SVM-based classification on AFM images of prostate cancer cells

Jianzhong Yu‡, Hanxing Gao‡, Xiaoxia Si, Hongqin Yang, Yuhua Wang*

Key Laboratory of Optoelectronic Science and Technology for Medicine of Ministry of Education, Provincial Key Laboratory for Photonics Technology, Fujian Normal University, Fuzhou 350007, China.

ABSTRACT

Prostate cancer is the 2nd most commonly occurring male cancer and the 4th most common cancer overall. Early detection and diagnosis are important for clinical treatment. Atomic force microscopy (AFM)-based techniques have been shown to have potential in detecting malignant cancers and artificial intelligence can improve the accuracy of diagnostic and prognostic prediction tests. In this study, the classification of AFM images of prostate cells was performed using machine learning. For early prediction, we used the support vector machine (SVM) to classification prostate cells and compare the classification performance with the remaining four conventional classifiers such as logistic regression (LR), stochastic gradient descent (SGD), K-nearest neighbours (KNN), random forest (RF). Most of the classifiers did well after using the feature selection method (BorutaShap). The results show that the accuracy (ACC) of the features selected using the BorutaShap algorithm combined with the SVM classifier can reach 82.5%. Our current study demonstrates that AFM imaging combined with machine learning can be used to identify prostate cancer cells with an effective classification performance and robustness.

Keywords: Prostate cancer, AFM images, SVM, Feature Selection, Classifications

1. INTRODUCTION

Prostate cancer is the second most common male cancer worldwide[1], about 1 in 41 men will die of prostate cancer. The death rate of prostate cancer increases with age, and nearly 55% of people dead after 65 years old. Even though, it is a serious condition only about 1 in 8 men will be diagnosed with prostate cancer during their lifetime[2.]. Clinicians treating prostate cancer face challenges in terms of early and accurate diagnosis of different stages of prostate cancer. An accurate and early diagnosis can help in deciding a better treatment plan which can be more effective and increase the survival rate in patients suffering from prostate cancer.

Force volume images in atomic force microscopy, also known as AFM nanometer resolution shapes imaging has great potential in the study of cell mechanics. AFM can be used to distinguish cancer from normal cells and tissues[3], but measuring the mechanics of cells remains uncertain [4] and difficult to identify activated cells and tissues in a clinical setting. For the remedy of these deficiencies, in addition to the aforementioned improvement method combined with computational models, the nanometer-resolution topography imaging capabilities of AFM also have great potential. The topography capabilities of AFM combined with machine learning methods, enable the vast amount of information contained in nanotopography maps to improve their performance and range of applications.

Currently, using artificial intelligence and machine learning for cancer prevention is a practical approach[5]. Artificial intelligence (AI) is an important technology that supports daily social life, economic activities, and also the health sector [6]. Machine learning applications have solved many problems, i.e., the prediction of cancer patients and predictions of corporate bankruptcy [7 8]. Cancer data has many features that contain information about cancer itself. Unfortunately, many

* Corresponding author: yuhwang@fjnu.edu.cn

‡ Equal contributors

SPIE-CLP Conference on Advanced Photonics 2022, edited by Xu Liu, Anatoly Zayats, Xiaocong Yuan, Proc. of SPIE Vol. 12601, 1260106
 · 0277-786X · doi: 10.1117/12.2667179

clinical data (including prostate cancer) have many irrelevant features, and this redundant information is negative for machine learning. Therefore, we choose to use feature selection to remove redundant information and select the best subset of features. The benefit of feature selection in machine learning is reducing the amount of data needed to reach the learning stage, increasing the value of predictive accuracy, more concise and easy-to-understand data, and reducing execution time [9 10 11]. The main contribution of this paper is to combine feature selection with machine learning and apply it to AFM prostate cancer image for early diagnosis. In addition, we compared four common models to find the best, most simple and effective prostate diagnosis model.

2. MATERIALS AND METHODS

Figure 1 depicts the workflow of the proposed work, highlighting the overall steps taken in this work. First, we preprocess and standardize the data. Second, feature selection techniques were used to obtain the best subset of data features, then the chosen subset is trained and tested with specified models. Finally, we evaluate the performance of our results and predict prostate cancer. Python 3 was used to carry out this task. The objection of this study is to predict and diagnose prostate cancer by using ML models and evaluate the most effective based on six criteria: specificity, sensitivity, precision, ACC, F1-score and receiver operating characteristic curve. All work is done in the anaconda environment, which uses Python's NumPy and SciPy numerical and scientific libraries, and pandas and matplotlib[12].

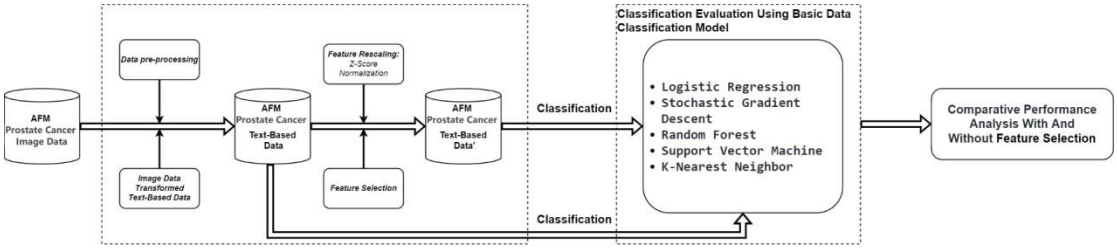

Figure 1. Comparative Analysis of Prostate Cancer Prediction Using Machine Learning Techniques.

2.1 Prostate Cancer Data

The data used in this study were data on prostate cancer, there are two types of cells, one is prostate cells (PZ-7) and the other is prostate cancer cells (PC-3). We selected the most stable Heigh channel to image the prostate cells and prostate cancer cells. After obtaining the cell image captured by AFM, we convert the image data of prostate cells into text-based data through Gwyddion (Version 2.57). Gwyddion is an open source free general scanning microscope image processing software, with a small size, a small interface and a very comprehensive processing tool. The prostate cancer dataset has 18 features like 'Maximum peak height (Sp)', 'Maximum pit depth (Sv)', 'Maximum height (Sz)', 'Projected area', 'Surface area', 'Volume', etc. In this paper, label 0 is used for non-cancer patients and label 1 for cancer patients.

2.2 Feature Selection

Feature selection is a process to reduce the number of attributes; it also involves the selection of a subset of original features. The main goal of feature selection is to reduce the dimensionality of data and to provide a better classifier model to improve classification accuracy. In this paper, we chose to use BorutaShap[13] for feature selection.

BorutaShap is a wrapper feature selection method that combines both the Boruta feature selection algorithm with Shapley values. This combination has proven to outperform the original permutation importance method in both speeds, and the quality of the feature subset produced. Not only does this algorithm provide a better subset of features, but it can also simultaneously provide the most accurate and consistent global feature rankings.

The algorithm steps are as follows [13]: Start by creating new copies of all the features in the data set and name them shadow + feature name, and shuffle these newly added features to remove their correlations with the response variable. Run a classifier on the extended data with the random shadow features included. Then rank the features using a feature importance metric the original algorithm used permutation importance as it's metric of choice. Create a threshold by using the maximum importance score from the shadow features. Then assign a hit to any feature that had exceeded this threshold. For every unassigned feature preform a two-sided T-test of equality. Attributes which have an importance significantly

lower than the threshold are deemed 'unimportant' and are removed them from process. Deem the attributes which have importance significantly higher than the threshold as 'important'. Remove all shadow attributes and repeat the procedure until an importance has been assigned for each feature, or the algorithm has reached the previously set limit of runs.

2.3 Prostate Cancer Classification

After the completion of feature selection, the selected feature passed through to the machine learning classifiers for the classification task. The proposed work considers five classifiers for the analysis of performance comparison (logistic regression (LR), stochastic gradient descent (SGD), K-nearest neighbours (KNN), random forest (RF), support vector machine (SVM)). The SVM method has been used for classification of Schizophrenia, insurance, and classification of Hyperspectral imagery [14 15 16]. However, only the best results were discussed in this paper. It can be seen that SVM allows decision bounds to be complex and can perform well even on low-dimensional data.

SVMs are set of related supervised learning methods used for classification and regression [17]. They belong to a family of generalized linear classification. A special property of SVM is, SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM is called Maximum Margin Classifier. SVM is based on the Structural risk Minimization (SRM). SVM map input vector to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separate the data. The separating hyperplane is the hyperplane that maximize the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be [18 19].

Suppose there is a dataset D, $\boldsymbol{x_i}$, $\boldsymbol{y_i}$where i = 1, …D, the set of training data in the dataset D that has two classes consist of N input vectors $\boldsymbol{x_i}$, … $\boldsymbol{x_n}$ and $\boldsymbol{y_i} \in -1,1$ with $\boldsymbol{y_i}$ being the class label from the dataset (malignant cancer or benign cancer). The hyperplane to be formed is defined as follows:

$$\boldsymbol{y(x) = w^T x_i + b} \tag{1}$$

Optimal Canonical Hyperplane (OCH) is a canonical Hyperplane having a maximum margin. For all the data, OCH should satisfy the following constraints:

$$\boldsymbol{y_i(w^T x_i + b) \geq 1, \forall i = 1, \ldots, N} \tag{2}$$

$$\boldsymbol{\min_{w,b} \frac{|w|^2}{2}} \tag{3}$$

2.4 Performance Evaluation of Model

Classification models assign data to predicted categories. There are four options. If the data is classified as positive with a positive label, it is considered a true positive (TP); if it is scored as negative, it is considered a false negative (FN). Data are considered true negatives (TN) if they have a negative label and are classified as negative, and false positives (FP) if it is classified as positive. From a classification model (classifier) and a data set, 2 × 2 confusion matrix can be formed and state the disposition of the dataset.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$\text{F1–score} = 2\frac{Precision \bullet Recall}{Precision + Recall} \tag{9}$$

with TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative. Precision is a measure of how accurately a model is predicted to be positive, or how much of the model is actually positive. Recall is used to count the number of true positives captured by the model and labelled as positive.

3. RESULTS AND DISCUSSION

Although there are other channels such as the Slope channel and the Adhesion channel in the AFM imaging mode, we selected the most stable height channel to image the prostate cells and prostate cancer cells (Figure 2) for subsequent discriminant models. Then we use Gwyddion to obtain 18 features using Gwyddion for AFM image processing.

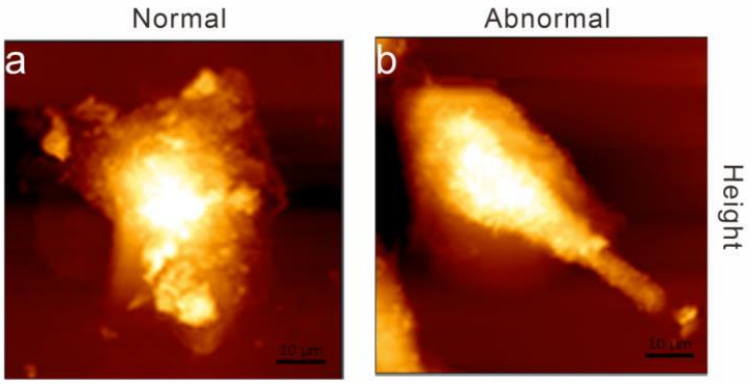

Figure 2. Atomic force tomography image. (a) PZ-7 normal prostate cells,(b) PC-3 prostate cancer cells. Where (a&b) are obtained through the Height channel in the AFM.

Then, the feature selection method is implemented for 18 features, In the end, we extracted the top 9 features of importance from the original 18 features (Figure 3). We exclude the remaining 9 features, retain the extracted 9 feature data for classification as an experimental dataset, and retain the original dataset for classification as a control experiment.

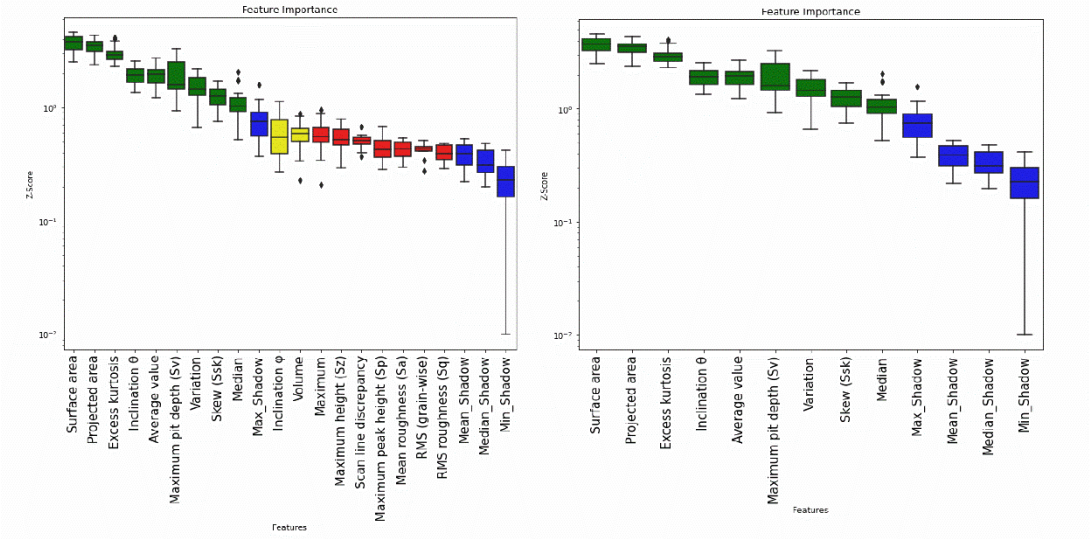

Figure. 3. Feature importance by BorutaShap (Left: without feature selection. Right: with feature selection. Red represents feature columns).

Table 1 shows a comparison of results using four classifiers without feature selection and feature selection on the AFM Prostate cancer dataset. All classifiers except KNN and SDG give better results after using BorutaShap. The SVM classifier provides the highest accuracy of 82.5% and reduces the time while improving the classification accuracy with feature

selection (From 380.79 s to 307.96 s). Without using feature selection, KNN classifier provides the best classification accuracy of 78.7%, but still lower than the SVM after using feature selection.

In this study, the shortest time of the five classifiers we used was SGD. The time required to classify using feature selection and not using feature selection is the same, both are 0.08s. But the accuracy of classification obtained by using feature selection is reduced, which from 62.0% to 52.2%.

In degree of the ascension of classification accuracy, SVM classifiers increased from 72.9% without feature selection to 82.5% with feature selection, an increase of 9.6%, It is the most improved of the seven classifiers. RF is the second most improved classifier. The experimental results show that our method can not only improve the classification performance, but also reduce the time in the identification and classification of AFM prostate cancer cells, which has substantial significance for the early prediction and diagnosis of prostate.

Table 1. Comparative Analysis of Prostate Cancer Prediction Using Machine Learning Techniques.

Models	Without feature selection					With feature selection				
	Precision	Recall	F1-score	Acc(%)	TIME(S)	Precision	Recall	F1-score	Acc(%)	TIME(S)
LR	0.75	0.85	0.79	77.0	0.70	0.77	0.93	0.83	**78.7**	0.66
KNN	0.83	0.85	0.82	**78.7**	6.79	0.83	0.85	0.81	78.5	1.06
SGD	0.575	0.68	0.59	**62.0**	0.08	0.50	0.51	0.41	52.2	0.08
RF	0.77	0.81	0.77	74.5	34.38	0.77	0.96	0.84	**78.3**	26.50
SVM	0.70	0.87	0.77	72.9	380.70	0.88	0.80	0.81	**82.5**	307.96

4. CONCLUSIONS

In summary, AFM cell imaging combined with SVM can be used to identify prostate cancer cells with considerable accuracy, and the use of BorutShap can also improve diagnostic efficiency. This research demonstrates that the combination of the AFM technique and machine learning will be significant for AFM application in cancer detection.

ACKNOWLEDGEMENTS

This work was partly supported by the Natural Science Foundation of Fujian Province under grant no. 2021J02028, the Program for Changjiang Scholars and Innovative Research Team in University under grant no. IRT_15R10.

REFERENCES

[1] "Prostate cancer statistics | World Cancer Research Fund International.", <https://www.wcrf.org/cancer-trends/prostate-cancer-statistics/> (1 December 2022).
[2] "Key Statistics for Prostate Cancer | Prostate Cancer Facts.", <https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html> (1 December 2022).
[3] Plodinec, M., Loparic, M., Monnier, C. A., Obermann, E. C., Zanetti-Dallenbach, R., Oertle, P., Hyotyla, J. T., Aebi, U., Bentires-Alj, M., Lim, R. Y. H. and Schoenenberger, C.-A., "The nanomechanical signature of breast cancer," 11, Nat. Nanotechnol. 7(11), 757–765 (2012).

[4] Wu, P. H., Aroush, D. R. B., Asnacios, A., Chen, W. C., Dokukin, M. E., Doss, B. L., Durand-Smet, P., Ekpenyong, A., Guck, J., Guz, N. V., Janmey, P. A., Lee, J. S. H., Moore, N. M., Ott, A., Poh, Y. C., Ros, R., Sander, M., Sokolov, I., Staunton, J. R., et al., "A comparison of methods to assess cell mechanical properties," Nat. Methods 15(7), 491–498 (2018).

[5] Lu, H., Li, Y., Uemura, T., Kim, H. and Serikawa, S., "Low illumination underwater light field images reconstruction using deep convolutional neural networks," Future Gener. Comput. Syst. 82(MAY), 142–148 (2018).

[6] Lu, H., Li, Y., Chen, M., Kim, H. and Serikawa, S., "Brain Intelligence: Go beyond Artificial Intelligence," undefined (2018).

[7] Rustam, Z. and Yaurita, F., "Insolvency Prediction in Insurance Companies Using Support Vector Machines and Fuzzy Kernel C-Means," undefined (2018).

[8] "Prediction of lung cancer patient survival via supervised machine learning classification techniques.", Int. J. Med. Inf. 108, 1–8 (2017).

[9] "A feature selection method based on improved fisher's discriminant ratio for text sentiment classification.", Expert Syst. Appl. 38(7), 8696–8702 (2011).

[10] Hall, M., "Correlation-based Feature Selection for Machine Learning," undefined (2003).

[11] Rahman, M. A. and Muniyandi, R. C., "Feature selection from colon cancer dataset for cancer classification using artificial neural network," Int. J. Adv. Sci. Eng. Inf. Technol. 8(4–2), 1387–1393 (2018).

[12] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., et al., "SciPy 1.0: fundamental algorithms for scientific computing in Python," 3, Nat. Methods 17(3), 261–272 (2020).

[13] Keany, E., "BorutaShap : A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values." (2020).

[14] Rampisela, T. V. and Rustam, Z., "Classification of Schizophrenia Data Using Support Vector Machine (SVM)," J. Phys. Conf. Ser. 1108(1), 012044 (2018).

[15] Rustam, Z. and Ariantari, N. P. A. A., "Support Vector Machines for Classifying Policyholders Satisfactorily in Automobile Insurance," J. Phys. Conf. Ser. 1028(1), 012005 (2018).

[16] "Classification of Hyperspectral Imagery based on spectral gradient, SVM and spatial random forest.", Infrared Phys. Technol. 95, 61–69 (2018).

[17] The Nature of Statistical Learning Theory.

[18] Srivastava, D. and Bhambhu, L., "DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE," undefined (2009).

[19] Akinnuwesi, B., Macaulay, B. O. and Aribisala, B. S., "Breast cancer risk assessment and early diagnosis using Principal Component Analysis and support vector machine techniques," undefined (2020).