Cancer Prediction using Machine Learning Algorithms

Anh Dang
aqdang16@earlham.edu
Department of Computer Science
Earlham College
Richmond, Indiana

ABSTRACT

The popularity of machine learning and artificial intelligence is increasing as those are now being applied to various fields but not only in pure Computing and Computer Science. In the last 15 years, machine learning has been used in Medical and Healthcare to study and analyze the large amount of healthcare data such as medical image, electronic health record, and patient genomics. However there are still few applications of those data that directly improve the health service for people. This paper proposes an analyzation of machine learning algorithms by using different models to predict the risks of having cancer. Different algorithms will have different accuracy based on the data type and configure. Models after training will be evaluated using cross-validation over the training data set.

KEYWORDS

machine learning, neural networks, disease prediction

1 INTRODUCTION

Over the past decades, cancer has been one of the diseases which causes most death out to people around the world. There were an estimated 18 million cancer cases around the world in 2018, of which 9.5 million cases were in men and 8.5 million in women [4]. In the United States, in 2016, cancer causes 22.5% of death for male and 21.1% for female [5]. In 2019, there is an estimation that there will be 1,762,450 new cancer cases diagnosed and 606,880 cancer deaths in the United States. With this growing global burden, prevention of cancer has been one of the most significant public health challenges of the 21st century. Predicting cancer can help people to prevent the disease and reduce the number of deaths cause by cancer.

With the rapid development of machine learning, it is possible to use machine learning algorithms to predict the risk of having a particular disease. Payan et al. used deep learning methods to build a convolutional neural networks that can predict the Alzheimer's disease status based on an MRI scan of the brain [10]. In the cancer area, Mobadersany et al. proposed a method of predicting cancer outcomes from histology and genomics using convolutional networks [9].

To further make use of medical data and make an effort to improve healthcare service, in this paper, I propose a project of analyzing models to predict risk of having cancer base on numerical data and medical image data. From that analyzation, a system will be developed where people can know their risk of having cancer with inputs such as images or numerical data similar to the training data set. Each model will be built with different machine learning algorithms and train with the appropriate data type.

2 RELATED WORKS

In this section, I reviewed the algorithms in diseases prediction in general, and papers that focused on cancer prediction. Some background about machine learning algorithms will also be introduced.

2.1 Disease Prediction

Machine Learning techniques have been used in disease prediction in various areas, including plant diseases and human diseases.

Kaundal et al. introduced a new disease prediction approach based on support vector machines for developing weather-based prediction models of plant diseases [7]. They fed the dataset into various machine learning algorithms and looked at the results to decided what is the best methods. The author had compared multiple regression, artificial neural network, feed-forward backpropagation neural network, generalized regression neural network, support vector machines and support vector regression. The results were verified using cross validation. They stated that within the neural networks, the generalized regression neural network outperformed the backpropagation neural networks by about 20–30%. Also, the higher predictive accuracy by latest machine learning techniques like support vector machines will generate more efficient prediction models.

Austin et al. used various machine learning algorithms in examining classification of heart failure subtypes [1]. They predicted the probability of having heart failure with preserved ejection fraction. The methods were used for classification were classification trees, bagged classification trees, random forests, boosted classification trees, and support vector machines. The authors used logistic regression, regression trees, bagged regression trees, random forests, and boosted regression trees for prediction. They found that that those modern classification methods had better performance over conventional classification trees for classifying heart failure patients according to disease subtype. When predicting, conventional regression trees had lower predictive accuracy compared with all other methods and logistic regression had the best predictive accuracy.

Chen, et al. streamlined machine learning algorithms for effective prediction of chronic disease outbreak in disease-frequent communities [2]. After using latent factor model to reconstruct the missing data, the authors proposed a new convolutional neural network (CNN)-based multimodal disease risk prediction algorithm using structured and unstructured data from hospital. They also compared their algorithms with other typical prediction algorithms and their new algorithms had a higher accuracy with converge speed. The author concluded that their new proposed algorithm had an accuracy of 94.8% and a convergence speed which is faster than that of the CNN-based unimodal disease risk prediction algorithm.

Although many machine learning algorithms were used in these papers, neural network and its variations were the most popular and have the best accuracy. Therefore, in this proposal, artificial neural network is one of the machine learning models that will be implemented for cancer prediction.

2.2 Cancer Prediction

The trend is also similar in specific cancer prediction that neural network and its variations are still widely used and have a high enough accuracy.

Cruz et al. reviewed published studies employing machine learning methods which use in cancer prediction and prognosis [3]. In the paper, several machine learning algorithms were used such as decision trees, naive bayes, k-nearest neighbor, neural network, support vector machines and genetic algorithm. After analysing those methods, the authors had summarized their benefits, assumptions and limitations. Using that information, they explain, compare and assess the performance of different machine learning that are being applied to cancer prediction and prognosis. Cruz et al. mentioned that artificial neural network was still predominate but there was evident that a growing variety of alternate machine learning strategies were being used and that they were being applied to many types of cancers to predict at least three different kinds of outcomes. They also believed that machine learning methods would generally improve the performance or predictive accuracy of most prognoses, especially when compared to conventional statistical or expert-based systems.

Korou et al. discussed the concepts of machine learning while outlining their application in cancer prediction and prognosis [8]. With three case studies, the authors examined several machine learning methods such as artificial neural network, decision trees, support vector machines and naive bayes. The results showed that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and classification can provide promising tools for inference in the cancer domain. Korou et al. showed the accuracy of different machine learning methods ranging from 68% to 100% (small number of patients). They found that among the most common applied machine learning algorithms relevant to the prediction outcomes of cancer patients, support vector machines and artificial neural network were widely used. However, the choice of the most appropriate algorithm depends on many parameters including the types of data collected, the size of the data samples, the time limitations as well as the type of prediction outcomes.

There are some others models that can be taken into consideration and will also be implemented in this proposal which are Support Vector Machine, Naive Bayes, Classification Tree and Genetic Algorithm.

3 METHODOLOGY

There will be two parts, prediction using numerical data and prediction using image data (see Figure 1). First, the data sets will be described and processed before training. After that, data sets will be fit in various models to produce prediction.

3.1 Numerical data

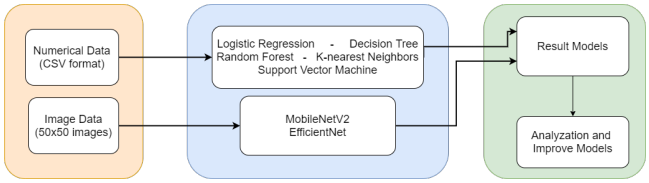

Figure 1: Design framework of the project.
Red: Data Blue: Models Green: Results

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280

Figure 2: Sample of numerical data set.

3.1.1 Data set and preprocessing. The numerical data set is the Breast Cancer Wisconsin (Diagnostic) Data Set as shown in Figure 2 [18]. There are 33 columns with features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The labels are M (malignant) and B (benign).

For numerical data set, process includes cleaning the data, dropping irrelevant columns, replacing missing values with appropriate values (mean, zero) depends on each feature.

3.1.2 Models and training. We will use Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machines, and K-Nearest Neighbors as classifying models for the numerical data set. The training process is written on Python with models import from Scikit Learn library.

First, we find the correlation between the target column "diagnosis" and other features. Then we choose features with correlation greater than 0.6 to be training features. Models will be trained and test by cross-validation with 5 folds.

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes [15]. In Machine Learning, it is an algorithm which is used for the classification problems, a predictive analysis algorithm and based on the concept of probability. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.

Decision tree classifier is a predictive modeling approach used in statistics, data mining and machine learning [13]. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). The target variable will take a discrete set of values. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [16]. Random decision forests correct for decision trees' habit of overfitting to their training set.

Support Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis [17].

K-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression [14]. The input consists of the k closest training examples in the feature space. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

3.2 Image data

Figure 3: IDC negative [6]

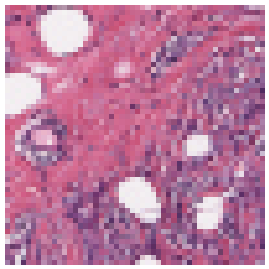

Figure 4: IDC positive [6]

3.2.1 Data set and preprocessing. The image data set is the Breast Histopathology Images Dataset with samples shown in Figure 3 and Figure 4 [6]. The data set has image of Invasive Ductal Carcinoma (IDC) which is the most common subtype of all breast cancers. The data set contains 277,524 patches of size 50 x 50 (198,738 IDC negative and 78,786 IDC positive).

For the image data set, processing steps include cropping, resizing and normalizing the image.

3.2.2 Models and training. We will use Artificial Neural Networks (ANN) and more specificly, MobileNetv2 and EfficientNet to train the image data set. The models are written using PyTorch library based on there papers [11, 12]. Each model is trained for 90 epochs with decreasing loss every 5 epochs.

MobileNetV2 is a family of general purpose computer vision neural networks designed with mobile devices in mind to support classification, detection and more [11]. The ability to run deep networks on personal mobile devices improves user experience, offering anytime, anywhere access, with additional benefits for security, privacy, and energy consumption. MobileNetV2 pushes the state of the art for mobile visual recognition including classification, object detection and semantic segmentation.

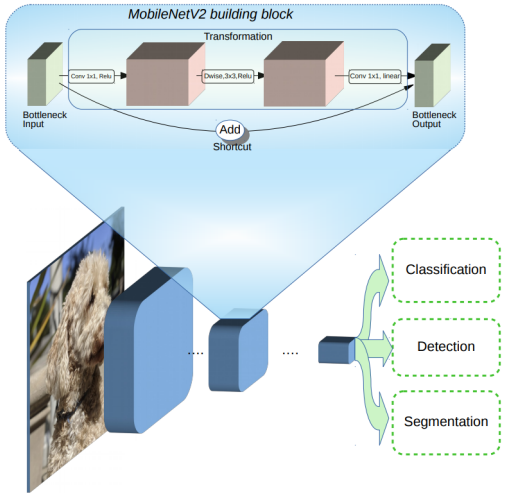

Figure 5: MobileNetV2 [11]

Input	Operator	t	c	n	s
$224^2 \times 3$	conv2d	-	32	1	2
$112^2 \times 32$	bottleneck	1	16	1	1
$112^2 \times 16$	bottleneck	6	24	2	2
$56^2 \times 24$	bottleneck	6	32	3	2
$28^2 \times 32$	bottleneck	6	64	4	2
$14^2 \times 64$	bottleneck	6	96	3	1
$14^2 \times 96$	bottleneck	6	160	3	2
$7^2 \times 160$	bottleneck	6	320	1	1
$7^2 \times 320$	conv2d 1x1	-	1280	1	1
$7^2 \times 1280$	avgpool 7x7	-	-	1	-
$1 \times 1 \times 1280$	conv2d 1x1	-	k	-	

Figure 6: Architecture of MobileNetV2 [11]

EfficientNet is a scaling method that uniformly scales all dimensions of depth/width/resolution using a simple yet highly effective compound coefficient. It will not only focus on improving the accuracy, but also the efficiency of models. In this paper, we are scaling up MobileNets.

4 RESULT AND EVALUATION

After training, the models can be used to test prediction. The numerical data set will be tested using cross-validation with 5 folds so it is a division of 4 to 1 for train set and test set. The image data set will be tested by splitting the original data set to train set and test set by 80% and 20% respectively.

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x5	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	28×28	112	3
7	MBConv6, k5x5	14×14	192	4
8	MBConv6, k3x3	7×7	320	1
9	Conv1x1 & Pooling & FC	7×7	1280	1

Figure 7: Architecture of EfficientNet [12]

4.1 Numerical Data

The results of Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), K-nearest Neighbors (KNN) and there cross-validation scores are shown in the following table.

	Accuracy	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
LR	96.231	95.000	95.625	95.000	94.984	94.975
DT	100.000	90.000	92.500	94.167	95.309	95.234
RF	96.482	93.750	93.750	94.583	94.988	94.978
SVM	92.211	91.250	90.625	90.417	91.230	91.972
KNN	94.975	93.750	94.375	92.917	93.422	93.978

Decision Tree has 100% accuracy but it is overfitting because the cross-validation scores are only about 92-93%. Logistic Regression has a really good acuracy and high cross validation score. Random Forest is also a good model with an accuracy of 96.5% and about 94.5% for the average cross-validation score. Support Vector Machine and K-nearest Neighbors have accuracy above 90% but not as good as the other models.

4.2 Image Data

The training and testing accuracy of the image data set using MobileNetV2 (MB) and EfficientNet (EN) are shown in the following table.

	Training Accuracy	Testing Accuracy
MobileNetV2	96.094	92.352
EfficientNet	95.312	91.021

Both models has high accuracy above 90%. MobileNetV2 is slightly better than EfficientNet in both training accuracy and testing accuracy.

5 DISCUSSION

For numerical data's models, we use GridSearchCV to tune the hyperparameters of each model. GridSearchCV will take input of our model and a dictionary of all possible values for each hyperparameter and return the hyperparameter setting for the best possible model.

	Accuracy	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
DT	96.231	93.750	95.000	94.583	94.672	94.472
RF	97.236	92.500	93.125	94.167	94.992	94.728
SVM	96.734	93.750	94.375	95.000	95.934	95.987
KNN	94.975	93.750	94.375	92.917	93.422	93.978

With the tuned hyperparameter, Decision Tree model is not overfitting with lower accuracy but better cross-validation score. Random Forest has better and the highest accuracy of 97.236%. Support Vector Machine also has a much higher accuracy of 96.734%. K-nearest Neighbors remains the same as the original hyperparameter is the best parameter for this model.

For image data's models, we tried two optimizer Adaptive Moment Estimation (Adam) and Stochastic Gradient Descent (SGD) and the combination of both with Adam for the first few epochs and SGD for later epochs. The result is that using only SGD as we did in ou method will give the best accuracy (> 90%) compare to only Adam (83%) and combination of Adam and SGD (85-87%).

This project aims to provide a fast and simple way for people to check if they are possible of having cancer without the need of going to hospitals. Therefore, cancer can be diagnosed early and people will have a higher chance to survive.

6 FUTURE WORK

We aim to combine our two models, numerical and image, by using another model which will produce the numerical data we need given the image. Therefore, we can predict cancer with just image using both models and the confidence level will be higher.

If there are enough data, we want to expand our models to predict other types of cancer but not just breast cancer.

7 ACKNOWLEDGEMENTS

This project was supported by the Earlham College Department of Computer Science as the Senior Capstone.
I would like to thank Dr. Charlie Peck and Dr. Igor Minevich for providing detailed feedbacks and helping me finish this project.

REFERENCES

[1] Peter C Austin, Jack V Tu, Jennifer E Ho, Daniel Levy, and Douglas S Lee. 2013. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of clinical epidemiology* 66, 4 (2013), 398–407.
[2] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang. 2017. Disease Prediction by Machine Learning Over Big Data From Healthcare Communities. *IEEE Access* 5 (2017), 8869–8879. https://doi.org/10.1109/ACCESS.2017.2694446
[3] Joseph A Cruz and David S Wishart. 2006. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics* 2 (2006), 117693510600200030.
[4] World Cancer Research Fund. 2019. *Worldwide cancer data, Global cancer statistics for the most common cancers.* https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data
[5] Melonie Heron. 2016. Deaths: Leading Causes for 2016.
[6] Kaggle. 2018. *Breast Histopathology Images.* https://www.kaggle.com/paultimothymooney/breast-histopathology-images
[7] Rakesh Kaundal, Amar S Kapoor, and Gajendra PS Raghava. 2006. Machine learning techniques in disease forecasting: a case study on rice blast prediction. *BMC bioinformatics* 7, 1 (2006), 485.
[8] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. 2015. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal* 13 (2015), 8–17.

[9] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. 2018. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences* 115, 13 (2018), E2970–E2979.

[10] Adrien Payan and Giovanni Montana. 2015. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. *arXiv preprint arXiv:1502.02506* (2015).

[11] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[12] Mingxing Tan and Quoc V Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946* (2019).

[13] Wikipedia. 2019. *Decision Tree Learning*. https://en.wikipedia.org/wiki/Decision_tree_learning

[14] Wikipedia. 2019. *k-nearest neighbors algorithm*. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[15] Wikipedia. 2019. *Logistic regression*. https://en.wikipedia.org/wiki/Logistic_regression

[16] Wikipedia. 2019. *Random forest*. https://en.wikipedia.org/wiki/Random_forest

[17] Wikipedia. 2019. *Support Vector Machine*. https://en.wikipedia.org/wiki/Support-vector_machine

[18] Olvi L. Mangasarian William H. Wolberg, W. Nick Street. 1995. *Breast Cancer Wisconsin (Diagnostic) Data Set*. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)