ANALYSIS OF RESONANCE RAMAN SPECTRA OF CANCEROUS SKIN TISSUES USING ARTIFICIAL NEURAL NETWORKS

BY

JACOB BAILIN

An Honors Thesis Submitted to the Department of Physics

Southern Connecticut State University
New Haven, Connecticut
April 21st, 2017

This honors thesis was prepared under the direction of the candidate's thesis advisor, Dr. Binlin Wu, and it has been approved by the members of the candidate's thesis committee. It was successfully defended and accepted by the University Honors Thesis Committee.

____
Dr. Binlin Wu
Thesis Advisor

____
Dr. Elliott Horch
University Reader

____
Dr. L. Evan Finch
Second Reader

____
Dr. Matthew Enjalran
Department Chairperson

____
Date

1 Introduction

Skin cancer is by far the most common of all types of cancer diagnoses made each year. Approximately 5.4 million instances of basal cell and squamous cell skin cancers are diagnosed each year in America. Of these diagnoses, about 8 out of 10 are basal cell carcinoma (BCC)[1] BCC is rarely malignant and does not typically metastasize to other areas of the body, but it grows rapidly and can be permanently disfiguring. Because of the extremely high incidence rate of this type of skin cancer, fast detection, accurate diagnosis, and rapid treatment is very important.

Currently, the gold standard for BCC diagnosis is through tissue biopsy and H&E histopathology. This indirect method of diagnosis is very invasive and requires experts in the field in order to make a diagnosis, and the diagnostic accuracy is quite low at around 80%. The invasiveness of the procedure, the dependency on highly specialized experts in histopathology, and the relatively low diagnostic accuracy is more than enough cause for the need for a more effective method of diagnosis.

Optical Biopsy (OB) techniques provide a much less invasive and highly quantifiable form of analysis of BCC tissues. Types of OB techniques include Fluorescence Spectroscopy (FS), Elastic Scattering Spectroscopy (ESS) and Raman Spectroscopy (RS). These OB techniques can be performed in vivo, in situ, and in real time,[2] making them superior methods of analysis than traditional biopsy. RS possesses the useful ability to distinguish intrinsic biomarkers within samples, making it a good selection for an OB technique.

Types of RS include Near Infra-Red (NIR) RS, as well as Resonance Raman (RR). NIR data possesses a low signal to noise ratio and a high background, and it is only able to detect non-resonant bio-molecular components, making its practical application limited when compared with RR. In RR, incident light with a frequency very near the energy of electronic transitions of compounds within the sample is used, allowing for a much greater signal to noise ratio, lower background, and much sharper more distinguished peaks within the gathered data. Certain peaks within the gathered data correspond to specific biological structures in normal and BCC afflicted tissues, and are what allow for discrimination between the two.

Dr. Binlin Wu and his colleagues have developed a RR technique using a visible wavelength of 532 nm as the excitation light source, which allows for much more clear and less noisy data than NIR and RR techniques previously used, called the Visible Resonance Raman (VRR) method.[3] The data can be classified as cancerous or normal using peak analysis only, but this method may not be robust. Depending on experimental conditions when gathering data with VRR, the absolute values of the peaks as well as the background can vary greatly. A much more robust method is to compare the relative difference in size between the peaks, which can be made consistent by using machine learning for dimensional reduction and classification as opposed to traditional peak analysis.

Using the VRR technique, multiple data samples of BCC and normal skin tissues have been gathered, and different forms of statistical analysis and machine learning techniques have been used to classify the data. Dr. Wu has used Principal Component Analysis (PCA)[4] for dimension reduction of the data in conjunction with Support Vector Machines (SVM)[5] to classify the data based on the scores of the individual Principal Components (PCs), with good results. Another method our group has been investigating is nonnegative matrix factorization (NMF).[5] That method uses NMF to decompose the spectral data like PCA, and the scores of NMF components are used for classification with SVM.

Based on our preliminary study, it is not for certain if PCA and NMF can robustly detect characteristic information in the RR spectral data, and both these methods have their limitations. For example, PCA provides negative values in the components, and it is difficult to interpret the basis spectra. NMF generates components that carry common spectral peaks and are also difficult to interpret. Both of these methods have linearity assumption. Therefore, it is worth trying other methods for comparison.

Artificial Neural Network (ANN) is another well-known machine learning method, and has been successfully used to dimensionally reduce and classify non-linear data a number of times in many different fields.[6-7] Based on the high dimensionality of the data sets as well as the possibility of non-linearity within the data, it is believed that the implementation of an ANN as a method of dimensional reduction may be able to detect characteristic information which could be subtle, and produce results that are similar or even better than those provided by PCA and NMF. This hypothesis will be tested through experimentation on the data with implementation of an ANN, and comparison with previous results of PCA and NMF.

2 Resonance Raman Spectroscopy

In order to understand Resonance Raman (RR) spectroscopy, traditional Raman Spectroscopy (RS) must be understood. When a monochromatic light source is shined on a sample, the sample scatters the light. The vast majority of light scattered by the sample is of the same wavelength as the incident light; this process is known as Rayleigh or elastic scattering.[9] In addition to Rayleigh scattering, a very small amount of incident light is scattered at a slightly lower wavelength than the incident light, due to the molecules of the sample either absorbing energy from the photons (Stokes shift) or giving up energy to the photons (anti-Stokes shift). These shifts in frequency correspond to the energy associated with transitions between rotational and vibrational states of the molecules in the sample, and by plotting a graph of the intensity of the shifted light versus frequency, information can be gleaned on the contents of the sample.[10-11]

Atomic bonds have vibrational states associated with them, and in RS these vibrational states will be represented by the energy shift in the scattered photons. Because specific atomic bonds have unique vibrational states, accurate and precise characterization of the molecular structure of a sample is possible using RS.[12]

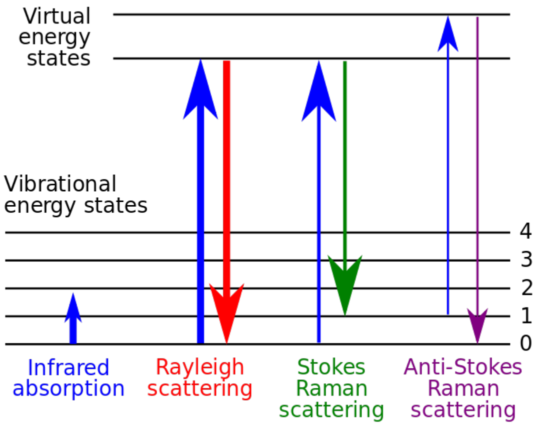


Figure 1: A representation of the difference between Rayleigh scattering and both Stokes Raman and Anti-Stokes Raman scattering[12]

RR is a specific form of RS where the excitation light is of a frequency very near the frequency of an electronic transition of the sample under analysis. RR is much more sensitive than non-resonance spectroscopy, and is much more useful for analysis of molecules of a greater size such as biomolecules.[13]

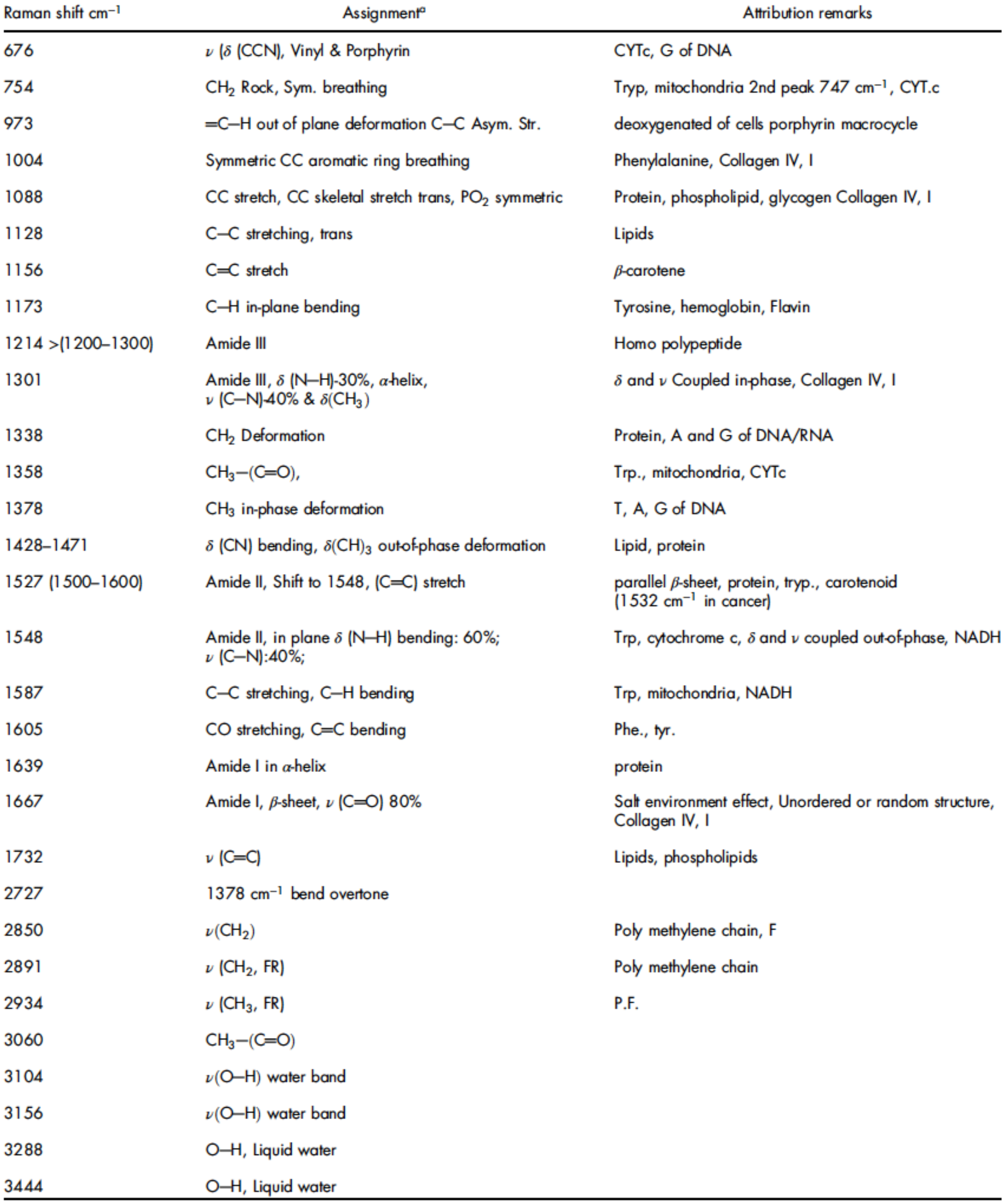
The ability of RS to uniquely identify molecules within a sample and the sensitivity of the RR technique is what makes RR a good choice as an OB technique for BCC diagnosis.

3 Methods of Analysis

A number of methods of analyzing the data will be implemented in order to find the most efficient and accurate way of classifying the data possible, but first it is essential to understand the traditional method of RR peak analysis that is currently commonly used for identifying representation of biological structures within RR data.

3.1 Raman Peak Analysis

Peaks at specific wavenumbers within RR spectra correlate directly to specific biological structures and biomolecules within biological samples. By performing RR on different biomolecules in isolation, the resulting peaks at specific wavenumbers that are characteristic of the bonds within the biomolecule can be identified and recorded, as shown in Figure 2.[5] Analysis of individual peaks or ratios between different peak values have been commonly used in Raman spectroscopic studies for cancer diagnosis.[5] Since Raman spectra contains rich and complex information due to the presence of so many biomolecules, the Raman peaks or peak ratios that can serve as good criteria for cancer diagnosis is still under investigation. Since raw Raman spectral data usually carries a strong background signal, baseline removal is commonly performed to preprocess the data. The peak intensities of Raman spectra can change quite significantly with variable experimental conditions and different baseline removal methods. In addition, individual peaks may be contributed to by multiple biomolecules due to spectral overlapping; therefore, analysis using individual Raman peaks may lead to false positives and false negatives. In contrast, machine learning methods such as PCA, NMF and ANN that consider the entire profile for each Raman spectrum may allow for a more robust analysis of the data. In my thesis work, I have used these different machine learning methods, particularly ANN methods, to analyze RR spectral data, and compared the efficacy of these methods in distinguishing normal and cancerous skin cancer tissue samples.



Raman shift cm^{-1}	Assignment[a]	Attribution remarks
676	ν (δ (CCN), Vinyl & Porphyrin	CYTc, G of DNA
754	CH_2 Rock, Sym. breathing	Tryp, mitochondria 2nd peak 747 cm^{-1}, CYT.c
973	=C–H out of plane deformation C–C Asym. Str.	deoxygenated of cells porphyrin macrocycle
1004	Symmetric CC aromatic ring breathing	Phenylalanine, Collagen IV, I
1088	CC stretch, CC skeletal stretch trans, PO_2 symmetric	Protein, phospholipid, glycogen Collagen IV, I
1128	C–C stretching, trans	Lipids
1156	C=C stretch	β-carotene
1173	C–H in-plane bending	Tyrosine, hemoglobin, Flavin
1214 >(1200–1300)	Amide III	Homo polypeptide
1301	Amide III, δ (N–H)-30%, α-helix, ν (C–N)-40% & δ(CH_3)	δ and ν Coupled in-phase, Collagen IV, I
1338	CH_2 Deformation	Protein, A and G of DNA/RNA
1358	CH_3–(C=O),	Trp., mitochondria, CYTc
1378	CH_3 in-phase deformation	T, A, G of DNA
1428–1471	δ (CN) bending, δ$(CH)_3$ out-of-phase deformation	Lipid, protein
1527 (1500–1600)	Amide II, Shift to 1548, (C=C) stretch	parallel β-sheet, protein, tryp., carotenoid (1532 cm^{-1} in cancer)
1548	Amide II, in plane δ (N–H) bending: 60%; ν (C–N):40%;	Trp, cytochrome c, δ and ν coupled out-of-phase, NADH
1587	C–C stretching, C–H bending	Trp, mitochondria, NADH
1605	CO stretching, C=C bending	Phe., tyr.
1639	Amide I in α-helix	protein
1667	Amide I, β-sheet, ν (C=O) 80%	Salt environment effect, Unordered or random structure, Collagen IV, I
1732	ν (C=C)	Lipids, phospholipids
2727	1378 cm^{-1} bend overtone	
2850	ν(CH_2)	Poly methylene chain, F
2891	ν (CH_2, FR)	Poly methylene chain
2934	ν (CH_3, FR)	P.F.
3060	CH_3–(C=O)	
3104	ν(O–H) water band	
3156	ν(O–H) water band	
3288	O–H, Liquid water	
3444	O–H, Liquid water	

Figure 2: Table of known associations between wavenumbers (Raman shift) and chemical bonds/biomolecules.[5]

3.2 Principal Component Analysis

PCA is a technique that uses linear algebra principles in order to find what are known as the Principal Components (PCs) of a given data set. Often in experimentation, much more data is gathered than is necessary or informative for answering the problem at hand. An experimenter does not know what information is useful for analysis at the time of experimentation, so it is much more practical to measure as many components as possible and extract the important information from the data at a later time. The goal of PCA is to find the PCs that characterize the data most strongly, while avoiding redundancy. This is framed as a "change of basis", as known in linear algebra. PCA extracts PCs from data by finding the dimensions with greatest variance, calculating the covariance between these dimensions to remove redundant data, and finally solving for the eigenvalues of the covariance matrix to extract the PCs.[4]

While PCA efficiently finds the components of greatest variance, these components may or may not be the most characteristic components of the experimental data set. PCA will extract components that are linear mixtures of the experimental data set and even contain negative values, which may be difficult to correlate with specific biological features within the sample. Comparison with NMF and AANN for dimensional reduction is crucial for these reasons.

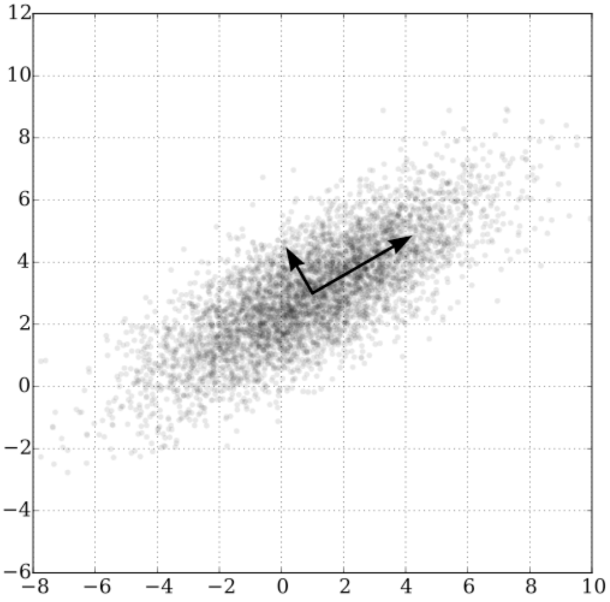


Figure 3: PCA performed on a multivariate Gaussian distribution.

PCA makes a few assumptions about the data in order to adhere to the laws of linear algebra, and to be able to reduce dimensions by using the concept of change of basis. Linearity is assumed, in order to be able to use the concept of change of basis to find linearly independent basis vectors (dimensions) within the data. PCA also assumes that dimensions with a larger variance are more important dimensions, while dimensions with low variance are more likely to be noise.

In this study, after application of PCA, SVM will be used to classify the data based on the PCs.

3.3 Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) is a group of multivariate analysis algorithms that factorize a matrix *X* into *W* and *H*, provided *W* and *H* are non-negative.[5]

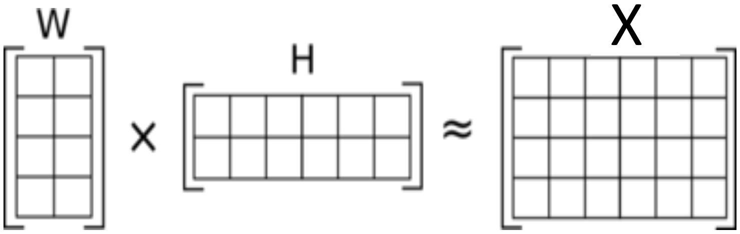


Figure 4: A pictorial representation of NMF performed on a matrix X.

NMF differs from PCA in that it does not imply any relationship between the extracted components; it only imposes a non-negativity constraint on both *W* and *H*. There are various algorithms that can solve NMF, such as the multiplicative update method and the alternating least squares method.

NMF has been successfully used for facial recognition to extract characteristic facial features from images of faces. The NMF algorithm is better at decomposing facial images into sets of characteristic features when compared with PCA, as PCA finds features that are linear mixtures of all of the facial features with the greatest variance, while NMF finds components that contain “parts” of faces,[5] as shown in Figure 5. In theory, NMF could extract features from the Raman spectral data that are characteristic of specific biological structures within the samples. Since it is already known that cancerous and non-cancerous skin tissue samples differ in their concentrations of certain identifiable biomolecules[27], this method could allow for accurate dimensional reduction before classification methods are tried.

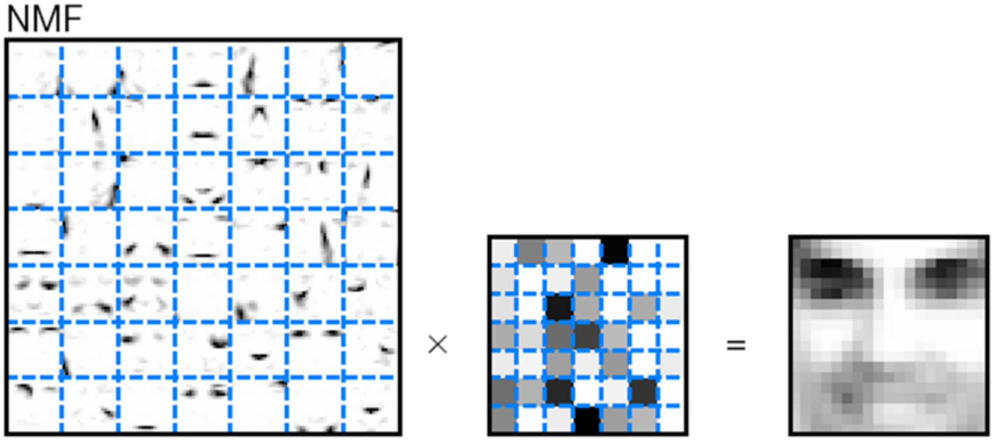


Figure 5: NMF performed on an image of a face for extraction of characteristic facial features. The matrix on the left displays characteristic facial features, and the matrix in the middle is representative of the abundance associated with these features.[5]

In this study, NMF is used to decompose RR spectral data, and then the scores of the NMF components are used to classify the data with SVM as was done with the PCA method.

3.4 Support Vector Machines

A Support Vector Machine (SVM) is what is known as a discriminative classifier, which uses an optimized hyperplane to categorize examples.[14] Using the case of 2-dimensional linearly separable points for simplicity, there are typically a number of lines which will separate the two classes of data, although they are not all optimal. The SVM algorithm finds the hyperplane that will give the greatest minimum distance between the training examples, or maximizes the margin of the data.

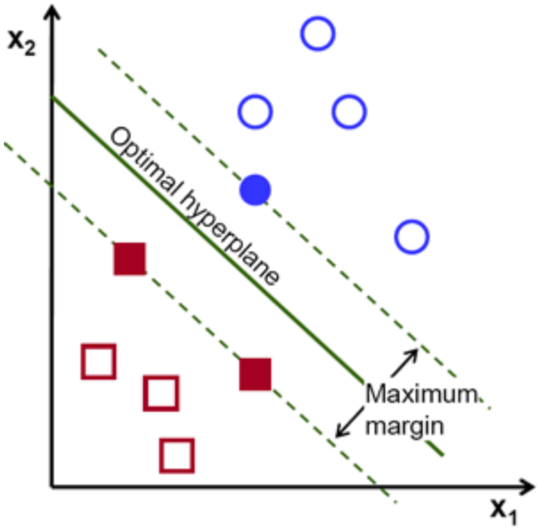


Figure 6: Representation of an optimal hyperplane linearly separating 2-dimensional data.[14]

The optimization is dependent on the data points that lie most closely to the decision boundary, known as the support vectors. The support vectors are the data points that would change the position of the decision boundary if they were to be removed.[14] An optimization algorithm is then used to determine the hyperplane that separates the two classes with a maximum margin between the support vectors of the different classes of data.

SVM can be expanded to work for non-linearly separable data sets as well, using what is known as the kernel function. This function is a transformation function that maps the data to a higher dimensional space where it becomes linearly separable.[15]

SVM is what is known as a "supervised" machine learning algorithm. Supervised learning uses a labeled data set in order to properly optimize, or "train", the machine learning algorithm. A data set consisting of known input and desired output pairs is analyzed by the algorithm, which then can make inferences about patterns in the data.[15]

In this study, SVM is used to classify the RR spectral data based on the scores of the components extracted from the RR data by different methods including PCA, NMF and ANN.

3.5 Overfitting

One potential error to consider with SVM or any machine learning algorithm is the possibility of overfitting the data set. Overfitting is when a statistical model becomes overly complex, and fits the training data set too specifically. A training data set that has been overfit will not be translatable to any test data, as the model is describing the random error of the training set as opposed to the true relationship of the data.[16]

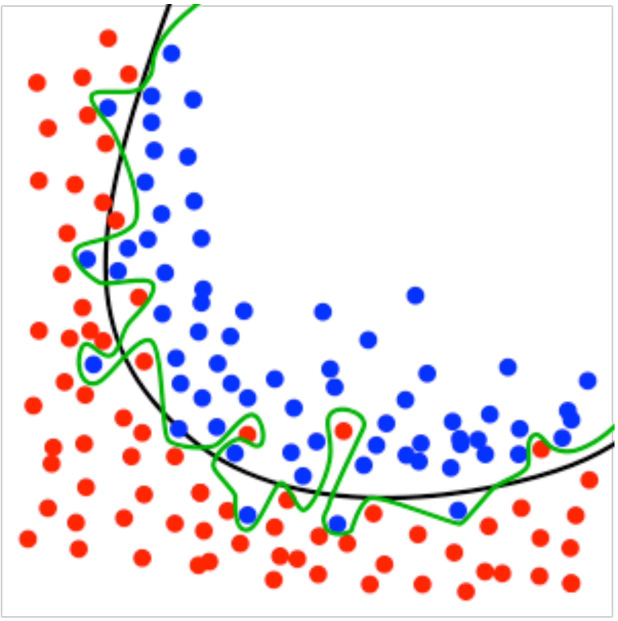


Figure 7: An example of an overfitted model (green), and a regularized model (black). While the green line perfectly classifies this data set, it does not characterize any deeper relationship amongst the data, and will not accurately classify a new test data set

Methods used to avoid overfitting include cross-validation and using regularization terms, which will be described in more depth later on.

3.6 Leave One Out Cross Validation

In order to know that a discriminative classifier such as SVM is properly modeling data and is not overfitted, a method of validation must be used. Leave One Our Cross Validation (LOOCV) takes the training data set, and uses all but one value to train the classifier, and one to test. The algorithm then cyclically uses every single value as a test value, while using the rest of the data set for training. This method ensures that the classifier is not overfitting the training data set.

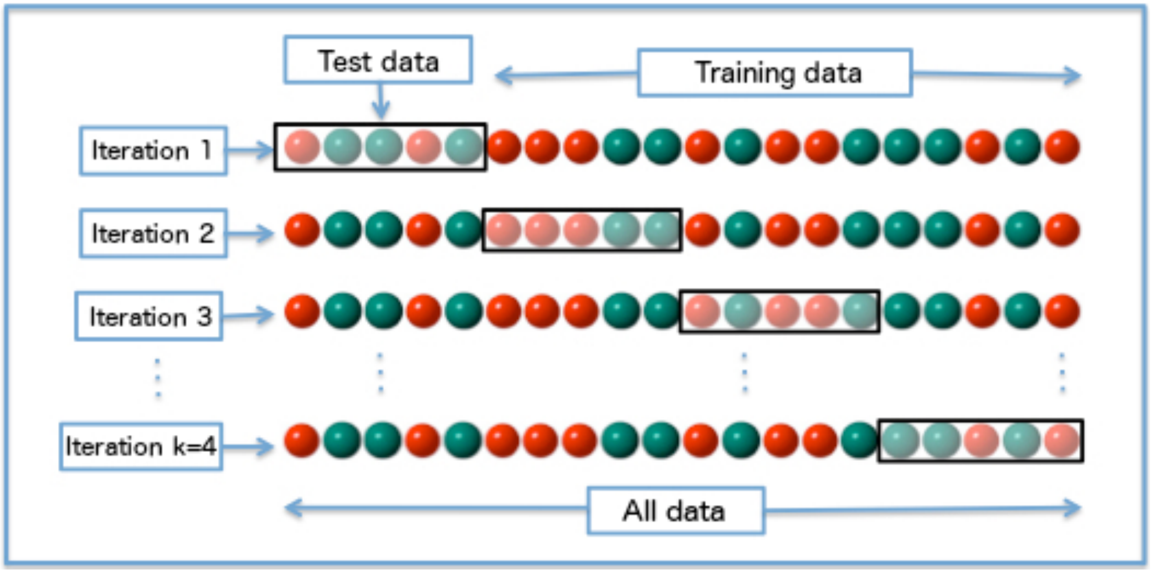


Figure 8: An example of leave five out cross validation. The algorithm uses five values to test, and the rest for training. It then cycles through using all values as a part of the test set, while using the remaining values for training.

3.7 Receiver Operating Characteristic Curve

A commonly used metric to determine how well a classifying algorithm has fit a specific data set is what is known as the Receiver Operating Characteristic Curve (ROC Curve). The ROC curve plots the true positive rate (known as the sensitivity) against the false positive rate (1-sensitivity).

Sensitivity and Specificity

In order to understand how a ROC curve can describe the fit of a model, the terms “sensitivity” and “specificity” must be defined. When performing supervised learning on a discriminative classifier, sensitivity and specificity describe the rate of “true positives” (sensitivity) and the rate of “true negatives” (specificity) identified. A true positive is a sample (data point) that is known to be positive, and is correctly identified as such. A true negative is a sample (data point) that is known to be negative, that is correctly identified as such. False positives and false negatives are the inverse; false positives are data points that are known to be positive and were improperly identified as negative, while false negatives are data points known to be negative that were improperly identified as positive.

“Positive” and “negative” simply refer to the two different possible classifications when discussing a binary classifier; they are arbitrary values. For example, positive corresponds to cancerous and negative corresponds to normal when in the context of the experimental data of this study.

The area under the ROC curve (AUROC) displays how well a particular model has been fit to a given data set. The AUROC will vary from 0.5 to 1. An AUROC of 0.5 is a completely random classification, while an AUROC of 1 is a perfectly performing classifier.

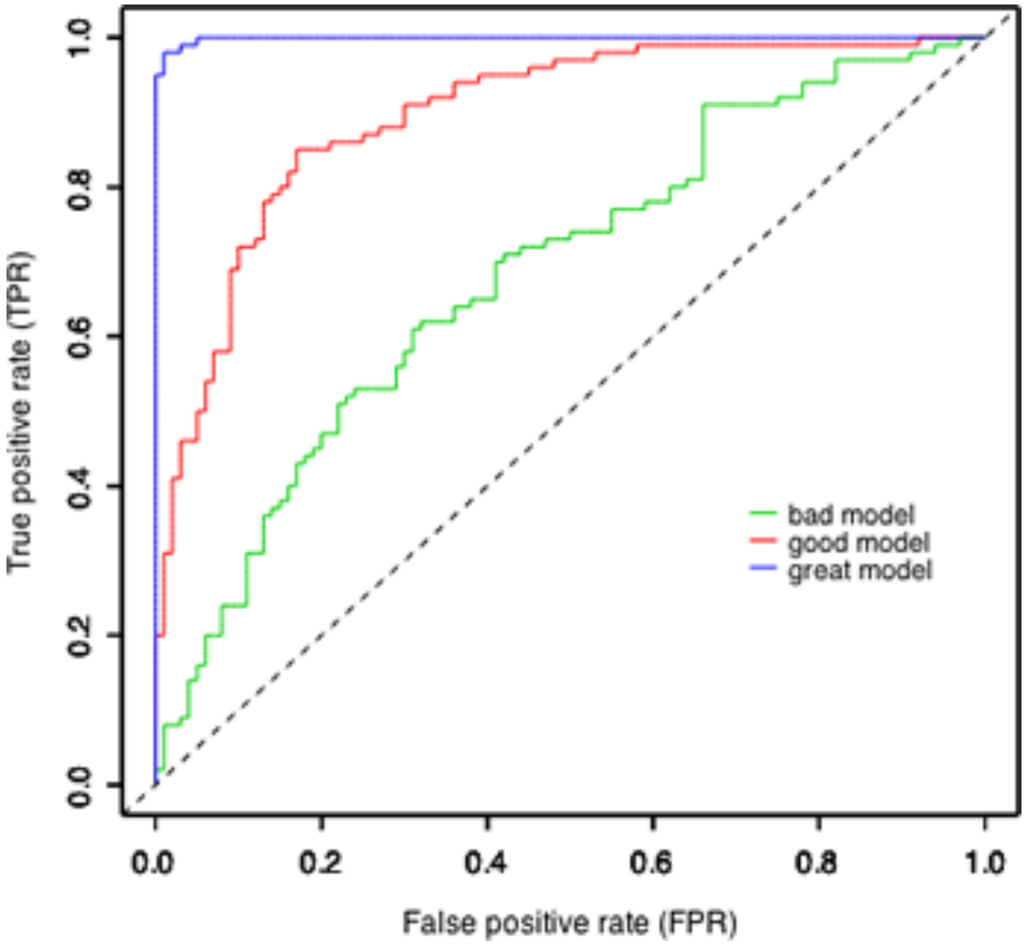


Figure 9: An example plot of several ROC curves of varying model fit.[17]

3.8 Artificial Neural Networks

An Artificial Neural Network (ANN) is a type of machine learning algorithm that is loosely modeled after the neuronal structure of the mammalian cerebral cortex.[18] ANNs are made up of a system of interconnected "neurons", each which contains what is known as an "activation function". Connections between different neurons within the network are weighted, and by adjusting the weights of these connections the output of the overall network can be varied. ANNs consist of many layers of neurons that can be interconnected in many different fashions in order for different types of data processing. The input layer, where the data is fed into the network, is connected to one or more "hidden layers", which are in turn connected to an output layer where the results of data processing are obtained.

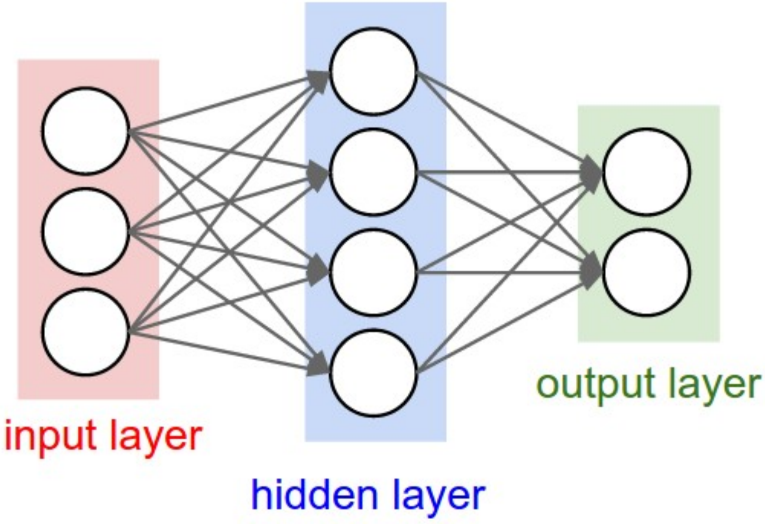


Figure 10: An example of the architecture of a 3-layer neural network, consisting of an input layer, a single hidden layer, and an output layer.

3.9 The Neuron

Artificial neurons found in ANNs are loosely modeled after the biological neuron found in the mammalian brain. Biological neurons operate by receiving signals from other neurons through connections known as synapses. When a combination of these signals received by a neuron are in excess of an "activation function" that is known by the neuron, the neuron will fire or activate, sending along a signal to the other neurons connected to it.[19] Artificial neurons operate on the exact same premise as biological neurons. A neuron will activate if the combined inputs feeding into the neuron are great enough to surpass the threshold of the activation function that neuron possesses.

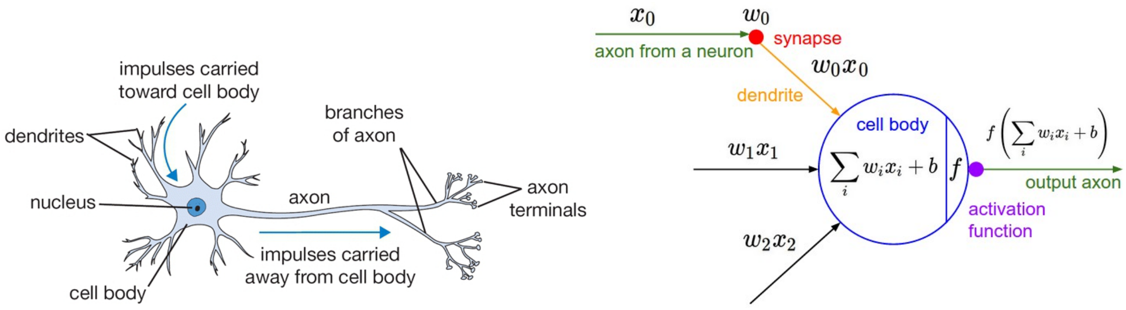


Figure 11: A drawing of a biological neuron (left) and the mathematical representation of an artificial neuron (right).[19]

Connections between neurons are assigned weights, in order to increase or decrease the strength of that signal within the network. An ANN is trained by altering these weights in an iterative fashion in order to achieve the output desired.

The inputs into a given neuron within a network are multiplied by a weight and then summed, as a linear combination:

$$\sum_i w_i x_i + b \tag{1}$$

Where w_i is the ith connection weight, x_i is the ith input value, and b is a bias value.

This linear combination is then fed through an activation function, which determines whether or not that particular neuron will "activate" and continue to pass information on through the network, like so:

$$f\left(\sum_i w_i x_i + b\right) \tag{2}$$

The Activation Function

The activation function is what determines if a specific neuron in an ANN will activate or not. Non-linear activation functions are used in order to allow the network to model a response variable that varies non-linearly (Using a linear activation function would only allow for linear approximations, as summing many linear layers would just result in another linear function). When the inputs into a given neuron sum up to a great enough value to surpass the activation function's threshold defined at that neuron, the neuron will pass on the numerical data on to the neurons that it is connected to ahead of it.

For example, consider the logistic sigmoid function:

$$f(x) = \frac{1}{1+e^{-x}} \tag{3}$$

If this is the chosen activation function within a network, the value of the linear combination of weights multiplied by inputs for each neuron will in essence be "squashed" through this function. A threshold value will determine if this neuron activates or not. Typically, values greater than or equal to 0.5 will cause the neuron to activate, and values less than 0.5 will cause the neuron to be inactive.

There are a number of different activation functions that are used for neurons within ANNs. Some commonly used activation functions are shown in Figure 12. There are advantages and disadvantages to the usage of different activation functions. In order for a gradient descent algorithm to be used for optimization of an ANN during training, the activation function must be continuously differentiable. During training, different activation functions will have different effects on the optimization process. For example, usage of the sigmoid function can lead to what is known as "saturation" during the optimization process. When a sigmoidal neuron's activation occurs near 0 or 1, the gradient of the function is very close to zero. Because of how the process of backpropagation works, this effectively stops this neuron from contributing to the optimization process. This will be discussed further in the section on backpropagation.

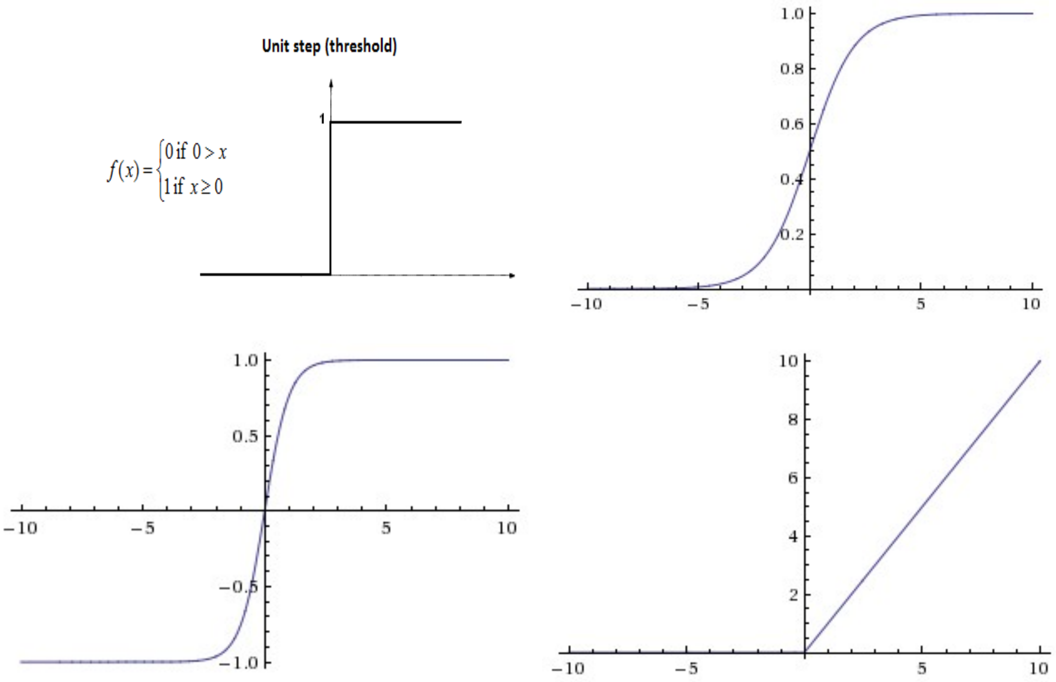


Figure 12: Some commonly used activation functions. Step function (upper left), sigmoid function (upper right), hyperbolic tangent function (lower left), rectified linear unit (lower right).

Activation functions that are non-differentiable, such as the step function above, are only used on an output layer of an ANN for classification purposes.

3.10 Training an ANN

There are a few different methods for training ANNs, although the most popular and efficient is using a process known as backpropagation in combination with gradient descent.

Backpropagation

Backpropagation is a widely used method to train ANNs. Recall that every connection between neurons within an ANN has an associated weight, in order to adjust that connection's significance in the network output. The goal of the process of backpropagation is to determine the weights throughout the network that allow for the smallest possible value of the "loss function", a function that calculates the difference between the expected values of a training set and the values of a training set calculated by the network. If we consider a network with *n* input and *m* output units and a training set {(**x**$_1$, **t**$_1$),…,(**x**$_p$,**t**$_p$)} consisting of *p* ordered pairs of *n* and *m* dimensional vectors, the loss function is defined as

$$E = \frac{1}{2}\sum_{i=1}^{p} ||o_i - t_i||^2 \tag{4}$$

Where o_i denotes the *i*-th calculated output and t_i denotes the *i*-th expected output.[19] The backpropagation algorithm works by choosing randomly initialized weights throughout the network, feeding the training data forward through the network, calculating the loss function and its gradient, and calculating the gradient with respect to each weight in the network backwards through the network. These gradients are used to then update the existing weights with new slightly more optimal weights, and this process is repeated iteratively until the gradient of the error function is minimized.

For a network with l neuron connections with associated weights, the gradient of the error function is defined as

$$\nabla E = (\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \dots, \frac{\partial E}{\partial w_l}) \tag{5}$$

Weights are updated incrementally, by multiplying the partial derivative of each weight by the negative of the "learning rate parameter". The learning rate parameter determines how large of a "step" the gradient descent algorithm takes when adjusting the weights of the network through every iteration. This parameter is chosen manually, and different factors determine what an optimal learning rate parameter is for a given situation. When the weights of the network are randomly initialized, a good choice would be to have a relatively larger step size in order to speed up convergence of the backpropagation algorithm. As the algorithm progresses, however, this large step size will slow progress as the gradient of the error function decreases. In practice, it is typical that the learning rate parameter is decayed throughout the optimization process, in order to have a quicker time to convergence throughout the whole process.

3.11 Autoencoder Artificial Neural Networks

An Autoencoder Artificial Neural Network (AANN) is an ANN with a specific structure, which is particularly good for dimensionality reduction. In an AANN, there exists at least one hidden layer with a smaller dimension size than the input and output layers. Most AANNs also possess input and output layers of the same dimensionality.

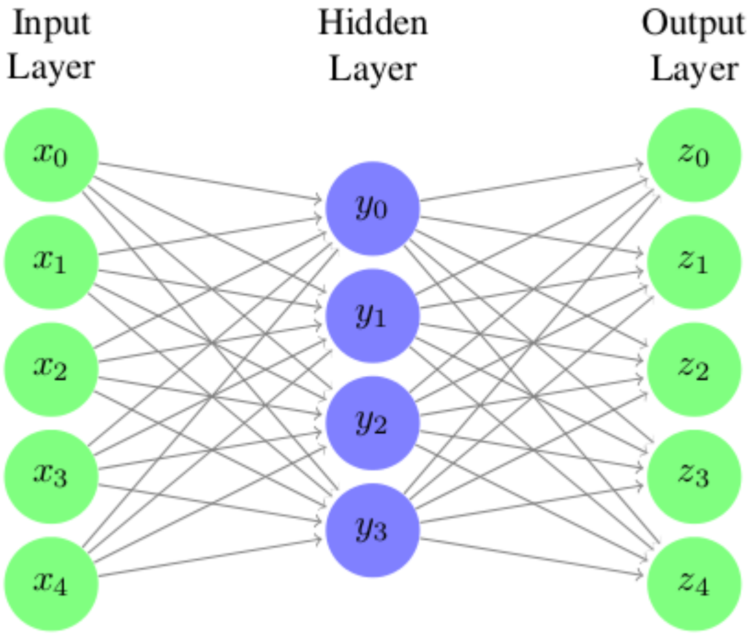


Figure 13: A representation of a simple 3-layer autoencoder artificial neural network. Notice the hidden layer has a lower dimensionality than the input and output layers.[20]

This structure forces the network to represent the input data within a lower dimensional space, allowing for effective dimensional reduction.

AANNs differ slightly in training to traditional ANNs. AANNs are capable of what is known as “unsupervised learning”. Unsupervised learning is when a network can be trained without the presence of training data at all. An AANN is still trained using the backpropagation algorithm like traditional ANNs, but with one major difference. Instead of seeking to minimize a loss function, the algorithm seeks to minimize the difference between the input layer values and the output layer values. An AANN network is optimized when the output layer resembles the input layer values as closely as possible.

By effectively representing the input data at the output layer unchanged, the hidden layer of a lower dimensionality (known as the bottleneck layer) is accurately representing the data in a lower dimensional space.

3.12 Training and Hyper Parameters within an AANN

3.12.1 Training an AANN

As previously stated, an AANN trains by minimizing the difference between the input data values and the reconstructed output data values. AANNs use what is known as the adjusted mean squared error function to determine when this value is minimized. The equation is as follows:

$$E = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}(x_{kn} - \hat{x}_{kn})^2 + \lambda * \Omega_{weights} + \beta * \Omega_{sparsity} \qquad (6)$$

Where x is the input data, $\hat{x}$ is the reconstructed output data, $\Omega_{weights}$ is the L2 regularization term, λ is the coefficient for the L2 regularization term, $\Omega_{sparsity}$ is the sparsity regularization term, and β is the coefficient for the sparsity regularization term.[21] The meaning and functions of these additional hyperparameters will be discussed in detail within this section.

3.12.2 Tolerance

The tolerance, or threshold, is a hyperparameter that determines when the network is sufficiently trained. In the case of an AANN, the tolerance can be chosen in two different ways. The first way is to establish a maximum number of iterations the network will perform in training. When the training process reaches this predetermined maximum number of iterations, training will cease.

This method is typically not the best way to establish a training tolerance. Different data sets and network configurations will require different numbers of iterations to reach.

A more effective method of tolerance implementation is defining a threshold value for the gradient descent algorithm. This way, the network will continue to train until it reaches the minimum gradient specified. Depending on the order of the input values this number will differ; for the purposes of our research, a threshold minimum gradient of 10^{-7} will be used.

3.12.3 L2 Weight Regularization

L2 weight regularization is one of two defined types of regularizing terms for AANNs. Regularizing terms are implemented to help prevent overfitting the network to the data. A common characteristic of overfitted networks is that the weights tend to become quite large. The L1 and L2 weight regularization terms prevent this by penalizing large weight values. L1 regularization penalizes the sum of the absolute values of the weights, while L2 regularization penalizes the sum of the squared values of the weights.[22] L2 regularization is mathematically defined as:

$$\Omega_{weights} = \frac{1}{2}\sum_{l}^{L}\sum_{j}^{n}\sum_{i}^{k}\left(w_{ji}^{(l)}\right)^{2} \quad (7)$$

Where $w^{(l)}$ is the weight in the lth hidden layer, L is the number of hidden layers, n is the number of examples, and k is the number of variables in the training data.

Adjusting the coefficient of the L2 regularizing term will influence how heavily penalized larger weights will be. During experimentation with our data, the coefficient for the L2 regularizing term was adjusted in order to find the most optimal result possible, while still maintaining sufficient regularization.

3.12.4 Sparsity Regularization

While an AANN uses the hidden layer in order to force a compressed representation of the input data, this is not the only constraint we can place in order to extract useful information.

Imposing what is known as a sparsity constraint can allow for useful extraction of information, even if the number of hidden units is large.[23]

Assuming the use of a logistic sigmoid function as the activation function, a neuron will be "active" if its output value is close to 1, and will be "inactive" if its output value is close to 0. After a forward pass through the network, an average output activation can be calculated on each neuron within the hidden layer. If a neuron has a low average output activation value, it is "active" in response to only a small number of the training examples. Using the sparsity parameter, we can constrain what the desired average output activation of the neurons within the hidden layer is.

If the average output activation is defined as:

$$\hat{\rho}_i = \frac{1}{n}\sum_{j=1}^{n} z_i^{(1)}\left(x_j\right) = \frac{1}{n}\sum_{j=1}^{n} h\left(w_i^{(1)T} x_j + b_i^{(1)}\right) \tag{8}$$

where n is the total number of training examples, x_j is the jth training example, $w_i^{(1)T}$ is the ith row if the weight matrix $W^{(1)}$, and $b_i^{(1)}$ is the ith entry of the bias vector $b^{(1)}$, the sparsity regularization term is defined as:

$$\Omega_{sparsity} = \sum_{i=1}^{D^{(1)}} KL(\rho||\hat{\rho}_i) = \sum_{i=1}^{D^{(1)}} \rho \log\left(\frac{\rho}{\hat{\rho}_i}\right) + (1+\rho)\log\left(\frac{1-\rho}{1-\hat{\rho}_i}\right) \tag{9}$$

Where ρ is the desired average output activation value, and $\hat{\rho}_i$ is the calculated average output activation value.[23] This sparsity regularization term is what is known as the Kullback-Leibler divergence,[23] and works in a similar way to L2 weight regularization; by penalizing large discrepancies between ρ and $\hat{\rho}_i$.

4 Data and Results

4.1 Experimental Data

In this study, fifty-five total spectra were collected from thirty-seven slices of tissue specimens, both normal and afflicted with BCC. Collection was done using a WITec alpha300R Raman microscope and imaging system, and an incident light wavelength of 532nm was used.

The human skin tissues from both the normal and BCC specimens were obtained from the National Disease Research Interchange (NDRI) in Philadelphia, PA. The experimental procedures were approved by the Institutional Review Board (IRB) office through the City College of CUNY, where the sample preparation and analysis was prepared. The skin specimen was kept under snap-frozen conditions with no chemical treatment, and it was thawed to room temperature for the spectroscopic studies. The sections of tissue were mounted of uncoated glass microscope slides using Leica CM1080 Cryostats at -20^{o}C.

Typical Raman spectra of both the normal and BCC afflicted tissue after baseline removal are shown below in Fig. 7:

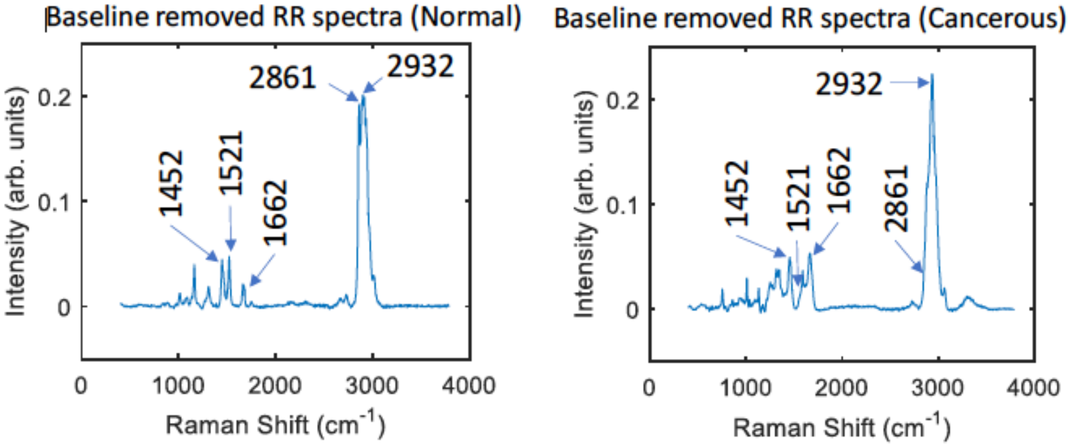


Figure 14: Typical resonance Raman spectra of normal (left) and cancerous (right) tissue, after baseline removal.

4.2 PCA and SVM

The combination of PCA as a method of dimensional reduction and SVM for classification yielded good results. With the dimension of the input data reduced to a dimension of 10, the first and third components were found to have the best classification results. Classifying the first and third components resulted in a sensitivity of 97.7%, a specificity of 75.0%, and an accuracy of 92.7%. A metric used for measuring the quality of these results is what is known as the area under the Receiving Operator Characteristic curve, or AUROC. These results had an AUROC of 0.96, with the maximum possible being 1.

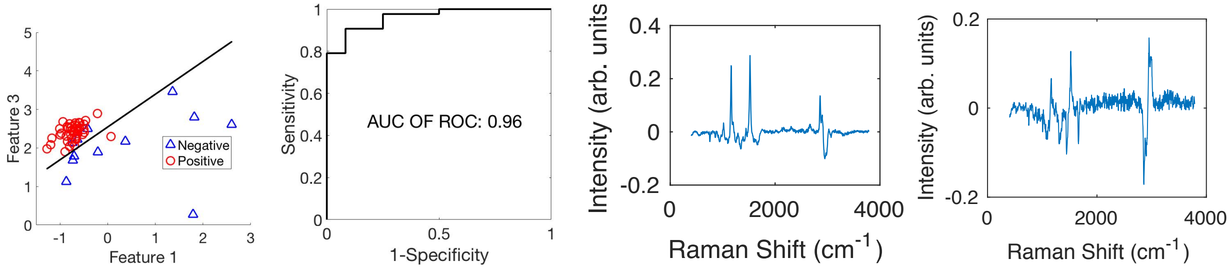


Figure 15: SVM classification based on the first and third PCA components. The basis graph of PC 1 is second from the right, and the basis graph of PC 3 is far right.[24]

4.3 NMF and SVM

Combining NMF and SVM obtained a higher degree of accuracy than PCA and SVM. NMF was used to reduce the dimensionality of the data set to 10, and the second and fifth components were found to have the best results when classified. Classification of the second and fifth components resulted in a sensitivity of 100.0%, a specificity of 91.7%, and an accuracy of 98.2%.

Notice how the NMF basis spectrum graphs do not have any negative portion; this is due to the non negativity constraint of NMF. In theory, this makes the NMF basis spectrum more easily correlated to specific biological structures within the samples than the PCA method of dimensional reduction.

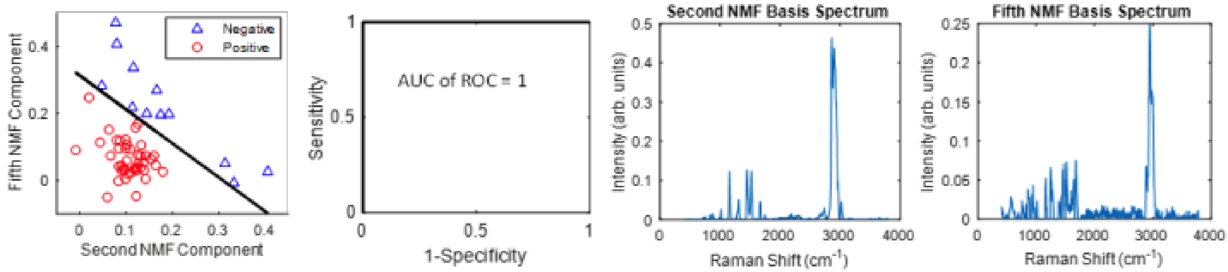


Figure 16: SVM classification based on the second and fifth NMF components. The basis graph of NMF component 2 is second from the right, and the basis graph of NMF component 5 is far right.

4.4 AANN and SVM

The combination of AANN and SVM yielded moderate results. The AANN was used to reduce the dimensionality to 10, and the first and fourth components were found to obtain the greatest results when classified. Classification of components one and four resulted in a sensitivity of 97.7%, a specificity of 76.8%, and an accuracy of 90.3%.

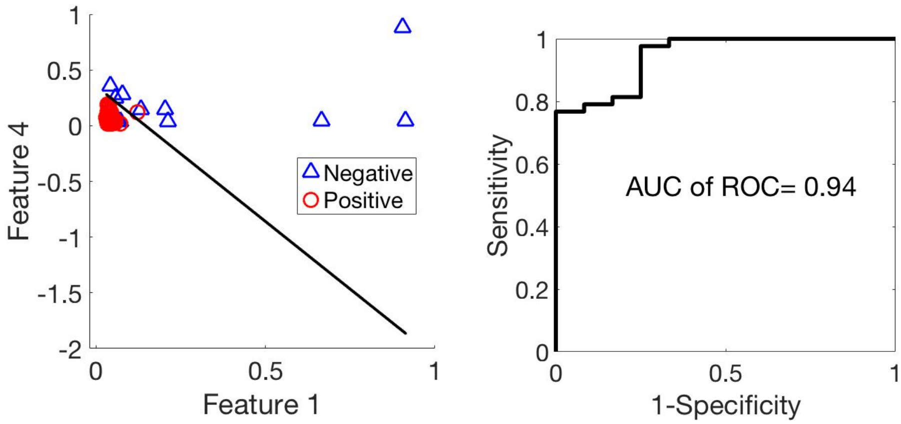


Figure 17: SVM classification based on the first and fourth AANN components.

Notice the dense clustering of the positive values within AANN classification compared to the two other methods. This may be a result of the normalization methods used when preprocessing the data. Further research is needed to find a more optimal form of normalization for data compressed using AANN.

Basis graphs were unable to be generated for the first and fourth AANN components, due to time constraints of the project. Going forward, these basis graphs will be generated and will be available for comparison with basis graphs resulting from PCA and NMF.

5 Discussion and Results

Using the AANN for dimensional reduction produced very similar results to using PCA. There was a 0.0% difference in sensitivity. AANN had a 1.8% advantage in specificity than PCA. PCA had 2.4% better accuracy than AANN. The AUC of the ROC was greater for PCA by a value of 0.04.

AANN performed very similarly to PCA in its ability to reduce the dimension of the experimental data and classify it. One major difference to note is the spread throughout the classification graphs. AANN clustered the data points very closely together. Going forward, further research must be completed in finding a more suitable form of normalization for raw data before using AANN to reduce the dimension of the data set. Improving this parameter may increase the accuracy of the classification performed by SVM after dimensional reduction through AANN.

Out of the three dimensional reduction methods used, NMF produced the best results. Compared to AANN, Performed 2.3% better in sensitivity, 14.9% better in specificity, and 7.9% better in accuracy.

There are a number of unknowns to consider when discussing the results of the research performed. As previously stated, directly correlating PCA components to biological structures within the samples will take more research and is a difficult task. Since the PCA components are linear mixtures of the RR spectra with the greatest variance, each PC may contain information about multiple biological structures. Isolating specific structures from raw RR spectra using PCA for dimensional reduction is a task that will require more study in the future.

NMF was used for dimensional reduction in order to theoretically overcome the downsides of PCA. When used for facial recognition, NMF was shown to be able to extract characteristic features from facial images successfully. Applying NMF to the RR spectral data is thought to theoretically produce similar characteristic feature extraction. The current issue lies in validating this theory; while NMF extracts characteristic features from facial images, it cannot be assumed that the same characteristic feature extraction will occur when applied to RR spectra from tissue samples. More future work will need to be completed on validating this theory before this method can be used with a high degree of confidence.

Using the AANN for feature extraction worked with a high degree of accuracy, but similar to the other methods of dimensional reduction, more research is needed before the components extracted by this method can be correlated to structures within the raw RR data. Interpreting the meaning of the reduced dimensional space is more difficult and ambiguous with the AANN, due to the non-linear process by which the dimensional reduction is performed.

More information would be able to be gathered if there was a larger sample size to work with, which will be dependent on the ability of Dr. Wu's colleagues to obtain more normal and BCC skin tissue samples.

Going forward, a more thorough understanding of how the reduced dimensional space of all three of the dimensional reduction methods used relates to the biological structures of the biomolecules within the samples must be found.

6 Conclusion

The research conducted throughout this study provided promising results for the improvement of traditional diagnosis procedures of BCC. All three methods investigated performed with a higher degree of accuracy than traditional tissue biopsy and H&E histopathology by a diagnostician, which is the current gold standard for BCC diagnosis. These methods show the potential for a less invasive, faster, and more robust diagnostic process.

More research must be done into the correlation between the classified features from each dimensional reduction method and their relationship to the biological structures found within the samples, but the results obtained from this study show with confidence that machine learning algorithms such as PCA, NMF, and AANNs when combined with SVM for classification can successfully discriminate between RR spectra from cancerous and normal skin tissue samples.

References

1. Skin Cancer: Basal and Squamous Cell. (n.d.). Retrieved October 24, 2016, from http://www.cancer.org/cancer/skincancer-basalandsquamouscell/detailedguide/index
2. Optical Biopsy O - MIT. (n.d.). Retrieved October 24, 2016, from http://web.mit.edu/hst.035/ReadingSupplement/05_03_Bigio/Bigio-Mourant_Encyclo_chap.pdf
3. Liu, C., Boydston-White, S., Wu, B., Sriramoju, V., Zhang, C., Beckman, H., . . . Alfano, R. R. (n.d.). Depth Dependence of Resonance Raman Spectra of Basal Cell Carcinoma and Normal Human Skin Tissues (Publication).A Tutorial on PCA paper
4. Shlens, J. (n.d.). A Tutorial on Principal Component Analysis (Rep.).
5. Liu CH, Sriramoju V, Boydston-White S, Wu B, Zhang C, Pei Z, Sordillo L, Beckman H, Alfano RR. Proc. SPIE 2017;10060:100601B
6. Yuan, Hui, Cynthia F. Van Der Wiele, and Siamak Khorram. "An Automated Artificial Neural Network System for Land Use/Land Cover Classification from Landsat TM Imagery." *MDPI* (2009): 243-65. Web.
7. Tayel, Mazhar B., and Mohamed E. El-Bouridy. "ECG Images Classification Using Artificial Neural Network Based on Several Feature Extraction Methods." *2008 International Conference on Computer Engineering & Systems* (2008): n. pag. Web.
8. InPhotonics: What is Raman spectroscopy? (n.d.). Retrieved October 18, 2016, from http://www.inphotonics.com/raman.htm
9. THE RAMAN EFFECT - physics.rutgers.edu. (n.d.). Retrieved October 16, 2016, from https://www.physics.rutgers.edu/grad/506/raman/raman.pdf
10. Raman Effect. (n.d.). Retrieved October 11, 2016, from https://www.britannica.com/science/Raman-effect
11. Raman Tutorial - Raman Scattering - HORIBA. (n.d.). Retrieved October 12, 2016, from http://www.horiba.com/scientific/products/raman-spectroscopy/raman-academy/raman-tutorial/raman-scattering/
12. Resonant vs. Nonresonant Raman Spectroscopy. (n.d.). Retrieved October 12, 2016, from http://chem.libretexts.org/Core/Physical_and_Theoretical_Chemistry/Spectroscopy/V

ibrational_Spectroscopy/Raman_Spectroscopy/Resonant_vs._Nonresonant_Raman_Spectroscopy

13. Introduction to Support Vector Machines — OpenCV 2.4.13.1 ... (n.d.). Retrieved October 16, 2016, from http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html
14. Enhancement of SVM based MRI Brain Image Classification ... (n.d.). Retrieved October 23, 2016, from http://www.indjst.org/index.php/indjst/article/view/91042
15. Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) *Foundations of Machine Learning*, The MIT Press ISBN 9780262018258.
16. Hawkins, Douglas M. "The problem of overfitting." Journal of chemical information and computer sciences 44.1 (2004): 1-12.
17. "Lecture 22—Wednesday, November 10, 2010." *Lecture 22—Monday, November 8, 2010*. University of North Carolina, n.d. Web. 28 Apr. 2017.A Guide to Support vector machines (SVMs). (n.d.). Retrieved October 21, 2016, from http://www.cs.ucf.edu/courses/cap6412/fall2009/papers/Berwick2003.pdf
18. Gordon, G. "Support Vector Machines And Kernel Methods." *Machine Learning for Spatial Environmental Data* (2009): 247-346. Carnegie Mellon University. Web.Basic Concepts for Neural Networks - cheshireeng.com. (n.d.). Retrieved October 22, 2016, from https://www.cheshireeng.com/Neuralyst/nnbg.htm
19. Karpathy, A. (n.d.). Convolutional Neural Networks for Image Recognition. Retrieved October 20, 2016, from http://cs231n.github.io/neural-networks-1/
20. Musical Audio Synthesis Using Autoencoding Neural Networks ... (n.d.). Retrieved October 23, 2016, from http://bregman.dartmouth.edu/content/musical-audio-synthesis-using-autoencoding-neural-networks
21. Olshausen, B. A. and D. J. Field. "Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1."*Vision Research*, Vol.37, 1997, pp.3311–3325.
22. McCaffrey, James D. "L1 and L2 Regularization for Machine Learning." *James D. McCaffrey*. N.p., 12 Feb. 2015. Web. 28 Apr. 2017.
23. "Autoencoders and Sparsity." *Autoencoders and Sparsity - Ufldl*. Stanford University, n.d. Web. 28 Apr. 2017.

24. Jason Smith, Jacob Bailin, and Binlin Wu**, “**Characterization and discrimination of basal cell carcinoma and normal human skin tissues using resonance Raman spectroscopy,” presented at CMOC, University of Connecticut, Storrs, CT, April 2017.