MACHINE LEARNING METHODS IN CANCER DIAGNOSIS AND PROGNOSIS

Mehrdad Zandigohar
Mechanical Engineering Department
Southern Illinois University Edwardsville
Edwardsville, Illinois 62026
December, 2019

Abstract: Cancer is one of the most dangerous diseases in the world, and still, many people cannot survive. Further, there still is no cure for cancer stages of 4, hence, early detection of cancer plays a significant role in human survival. When cancer is detected early, treatment is more likely to be successful. Recent researches on machine learning methods and algorithms in cancer diagnosis have shown that ML can help diagnose people with cancer in early stages. Machine learning methods are even used for cancer susceptibly prediction including artificial neural networks (ANNs), support vector machine (SVMs), Bayesian network (BNs), graph-based semi-supervised learning (Graph-based SSL), SSL Co-training algorithm and decision trees (DTs). The mentioned techniques have been used to model the progression and treatment of cancerous cells. ML is able to detect key features from complex datasets and in this project, a comprehensive analysis of those ML techniques will be utilized in order to find the accuracy of each method. A good picture of the link between ML algorithms and breast cancer detection would be outlined, and the significant impact of AI in dangerous disease treatments would be depicted. Each method were ranked and the important features in cancer detection was outlined.
Keywords: Cancer, Breast Cancer, Machine Learning, Supervised Learning

1. Introduction

Cancer can result from abnormal proliferation of any of the different kinds of cells in the body, so there are more than a hundred distinct types of cancer, which can vary substantially in their behavior and response to treatment. The most important issue in cancer pathology is the distinction between benign and malignant tumors. A tumor is any abnormal proliferation of cells, which may be either benign or malignant. A benign tumor, such as a common skin wart, remains confined to its original location, neither invading surrounding normal tissue nor spreading to distant body sites. A malignant tumor, however, is capable of both invading surrounding normal tissue and spreading throughout the body via the circulatory or lymphatic systems (metastasis). Only malignant tumors are properly referred to as cancers, and it is their ability to invade and metastasize that makes cancer so dangerous. Whereas benign tumors can usually be removed surgically, the spread of malignant tumors to distant body sites frequently makes them resistant to such localized treatment. (Cooper GM. 2000) Doctors combine the T, N, M results and other factors specific to the cancer to determine the stage of cancer for each person. Most types of cancer have four stages from stage one to four. Some cancers also have a stage zero. Stage 0 describes cancer in situ, which means "in place." Stage 0 cancers are still located in the place they started and have not spread to nearby tissues. This stage of cancer is often highly curable, usually by removing the entire tumor with surgery. Stage 1 is usually a small cancer or tumor that has not grown deeply into nearby tissues. It also has not spread to the lymph nodes or other parts of the body. It is often called early-stage cancer. Stage II and Stage III, in general, indicate larger cancers or tumors that have grown more deeply into nearby tissue. They may have also spread to lymph nodes but not to other parts of the body. Stage IV means that the cancer has spread to other organs or parts of the body. It may also be called advanced or metastatic cancer. (cancer.net, 2019) machine learning methods have become a popular tool for medical researchers, further, the accurate prediction of a disease outcome is one of the most interesting and challenging tasks for physicians. As a result, machine learning methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relationships between them, from complex datasets, while they are able to effectively predict future outcomes of a cancer type. (K. Kourou, et al. 2015) Great advances in AI can save lives!

K Nearest neighbors (k-NN) determines training points that are closely related to an input vector through an appropriate similarity metric. Nearest neighbor algorithms are attractive because they are easy to implement, nonparametric, and learning-based. (Ni, et al. 2009) Under many circumstances, the k-nearest neighbor algorithm is used to perform the classification. This decision rule provides a simple nonparametric procedure for the assignment of a class label to the input pattern based on the class labels represented by the k-closest (say, for example, in the Euclidean sense) neighbors of the vector. The k-NN rule is a suboptimal procedure. However, it has been shown that in the infinite sample situation, the error rate for the 1-NN rule is bounded above by no more than twice the optimal Bayes error rate and, that as k increases, this error rate approaches the optimal rate asymptotically. (Keller, et al. 1985) A support vector machines (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 dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either

side. (Patel, et al.) In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. (Hearst, et al, 1998) Gaussian process are a powerful algorithm for both regression and classification. Their greatest practical advantage is that they can give a reliable estimate of their own uncertainty. In other words, a Gaussian process is a probability distribution over possible functions. Since Gaussian processes let us describe probability distributions over functions we can use Bayes' rule to update our distribution of functions by observing training data. A key benefit is that the uncertainty of a fitted GP increases away from the training data and this is a direct consequence of GPs roots in probability and Bayesian inference. When you're using a GP to model your problem you can shape your prior belief via the choice of kernel. This lets you shape your fitted function in many different ways. Gaussian processes are non-parametric (although kernel hyper parameters blur the picture) they need to take into account the whole training data each time they make a prediction. This means not only that the training data has to be kept at inference time but also means that the computational cost of predictions scales with the number of training samples. (Knagg, 2019) A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Further, a decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression). It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter/differentiator in input variables. Dividing efficiently based on maximum information gain is key to decision tree classifier. However, in real world with millions of data dividing into pure class in practically not feasible (it may take longer training time) and so we stop at points in nodes of tree when fulfilled with certain parameters (for example impurity percentage). Decision tree is classification strategy as opposed to the algorithm for classification. It takes top down approach and uses divide and conquer method to arrive at decision. We can have multiple leaf classes with this approach. (Brid, 2018) Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The fundamental concept behind random forest is a simple but powerful one: the wisdom of crowds. In data science speak, the reason that the random forest model works so well is that A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this

wonderful effect is that the trees protect each other from their individual errors (as long as they don't constantly all error in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction. (Yiu, 2019) Modeled loosely on the human brain, a neural net consists of thousands or even millions of simple processing nodes that are densely interconnected. Most of today's neural nets are organized into layers of nodes, and they're "feed-forward," meaning that data moves through them in only one direction. An individual node might be connected to several nodes in the layer beneath it, from which it receives data, and several nodes in the layer above it, to which it sends data. To each of its incoming connections, a node will assign a number known as a weight. When the network is active, the node receives a different data item (a different number) over each of its connections and multiplies it by the associated weight. It then adds the resulting products together, yielding a single number. If that number is below a threshold value, the node passes no data to the next layer. If the number exceeds the threshold value, the node fires, which in today's neural nets generally means sending the number (the sum of the weighted inputs) along all its outgoing connections. (Hardesty, 2017) A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN). The term MLP is used ambiguously, sometimes loosely to refer to any feedforward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptrons (with threshold activation). Multilayer perceptrons are sometimes colloquially referred to as vanilla neural networks, especially when they have a single hidden layer. (Hastie, et al. 2009) It at least has 3 nodes, one hidden node in between and two input/output nodes. Adaptive boosting (AdaBoost), focuses on classification problems and aims to convert a set of weak classifiers into a strong one. For any classifier with accuracy higher than 50%, the weight is positive. The more accurate the classifier, the larger the weight. While for the classifer with less than 50% accuracy, the weight is negative. It means that we combine its prediction by flipping the sign (towardsdatascience.com, 2017). Naïve Bayes (NB) classifiers are a set of classifiers which probabilistic classifiers based on Bayes rule (Maximum A Posteriori decision rule in a Bayesian setting). There is very little explicit training in Naive Bayes compared to other common classification methods. The only work that must be done before prediction is finding the parameters for the features' individual probability distributions, which can typically be done quickly and deterministically. This means that Naive Bayes classifiers can perform well even with high-dimensional data points and/or a large number of data points. (Soni, 2018 and McCallum 2019) Quadratic discriminant analysis (QDA) is closely related to linear discriminant analysis (LDA), where it is assumed that the measurements are normally distributed. Unlike LDA however, in QDA there is no assumption that the covariance of each of the classes is identical. To estimate the parameters required in quadratic discrimination more computation and data is required than in the case of linear discrimination. If there is not a great difference in the group covariance matrices, then the latter will perform as well as quadratic discrimination. Quadratic Discrimination is the general form of Bayesian discrimination. (rapidminer.com, 2019)

According to the literature review, it is expected that these ML methods show good accuracies and provide some features importance. SVM, DTs, BN, and ANN was predicted to be the best choices.

2. Methods

Our machine learning questions here is whether the tumor found via various methods is either benign or malignant? As a result, the final output diagnosis model is a binary result stating as Yes/No for the tumor being malignant.

The dataset was gathered from Kaggle website and for learning, sklearn library in Python was used. For visualization matplotlib library was selected and some figures were plotted in MATLAB. The clinical data of our study was extracted from breast cancer Wisconsin (diagnostic) data set including clinical quantitative information of 569 patients. Every instance in any dataset used by machine learning algorithms is represented using the same set of features. The features may be continuous, categorical or binary. If instances are given with known labels (the corresponding correct outputs) then the learning is called supervised, in contrast to unsupervised learning, where instances are unlabeled, researchers hope to discover unknown, but useful, classes of items (Jain et al., 1999). Features are 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 3-dimensional space is that described in a previous research done by K. P. Bennett et al, 1992.

The dataset properties includes ID number, diagnosis (M = malignant, B = benign) and 10 features. Ten real-valued features are computed for each cell nucleus as below:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness ($\frac{perimeter^2}{area}$ - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, and field 23 is Worst Radius. All feature values are recoded with four significant digits. There are no missing attribute values for the dataset. Class distribution consists of 357 benign cases and 212 malignant cases.

The machine learning method in supervised learning for cancer data is a classification problem. The classifiers used for learning were nearest neighbors, linear SVM, Gaussian process, decision tree, random forest, neural network, adaptive boosting, naïve Bayesian networks and quadratic discriminant analysis. The accuracy of each method for the dataset was studied. For decision tree classifier and random forest classifier, the maximum depth was set to 5 and for support vector machine learning the kernel was set to linear. For K-nearest neighbor, K was chosen to be 3.

The dataset was divided to two subsets: 80% of the data for training and 20% of the data for testing.

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The random forest classifier was chosen for feature importance extraction and the results were generated in the code output.

All in all, the most common algorithms used for diagnosis were chosen and the general parameters were set into the study.

3. Results

The accuracy of each classification method was computed and the results are as follows:

Classification Method	Accuracy
Nearest Neighbors	93.859649 %
Linear SVM	94.736842 %
Gaussian Process	96.491228 %
Decision Tree	94.736842 %
Random Forest	95.614035 %
Neural Network	92.982456 %
AdaBoost	96.491228 %
Naive Bayes	96.491228 %
QDA	97.368421 %

The top three best methods in the study were QDA, Gaussian process and adaptive boosting. The complete comparison between all the methods are available in figure 1 in the bar diagram done in MATLAB software.

The next step is to demonstrate the key features of the study. The impact of each feature are demonstrated in the following list:

1. fractal_dimension_se feature 27 (0.123613)
2. fractal_dimension_worst feature 20 (0.102393)
3. compactness_se feature 23 (0.090718)
4. concavity_se feature 22 (0.073362)
5. concavity_mean feature 7 (0.069074)
6. concave points_worst feature 0 (0.057682)
7. area_worst feature 2 (0.057616)
8. perimeter_mean feature 6 (0.055987)
9. perimeter_worst feature 3 (0.052329)
10. symmetry_se feature 26 (0.042248)
11. fractal_dimension_mean feature 25 (0.027331)
12. concave points_se feature 21 (0.024081)
13. radius_worst feature 1 (0.023992)

14. area_se feature 13 (0.023794)
15. texture_mean feature 12 (0.021898)
16. symmetry_mean feature 24 (0.019354)
17. compactness_worst feature 10 (0.017798)
18. radius_mean feature 5 (0.017481)
19. smoothness_se feature 28 (0.014939)
20. concave points_mean feature 4 (0.011544)
21. texture_se feature 29 (0.010375)
22. compactness_mean feature 17 (0.009089)
23. radius_se feature 16 (0.008688)
24. smoothness_worst feature 15 (0.007596)
25. area_mean feature 8 (0.007226)
26. concavity_worst feature 9 (0.006757)
27. symmetry_worst feature 18 (0.006067)
28. smoothness_mean feature 19 (0.005820)
29. perimeter_se feature 14 (0.005708)
30. texture_worst feature 11 (0.005441).

Using MATLAB library (matplotlib), the bar diagram of the feature significance analysis is also available in figure 2.

According to the assumptions made in the introduction part, classifiers showed good accuracy but not in the expected order. This could be due to many reasons discussed in the discussion section including the data type, the parameters set for the algorithms, etc.

4. Discussions

A real dataset of breast cancer was successfully studied by means of different algorithms of supervised machine learning which here was a classification problem. The algorithms that was used shown high accuracy. There are multiple reasons for having high success rate of prediction including homogeneity of data, binary output and data size. This study, along numerous other studies, prove that data science can play major role in human health. In this section different algorithms and features are discussed.

Algorithms

ANNs handle a variety of classification or pattern recognition problems. They are trained to generate an output as a combination between the input variables. Multiple hidden layers that represent the neural connections mathematically are typically used for this process. Even though ANNs serve as a gold standard method in several classification tasks they suffer from certain drawbacks, for instance, their generic layered structure proves to be time-consuming while it can lead to very poor. SVMs are a more recent approach of ML methods applied in the field of cancer prediction/prognosis. Initially SVMs map the input vector into a feature space of higher dimensionality and identify the hyperplane that separates the data points into two classes. The marginal distance between the decision hyperplane and the instances that are closest to boundary is maximized. The resulting classifier achieves considerable generalizability and can therefore be used for the reliable classification. DTs follow a tree-

structured classification scheme where the nodes represent the input variables and the leaves correspond to decision outcomes. DTs are one of the earliest and most prominent ML methods that have been widely applied for classification purposes. Based on the architecture of the DTs, they are simple to interpret and quick to learn. When traversing the tree for the classification of a new sample we are able to conjecture about its class. The decisions resulted from their specific architecture allow for adequate reasoning which makes them an appealing technique. BN classifiers produce probability estimations rather than predictions.
As their name reveals, they are used to represent knowledge coupled with probabilistic dependencies among the variables of interest via a directed acyclic graph. BNs have been applied widely to several classification tasks as well as for knowledge representation and reasoning purposes.

Analysis and comparison
The size of the data should be sufficiently large to get accurate results. Besides, ML algorithms work better when the dimensionality is lower. However, a growing trend was noted in the studies published the last few years that applied semi-supervised ML techniques for modeling cancer survival. Overall, SVM and ANN classifiers were widely used. ANNs have been used extensively for nearly 30 years. SVMs constitute a more recent approach in the cancer prediction/ prognosis and have been used widely due to its accurate predictive performance. On the other hand, here quadratic discriminant analysis (QDA) was identified as the best algorithm within different classifiers that was studied. All of the classifiers performed well (accuracy>90%) and were not computationally expensive.

Best algorithm
The best algorithm to use for learning highly depends on the types of data collected, size of the data samples, time limitations and type of prediction outcomes. As a result, there is not a unique method to say is the possible algorithm for diagnosis. Here we found QDA to be the best, however this was found by trying various learning techniques and comparing them to each other.

Features
Taking a more in-depth view into feature importance results concludes that fractal dimensions' standard error and worse case are the most important features in breast cancer diagnosis. Then, compactness and concavity play major roles in cancer diagnosis. The top 6 features are proven here to have around 53% impact on breast cancer diagnosis, which means we found 6 features out of 30 feature that have more than 50% impact on the entire diagnosis. This finding can potentially be beneficial in my future work on this area.

Still, there are lots of things needs to be improved. First, the accuracy of the learning needs to approach to 100% in order for the method to be used in real-life application in medical diagnosis. Second, the method needs to be reliable and should be true in all conditions. Our future work is to overcome these limitations.

Overall, data science is one of the most helpful methods to improve human wellbeing and I hope one day there would be no malign diseases thanks to improvements in bioinformatics and other related fields.

References

Jain, A.K., Murty, M. N., and Flynn, P. (1999), *"Data clustering: A review, ACM Computing Surveys"*, 31(3): 264–323.

K. P. Bennett and O. L. Mangasarian: *"Robust Linear Programming Discrimination of Two Linearly Inseparable Sets"*, Optimization Methods and Software 1, 1992, 23-34

https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/stages-cancer

The Cell: A Molecular Approach. 2nd edition. Cooper GM. Sunderland (MA): Sinauer Associates; 2000.

https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/data

Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos , Michalis V. Karamouzis, Dimitrios I. Fotiadis, "*Machine learning applications in cancer prognosis and prediction*", Computational and structural biology journal, 2015.

https://scikit-learn.org/

Karl S. Ni, and Truong Q. Nguyen, *"An Adaptable -Nearest Neighbors Algorithm for MMSE Image Interpolation"*, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 18, NO. 9, SEPTEMBER 2009.

Savan Pater, https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

JAMES M. KELLER, MICHAEL R. GRAY, AND JAMES A. GIVENS, JR., *"A Fuzzy K-Nearest Neighbor Algorithm"*, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, VOL. SMC-15, NO. 4, JULY/AUGUST 1985

M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998.

Oscar Knagg, "An intuitive guide to Gaussian processes", https://towardsdatascience.com/an-intuitive-guide-to-gaussian-processes-ec2f0b45c71d

Rajesh S. Brid, Decision trees: a simple way to visualize a decision, https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

Larry Hardesty, Neural networks: Ballyhooed artificial-intelligence technique known as "deep learning" revives 70-year-old idea, MIT News Office, April 14, 2017.

Boosting algorithm: AdaBoost, https://towardsdatascience.com/boosting-algorithm-adaboost-b6737a9ee60c, 2017.

Devin Soni, Introduction to Naive Bayes Classification, towards data science, 2018.

McCallum, Andrew. "Graphical Models, Lecture2: Bayesian Network Represention" (PDF), 2019.

Appendix

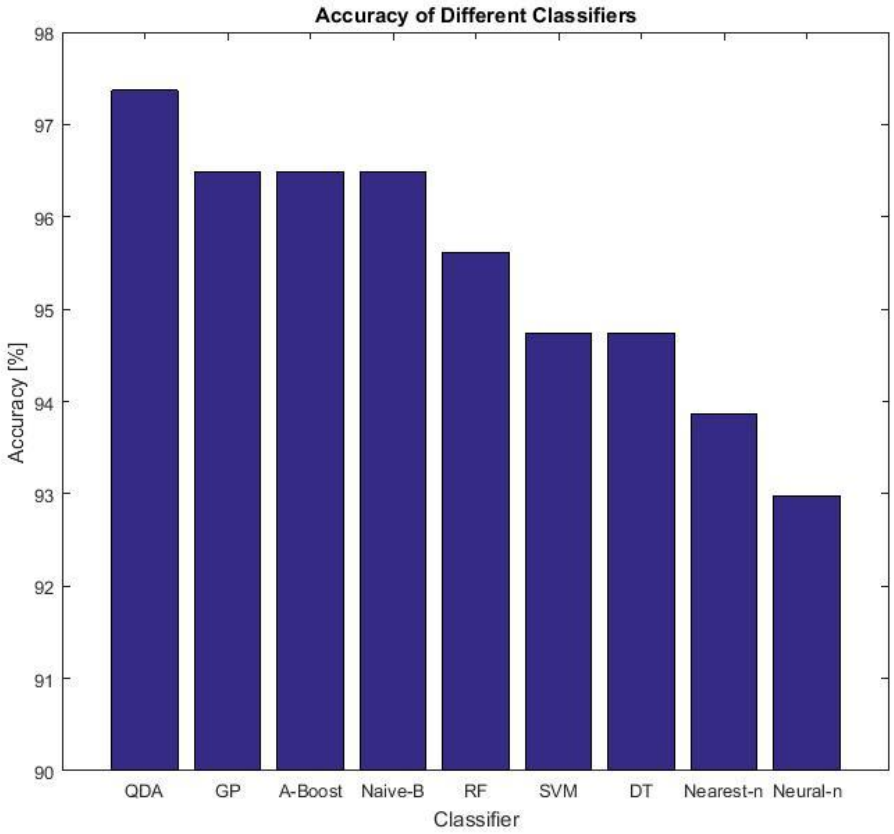

Figure 1: Classifier accuracies sorted in ascending order.

Classification Methods:

Nearest Neighbors
Linear SVM
Gaussian Process
Decision Tree
Random Forest
Neural Network
AdaBoost
Naive Bayes
QDA

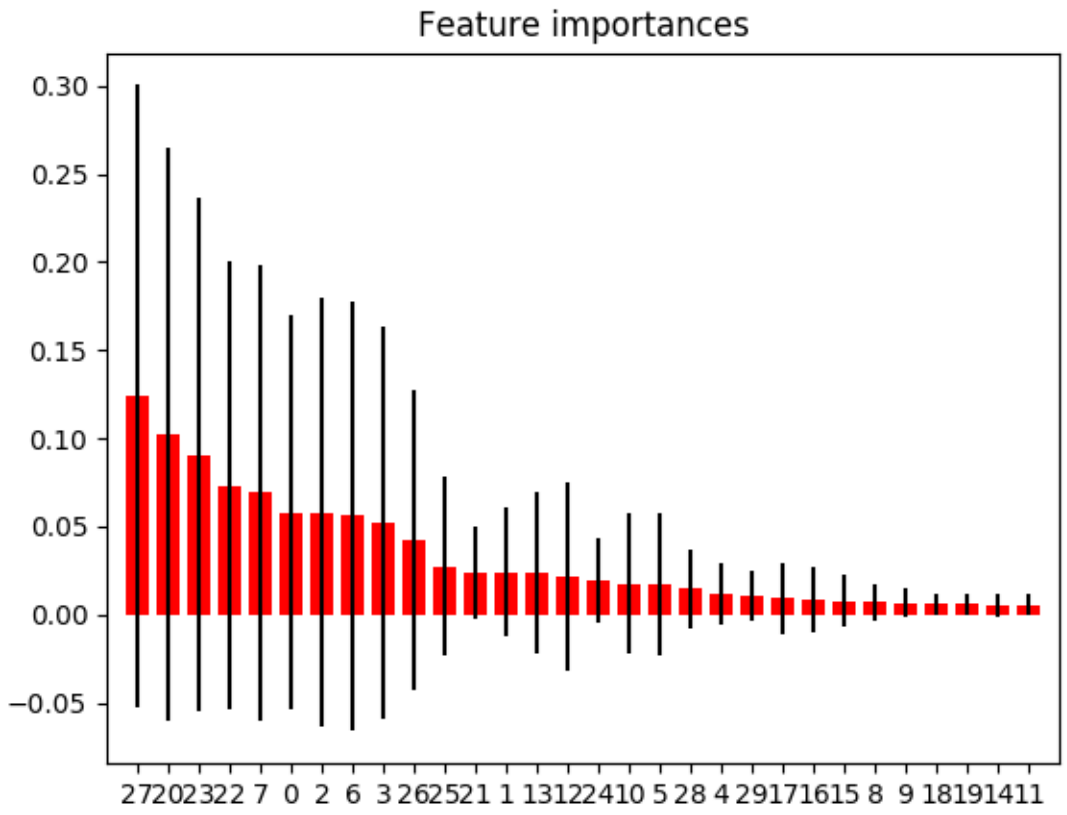

Figure 2: Feature importance bar diagram sorted in ascending order; each feature number is available on x-axis and the impact of each feature is on y axis

fractal_dimension_se feature 27 (0.123613), fractal_dimension_worst feature 20 (0.102393), compactness_se feature 23 (0.090718), concavity_se feature 22 (0.073362), concavity_mean feature 7 (0.069074), concave points_worst feature 0 (0.057682), area_worst feature 2 (0.057616), perimeter_mean feature 6 (0.055987), perimeter_worst feature 3 (0.052329), symmetry_se feature 26 (0.042248), fractal_dimension_mean feature 25 (0.027331), concave points_se feature 21 (0.024081), radius_worst feature 1 (0.023992), area_se feature 13 (0.023794), texture_mean feature 12 (0.021898), symmetry_mean feature 24 (0.019354), compactness_worst feature 10 (0.017798), radius_mean feature 5 (0.017481), smoothness_se feature 28 (0.014939), concave points_mean feature 4 (0.011544), texture_se feature 29 (0.010375), compactness_mean feature 17 (0.009089), radius_se feature 16 (0.008688), smoothness_worst feature 15 (0.007596), area_mean feature 8 (0.007226), concavity_worst feature 9 (0.006757), symmetry_worst feature 18 (0.006067), smoothness_mean feature 19 (0.005820), perimeter_se feature 14 (0.005708), texture_worst feature 11 (0.005441).