K-nearest neighbors algorithm (KNN) and artificial neural networks (ANN) accurately predicting malignancy of breast cancer (BC) tissue solely based of features acquired from imaging modalities.

Bailey O'Shea, *George Mason University, Dept. of Bioengineering*

***Abstract*—With the reoccurrence of unnecessary open surgeries on potential malignant tissue, there is a need for additional non-invasive tools oncologists and radiologists can utilize to help argue the reason behind performing surgical biopsies. Thus, machine learning algorithms (MLAs) have seen a great deal of attention to the classification of tissue malignancy. One major benefit rises in having the ability to utilize past accessible datasets to accurately predict/classify new data/patients with similar features. The purpose of this paper was to apply and assess two MLAs—k-nearest neighbor (KNN) and artificial neural network (ANN)—on classification accuracy of breast cancer (BC) malignancy. Importantly, features used for the MLAs are acquired from imaging modalities, solely. For this particular dataset, features seen to be extracted from medical images include clump thickness, uniformity of cell size, uniformity of cell shape and marginal adhesion. The optimal k-nearest neighbor and ANN hidden layer will be reported. After implementing and testing the two MLAs, the accuracy for the KNN and ANN were 100% at 132-nearest neighbors and 95.24% ± 0.224 respectively. Considering the performance across both MLAs, the optimal classification algorithm for this dataset is the KNN algorithm. Thus, allowing for the possibility of clinical use as an additional consultation tool.**

I. INTRODUCTION

Breast cancer (BC) has become a serious health concern as it affects roughly 13 percent (1 in 8) of women within the United States; BC is the second leading cause of death for women (lung cancer being the first). Additionally, American Cancer Society has predicted that for 2020, about 276,480 new cases of invasive BC will be diagnosed with roughly 48,530 of those cases being carcinoma in situ [1]. Current methods for detection include imaging modalities such as mammography, magnetic imaging resonance (MRI), nuclear imaging, ultrasound (US) and computed tomography (CT). In the case of microcalcification (small calcium deposits that result in abnormal tissue on images) or tumor characteristics (shady contrasts within the image), oncologists or radiologists call for additional testing by performing a biopsy—a process to extract tissue samples for definite conformation. Like any cancer, BC can be categorized into two well-defined classes: malignant or benign. Malignant tumors refers to the cancer cells rapidly proliferating while spreading and damaging nearby breast tissue; in contrast, benign BC refers to cancer cells that do not invade surrounding tissue, causing it to be harmless to the patient. Having to make the choice to operate on a patient to determine malignancy does have drawbacks; (1) if the tumor was benign and thought to be malignant, the surgery was unnecessary but useful; (2) the possibility of tissue loss, fibrosis and symmetry; (3) recovery and potential medication required for patients post-surgery; (4) in rare instances, the possibility of a injurious or noxious episode to occur on the patient [2]. Hence, the need for a completely non-invasive and reliable assessment tool solely based off features acquired from imaging modalities that will aid oncologists and radiologists in definitive assurance and reason for surgery.

There are futuristic methods being used to assess the application on accurately predicting the difference between benign and malignant BC tissue given various features solely from non-invasive techniques. One method may include using machine learning algorithms (MLAs) to accurately assess the malignancy of BC tissue from features extracted from imaging modalities. MLAs refer to algorithms that utilize datasets with various features and class labels to "learn" and accurately predict class labels given specific features. For BC MLAs, features may include tumor thickness, cell uniformity, epithelial cell size and marginal adhesion with class labels possibly being malignancy or survival. To further quantify the two elements, features and class labels have the option to be a range of discrete or continuous values or binary. For example, considering the tumor thickness with BC patients, this feature can range greatly from patient-to-patient, whereas the class label for malignancy could be represented in binary: benign (0) or malignant (1).

Two well-known MLAs include k-nearest neighbors (KNN) algorithm and artificial neural networks (ANN), each having their own benefits, drawbacks and applications. KNN is known to be one of the simplest classification methods by utilizing its closest *k*-nearest neighbors to accurately predict the class given an unclassified data point with similar features (**Fig. 1**) [3].

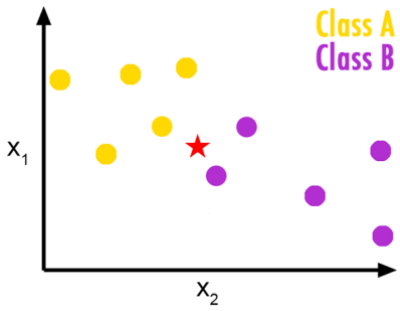

Figure 1: This figure illustrates the KNN algorithm with a given unknown data point (red star) and its correctly classified *k* neighbors. Permission granted for use of schematic from [4].

The KNN algorithm uses its training samples to classify data points. One benefit from this algorithm is its ability to have no training period, giving it the alias *lazy learning algorithm*. However, prior to determining the closest *k* neighbors, the distance from the unclassified data point or points need to be calculated. There are various distance equations the user could implement: Euclidean distance (**1**), Minkowski distance (**2**), Cityblock distance (**3**) or Cosine distance (**4**).

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (x_{ir} - x_{jr})^2} \qquad (\mathbf{1})$$

$$d(x_i, x_j) = \sqrt[p]{\sum_{r=1}^{n} |x_{ir} - x_{jr}|^p} \qquad (\mathbf{2})$$

$$\text{where } p = 1, 2, \dots, \infty$$

$$d(x_i, x_j) = \sum_{r=1}^{n} |x_{ir} - x_{jr}| \qquad (\mathbf{3})$$

$$d(x_i, x_j) = 1 - \frac{\sum_{r=1}^{n} x_{ir} x_{jr}}{\sqrt{\sum_{r=1}^{n} x_{ir}^2} \sqrt{\sum_{r=1}^{n} x_{jr}^2}} \qquad (\mathbf{4})$$

Another benefit from using the KNN algorithm is the uneducated property it bears where no prior information about the dataset is needed for the algorithm to accurately predict new data labels. However, drawbacks from implementing this learning method may include memory space needed for large datasets and being computationally expensive.

Similarly, ANNs are used for classification purposes when features of the corresponding labels are given. However, the process in which classification is determined is different from the KNN learning method. ANNs are referred to as connected processing units, called nodes, that can store and utilize information to accurately predict one or more labels of given features. Moreover, each individual node carries a numeric weight/parameter, showing the strength between units. A major benefit to ANN is when considering the various ways neural networks (NNs) are created; there are simpler and complex NNs such as the single-layer perceptron and Kohonen's SOM respectively (**Fig. 2**). The direction the outputs from each nodes correspond to other nodes or themselves determine which category the NN belongs to, feed-forward networks or recurrent/feedback networks. For example, if the output of one node is the input to a previous node or itself, it's considered to be feedback, whereas, the output of nodes going solely into the node ahead of them is feed-forward.

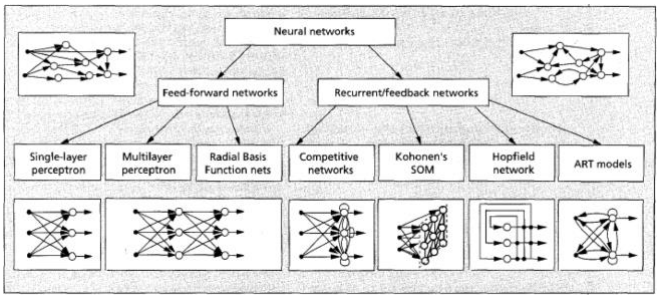

Figure 2: This figure displays the two types of neural networks, feed-forward and recurrent/feedback, and the various types of architectures. Permission granted for use of schematic from [5].

To explain the process of how NNs learn, there initially needs to be weights assigned each node. Assigning the weights numerical values does not matter as they will be updating as the NN processes the data. Secondly, the use of a learning algorithm needs to be implemented along with a training dataset for the network to "learn" what features correspond to what outputs, this refers to validation. Lastly, a testing dataset is utilized to assess the overall performance and accuracy in label prediction of the NN.

The overall purpose of this research is to utilize two machine learning algorithms, KNN and ANN, for the prediction of malignancy within BC tissue solely from features acquired from medical imaging modalities. Additionally, this research serves to further the knowledge of machine learning algorithms while hoping it will encourage clinical application.

II. Methods and Materials

Data corresponding to BC patients was utilized from the University of Wisconsin Hospitals, Madison. The Wisconsin Breast Cancer Database includes 699 patients with 10 attributes/features corresponding to the BC tissue (**Table 1**). Each feature's domain was provided a range of discrete values scaling the feature across all 699 patients. The given class label was benign (2) or malignant (4) for all patients. For those patients that have missing features, their features were awarded a '?'; there were 16 instances where features were missing. However, those 16 corresponding patients were still considered as their features could not be acquired from medical imaging modalities.

Initially, the determination of how many patients correspond to each dataset—training, validation and testing—was required. Thus, out of the 699 patients, the first 490 (70%) of patients will correlate to the training dataset while 105 (15%) of patients will be within the validation and testing dataset each, applying 699 patients to the MLAs. Howevern the two learning methods need further acknowledgement on what will be implemented:

KNN

Within the KNN algorithm, the training and validation dataset will serve to fine tune *k*; this will allow for the algorithm to iterate over a range of *k* neighbors, 1 to 250. Additionally, this gives the opportunity to plot the classification error percentage of the validation and testing set as a function of *k*. The error will be defined as the number of misclassified data points divided by the total number evaluated. After determining the optimal *k* value, this value will then be applied to the testing dataset, showing if *k* nearest neighbors correctly predicted a new data point.

ANN

The ANN architectures from **Fig. 2** implemented for this algorithm are the basic single-layered and multi-layered perceptron. Initially, random weights will be assigned to *n* number of nodes in which a threshold function will predict the output based off a variety of parameters. Once the predictions are made, they will then be compared to the corresponding true class label, being benign (2) or malignant (4). However, the respected class labels need to be translated into binary vectors, meaning, benign (2) and malignant (4) labels will translate to [1 0] and [0 1] respectively. If only the prediction does not match its true label, then will the algorithm update the weights. This process of checking prediction against the true label and updating weights will repeat itself until the algorithm has consistently predicted correct labels. Afterwards, assessment by reporting the best validation performance at the respected *epoch* will be illustrated. Comparison of performance across various hidden layers (1-5) with the Levenberg-Marquardt training algorithm will be implemented.

Features/Attributes

Based from the patient's tumor information and characteristics, features that can be determined from current and widely used medical imaging modalities, such as quantitative ultrasound (QUS), are clump thickness, uniformity of cell size, uniformity of cell shape and marginal adhesion [6,7]. Thus, only those four respected features will be applied to the two MLAs.

Coding

Application for this research and algorithm implementation will be used via Mathworks' MATLAB software and Deep Learning Toolbox's Applications.

III. Results

KNN

The following graphs show the classification error plot as a function of *k*:

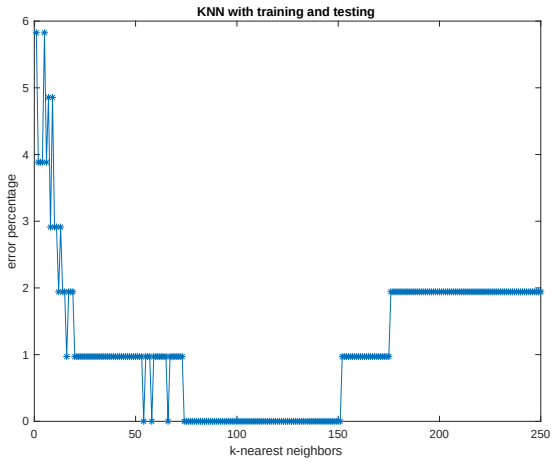

Figure 3: This figure shows the classification error plot as a function of *k* for the KNN classifier using the Euclidean distance metric excluding validation. As presented by the graph, this classifier reports numerous optimal *k* values. (i.e., when *k* = 75-150, error = 0%).

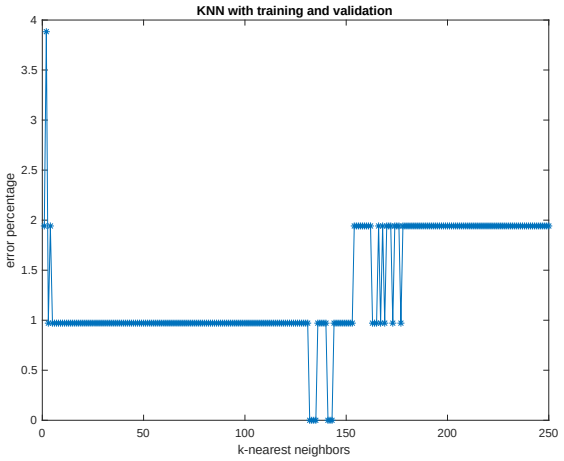

Figure 4: This figure shows the classification error plot as a function of *k* for the KNN classifier using the Euclidean distance metric including validation. As presented by the graph, this classifier reports only several optimal *k* values. (i.e., when *k* = 132-136, error = 0%).

ANN

The following graph show the performance and confusion matrix from the optimal three hidden layer NN using the Levenberg-Marquardt algorithm:

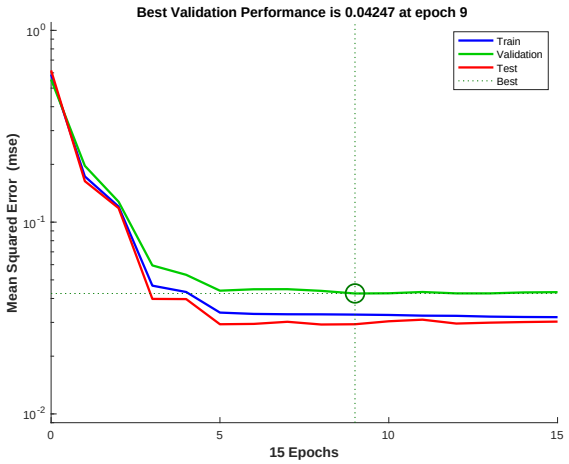

Figure 9: This figure shows the ANN classifier performance at three hidden layers with Levenberg-Marquardt backpropagation to be 0.04247 at epoch 9 out of 1000.

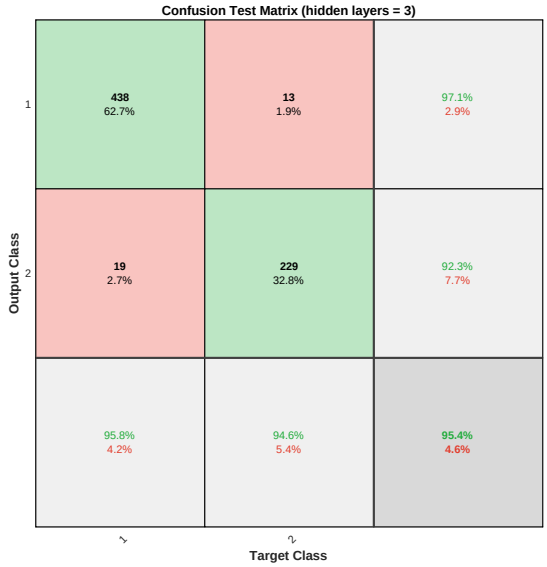

Figure 10: This figure shows the ANN classifier test confusion matrix accuracy at three hidden layers with Levenberg-Marquardt backpropagation to be 95.4%.

Table 1 summarizes the ANN performance and accuracy (MEAN and s.d.) from the various hidden layers across 10 iterations.

IV. Discussion

How MLAs perform and implemented are partially affected by data type. Meaning, having real values compared to binary can result in applying a MLA that handles one better than the other. In this instance, binary labels were provided, allowing for a much simpler implementation of MLAs, such as KNN and ANN. With such linear data, accuracies from these algorithms can be predicted to high (>90%).

KNN

Applying this logic to the KNN implemented above, by choosing the 132-nearest neighbors with or without a validation set, the resulting accuracy is 100%. Meaning, with this particular dataset and *k*-nearest neighbors, the KNN accurately classified binary malignancy (benign (0) or malignant(1)) of the entire test dataset (105 patients). However, in clinical practice, this method of machine learning and prediction, solely, to confer patients with malignant or benign cancer tissue would not be exploited. Rather, this application, like others, would be utilized as an aid for oncologists and radiologists in clinical situations. Moreover, if individuals were to utilize the KNN algorithm, it can become computationally expensive and unreliable depending on the amount of data/patients and features respectively.

ANN

The same concept of clinical application for ANN classification applies; this method of accurately classifying new data would be used as another tool for consultation. Moreover, the level of accuracy within the ANN needs to be significant (>95%) for clinical use. Thus, when considering the accuracy and overall best performance across the various hidden layered ANNs, the optimal ANN hidden layer is three (**Table 1)**; this three hidden layer NN reported a performance and accuracy of 0.031439 ± 0.0105 and 95.24% ± 0.224 respectively. However, while ANNs lack the ability to provide explained behavior behind solutions, this may cause reduced trust in the algorithm. Additionally, by experimenting with various NN architecture types, it's possible find an optimal network structure for particular datasets. Thus, achieving the proper NN is by trial and error, as seen above.

V. Conclusion

The overall purpose of this research was to implement and assess two machine learning algorithms (MLAs), k-nearest neighbors and artificial neural networks, for the classification of malignancy (benign = 0 and malignant = 1) within breast cancer tissue. Only the following BC tissue features acquired from medical imaging modalities were utilized: clump thickness, uniformity of cell size, uniformity of cell shape and marginal adhesion. After implementation, the optimal *k*-nearest neighbor value and ANN hidden layer was 132 and three respectively; KNN of 132 neighbors reported an accuracy of 100% in classifying malignant versus benign tissue, whereas, ANN with three hidden layers reported an accuracy of 95.24% ± 0.224. Thus, for this particular dataset and potentially another consultation tool in clinical practice, the ideal MLA used would be KNN.

Acknowledgements

This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

REFERENCE(S)

[1] "How Common Is Breast Cancer?: Breast Cancer Statistics," *American Cancer Society*. [Online]. Available: https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html. [Accessed: 29-Mar-2020].

[2] "Risks of Cancer Surgery," *American Cancer Society*. [Online]. Available: https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/surgery/risks-of-cancer-surgery.html. [Accessed: 30-Mar-2020].

[3] P. Cunningham, and S. Delany. "k-Nearest Neighbour Classifiers," *Mult Classif Syst.* [Online]. Available: https://www.researchgate.net/publication/228686398_k-Nearest_neighbour_classifiers [Accessed: 29-Mar-2020].

[4] Q. Wei, Class Lecture, Topic: "Data II, KNN." BENG420/520, Volgenau School of Engineering, George Mason University, Fairfax, VA., Mar., 2020.

[5] Q. Wei, Class Lecture, Topic: "Data IV, ANN I." BENG420/520, Volgenau School of Engineering, George Mason University, Fairfax, VA., Mar., 2020.

[6] L. Sannachi, et al. "Breast Cancer Treatment Response Monitoring Using Quantitative Ultrasound and Texture Analysis: Comparative Analysis of Analytical Models," *Transl Oncol*, vol. 12, no. 10, p. 1271-81, Oct. 2019.

[7] L. Sannachi, et al. "Quantitative Ultrasound Monitoring of Breast Tumour Response to Neoadjuvant Chemotherapy: Comparison of Results Among Clinical Scanners," *Ultrasound in Medicine & Biology*, vol. 46, no. 5, p. 1142-57, May 2020.

ANN Classifier		
Hidden Layers	**Best Performance (Avg & s.d.)**	**Confusion Matrix Accuracy (Avg & s.d.)**
1	0.037451 ± 0.119	95.16 ± 0.335
2	0.032415 ± 0.00980	95.13 ± 0.303
3	0.031439 ± 0.0105	95.24 ± 0.224
4	0.041338 ± 0.00999	95.35 ± 0.323
5	0.036995 ± 0.00964	95.16 ± 0.313

Table 1: This table shows the average and standard deviation across the hidden layers. Each layer was ran 10 times for avg and s.d..