Detection of Hepatoma based on Gene Expression using Unitary Matrix of Singular Vector Decomposition

Lailil Muflikhah[1], Nashi Widodo[2], Wayan Firdaus Mahmudy[3], Solimun[4], Ninik Nihayatul Wahibah[5]
Faculty of Computer Science[1,3]
Faculty of Mathematics and Natural Sciences[2,4,5]
Brawijaya University, Malang, Indonesia[1,2,3,4]
University of Riau, Pekanbaru, Riau, Indonesia[5]

***Abstract*—Hepatoma is a long-term disease with a high risk of mortality. However, the disease is late detected, at the fourth level stadium due to silent symptoms. The infected hepatitis B virus gene *HBx* is a genome virus to trigger liver disease. This virus inserts material genetic into the host and disturbs the cell cycle. The regulation of gene expression is blocked to make work abnormal, especially for repairing and degrading. A microarray is a tool to quantify the RNA gene expression in huge volumes without any information for the related potential gene. Therefore, this study is proposed a feature extraction method using a unitary singular matrix for simplifying the classification model of hepatoma detection. Principally, the feature is decomposed using a singular vector to get the *k*-rank value of pattern. This matrix is applied to the representative machine learning algorithm, including KNN, Naïve Bayes, C5.0 Decision Tree, and SVM. The experimental result achieved high performance with Area under the Curve (AUC) of above 90% on average.**

Keywords*—Hepatoma; gene expression; feature extraction; unitary matrix

I. Introduction

Hepatoma is a liver disorder disease and progresses from chronic, acute, cirrhosis to over a long period of about 30-40 years to liver cancer. This type of disease has a high number compared to cancer as a cause of death. Many studies carried out an early detection because of silent symptoms. The disease is detected after the third to fourth stadium stages. In the biology field, the relationship between mutations or genetic variations of the hepatitis B virus and liver cancer is open research and is still being debated.

Another hand, the infected virus affects abnormal gene expression regulation. An infected virus is a cause of Hepatoma disease and blocks the expression for reparation or destruction in the cell cycle. However, the new genes are produced uncontrol-up to trigger the oncogene in high volume. A microarray is a tool to investigate and quantify gene expression. The large number of genes involved in screening requires further analysis to get any information inside the genes.

The large volume of gene expression effected to high dimensional data. It required high space memory and high-speed in computation time to construct a classifier model for hepatoma detection. Therefore, many studies applied feature reduction methods to construct modeling data for the disease detection. The methods are including clustering, hybrid SFS and LASSO, Random Forest, and Dynamic Bayesian Network or using bioinformatics tools with a statistical approach to obtain information on significant differences in gene expression and then used as features to implement in classifier model of machine learning algorithms for detection [1]–[8].

Basically, there are two methods for reducing data volume including feature extraction and feature selection to simplify the machine learning model. Many studies on hepatoma detection are based on gene expression using feature selection for dimension reduction. However, the feature selection affects loss information. Therefore, this research aims to reduce the dimensional data using feature extraction method through the singular vector decomposition to get a unitary matrix. The matrix contains eigen value and indicates important information of pattern of data collection with certain *k*-rank value.

This paper consists of five sections. The first section is provided the background of this study. Then, the previous related research and research gap are described in Section 2. The proposed method and basic concept are performed in Section 3. The result and discussion are provided in Section 4. Last, the conclusion and future work are prepared in Section 5.

II. Related Work

Research on hepatoma disease detection using gene expression is related to the high dimensional data. Many studies applied to reduce the dimensional data to simplify learning model for detection. The maximum redundancy minimum relevance (mRMR) is a method to identify the significant gene as biomarker in hepatoma mechanism [6]. Then. Markov clustering method was applied to identify liver cancer module biomarkers from gene expression GSE20948 and achieved the AUC rate of 0.875 [3]. Also, a dynamic Bayesian network feature selection was applied to SVM classifier for the diagnosis of liver cancer using data set under geo access number GSE17856 and achieved high accuracy [2]. Another research for feature selection method was Hybrid Forward Selection of the LASSO technique was applied to the SVM algorithm for liver cancer disease classification. It achieved an accuracy rate of 98.2% [1]. Then, Zhang et.al (2020) researched gene expression microarray data including GSE54236, GSE6404, GSE121248 for early diagnosis of liver cancer using an SVM classifier that combined Maximum Redundancy and Minimum Relevance (mRMR) feature

selection method, and the results achieved a high performance [6]. Recently, the reduced data is also applied by removing unrelated features and identifying the significant gene expression using machine learning method and statistical approach [7]–[9]. However, all those studies aimed to select the feature for dimension reduction. The omitted data is an effect to integrated information. Therefore, this research is proposed feature extraction by transforming the gene expression values in unitary matrix using the singular vector decomposition approach. The matrix contains important information and delineates the data collection.

III. Material and Research Method

A. Data Set

The data set of human gene expression in this study is taken from the data bank National Center for Biotechnology Information (NCBI) at URL: https://www.ncbi.nlm.nih.gov/geo/. They used a blood cell platform (GPL15491) and a liver tissue platform (GPL570) with GEO access numbers GSE114783, GSE55092, and GSE121248. Refer to the biological network approach, the data GSE114783 is addressed to investigate potential mechanisms and biological markers of every stage from HBV infection to hepatoma. Global gene profiling methods of healthy individuals (HC), HBV carrier (HBVC), chronic hepatitis B (CHB), liver cirrhosis (LC), and hepatoma (HCC) patients were analyzed by sequencing gene [10].

The different gene expressions were found by corrective RVM (Random variance model) analysis of ANOVA and STC (Series Test of Cluster). Mononuclear blood cells (PBMC) from three healthy people (HC), three HBV carriers (HBVC), three chronic hepatitis B (CHB) patients, three liver cirrhosis (LC), and three hepatomas (HCC) samples as the details shown in Table I, two data sets are under platform liver tissue (GPL 570) including GSE55092 and GSE121248. The sample GSE55092 is identified molecular and genomic of Whole Liver Tissue (WLT) of 17 samples and Laser Capture-misdirected (LCM) of 11 samples. In these samples, the gene was applied to profiling the WLT at any distance from the centroid of the tumor. They were taken from 11 patients' liver cancer using the selected LCM samples [11]. Another sample GSE121248 consists of the profiled gene expression under platform blood cell [12]. The two kinds of these samples are taken from liver tissue, either liver cancer-induced by Hepatitis B chronic or normal tissue in the adjacency of Affymetrix construction.

TABLE I. The Data Set Details

Data sets	Total features	Class (number of data)
GSE114783 (GPL15491)	30142	Liver cirrhosis (10), healthy control (3), chronic hepatitis B(10), hepatitis B virus(3), HCC (10)
GSE55092 (GPL 570)	54.676	Whole Liver Tissue (120), malignant hepatocytes (10), non-malignant hepatocytes (10)
GSE121248 (GPL 570)	54.676	Adjacent Normal sample (37), tumor sample (70)

B. Research Method

Research on Hepatoma detection through machine learning methods using gene expression microarray data is a big data problem. The large size of the gene as a feature affects data modeling in building classification algorithms to make it so complicated and time-consuming. The number of gene expressions (*m)* is much more than the number of data (*n*). In another word, we can notate as (m>>n) as illustrated in the following matrix.

$$\begin{bmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots & Anm \end{bmatrix} \quad (1)$$

Therefore, this research is proposed to reduce dimensional data by feature extraction. The extraction is carried out to find patterns of gene expression differences in the hepatoma mechanism. The related method used in this research is dimension reduction through feature extraction using singular vector decomposition (SVD) and principal component analysis (PCA). Generally, the main stages of the proposed method are described in Fig. 1.

1) Singular Value Decomposition (SVD): A microarray of gene expression is represented by a matrix of associated genes from the host. For example, there are m gene expressions and n hosts (samples), it can be made m x n sample-genes as matrix A for the total RNA formed. SVD is a Latent Semantic Indexing method to find patterns in a matrix and identify gene expressions that are similar to one another. This section describes some of the basic components of the SVD used as a dimensional reduction method. Making a new matrix from matrix A with m gene expression x n hosts which is a matrix of U, $\sum$ and V so that $A = U\sum V^T$ can be illustrated as in Fig. 2. U and V are unitary and orthogonal matrices that have unit columns so that $U^TU=I_m$ [8].

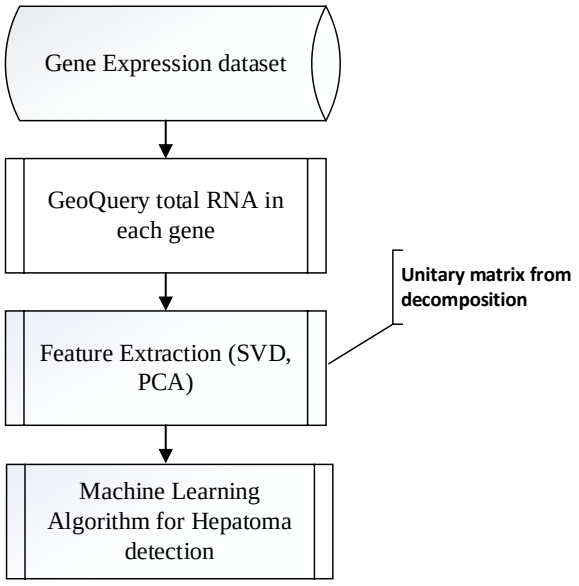

Fig. 1. Flowchart of General System for Hepatoma Detection.

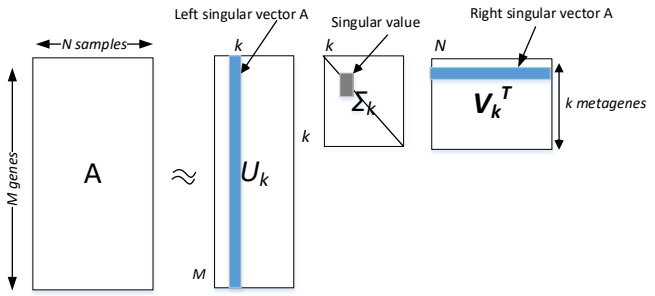

Fig. 2. Reduced Dimension Representation of the Genes-Samples Matrix.

Then, the proposed method of feature extraction using unitary matrix in SVD has several steps are as follows:

a) Compute its transpose A^T and A^TA.

b) Get the eigenvalues of A^TA and arrange these in descending order, in the absolute sense. Then, compute the square roots to obtain the singular values of *A*.

c) Construct a diagonal matrix $\sum$ by placing singular values in descending order along its diagonal. After that, compute its inverse, $\sum^{-1}$.

d) Use the ordered eigenvalues from step b and compute the eigenvectors of A^TA. Put these eigenvectors along with the columns of *V* and count its transpose, V^T.

e) Compute membership matrix *U* as $U=AV\sum^{-1}$.

f) Matrix *U* with *k*-rank as dimension is used as data set to construct classifier model of machine learning algorithm.

2) Principle Component Analysis (PCA): Another way to reduce the matrix dimension is of Latent Semantic Indexing (LSI) is to use principal component analysis (PCA). The main goal of PCA is to acquire a new set of dimensions (features) that better capture data variability. The first dimension is chosen to capture as much variability as possible. The second dimension is orthogonal with the first dimension capturing as much of the remaining variability as possible, and so on. Hence, the strongest pattern is found in the first dimension as [13]. As an illustration, the PCA is implemented to decompose gene expression-samples matrix for reduction as shown in Fig. 3.

Generally, this method obtains the eigenvector and eigenvalues from the covariance of matrix *A*, as stated in detail below:

a) Construct an $N \times d$ document-term matrix *A*, with one-row vector A_n per data point.

b) Then matrix *A* subtract mean is multiplied from each row vector A_n in *A*.

c) Get the covariance matrix *Y* of *A*.

d) Find eigenvector and eigenvalues of *Y*.

e) The principal component is obtained from *M* eigenvectors with the largest eigenvalues.

The PCA is known for applying Singular Vector Decomposition (SVD) on the covariance matrix. Here, the illustration of PCA for document-term matrix *A* is shown in (2).

$$\begin{aligned} A &\rightarrow Y, \text{ where } Y = A_i - \mu_j \\ Y &\rightarrow Y^T \\ 1/(n)\, Y^TY &\rightarrow A \\ A &\rightarrow U\sum V^T \end{aligned} \quad (2)$$

3) Machine learning algorithm: A machine learning algorithm is a generated method from data collection to construct a pattern for prediction or description. The algorithm is a way to make computer programs that increase performance based on its experience [14]. In this research, the method is addressed for the classifier model for hepatoma detection. Some various representative machine learning algorithms are applied including hyperplane function (Support Vector Machine), probability-based (Naïve Bayes), similarity-based (*k*-Nearest Neighbor), entropy-based (C5.0 Decision Tree), an ensemble method for aggregation (Random Forest).

a) Support Vector Machine (SVM): Support Vector Machine (SVM) algorithm is a supervised machine learning method for classification that desires to get the optimal hyperplane function. Initially, the function is to define two classes (binary class) in a linear function. Then, it was developed into non-linear classifiers by involving kernel tricks in the high dimension. The data is transformed into a high dimension of vector space [15].

b) Naïve Bayes Classifier: Another supervised learning method based using the statistical approach is the Naïve Bayes algorithm. This method used probability theory. Naïve Bayes Classifier is a simple classification method based on the Bayesian probability theorem. The main character is a very strong (naïve) assumption of independence from each event. The model is easy to create using this formula as in (2) [16].

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} \quad (3)$$

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

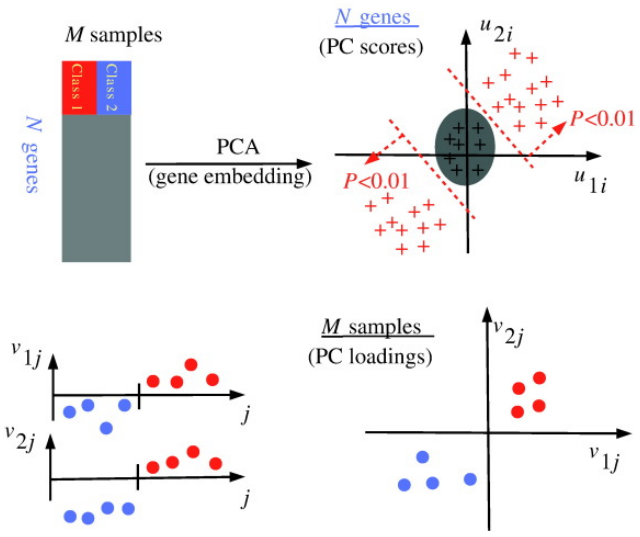

Fig. 3. Principle Component Analysis of Gene Expression Data Set.

c) K-Nearest Neighbor: K-Nearest Neighbor is a supervised learning method based on the distance or similarity of an object's characteristics. The algorithm decides the class of data points from the training data whose similar enough. The class of new data object is chosen in *k* closet data and is taken in the majority vote of class from training data [17].

d) C5.*0 Decision Tree:* C5.0 is a supervised learning algorithm that constructs a tree based on entropy value to build decision rule. This algorithm is an extension of Decision Tree 4.5 with a simpler tree (rule set) that is built so that the steps taken are more concise. This method is better than the previous Decision Tree method, ID3, and C4.5 for pruning and memory allocation (space complexity) [18]. The difference between C5.0 and C4.5 is the boosting and voting processes to determine the class based on the calculation of a combination of several trees.

e) Random Forest: Random forest (RF) is an enhanced method of a decision tree that is built using aggregating several trees. The trees are grown without pruning during training [19]. The algorithm is a decision tree method similar to the Classification and Regression Tree (CART) method with maximum size without pruning. The scheme resembles bagging in the training data set to build a new tree. To predict the new data, it collects the class from several trees.

4) K-Fold Cross-Validation: K-fold cross-validation is a method used to evaluate the performance of experimental results. The whole data are divided by *k* parts, then they are iterated for *k* iterations in the different folds. In this research, the total number of folds (*k=10*) will be divided into two parts, namely training and testing data. In the testing data used as *m* fold, and in the training, data used as *k-m* fold. Each fold will be filled with class +1 and class -1 data within a proportion [20].

IV. Result and Discussion

A. Experimental Result

Several representative machine learning algorithms are used to evaluate the performance of feature extraction. The representative machine learning algorithms are including KNN, Naïve Bayes, SVM, C5.0 Decision Tree, and Random Forest. Three data sets GSE114783, GSE55092, and GSE121248 were applied to decompose the matrix using the proposed method (SVD and PCA) for dimension reduction. Thus, they were applied to the representative machine learning using k-fold cross-validation, and as a result, the performance including accuracy, sensitivity, specificity, and AUC was shown in Table II, Table III, and Table IV.

In Table II, the SVM and RF are stable and achieve the highest performance of 1 (100%). It means that both algorithms do not depend on the data distribution and the number of data sets. However, the worst classification is using the K-NN algorithm with *k*-nearest value =3. The accuracy and sensitivity achieved the lowest of 58% using original data (without dimension reduction). It shows that the performance proposed method is dominant in any representative machine learning algorithms.

In Table III, the performance result including accuracy, sensitivity, and specificity is highest at SVM and RF algorithms. In contrast, the specificity of KNN is the lowest. It means that the ability to classify in a negative class is not good. The reduction using PCA method is slightly better than the others.

TABLE II. Performance Comparison of the Proposed Method using Representative Machine Learning Algorithms on GSE114783

Perform. Measure	Reduct Method	Machine Learning Algorithm				
		SVM	NB	RF	K-NN	C5.0
Accuracy	SVD full	1	0,94	1	0,61	0,89
	SVD k=20	1	0,83	1	0,69	0,86
	PCA	1	0,94	1	0,61	0,89
	no-FS	1	0,72	1	0,58	0,97
Sensitivity	SVD full	1	0,94	1	0,61	0,89
	SVD k=20	1	0,83	1	0,69	0,86
	PCA	1	0,94	1	0,61	0,89
	No-FS	1	0,72	1	0,58	0,97
Specificity	SVD full	1	0,98	1	0,86	0,96
	SVD k=20	1	0,94	1	0,9	0,94
	PCA	1	0,98	1	0,87	0,96
	No-FS	1	0,89	1	0,86	0,99
AUC	SVD full	1	0,95	1	0,73	0,92
	SVD k=20	1	0,89	1	0,8	0,90
	PCA	1	0,96	1	0,74	0,92
	No-FS	1	0,81	1	0,72	0,98

TABLE III. Performance Comparison of the Proposed Method using Representative Machine Learning Algorithms on GSE55092

Performance Measure	Reduction Method	Machine Learning Algorithm				
		SVM	NB	RF	KNN	C5.0
Accuracy	SVD full	1	0,9	1	0,86	0,97
	SVD k=10	1	0,88	1	0,93	0,98
	PCA	1	0,9	1	0,93	0,97
	no-FS	1	0,957	1	0,921	0,993
Sensitivity	SVD full	1	0,897	1	0,857	0,969
	SVD k=10	1	0,879	1	0,929	0,977
	PCA	1	0,897	1	0,929	0,969
	No-FS	1	0,957	1	0,921	0,993
Specificity	SVD full	1	0,992	1	0,143	0,913
	SVD k=10	1	0,779	1	0,571	0,914
	PCA	1	0,992	1	0,614	0,957
	No-FS	1	0,997	1	0,529	0,99
AUC	SVD full	1	0,945	1	0,500	0,941
	SVD k=10	1	0,829	1	0,750	0,945
	PCA	1	0,945	1	0,771	0,962
	No-FS	1	0,977	1	0,725	0,975

Then, in the last experiment of data set GSE121248, two algorithms including SVM and RF are stable and achieve the highest performance of 100% (1). In contrast, the lowest performance is at KNN algorithm in SVD without using k-rank for accuracy and the Area Under the Curve (AUC). However, the proposed method using SVD with *k*-rank=10 is high performance in accuracy, sensitivity, and specificity as shown in Table IV.

Furthermore, the proposed method of dimension reduction using SVD with *k*-rank value has maximum the number of a dataset (SVD full). It means that there are *k* patterns of gene expression data collection. In the data set GSE55902 and GSE121248, the variance pattern is notated as eigenvalue and converges at the first 10 singular values as shown in Fig. 4. However, the eigen value convergence of GSE114783 is in the first 20 values as shown in Fig. 5, The variance pattern values indicate significant feature values.

Another hand, in PCA method, the characteristic of dimension values has convergence starting from 10 variances data as shown in Fig. 6. The larger the value of *k*-rank, the smaller the data variance is.

Then, the computational time required for the hepatoma detection using the machine learning algorithm representation in this study appears so much short, due to a large number of reduced features. The reduction of computation time is very significant in the Naïve Bayes algorithm, then C5.0 Decision Tree and Random Forest. The comparison of computational time on the machine learning algorithm using the proposed method is shown in Fig. 7, Fig. 8, and Fig. 9.

TABLE IV. PERFORMANCE COMPARISON OF THE PROPOSED METHOD USING REPRESENTATIVE MACHINE LEARNING ALGORITHMS ON GSE121248

Performance Measure	Reduction Method	Machine Learning Algorithm				
		SVM	NB	RF	KNN	C5.0
Accuracy	SVD full	1	0,94	1	0,78	0,98
	SVD k=10	1	0,96	1	0,95	0,98
	PCA	1	0,91	1	0,92	0,97
	no-FS	1	0,953	1	0,907	0,991
Sensitivity	SVD full	1	0,95	1	0,97	0,97
	SVD k=10	1	0,97	1	0,97	0,97
	PCA	1	0,84	1	0,97	0,92
	No-FS	1	0,973	1	0,973	0,973
Specificity	SVD full	1	0,94	1	0,89	0,99
	SVD k=10	1	0,96	1	0,94	0,99
	PCA	1	0,94	1	0,89	1
	No-FS	1	0,943	1	0,871	1
AUC	SVD full	1	0,945	1	0,73	0,98
	SVD k=10	1	0,965	1	0,955	0,98
	PCA	1	0,89	1	0,93	0,96
	No-FS	1	0,963	1	0,939	0,982

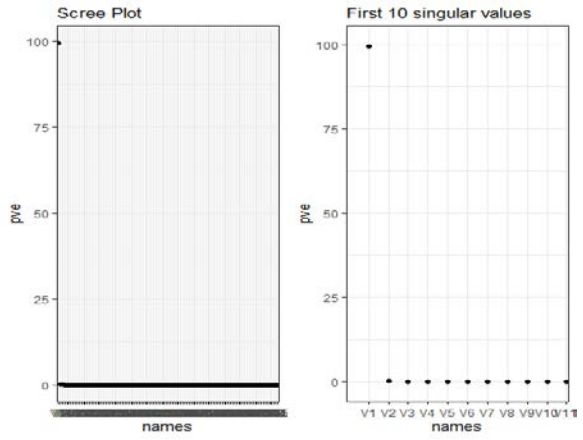

Fig. 4. Eigen Value with *K*-Rank = 10 of SVD Decomposition on GSE550922 and GSE121248.

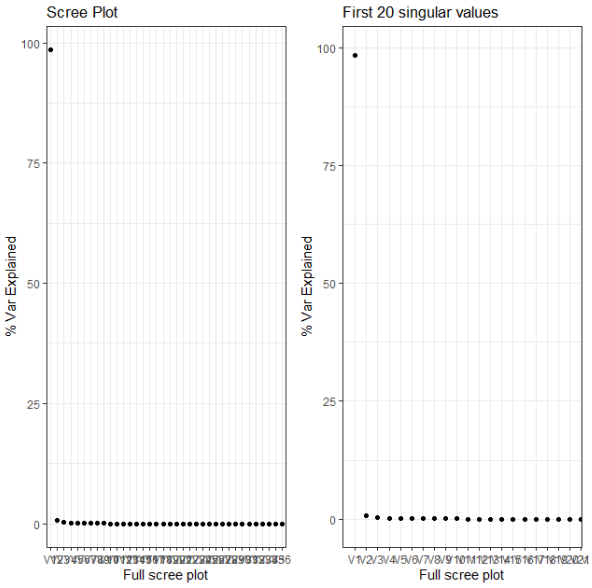

Fig. 5. Eigen Value with *K*-Rank = 20 of SVD Decomposition on GSE114783.

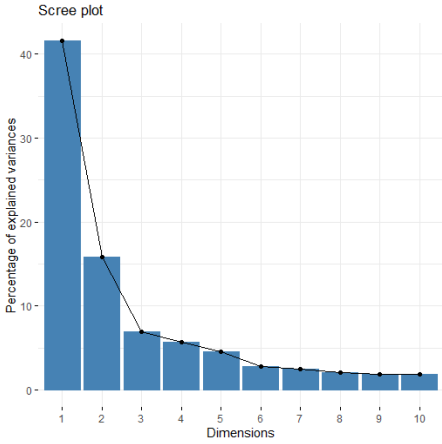

Fig. 6. The Proportion of Variance Explained for each Principal Component in Feature Reduction GSE114873

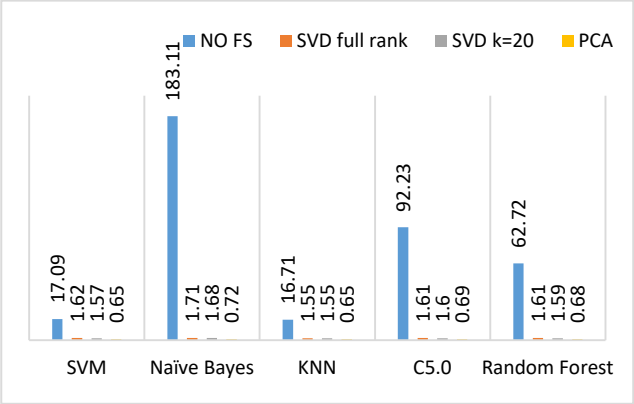

Fig. 7. Comparison of Time Computation GSE114783 using non-Negative Matrix Factorization (NMF-) Dimension Reduction.

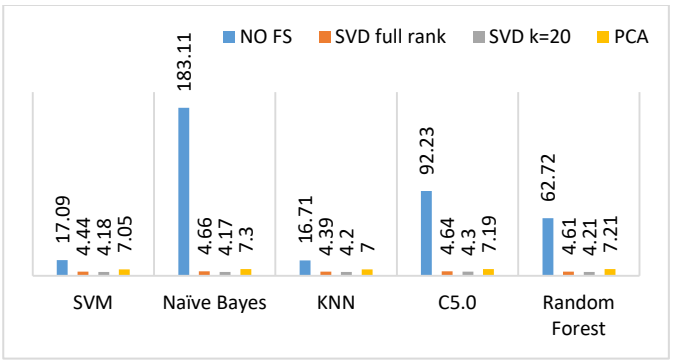

Fig. 8. Comparison of Time Computation GSE55092 using non-Negative Matrix Factorization (NMF-) Dimension Reduction.

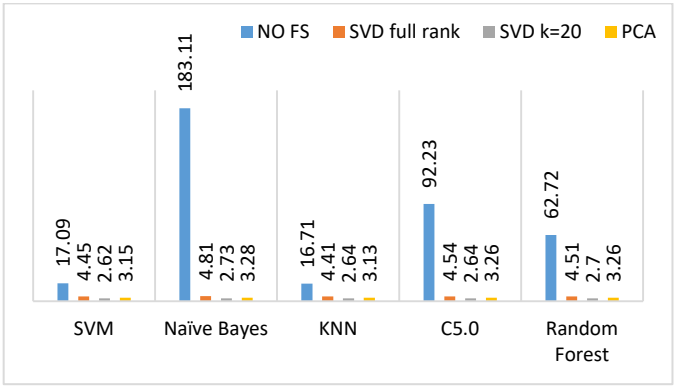

Fig. 9. Comparison of Time Computation GSE121248 using non-Negative Matrix Factorization (NMF-) Dimension Reduction.

V. Conclusion

In this research, feature extraction of the gene expression was applied to reduce high dimensional data using non-negative matrix factorization. Decomposing the data is addressed to get the significant information from the data collection. A unitary matrix consists of singular value with *k*-rank and is produced from decomposition non-negative matrix factorization. The *k-rank* is representative of the number of patterns from value data collection. The maximum dimensional data size after decomposition is the amount of data. Therefore, the proposed method applied unitary matrix from singular vector decomposition (SVD) and Principal Component Analysis (PCA) as data representative to build a classifier model for detection.

The experimental result showed that the reduced data was implemented to the representative supervised learning algorithms for detection and achieved high performance in time and space complexity. The accuracy, sensitivity, and specificity rates are very high, especially for SVM and Random Forest method of 100%. Furthermore, the computation time is very short including decomposing process.

VI. Future Work

A unitary matrix of gene expression data decomposition has k-rank value that indicates the number of patterns in data collection. However, the k-rank is not fixed value for all data set, but it is determined based on eigenvalue convergence. Therefore, it needs to develop method to get the optimum k value.

Acknowledgment

This research is financially supported by the Minister of Research and Technology, the Republic of Indonesia in a program of Doctoral Dissertation grant under contract number: 023/SP2H/LT/DRPM/2021, which was dated March 3, 2021.

References

[1] M. J. Abinash and V. Vasudevan, “A Hybrid Forward Selection Based LASSO Technique for Liver Cancer Classification,” in Nanoelectronics, Circuits and Communication Systems, Singapore, 2019, pp. 185–193. doi: 10.1007/978-981-13-0776-8_17.

[2] A. Akutekwe, H. Seker, and S. Iliya, “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), Oct. 2014, pp. 1–6. doi: 10.1109/ICASTECH.2014.7068140.

[3] H. Kaur, A. Dhall, R. Kumar, and G. P. S. Raghava, “Identification of Platform-Independent Diagnostic Biomarker Panel for Hepatocellular Carcinoma Using Large-Scale Transcriptomics Data,” Frontiers in Genetics, vol. 10, p. 1306, 2020, doi: 10.3389/fgene.2019.01306.

[4] X. Lin et al., “The Robust Classification Model Based on Combinatorial Features,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 2, pp. 650–657, Mar. 2019, doi: 10.1109/TCBB.2017.2779512.

[5] C. Shen and Z. Liu, “Identifying module biomarkers of hepatocellular carcinoma from gene expression data,” in 2017 Chinese Automation Congress (CAC), Oct. 2017, pp. 5404–5407. doi: 10.1109/CAC.2017.8243741.

[6] Z.-M. Zhang, J.-X. Tan, F. Wang, F.-Y. Dao, Z.-Y. Zhang, and H. Lin, “Early Diagnosis of Hepatocellular Carcinoma Using Machine Learning Method,” Front. Bioeng. Biotechnol., vol. 8, 2020, doi: 10.3389/fbioe.2020.00254.

[7] D. J. Kim et al., “Comprehensive Metabolomic Search for Biomarkers to Differentiate Early Stage Hepatocellular Carcinoma from Cirrhosis,” Cancers (Basel), vol. 11, no. 10, p. E1497, Oct. 2019, doi: 10.3390/cancers11101497.

[8] G. H. Golub and C. F. Van Loan, Matrix computations, Fourth edition. Baltimore: The Johns Hopkins University Press, 2013.

[9] X. Gan et al., “Identification of Gene Signatures for Diagnosis and Prognosis of Hepatocellular Carcinomas Patients at Early Stage,” Frontiers in Genetics, vol. 11, p. 857, 2020, doi: 10.3389/fgene.2020.00857.

[10] Y. Lu et al., “Dynamic edge-based biomarker non-invasively predicts hepatocellular carcinoma with hepatitis B virus infection for individual patients based on blood testing,” J Mol Cell Biol, vol. 11, no. 8, pp. 665–677, 19 2019, doi: 10.1093/jmcb/mjz025.

[11] M. Melis et al., “Viral expression and molecular profiling in liver tissue versus microdissected hepatocytes in hepatitis B virus - associated hepatocellular carcinoma,” J Transl Med, vol. 12, p. 230, Aug. 2014, doi: 10.1186/s12967-014-0230-1.

[12] S. M. Wang, L. L. P. J. Ooi, and K. M. Hui, “Identification and Validation of a Novel Gene Signature Associated with the Recurrence of Human Hepatocellular Carcinoma,” Clin Cancer Res, vol. 13, no. 21, pp. 6275–6283, Nov. 2007, doi: 10.1158/1078-0432.CCR-06-2236.

[13] J. Gao and J. Zhang, “Clustered SVD strategies in latent semantic indexing,” Information Processing & Management, vol. 41, no. 5, pp. 1051–1063, Sep. 2005, doi: 10.1016/j.ipm.2004.10.005.

[14] T. M. Mitchell, Machine Learning. New York: McGraw-Hill, 1997.

[15] S. Vijayakumar and S. Wu, “Sequential Support Vector Classifiers and Regression,” 1999.

[16] Data Mining. Elsevier, 2012. doi: 10.1016/C2009-0-61819-5.

[17] O. Sutton, “Introduction to k Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction,” p. 10.

[18] R. Pandya and J. Pandya, “C5. 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning,” IJCA, vol. 117, no. 16, pp. 18–21, May 2015, doi: 10.5120/20639-3318.

[19] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[20] V. Pestov, “Is the k-NN classifier in high dimensions affected by the curse of dimensionality?,” Computers & Mathematics with Applications, vol. 65, no. 10, pp. 1427–1437, May 2013, doi: 10.1016/j.camwa.2012.09.011.