An Empirical Study of Machine Learning Algorithms for Cancer Identification

Turki Turki
King Abdulaziz University
Department of Computer Science, P.O. Box 80221, Jeddah 21589, Saudi Arabia
tturki@kau.edu.sa

***Abstract*—Predicting cancer disease state is an important problem in the cancer discovery process. For example, discriminating between benign and malignant tumors improves the medical diagnosis of cancer. Although technological advancing led to the generation of data pertaining to patients with different disease states, evaluating the prediction performance of machine learning algorithms would be an important step. In this paper, we propose using machine learning algorithms such as a variant of AdaBoost, deepboost, xgboost and support vector machines and evaluate them using area under curve and accuracy on real clinical data related to thyroid cancer, colon cancer and liver cancer. Experimental results show the good performance of SVM.**

Keywords—Boosting, Machine Learning, Cancer Genomics, Medical Diagnosis

I. INTRODUCTION AND RELATED WORK

Cancer rates could increase to 15 million by 2020 [1]. Providing healthcare providers with effective tools to identify cancer cases would be an important step in the medical diagnosis for cancer patients. To achieve this goal, researchers from different domains developed computational approaches, which improve the medical diagnosis process. For example, to differentiate malignant from benign thyroid nodules, Stokowy *et al*. [2] proposed a computational approach based on gene expression data, to improve the pre-surgical diagnostics of thyroid nodules. Zhang *et al*. [3] proposed using extreme learning machine (ELM) applied to several microarray data including lung data, lymphoma data and other benchmark dataset for cancer diagnosis. Compared to other variants of support vector machines, the experimental results demonstrated the good performance of ELM when considering the accuracy as a performance measure. Wang *et al.* [4] proposed a machine learning approach coupled with feature selection for tumor classification. The proposed approach compared with others using data generated via microarray technology. Recent machine learning approaches have been proposed to improve the cancer diagnosis process [5, 6].

Although the previous approaches have contributed to improve the cancer diagnosis process, these approaches depend heavily on the computational methods to improve the prediction performance. As various machine learning researchers have recently proposed computational methods to improve the performance, there is a need to apply these recent machine learning algorithms for clinical data generated via various technologies. In this paper, we present using (1) DeepBoost, which is a new ensemble learning algorithm [7]; (2) xgboost, which is scalable end-to-end tree boosting system [8]; (3) a variant of Adaboost [9]; and (4) support vector machines (SVM) [10, 11] . We apply these machine learning algorithms for cancer identification. Experimental results on real data pertaining to thyroid cancer, colon cancer, and liver cancer show that SVM outperforms the previously mentioned algorithms.

The rest of the paper is organized as follows. Section II reports experimental results of machine learning algorithms on real data pertaining to thyroid cancer, colon cancer and liver cancer. Section III concludes the paper and points out future work.

II. EXPERIMENTS AND RESULTS

We empirically assess the performance of machine learning algorithms on thyroid cancer data, colon cancer data, and liver cancer data. This section first describes the datasets and experimental methodology and presents the experimental results.

A. Datasets

1) Data Pertaining to Thyroid Cancer

The dataset $S \in \mathbb{R}^{25 \times 1147}$ consisted of 25 samples, 1146 features, and a column vector of labels. 8 out of the 25 patients had follicular adenomas (i.e., benign tumor) while the remaining 17 patients had follicular carcinomas (i.e., malignant tumor). The dataset was downloaded from the Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/) with the accession number GSE62054 or available at the address (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62054).

978-1-5386-5053-0/18/$31.00 ©2018 IEEE

2) Data Pertaining to Colon Cancer

The dataset $S \in \mathbb{R}^{21\times5608}$ consisted of 21 samples, 5607 features, and a column vector of labels. 7 out of the 25 patients had normal colon while the remaining 14 patients had tumor colon. The dataset was downloaded from the Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/) with the accession number GSE98406 or available at the address (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE98406).

3) Data Pertaining to Liver Cancer

The dataset $S \in \mathbb{R}^{21\times5608}$ consisted of 21 samples, 5607 features, and a column vector of labels. 7 out of the 25 patients had normal liver while the remaining 14 patients had liver metastases. The dataset was downloaded from the Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/) with the accession number GSE98406 or available at the address (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE98406).

B. Experimental Methodology

We considered four machine learning algorithms, namely DeepBoost [7], xgboost [8], a variant of Adaboost (also called Boost I) [9], and support vector machines (SVM) [10].

To evaluate the machine learning algorithms, we partitioned each dataset into five folds, where we assigned different folds to the training and test sets. Specifically, for each run $j \in \{1,2,3,4,5\}$, fold j was assigned to the test while the remaining folds were assigned to the training set. In each run, we calculated the area under curve (AUC) and accuracy (ACC) and reported the standard deviation on the test set. Then, the average over all five runs was reported, where the average result corresponds to the performance result of utilizing 5-fold cross-validation.

To assess the stability of machine learning algorithms on a cancer identification task, we repeated the 5-fold cross-validation three times on each dataset, where in each time the dataset is randomly partitioned and the performance results of all machine learning algorithms are reported. Then, we reported the average result over the three times of 5-fold cross-validation.

As each dataset consisted of different number of positive and negative samples, the AUC evaluation measure was utilized. Also, when we trained the machine learning algorithms we balanced the number of positive and negative examples via undersampling from the class with majority number of examples. Then, we provided the same number of positive and negative examples to each machine learning algorithm during the training step.

In this work, the software used included DeepBoost [12], xgboost [13], and SVM with a linear kernel [10]. We performed and ran the experiments using R.

C. Experimental Results

1) Predicting Malignancy of Thyroid Nodules

Table 1 shows the performance of four machine learning algorithms on the thyroid cancer dataset. Each result obtained from running the 5-fold cross-validation on a thyroid dataset corresponds to the mean area under curve (MAUC). AMAUC stands for the average of MAUC, where the reported AMAUC is the average of three MAUCs obtained via running 5-fold cross-validation three times. In each run of 5-fold cross-validation, we randomly partition the dataset into five folds and then report the MAUC. The same process is utilized for AMACC. The higher AMAUC an algorithm has, the better its performance is. Similarly, higher AMACC scores indicate the good performance of the machine learning algorithm. AUC is an evaluation measure used for the imbalanced classification to measure the performance. It can be shown from Table 1 that xgboost yields the highest AMAUC score (result is colored in red). When the number of correctly predicted examples matters without considering the nature of imbalanced classification problem, the accuracy is employed as a performance measure, where xgboost yields the best AMACC score.

Table 1: AMAUC and AMACC scores of machine learning algorithms on the task of predicting malignancy in thyroid nodules. The algorithm with the highest AMAUC is colored in Red. The algorithm with the highest AMACC is shown in bold. SD is the standard deviation, which is the average standard deviations of three runs of 5-fold cross-validation. AMAUC, Average MAUC. AMACC, Average MACC. MAUC, mean area under curve obtained from utilizing 5-fold cross-validation. MACC, mean accuracy obtained from utilizing 5-fold cross-validation

	SD	AMAUC	SD	AMACC
SVM	0.252	0.750	0.213	0.726
Xgboost	0.224	0.811	0.226	**0.798**
DeepBoost	0.194	0.758	0.192	0.744
Boost I	0.000	0.500	0.103	0.658

Figure 1 shows the mean area under curve (MAUC) results of machine learning algorithms used in calculating AMAUC results in Table 1. MAUC-1 denotes the mean area under curve result obtained from the first run of 5-fold cross-validation. MAUC-2 stands for the mean area under curve result obtained from the second run of 5-fold cross-validation. MAUC-3 stands for the mean area under curve result obtained from the third run of 5-fold cross-validation. It can be shown from Figure 1 that xgboost yields the highest MAUC under two runs of 5-fold cross-validation.

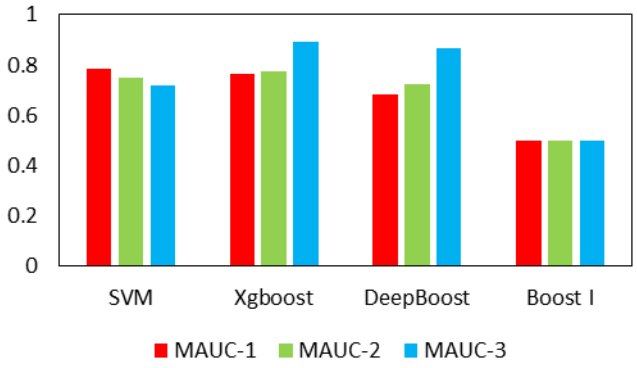

Figure 1: Mean Area Under Curve (MAUC) scores for machine learning algorithms obtained from running 5-fold cross-validation three times, where the thyroid cancer dataset was randomly partitioned each time into 5 folds.

Figure 2 shows the mean accuracy (MACC) results of machine learning algorithms used in calculating AMACC results in Table 1. MACC-1 stands for the mean accuracy result obtained from the first run of 5-fold cross-validation. MACC-2 stands for the mean accuracy result obtained from the second run of 5-fold cross-validation. MAUC-3 stands for the mean accuracy result obtained from the third run of 5-fold cross-validation. It can be shown from Figure 2 that xgboost yields the highest MACC under all three runs of 5-fold cross-validation.

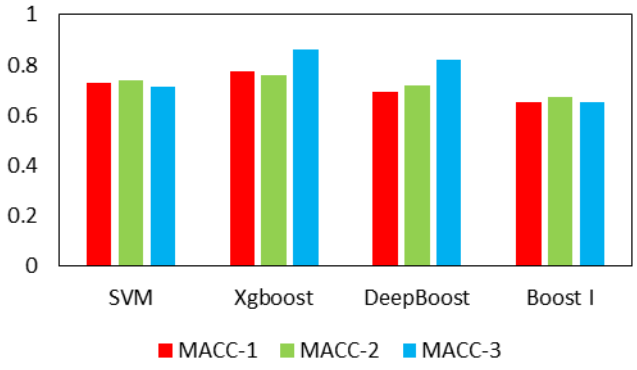

Figure 2: Mean Accuracy (MACC) scores for machine learning algorithms obtained from running 5-fold cross-validation three times, where the thyroid cancer dataset was randomly partitioned each time into 5 folds.

2) Predicting Tumor State in Colon Cancer Patients

According to the AUC performance measure for imbalanced classification, Table 2 shows that SVM yields the highest AMAUC score (result is colored in red). Also, SVM yields the best AMACC score.

Table 2: AMAUC and AMACC scores of machine learning algorithms on the task of predicting tumor state in colon cancer patients. The algorithm with the highest AMAUC is colored in Red. The algorithm with the highest AMACC is shown in bold. SD is the standard deviation, which is the average standard deviations of three runs of 5-fold cross-validation. AMAUC, Average MAUC. AMACC, Average MACC. MAUC, mean area under curve obtained from utilizing 5-fold cross-validation. MACC, mean accuracy obtained from utilizing 5-fold cross-validation

	SD	AMAUC	SD	AMACC
SVM	0.153	0.897	0.159	**0.890**
Xgboost	0.178	0.872	0.190	0.857
DeepBoost	0.267	0.822	0.252	0.833
Boost I	0.000	0.500	0.103	0.694

Figure 3 shows the mean area under curve (MAUC) results of machine learning algorithms used in calculating AMAUC results in Table 2. MAUC-1 represents the mean area under curve result obtained from the first run of 5-fold cross-validation. MAUC-2 stands for the mean area under curve result obtained from the second run of 5-fold cross-validation. MAUC-3 stands for the mean area under curve result obtained from the third run of 5-fold cross-validation. As illustrated from Figure 3, SVM yields the highest MAUC under two runs of 5-fold cross-validation.

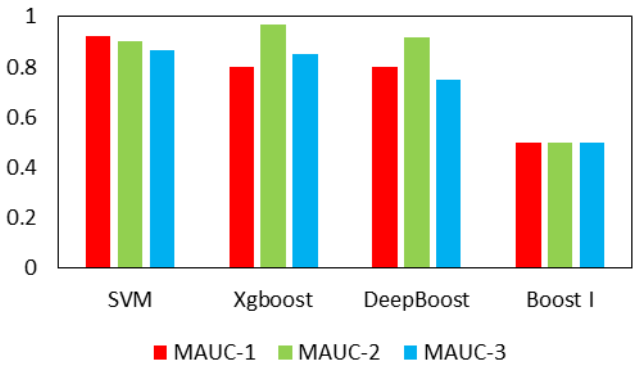

Figure 3: Mean Area Under Curve (MAUC) scores for machine learning algorithms obtained from running 5-fold cross-validation three times, where the colon cancer dataset was randomly partitioned each time into 5 folds.

Figure 4 shows the mean accuracy (MACC) results of machine learning algorithms used in calculating AMACC results in Table 2. MACC-1 is the mean accuracy result obtained from the first run of 5-fold cross-validation. MACC-2 denotes the mean accuracy result obtained from the second run of 5-fold cross-validation. MAUC-3 is the mean accuracy result obtained from the third run of 5-fold cross-validation. As shown from Figure 4, SVM yields the highest MACC under two runs of 5-fold cross-validation. Xgboost yields a high MACC result (see MACC-2).

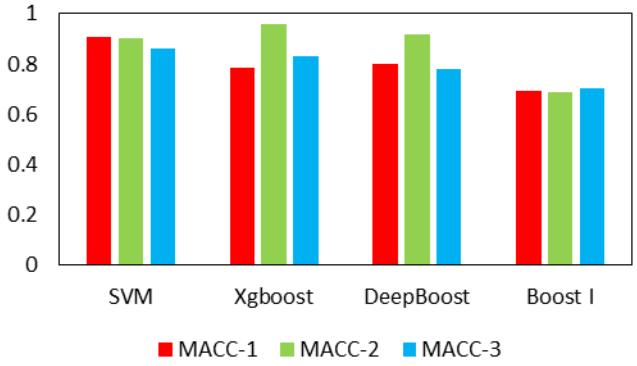

Figure 4: Mean Accuracy (MACC) scores for machine learning algorithms obtained from running 5-fold cross-validation three times, where the colon cancer dataset was randomly partitioned each time into 5 folds.

3) Predicting Tumor State in Liver Cancer Patients

Table 3 shows that SVM yields the highest AMAUC score (result is colored in red). Also, SVM has the best AMACC score (result is shown in bold).

Table 3: AMAUC and AMACC scores of machine learning algorithms on the task of predicting tumor state in liver cancer patients. The algorithm with the highest AMAUC is colored in Red. The algorithm with the highest AMACC is shown in bold. SD is the standard deviation, which is the average standard deviations of three runs of 5-fold cross-validation. AMAUC, Average MAUC. AMACC, Average MACC. MAUC, mean area under curve obtained from utilizing 5-fold cross-validation. MACC, mean accuracy obtained from utilizing 5-fold cross-validation

	SD	AMAUC	SD	AMACC
SVM	0.151	0.897	0.143	**0.895**
Xgboost	0.199	0.797	0.222	0.775
DeepBoost	0.203	0.791	0.218	0.786
Boost I	0.000	0.500	0.092	0.652

Figure 5 shows the mean area under curve (MAUC) results of machine learning algorithms used in calculating AMAUC results in Table 3. MAUC-1 stands for the mean area under curve result obtained from the first run of 5-fold cross-validation. MAUC-2 stands for the mean area under curve result obtained from the second run of 5-fold cross-validation. MAUC-3 stands for the mean area under curve result obtained from the third run of 5-fold cross-validation. Figure 5 shows that SVM yields the highest MAUC under all runs of 5-fold cross-validation.

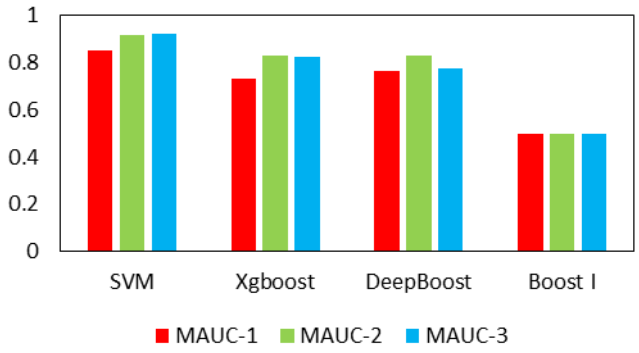

Figure 5: Mean Area Under Curve (MAUC) scores for machine learning algorithms obtained from running 5-fold cross-validation three times, where the liver cancer dataset was randomly partitioned each time into 5 folds.

Figure 6 shows the mean accuracy (MACC) results of machine learning algorithms used in calculating AMACC results in Table 3. MACC-1 is the mean accuracy result obtained from the first run of 5-fold cross-validation. MACC-2 denotes the mean accuracy result obtained from the second run of 5-fold cross-validation. MAUC-3 is the mean accuracy result obtained from the third run of 5-fold cross-validation. It can be shown from Figure 6 that SVM yields the best MACC results (see MACC-1, MACC-2 and MACC-3).

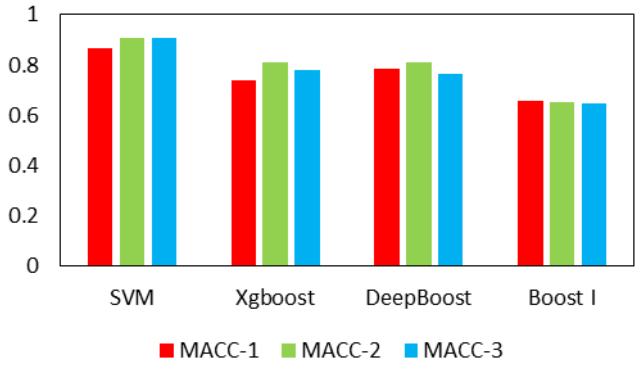

Figure 6: Mean Accuracy (MACC) scores for machine learning algorithms obtained from running 5-fold cross-validation three times, where the liver cancer dataset was randomly partitioned each time into 5 folds.

It is worth noting that Boost I yielded poor performance results when applied to all datasets. Similar poor performance results were also reported in [9].

III. CONCLUSION AND FUTURE WORK

We propose using machine learning algorithms including DeepBoost, xgboost, a variant of AdaBoost and SVM for the task of cancer identification. Experimental results on real clinical data pertaining to thyroid cancer, colon cancer and liver cancer show the good performance of support vector machines.

In future work, we plan to (1) improve the performance of machine learning algorithms via employing transfer learning techniques as in [14]; (2) boost the performance of the machine learning used in this study; and (3) apply feature learning techniques coupled with the machine learning employed in this study.

ACKNOWLEDGMENT

The author thanks the anonymous reviewers for their useful comments that helped improve the paper considerably. The author also thanks King Abdulaziz University for the financial support.

REFERENCES

[1] W. H. Organization, "Global cancer rates could increase by 50% to 15 million by 2020," *Global cancer rates could increase by 50% to 15 million by 2020*, 2003.

[2] T. Stokowy, B. Wojtaś, J. Krajewska, E. Stobiecka, H. Dralle, T. Musholt, S. Hauptmann, D. Lange, L. Hegedüs, and B. Jarząb, "A two miRNA classifier differentiates follicular thyroid carcinomas from follicular thyroid adenomas," *Molecular and cellular endocrinology,* vol. 399, pp. 43-49, 2015.

[3] R. Zhang, G.-B. Huang, N. Sundararajan, and P. Saratchandran, "Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),* vol. 4, no. 3, pp. 485-495, 2007.

[4] Y. Wang, I. V. Tetko, M. A. Hall, E. Frank, A. Facius, K. F. Mayer, and H. W. Mewes, "Gene selection from microarray data for cancer classification—a machine learning approach," *Computational biology and chemistry,* vol. 29, no. 1, pp. 37-46, 2005.

[5] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," *Computer methods and programs in biomedicine,* vol. 153, pp. 1-9, 2018.

[6] Y. Yamamoto, A. Saito, A. Tateishi, H. Shimojo, H. Kanno, S. Tsuchiya, K.-i. Ito, E. Cosatto, H. P. Graf, and R. R. Moraleda, "Quantitative diagnosis of breast tumors by morphometric classification of microenvironmental myoepithelial cells using a machine learning approach," *Scientific Reports,* vol. 7, 2017.

[7] C. Cortes, M. Mohri, and U. Syed, "Deep boosting," *Proceedings of the 31st International Conference on Machine Learning*, 2014. pp. 1179-1187.

[8] T. Chen, and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016. pp. 785-794.

[9] T. Turki, and J. T. L. Wang, "Reverse Engineering Gene Regulatory Networks Using Sampling and Boosting Techniques," *Machine Learning and Data Mining in Pattern Recognition: 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings*, P. Perner, ed., pp. 63-77, Cham: Springer International Publishing, 2017.

[10] C.-C. Chang, and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 2, no. 3, pp. 27, 2011.

[11] J. Zhang, S. O. Williams, and H. Wang, "Intelligent computing system based on pattern recognition and data mining algorithms," *Sustainable Computing: Informatics and Systems*, 2017.

[12] D. Marcous, and Y. Sandbank, "deepboost: Deep Boosting Ensemble Modeling," *R package version 0.1.6*.

[13] T. Chen, T. He, and M. Benesty, "xgboost: extreme gradient boosting (2015)," *R package version 0.3-3*.

[14] T. Turki, Z. Wei, and J. T. Wang, "Transfer Learning Approaches to Improve Drug Sensitivity Prediction in Multiple Myeloma Patients," *IEEE Access*, 2017.