Final project report

Using four different models to predict the risk of breast cancer

Dan Xi, Pid: 4884434, Group9

1. Project Definition

1.1 Motivation

The diet of modern life is irregular, the work pressure is high, genetically modified foods, and cancer is commonplace in our lives. Breast cancer is the most frequently encountered cancer type in women and the second most terminal one after lung cancer. Due to the clinical experience of the doctors based on the doctor's work experience, relatively young doctors are not experienced enough to find early symptoms. It is very helpful if employed the artificial intelligent technique to assist doctors in diagnosing patients.

In this project, we used four method to diagnose patients who had breast cancer. In order to train the model, we use the data from UC Irvine "Breast Cancer Wisconsin" data set. In this dataset, there are 569 entries, with 357 benign cases and 212 malignant cases. Each column contains 33 features, but there are some useless features. After comparison study, we found that Deep Learning could achieve better result without feature selection.

Though these models cannot diagnose cancer conclusively, it could help physicians in deciding whether a biopsy is required by providing information about whether the patient has breast cancer or not. Using our models can help doctors prepare for diagnosis to provide effective information and improve the cure rate of early patients. Studies have shown that using machine learning methods to help doctors diagnose can give a more accurate result.

1.2 Problem Statement/Project Overview

The traditional method needs to check various physical indicators through instruments, combined with some external conditions (such as eating habits, family history), plus the clinical diagnosis experience of the doctor to determine whether the patient is sick.

Now we can use some models of machine learning, through the data and image information collected in advance, through training and testing of these data sets, we can get a relatively accurate prediction result, thereby helping doctors work in early diagnosis, as Patients avoid risks.

Machine learning is becoming a highly active industrial research topic; both high-tech IT companies and other traditional companies. Deep Learning is changing the way we look at technologies. There is a lot of excitement around Artificial Intelligence (AI) along with its branches namely Machine Learning (ML) and Deep Learning at the moment.[1]

Deep Learning using in Healthcare field likes Breast or Skin-Cancer diagnostics, Mobile and Monitoring Apps, or prediction and personalized medicine on the basis of Biobank-data. AI is completely reshaping life sciences, medicine, and healthcare as an industry. Innovations in AI are advancing the future of precision medicine and population health management in unbelievable ways. Computer-aided detection, quantitative imaging, decision support tools and computer-aided diagnosis will play a big role in years to come.[1]

In this project, we use four different models (SVM, Random Forest, KNN, Deep Learning) to train and test the dataset and predict the risk of breast cancer. We do features selection and put whole data to get results and do the comparison in same model.

1.3 Metrics

To make it easier to tune the model, a comprehensive metric was used to measure the performance of the model. The metric is determined by counting simulated human interventions. In this dataset, there is 33 features in total, we do the data clean and delete some. First I put the whole data in to four models to train and test. Second I do the features selection and I choose six features put to models. Compare the result of the same model when with the feature or without the feature.

2. Analysis

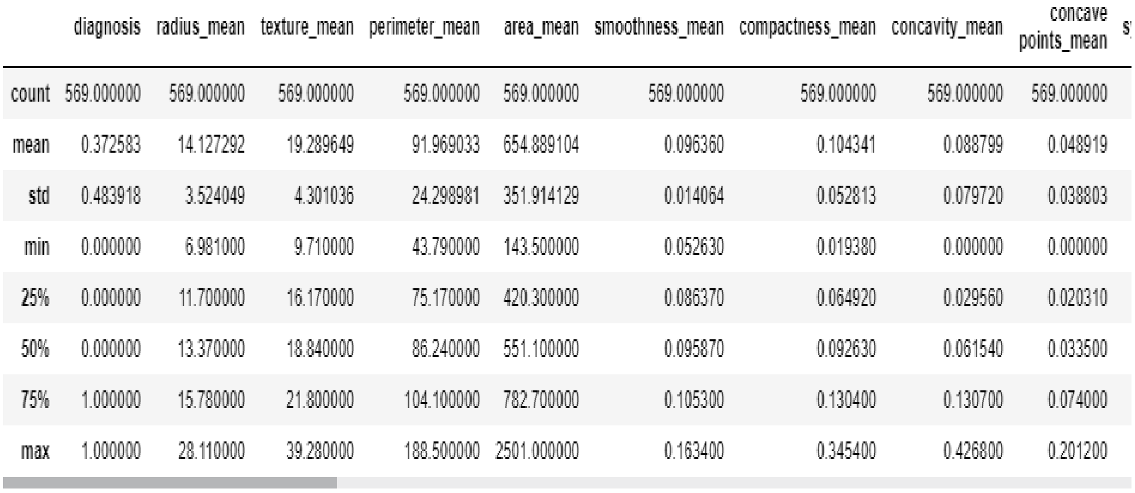
2.1 Data Exploration

We assume that the features collected from patients have enough information to diagnose breast cancer.

The data we used is from UC Irvine "Breast Cancer Wisconsin" data set. There are 569 entries, with 357 benign cases and 212 malignant cases. Each column contains 33 features. The diagnosis of breast cancer sets 0 for benign and 1 for malignant. In this step we do the data describe and data visualization.

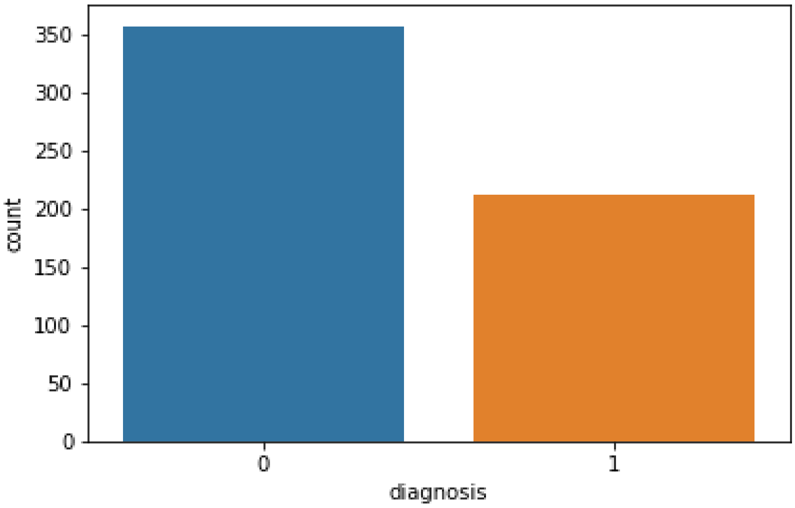
2.2Exploratory Visualization

2.2.1. Data describe:

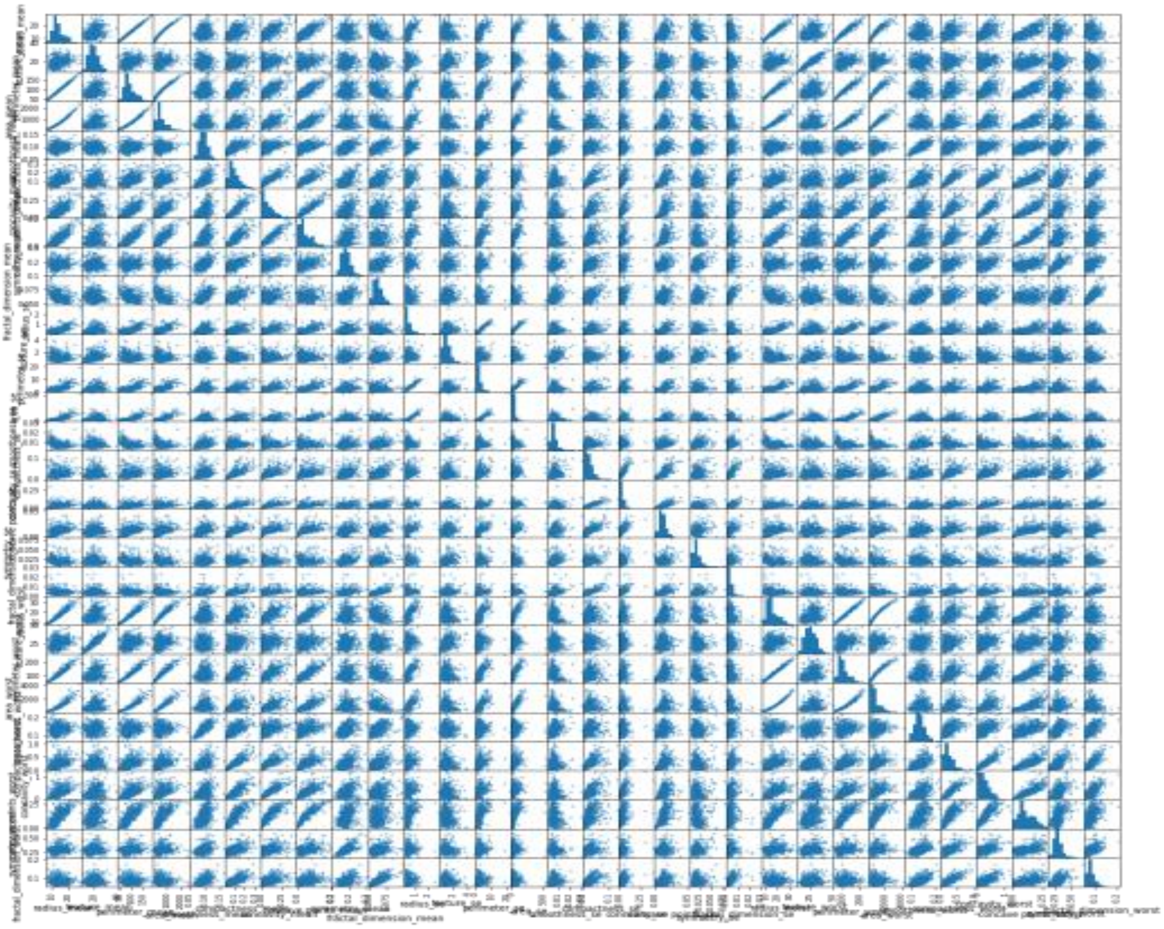


	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	0.372583	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919
std	0.483918	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803
min	0.000000	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000
25%	0.000000	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310
50%	0.000000	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500
75%	1.000000	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000
max	1.000000	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200

2.2.2Visualizing diagnostic results



2.2.3. Scatter plot matrix



2.3 Algorithms and Techniques

2.3.1Support Vector Machine

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.[2]

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

The learning of the hyperplane in linear SVM is done by transforming the problem using some linear algebra. This is where the kernel plays role. For **linear kernel** the equation for prediction for a new input using the dot product between the input (x) and each support vector (xi) is calculated as follows: f(x) = B(0) + sum(ai * (x, xi)).[3]

This is an equation that involves calculating the inner products of a new input vector (x) with all support vectors in training data. The coefficients B0 and ai (for each input) must be estimated from the training data by the learning algorithm. [3]

2.3.2.Random Forest

The random forest is a model made up of many decision trees. Rather than just simply averaging the prediction of trees (which we could call a “forest”), this model uses two key concepts that gives it the name random: Random sampling of training data points when building trees

$$I_G(n) = 1 - \sum_{i=1}^{J} (p_i)^2$$

2.3.3. KNN

K-Nearest Neighbors, or KNN for short, is one of the simplest machine learning algorithms and is used in a wide array of institutions. KNN is a non-parametric, lazy learning algorithm. When we say a technique is non-parametric, it means that it does not make any assumptions about the underlying data. In other words, it makes its selection based off of the proximity to other data points regardless of what feature the numerical values represent. Being a lazy learning algorithm implies that there is little to no training phase. Therefore, we can immediately classify new data points as they present themselves.[4]

2.3.4Deep Learning

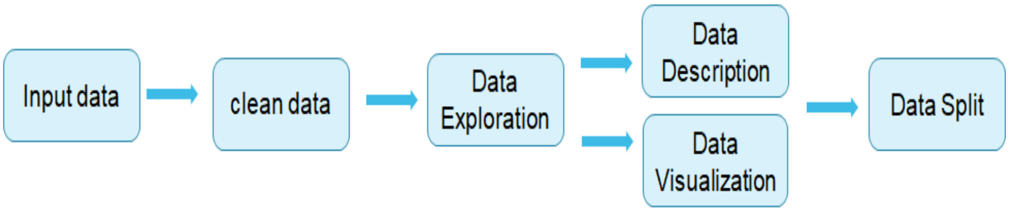
Deep learning is an increasingly popular subset of machine learning. Deep learning models are built using neural networks. A neural network takes in inputs, which are then processed in hidden layers using weights that are adjusted during training. Then the model spits out a prediction. The weights are adjusted to find patterns in order to make better predictions. The user does not need to specify what patterns to look for — the neural network learns on its own.

Deep feedforward networks, also often called feedforward neural networks, or multilayer perceptrons(MLPs), are the quintessential deep learning models. The goal of a feedforward network is to approximate some function f*. For example, for a classifier, y = f*(x) maps an input x to a category y. A feedforward network defines a mapping y = f(x; θ) and learns the value of the parameters θ that result in the best function approximation.[5]

3.Methodology

3.1 Data Preprocessing

Preprocessing Pipeline



3.2 Implementation

Hardware	1 server: 32G RAM, 16 CPU ,1080 Ti GPU
Jupyter	
Platform and Libraries	Numpy,Kras,Tensorflow,sk-learn, pandas,

3.3 Refinement

In order to refine the training process and improve the performance of our model, we take the following measures during our project.

Data preprocessing

In order to get good accuracy, we do the data preprocessing, clean the data, do the features selection. We put whole data to the model, we scaling the dataset.

4 Results

4.1 Result

Input data with feature selection

Model	Accuracy
SVM	92.98%
Random Forest Classifier	90.64%
KNN	92.40%
Deep Learning*	98.42%*

After feature selection, compare with other method, Deep Learning based method achieved the best performance. The reason behind the result is that deep learning-based method could be regarded as a feature learner.

5. Conclusion

5.1Conclsuion

In this project, we first do research survey the possible solutions to solve breast cancer diagnosing problem. In order to get the best performance, we selected four different methods: SVM, Random Forest, KNN, and Deep Learning-based method. After comparison experience, we found that performance of traditional based methods (SVM, KNN, Random Forest) depend on the feature selection, and deep learning method as end-to-end solution that is not sensitive with that.

In collusion, we do the result compassion, we can get below:

Comparison with the results, through the accuracy we can know that the Deep Learning can get better predict result than other models. I think the neural network has deep research value in the future. Because this model can handle the big data.

5.2 Future work

In the future, I will use this Deep learning model to train some bigger datasets or create more layer and compassion the result.

Reference:

[1].Top 15 Deep Learning applications that will rule the world in 2018
https://medium.com/breathe-publication/top-15-deep-learning-applications-that-will-rule-the-world-in-2018-and-beyond-7c6130c43b01

[2] Support Vector Machine — Introduction to Machine Learning Algorithms
https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[3] Chapter 2 : SVM (Support Vector Machine) — Theory, Savan Patel 2017
https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

[4] K Nearest Neighbor Algorithm In Python, Cory Maklin 2019
https://towardsdatascience.com/k-nearest-neighbor-python-2fccc47d2a55

[5] Deep Learning An MIT Press book Ian Goodfellow and Yoshua Bengio and Aaron Courville 2016
http://www.deeplearningbook.org/

[6] Wolberg, WIlliam H. Breast Cancer Wisconsin (Original) Data Set. UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set, 15 July 1994

[7] Home - Keras Documentation. (2018). Retrieved from https://keras.io/

[8] Scikit-learn Machine Learning in Python . https://scikit-learn.org/stable/

INDIVI DUAL CONTRIBUTION

Member	**Task**	**contribution**
Dan Xi	Investigation and Analysis Data preprocessing Design model approach Experiment Prepare the presentation Write the final report	**100%**