An Evaluation of Machine Learning Techniques to Diagnose Certain Types of Cancer in Early Stages

Sachdeva, Saarthak
saarthaksachdeva@icloud.com

Anand, Mayank
mayank.anand1998@gmail.com

October 10, 2015

1 Abstract

Conventional cancer detection techniques have proven to be ineffective in early stage detection for most types of cancer. Previous studies indicate that about 60% of all cancerous tumors are detected in later stages, which significantly reduces the survival probability of a patient. The objective of this paper is to provide a detailed evaluation of different machine learning techniques to detect different types of cancer in early stages using publicly available data containing a number of extrinsic parameters, and to provide details of development of an algorithm based on Convolutional Neural Networks to perform automated classification of multiple types of cancer (Melanoma, Squamous and Basal Cell Carcinomas, Retinoblastoma, Lymphoma, Breast Cancer, and Glioma) as benign or malignant to a high degree of accuracy in an efficient way.

2 Introduction

An average of 51% [1] of cancer cases are diagnosed after the local stage (in later stages, after the cancer cells have metastasized). Many types of cancer are almost completely treatable in early stages, but the probability of a patient surviving for five years after diagnosis decreases [1] drastically with each stage. [2] [3] [4] It is estimated that about 3.7 million lives can be saved every year by early diagnosis of cancer. Machine learning is a promising branch of computer science that, out of a lot of other applications, identify patterns. Previous studies have shown that certain types of cancer show minute, almost indistinguishable, but consistent symptoms in early stages, which a machine learning algorithm can be trained to identify as a pattern. It can be utilised for predicting the probability of a new case being malignant or benign based on appropriate data input, with a high degree of accuracy. There are many machine learning algorithms applicable to this scenario, and it is important to analyze and understand the most optimal one, and utilize the same for development to provide a consistent, reliable, and efficient diagnosis method, which is undertaken in the first part of "Experimentation".

3 Research

Background Research for this project was divided into two major sections

- Current Cancer Detection Methods
- Analysis of Machine Learning Techniques for Cancer Diagnosis

3.1 Exploring Current Diagnosis Methods

3.1.1 Melanoma

It is a type of skin cancer caused by uncontrollable melanocyte division. The tumors are similar to benign moles in appearance, which delays diagnosis.

Dermatologists use a method known as the A.B.C.D. Principle, i.e. Asymmetry, Border Irregularity, Color Variation, and Diameter to detect if a tumor is malignant [5]. If a mole is satisfies a few of the given criteria, a biopsy is recommended for confirming diagnosis.

Even though this method is fairly reliable because of the distinct recurring patterns, many cases of melanoma go undetected because of the difficulty in differentiating a malignant tumor from a benign one with the naked eye or a Dermatoscope.

Distinctive Features: Distinctive Asymmetry and Irregularity

3.1.2 Squamous Cell Carcinoma / Basal Cell Carcinoma

Despite their different external appearances, the conventional pre-biopsy diagnosis methods for basal cell carcinoma and squamous cell carcinoma are largely similar, and primarily based on visual and dermatoscopic inspection of a lesion on the skin. Squamous Cell Carcinoma appears as a non-healing ulcer or lesion on the skin, initially a small nodule, that enlarges over time, developing into a pink ulcer with hard, keratinous edges. Basal Cell Carcinoma also develops into an ulcerous tissue that spreads very slowly.

Reliable SCC/BCC detection is extremely hard without performing an invasive procedure like a biopsy, because of lack of human-recognisable patterns, and wildly varying appearances.

Distinctive Features: Tumors are pearly and translucent in early stages with indurated and pale edges and a red, ulcerated center.

3.1.3 Retinoblastoma

It is a type of ocular cancer that mostly occurs in children, leading to vision loss in thousands of children every year.[6]

It is difficult to detect without a genome sequencing test that tests the presence of the RB1 gene, which is both inefficient and expensive. Normally detected after observing a layer over ocular blood vessels, or reports of ocular discomfort. In both cases, it is often found that the cancer has progressed enough to cause significant damage to the patient's eyes. There are no widely recommended screening tests, because retinoblastoma is a relatively rare cancer. But in early stages, in a bright light, the pupil of the eye almost always appears white or pink. Combined with factors like discomfort or vision problems, it can serve as an accurate metric for early stage diagnosis. **Distinctive Features:** Photographs of the eye with flash will appear different.

3.1.4 Breast Cancer

Breast cancer is cancer that develops from breast tissue. Hard to detect in early stages, but easy to cure if detected early.

Most types of breast cancer show distinctive visual symptoms, along with tactile symptoms which can be a lump in the breast, rashes, pain in the breast area etc. Current methods for diagnosis - mammograms, X-Rays, etc. all are performed to check for lumps which are definitive indicators of a tumor.

Distinctive Features: Change in shape of breast, presence of lump, texture change, breast pain, armpit lymph nodes visibility

3.1.5 Lymphoma

It's a name applied to a group of blood cell tumors that develop from lymphatic cells.

It's primary symptom is swelling of lymph nodes, along with other secondary symptoms like fever. Because of the relatively common symptoms, it is hard to diagnose until it has progressed significantly. Currently, after visual prognosis, a biopsy (core needle, excisional or incisional) or peritoneal fluid sampling is recommended to confirm diagnosis.

Distinctive Features: Size of swollen lymph node, number of swollen nodes, location of swollen lymph nodes

3.1.6 Brain Tumor

A type of tumor that occurs upon abnormal cell division within the brain, and may be malignant or benign. Glioma and Glioblastoma are two of the most common types of malignant brain tumors and have definite characteristics which can help in early diagnosis.

Conventionally, MRI, CT, CAT, Electroencephalography are the recommended procedures after a prognosis is made by studying a patient's cognitive / brain functioning. The results of these are studied manually, which may result in failure to detect the tumor in a local stage.

Distinctive Features: Aberrations in Angiogram, EEG, or CT Scan, self and third person reports of cognitive and motor functioning.

3.2 Evaluation of different machine learning techniques for cancer diagnosis

For the purposes of this project, the following machine learning techniques were studied and evaluated.

- Support Vector Machine
- Convolutional Neural Network
- Decision Tree
- Bayesian Network
- k-Nearest Neighbor Clustering

Apart from these, sliding window based recurrent neural networks and fuzzy logic based classifiers were also evaluated but were found to be inappropriate for this application based on the receiver operating characteristics (ROC).

4 Experimentation

4.1 Selection of optimal classification features

For each type of cancer considered, feature vectors are constructed using results from previous studies as well as localized featured descriptors extracted using a Single Layer Perceptron (SLP). The following numbers of features were used to train the algorithms (with the exception of the convolutional neural network):

- **Melanoma:** 14
- **Squamous Cell Carcinoma:** 12
- **Basal Cell Carcinoma:** 12
- **Retinoblastoma:** 7
- **Lymphoma:** 8
- **Breast Cancer:** 10
- **Brain Tumor:** 5

The entire feature selection process is the subject of a separate paper.

4.2 Selection of an optimal ML algorithm

After research, we realized that using characteristics commonly used by oncologists to detect signs of cancer as data points for classification datasets was not ideal, since sometimes features tended to vary greatly. Therefore, we implemented a Deep Belief Network was implemented with contrastive divergence to select ideal classification features for each type of cancer amongst the given set.

The machine learning techniques mentioned in the previous section were implemented in Torch and trained on a dataset of features for each type of cancer, which was split into a 70:25:5 Training:Validation:Testing ratio, and confusion matrices for each of them were computed

4.3 Development of the EarlyDetect platform

For each type of cancer, certain pre-prognosis processes that aid feature extraction were carried out

- **Melanoma:** Gaussian Blurring and Morphological Closing
- **Squamous Cell Carcinoma:** Periodic feature elimination by Fourier transform
- **Basal Cell Carcinoma:** Periodic feature elimination by Fourier transform
- **Retinoblastoma:** Brightness and contrast adjustments, minor feature elimination by morphological closing
- **Lymphoma:** Minor input normalization using a Gaussian distribution

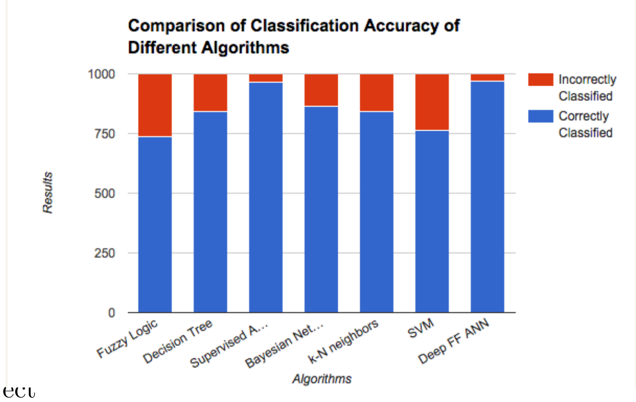

Figure 1: Comparison of Different Machine Learning Techniques

- **Breast Cancer:** Periodic feature elimination by Fourier transform, input normalization
- **Brain Tumor:** Minor input normalization and noise filtering on EEG data

After determining that a convolutional neural network was an ideal machine learning technique for diagnosis of cancer, eight 15 layer deep convolutional neural networks on the following specifications were implemented in Torch

The neural networks were deployed as a unified API for earlyDetect, which is used by the smartphone and desktop apps. Using hive computing, it is planned that resources from idle computers across the world will be utilized for running and retraining the large neural networks.

5 Results

The results shown in 1 were obtained after experimentation with different types of machine learning algorithms.

As clearly evident from the data, a convolutional neural network (labeled Deep FF ANN) provided the best classification accuracy for the melanoma dataset. Other datasets too showed very similar results.

After implementation of the deep convolutional neural network, the classification accuracy results shown in 2 were obtained for each type of cancer.

Contrary to expectations, the neural network showed abnormally high classification accuracy for Melanoma, with an extremely low (0.001%) false

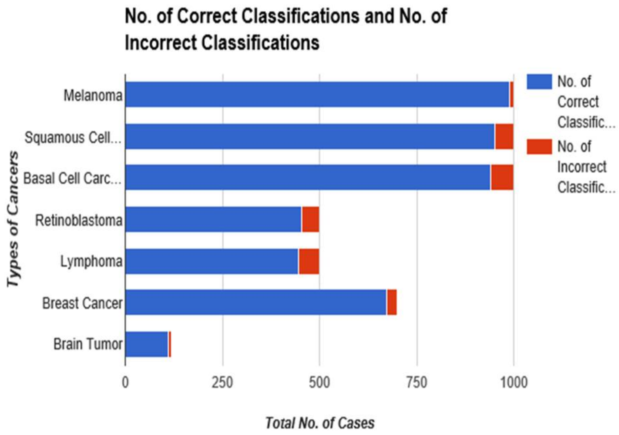

Figure 2: Classification Accuracy for All Cancer Types

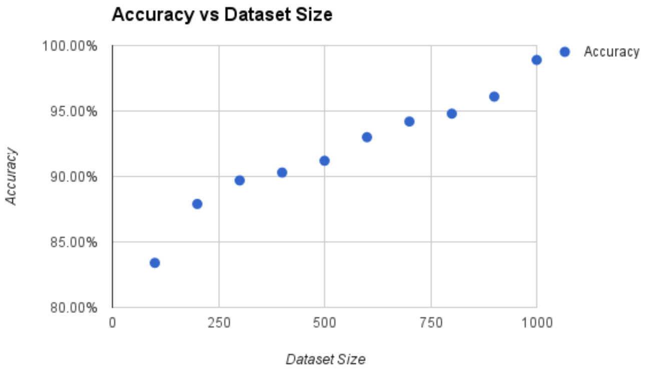

Figure 3: Correlation of training dataset size and classification accuracy

positive rate. For all types of cancer considered, we were able to achieve a classification accuracy of more than 85%, with very low false positive and false negative rates (1% on average)

Figure 3 shows a graph between training dataset size and classification accuracy, which shows a clear, nearly linear correlation between the two quantities.

The average diagnosis time was found to be 110.7s, with the maximum time taken for breast cancer and lymphoma classifications, at an average of 290s, and the minimum taken for melanoma classification at less than 30s. As the dataset size increases, the diagnosis time grows logarithmically and plateaus about 10000 samples for specific type of cancer

6 Conclusion

We were able to successfully identify an efficient machine learning technique for early diagnosis of cancer, and were able to implement it successfully in an open source platform (earlyDetect). The following conclusions were drawn from the data.

- The average diagnosis time was considerably lower than conventional diagnosis methods, at 110.7s
- The average accuracy is higher than the accuracy of conventional methods, at 93.9%
- As expected, there is a clear correlation between training dataset size and classification accuracy. This implies that the algorithm can be expected to become more accurate as more diagnoses are added to it's dataset.

References

[1] Cancer Facts and Figures 2015, American Cancer Society.

[2] Breast cancer survival rates, by stage, American Cancer Society.

[3] James A Keir et al. 2007. Outcomes in Squamous Cell Carcinoma with Advanced Neck Disease

[4] S McPhail, S Johnson, D Greenberg, M Peake and B Rous. 2015. Stage at diagnosis and early mortality from cancer in England

[5] Naheed R. Abbasi, Helen M. Shaw, Darrell S. Rigel, Robert J. Friedman, William H. McCarthy, Iman Osman, Alfred W. Kopf, David Polsky. 2004. Early Diagnosis of Cutaneous Melanoma, Revisiting the ABCD Criteria

[6] DJ DerKinderen et al. 1985. Early diagnosis of bilateral retinoblastoma reduces death and blindness

[7] Yann LeCun, John Denker. 1993. Convolutional Neural Networks