Identification of Potential Gene Expression in Hepatitis B Virus-induced Liver Cancer Using the Simulated Annealing Algorithm

Lailil Muflikhah (lailil@ub.ac.id)
Brawijaya University
Widodo
Brawijaya University
Wayan Firdaus Mahmudy
Brawijaya University
Solimun
Brawijaya University

Research Article

Keywords: potential genes, hepatitis B infection, Liver cancer, simulated annealing algorithm

Posted Date: February 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-155316/v1

Identification of Potential Gene Expression in Hepatitis B Virus-induced Liver Cancer Using the Simulated Annealing Algorithm

Lailil Muflikhah[1,2]*, Widodo[2], Wayan Firdaus Mahmudy[1], Solimun[2]

[1]Faculty of Computer Science, Brawijaya University, Malang, Indonesia [2]Faculty of Mathematics and Natural Science, Brawijaya University, Malang, Indonesia *Corresponding author's email: lailil@ub.ac.id

Abstract

Background

Liver cancer is often associated with hepatitis B infection, and there is a progression from chronic hepatitis to hepatocellular carcinoma. The replication of the virus affects the cell cycle of the host. Many oncogenes that express themselves in the formation of proteins can be highly expressed. A microarray technique can be used as a high throughput measure to determine gene expression in the mechanism of hepatoma progression. However, it has lacks important information on the potential genes to trigger the disease. The information is about the produced mRNA and has a difference to the normal cell line as control.

Results

This study aimed to identify the potential genes that could be used as predictors to detect liver cancer using a heuristic algorithm, simulated annealing optimization. The basic idea of this algorithm was to overcome the combinatorial problem using probability values to select the significant features as a predictor for the classifier model in the representative machine learning algorithms, such as SVM, KNN, Naïve Bayes. C5.0 Decision Tree, and Random Forest. The experimental results showed quite high performance, more than 90% on average. However, using the simulated annealing algorithm requires substantial computation time to identify genes that could be used for detecting liver cancer.

Conclusion

A large amount of gene expression data can be reduced by identifying the potential gene as a predictor in the liver cancer mechanism. Using simulated annealing optimization, the experiment result obtained almost all the selected gene belongs to the type of protein-coding in liver tumor progression. Using cross-validation, the result achieved high performance even though the required time for selection is higher than classification using the machine learning method.

Background

Liver cancer can be caused by hepatitis B infection, progressing from the chronic, cirrhosis stage to hepatocellular carcinoma over a relatively long time. However, most patients do not realize that the disease has reached a severe stage. The hepatitis B virus can induce cancer by inserting its genetic load into the host body, affecting the performance of the cell cycle. Many genes or proteins are involved in liver cancer, and they disrupt the regulation of gene expression. For example, in patients with liver cancer, a regulator of the cell cycle (Cdk4, Cyclin I, Cyclin I, D3, CIP2) increases to cause uncontrolled division, without first repairing. Likewise, the tumor suppressor protein disrupts the regulation of gene expression and induces the formation of malignant tumor cells (Feng *et al.,* 2010; Lee *et al.,* 2000; Lim *et al.,* 2010; Wu *et al.,* 2001).

The use of the gene expression microarray is a substantial data analysis challenge of the biological computation approach to detect the disease. The large number of genes involved in screening requires substantial computational power to construct a classifier model for identification. The feature selection method preprocesses data to obtain significant features that are used as predictors to simplify the model. Many related studies on liver cancer prediction have been conducted using gene expression data and feature selection methods. Feature selection using the Markov clustering method was applied to the support vector machine (SVM) classifier algorithm for identifying HCC module biomarkers from gene expression GSE20948 and achieved an AUC rate of 0.875 (Shen and Liu, 2017). Additionally, dynamic Bayesian network feature selection was applied to the SVM classifier algorithm to diagnose HCC using a dataset with geo access number GSE17856, achieving an accuracy rate of 100% (Akutekwe *et al.,* 2014). Another approach used for feature selection is Hybrid Forward Selection based on the Least Absolute Shrinkage and Selection Operator (LASSO) technique, and this has been applied to the SVM algorithm for HCC disease classification with an accuracy rate of 98.2% (Abinash and Vasudevan, 2019). Recently, Zhang et al. (2020) examined gene expression microarray data, including GSE54236, GSE6404, and GSE121248, for early diagnosis of HCC using an SVM classifier that combined the Maximum Redundancy and Minimum Relevance (mRMR) feature selection method, achieving an accuracy rate of 100%, sensitivity rate of 100%, and specificity rate of 100% (Zhang *et al.,* 2020). However, other research on the classification of liver cancer using a combination of the selected feature using the constrict feature combinations and classification model (CFC-CM) achieved an accuracy rate of 88.73% ± 2.06% (Lin *et al.,* 2019).

Therefore, this research aimed to optimize the combination of features in predictor selection in liver cancer mechanisms using the simulated annealing feature selection algorithm. By using RNA quantification of gene expression, the significant genes impacting the status of liver disease were obtained using a machine learning

algorithm. Lastly, this paper is organized as follows. In Section 2, the materials and methods are presented. Section 3 provides the experimental results and performance evaluation. Section 4 provides the conclusions, and Section 5 provides recommendations for further studies. This study provides a new approach that can be used to identify genes and to indicate liver cancer in patients.

Method

Generally, this research process consisted of four sub-procedures, as shown in Figure 1. First, data acquisition of gene expression from NCBI was addressed to obtain the total RNA. All RNA data were used as inputs into the simulated annealing algorithm, and the particular genes were identified by optimizing the combination of selected features. The importance of features was ranked using a system of descendants. A high importance value was used to construct the classifier model of a machine learning algorithm. It showed the number of features reduction significantly for representative gene expression in liver cancer detection.

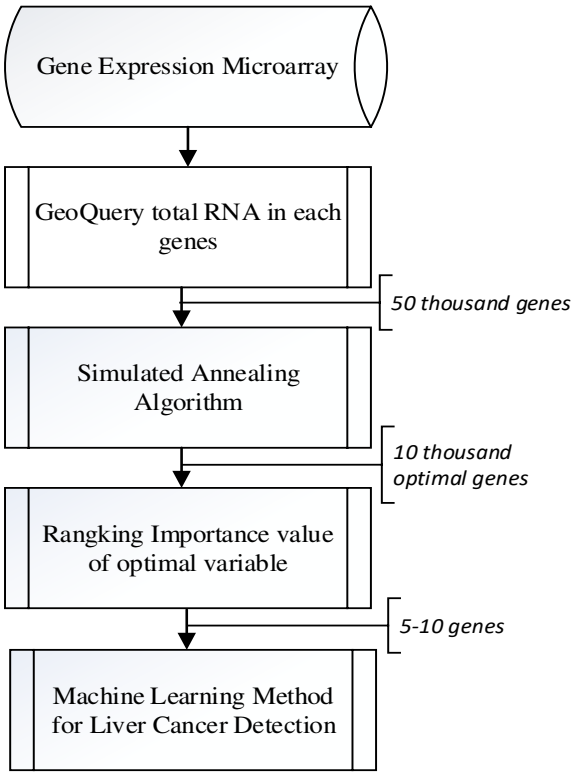

Figure 1. General stages of potential gene identification in the progression of liver cancer.

The dataset that was used in the liver tissue platform was GPL570, with samples in series GSE55092 and GSE121248. In the first data set, GSE55092, the samples were taken from 11 patients at various distances from the center of the tumor. Whole Liver Tissue (WLT) was compared to a Laser Capture-misdirected (LCM) sample of malignant and non-malignant hepatocytes in the same liver. In the sample, profiling gene expression was performed on 17 WLT specimens at any distance from the tumor center from 11 patients of HCC and the selected

LCM sample. Another data set, GSE121248, consisted of profiling gene expression in an array. It was taken from liver tissue of either hepatitis B chronic induced HCC or healthy, adjacent liver tissue using Affymetrix gene construction.

Simulated Annealing Feature Selection

Simulated annealing (SA) is a statistical mechanics method with a random search technique that analogizes the annealing process of metal, is like cooling and freezing that carried out into a strong crystal structure with minimum energy. The SA method is used as the basis for general optimization techniques for combinatorial problems to obtain global optimum. It was developed by S. Kirkpatrick in 1983 deals with very nonlinear problems and was implemented in the case of Traveling Salesman Problem (TSP) and computer hardware design. Two problems were required a complex space as much as NP-complete, to obtain the optimum solution. The stages of the Simulated Annealing algorithm are as shown in the pseudocode of Figure 2.

Input: Training data set; Output: Features combination (S_c)	
1	$S_c \leftarrow Null$
2	$Temp = 100000$
3	$r \leftarrow 0.9$
4	Generate an initial solution, S_0
5	$S_c \leftarrow S_o$
6	Calculate the cost of initial solution, Cost (S_c)
7	**while** *(Temp> 0.001)* **do**
8	**Begin**
9	Find the current solution randomly, S_n, of S_c whose a few different from V_b
10	if *(Cost(S_c) == Cost (Sn))*
11	$S_c \leftarrow S_n$
12	**Else**
13	Generate uniformly random number q in the range (0, 1)
14	If ($q =< e^{-\frac{\text{Cost}(S_n) - \text{Cost}(S_C)}{T}}$)
15	$S_c \leftarrow S_n$
16	$Temp \leftarrow r \; x \; Temp$
17	**end** // for while loop

Figure 2. Pseudocodes of Simulated Annealing.

The initial step of SA has randomly selected an initial solution in the cost function. It is also assumed to be the optimal solution. Then, the next solution is chosen as the current optimal solution, and the cost value is calculated while the temperature T does not meet the desired criteria. However, if the current optimal solution is greater than the new solution, then there is no chance of the solution. Unless, it is generated a random value q in the range of

(0, 1). In this case, the replacement of the optimal solution was only permitted if q value is less than $e^{-\frac{\text{Cost}(V_n)-\text{Cost}(V_b)}{T}}$. Furthermore, the temperature T has been decreased to obtain the termination conditions.

K-Nearest Neighbor Classifier

The K-Nearest Neighbor (KNN) classifier is a supervised learning algorithm for classification that uses the concept of similarity or difference for the data points that are divided into some classes to predict the label or class of the new data object. This algorithm is known as a lazy classification method due to there being no construction to construct a classifier model from training data. The algorithm decides the class of new data points from the training that is similar enough. The class of data objects is then chosen in *k* closet data and is taken in the majority vote of class from training data (Sutton, 2012.).

Support Vector Machine Classifier

The SVM classifier is a supervised learning method for classification to seek the optimal hyperplane function. Initially, the function splits the input space into two classes. Then it develops non-linear classifiers by involving kernel tricks within the high volume of dimension. Next, the data are transformed into a high dimension of vector space. Several kernel functions are applied to multiple classes with non-linear classifiers using Polynomial function, Radial Bases Gaussian Function (RBF), and Sigmoid Function. Each class is labelled y-i e {−1, +1} for i = 1, 2, n, where n is total data. The label is divided from the hyperplane (support vector), as defined in Eq. (1):

$$w.x + b = 0 \qquad (1)$$

The point data *x*i is set to −1, as shown in Eq. (2):

$$w.x_i + b \leq -1 \qquad (2)$$

The point data *x*i is set to +1, as shown in Eq. (3):

$$w.x_i + b \geq +1 \qquad (3)$$

The maximum margin is considered the maximum distance of the hyperplane to close the data object.

$$\frac{1}{\|w\|} \qquad (4)$$

The basic concept of the SVM classifier is to transform the data x to the high dimensional vector space function < D (x). Thus, the new vector space data is represented in the objective function. The training data is a learning process to seek support vectors as hyperplane by the dot product among the new vector space data using a kernel function as is defined in Eq. (5):

$$K(x_i.x_j) = \Phi(x_i).\Phi(x_j) \qquad (5)$$

The Radial Basis Function (RBF) is a kernel trick used in this study, as shown in Eq. (6):

$$K(X_i \cdot X_j) = \exp\left(-\left(\frac{\|X_i - X_j\|^2}{2\sigma^2}\right)\right) \quad (6)$$

Then, the next step is applying the sequential SVM algorithm to construct predictions, such as the Hessian matrix, iteration to obtain the greatest of the least error rate, or Max (| 5a |) <e. Then, the test data and training data are calculated to obtain the bias and similarities between each of these datasets. As a result, positive or negative classes are obtained (Vijayakumar and Wu, 1999).

Naive Bayes Classifier

Naive Bayes Classifier is a classification method based on the Bayesian probability theorem. The primary character is a very strong (naive) assumption of independence from each event. The model is easy to create using Eq. (7) (Data Mining, 2012)

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} \quad (7)$$

where *X* is attributed, and *C* is class.

The Naive Bayes classifier is sought to maximize the probability value of each class, and it is expressed as the Hypothesis Maximum Posteriori, as seen in Eq. (8):

$$H_{MAP} = \arg\max P(C|x_1, x_2, \ldots x_n) = \arg\max \mathrm{P(C)} \prod \prod_{i=1}^{n} P(x_i|C) \quad (8)$$

C5.0 Decision Tree Classifier

A decision tree classifier is a method of building a model of the structured tree as a recursive division. This method is known as 'divide and conquers' because it uses the feature values to divide the data into smaller subsets of similar classes. The data are divided into branches that designate the selected decision. The tree is completed by leaf nodes as the termination of the decision. It is used to define the result following a combination of decisions. The C5.0 decision tree classifier model is an extended C4.5 decision tree algorithm, as proposed by Quinlan in 1993 (Pandya *et al.*, 2015). The C5.0 trees are simpler than the C4.5. Boosting, winnowing, and asymmetric costs for specific errors are applied to this model (Ojha *et al.,* 2017).

Random Forest

Random Forest (RF) is an aggregation of tree predictors with a uniform distribution in one forest. This tree is built as Classification and Regression Trees (CART) decision tree without pruning. The RF is a supervised learning method to be addressed for evaluating the performance of feature selection and reduction in liver cancer detection. The RF method is built from several decision trees by selecting the number of F features randomly. This means

that not all features are used to build the tree. The value of F affects the performance of the RF. If the F value is too small ($F << M$ attribute), the correlation value from the tree will tend to be a small value (small correlation). Otherwise, if the F value is too large ($F >> M$ attribute), the correlation value from the tree will tend to be a great value (strong correlation) (Breiman, 2001). Furthermore, the F value and the number of all attributes, M, can be determined by equation ()

$$F = \log_2(M + 1) \qquad (9)$$

In the RF method, the number of training data is selected using a bootstrapping technique. The bootstrap aggregating method in sampling is implemented in building each decision tree with previously selected candidate attributes. Broadly, the stages of the Random Forest algorithm are as follows:

- The samples are taken bootstrap from training data
- Building the decision tree
- Feature selection m < M in splitting nodes
- Each bootstrap sample creates a decision tree
- The class is taken in the majority vote

K-fold cross-validation

K-fold cross-validation is a method used to evaluate the performance of experimental results. The entire dataset was divided into *k* parts. Then, the parts were iterated for *k* iterations in a different fold. The total number of folds (k) was divided into two parts, namely, training, and testing data. The testing data were used as m fold, and the training data were used as k-m fold. Each fold was filled with class +1 and class −1 data at a proportion of 50%: 50% (Refailzadeh, 2008).

Figure 3. An illustration of *K-fold cross-validation*.

The K-fold cross-validation is illustrated in Figure 3. The first data collection was used as test data, and the rest of the data were used as training data in the first iteration. Then, the second data collection was used as test data, and the rest of the data were used as training data. Substitution of test data and training data was conducted as the number of iterations. Each iteration obtained an accuracy rate value within ten iterations, and then the average value of accuracy was calculated. This study used 10-fold cross-validation to evaluate the performance of the representative machine learning implementation upon the selected gene expression.

Results and Discussion

The platform of liver tissue, GPL570, had an unbalanced data distribution in class as shown in Table 1.

Table 1. Details of the datasets used in the study.

Data sets	Total data	The number of genes	The number of data in each class
GSE55092 (GPL 570)	140	54676	Whole Liver Tissue (120), malignant hepatocytes (10), non-malignant hepatocytes (10)
GSE121248 (GPL 570)	107	54676	Adjacent Normal sample (37), Tumor sample (70)

The box plot in Figures 4 and 5 show the distribution of numerical data of gene expression in mRNA to be performed in the data quartiles (or percentiles) and the averages. The expression values of GSE55092 and GSE121248 were in the range of (2-12). Additionally, the scatter plots of Figure 4 and Figure 5 show the relationship between the average log-expression value and the expression log-ratio for GSM139321 and GSM3428716 as described them.

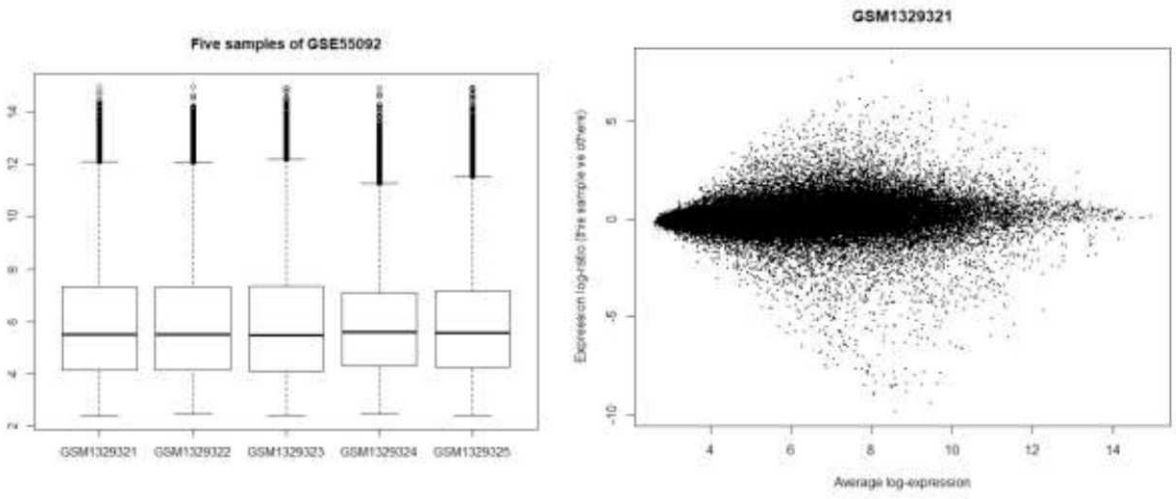

Figure 4. Boxplot in-range value of a series GSE55092 and scatter plot of the sample GSM1329321 in log-ratio of expression value.

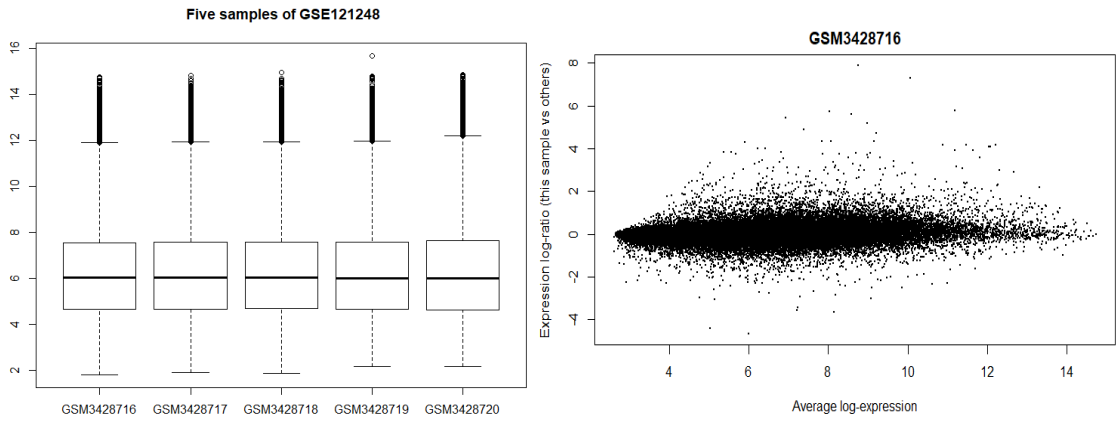

Figure 5. Boxplot in-range value of a series GSE121248 and scatter plot of the sample GSM3428716 in log-ratio of expression value

As an illustration, two datasets using the same platform GPL570 have referential integrity and can be plotted in entity-relationship (ER). An entity, in this context, is an object, component of data, or collection of similar entities. These entities can have attributes that define their properties. The entity-relationship diagram in Figure 6 shows the relationships between the datasets of gene expression information (GSE55092 and GSE121248) that were stored in the database. By defining the entities, their attributes, and showing the relationships between them, the ER diagram illustrates the logical structure of the databases. There are some related entities, including the entities of the sample GSM, serial expression GSE, and platform of gene GPL.

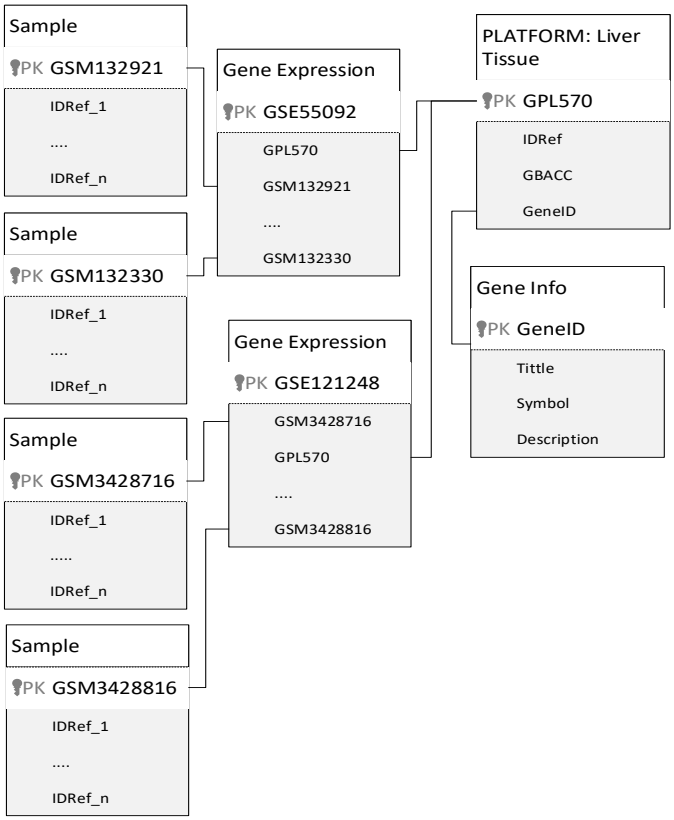

Figure 6. The entity-relationship diagram of the GSE55092 and GSE121248 gene expression datasets.

After applying the Simulated Annealing optimization method and ranking based on the variable importance values, then ten features were selected of GSE55092, including SpotID: “216765_at,” “236985_at,” “FFX.PheX.5_at” “238611_at,” “205386_s_at,” “216604_s_at,” “209740_s_at,” “209711_at,” “227073_at,” and “202596_at.” However, some SpotID were not identified in the gene database, and all types of protein were coded, as shown in Table 2. The selected genes are related to either direct or indirect liver cancer progression. An expression of the gene *ENSA* can reduce tumor propagation of liver cancer. The overexpressed *ENSA* is correlated to the growth of suppressor tumors concern. (Chen et al., 2013). *MDM2* is also an important role in tumor progression (*MDM2–P53 Pathway in Hepatocellular Carcinoma | Cancer Research*, 2014). Another selected gene, *MAP3K2,* is identified to directly lead to the cycle cell of HCC progression (Shi et al., 2020).

Table 2. The feature selection result of gene expression GSE55092 using the simulated annealing algorithm

GenelD	**Gene Symbol**	**Description**	**Type of Gene**
2029	*ENSA*	endosulfine alpha	Protein –coding
4193	*MDM2*	MDM2 proto-oncogene	Protein –coding
5607	*MAP2K5*	mitogen-activated protein kinase kinase 5	Protein –coding
8228	*PNPLA4*	patatin like phospholipase domain containing 4	Protein -coding
10746	*MAP3K2*	mitogen-activated protein kinase kinase kinase 2	Protein -coding
23169	*SLC35D1*	solute carrier family 35 member D1	Protein -coding
23428	*SLC7A8*	solute carrier family 7 member 8	Protein -coding

Then, by the same method, the other selected features of GSE121248 are including 209365_s_at, 203358_s_at, 200882_s_at, 1568638_a_at, 227654_at, 222077_s_at, 202243_s_at, 211609_x_at, 223516_s_at, 200783_s_at. They have the related information of genes as shown in Table 3. Almost all the selected genes belong to the type of liver cancer progression. *Indoleamine 2,3-dioxygenase (IDO)* increased the regulation in hepatocellular carcinoma and effects impede locally immune and metastasis promotion (Shibata et al., 2016). The upregulated gene *Stathmin 1 (STMN1)* facilitated and triggered the liver cancer pathway (R. Zhang et al., 2020). Also upregulation of *Proteasome 26S subunit, non-ATPase 4* can trigger liver cancer (Cai et al., 2019). MicroRNA-137 has a suppressive role in liver cancer via targeting EZH2(Shi et al., 2020). Extracellular matrix protein 1, a novel prognostic factor, is associated with the metastatic potential of hepatocellular carcinoma(H. Chen et al., 2011). Upregulation of Rac GTPase-activating protein 1 is significantly associated with the early recurrence of human hepatocellular carcinoma(Wang et al., 2011).

Table 3. The Feature selection result of gene expression GSE121248 using simulated annealing algorithm

GeneID	Gene Symbol	Description	Type of Gene
169355	*IDO2*	indoleamine 2,3-dioxygenase 2	Protein -coding
3925	*STMN1*	stathmin 1	Protein -coding
5710	*PSMD4*	proteasome 26S subunit, non-ATPase 4	Protein -coding
5692	*PSMB4*	proteasome 20S subunit beta 4	Protein -coding
2146	*EZH2*	enhancer of zeste 2 polycomb repressive complex 2 subunit	Protein -coding
1893	*ECM1*	extracellular matrix protein 1	Protein -coding
29127	*RACGAP1*	Rac GTPase activating protein 1	Protein -coding
29964	*PRICKLE4*	prickle planar cell polarity protein 4	Protein -coding
100188893	*TOMM6*	translocase of outer mitochondrial membrane 6	Protein -coding
140876	*FAM65C*	RIPOR family member 3	Protein -coding

Performance Measure

When the selected feature results were applied to the machine learning algorithms, and their performance was evaluated using the confusion matrix to get information for accuracy performance measures for sensitivity, specificity, and Area Under the Curve were obtained as presented in Table 4 (Baratloo et al., 2015)(Jiao & Du, 2016).

Table 4. The performance measure metrics.

Measure	Formula	Definition
Accuracy	$\frac{tp+tn}{tp+fn+fp+tn}$	The correctness of a classifier
Sensitivity	$\frac{tp}{tp+fn}$	The Effectiveness for identifying the positive liver cancer classification
Specificity	$\frac{tn}{tn+fp}$	Effectiveness for identifying negative liver cancer (normal) classification
AUC	$\frac{1}{2}\left(\frac{tp}{tp+fn}+\frac{tn}{tn+fp}\right)$	The ability of the classifier to avoid misclassification

Remark:

- True positive *(tp):* cases predicted as liver cancer, and they are liver cancer.

- True negative *(tn):* cases are predicted as healthy, and they are healthy.

- False-positive *(fp):* cases are predicted as liver cancer, but they are healthy.

- False-negative *(fn):* cases are predicted as healthy, but they are liver cancer.

3.3. Comparison of machine learning algorithms

Table 5 and Table 6 show the performance results of when we compared using feature selection using the simulated annealing algorithm and not using feature selection to find the potential gene expression to construct a

classifier model of machine learning algorithms including SVM, Naive Bayes (NB), Random Forest (RF), K-Nearest Neighbor (KNN), and Decision Tree C5.0.

Table 5. Comparison of the performance of GSE55092 using classifier methods.

Performance measurement		***Algorithm***				
		SVM	***NB***	***RF***	***KNN***	***C5.0***
Accuracy	SAFS	1	1	1	0.99	0.99
	no-SAFS	1	0.95	1	0.91	0.99
Sensitivity	SAFS	1	1	1	0.99	0.99
	No-SAFS	1	0.97	1	0.97	0.97
Specificity	SAFS	1	1	1	0.99	0.99
	No-SAFS	1	0.94	1	0.87	1
AUC	SAFS	1	1	1	0.99	0.99
	No-SAFS	1	0.96	1	0.94	0.99

Table 6. Comparison of Result Performance GSE121248 Using Classifier Methods

Performance measurement		***Algorithm***				
		SVM	***NB***	***RF***	***KNN***	***C5.0***
Accuracy	SAFS	1	0.95	1	0.95	0.97
	no-SAFS	1	0.95	1	0.91	0.99
Sensitivity	SAFS	1	0.97	1	0.97	0.97
	No-SAFS	1	0.97	1	0.97	0.97
Specificity	SAFS	1	0.95	1	0.95	0.97
	No-SAFS	1	0.94	1	0.87	1
AUC	SAFS	1	0.96	1	0.96	0.97
	No-SAFS	1	0.96	1	0.94	0.98

We found that feature selection using the simulated annealing algorithm required a computation time for GSE55092 of 6.582 minutes. Also, the required computation time of GSE121248 is 4.65851 minutes. These findings show the drawback of the simulated annealing algorithm is the substantial computation time required to solve the combinatorial problem (Busetti, 2003). The comparison of computation time required for representative machine learning is shown in Table 7.

Table 7. Comparison of computation time using machine learning classifier (in the second)

Algorithm	GSE55092		GSE121248	
	non-SAFS	SAFS	non-SAFS	SAFS
SVM	16.13	0.17	12.16	0.051
KNN	20.77	0.34	11.85	0.038
C5.0	104.36	0.10	80.59	0.096
RF	62.97	0.38	55.53	0.043
NB	225.99	0.05	262.48	0.107

Furthermore, the total computation time required for detecting liver cancer using representative machine learning is as shown in Figure 7. The required time for identifying potential gene expression using simulated annealing feature selection is much higher than classifier modeling in machine learning methods.

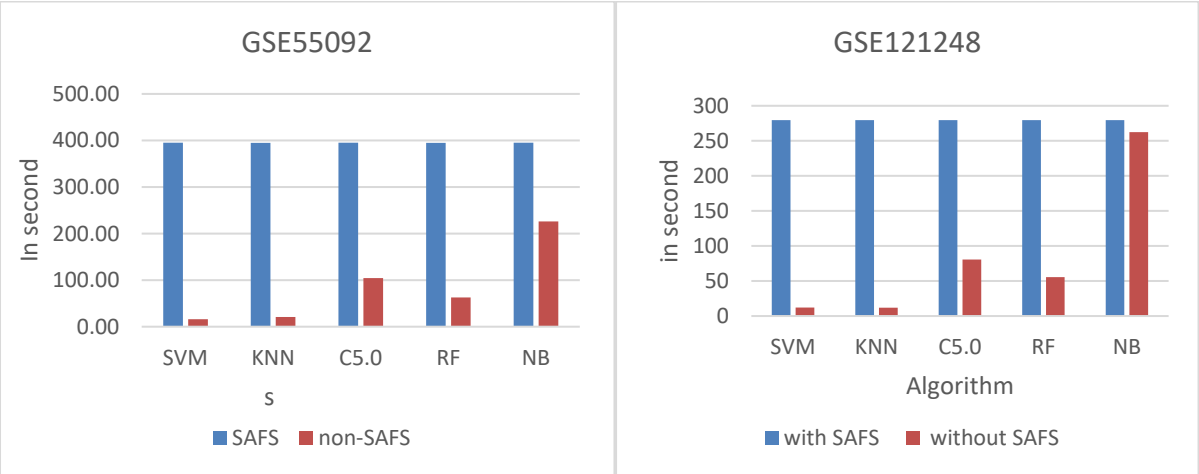

Figure 7. Computation Time Comparison of Liver Cancer Detection for GSE55092 and GSE121248

Conclusions

Simulated annealing feature selection was implemented to identify potential genes that could be used to detect liver cancer mechanisms. All kinds of selected genes are protein-coding and almost all are related to liver cancer progression. The genes were constructed to simplify the classifier model using a machine learning algorithm and achieved high-performance measures of 90% above, including accuracy, sensitivity, and specificity. The performance rates increase in the range of 1.4% - 3.4% on average. However, the computation time required was high in the feature selection method. Therefore, in future work, it is necessary to modify the simulated annealing algorithm to reduce the computation time.

Acknowledgment

We are thanking you to World Class University program that facilitate proofreading by Enago.

Authors' Contributions

LM wrote this article and implementation of the program, W created the concept and validated the dataset, WFM designed algorithm, S simulated and evaluated testing data. All authors reviewed, and approved the final manuscript.

Funding

This research is financially supported by Kementerian Riset, Teknologi dan Pendidikan Tinggi,through a program of doctoral dissertation grant under contract number: 293.14/UN10.C10/PN/2020

Availability of data and material

The data set used and analyzed of this study is taken from gene bank, National Center for Biotechnology Information (NCBI), and is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55092 and https://www.ncbi.nlm.nih.gov/geo/query/ acc.cgi?acc=GSE121248

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

'Faculty of Computer Science, Brawijaya University, Malang, Indonesia, [2]Faculty of Mathematics and Natural Science, Brawijaya University, Malang, Indonesia

References

Abinash, M.J., Vasudevan, V., 2019. A Hybrid Forward Selection Based LASSO Technique for Liver Cancer Classification, in: Nath, V., Mandal, J.K. (Eds.), Nanoelectronics, Circuits and Communication Systems, Lecture Notes in Electrical Engineering. Springer, Singapore, pp. 185-193. https://doi.org/10.1007/978-981-13-0776-8_17

Akutekwe, A., Seker, H., Iliya, S., 2014. An optimized hybrid dynamic Bayesian network approach using differential evolution algorithm for the diagnosis of Hepatocellular Carcinoma, in: 2014 IEEE 6th International Conference on Adaptive Science Technology (ICAST). Presented at the 2014 IEEE 6th International Conference on Adaptive Science Technology (ICAST), pp. 1-6. https://doi.org/10.1109/ICASTECH.2014.7068140

Baratloo, A., Hosseini, M., Negida, A., & El Ashal, G. (2015). Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. Emergency, 3(2), 48–49.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5-32. https://doi.org/10.1023/A:1010933404324

Cai, M.-J., Cui, Y., Fang, M., Wang, Q., Zhang, A.-J., Kuai, J.-H., Pang, F., & Cui, X.-D. (2019). Inhibition of PSMD4 blocks the tumorigenesis of hepatocellular carcinoma. Gene, 702, 66–74. https://doi.org/10.1016/j.gene.2019.03.063

Chen, H., Jia, W.-D., Li, J.-S., Wang, W., Xu, G.-L., Ma, J.-L., Ren, W.-H., Ge, Y.-S., Yu, J.-H., Liu, W.-B., Zhang, C.-H., & Wang, Y.-C. (2011). Extracellular matrix protein 1, a novel prognostic factor, is associated with metastatic potential of hepatocellular carcinoma. Medical Oncology (Northwood, London, England), 28 Suppl 1, S318-325. https://doi.org/10.1007/s12032-010-9763-1

Chen, Y.-L., Kuo, M.-H., Lin, P.-Y., Chuang, W.-L., Hsu, C.-C., Chu, P.-Y., Lee, C.-H., Huang, T. H.-M., Leu, Y.-W., & Hsiao, S.-H. (2013). ENSA expression correlates with attenuated tumor propagation in liver cancer. Biochemical and Biophysical Research Communications, 442(1–2), 56–61. https://doi.org/10.1016/j.bbrc.2013.10.165

Data Mining, 2012.. Elsevier. https://doi.org/10.1016/C2009-0-61819-5

Evaluating a Classification Model [WWW Document], n.d.. ritchieng.github.io. URL http://www.ritchieng.com/machine-learning-evaluate-classification-model/ (accessed 11.1.19). Feng, H., Li, X., Niu, D., Chen, W.N., 2010. Protein profile in HBx transfected cells: a comparative iTRAQ-coupled 2D LC-MS/MS analysis. J Proteomics 73, 1421-1432. https://doi.org/10.1016/jjprot.2009.12.004

Huynh, H., Nguyen, T. T. T., Chow, K.-H. K.-P., Tan, P. H., Soo, K. C., & Tran, E. (2003). Over-expression of the mitogen-activated protein kinase (MAPK) kinase (MEK)-MAPK in hepatocellular carcinoma: Its role in tumor progression and apoptosis. BMC Gastroenterology, 3(1), 19. https://doi.org/10.1186/1471-230X-3-19

Jiao, Y., & Du, P. (2016). Performance measures in evaluating machine learning-based bioinformatics predictors for classifications. Quantitative Biology, 4(4), 320–330. https://doi.org/10.1007/s40484-016-0081-2

Lee, S.W., Lee, Y.M., Bae, S.K., Murakami, S., Yun, Y., Kim, K.W., 2000. Human hepatitis B virus X protein is a possible mediator of hypoxia-induced angiogenesis in hepatocarcinogenesis. Biochem. Biophys. Res. Commun. 268, 456-461. https://doi.org/10.1006/bbrc.2000.2093

Lim, W., Kwon, S.H., Cho, Hyeseon, Kim, S., Lee, S., Ryu, W.S., Cho, H., 2010. HBx targeting to mitochondria and ROS generation are necessary but insufficient for HBV-induced cyclooxygenase-2 expression. J. Mol. Med. 88, 359369. https://doi.org/10.1007/s00109-009-0563-z

Lin, X., Huang, X., Zhou, L., Ren, W., Zeng, J., Yao, W., Wang, X., 2019. The Robust Classification Model Based on Combinatorial Features. IEEE/ACM Transactions on Computational Biology and Bioinformatics 16, 650-657. https://doi.org/10.1109/TCBB.2017.2779512

Nguyen, T.T., Huang, J.Z., Wu, Q., Nguyen, T.T., Li, M.J., 2015. Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. BMC Genomics 16, S5. https://doi.org/10.1186/1471-2164-16-S2-S5

Ojha, U., Jain, M., Jain, G., Tiwari, R.K., 2017. Significance of important attributes for decision making using C5.0, in: 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). Presented at the 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1 -4. https://doi.org/10.1109/ICCCNT.2017.8204031

Pandya, R., Pandya, J., Infotech, K.P.D., n.d. C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning.

Shen, C., Liu, Z., 2017. Identifying module biomarkers of hepatocellular carcinoma from gene expression data, in: 2017 Chinese Automation Congress (CAC). Presented at the 2017 Chinese Automation Congress (CAC), pp. 5404-5407. https://doi.org/10.1109/CAC.2017.8243741

Shi, X., Liu, T.-T., Yu, X.-N., Balakrishnan, A., Zhu, H.-R., Guo, H.-Y., Zhang, G.-C., Bilegsaikhan, E., Sun, J.-L., Song, G.-Q., Weng, S.-Q., Dong, L., Ott, M., Zhu, J.-M., & Shen, X.-Z. (2020). MicroRNA-93-5p promotes hepatocellular carcinoma progression via a microRNA-93-5p/MAP3K2/c-Jun positive feedback circuit. Oncogene, 39(35), 5768–5781. https://doi.org/10.1038/s41388-020-01401-0

Shibata, Y., Hara, T., Nagano, J., Nakamura, N., Ohno, T., Ninomiya, S., Ito, H., Tanaka, T., Saito, K., Seishima, M., Shimizu, M., Moriwaki, H., & Tsurumi, H. (2016). The Role of Indoleamine 2,3-Dioxygenase in Diethylnitrosamine-Induced Liver Carcinogenesis. PLOS ONE, 11(1), e0146279. https://doi.org/10.1371/journal.pone.0146279

Simulated annealing - an overview | ScienceDirect Topics [WWW Document], n.d. URL https://www.sciencedirect.com/topics/computer-science/simulated-annealing (accessed 10.22.20).

Sutton, O., 2012. Introduction to k Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction 10.

Vijayakumar, S., Wu, S., 1999. Sequential Support Vector Classifiers and Regression, in: IIA/SOCO.

Wang, S. M., Ooi, L. L. P. J., & Hui, K. M. (2011). Upregulation of Rac GTPase-activating protein 1 is significantly associated with the early recurrence of human hepatocellular carcinoma. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 17(18), 6040–6051. https://doi.org/10.1158/1078-0432.CCR-11-0557

Wu, C.G., Salvay, D.M., Forgues, M., Valerie, K., Farnsworth, J., Markin, R.S., Wang, X.W., 2001. Distinctive gene expression profiles associated with Hepatitis B virus x protein. Oncogene 20, 3674-3682. https://doi.org/10.1038/sj.onc.1204481

Xie, J., Zhu, X. Y., Liu, L. M., & Meng, Z. Q. (2018). Solute carrier transporters: Potential targets for digestive system neoplasms. Cancer Management and Research, 10, 153–166. https://doi.org/10.2147/CMAR.S152951

Zhang, R., Gao, X., Zuo, J., Hu, B., Yang, J., Zhao, J., & Chen, J. (2020). STMN1 upregulation mediates hepatocellular carcinoma and hepatic stellate cell crosstalk to aggravate cancer by triggering the MET pathway. Cancer Science, 111(2), 406–417. https://doi.org/10.1111/cas.14262

Zhang, Z.M., Tan, J.X., Wang, F., Dao, F.Y., Zhang, Z.Y., Lin, H., 2020. Early Diagnosis of Hepatocellular Carcinoma Using Machine Learning Method. Front. Bioeng. Biotechnol. 8 https://doi.org/10.3389/fbioe.2020.00254

Figures

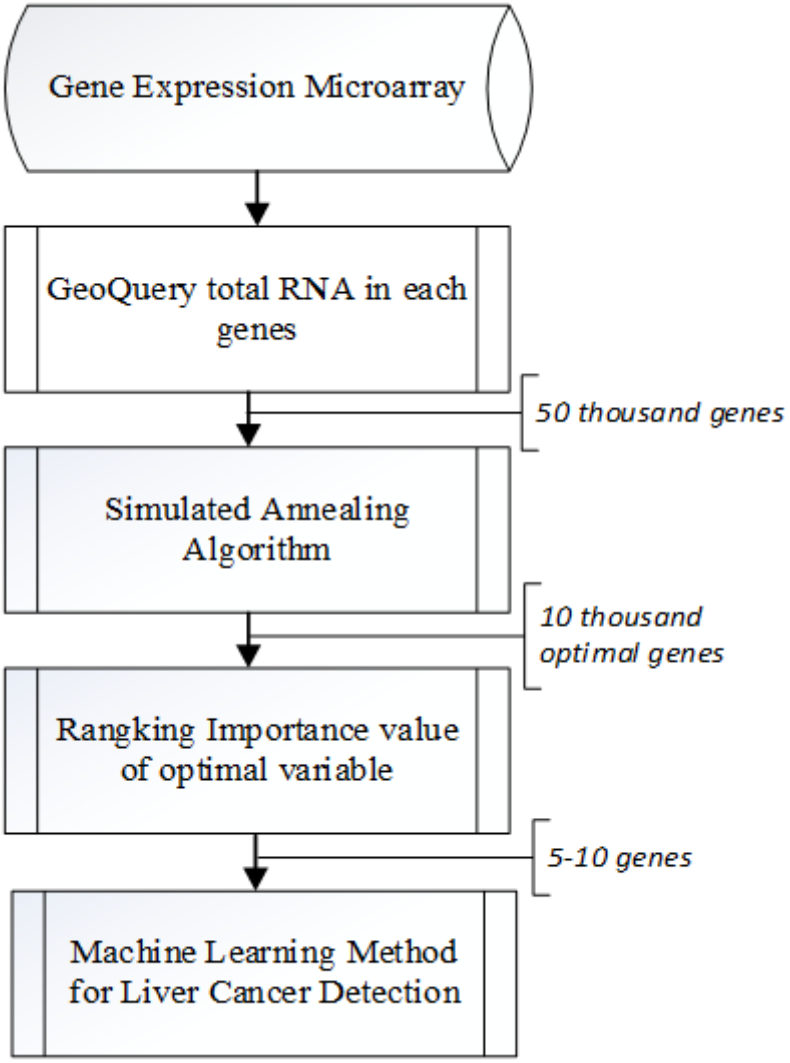

Figure 1

General stages of potential gene identification in the progression of liver cancer.

Input: Training data set; Output: Features combination (S_c)	
1	$S_c \leftarrow Null$
2	$Temp = 100000$
3	$r \leftarrow 0.9$
4	Generate an initial solution, S_0
5	$S_c \leftarrow S_o$
6	Calculate the cost of initial solution, Cost (S_c)
7	**while** $(Temp > 0.001)$ **do**
8	**Begin**
9	Find the current solution randomly, S_n of S_c whose a few different from V_b
10	if $(Cost(S_c) == Cost(Sn))$
11	$S_c \leftarrow S_n$
12	**Else**
13	Generate uniformly random number q in the range (0, 1)
14	If $(q =< e^{-\frac{Cost(S_n) - Cost(S_c)}{T}})$
15	$S_c \leftarrow S_n$
16	$Temp \leftarrow r \times Temp$
17	**end** // for while loop

Figure 2

Pseudocodes of Simulated Annealing.

Figure 3

An illustration of K-fold cross-validation.

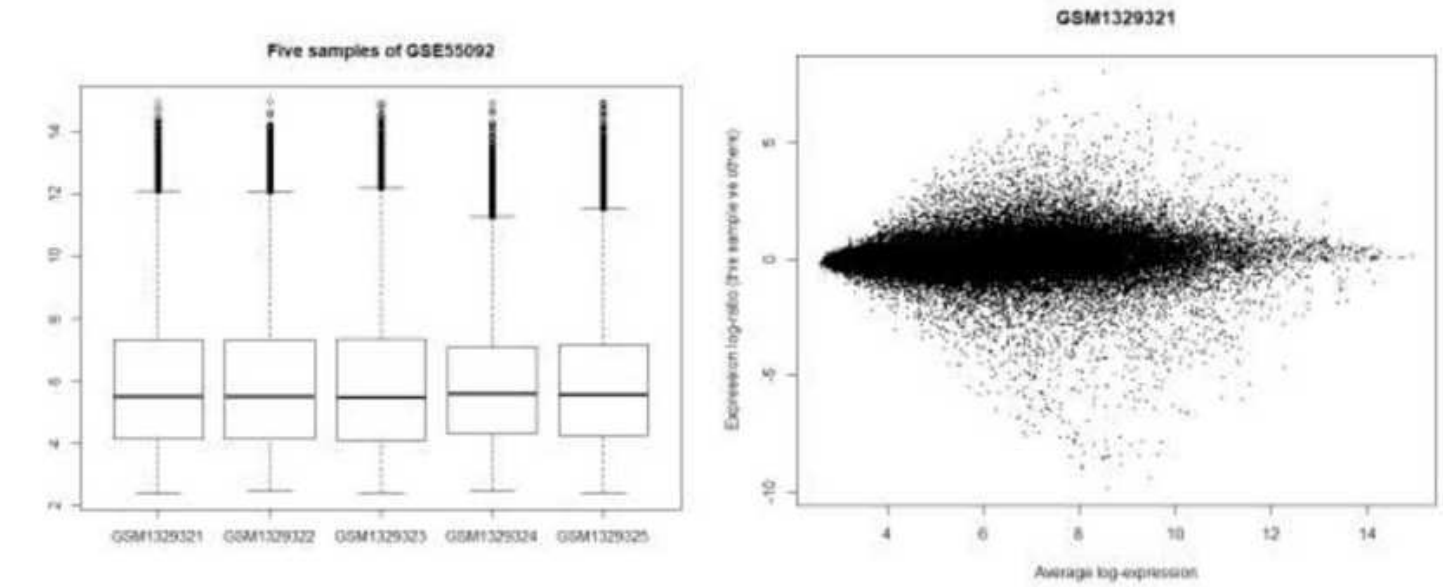

Figure 4

Boxplot in-range value of a series GSE55092 and scatter plot of the sample GSM1329321 in log-ratio of expression value.

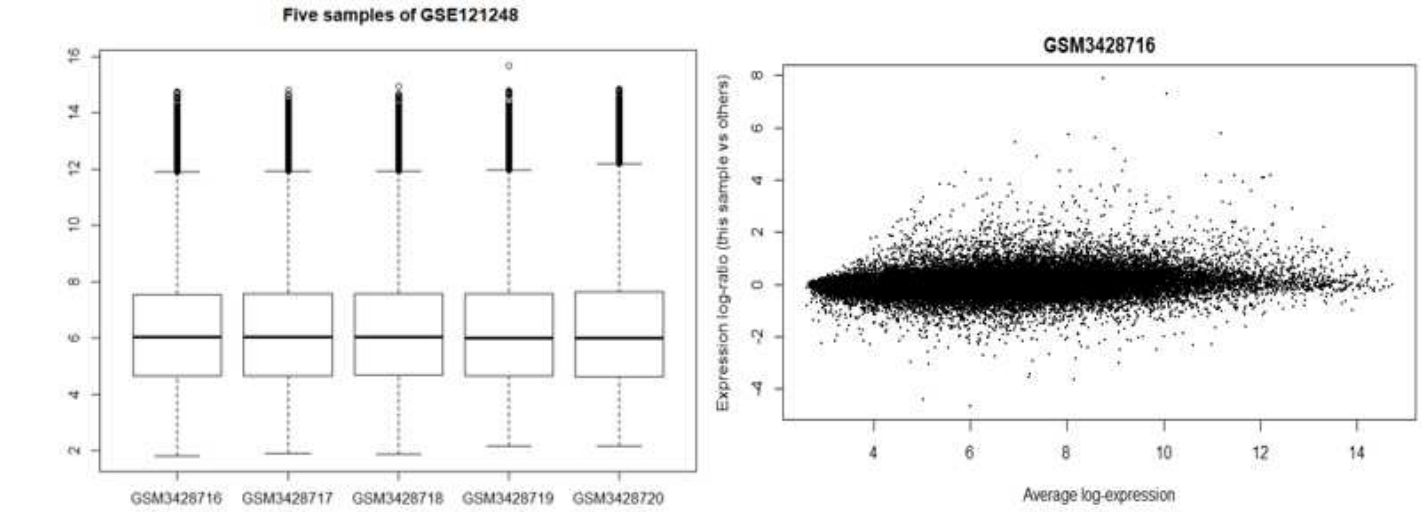

Figure 5

Boxplot in-range value of a series GSE121248 and scatter plot of the sample GSM3428716 in log-ratio of expression value

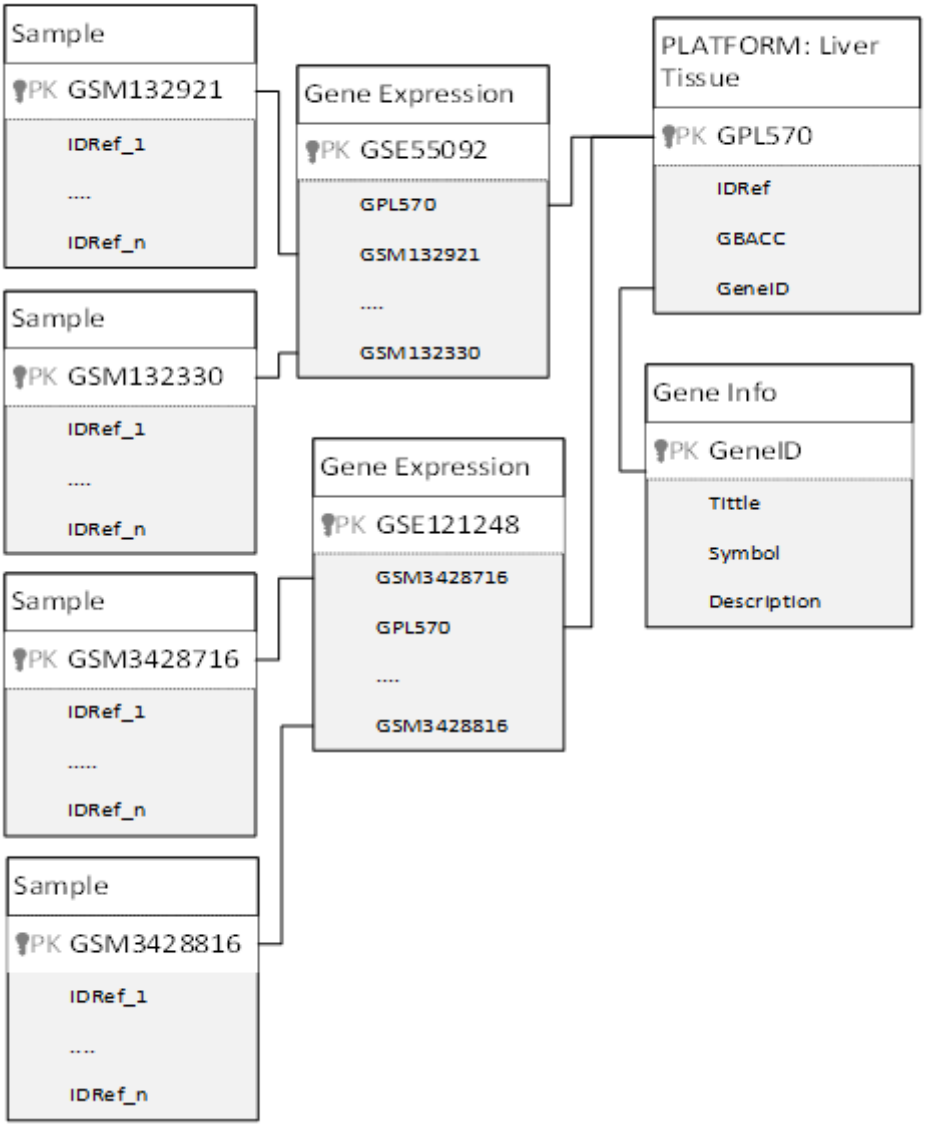

Figure 6

The entity-relationship diagram of the GSE55092 and GSE121248 gene expression datasets.

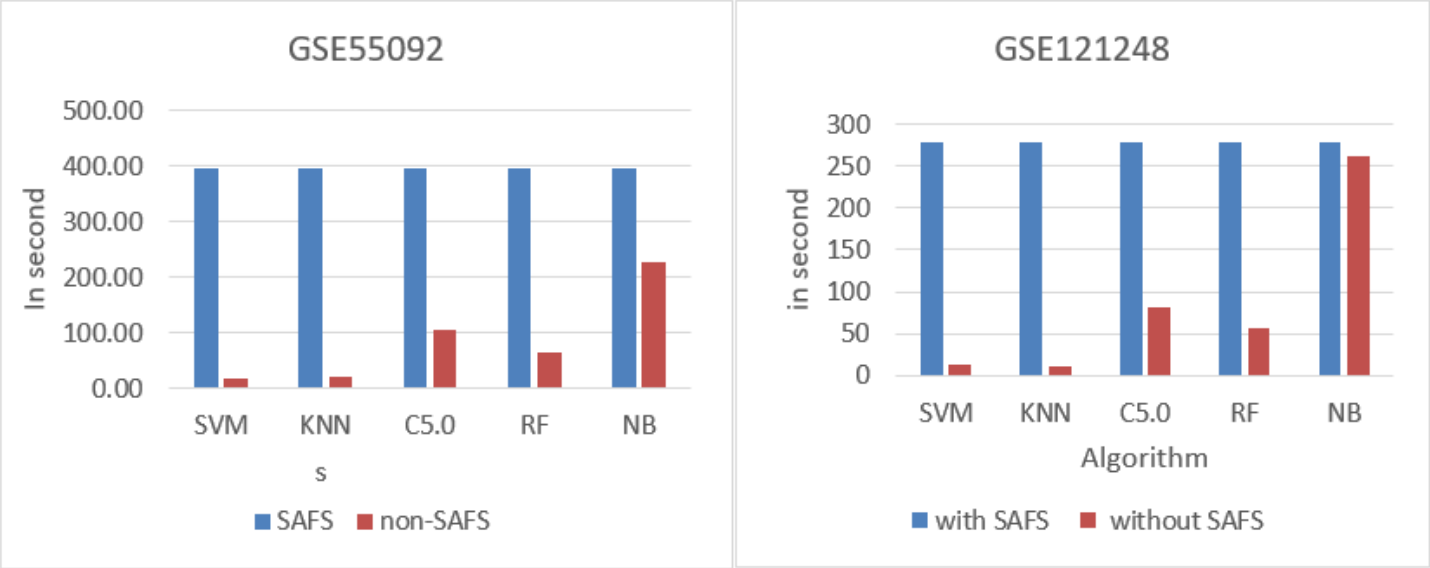

Figure 7

Computation Time Comparison of Liver Cancer Detection for GSE55092 and GSE121248