An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer

[1]Muhammad Imran Faisal, [2]Saba Bashir, [3]Zain Sikandar Khan, [4]Farhan Hassan Khan

[1,2,3]Department of Computer Science, Federal Urdu University of Arts, Science &Technology, Islamabad, Pakistan

[4]Knowledge & Data Science Research Center, Department of Computer & Software Engineering,
College of Electrical & Mechanical Engineering, National University of Sciences & Technology (NUST), Islamabad, Pakistan

{[1]faisii700, [2]saba.bashir3000, [3]zain.sikander50}@gmail.com, [4]farhan.hassan@ceme.nust.edu.pk

Abstract – Researchers have widely used statistical and machine learning techniques to construct prediction models in several domains such as prediction of software faults, spam detection, disease diagnosis, and financial fraud identification. The prediction of patients prone to lung cancer can help doctors in their decision making regarding their treatments. In this regard, this research paper attempts to evaluate the discriminative power of several predictors in the study to increase the efficiency of lung cancer detection through their symptoms. A number of classifiers including Support Vector Machine (SVM), C4.5 Decision tree, Multi-Layer Perceptron, Neural Network, and Naïve Bayes (NB) are evaluated on a benchmark dataset obtained from UCI repository. The performance is also compared with well-known ensembles such as Random Forest and Majority Voting. Based on performance evaluations, it is observed that Gradient-boosted Tree outperformed all other individual as well as ensemble classifiers and achieved 90% accuracy.

Keywords: Machine Learning, Prediction, Performance, Ensembles

1. Introduction

In 2012, a survey was conducted, which reports [1] 1.6 million deaths and 1.8 million new cases of lung cancer patients. Lung cancer is common in both gender of US and reported as more dangerous as compared to other types of cancer. Only 15% of cases are detected at the early stage. The most common symptom, i.e. smoking, is reported for lung cancer patients but not all patients involve this symptom. However, several symptoms of lung cancer patients such as their smoking rate can help to detect the lung cancer patient at the early stage. Though, the research community has used certain machine learning techniques such as GEP [2], Fuzzy deep learning [3], and SVM [5]. Wender et al. [3] reported that lung cancer as a serious killer disease in the world mainly in America and East Asia. Moreover, the authors presented that lung cancer patients are 25% higher than patients of other cancer types such as breast cancer and blood cancer. The tumor movements are divided into two parts intra-fractional variation and inter-fractional variation. Intra-fractional works in single treatment sessions and inter-fractional arises between different sessions. Consequently, we evaluate the discriminative power of certain predictors used to improve the accuracy of the prediction model. The layout of an ensemble-based approach is shown in Figure 1.

The dataset was retrieved from the UCI repository. Firstly, the effectiveness of Naïve Bayes (NB), Random Forrest (RF), Support Vector Machine (SVM), and MLP is assessed in terms of accuracy and f-measure. Secondly, a majority voting based ensemble of top-3 best performing classifiers is constructed to predict lung cancer.

Research Contributions:

The major contributions of this research are as follows

- Identification of well-known classifiers and ensemble approaches utilized for lung cancer prediction
- Computation of results over benchmark dataset obtained from UCI repository
- Proposed a majority voting based ensemble based on top-3 best performing individual classifiers
- Comparison and evaluation of results which show that Gradient-boosted Tree outperformed all another individual as well as ensemble classifiers
- Achieved 90% accuracy for lung cancer identification

The remaining part of the study is structured in four sections. In the second section, we present related work. The third section presents the experimental procedure whereas the results are discussed in the fourth section. Finally, in the fifth section, we present conclusions of our work.

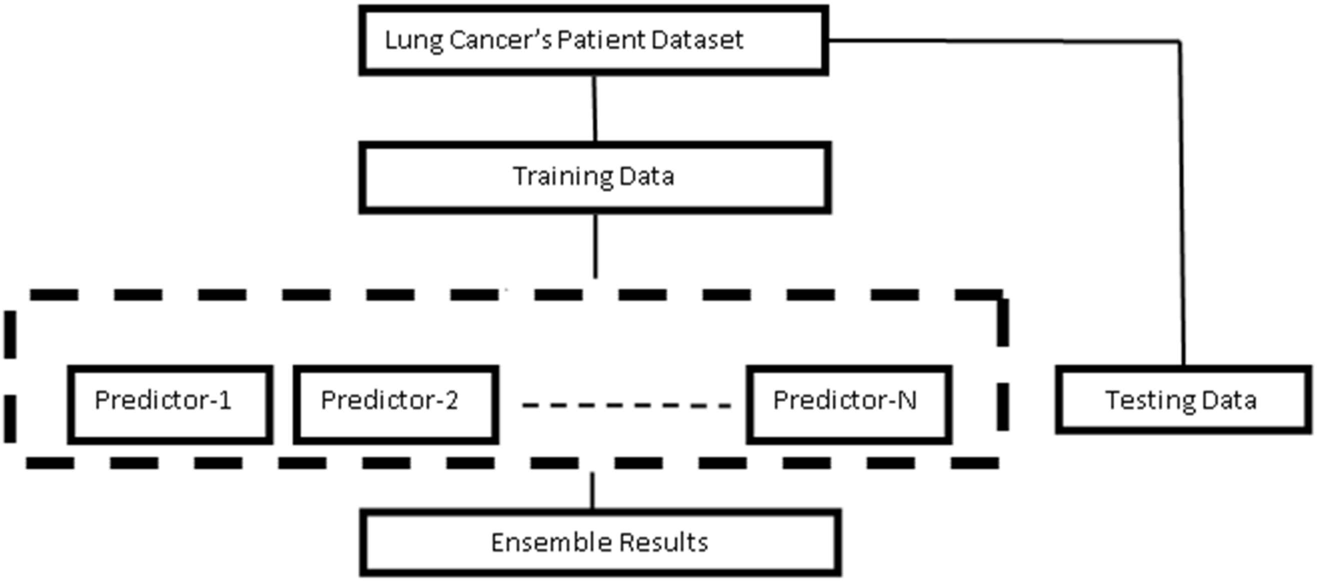

Figure 1: Overview of Ensemble-based Approach

2. RELATED WORK

Like other domains, researchers have successfully implemented the machine and statistical learning techniques to construct prediction models in the context of certain diseases such as lungs cancer.

For diagnosing lung cancer, authors [2] performed some tests on DNA and proteins to identify a tumor in earlier stages. In this study, the author reports the best results of GEP as compared to other classifiers. Li et al. [4] reported that lung cancer may affect one or both lungs. Moreover, the authors list the symptoms of lung cancer as chest pain, chronic cough, difficulty breathing and sudden weight loss. These symptoms can be diagnosed in an early stage. If treatment can be started in an early stage, it can help patients recover. Like other studies, Kourou et al. [5] also investigated that lung cancer is an important reason for cancer-related deaths around the world. If it is identified in the early stage then patients can be treated effectively. Subsequently, authors reported that Microarray technology is the best way to diagnose the lung cancer. In this regard, the authors proposed a model for the diagnosis of lung cancer from microarray data. In GEP model, SVM and Neural network models were used to predict lung cancer and effective results were produced.

Wender et al. [6] used fuzzy deep learning to predict tumor movements which helped to increase the delivery, reduce and accurate application of the radiation and less damage to the healthy tissues during the radiography [7-8]. Karakach et al. [9] suggest microarray analysis as a tool to diagnose diseases. This tool becomes functional by implementing SVM. Due to variations in lung cancer symptoms, its treatment becomes difficult for patients. Data mining or machine learning tool can be more effective to use when these symptoms are utilized as features and predict lung cancer patients in the early stage [10].

It is evident from the state of the art literature review that lung cancer has achieved much attention from the researcher community. A number of approaches have been proposed which need to evaluated and compared for lung cancer prediction [11-14]. This research is focused on assessing such approaches in order to conclude the best methodology for lung cancer detection.

3. PROPOSED METHODOLOGY

The proposed methodology starts with data acquisition which is followed by pre-processing. The selected classifiers are then trained and tested on the benchmark dataset using standard 10-fold cross-validation approach. The results are computed and evaluated to identify the best methodology for lung cancer detection. An overview of the proposed approach is given in Figure 2.

3.1 Data Acquisition

In this paper, we used a dataset namely Lung Cancer which is retrieved from the UCI online repository. The dataset has 32 instances and 57 attributes, 1 class attribute, and 56 predictive attributes.

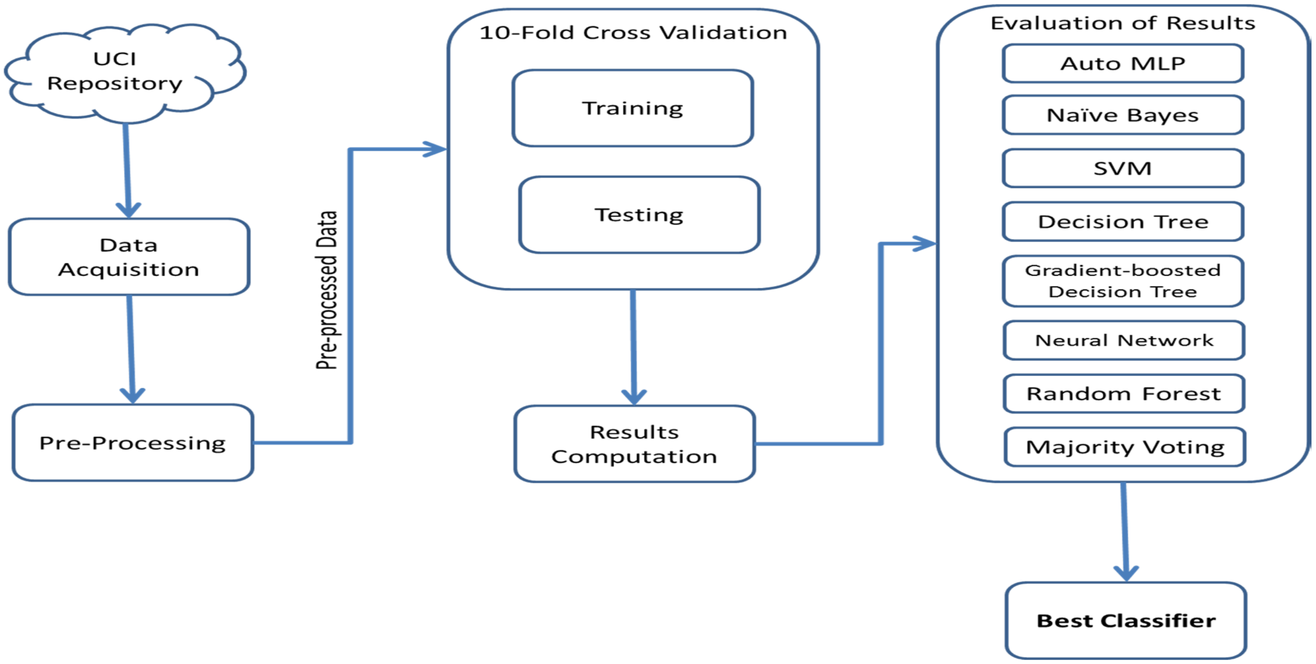

Figure 2: Overview of the Proposed Approach

The main objective of our proposed study is to investigate and measure the effectiveness of different base level predictors and well-known ensemble methods such as Random Forest and Majority Voting. In this study, RapidMiner tool was utilized to perform experiments with base level classifiers namely Naïve Bayes (NB), Support Vector Machine (SVM), Neural Networks, Multi-Layer Perceptron, Decision Tree and Gradient-boosted Decision Tree. Random Forest and Majority Voting based ensembles are also evaluated using the same tool.

3.2 Base level Classifier/Predictors

The selection of these base-level classifiers depends on certain aspects, such as they're widely used in the context of disease prediction. The short description of base level classifiers namely Naïve Bayes (NB), C4.5 Decision Tree, and Support Vector Machine (SVM) [11, 12] is as follows.

- Naïve Bayes (NB): NB classifier is based on the Naïve Bayes theorem and works according to the probability of events. The core assumption of NB is that all attributes are independent.
- Support Vector Machine (SVM): SVM is a classifier which constructs a decision area beside a margin via the nearest data points.
- C4.5 Decision Tree: It produces a set of rules. Classification decisions can be functional via these rules. This classifier utilizes the concept of information entropy.

3.3 Majority Voting based Ensemble method

We construct our approach via the use of a widely use ensemble technique namely majority voting. The voting ensemble technique is a common example of the multi-expert approach, which helps to combine the classifiers in a parallel fashion. Subsequently, each classifier trained on all data and contributes to a decision. Finally, the voting technique helps to generate the final solution [13, 14].

4. Results, Evaluation, and Discussion

Firstly, we compare the performance of base-level classifiers to each other. The outcomes are shown in Table 1.

4.1 Performance and Validation measures

In this paper, we perform 10-fold cross-validation and used performance measures such as Precision, Recall, F-measure, and Accuracy to determine the effectiveness of the proposed approach.

4.2 Workflow

We performed several experiments. However, workflow of experiments is generalized and can be shown through the following steps

Step-1: Extraction of datasets through an online repository.

Step-2: Application of pre-processing for data cleaning.

Step-3: Standard 10-fold cross validation is applied for training and testing.

Step-4: Computation of results for all individual classifiers.

Step-5: Select top-3 classifiers based on the performance measure such as accuracy and compose majority voting based ensemble.

Step-6: Compute results for Random Forest and Majority Voting based ensemble.

Step-7: Performance comparison is conducted for all individual as well as ensemble classifiers in order to identify the best classifier for lung cancer detection.

Table 1: Results and Performance Evaluation

Predictors	Acc. %	Prec. %	Rec. %	F-M. %
Auto MLP	78.33	68.75	70.57	69.65
Naïve Bayes	85.00	77.08	79.71	78.37
SVM	79.17	66.07	60.28	63.04
Decision Tree	78.33	68.75	70.57	69.65
GradientBoosted Tree	90.00	87.82	83.71	85.71
Nueral Network	71.67	63.18	66.57	64.83
Random Forest	79.17	39.15	50.00	43.91
Majority Voting MLP+GBT+SVM	**88.57**	**84.44**	**76.57**	**80.31**

The results indicate that SVM and C4.5 Decision Tree is outperformed with a minor difference in terms of F-measure. Similarly, in terms of accuracy, SVM outperformed the rest of predictors.

Subsequently, to achieve our research aim and investigate the performance of ensemble methods in the context of lungs cancer prediction, the results are computed for Random Forest and Majority Voting based ensemble. The results indicate that the majority voting based ensemble outperformed all other classifiers except Gradient Boosted Tree which performed best overall to achieve 90% accuracy.

In this research, we also found some threats. The first threat is related to a generalization of results since we performed our experiments on a single dataset. Consequently, the result may vary if we consider several experiments with different datasets. The secondly related threat to the selection of one ensemble technique. We report the results according to the functionality of the majority voting method.

5. Conclusions and Future Work

The focus of this research is an evaluation of machine learning classifiers as well as ensembles for lung cancer detection. For this purpose, individual classifiers including MLP, Neural Network, Decision Tree, Naïve Bayes, Gradient Boosted Tree, and SVM are assessed. Random forest and majority voting based ensembles are also analyzed for lung cancer prediction. It is observed that Gradient Boosted Tree outperformed all other individual and ensemble classifiers. In the future, we plan to evaluate other lung cancer and different disease datasets. Similarly, other ensemble technique like Stacking, Adaboost, and Bagging will be analyzed.

References

[1] Cabrera, J., Dionisio, A., & Solano, G. (2015, July). Lung cancer classification tool using microarray data and support vector machines. In Information, Intelligence, Systems, and Applications (IISA), 2015

[2] Yu, Z., Chen, X. Z., Cui, L. H., Si, H. Z., Lu, H. J., & Liu, S. H. (2014). Prediction of lung cancer based on serum biomarkers by gene expression programming methods. Asian Pacific Journal of Cancer Prevention, 15(21), 9367-9373.

[3] Wender, R., Sharpe, K. B., Westmaas, J. L., & Patel, A. V. (2016). The American Cancer Society's approach to addressing the cancer burden in the LGBT community, 3(1), 15-18.

[4] Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., & Cui, Q. (2013). HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. Nucleic acids research, 42(D1), D1070-D1074.

[5] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.

[6] Wender, R., Sharpe, K. B., Westmaas, J. L., & Patel, A. V. (2016). The American Cancer Society's approach to addressing the cancer burden in the LGBT community. LGBT health, 3(1), 15-18.

[7] Hussain. S, Keung. J, Khan. A. A., Performance evaluation of ensemble methods for software fault prediction, An experiment, Proceeding of ASWEC, 2015.

[8] Hussain. S, Asghar. Z, Ahmad. B, Ahmad. S., A step towards software corrective maintenance using RCM model, International Journal of Computer Science and Information Security, 4(1), 2009.

[9] Karakach, T. K., Flight, R. M., Douglas, S. E., & Wentzell, P. D. (2010). An introduction to DNA microarrays for gene expression analysis. Chemometrics and Intelligent Laboratory Systems, 104(1), 28-52.

[10] Marusyk, A., Almendro, V., & Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer?. Nature Reviews-Cancer, 12(5),323.

[11] Hussain. S, Saqib. S. M., Ahmad. B, Ahmad. S., Mapping of SOA and RUP : DOA as case study, Journal of Computing, 2(1), 2010.

[12] Hussain. S, Threshold analysis of design metrics to detect design flaws, Proceeding of ACM SAC, 2016.

[13] Zeng, X., Zhang, X., & Zou, Q. (2015). Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. Briefings in bioinformatics, 17(2), 193-203.

[14] Yu, H. L., Gao, S., Qin, B., & Zhao, J. (2012). Multiclass microarray data classification based on confidence evaluation. Genetics and molecular research: GMR, 11(2), 1357-1369.