HEPATOCELLULAR CARCINOMA PREDICTION THROUGH DIMENSIONALITY REDUCTION AND OPTIMIZED CLASSIFIERS

Parvathy S. Menon[1, a)], Devi C. Arati[1, b)]

[1]Center for Computational Engineering & Networking, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham 641112, India.

[a)] cb.en.p2dsc22011@cb.students.amrita.edu, [b)] cb.en.p2dsc22003@cb.students.amrita.edu

ABSTRACT

Hepatocellular carcinoma (HCC) is a common type of liver cancer worldwide, that causes the death of about 600,000 patients every year. Patients with HCC have rare chances of survival. The chances of survival increase, if the cancer is diagnosed early. The use of machine learning for HCC survival prediction is motivated by two key factors – enhancing the patient's quality of life with a quick and accurate diagnosis, and identifying relevant features to improve accuracy of the prediction model. Dimensionality reduction-based methods have shown state-of-the-art performance on many disease detection problems, which motivates the development of machine learning models based on reduced features dimension. The objective of this work is to determine the best combination of dimensionality reduction and optimization technique for classification of the data. Experimental results on publicly available HCC dataset indicate that a combination of PCA-GA-SVM shows improvement in the HCC prediction accuracy, with reduced computational time, considering a reduced dataset. Apart from performance improvement, the proposed method also shows lower complexity from two aspects, i.e., reduced processing time in terms of hyperparameters optimization and training time. The proposed method achieved accuracy of 75.61% and AUC of 0.7548.

INTRODUCTION

As per World Health Organization (WHO) reports, about 14.1 million new cancer patients and 8.2 million deaths are caused by cancer worldwide. Hepatocellular carcinoma (HCC), which is the malignancy of liver and is caused by chronic liver disease and cirrhosis, is one type of cancer. Recent research shows that the deadliest cancer around the world is HCC that is causing around 600,000 deaths each year. Moreover, liver cancer is ranked as the sixth commonly diagnosed cancer all over the world. These facts evidently show the impact of HCC on human life worldwide. Populations in East Asia and Pacific, South Asia, and parts of Sub-Saharan Africa are more susceptible to HCC, largely due to the outbreak of infection decades ago. It is possible to lower down the deaths caused by HCC if is diagnosed at the early stages. In case of advanced stage of disease, it cannot be cured but medications can help to support and prolong the life. In order to meet this objective, we need to exploit different techniques of data mining and machine learning to design an automated diagnostic system for efficient HCC prediction.

Hepatocellular carcinoma (HCC) is among the world's most common malignancies, accounting to more than 750,000 new cases every year, as reported by the Mayo Clinic, USA. The mortality rate of HCC is the third among all cancers. HCC resection and liver transplantation help patients with early HCC. Liver transplantation is seldom performed because of limited liver donors; therefore, liver resection remains the most widely used radical HCC treatment. Considering the seriousness of the situation, it is important that studies are done that avails prediction of this condition at its early stages. Previous research on this topic is summarized in this report.

Ding Y (2006) proposed two pre-processing methods for missing and heterogeneous data and used k-means clustering. Dong R. et.al. (2019) used a dataset of 4000 chronic hepatitis C patients diagnosed at Cairo University's multidisciplinary hospital is used with linear regression. The dataset is balanced using Synthetic Minority Over-sampling Technique (SMOTE) methods. The performance of LR and Neural Networks (NN) is 75.2% and 73%, respectively. CART, AD Tree, and REP-Tree models give an excellent area under the receiver-operating characteristic curve (AUROC), ranging between 95.5% and 99%. The high accuracy of HCC diagnosis ranges between 93.2% and 95.6%.

Santos et al. (2015) in their paper, introduced a new cluster-based oversampling method to improve survival prediction of hepatocellular carcinoma patients. The approach was robust to small and imbalanced datasets. Pre-processing procedures of this work included data imputation methods, that dealt with heterogeneous and missing data (HEOM). The machine learning classifier, K-Means was used to classify the underlying patient group. The final approach is applied in order to diminish the impact of underlying patient profiles with reduced sizes on survival prediction. It is based on K-means clustering and the SMOTE algorithm to build a representative dataset and use it as training example for different machine learning procedures (logistic regression and neural networks).

Cheng et al. (2006) optimized the parameters without degrading the SVM classification accuracy.

In the study conducted by Liaqat Ali et al. (2021) feature extraction was the point of initial focus, followed by optimization of the ML classifier. In the paper the authors proposed a new hybrid intelligent system that hybridized three algorithms, i.e., linear discriminant analysis (LDA) for dimensionality reduction, support vector machine (SVM) for classification and genetic algorithm (GA) for SVM optimization. The three models were hybridized to a black box model called LDA-GA-SVM. The experiment resulted in improved accuracy score. The proposed method achieved accuracy of 90.30%, sensitivity of 82.25%, specificity of 96.07% and Matthews Correlation Coefficient (MCC) of 0.804.

Vikas J. et.al (2019) proposed a method which used data science and machine learning to build a system to find how much the treatment can be successful instead of predicting how long one will survive. The authors also performed a comparative study on various ML classifiers, out of which the Logistic Regression Algorithm was found to offer the best performance with an accuracy of 99.49%. Valarmathi et.al.(2021) analysed the GridSearch, randomised search and genetic algorithm for tuning the hyperparameters of the RF and XGBoost classifiers for Cleveland dataset. By tuning the parameters for RF and XGBoost classifiers, better results were obtained for RF classifiers, particularly for RF-GA.

Our study focuses on reducing the computational complexity of classifiers using dimensionality reduction of the dataset. Dimensionality reduction can negatively affect the classification accuracy, which is undesirable. Hyperparameters of the classifier can be optimized to address the performance deterioration.

In this paper, Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are the two techniques used for dimensionality reduction and the classifier used is Support Vector Machine (SVM). The algorithms selected to optimize the hyperparameters of the classifiers are Genetic Algorithm (GA), Bayesian Optimization with Gaussian Process (BO-GP), and Bayesian Optimization with Tree structured Parzan Estimator (BO-TPE). The organization of the paper is as follows: Section II discusses the details of analysing and pre-processing the dataset along with the algorithms used for the study; Section III highlights the significance of the obtained results, provides comparative analysis, and comparison with the reported works from literature; Section IV provides an overall summary and outlines the future scope of the current work.

MATERIALS AND METHODS

Dataset Details and Pre-processing

The dataset used in this paper for HCC prediction is adopted from UCI machine learning repository. The dataset was collected at Coimbra's Hospital and University Centre (CHUC), Portugal, and contains samples collected from 204 subjects. The dataset contains 49 features in total, which can be subdivided into two groups, i.e., quantitative features and qualitative features. The number of quantitative features is equal to 23, and the number of qualitative features is equal to 26. The label of the dataset denotes survival at one year and can assume a value of 0 (dies) or 1 (lives/survives). For classification task, the dataset is separated into two parts, where 80 percent of the dataset is considered for training the model and 20 percent is considered for testing. Since the magnitude of data points vary significantly in this dataset, it is essential to normalize all the values to the same magnitude. Normalization is achieved using the relation,

$$X' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Here X' is the scaled data in the range (0,1), x is the original data, x_{max} and x_{min} are the maximum and minimum values in the data respectively.

Proposed Method

To improve survival classification accuracy, many data mining algorithms have been utilized for feature pre-processing. In this work, two feature extraction methods (dimensionality reduction techniques) are exploited – Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA). Both

methods are used to reduce the number of features in a dataset while retaining as much information as possible. LDA is used to transform the original features set into a reduced dimension, to improve the predictive capabilities of machine learning-based predictive models. The job of LDA is to maximize the fisher ratio which will result in minimum within-class scatter and maximum between-class scatter. PCA works by identifying the directions (components) that maximize the variance in a dataset. It seeks to find the linear combination of features that captures as much variance as possible. The first component is the one that captures the maximum variance, the second component is orthogonal to the first and captures the remaining variance, and so on.

The classification of patients' data is performed using a Support Vector Machine (SVM) classifier and a Random Forest (RF) classifier, where the input is the dimensionally reduced feature set. To select one among the two dimensionality reduction techniques, reduced feature vectors after performing PCA and LDA are separately fed into the SVM and RF and compared. Once dimensionality reduction is carried out on the input, the classifier performance would be deteriorated to a certain extent. To confront this issue, the parameters of the classifier models need to be optimized (Figure 1 shows the workflow of the study). Grid search is the most used method to meet this objective. However, since grid search is computationally very expensive, three hyperparameter optimization algorithms – Genetic Algorithm (GA), Bayesian Optimization with Gaussian Processes (BO-GP), and Bayesian Optimization with Tree structured Parzan Estimator (BO-TPE), are compared to tune the hyperparameters, 'gamma', 'C' and kernel, of SVM. The three algorithms are used to tune the six hyperparameters of the random forest classifier, namely criterion, max_depth, max_features, min_samples_leaf, min_samples_split and n_estimators.

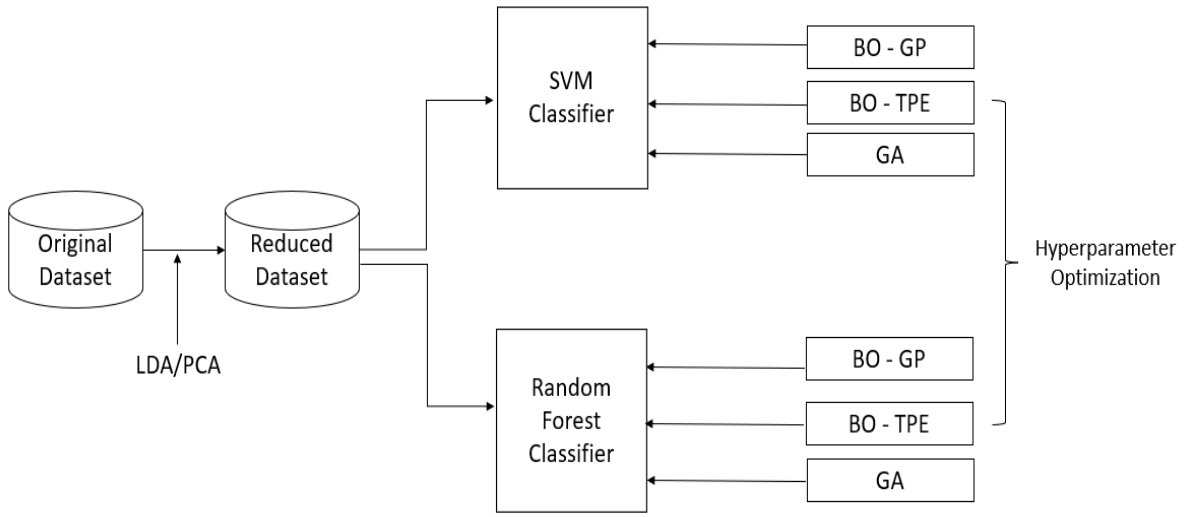

Figure 1: Workflow Diagram

GA randomly generates initial population consisting of chromosomes. The values of the three hyperparameters are directly coded in the chromosomes. To assess the performance of each

chromosome, a fitness function is designed. Bayesian Optimization (BO) is a sequential design strategy for global optimization of black-box functions that does not assume any functional forms. It is usually employed to optimize expensive-to-evaluate functions. The main idea behind Bayesian optimization for such a problem is to use all of the information gathered in previous iterations for performing the next step.

EXPERIMENTAL RESULTS AND DISCUSSION

The original dataset which contains 204 subjects and 49 features is reduced by PCA to 10 principal components, each with 204 data points. This means that the original data is reduced to a new dataset, which is almost a fifth its size. Classification and prediction tasks are then carried out using this dataset, which makes it computationally less tedious. The experimental results indicate that SVM performs better with an accuracy of 73.17% and 75.6% when PCA is used for dimensionality reduction when compared to the performance after dimensionality reduction using LDA, with the accuracy of 63.41% and 70.48% (Table 1).

Algorithm	f1-score	precision	recall	accuracy	AUC
LDA-SVM	0.6341	0.6345	0.6345	0.6341	0.6345
PCA-SVM	0.7317	0.7321	0.7321	0.7317	0.7321
LDA-RF	0.7028	0.7150	0.7047	0.7073	0.7048
PCA-RF	0.7524	0.7662	0.7535	0.7560	0.7786

Table 1: Performance comparison of LDA and PCA

After dimensionality reduction, when the dataset is fed into an SVM classifier, the classification accuracy dropped. From Table 2, it is clear that the model which is a combination of PCA-GA-SVM significantly improved the prediction accuracy to 75.61%, from the previous 73.21%. But when the optimization alorithms are used to tune the hyperparameters of the RF classifier, for the combination PCA-BO-GP-RF the accuracy increased to 80.48%. A combination of Area Under Curve (AUC) and Accuracy is used to evaluate the models. Receiver Operating Characteristic Area Under Curve (ROC-AUC) is a curve that maps the relationship between the True Positive Rate and the False Positive Rate of the model across different thresholds. Accuracy is one of the most common and simplest metrics used for validation in machine learning applications, that determines the percentage of correct prediction by any model. Table 2 shows that PCA-BO-GP-RF is the one model that performs best among all models considered, in terms of accuracy (80.48%) and AUC (0.8024). Table 3 shows the runtime for different combinations of optimizers with classifiers. It can be observed that the run-time for PCA-BO-GP-RF is around 16 seconds while for PCA-GA-SVM it is 6 seconds. But the accuracy and AUC value is the highest for PCA-BO-GP-RF. Based on the requirement of either higher accuracy

or lesser runtime either of the two can be chosen. the Figure 2 shows the ROC plots for all models considered for the study.

Algorithm	f1-score	precision	recall	accuracy	AUC
PCA-SVM-Gridsearch	0.7291	0.7365	0.7297	0.7317	0.7298
PCA-BO-GP-SVM	0.7291	0.7365	0.7298	0.7317	0.7548
PCA-BO-TPE-SVM	0.6829	0.6833	0.6833	0.6829	0.6833
PCA-GA-SVM	0.7559	0.7559	0.7559	0.7561	0.7560
PCA-RF-Gridsearch	0.7524	0.7662	0.7536	0.7561	0.7536
PCA-BO-GP-RF	0.8019	0.8175	0.8023	0.8048	0.8024
PCA-BO-TPE-RF	0.32786	0.2439	0.5	0.4878	0.5
PCA-GA-RF	0.7784	0.7868	0.7785	0.7805	0.7786

Table 2: Performance evaluation of SVM hyperparameter optimization algorithms after PCA

Combination	Duration (HR:MM:SS)
LDA-SVM	Duration: 0:00:00.795599
PCA-SVM	Duration: 0:00:00.679973
PCA-BO-GP-SVM	Duration: 0:00:35.308632
PCA-BO-TPE-SVM	Duration: 0:00:00.962403
PCA-GA-SVM	Duration: 0:00:06.859140
PCA-BO-GP-RF	Duration: 0:00:16.678570
PCA-BO-TPE-RF	Duration: 0:06:42.260743
PCA-GA-RF	Duration: 0:21:22.500624

Table 3: Run time for different combinations of optimizers with classifiers

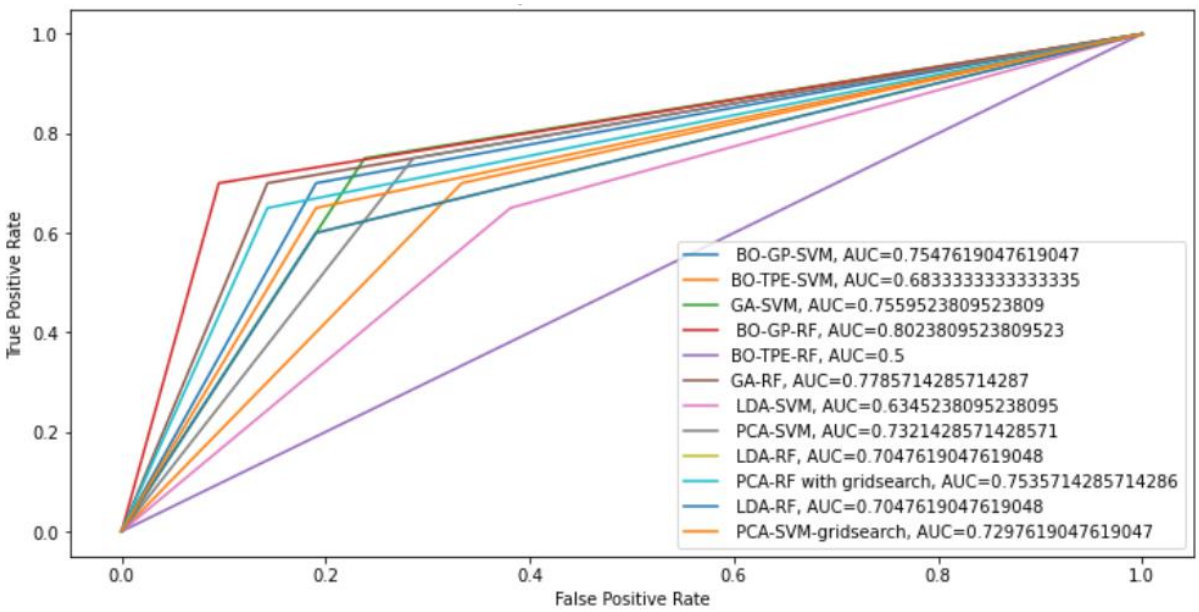

Figure 2: Comparison of ROC Charts

CONCLUSION

In this work, different techniques were used to optimize the hyperparameters of SVM classifier, used on a dataset that was dimensionally reduced. This enabled the prediction of chance of survival of an HCC patient computationally less complex. Although dimensionality reduction of the dataset could be done using PCA or LDA, PCA produced better results with the classifier selected (SVM), with an accuracy of 73.17% for SVM and 75.6 for RF classifiers. The analysis showed that while working with SVM, GA would be the most suitable hyperparameter optimization technique and while working with RF, BO-GP would be the most suitable hyperparameter optimization technique. The combination of dimensionality reduction using PCA, classification using RF, with the hyperparameters optimized using BO-GP produced a prediction model with an accuracy of 75.61% and AUC of 0.7548. Based on the requirement of lesser runtime or higher accuracy either PCA-GA-SVM or PCA-BO-GP-RF can be chosen. From This prediction is expected to improve the oncologists' quality of decision-making during diagnosis of HCC patients. It was also observed that the proposed method shows lower complexity in terms of processing time. The lower complexity was observed from two aspects, i.e., hyperparameters optimization and training time.

REFERENCES

1. Ding, Y.;Wilkins, D. Improving the Performance of SVM-RFE to Select Genes in Microarray Data. In BMC Bioinformatics; Springer: Berlin/Heidelberg, Germany, 2006. [CrossRef]
2. Dong, R.; Yang, X.; Zhang, X.; Gao, P.; Ke, A.; Sun, H.-C.; Zhou, J.; Fan, J.; Cai, J.; Shi, G. Predicting overall survival of patients with hepatocellular carcinoma using a three-category method based on DNA methylation and machine learning. J. Cell. Mol. Med. 2019, 23, 3369–3374. [CrossRef]
3. Miriam Seoane Santos, Pedro Henriques Abreu, Pedro J. García-Laencina, Adélia Simão, Armando Carvalho, A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients, Journal of Biomedical Informatics, Volume 58, 2015, Pages 49-59.
4. Ali, M.A.S.; Orban, R.; Rajammal Ramasamy, R.; Muthusamy, S.; Subramani, S.; Sekar, K.; Rajeena P. P., F.; Gomaa, I.A.E.; Abulaigh, L.; Elminaam, D.S.A. A Novel Method for Survival Prediction of Hepatocellular Carcinoma Using Feature-Selection Techniques. *Appl. Sci.* 2022, *12*, 6427.
5. Valarmathi et.al., Heart disease prediction using hyper parameter optimization (HPO) tuning.", Biomedical Signal Processing and Control, 2021.
6. Ali, L., Wajahat, I., Amiri Golilarz, N. et al. LDA–GA–SVM: improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine. Neural Comput & Applic 33, 2783–2792 (2021).

7. Yan Wei, Ni Ni, Dayou Liu, Huiling Chen, Mingjing Wang, Qiang Li, Xiaojun Cui, Haipeng Ye, "An Improved Grey Wolf Optimization Strategy Enhanced SVM and Its Application in Predicting the Second Major", Mathematical Problems in Engineering, vol. 2017, Article ID 9316713, 12 pages, 2017.
8. Cheng-Lung Huang, Chieh-Jen Wang, A GA-based feature selection and parameters optimization for support vector machines, Expert Systems with Applications, Volume 31, Issue 2, 2006, Pages 231-240, ISSN 0957-4174.