2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Identification of Significantly Expressed Gene Mutations for Automated Classification of Benign and Malignant Prostate Cancer

Robert B. Eshun
Dept. of Comp. Data Sci. & Eng.
North Carolina A&T State University
Greensboro, North Carolina, USA
rbeshun@aggies.ncat.edu

A.K.M. Kamrul Islam
Dept. of Comp. Data Sci. & Eng.
North Carolina A&T State University
Greensboro, North Carolina, USA
akislam@ncat.edu

Marwan U. Bikdash
Dept. of Comp. Data Sci. & Eng.
North Carolina A&T State University
Greensboro, North Carolina, USA
bikdash@ncat.edu

ABSTRACT

Among males, prostate cancer (Pca) is the cancer type with the highest prevalence and the second leading cause of cancer deaths. The current screening methods for prostate cancer lack effectiveness such as prostate-specific antigen (PSA) and digital rectal exam (DRE). Machine learning models have been used to predict Pca progression, Gleason score, and laterality. In this research paper, we have employed novel Machine learning techniques such as Bayesian approach, Support vector machines (SVM), Decision Trees, Logistic Regression, K-Nearest Neighbors, Random Forest and AdaBoost for detecting malignant prostate cancers from benign ones. Moreover, different feature extracting strategies are proposed to improve the detection performance and identify potential genomic biomarkers. The results show the Lasso feature set yielded high performance from the models with SVM achieving exemplary classification accuracy of 97%. The Lasso and SVM combination reported many significant biomarker genes and gene mutations including but not restricted to CA2320112, CA2328529, and CA2436168.

CCS CONCEPTS

• **Applied computing** → *Life and medical sciences*; • **Life and medical sciences** → Bioinformatics.

KEYWORDS

Prostate Cancer (Pca), Machine Learning (ML), Feature Selection, Gene Expression Profiles, Lasso

ACM Reference Format:
Robert B. Eshun, A.K.M. Kamrul Islam, and Marwan U. Bikdash. 2021. Identification of Significantly Expressed Gene Mutations for Automated Classification of Benign and Malignant Prostate Cancer . In *Proceedings of ACM Conference (Conference'21).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'21, August 2021, Washington, DC, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Prostate cancer (Pca) is one of the most common forms of cancer and the third leading cause of cancer death in North America [10, 27]. The prediction of the pathologic stage of PCa before an intervention enables improved patient prognosis and management for treatment planning [7]. Discovery of new diagnostic biomarkers for effective prostate cancer detection and management strategies for prostate cancer is therefore highly sought after. Traditional methods for detecting prostate cancer such as prostate specific antigen (PSA) blood test, trans-rectal ultrasound image (TRUS) guided biopsy, and digital rectal exam (DRE) have inherent limitations. PSA blood test statistical results shows a specificity of 61% and a low sensitivity of 34.9% [9]. The recent advance in artificial intelligence (AI) and computational capabilities have facilitated robust pattern recognitions from diverse sources of information including large heterogenous data sets, images and genomic data. Although the application of AI in medicine remains in its early stages, some studies have introduced different prediction models for advanced PCa using conventional machine learning [7]. Presently there is growing interest in medical research in investigating genomic alterations in patients diagnosed with benign and malignant tumors to identify measurable changes distinguishing the two patient groups, and application of ML to genomics data to accurately distinguish the two patient groups. This study aims to: (a) Apply feature selection methods to decrease the dimensions of the gene expression data to (b) identify a small set of predictors or potential genomic biomarkers to distinguish the histological subtypes. (c) Apply machine learning algorithms to construct computational models that differentiate between malignant and benign tumor types (d) Evaluate the performance of the machine learning models on the significantly expressed genes. This paper is organized as follows. Section 2 presents a review of literature on applications of ML and simple DL models for prostate cancer prediction-based on genomic data, and Section 3 provides descriptions of the data preparation methods, feature selection and classifier techniques used to develop the predictive models. In Section 4, the results of the experiment are discussed in details followed by the concluding remarks in Section 5.

2 RELATED WORK

Critical aspects of the biology and molecular basis for prostate malignancy remain poorly understood. To reveal fundamental differences between benign and malignant growth of prostate cells, gene expression profiling of prostate cancer using cDNA microarrays consisting of 6500 human genes revealed for the first time that significant and widespread differences in gene expression patterns exist between benign and malignant growth of the prostate gland [17]. [21] performed a comprehensive gene expression analysis

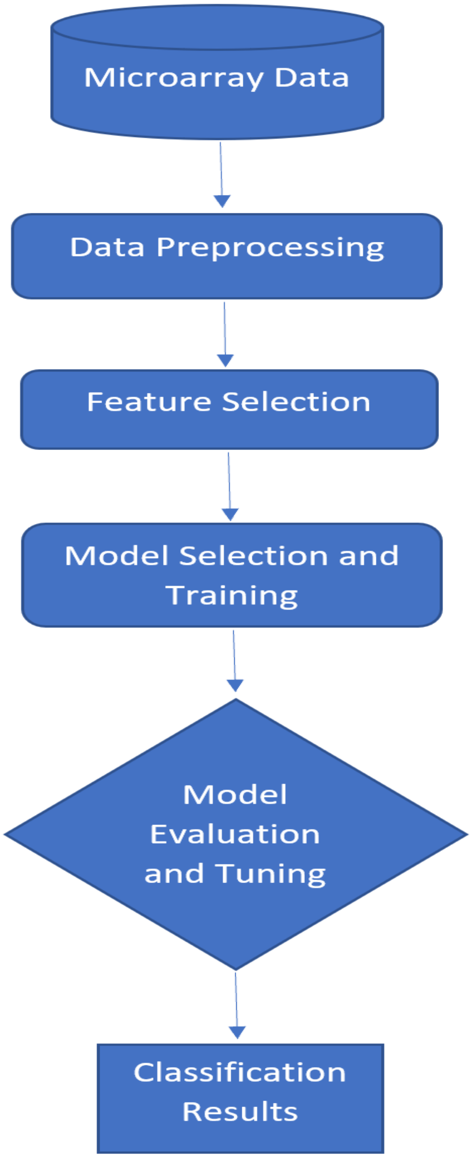

Figure 1: Workflow of Model

on samples including prostate cancer tissues, prostate tissues adjacent to tumor and found that gene expression patterns can be used to predict the aggressiveness of prostate cancer [28]. Previous studies on the 11-tumor database have shown that machine learning consistently performs well in multi-cancer type scenarios and demonstrated that Logistic Regression achieves efficient and accurate tumor classification based on gene expression (microarray) data with accuracy of 90.6% [25] In their paper, [1] present and validates various classification techniques on supervised machine learning (ML) for predicting prostate cancer. A modified Logistic Regression (LR) classifier was proposed based on both clinical and tumor stage characteristic showed improvement in accuracy and positive prediction value (PPV) as compared to existing classifiers. Machine learning models were recently utilized to predict the outcomes of Pca, and to find potential biomarkers for the clinical features of the disease. In their review of machine learning methods, [2] reported high performance of the genomic data-based models with an accuracy of more than 90%, and identification of many biomarkers genes and genes transcripts including but not restricted to CARNA22 in prediction of including Pca progression, Gleason score, and laterality. A method that uses machine learning techniques to identify transcripts that correlate with prostate cancer development and progression was applied by [14]. They isolated transcripts that have the potential to serve as prognostic indicators and may have tremendous value in guiding treatment decisions. Analysis of normal versus malignant prostate cancer data sets indicates differential expression of the genes HEATR5B and DDC [3]

3 METHODS

3.1 Data Preparation

The dataset for the study was accessed from the Gene Expression Omnibus database. The dataset of GSE94767 consists of 236 samples from fresh frozen tissue from the prostatectomies of 154 prostate cancer (Pca) patients. 185 samples were malignant tissue and 51 samples morphologically benign tissue. Gene-level signal estimates were derived from CEL files generated from Affymetrix GeneChip Exon 1.0 ST arrays using the robust multiarray analysis algorithm8 implemented in the Affymetrix Expression Console software package [16]. The dataset was pre-processed to remove the insufficient fibroblast samples and the few instances of null data imputed with the feature mean values. We separated the cohort of 236 samples into two different sets, the training set with 142 samples for training and validation, and the test set with 95 samples where the training/ test ratio was 60%, 40%.

3.2 Feature Extraction

The goal of supervised feature selection is to find a subset of input features that are responsible for predicting output values. The large number of features increases the computational costs and leads to the problem of curse of dimensionality [8, 22]. For this project we used four feature selection methods to avoid the curse and enhance the generalization ability of the model [20, 21]. The features selection methods implemented are, specifically, the fisher-score metric, Extra-Tree classifier, Hilbert Schmidt Independence Criterion Lasso (HSIC) and Lasso regularization. Fisher-score methods are feature selection methods based on similarity that assess feature importance in terms of the ability to preserve data similarity. The Lasso methods are based on sparse learning and employ regularization terms to reduce the weights of unimportant features in the model [11]. The HSIC is a nonlinear feature selection method considering the nonlinear input and output relationship. HSIC Lasso employs the HSIC to measure dependency between variables [5]. The Extra Trees Forest ensemble method which is a decision tree based feature selector using the Gini index criteria. [18] A feature selection repository of Python named “scikit feature” was used to implement the fisher-score method. This feature selection models reduced the

Table 1: Accuracy of Models on Feature Sets Including Baseline

			Accuracy (%)		
Type	**Lasso**	**Fisher-Score**	**Extra-Tree**	**HSIC**	**All Features**
Naive Bayes	85.3	76.8	75.8	81.1	71.6
Logistic Reg.	95.8	77.9	80.0	83.2	72.6
KNN	83.2	77.9	82.1	83.2	72.6
CART	76.8	78.9	80.0	75.8	72.6
SVM	96.8	77.9	80.0	83.2	74.7
RF	85.3	80.0	84.2	80.0	78.9
AdaBoost	88.4	72.6	83.2	86.3	73.7

Table 2: Classification Performance of Models on Lasso Feature Set

Type	Precision	Recall	Specificity	F1	AUC
Naive Bayes	0.87	0.85	0.61	0.86	0.92
Logistic Reg.	0.96	0.96	0.86	0.96	0.99
KNN	0.82	0.83	0.70	0.81	0.81
CART	0.78	0.77	0.24	0.77	0.52
SVM	0.97	0.97	0.90	0.97	0.99
RF	0.86	0.85	0.87	0.83	0.85
AdaBoost	0.88	0.88	0.85	0.87	0.88

number of genes from approximately 22,000 to four different sets with 50 genes.

3.3 Classifiers

Multiple classification methods were applied on the data to identify which methods separate the locations better. NB is based on Bayes's rule of conditional probability. It uses all attributes and allows them to make contributions to the decision as if they were all equally important and independent of one another, with the probability denoted by (1).

$$P(H|E) = \frac{\prod_{i=1}^{n} P(E_i|H) \times P(H)}{P(E)} \tag{1}$$

where $P(H)$ denotes the probability of event H, $P(H|E)$ denotes the probability of event H with the condition of occuring event E, n is the n^{th} attribute of the instance, H is the outcome in question, and E is the combination of all the attribute values [23]. SVM maps the input into a high-dimensional feature space and finds a separating hyperplane that maximizes the margin between two classes in this space [13]. The solution of the optimal hyperplane can be written as a combination of a few input points that are called support vectors [12]. Logistic Regression (LR) forms a predictor variable that is a linear combination of the feature variables. The values of this predictor variable are then transformed into probabilities by a logistic function. This method is widely used for 2-class prediction in biostatistics. [6] K-Nearest Neighbors (KNN) is another classification technique that works on the idea of assigning the label of a classified data point to an unclassified data point nearest to it. Starting with the unclassified class data point as the input vector in the feature space, it is assigned to the class in which majority of its K nearest data points belong to [12] Decision trees (with the CART algorithm) interpolates learned knowledge from a dataset into a tree which is governed by if-then rules . Each node in the tree represents the learning variable which recursively checks how accurately can each node classify the labeled data, by calculating the information gain and entropy of each node. This learning process leads to selecting the best node as the parent node and the children nodes carry the possible values of the selected input data [6]. Random forest (RF) is an ensemble learning approach, where many decision trees are generated during the training stage, with each tree based on a different subset of features and trained on a different part of the same training set. During the classification of unseen examples, the predictions of the individually trained trees are then agglomerated using the majority vote. This bootstrapping procedure is found to efficiently reduce the high variance that an individual decision tree is likely to suffer from [4, 14]. The RF operates by constructing a multitude of decision trees on various subsamples of the dataset and results in a mean prediction of decision trees to improve accuracy and avoid over-fitting [19]. AdaBoost uses the complete training dataset to train the weak learners, where the training examples are reweighted in each iteration to build a strong classifier that learns from the mistakes of the previous weak learners in the ensemble [24]. In order to evaluate the model, certain metrics are used, such as accuracy which calculates the ratio of correctly classified samples against the total number of samples. Two other metrics used are sensitivity and specificity, where sensitivity indicates, how well the test predicts one category and specificity measures how well the test predicts the other category. Another important metric is the AUC (Area Under The Curve), where the curve is the ROC (Receiver Operating Characteristics) curve. It is a graph that shows the performance of a classification model by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR).

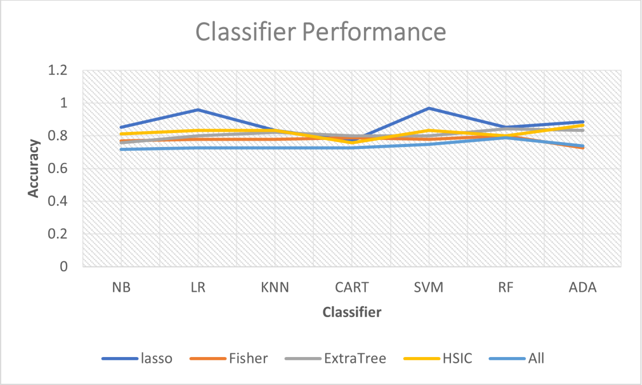

Figure 2: Comparison of Model Accuracy on Different Feature Sets

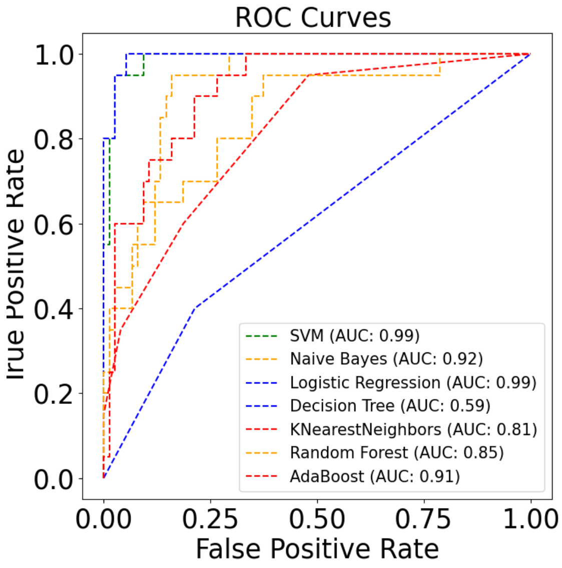

Figure 3: Area Under ROC Curves of Models on Lasso Predictors

3.4 Performance Evaluation

The analysis uses the sensitivity, specificity and precision as performance metrics to measure the efficacy of the models. The analysis is experimented across all classes of the positive and negative examples to obtain a more balanced assessment of the true positive and true negative predictions. The primary performance metrics are the area under Receiver Operator Characteristic curve (AUC), accuracy and F1 score. Given that the dataset is unbalanced, the focus is on the AUC and F1 score.

4 RESULTS & DISCUSSION

The four feature selection methods consisting of Fisher-score, HSIC methods, Extra-tree classifier and Lasso model were each used separately to analyze the dataset and identify the fifty most significant features based on the selection criteria. The extracted feature separate sets (each consisting of 50 features or genomic profiles) were then used as input to evaluate how well the supervision-based machine learning models predict the benign/malignant groups. The results of the performance accuracy of the models based on the separate feature sets is presented in Table 1. The performance of the various classification algorithms with all features is provided as a benchmark. The accuracy results show all the models performing with similar accuracy of average 75% based on all features. This average score indicates good predictive performance on all features which is better than a random guess (50%). Noticeably, the models achieved marginal improvement or decrement in the scores on the fisher-score selected feature set compared to the baseline with the lowest scoring ADA model recording an increase of 1% points to 72.6% and the RF and BAG increasing their scores by 1% to 80%. The decision tree classifier (CART) achieved the best improvement in performance with a 6% increase to 78.9%. Majority of the models produced accuracy scores of >= 80% on the feature set selected with the Extra-tree method. The least performing on the dataset was the Naïve Bayes classifier that scored 75.8% accuracy and the best results were observed for the Random Forest classifier with accuracy of 84.2%. The highest accuracy score on the HSIC selected feature set was 86.3% produced by the ADA classifier and the three classifiers SVM, CART and LR yielded the same accuracy score of 83.2%. The best results from the reduced datasets was produced from the Lasso selected features with the top performers achieving >10% improvement in accuracy over the scores recorded in the next best feature (HSIC) set. The best performer was observed to be the SVM classifier which achieved very high accuracy of 96.8% on the Lasso dataset and which represented a > 20% improvement on its baseline score. Following closely with similar robust prediction capability was the Logistic Regression estimator that produced an accuracy of 95.8% which was also 20% greater than the baseline. The other classifiers were observed to yield very good performance on the dataset with the ADABoost, RF, NB, and KNN scoring accuracies of 88.4%, 85.3%, 85.3% and 83.2% respectively. The decision tree (CART) classifier was the least performing on the dataset with accuracy of 76.8%. The comparison of the accuracy scores is illustrated in Figure 2. The robust performance of the SVM and LR classifiers on the Lasso dataset were demonstrated on their accurate predictions of the positive and negative classes with the SVM yielding precision, sensitivity and F1 values of 0.97 and exemplary AUC scores of 0.99 as shown in Table 2). The LR similarly scored 0.96 on precision, recall and F1 and achieved AUC of 0.99. The performance scores on the evaluation metrics of precision, recall and specificity are found on Table 2 and the area under ROC curves are illustrated in Figure 3. The Lasso method was found to have produced the most significant features due to the capability of the selected features or biomarker genes to produce robust accurate prediction on data. The heatmap of Figure 5 shows that the selected features have low correlation between the members. The 50 canonical alleles identified by the Lasso feature selection method, including CA2320112, CA2328529, CA2436168, CA2841386, CA2844385, CA2946225, CA3265175, CA3275386, 3275392, CA3275524, CA3551874, CA3556723, CA3819651, CA3819695, CA3823379, and CA3980930 were observed to demonstrate a significant association with whether a prostate cancer tumor was malignant or benign and are potential

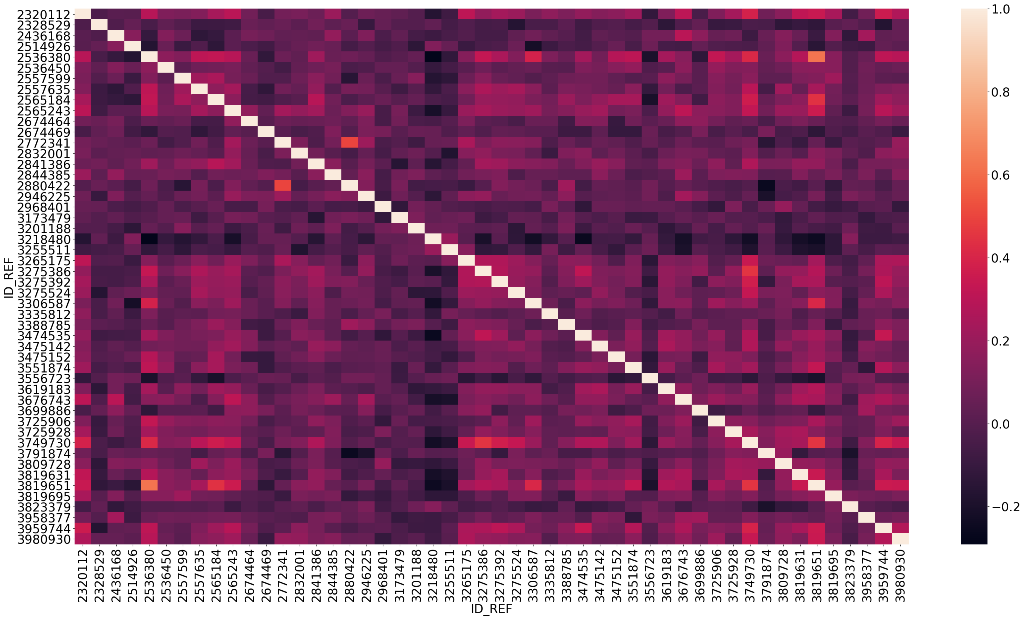

Figure 4: Heatmap Showing Correlation between Gene Variants of Lasso Set

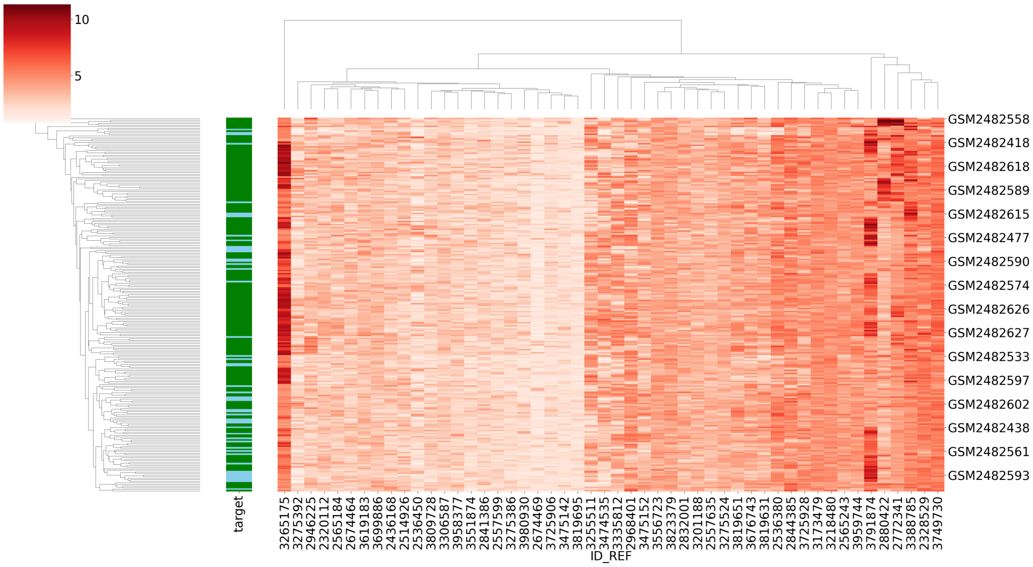

Figure 5: Clustermap of Gene Variants Indicating Target Group of Samples

informative targets for the treatment considerations and diagnosis. The cluster map of the gene and gene mutations is presented in Figure 5.

4.1 Discussion

The fisher-score metric produced predictors that resulted in the classifier models scoring average accuracy of 77% (with RF and

AdaBoost scoring the highest and lowest precision on the predictors at 80% and 72.6% respectively) which was marginally better than the baselines. There was an improvement in accuracy on the Extra-Tree predictor set with the RF and AdaBoost models now observed to obtain a high score of 84.2% and 83.2% respectively and the NB trailing in precision in the group with 75.8% accuracy. All models scored >80% on the HSIC predictor set except the Decision Tree classifier which recorded classification accuracy of 75.8%. The AdaBoost methods showed robust performance on these predictors with 86.3% prediction correctness, and the LR, SVM and KNN models produced good capability in differentiating samples with 83.2% accuracy. The Lasso method identified predictors that resulted in the SVM and LR models scoring an exemplary 96.8% and 95.8% accuracy respectively on the dataset (which is >20% on the baselines and 10% more than the best precision from the HSIC feature set.) Given high dimensional data with several highly correlated variables, all of which are related to some extent to the response variable, lasso tends to pick only one or a few of them and shrinks the rest to 0 [26]. The gene variants or alleles identified by the Lasso method The LASSO is noted to be well suited where the number of predictors may be large relative to the sample size, and the predictors may be correlated [29]. Given the high dimensional data, the lasso estimator seems to pick only the significant gene variants which are highly associated with the outcomes and shrinks the other relevant highly correlated genes to zero. The heatmap of Figure 4 shows that the selected features have low correlation between the members. This is also suggests the highly correlated genes, which normally share one common biological pathways, may not be relevant genes in the determination of prostate tumor malignancy. The performance of SVM degrades with gene expression profiles are noisy due to both biological and technical variations in the data and with the Lasso feature set devoid of such challenges, the SVM performs flawlessly. Logistic regression on the other hand achieves excellent results on the lasso features due to the low correlations among the predictors and reduced incidence of influential outliers [15].

5 CONCLUSIONS

The machine learning models in general yielded higher prediction scores on the Lasso feature sets by a wide margin. The combination of the SVM and LR classifiers applied to features selected from the Lasso method produced the most robust performance in prediction with accuracy scores of 97% and 96% respectively. The models detected many biomarker genes or alleles including CA2320112, CA2328529, CA2436168, CA2841386 and CA2844385 and showed they are highly correlated with the classes studied. The empirical results indicated that the Lasso method was able to identify the most significant set of genes and showed that the selected genes can be used to differentiate the tumor stages. This study revealed the proposed method can learn discriminative genes in prostate tumor and classify the malignant or benign cancer accurately. The classification model could be further applied in the clinical practice to provide valuable information for cancer treatment and precision medicine. The limitation of this study is that it assesses the prostate tissue samples in the dataset without considering the fibroblasts. Future research of this study is to conduct a multi-class classification of the prostate samples including fibroblasts and explore modifications of the experimental processes for the application of the deep learning models.

REFERENCES

[1] Mansoor Alam, Mansour Tahernezhadi, Hari Kiran Vege, P Rajesh, et al. 2020. A Machine Learning Classification Technique for Predicting Prostate Cancer. In *2020 IEEE International Conference on Electro Information Technology (EIT)*. IEEE, 228–232.

[2] Abedalrhman Alkhateeb, Govindaraja Atikukke, and Luis Rueda. 2020. Machine learning methods for prostate cancer diagnosis. J. *Journal of Cancer* 1, 3 (2020), 70–75.

[3] Abedalrhman Alkhateeb, Iman Rezaeian, Siva Singireddy, Dora Cavallo-Medved, Lisa A Porter, and Luis Rueda. 2019. Transcriptomics Signature from Next-generation Sequencing Data Reveals New Transcriptomic Biomarkers Related to Prostate Cancer. *Cancer informatics* 18 (2019), 1176935119835522.

[4] Usman Bashir, Bhavin Kawa, Muhammad Siddique, Sze Mun Mak, Arjun Nair, Emma Mclean, Andrea Bille, Vicky Goh, and Gary Cook. 2019. Non-invasive Classification of Non-small Cell Lung Cancer: A Comparison Between Random Forest Models Utilising Radiomic and Semantic Features. *The British journal of radiology* 92, 1099 (2019), 20190159.

[5] Héctor Climente-González, Chloé-Agathe Azencott, Samuel Kaski, and Makoto Yamada. 2019. Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data. *Bioinformatics* 35, 14 (2019), i427–i435.

[6] Chris Ding and Hanchuan Peng. 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology* 3, 02 (2005), 185–205.

[7] Okyaz Eminaga, Omran Al-Hamad, Martin Boegemann, Bernhard Breil, and Axel Semjonow. 2020. Combination possibility and deep learning model as clinical decision-aided approach for prostate cancer. *Health informatics journal* 26, 2 (2020), 945–962.

[8] Robert Eshun, AKM Kamrul Islam, and Marwan Bikdash. 2021. Histological classification of non-small cell lung cancer with RNA-seq data using machine learning models. In *2021 ACM Southeast Conference*. ACM.

[9] Osama Hamzeh, Abedalrhman Alkhateeb, Julia Zheng, Srinath Kandalam, and Luis Rueda. 2020. Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data. *BMC bioinformatics* 21, 2 (2020), 1–10.

[10] Osama Hamzeh, Abedalrhman Alkhateeb, Julia Zhuoran Zheng, Srinath Kandalam, Crystal Leung, Govindaraja Atikukke, Dora Cavallo-Medved, Nallasivam Palanisamy, and Luis Rueda. 2019. A hierarchical machine learning model to discover gleason grade-specific biomarkers in prostate cancer. *Diagnostics* 9, 4 (2019), 219.

[11] Yong Han, Yuan Ma, Zhiyuan Wu, Feng Zhang, Deqiang Zheng, Xiangtong Liu, Lixin Tao, Zhigang Liang, Zhi Yang, Xia Li, et al. 2021. Histologic Subtype Classification of Non-small Cell Lung Cancer Using PET/CT Images. *European journal of nuclear medicine and molecular imaging* 48, 2 (2021), 350–360.

[12] Samuel H Hawkins, John N Korecki, Yoganand Balagurunathan, Yuhua Gu, Virendra Kumar, Satrajit Basu, Lawrence O Hall, Dmitry B Goldgof, Robert A Gatenby, and Robert J Gillies. 2014. Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features. *IEEE access* 2 (2014), 1418–1426.

[13] AKM Kamrul Islam and Saeid Belkasim. [n.d.]. Ensemble of SVM for Colorectal Cancer Classification from Microarray Gene Expression Data. ([n. d.]).

[14] Nathan T Johnson, Andi Dhroso, Katelyn J Hughes, and Dmitry Korkin. 2018. Biological Classification with RNA-seq Data: Can Alternatively Spliced Transcript Expression Enhance Machine Learning Classifiers? *Rna* 24, 9 (2018), 1119–1132.

[15] Alboukadel Kassambara. 2018. *Machine learning essentials: Practical guide in R*. Sthda.

[16] Bogdan-Alexandru Luca, Daniel S Brewer, Dylan R Edwards, Sandra Edwards, Hayley C Whitaker, Sue Merson, Nening Dennis, Rosalin A Cooper, Steven Hazell, Anne Y Warren, et al. 2018. DESNT: a poor prognosis category of human prostate cancer. *European urology focus* 4, 6 (2018), 842–850.

[17] Jun Luo, David J Duggan, Yidong Chen, Jurga Sauvageot, Charles M Ewing, Michael L Bittner, Jeffrey M Trent, and William B Isaacs. 2001. Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer research* 61, 12 (2001), 4683–4688.

[18] Oskar Maier, Matthias Wilms, Janina von der Gablentz, Ulrike M Krämer, Thomas F Münte, and Heinz Handels. 2015. Extra tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences. *Journal of neuroscience methods* 240 (2015), 89–100.

[19] Yashwanth Karthik Kumar Mamidi and Tamjidul Hoque. 2020. Classification of Prostate Cancer Patients into Indolent and Aggressive Using Machine Learning. (2020).

[20] Hiroshi Motoda and Huan Liu. 2002. Feature selection, extraction and construction. *Communication of IICM (Institute of Information and Computing Machinery, Taiwan) Vol* 5, 67-72 (2002), 2.

[21] Khurram Rabby, AKM Kamrul Islam, Saeid Belkasim, and Marwan Bikdash. 2021. Wavelet transform-based feature extraction approach for epileptic seizure

classification. In *2021 ACM Southeast Conference*. ACM, 164–169.

[22] Md Khurram Monir Rabby, AKM Kamrul Islam, Saeid Belkasim, and Marwan U Bikdash. 2021. Epileptic Seizures Classification in EEG Using PCA Based Genetic Algorithm Through Machine Learning. In *Proceedings of the 2021 ACM Southeast Conference*. 17–24.

[23] Sterling Ramroach, Ajay Joshi, and Melford John. 2020. Optimisation of Cancer Classification by Machine Learning Generates an Enriched List of Candidate Drug Targets and Biomarkers. *Molecular omics* 16, 2 (2020), 113–125.

[24] Sebastian Raschka and Vahid Mirjalili. 2017. Python Machine Learning: Machine Learning and Deep Learning with Python. *Scikit-Learn, and TensorFlow. Second edition ed* (2017).

[25] Reinel Tabares-Soto, Simon Orozco-Arias, Victor Romero-Cano, Vanesa Segovia Bucheli, José Luis Rodríguez-Sotelo, and Cristian Felipe Jiménez-Varón. 2020. A Comparative Study of Machine Learning and Deep Learning Algorithms to Classify Cancer Types Based on Microarray Gene Expression Data. *PeerJ Computer Science* 6 (2020), e270.

[26] Sijian Wang, Bin Nan, Saharon Rosset, and Ji Zhu. 2011. Random lasso. *The annals of applied statistics* 5, 1 (2011), 468.

[27] Sunghwan Yoo, Isha Gujrathi, Masoom A Haider, and Farzad Khalvati. 2019. Prostate cancer detection using deep convolutional neural networks. *Scientific reports* 9, 1 (2019), 1–10.

[28] Yan Ping Yu, Douglas Landsittel, Ling Jing, Joel Nelson, Baoguo Ren, Lijun Liu, Courtney McDonald, Ryan Thomas, Rajiv Dhir, Sydney Finkelstein, et al. 2004. Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *Journal of clinical oncology* 22, 14 (2004), 2790–2799.

[29] Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* 67, 2 (2005), 301–320.