was set to 0.4 and λ was set to 0.02216993 based on the grid search. For the SVM classifier, the parameters were the same as the previous experiment. The experimental results are shown in Table 5.

The 2-category recall for the RF and SVM classifiers did not change with the addition of the FFT features. However, the 2-category recall for the EN improved by 2% to about 90%.

Table 5. Classification performance of the spatially-invariant features including FFT. (GAS, GAO, GLCM, SC, CM, FFT)

		2-Category			3-Category		
Selector	Classifier	Accuracy	Precision	Recall	Accuracy	Precision	Recall
none	EN	84.48%	87.13%	89.66%	81.07%	0.00%	0.00%
none	RF	83.18%	86.75%	88.23%	79.78%	100.00%	0.73%
none	SVM	56.93%	68.11%	68.25%	53.88%	1.34%	0.49%

Next, we repeat the experiment in Table 5 with automatic feature selection. The same spatial invariant feature set was chosen for this experiment. In this case, the EN feature selector was applied to reduce the number to 108 features. The breakdown of the remaining features was GLCM(6), shape context(11), corner matrix(1), average Gabors(5) and FFT(85). For the EN classifier, α was set to 0.4 and λ was set to 0.02216993 based on the grid search. For the SVM classifier, the parameters remained the same. The experimental results are shown in Table 6.

The 2-category recall for the SVM classifier continued to remain insensitive to changes in the FFT features. The RF and EN classifiers behaved differently with the reduction in FFT features. The RF classifier recall improved by 1% by eliminating possible overfitting. In contrast, the EN classifier recall worsened by 2% as the eliminated features apparently carried useful information.

Table 6. Classification performance of the spatially-invariant features including FFT, after feature selection with EN. (GAS, GAO, GLCM, SC, CM, FFT)

		2-Category			3-Category		
Selector	Classifier	Accuracy	Precision	Recall	Accuracy	Precision	Recall
EN	EN	82.99%	87.05%	87.74%	79.57%	0.00%	0.00%
EN	RF	84.31%	87.10%	89.45%	80.92%	100.00%	0.25%
EN	SVM	61.20%	78.51%	68.69%	58.80%	0.00%	0.00%

5.2 Spatially-variant features with single stage classifier

The poor 3-category recall performance for the spatially-invariant feature sets suggested that there was some aspect of the crystal growth that may be dependent on the position within the well. To test this hypothesis, we replaced the Gabor average features with spatially-variant features such as the Gabor filter and their marginals with respect to scale and orientation. Considering that the total number of features became 83835, feature selection was required to prevent overfitting. Due to its poor 2-category recall, the SVM classifier was replaced by a NB and a LD classifier. The next two experiments compare the performance of the RF and EN feature selectors.

Application of the RF feature selector reduced the number to 1321 features. The breakdown of the remaining features was GF(909), GMS(356), GMO(54), GLCM(0) and SC(2). Application of the EN feature selector reduced the number to 2042 features. The breakdown of the remaining features was GF(1278), GMS(367), GMO(384), GLCM(1) and SC(12).

The experimental results for the RF selector are shown in Table 7 while the experimental results for the EN selector are shown in Table 8. The switch to the EN feature selector clearly improved both the 2-category and 3-category performance of the NB and LD classifiers. The RF classifier seemed insensitive to the choice of feature selector. The 2-category recall for the RF classifier was 2% better than the spatially-invariant result in Table 5, supporting the need for spatially-variant features. The 2-category recall for the NB classifier was the best at 92%, but the low precision suggests that some images are classified incorrectly as crystal images. In terms of 2-category accuracy alone, the combination of RF feature selector and classifier yielded the best result of 85.37%. In terms of 2-category recall alone, the combination of EN feature selector and NB classifier yielded the best result of 94.11%. However, the low precision again suggests misclassification. For both the RF and EN classifiers, the best combination of accuracy and recall was achieved by choosing the matching algorithm for feature selection. Both the RF and EN classifiers showed better accuracy and recall than the spatially-invariant results in Table 6.

Table 7. Classification performance of the spatially-variant features, after feature selection with RF. (GF, GMS, GMO, GLCM, SC)

		2-Category			3-Category		
Selector	Classifier	Accuracy	Precision	Recall	Accuracy	Precision	Recall
RF	NB	73.33%	66.16%	92.39%	57.13%	12.68%	69.83%
RF	LD	75.45%	79.00%	84.46%	69.74%	17.99%	21.45%
RF	RF	85.37%	87.68%	90.47%	82.02%	33.33%	0.50%

Table 8. Classification performance of the spatially-variant features, after feature selection with EN. (GF, GMS, GMO, GLCM, SC)

		2-Category			3-Category		
Selector	Classifier	Accuracy	Precision	Recall	Accuracy	Precision	Recall
EN	NB	77.22%	70.89%	94.11%	66.35%	18.05%	69.33%
EN	LD	77.78%	81.47%	85.16%	73.09%	24.33%	22.69%
EN	RF	85.02%	87.41%	90.22%	81.67%	0.00%	0.00%
EN	EN	85.31%	86.18%	91.69%	81.96%	56.76%	5.24%

Recall that the high value of crystal images in this application means any model that falsely rules out crystals is not desirable. However, this involves a trade-off between recall and precision since the original goal is to reduce the search space. We would like high accuracy combined with a low false negative rate. The combination of RF feature selector and classifier yielded the best 3-category accuracy of 82.02% with the combination of EN feature selector and classifier just behind at 81.96%. However, Table 9 clearly shows the EN-EN combination had fewer false negatives than the RF-RF combination.

Table 9. Number of crystals incorrectly classified as clear in 3-category experiments for both RF reduced and EN reduced datasets

	Classifier			
Selector	NB	LD	RF	EN
RF	40	35	13	N/A
EN	19	38	11	5

5.3 Spatially-variant features with cascade classifier

The combination of a classifier with better recall and a classifier with better precision is a logical conclusion from the previous experiments. To exploit this observation, this section explores two-tier cascade classifiers as our model. A cascade classifier is constructed to first detect clear images. The second stage sorts through the remaining images to separate the precipitate class from the crystal class. Since the choice of feature selector is not conclusive from the previous experiment, both the RF and EN selectors were compared again with the cascade classifiers.

The experimental results for the RF selector are shown in Table 10 while the experimental results for the EN selector are shown in Table 11. In comparison with the RF selector, the EN selector clearly improved the 3-category accuracy and precision of all of the cascaded classifiers without negatively affecting the recall. In terms of 2-category accuracy alone, the combination of RF feature selector and the LD-RF cascade classifier yielded the best result of 85.35%. In terms of 2-category recall alone, the combination of EN feature selector and RF-NB classifier yielded the best result of 95.88%, but the low precision continued to be a drawback. For both selectors, the cascades that ran the RF classifier last provided the best combination of accuracy and recall, suggesting that RF is a potentially good choice for identifying false negative crystals among cases classified as precipitate.

Table 10. Classification performance of cascade classifiers with spatially-variant features, after feature selection with RF. (GF, GMS, GMO, GLCM, SC)

			2-Category			3-Category		
Selector	Classifier 1	Classifier 2	Accuracy	Precision	Recall	Accuracy	Precision	Recall
RF	RF	NB	74.31%	65.68%	94.92%	58.25%	12.85%	68.33%
RF	NB	RF	85.07%	87.95%	89.85%	68.58%	12.66%	69.83%
RF	RF	LD	79.36%	74.66%	93.05%	73.90%	18.18%	21.45%
RF	LD	RF	85.35%	87.70%	90.43%	79.52%	17.92%	21.45%

Table 11. Classification performance of cascade classifiers with spatially-variant features, after feature selection with EN. (GF, GMS, GMO, GLCM, SC)

			2-Category			3-Category		
Selector	Classifier 1	Classifier 2	Accuracy	Precision	Recall	Accuracy	Precision	Recall
EN	RF	NB	77.58%	69.98%	95.88%	66.95%	18.70%	68.33%
EN	NB	RF	84.87%	87.77%	89.72%	73.89%	18.05%	69.33%
EN	RF	LD	80.56%	76.43%	93.79%	75.95%	24.46%	22.44%
EN	LD	RF	85.02%	87.44%	90.19%	80.09%	24.33%	22.69%

6 DISCUSSION

This study developed an automated process for classifying protein crystallization images by applying Elastic Net (EN) for both feature selection and as a model of protein crystal image classification. The best performing regularized, logistic regression model utilizes EN to select the features and a two-tier cascade classifier using naïve Bayes followed by random forest algorithm, resulted in performance competitive to those reported by the previous studies.

The inclusion of spatially-variant features improved the performance of the model (see Table 6 and Table 8). This likely reflects the selection of the liquid drop as the region of interest (ROI). The drop is unlikely to be located concentric with the hemispheric well. The position of the drop in the well along with surface wetting factors alters the ratio of the surface area to volume, causing the precipitation probability to vary within the ROI. Interestingly, the RF feature selector removed most of the orientation-based Gabor marginals (GMO), while the EN selector kept them. For all single and cascade classifiers, the EN feature selector clearly performed superior to the RF feature selector.

There seems to be substantial overlap in the information captured by the Gabor filters (GF) and the FFT. As mentioned earlier, the GF can be considered a localized FFT. By forming marginals and averages, the GF has been transformed into a global frequency measure, similar to FFT. This theory is supported by the lack of improvement when FFT is added to the spatially-invariant feature set (see Table 4 and Table 5). When Gabor averages are present, the EN feature selector trimmed the raw FFT feature count from 16384 to 85.

Most of the prior studies developed models for datasets of limited size (see Table 1). The study by Cumbaa and Jurisica '05 used a dataset with a similar percentage of crystal images as our study, but much fewer features. In their 2010 follow-up study, they used a dataset approximately the same size as ours, but their percentage of crystal images was much higher. Their number of features was comparable, and they used a RF classifier. Their model reached a crystal/clear accuracy of 61.6% and recall of 80.2%, while we achieved a 3-category crystal accuracy of 73.89% and recall of 69.33% (see row 2 of Table 11). In order to compare with their precipitate/clear result, we can combine our crystal images and precipitate images to form a 2-category result. In this case, their model reached a precipitate/clear accuracy of 86.4% and recall of 88.7%, while we achieved a 3-category crystal accuracy of 84.87% and recall of 89.72%.

One of the most difficult challenges in this study is developing a model that can cope with the data imbalance between the three classes. The highest value, crystal class only comprised a small percentage of the data, making it difficult to simultaneously achieve high accuracy and high recall. Some classifiers, such as the SVM, have been reported to perform poorly on imbalanced datasets[1]. Future work on this application will search for methods to extend the performance of our model. For example, it is possible to change the relative weight of the classes, giving a larger weight to the crystal class due to its small size and high value. Rebalancing the dataset is another option to increase the presence of the crystal class. Cumbaa and Jurisica appeared to do this when they increased the percentage of the crystal images in their 2010 study. Lastly, collecting more expert evaluated data will benefit the entire community working on solutions to the protein crystallization problem.

ACKNOWLEDGEMENT

The authors would like to thank Genentech, Inc. for its invaluable support and providing access to the data used in this study. We also thank Kathryn Woods, Qixin Bei, Weiru Wang, Jeff Harris, Ilmi Yoon, and Dragutin Petkovic for their invaluable support. This research was also partially supported by the SFSU Center of Computing for Life Science.

REFERENCES

[1] Akbani, R., Kwek, S. and Japkowicz, N., "Applying support vector machines to imbalanced datasets," Proc. Machine Learning: ECML, 39-50 (2004).

[2] Belongie, S., Malik, J. and Puzicha, J., "Shape context: A new descriptor for shape matching and object recognition," Advances in Neural Information Processing Systems v2, 3-9 (2000).

[3] Ben Salem, Y. and Nasri, S., "Rotation invariant texture classification using support vector machines," 2011 Int. Conf. on Communications, Computing and Control Applications (CCCA), 1-6 (2011).

[4] Bern, M., Goldberg, D., Stevens, R.C. and Kuhn, P., "Automatic classification of protein crystallization images using a curve-tracking algorithm," Journal of Applied Crystallography 37(2), 279-287 (2004).

[5] Breiman, L., "Random forests," Machine Learning 45(1), 5-32 (2001).

[6] Clausi, D. A., "An analysis of co-occurrence texture statistics as a function of grey level quantization," Canadian Journal of Remote Sensing 28(1), 45-62 (2002).

[7] Collins, J., "Machine learning for large image datasets," Master's thesis, San Francisco State University (2013).

[8] Cortes, C. and Vapnik, V., "Support-vector networks," Machine learning 20(3), 273-297 (1995).

[9] Cumbaa, C. and Jurisica, I., "Automatic classification and pattern discovery in high-throughput protein crystallization trials," Journal of Structural and Functional Genomics 6(2-3), 195-202 (2005).

[10] Cumbaa, C. and Jurisica, I., "Protein crystallization analysis on the world community grid," Journal of Structural and Functional Genomics 11(1), 61-69 (2010).

[11] Duda, R. O., Stork, D. G. and Hart, P. E., [Pattern classification, 2nd ed.], John Wiley, New York (2001).

[12] Durbin, S.D. and Feher, G., "Protein crystallization," Annual Review of Physical Chemistry 47(1), 171-204 (1996).

[13] Haralick, R. M., Shanmugam, K., and Dinstein, I., "Textural features for image classification," IEEE Trans. on Systems, Man and Cybernetics SMC-3(6), 610-621 (1973).

[14] Harris, C. and Stephens, M., "A combined corner and edge detector," Proc. Alvey Vision Conf. v15, 50 (1988).

[15] Harris, Z., "Distributional structure," Word 10(23), 146-162 (1954).

[16] Huo, Q., Ge, Y. and Feng, Z. D., "High performance Chinese OCR based on Gabor features, discriminative feature extraction and model training," Proc. Acoustics, Speech, and Signal Processing, v3, 1517–1520 (2001).

[17] Kawabata, K., Saitoh, K., Takahashi, M., Asama, H., Mishima, T., Sugahara, M. and Miyano, M., "Evaluation of protein crystallization state by sequential image classification," Sensor Review, v28(3), 242-247 (2008).

[18] Majid, A. S., de Paredes, E. S., Doherty, R. D., Sharma, N. R. and Salvador, X., "Missed breast carcinoma: Pitfalls and pearls1," Radiographics 23(4), 881-895 (2003).

[19] Newland, O., "Highly loss-sensitive classification of protein crystallization images," Master's thesis, San Francisco State University (2013).

[20] Okada, K., Steffens, J., Maurer, T., Hong, H., Elagin, E., Neven, H. and von der Malsburg, C., [Face Recognition: From Theory to Applications], Springer, Berlin, 186-205 (1998).

[21] Russell, S. and Norvig, P., [Artificial Intelligence: A Modern Approach, 3rd ed.], Prentice-Hall, San Francisco, 808-809 (2010).

[22] Shi, J. and Tomasi, C., "Good features to track," Proc. Computer Vision and Pattern Recognition, 593-600 (1994).

[23] Soh, L. and Tsatsoulis, C., "Texture analysis of sar sea ice imagery using gray level co-occurrence matrices," IEEE Trans. on Geoscience and Remote Sensing 37(2), 780-795 (1999).

[24] Vigna, G., Valeur, F., Balzarotti, D., Robertson, W., Kruegel, C., and Kirda, E., "Reducing errors in the anomaly-based detection of web-based attacks through the combined analysis of web requests and SQL queries," Journal of Computer Security, 17(3), 305-329 (2009).

[25] Wilson, J., "Towards the automated evaluation of crystallization trials," Acta Crystallographica Section D: Biological Crystallography 58(11), 1907-1914 (2002).

[26] Yang, X., Chen, W., Zheng, Y., and Jiang, T., "Image-based classification for automating protein crystal identification," [Intelligent Computing in Signal Processing and Pattern Recognition], Springer, Berlin, 932-937 (2006).

[27] Zhou, M. and Wei, H., "Face verification using Gabor wavelets and Adaboost," in Proc. Conf on Pattern Recognition v1, 404 –407 (2006).

[28] Zou, H. and Hastie, T., "Regularization and variable selection via the Elastic Net," J. of the Royal Statistical Society: Series B (Statistical Methodology) 67(2), 301-320 (2005).