A Comprehensive Study On The Applications of Machine Learning For Diagnosis Of Cancer

Mehak Rajput, Mohnish Chakravarti, Tanay Kothari
{mehakrjpt, mohnishchakravarti7, tanaykothari1998} @ gmail.com

May 12, 2015
1 Abstract

Collectively, lung cancer, breast cancer and melanoma was diagnosed in over 535,340 people out of which, 209,400 deaths were reported [13]. It is estimated that over 600,000 people will be diagnosed with these forms of cancer in 2015. Most of the deaths from lung cancer, breast cancer and melanoma result due to late detection. All of these cancers, if detected early, are 100% curable. In this study, we develop and evaluate algorithms to diagnose Breast cancer, Melanoma, and Lung cancer. In the first part of the study, we employed a normalised Gradient Descent and an Artificial Neural Network to diagnose breast cancer with an overall accuracy of 91% and 95% respectively. In the second part of the study, an artificial neural network coupled with image processing and analysis algorithms was employed to achieve an overall accuracy of 93% A naive mobile based application that allowed people to take diagnostic tests on their phones was developed. Finally, a Support Vector Machine algorithm incorporating image processing and image analysis algorithms was developed to diagnose lung cancer with an accuracy of 94%. All of the aforementioned systems had very low false positive and false negative rates. We are developing an online network that incorporates all of these systems and allows people to collaborate globally.
2 Introduction

Breast cancer, lung cancer and melanoma together account for over 30% of the total cases of cancer reported in 2014. With over 200,000 deaths last year and an expected 300,000 deaths this year, it is imperative that measures are taken to prevent the rise in the deaths occurring collectively due to these diseases.

The 5-year survival rate for lung cancer is 54% However, if not diagnosed early, the chances of survival reduce drastically to 16%. The 5-year survival rate for breast cancer is 86% but it falls to 22% if not diagnosed early. Finally, the 5-year survival rate for melanoma is 95% but it falls to 22% if not diagnosed early. All of these diseases, if diagnosed early, are 100% curable. To diagnose these diseases early, we develop and apply novel machine learning algorithms.

In 2007, the Universiti Sains Malaysia and Universiti Malaysia Perlis collaborated to collect data similar to WBC. This data was utilized to build diagnostic programs. Seven implementations were tested and results proved encouraging; accuracy ranged between 75.38% and 100%. The purely AI based programs did not fare as well as conventionally programmed implementations, but showed promise. Over the years, further research has been done in AI and their applications to biomedicine ([15], [16], [17]).

There are several benefits to using Machine Learning algorithms in the field of biomedicine. It eliminates the added dimension of mistakes committed due to human carelessness. The extensive amount of data available online also makes it possible for the algorithms to train themselves and achieve near perfect accuracy - something that is very difficult for any human to achieve. Furthermore, it is easy to replicate and can be shared globally, thus reducing costs associated with logistics and labor force.

For this study, machine learning algorithms such as Support Vector Machine, Artificial Neural Network and Gradient Descent were used. Image processing and analysis was also used to obtain features from images in the case of lung cancer and melanoma.

The next step was feature selection. For breast cancer, features such as the smoothness, concavity, fractal dimension, etc. were used from the WBC database. For lung cancer, image processing and analysis was applied on CT images from the NCBI database to obtain features such as area, convex area, solidity, etc. For melanoma, the ABCD standard prescribed by dermatologists was used. The images were obtained through the PH$^2$ dataset.
Outline  The remainder of this article is organised as follows. Section 3 contains our research, methods and findings for detecting lung cancer. Section 4 contains our research, methods and findings for detecting melanoma. Section 5 contains our research, methods and findings for detecting breast cancer. Section 6 contains our conclusions. Section 7 and Section 8 present the acknowledgements and the references respectively.
3 Lung Cancer

Support Vector Machine along with image processing and image analysis algorithms were used in this part of the study.

3.1 Image pre-processing, processing and segmentation

Total variation denoising was used in the pre-processing stage. It is very effective at preserving edges while smoothing away noise in flat regions, even at low signal-to-noise ratios.

There are two types of pixels in chest CT images: 1) Pixels within the dense body and chest wall structures. 2) Low density pixels. The optimal thresholding proposed in [19] is applied to the pre-processed image to obtain a segmentation threshold.

\[
T_{i+1} = \frac{1}{2}(u_a + u_b)
\]

In this equation, \(T_i\) is the segment threshold after the \(i^{th}\) step. To choose a new segmentation threshold, we apply \(T_i\) to the image to separate the pixels into body and nobody pixels. Let \(u_a\) and \(u_b\) be the mean grey level of body and nobody pixels.

The pixels with a density lower than the threshold value are assigned the value 1 whereas other pixels are assigned the value of 0. The remaining non-body pixels are eliminated through morphological closing.
3.2 Feature Selection

After the ROI (Region Of Interest) was obtained, GLCM was used to extract the features. Gray Level Co-occurrence Matrix is a very popular way to describe the texture of an image. The features extracted and the methods used are shown in the table. (Table 1)

| Feature          | Formula | Function                                      |
|------------------|---------|-----------------------------------------------|
| Area (A)         | -       | Total number of pixels in the ROI             |
| Convex Area (CA) | -       | Total number of pixels in the convex region of the ROI |
| Equivalent Diameter (ED) | $ED = \frac{\sqrt{4A + A^2}}{A}$ | Area of circle equal to the area of the ROI     |
| Solidity (S)     | $S = \frac{A}{CA}$ | Ratio of Area to Convex Area                  |
| Energy (E)       | $E = \sum_0^N p^2(i,j)$ | Summation of the squared elements in the GLCM  |
| Contrast (C)     | $C = \sum_i^N \sum_j^N (i - j)^2 p(i,j)$ | Measure of contrast between intensity of adjacent pixels over the whole ROI |
| Eccentricity (EC)| -       | Ratio of distance between the foci of the ellipse and its major axis length |
| Homogeneity (H)  | $H = \sum_j^N \frac{p(i,j)}{1 + |i + j|}$ | Measure of closeness of the distribution of elements in the GLCM to the GLCM of each ROI |
| Correlation (CO) | $CO = \sum_i^N \sum_j^N \frac{p(i,j) - \mu_r \mu_c}{\sigma_r \sigma_c}$ | Measure of correlation of pixel to its neighbor over the ROI |
3.3 Support Vector Machine Development

A linear classifier was used. Best hyper plane is the one that represents the largest separation or margin between the two classes. So we chose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If such a hyper plane exists, it is known as the maximum margin hyperplane and the linear classifier it defines is known as a maximum classifier.

![Figure 1: An example of a maximum margin classifier](image)

The CT images used were obtained from NCBI’s online database.

3.4 Results

Over 197 images were tested on the SVM. The classifier only chose 2 out of the 9 features available for classification at any given time. It was seen that the most reliable features were Area and Convex Area (please refer to the table in the results section).

A breakup of 70 : 30 was used for training and testing of our algorithm. The data presented here is from the testing period of the algorithm.

Based on the classification, the tumour was either diagnosed as benign or malignant. The SVM delivered an accuracy of over 94%, mis-diagnosing only 11 of the images. 5 were false positive and 6 were false negative.

| Output | Significance |
|--------|--------------|
| 0      | Benign       |
| 1      | Malignant    |
The overall accuracy delivered by the SVM was, as noted earlier, 94%. The results have been summarised below (Figure 2).

| Overall Outcome | Overall Results |    |    |
|-----------------|-----------------|----|----|
|                 | Positive        | Negative |    |    |
| TP: 129         | FP: 5           | TN: 59 | FN: 6 |
| Inconclusive: 0 |                 |    |    |
| Positive Prediction Rate | 96.23% | Negative Prediction Rate | 99.77% |

Sensitivity 95.56%

Specificity 92.12%

Matthews Correlation Coefficient 0.874

Figure 2: Summary of the results from the Lung Cancer Classifier
4 Melanoma

Artificial Neural Networks along with image processing and image analysis algorithms were used in this part of the study.

4.1 Image pre-processing, processing and segmentation

Noise such as hair was removed using morphological closing. An accurate border was determined. The gradient function was used.

4.2 Feature Selection

Features such as asymmetry along the minor axes, asymmetry along the major axis, border irregularity, entropy, color variation, diameter, etc. were extracted using built-in or custom coded functions. The asymmetry of the lesion was determined by overlapping the halves of the images on their major and minor axes. The border was also calculated using in-built functions we used an innovative out-in search approach to accurately determine the border.

4.3 Artificial Neural Network Development

The neural network was developed using an in-built toolbox in MATLAB. The features extracted were given numerical values and used as input in the ANN.

4.4 Mobile Application

The ANN was converted into a standalone service and deployed on the web. It was then integrated into the mobile application.
The Melanomore application is currently under development and further
details will be released shortly.

When the users click on Get Results after uploading the image, they will be redirected
to a page where they will be able to see their complete report. The report will
automatically get uploaded to the database and should the user require to submit the
report to their dermatologist, they can do so with inbuilt functionality.

Figure 4: The flowchart of the processes of the application.

Figure 5: The mobile application for diagnosis of melanoma.
4.5 Results

The ANN employed in the prototype of the application had a promising accuracy of 92% for training and validating purposes over 450 images were used (400 from $PH^2$ dataset and 56 images sourced from other publicly available datasets). Throughout the training period, from an initial accuracy of 85% at use of 250 images, it gradually increased in accuracy to 93% at the end of 450 images used. Also, the FP, FN, TP, TN, and the inconclusive rates were low.

A breakup of 70 : 30 was used for training and testing of our algorithm. The data presented here is from the testing period of the algorithm.

Figure 6: Comparison of training size and the accuracy of the algorithm.

As the training size and the increased, the inconclusive diagnosis rate reduced too.

Figure 7: Comparison of training size and the inconclusive diagnosis rate of algorithm.
A T-test was taken to prove that the values calculated for asymmetry and border were actually different for both cases - malign as well as benign. The results of the T-test showed that both the groups were significantly different as both of them had extremely low probability values to be in the same group.

| Border      | Asymmetry  |
|-------------|------------|
| $6.19 \cdot 10^{-7}$ | $8.12 \cdot 10^{-3}$ |

The results have been summarised below (Figure 8).

| Overall Outcome | Overall Results | Sensitivity 93.23% | Specificity 91.28% | Matthews Correlation Coefficient 0.844 |
|-----------------|-----------------|---------------------|---------------------|---------------------------------------|
| Positive        | Negative        |                     |                     |                                       |
| TP : 234        | FP : 15         |                     |                     |                                       |
| TN : 157        | FN : 17         |                     |                     |                                       |
| Inconclusive : 6|                 |                     |                     |                                       |
| Positive Prediction Rate | 93.98% | Negative Prediction Rate | 90.23% | | | | |

Figure 8: Summary of the results from the Melanoma classifier.
5 Breast Cancer

Artificial Neural Network and Gradient Descent were used in this part of the study.

5.1 Feature Selection

The features were selected from the Wisconsin Breast Cancer Database[28].

| Feature                  | Details                                                      |
|--------------------------|--------------------------------------------------------------|
| Clump Thickness          | Assesses if cells are multi or monolayered                   |
| Uniformity of Cell Size  | Consistency in size of cells                                 |
| Uniformity in Cell Shape | Estimation of equality of cell shapes and identification of marginal variances |
| Marginal Adhesion        | Quantify the number of cells on the outside of epithelial that stick together |
| Single Epithelial Cell Size | Determines if epithelial cells are significantly enlarged   |
| Bare Nuclei              | Quantify the proportion of cells that are not surrounded by cytoplasm to those that are |
| Bland Chromatin          | Rates uniformity of texture of nucleus                       |
| Normal Nucleoli          | Determines whether nucleoli are small or large, barely visible or more visible |
| Mitoses                  | The level of mitotic activity                                |

5.2 Gradient Descent Development

A normalized gradient descent was developed in Python with custom libraries. Optimal results appeared at (number of iterations) \( T = 5000 \) and \( \sigma = 2 \cdot 10^{-9} \).

5.3 Artificial Neural Network Development

A custom made ANN consisting of 7 layers was coded in Octave to process the dataset and achieve optimal results [25].
5.4 Results

Both the algorithms that were designed for diagnosing breast cancer delivered a high level of accuracy. The ANN rose in accuracy from 87% to around 95% on increase in training size. The gradient descent delivered an accuracy of 91% after 5000 iterations. The false positive and false negative rate in this case, was low too (Figure 9, Figure 10).

A breakup of 55 : 45 was used for training and testing of our algorithm. The data presented here is from the testing period of the algorithm.

![Accuracy using an ANN](image)

Figure 9: Comparison of the training data size and the accuracy of the algorithm.

![Accuracy using Normalized Gradient Descent](image)

Figure 10: Comparison of the number of iterations and the accuracy of the algorithm.
The performance of the algorithms have been summarised (Figure 11).

![Table 1](image1.png)

Figure 11: Summary of the results from the Breast Cancer Classifier.

### 6 Conclusion

The algorithms achieved a high accuracy: The ANN used for detecting melanoma achieved an accuracy of 93%. The ANN and the Gradient Descent used for detecting breast cancer achieved an accuracy of 95% and 91% and the SVM used for detecting lung cancer achieved an accuracy of 94%. These algorithms also succeeded in achieving a very low false positive and false negative rate, indicating that our experiments were a success. All the algorithms used in this study were self developed. The ever learning nature of the algorithms make it possible for them to achieve near perfect accuracy on an increase in training size - something that an proposed online network facilitate.

Image processing and image analysis was used to obtain data directly from the CT scans and the skin images in the cases of lung cancer and melanoma. It was decided to go forward with using the WBC dataset because it has been historically proven to be reliable. Images of the melanoma skin lesions as well as CT scans were taken from online databases (primarily PH2 dataset and NCBI cancer imaging archive). We feel that the image processing algorithms gave us an advantage by allowing us to directly obtain data from the images. The success of our experiments also results due to our selection of features and methods for their calculations which were based on our exhaustive research of the work done in the past.

Although previous work has succeeded in building computer aided systems for diagnosis (CAD) of diseases, CAD of cancer has only recently
gained traction. The results achieved in this study are comparable to the seminal works in this field. In addition to the well-performing algorithms, image processing and analysis algorithms and feature selections, the construction of a naive mobile application in the case of melanoma diagnosis is something that people will greatly benefit from. The mobile application allows people directly access and take diagnostic tests on their mobile phones, hence eliminating costs associated with logistics and consultation. This is a boon in a world where consultation fees are sky-rocketing. In addition to this, a mobile application makes it feasible to carry out reliable tests in resources poor areas where taking exhaustive tests is not at all possible.

On observation, the algorithms used in this study also outperform several commercial software such as SkinSeg. With virtually no commercial software available for lung cancer and breast cancer diagnosis, we further plan to develop novel software that can be readily implemented in clinics all around the world. This will facilitate a global collaborative organisation that works on diagnosis research and also uses our software to actually diagnose diseases.

In the future, we plan to develop diagnostic algorithms for other diseases. We are also working on the online diagnosis software for collaborative research.

7 Acknowledgements

We would like to thank our family and friends for their tremendous support throughout the course of this project.

Mohnish would like to thank Giovanni Parmigiani, Chair and Professor, Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute for his invaluable advice.

Tanay would like to thank Philip Klein for providing the inspiration for this project through "Coding the Matrix: Linear Algebra through Computer Science Applications" and Andrew Ng for his instructional videos on "Machine Learning" on Coursera.

Finally, we would also like to thank the University of Porto, NCBI for letting us use their database.
[1] Teresa M, Pedro M. F, Jorge Marques et al. ”PH² - A dermoscopic image database for research and benchmarking" 35th International Conference of the IEEE Engineering in Medicine and Biology Society, pp.5437-5440, July 3-7, 2013, Osaka, Japan.

[2] Tomatis S, Bono A et al. ”Automated melanoma detection: multispectral imaging and neural network approach for classification” Med Phys. 2003 Feb;30(2): 212-21.

[3] Nasser H. Sweilam et al. ”Support vector machine for diagnosis cancer disease: A comparative study” Egyptian Informatics Journal, Dec 2010(11); 81–92

[4] Harald Ganster et al. ”Automated Melanoma Recognition” IEEE Transactions on Medical Imaging, March 2001 (20.3)

[5] http://www.skincancer.org. Facts and Statistics. Web

[6] Fikret Ercal et al. ”Neural Network Diagnosis of Malignant Melanoma From Color Images” IEEE Transactions on Biomedical Engineering, (41.9)

[8] Ayman El-Baz et al. 2002. ”Visualization and identification of Lung Abnormalities in Chest Spiral CT Scan: Phase-I” International Conference on Biomedical Engineering, Cairo, Egypt

[9] Armato, S.G., M.L. Giger and H. MacMahon, 2001. ”Automated detection of lung nodules in CT scans: Preliminary results” Med. Phys., 28: 1552-1561.

[10] Gurcan, M.N., B. Sahiner, N. Petrick, H. Chan and E.A. Kazerooni, et al. . 2002. ”Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer aided diagnosis system” Med. Phys., 29: 2552-2558.

[11] Hara, T. et al. 1999. ”Automated Lesion Detection Methods for 2D and 3D Chest X-Ray Images” Proceedings of the 10th International Conference on Image Analysis and Processing, Sep 27-29, Venice, Italy, pp: 768-773.

[12] Penedo, et al. 1998. ”Computer-aided diagnosis: a neural-network based approach to lung nodule detection” IEEE Trans. Med. Image., 17: 872-880. ISSN: 0278-0062

[13] http://www.cancer.org. Facts and Statistics. Web
[14] Ruey-Feng Chang, PhD, Wen-Jie Wu, MS et al. "Support Vector Machines for Diagnosis of Breast Tumors on US Images"

[15] Qeethara Kadhim Al-Shayea. 2011. "Artificial Neural Networks in Medical Diagnosis" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2

[16] T. Jayalakshmi and Dr. A. Santhakumaran. 2010. "Improved Gradient Descent Back Propagation Neural Networks for Diagnoses of Type II Diabetes Mellitus" Global Journal of Computer Science and Technology Vol. 9 Issue 5 (Ver 2.0),

[17] Filippo Amato, Alberto López. 2013. "Artificial neural networks in medical diagnosis" Journal of Applied Biomedicine Volume 11, Issue 2, Pages 47–58

[18] American Lung Association. Lung Cancer Fact Sheet. Web

[19] R.-E. Fan, P.-H. Chen, and C.-J. Lin. "Working set selection using the second order information for training SVM" Journal of Machine Learning Research 6, 1889-1918, 2005

[20] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 18.

[21] William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

[22] Tidke, Chakkarwar. "Classification of Lung Tumour Using SVM" IJCER, Vol 2 (5).

[23] O. L. Mangasarian, R. Setiono, and W.H. Wolberg. 1990. "Pattern recognition via linear programming: Theory and application to medical diagnosis in Large-scale numerical optimization”, Thomas F. Coleman and Yuying Li, editors, SIAM Publications, pp 22-30.

[24] K. P. Bennett O. L. Mangasarian. 1992. "Robust linear programming discrimination of two linearly inseparable sets”, Optimization Methods and Software 1, 23-34 (Gordon Breach Science Publishers).

[25] Rudy Setiono. "Extracting rules from pruned neural networks for breast cancer diagnosis"
