Siamese neural networks for the classification of high-dimensional radiomic features

Mahajan, Abhishaike, Dormer, James, Li, Qinmei, Chen, Deji, Zhang, Zhenfeng, et al.

Abhishaike Mahajan, James Dormer, Qinmei Li, Deji Chen, Zhenfeng Zhang, Baowei Fei, "Siamese neural networks for the classification of high-dimensional radiomic features," Proc. SPIE 11314, Medical Imaging 2020: Computer-Aided Diagnosis, 113143Q (16 March 2020); doi: 10.1117/12.2549389

SPIE.

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

Siamese neural networks for the classification of high-dimensional radiomic features

Abhishaike Mahajan[1,2], James Dormer[1], Qinmei Li[1,3], Deji Chen[3], Zhenfeng Zhang[3*], Baowei Fei[1,4,*]

[1]*Department of Bioengineering, University of Texas at Dallas, Richardson, TX*
[2]*Department of Cognition and Neuroscience, University of Texas at Dallas, Richardson, TX*
[3]*Department of Radiology, The Second Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, China*
[4]*Department of Radiology and Advanced Imaging Research Center, University of Texas Southwestern Medical Center, Dallas, TX*

Abstract

This study demonstrates that a variant of a Siamese neural network architecture is more effective at classifying high-dimensional radiomic features (extracted from T2 MRI images) than traditional models, such as a Support Vector Machine or Discriminant Analysis. Ninety-nine female patients, between the ages of 20 and 48, were imaged with T2 MRI. Using biopsy pathology, the patients were separated into two groups: those with breast cancer (N=55) and those with GLM (N=44). Lesions were segmented by a trained radiologist and the ROIs were used for radiomic feature extraction. The radiomic features include 536 published features from Aerts *et al.*, along with 20 features recurrent quantification analysis features. A Student T-Test was used to select features found to be statistically significant between the two patient groups. These features were then used to train a Siamese neural network. The label given to test features was the label of whichever class the test features with the highest percentile similarity within the training group. Within the two highest-dimensional feature sets, the Siamese network produced an AUC of 0.853 and 0.894, respectively. This is compared to best non-Siamese model, Discriminant Analysis, which produced an AUC of 0.823 and 0.836 for the two respective feature sets. However, when it came to the lower-dimensional recurrent features and the top-20 most significant features from Aerts *et al.*, the Siamese network performed on-par or worse than the competing models. The proposed Siamese neural network architecture can outperform competing other models in high-dimensional, low-sample size spaces with regards to tabular data.

Keywords: Neural Network, Siamese Network, Breast Cancer, Mastitis, Radiomics, Disease Classification, MRI, Machine Learning

1. Introduction

In the past, neural networks have shown extremely high representational power when it comes to modeling complex and non-linear data, such as images or audio signals. As an example, Convolutional Neural Networks (CNN) hold the state-of-the-art performance in a vast number of problems within medical image analysis[1]. However, while fully-connected neural networks should be also useful for the high-dimensional tabular features often found in radiomics, they are largely absent from the literature; likely due to the large amounts of data that neural networks often demand to work effectively. Indeed, neural network models seem to largely be reserved for medical research problems where the dataset is in the several thousands[2, 3], which is quite rare to find in highly specific medical image analysis problems, such as granulomatous lobular mastitis (GLM) versus breast cancer

Medical Imaging 2020: Computer-Aided Diagnosis, edited by Horst K. Hahn, Maciej A. Mazurowski, Proc. of SPIE Vol. 11314, 113143Q · © 2020 SPIE
CCC code: 1605-7422/20/$21 · doi: 10.1117/12.2549389

While models such as Random Forests (RF) and Support Vector Machines (SVM) are adept at dealing with low-sample, high-dimensional spaces, we believe that a variant of neural network holds promise in working in working in similar spaces with far more efficacy: Siamese neural networks. Due to Siamese networks not modeling the underlying distribution of all features in relation to the classes, but rather creating distance-based representations of classes, this neural architecture allows for problems with many classes and low amount of data per class to be solved very effectively[4]. This characteristic of the architecture has been exploited in many papers, such as state-of-the-art performance on the Omniglot dataset[5], matching duplicate questions on Quora[6], and, recently, for person re-identification in videos[7]. The specifics of the Siamese network architecture will be discussed in the methods section.

Recurrence quantification analysis (RQA) has previously been used to study the homogeneity and revisiting of phase-space within time-series data. While attempts have been made to generalize RQA to spatial-domain data, such as medical images, the computation time of such technique are usually intractable for images larger than 50 x 50, as two-dimensional RQA has an $O(N^5)$ time complexity in the worst case. We introduce the mathematical underpinnings of this generalized technique in the methods section to define the reasoning behind recurrence plots, but this technique alone was insufficient. As such, there have been no previous studies on applying RQA-based feature extraction for image classification. In this work, by considering T2 MRI images as simply a collection of time-series summed over a given dimension, we apply non-generalized RQA to such images. Furthermore, by using an approximation technique for traditional RQA, we reduce the computation time from $O(N^2)$ to $O(Nlog(N))$, allowing us to retain the original size of the input T2 MRI image without any resizing. This will allow us to study how radiomic features could be used to accurately distinguish granulomatous lobular mastitis from breast cancer in young women without using contrast enhancement or time series data.

The purpose of this study is to showcase the effectiveness of Siamese neural networks in tabular, high-dimensional, radiomic features. Furthermore, we establish that, using the radiomics features used by Aerts et al.[8] and additional recurrent features described in this work as input, Siamese networks outperform two conventional non-neural-network models (Support Vector Machines and Discriminant Analysis) on Area-Under-Curve (AUC), Accuracy, Sensitivity, and Specificity metrics for the highest dimensional feature sets.

2. Methods

2.1 Patient Cohort

Retrospectively, 99 women between the ages of 20 and 48 who were imaged using a 1.5T or 3T MRI to collect T2 images were selected for the study. These women were separated into two groups based on pathological diagnosis via biopsy. The first group contained women with breast cancer (N = 55) while the second group was diagnosed with granulomatous lobular mastitis (N=44). Lesions were segmented by an experienced radiologist to create masks used localize the region of interest (ROI) (Figure 1).

2.2 Aerts et al. Features

We use 536 radiomic features presented by Aerts *et al.* 2014, who used the features to determine tumor phenotype in patients with lung or head and neck cancer. These features consist of 8 shape and size-based features, 273 gray-level co-occurrence matrix (GLCM) features (25 calculations across 13 directions), 143 gray-level run length (GLRL) features (11 calculations across 13 directions), and 112 wavelet statistical features (11 calculations with 8 different filter sets).

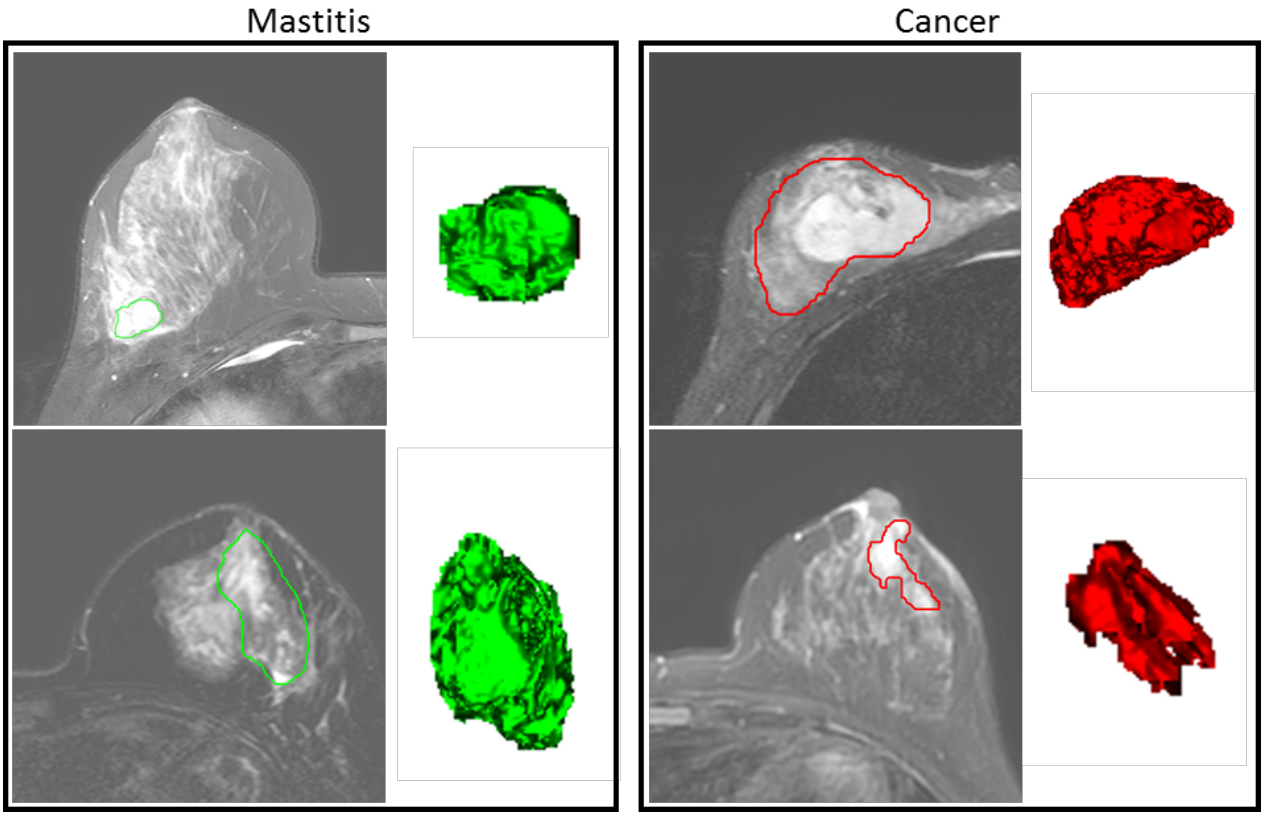

Figure 1. Example images of granulomatous lobular mastitis and breast cancer. The GLM segmentation by the radiologist is shown in red, while the cancer segmentation is shown in green. Using visual inspection alone, it is difficult to distinguish between GLM and breast cancer.

2.3 Recurrent Features

Furthermore, we use 20 additional radiomic features based on recurrent features. These features consist of 5 approximate Recurrence Quantification Analysis (RQA) features each created from 4 versions of a T2 MRI image. The 5 RQA features are Recurrence Rate (RR), Determinism (DET), Average Diagonal Length (ADL), Ratio (R), and Longest Line (LL), all of which were given using an approximation technique given by Schultz *et al.*[9]. For the calculation of these features, the input will be the ROI of a breast MRI, which is a matrix of S slices and of spatial dimensions H and W, or $MRI^{S\times H\times W}$. The regions of an MRI image, along with the rationale for using them, and the mathematical formulation below, with examples shown in Figure 2.

I. Summed alongside the slice dimension, allowing the approximation technique to search for recurrences within the horizontal dimension along the vertical dimension of the image: $MRI^{H\times W} = \sum_{i=1}^{S} MRI^{i\times H\times W}$.

II. Summed alongside the slice dimension and then transposed, allowing the technique to search for recurrences within the vertical dimension along the horizontal dimension of the image: $MRI^{W\times H} = (\sum_{i=1}^{S} MRI^{i\times H\times W})^{T}$.

III. Summed alongside the height dimension, allowing the technique to search for recurrences within the channel dimension alongside the width dimension. $MRI^{S\times W} = \sum_{i=1}^{H} MRI^{S\times i\times W}$.

IV. Summed alongside the width dimension, allowing the technique to search for recurrences within the channel dimension alongside the height dimension. $MRI^{S\times H} = \sum_{i=1}^{W} MRI^{S\times H\times i}$.

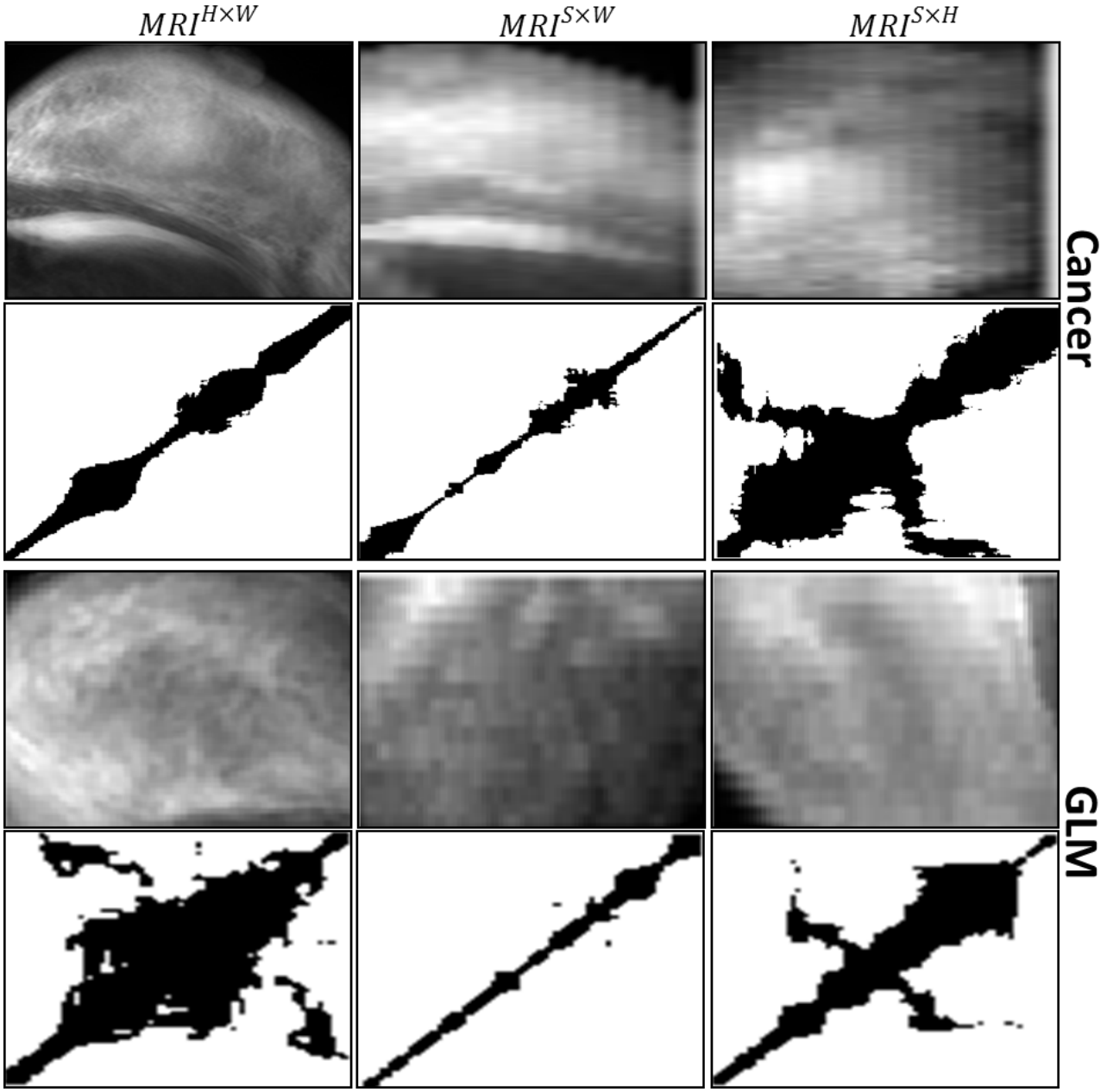

Figure 2: Examples of summed regions of interests (Rows 1 and 3) and their associated recurrence plots (Rows 2 and 4). Rows 1 & 2 represent breast cancer, while 3 & 4 represent GLM.

2.4 Siamese Neural Network

Siamese networks are a neural architecture which began to see wide-spread use in the 1990's for signature verification[10], which attempts to learn a distance-based representation of input features for the end goal of learning feature *similarity*, rather than feature classification. As mentioned earlier, Siamese networks are essentially a form of metric learning, similar to distance-based algorithms like K-Nearest Neighbors.

These networks typically consist of twin networks which share weight/parameter updates and accept inputs X_a and X_b, and produce an encoded representation of that input at the networks k^{th} layer, or $X_a^{(k)}$ and $X_b^{(k)}$. These inputs are creating by making all possible, non-repeating pairs from the training set. Thus, the total number of paired training points of a dataset of size M is given by $\frac{M(M+1)}{2}$.

The produced vectors of these twin network are then used to compute some distance metric vector, typically L2-distance, which is used as the output of the network and can be understand as the learned distance between the two inputs. There are a variety of loss functions to use with Siamese network outputs, but they all primarily revolve around maximizing/minimizing the distance between inputs that are dissimilar/similar. For example, the contrastive loss function[11] gives a non-zero loss when the distance D between the output and the ground truth y (which is either 1 or 0) are dissimilar by a number larger than a given margin:

$$yd^2 + (1-y) * \max(margin - d, 0)$$

An example of this prototypical Siamese network can be seen in Figure 3.

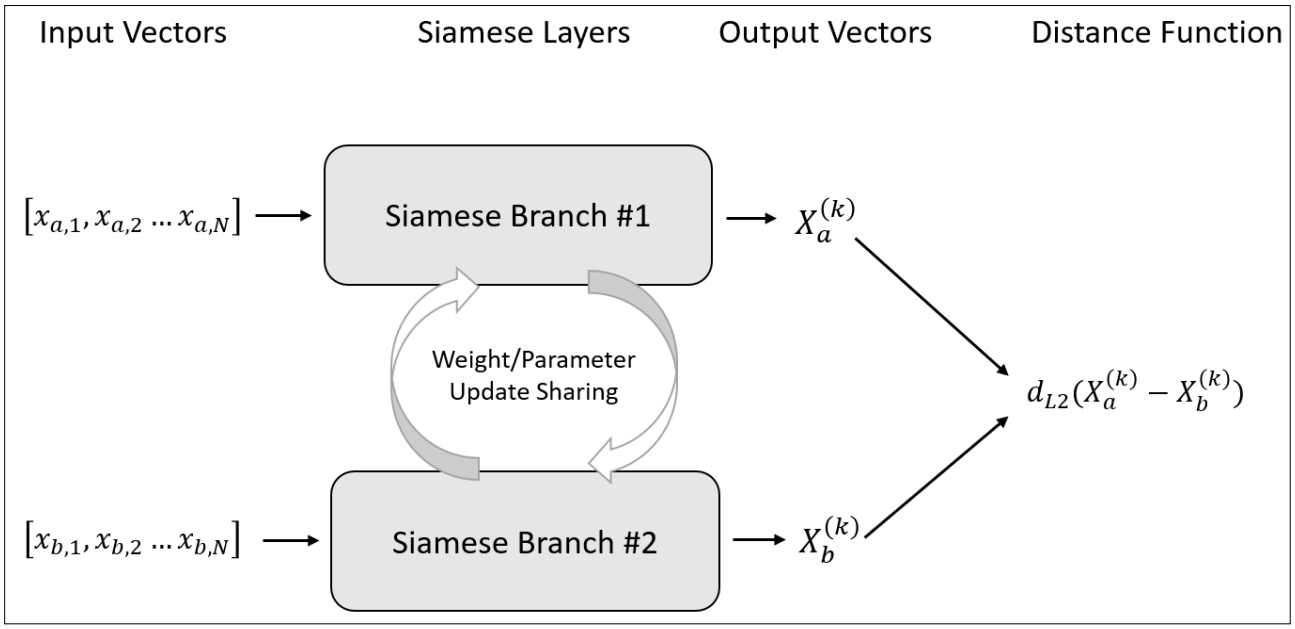

Figure 3. A typical Siamese network with two branches.

However, our experiments showed that a distance-based loss performed quite poorly. These results were similar to Homma et al.[6], who found that 'forcing' a particular distance metric upon the network, such as L2 or L1 distance, often resulted in mediocre results. Instead, we follow similar a similar methodology by removing the distance function entirely, and, instead,

use the outputs of the Siamese network to create a vector composed of the outputs $X_a^{(k)}$ and $X_b^{(k)}$themselves and the squared difference between them, all concatenated alongside the row dimension. This created vector is then fed into another neural network, which has two ReLU layers, and a final, sigmoid-activation layer to classify the inputs as a 1 (similar) or 0 (dissimilar). The network is then updated using binary cross-entropy. This essentially allows the Siamese network to create its own distance measure within its learned vector space.

It should be noted that while Homma *et al.* also included the Hadamard product of these two output vectors in this concatenated vector, we found that this addition decreased the overall accuracy of the network, so it was removed. An example of this modified Siamese network can be seen in Figure 4. It should be noted that modified Siamese network is no longer symmetric with the inputs. In practice, this ends up being a non-issue, as our method of pairwise input generation will not create any reciprocals of previous pairs. As in, if we have a pair of vectors $[A, B]$ as an input sample, we will never have a pair of $[B, A]$ in our dataset.

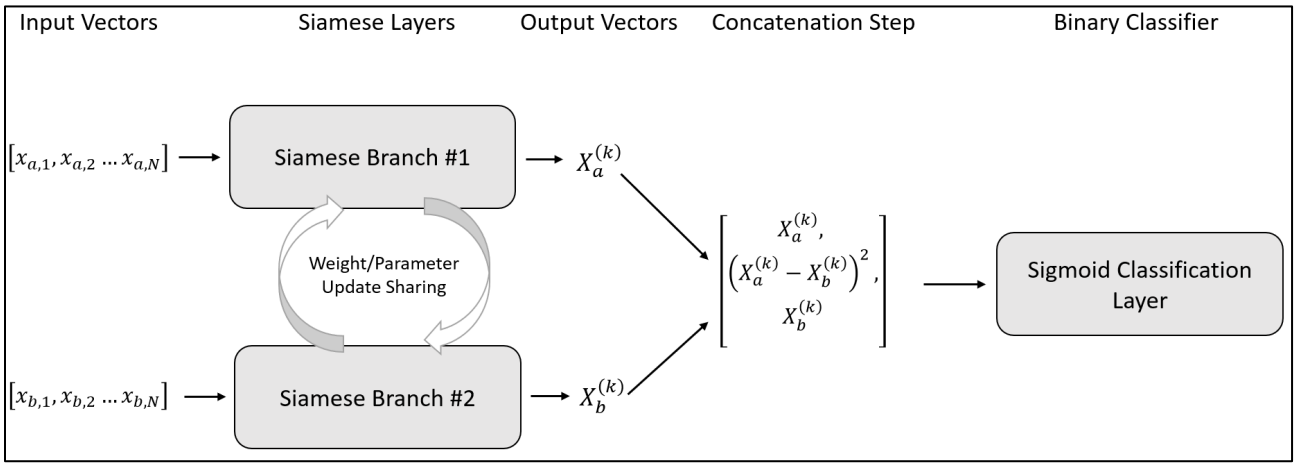

Figure 4. Siamese network weighting and parameter updates during training.

Finally, due to Siamese networks not assigning a singular class to any given input, any classification problem must use a comparison group which contains all classes within the problem to evaluate the accuracy of the network. For larger datasets, using K-Means to create representative centroids for each class would be a possible approach for this comparison group. Our method of creating a comparison group will be discussed in the validation section.

Finally, the architecture details are composed of the following: 5 layers in the Siamese branch, consisting of 512, 256, 128, 64, and 32 nodes, respectively. The sigmoid classification layer takes the concatenation of the Siamese branch (which is a vector of length 96) and passes it to layers consisting of 64, 32, 16, 8, and 1 nodes, respectively. All layers have ReLU activations placed after them, with the final layer having a sigmoid activation. Furthermore, all layers have L2 regularization applied to the weights. The model was trained using an ADAM optimizer with the learning rate set to 0.0002 over 300 epochs, and the most accurate model (given the validation loss) was used to evaluate the model.

2.5 Lesion Classification and Validation

As our results will be directly compared to our previously published work, we will use their methods to compare the Siamese architecture for certain feature groups to their used models (Support Vector Machine, or SVM, and Discriminant Analysis, or DA). There are four feature groups used to evaluate this model. Using an alpha value of 0.05, we select the significant features using the training group from 1) Aerts *et al*. features, 2) RQA features, 3) the combination of the former two, and 4) the top-20 most significant features from the Aerts et al. features. This final group was used as a direct comparison to match the total number of RQA features.

As noted earlier, Siamese networks require a comparison group to use for validating for classification problems. Due to relatively small size of our dataset/number of classes, we use the approach given by Koch *et al.*[5], where the comparison group is the entirety of the training set. As in, we create pairs between all points in the training set and all points in the testing set. From here, given the network-given similarity between each point in the training set and any given testing point, we simply select the class that has the highest normalized similarity as the predicted class.

3 Results

The Siamese network was implemented in TensorFlow[12] and deployed on an Nvidia Titan XP GPU. Across all the patients, with an alpha value of 0.05, 264 of the 536 Aerts *et al.* features (49%) were found to be statistically significant between the two groups when using a Student T-Test. For the RQA features across, across all patients, 18 of the 20 features (90%) RQA features were significant. Furthermore, we also used a cross-validation setup; each classifier was trained on 11 folds of data, where the features used were determined based on significance determined in the training data, with each fold containing 5 cancer patients and 4 mastitis patients for evaluation. This was repeated 25 times to create average AUC, accuracy, specificity, and sensitivity values. The patients used for training in each fold were identical for each model tested. The results from the network are shown in Table I.

Table I: Results from using the various features and models to distinguish mastitis from cancer in breast cancer patients.

Model	Features	Accuracy	Sensitivity	Specificity	AUC
Siamese	Aerts	0.783	0.823	0.752	0.853
Siamese	Aerts Top 20	0.719	0.659	0.765	0.719
Siamese	RQA	0.723	0.679	0.756	0.760
Siamese	Aerts & RQA	**0.847**	**0.849**	**0.845**	**0.894**
DA	Aerts	0.748	0.796	0.694	0.823
DA	Aerts Top 20	0.717	0.756	0.673	0.768
DA	RQA	0.748	0.778	0.717	0.836
DA	Aerts & RQA	0.737	0.784	0.694	0.836
SVM	Aerts	0.727	0.764	0.689	0.777
SVM	Aerts Top 20	0.659	0.699	0.612	0.768
SVM	RQA	0.717	0.737	0.691	0.836
SVM	Aerts & RQA	0.727	0.750	0.707	0.814

The highest AUC value was obtained when the features from Aerts *et al.* were combined with RQA features, with a value of 0.894. In comparison, the highest AUC value acquired when using DA and SVM was 0.836. For SVM, this was obtained using only the recurrent features, while DA performed equally well using the recurrent features alone as well as the combination of recurrent features and those from Aerts *et al.*.

4 Discussion

While deep learning has been previously used for medical images, it is almost always with raw images combined with convolutional neural networks, as discussed extensively in Lundervold *et al*.[1] When it comes to tabular radiomics data, the vast majority of studies use alternative models, such as support vector machines, random forests, and logistic regression[13-15]. This is likely due to traditional neural networks requiring a large amount of data to work effectively, which is something that is often lacking in many niche medical problems.

In this work, we applied a variant of a Siamese neural network to the problem of mastitis/cancer classification of tabular radiomics data. Despite having over 250 features in our Aerts and Aerts & RQA feature groups, which is 2.5x more features than samples (99), the network performed more effectively than the best two competing non-neural-network models (Discriminant Analysis and Support Vector Machines). Within this comparison, the Siamese network scored higher in accuracy, sensitivity, specificity, and area-under-curve metrics for the classification task.

This is the first study that uses a Siamese network for radiomic feature classification and compares it against non-neural network models. Our main contribution is demonstrating the usefulness of deep metric learning in tabular radiomics data and to show state-of-the-art results in the classification of GLM and malignant breast cancer.

Future improvements to this work are numerous. For one, it would be beneficial to incorporate more elaborate techniques to improve the efficiency of the Siamese network, such as smart negative mining[16], triplet losses[17], and others. While most of these techniques were created for computer-vision purposes, they would be just as likely to work for tabular data. Secondly, it is possible that the usage of Siamese networks could be extended to the convolutional case, allowing for raw-image classification with a small number of samples. Thirdly, and finally, it is interesting that fact that Siamese networks were far less accurate with the smaller feature groups than the DA and SVM models, which both had an AUC of 0.836 on the 18 significant RQA features, while the Siamese network only had an AUC of 0.760. A future study could focus on why metric learning seems to counter-intuitively suffer in a lower-dimensional space compared to higher dimensional spaces.

Finally, the primary limitation of this study is that the network used was not benchmarked on multiple high-dimensional radiomics datasets, especially in different imaging modalities, such as features from CT scans. In future studies, it would be worth exploring the impact that differing feature groups have on a Siamese networks data efficiency.

5 Conclusion

In this study, the usefulness of Siamese neural networks for high-dimensional, low sample problems was investigated using a GLM versus breast cancer classification task with tabular data from radiomic features. It was found that this variant of neural network performs exceedingly well compared to two leading non-neural-network models in classifying the given radiomics features. The highest patient-level accuracy of 84.7% was achieved was the Siamese network was combined with the collection of textural features and RQA features. As a major limitation in this study was the small number of patients (N = 99), future work will expand the number of patients to test the generalizability of the method described.

Acknowledgment

This research was supported in part by the U.S. National Institutes of Health (NIH) grants (R01CA156775, R01CA204254, R01HL140325, and R21CA231911) and by the Cancer Prevention and Research Institute of Texas (CPRIT) grant RP190588.

References

[1] Lundervold, A. S., and Lundervold, A., “An overview of deep learning in medical imaging focusing on MRI,” Zeitschrift für Medizinische Physik, (2018).

[2] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S., “Dermatologist-level classification of skin cancer with deep neural networks,” Nature, 542(7639), 115 (2017).

[3] Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., and Webster, D. R., “Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning,” Nature Biomedical Engineering, 2(3), 158 (2018).

[4] Horiguchi, S., Ikami, D., and Aizawa, K., “Significance of Softmax-based Features in Comparison to Distance Metric Learning-based Features,” arXiv preprint arXiv:1712.10151, (2017).

[5] Koch, G., Zemel, R., and Salakhutdinov, R., "Siamese neural networks for one-shot image recognition." 2.

[6] Homma, Y., Sy, S., and Yeh, C., “Detecting Duplicate Questions with Deep Learning,” (2016).

[7] Wu, L., Wang, Y., Gao, J., and Li, X., “Where-and-when to look: Deep siamese attention networks for video-based person re-identification,” IEEE Transactions on Multimedia, (2018).

[8] Aerts, H. J. W. L., Velazquez, E. R., Leijenaar, R. T. H., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M. M., Leemans, C. R., Dekker, A., Quackenbush, J., Gillies, R. J., and Lambin, P., “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” Nature Communications, 5, 4006 (2014).

[9] Schultz, D., Spiegel, S., Marwan, N., and Albayrak, S., “Approximation of diagonal line based measures in recurrence quantification analysis,” Physics Letters A, 379(14-15), 997-1011 (2015).

[10] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R., "Signature verification using a" siamese" time delay neural network." 737-744.

[11] Hadsell, R., Chopra, S., and LeCun, Y., "Dimensionality reduction by learning an invariant mapping." 1735-1742.

[12] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X., “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” arXiv preprint arXiv:1603.04467, (2015).

[13] E, L., Lu, L., Li, L., Yang, H., Schwartz, L. H., and Zhao, B., “Radiomics for Classification of Lung Cancer Histological Subtypes Based on Nonenhanced Computed Tomography,” Academic Radiology, (2018).

[14] Cho, H.-h., Lee, S.-h., Kim, J., and Park, H., “Classification of the glioma grading using radiomics analysis,” PeerJ, 6, e5982 (2018).

[15] Chen, C.-H., Chang, C.-K., Tu, C.-Y., Liao, W.-C., Wu, B.-R., Chou, K.-T., Chiou, Y.-R., Yang, S.-N., Zhang, G., and Huang, T.-C., “Radiomic features analysis in computed tomography images of lung nodule classification,” PloS one, 13(2), e0192002 (2018).

[16] Harwood, B., Kumar, B., Carneiro, G., Reid, I., and Drummond, T., "Smart mining for deep metric learning." 2821-2829.

[17] Schroff, F., Kalenichenko, D., and Philbin, J., "Facenet: A unified embedding for face recognition and clustering." 815-823.