Title

Liver cancer biomarkers Identification using gene expression data to facilitate early diagnosis and prognosis of the disease.

Abstract

Cancer is a terrible heterogeneous illness that refers to the abnormal proliferation of cell tissue. As a result, precisely recognizing the presence of cancer cells becomes critical. Given the number of individuals affected by cancer, we need to develop better methods of detection and treatment. Artificial intelligence (AI), particularly machine learning algorithms, is a hot study topic in medical science. It is capable of making conclusive interpretations of "big"-sized complicated data and so looks to be the most effective instrument for the analysis and comprehension of multi-omics data for patient-specific observations. Many of these techniques, including Decision Trees (DTs), Neural Networks (NN), and Support Vector Machines (SVM), have been widely used in cancer diagnosis. This paper uses gene expression data to identify liver cancer biomarkers to aid in early detection and disease prognosis monitoring.

Introduction

Hepatocellular carcinoma (HCC) is the fifth leading cancer-related death cause globally, and its prevalence is increasing. Despite the fact that there are several therapeutic alternatives. HCC diagnosis and management necessitate a comprehensive approach including variety of clinical disciplines. High-quality imaging is the cornerstone of diagnosis, with MRI being the preferred test. When an MRI is inconclusive, some patients require a guided biopsy. Treatment choices determined by the stage of the tumor and the degree of underlying synthetic dysfunction, therefore early detection is critical to patient survival, timely diagnosis is essential for effective therapy, and early discovery dramatically raises the five-year survival rate. Normally, cells proliferate and die in an orderly fashion, with each new cell replacing one that has died. However, cells can become aberrant and continue to develop. The fundamental cause of HCC pathogenesis is a multi-step pathway that involves an accumulation of gene changes that lead to different molecular and cellular abnormalities. Cancerous or malignant tumors can spread. They may infiltrate adjacent tissue and damage regular cells. Cancer cells can break free and move to other regions of the body via the bloodstream or lymph arteries. Data from microarray technology combined with bioinformatics methods offer a major molecular technique for a thorough investigation of gene expression dysregulation between cancer and normal patient samples. In recent years, large-scale microarray data have been released in numerous databases, and using these data to conduct an integrated analysis is a useful technique to monitor the patient's cancer growth. The search for possible biomarkers has been conducted in various tumors using projects based on this technique. Data from microarray technology combined with bioinformatics methods offer a major molecular technique for a thorough investigation of gene expression dysregulation between cancer and normal patient samples. In recent years, large-scale microarray data have been released in numerous databases, and using these data to conduct an integrated analysis is a useful technique to monitor the patient's cancer growth. The search for possible biomarkers has been conducted in various tumours Through the examination of gene

expression samples from human cancer and non-cancerous tissue types, machine-learning models were discovered that successfully discriminate malignant tissue from normal tissue as well as various malignant tissue types from one another. The known tissue-specific cancer-related pathways were verified using functional characterization and pathway analysis, and new cancer-related pathways as well as functional groups for each of the tissue-specific anticipated biomarkers were found. The diagnostic abilities of the biomarkers predicted by the techniques in this study (and then evaluated by comparing their sensitivity and specificity to the sensitivity and specificity of known biomarkers for all tissue types) showed notable improvements over current biomarkers. Using Proposed models SVM and Naive Bays we will perform prognosis while the patient is receiving therapy, the selected biomarkers will be checked; if their values fall into the normal range, the treatment is effective. and diagnosis which is every new patient's liver tissue sample will be examined for the identified biomarkers; if they are present, the patient may develop liver cancer.

Methods

System Architecture

Describe the best feature selection method and ml model.

Dataset

NCBI Gene Expression Omnibus[1] is an international repository that contains multiple forms of genomic data collected by researchers. It contains microarray data of different tissue samples, but each data is collected from a different experiment under different experimental conditions, also the data is usually in its raw format and needs further processing to be used by machine learning models. Since each microarray contains a huge number of probes, which may sometimes reach 2000[2]

The selected dataset was provided by “CuMiDa” project[[3] and hosted on Kaggle, it is the processed format of microarrays raw data with succession id = “GSE14520” that was collected from NCBI Gene Expression Omnibus website.

The dataset included 375 samples of healthy and cancerous tissues, and the number of probes is 22279. The columns represent probe ids and each record represents a tissue sample.

Dataset Visualization

A random sample of 5 probe ids were selected to check common statistical values, it can be seen in **figure 1** that all their value falls in the range of 1 to 16 which is the expected range for the normalized reading values.

	203380_x_at	211224_s_at	221092_at	219279_at	205605_at
count	357.000000	357.000000	357.000000	357.000000	357.000000
mean	9.017054	4.695510	3.449816	3.759549	3.582705
std	0.459521	0.967952	0.197730	0.459393	0.209077
min	7.385061	3.037065	3.083345	2.998020	3.148123
25%	8.778733	3.830298	3.327087	3.452017	3.439907
50%	9.120436	4.628020	3.415459	3.666093	3.542227
75%	9.333791	5.378825	3.537291	4.003379	3.709221
max	10.100178	7.050719	4.985782	5.661041	4.739721

Figure 1 - Statistical values of a 5 probe ids random sample

The dataset is balanced and contains almost the same number of samples for every class as shown in **figure 2** of the class count.

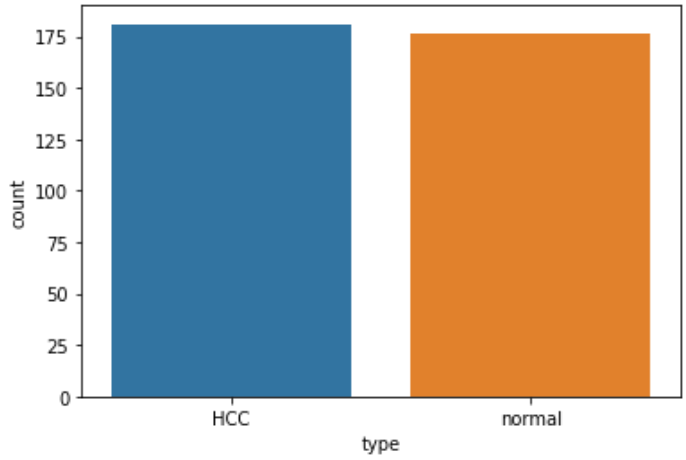

Figure 2 - Dataset class distribution

The dataset dimensions were reduced to 2 dimensions using t-sne algorithm and the transformed data was plotted to study the data distribution in **figure 3**. It is noticed that the two data classes are separated except for a few points.

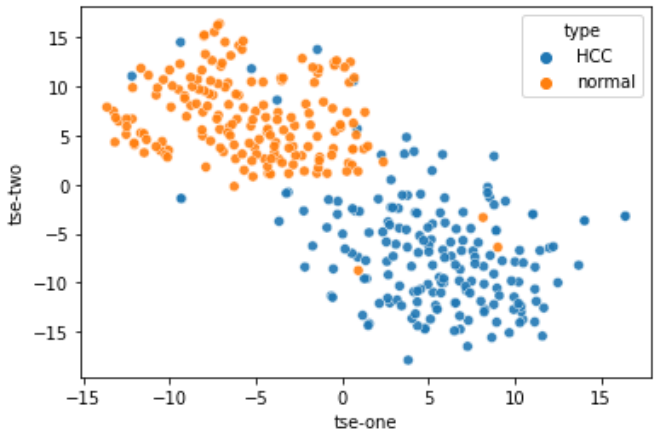

Figure 3 - scatter plot of dataset after applying t-sne

Data preparation

Initially the dataset column names were the probe id of the microarray chip used in analyzing the tissue samples, to identify potential biomarkers these probe ids were mapped to corresponding gene name using data acquired from ensemble website[4] for the microarray chip used in the collection of the raw data. The samples column was dropped since it contained sample id which won't be a useful feature for the machine learning models.

Feature Engineering

Since the number of features in the data set is 20028 features. Feature engineering is a crucial step to build an accurate machine learning model from microarray gene expression data. It is used to select, manipulate, and transform raw data into features that can be used in supervised learning. The best-selected features (Genes) represent the genes that help in the prediction of cancer and normal samples. A comparison between two feature reduction approaches were made to compare the performance of each approach. The first approach was feature extraction which reduces the dimensionality of features and finds a smaller set of new features by combining original features. The second approach is feature selection that selects the most relevant features.

PCA

Principal Component Analysis (PCA) is a feature extraction method that is frequently used to decrease the number of dimensions in large data sets by condensing a large set of variables into a smaller set that retains most of the original set's information. In biological sciences, PCA is frequently used for the study of omics datasets. It can be used to look at similarities across different samples because it provides unsupervised information on the main directions of highest variability in the data. After applying PCA on the training data set the results show that the best number of dimensions are 75 features which are considered potential biomarkers **figure 4**.

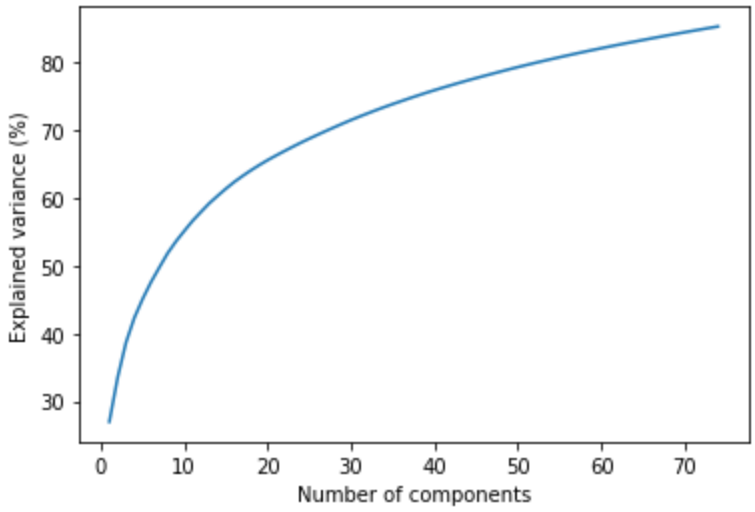

Figure 4- applying PCA on the training data

After using PCA on the training dataset tsne diagram is used to plot the data distribution between three PCs PCA1, PCA2 and PCA3 **figure 5**.

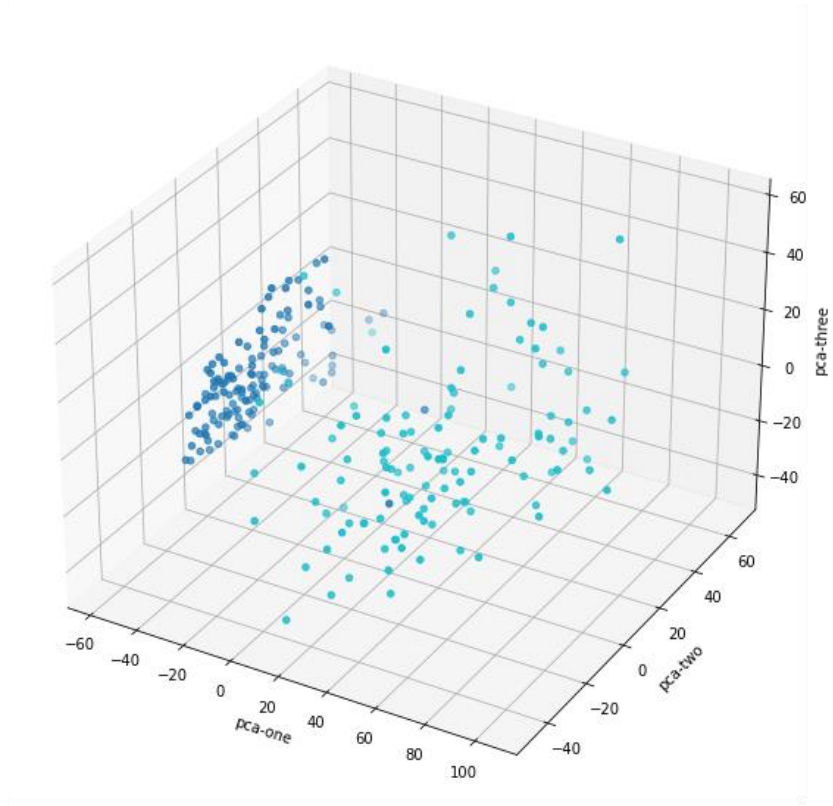

Figure 5- Data distribution using t-sne

Feature selection using filter method

Feature selection is the process of reducing the number of features by selecting a subset of the most relevant features to be used as an input to machine learning model. The filter method is one of feature selection approaches that ranks each feature based on feature subset relevance, the assessment of filter method is by using statistical tests[8]. Since the labeled data is nominal data chi-square is used as the best statistical approach to determine the dependency of two variables. After applying chi-square on gene expression features and SVM model is applied on the training data, the results show that the best number of features is 200 feature which are considered potential biomarkers **figure 6,7** .

	CCL5	ESRRAP1	PXN	SEC11A	TOP2A	NQO1	ACSL3.2	SIGMAR1	EGR1	FAM3C	...	MCTP2	TMOD3	FIP1L1	ALDH6A1.3
0	3.654116	6.720586	5.015457	10.373907	6.487182	3.484757	7.443709	7.513818	4.234161	9.108184	...	3.364998	3.865661	5.785655	8.765856
1	5.137159	5.246931	4.539729	10.863529	5.809140	3.617111	9.126945	6.978191	4.575328	6.651637	...	3.468009	3.465546	5.088006	6.500905
2	4.515175	6.121159	4.862556	11.232235	4.315457	3.696638	7.167784	7.717214	3.935277	6.839798	...	3.658915	3.714477	5.403839	7.550403
3	5.192624	6.275763	4.661036	10.229783	4.940407	4.399711	7.945846	7.484491	5.173549	7.877896	...	3.276052	3.681416	5.159395	8.171625
4	4.961625	6.216846	5.121474	9.978668	5.830239	5.780928	7.744503	7.694174	4.720321	7.544878	...	3.699457	3.679710	5.372327	7.524524
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
352	5.193377	6.183175	4.639223	9.727213	3.591141	3.608490	6.673683	8.079547	7.159559	6.622838	...	3.510749	3.265976	5.230712	10.793508
353	5.704730	6.224405	4.457951	9.842509	3.271005	3.340908	6.868325	8.201286	9.975128	8.337508	...	3.465303	3.452212	5.755963	10.612945
354	4.284763	5.688998	4.666346	10.063517	7.249937	4.928347	8.047757	7.784594	5.122428	9.072875	...	3.346077	3.889923	6.656467	9.252310
355	5.472988	6.136591	4.352139	9.620298	3.840579	3.549621	6.299182	8.046369	8.747951	7.092517	...	3.296420	3.368441	5.786661	10.689106
356	5.598791	5.924060	4.621681	9.654952	3.685207	3.607251	6.633805	7.862495	8.482645	7.526601	...	3.681690	3.386870	5.875606	10.469765

357 rows × 200 columns

Figure 6- Selected features

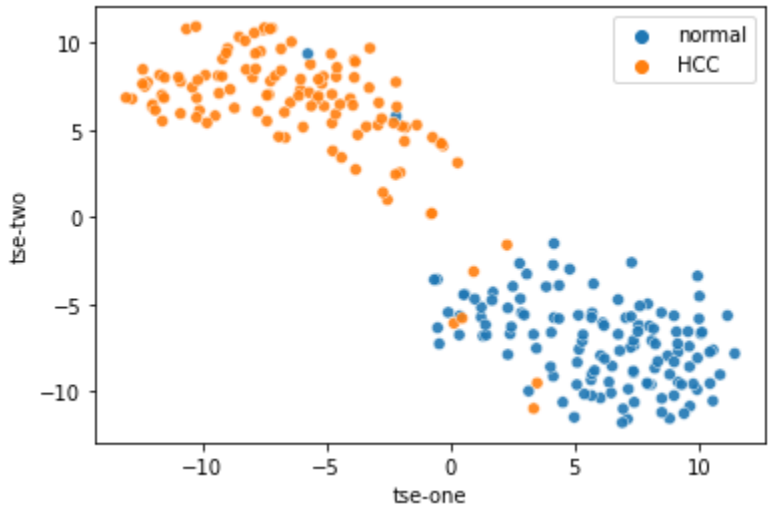

Figure 7- Data Visualization after feature selection

Classification Machine Learning Model

The main problem is to classify tissue samples into normal or cancer samples hence, it is a binary classification problem. For this reason, Support Vector Machines (SVMs) and Naïve Bayes models were used in this classification problem. These machine-learning methods were chosen because of their extensive and successful applications to datasets from genomic and proteomic domains [5].

SVM

Several theoretical reasons explain the superior empirical performance of SVMs in microarray data: e.g., they are robust to the high variable-to-sample ratio and large number of variables, they can learn efficiently complex classification functions, and they employ powerful regularization principles to avoid overfitting [5]. SVM is a classification technique used for both linear & nonlinear data. SVM algorithms create the best hyperplane that separates future data points into classes. A hyperplane which is a wide gap that separates data points belonging to either class. New data points to be predicted are assigned classes based on which side of the hyperplane they fall into [9].

Naïve Bias

The Naïve Bayes classifier (NBC) is one of the most popular classifiers for class prediction or pattern recognition from microarray gene expression data (MGED). The Naive Bayes classifiers (NBCs) are a family of probabilistic classifiers depending on the Bayes' theorem with independence and normality assumptions among the variables. The common rule of NBCs is to pick the hypothesis that is most probable; this is known as the maximum a posteriori (MAP) decision rule [6].

Performance Evaluation

10-fold cross-validation is employed for model selection and accuracy estimation. This method divides the data into 10 parts, uses nine parts to create the models, and uses the remaining one component to create and assess the predictions. Then, each portion (internal test set) is tested against the other nine parts ten times by repeating this method (internal train set). The model's performance is estimated impartially using the average performance over the 10 accuracies [7].

Apply Cross Validation on SVM and Naïve Bayes models using PCA data

Accuracy of SVM model with PCA data: **0.97 figure 8**

Variance: +/- 0.04

#K	1	2	3	4	5	6	7	8	9	10
K score	**1.**	**0.96**	**0.96**	**0.96**	**1.**	**1.**	**0.96**	**0.96**	**0.96**	**0.95**

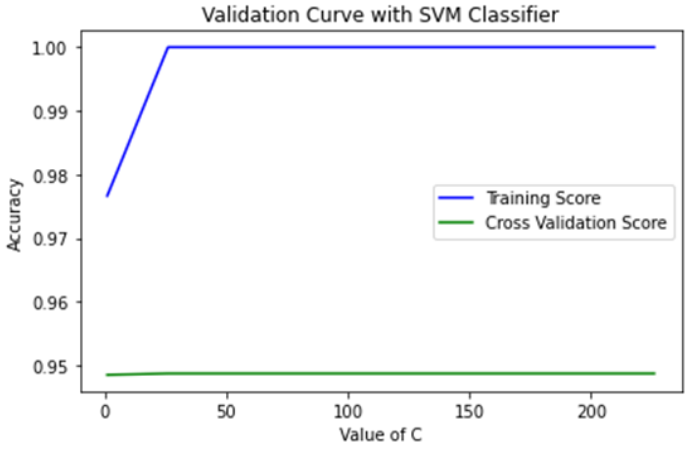

figure 8

Accuracy of SVM model with feature selection filter method data: **0.97 figure 9**

Variance: +/- 0.03

#K	1	2	3	4	5	6	7	8	9	10
K score	**1.**	**0.96**	**0.96**	**0.96**	**1.**	**0.96**	**0.96**	**0.96**	**0.96**	**0.95**

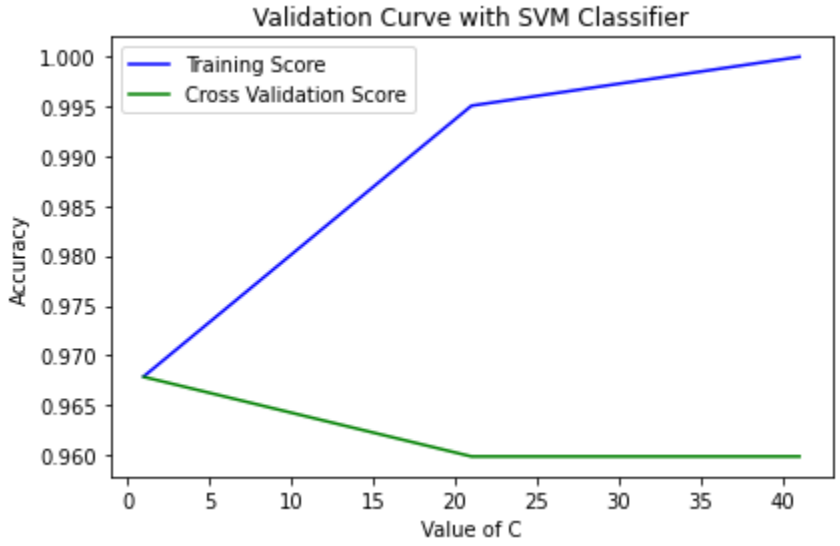

figure 9

Accuracy of Naïve Bayes model with PCA data: **0.91 figure 10.**

Variance: **+/- 0.08**

#K	1	2	3	4	5	6	7	8	9	10
K score	**0.88**	**0.88**	**0.96**	**0.88**	**0.96**	**0.92**	**0.96**	**0.92**	**0.84**	**0.91**

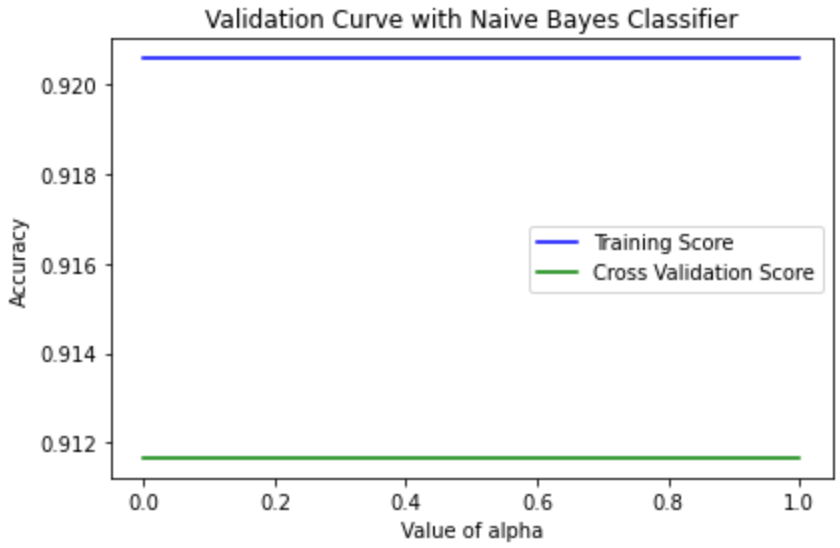

figure 10

Accuracy of Naïve Bayes model with feature selection filter method data data: **0.95 figure 11**

Variance: +/- **0.08**

#K	1	2	3	4	5	6	7	8	9	10
K score	**1.**	**0.88**	**0.88**	**0.96**	**1.**	**0.96**	**0.96**	**0.96**	**0.96**	**0.916**

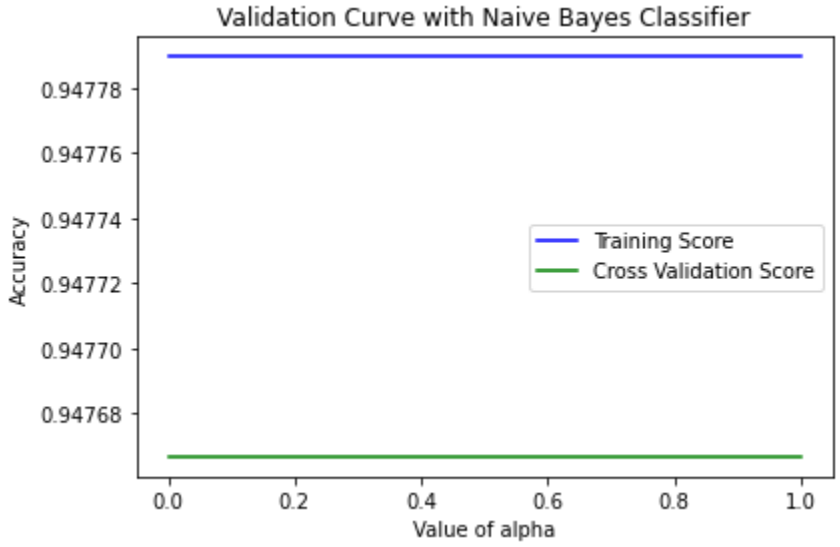

figure 11

Results and Discussion

After evaluating the four models we noticed that Naïve Bayes models are sensitive to outliers and had high variance which caused overfitting.

Accuracy: 0.91 (+/- 0.08) K-fold cross validation score: for each [0.88 0.88 0.96 0.88 0.96 0.92 0.96 0.92 0.84 0.91666667]

By comparison, we chose the SVM model using PCA data as a champion model and applying it to test data provided the following results **figure 12**.

	precision	recall	f1-score	support
0	0.96	0.96	0.96	47
1	0.97	0.97	0.97	61
accuracy			0.96	108
macro avg	0.96	0.96	0.96	108
weighted avg	0.96	0.96	0.96	108

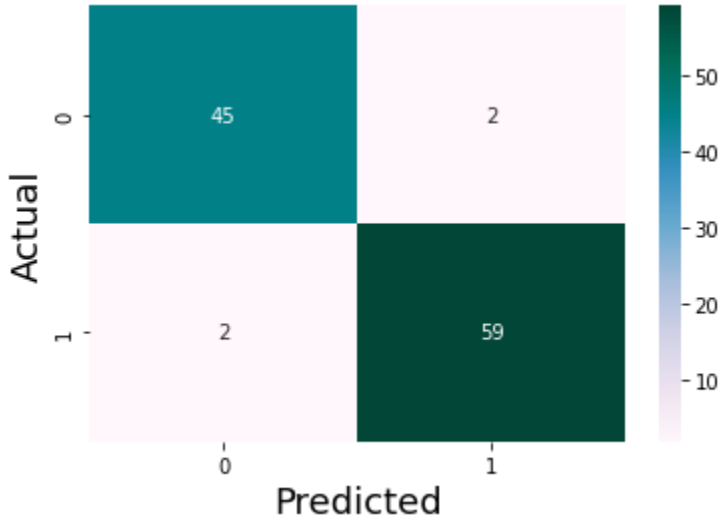

Figure 12- Confusion matrix of SVM model using PCA

Summary and Conclusion

The search for genes, proteins, and other molecules that can provide information about cancer is done through biomarker testing. A distinct pattern of biomarkers characterizes each person's cancer. Some biomarkers influence the efficacy of specific cancer therapies. Machine Learning uses biomarkers to detect early diseases, so after applying our algorithms to data with ID “GSE14520” from the “CuMiDa” project, we choose the highest algorithm accuracy that can early detect liver cancer. From SVM and Naïve Bias algorithms with PCA and Feature Selection using Chi-square, we choose SVM with PCA of accuracy 96%

References

Provide citation information for all the previous publications referred to in your paper. Cite only those references that directly support your work.

[1] T. Barrett *et al.*, “NCBI GEO: archive for functional genomics data sets—update,” *Nucleic Acids Research*, vol. 41, no. D1, pp. D991–D995, Jan. 2013, doi: 10.1093/NAR/GKS1193.

[2] Z. He, L. Wu, M. W. Fields, and J. Zhou, “Use of Microarrays with Different Probe Sizes for Monitoring Gene Expression,” *Applied and Environmental Microbiology*, vol. 71, no. 9, p. 5154, Sep. 2005, doi: 10.1128/AEM.71.9.5154-5162.2005.

[3] B. C. Feltes, E. B. Chandelier, B. I. Grisci, and M. Dorn, “CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research,” *https://home.liebertpub.com/cmb*, vol. 26, no. 4, pp. 376–386, Apr. 2019, doi: 10.1089/CMB.2018.0238.

[4] F. Cunningham *et al.*, “Ensembl 2022,” *Nucleic Acids Research*, vol. 50, no. D1, pp. D988–D995, Jan. 2022, doi: 10.1093/NAR/GKAB1049.

[5] A. Statnikov, L. Wang, and C. F. Aliferis, “A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification,” *BMC Bioinformatics*, vol. 9, 2008, doi: 10.1186/1471-2105-9-319.

[6] M. S. Ahmed, M. Shahjaman, M. M. Rana, and M. N. H. Mollah, “Robustification of Naïve Bayes Classifier and Its Application for Microarray Gene Expression Data Analysis,” *BioMed Research International*, vol. 2017, 2017, doi: 10.1155/2017/3020627.

[7] A. Mohammed, G. Biegert, J. Adamec, and T. Helikar, “Identification of potential tissue-specific cancer biomarkers and development of cancer versus normal genomic classifiers,” 2017. [Online]. Available: www.impactjournals.com/oncotarget/

[8] DTI5126: Fundamentals for Applied Data Science/ Feature Engineering / Bisi Runsewe

[9] DTI5126: Fundamentals for Applied Data Science/ Supervised MLTechniques, Evaluation & Tuning/ Bisi Runsewe