based on the extracted tumor features. Feature extraction and selection are critical to the quality of classifiers founded through data mining methods. To extract useful information and diagnose the tumor, a hybrid of K-means and support vector machine (K-SVM) algorithms is developed. The K-means algorithm is utilized to recognize the hidden patterns of the benign and malignant tumors separately. The membership of each tumor to these patterns is calculated and treated as a new feature in the training model. Then, a support vector machine (SVM) is used to obtain the new classifier to differentiate the incoming tumors. Based on 10-fold cross validation, the proposed methodology improves the accuracy to 97.38%, when tested on the Wisconsin Diagnostic Breast Cancer (WDBC) data set from the University of California – Irvine machine learning repository. Six abstract tumor features are extracted from the 32 original features for the training phase. The results not only illustrate the capability of the proposed approach on breast cancer diagnosis, but also shows time savings during the training phase. Physicians can also benefit from the mined abstract tumor features by better understanding the properties of different types of tumors [23].

III. METHODS

In this paper, we have investigated two data mining techniques: Clustering and Classification. In this paper, we used these algorithms to predict the survivability rate of breast cancer data set. We have selected these two clustering and classification techniques to find the most suitable one for predicting cancer survivability rate [24]. The 24 of 76 features used in our study are listed in Table I.

TABLE I. FEATURE

parameter	*Explain*	*Parameter*	*Explain*
F	Family history	T	Size of the original tumor
MS	Marital status	N	lymph nodes
S	Smoking	N+	Cancer has been found in the lymph nodes
C	Childbirth	STAGE	A number on a scale of 0 through IV
P	Pregnancy	PATH	The type of pathology
TA	Type of abortion	GRADE	Grading is a way of classifying cancer cells
NA	Number of abortion	LVI	Lymphovascular invasion
B	Breastfeeding	ER	Estrogen receptor
H	Hormones (estrogen and progesterone)	PR	Progesterone receptor
DH	Duration of hormone use	HER2	Human epidermal growth factor receptor 2
CT	Computed tomography	P53	Tumor protein
RT	Radio therapy	KI67	Antigen identifiedby monoclonal antibody

In order to make the gathered data being in hospital in Tehran in numerical format, a coding scheme is used. This coding is depicted in Table II for coding of variables describing the general and Table III for variables coding cancer.

TABLE II. CODING OF VARIABLES DESCRIBING THE GENERAL

Gender	*Education*	*Type AB*	*FH*
Female 0	Collegiate 1	Criminal 1	Frist degree 1
Male 1	Diploma 2	Medical 2	Second degree 2
	School 3	C/M 3	Yes/unknown 3
	Guidance 4	Unknown 6	Unknown 6
	Illiterate 5	No 9	No 9

Smoking	*Fat*	*Married*	*Menopause*
Yes 1	Yes 1	Singel 1	Histectomy 1
Yes/no 2	Yes/no 2	Married 2	Natural 2
Unknown 6	Unknown 6	Divorce 3	
No 9	No 9	Widowed 4	
		Unknown 6	

TABLE III. VARIABLES CODING CANCER

Surgery	*Armpit*	*CT*	*H.Name*	*Path*
Bcs 1	AXLND/padding 1	Yes 1	Tamoxifin 1	IDC 1
Mrm 2	AXLND/darrinage 2	Neo 2	Letrozol 2	DCIS 2
Bcs/Mrm 3	SLN/darrinage 3	Unknown 6	Aromysin 3	IDC/DCIS 3
Unknown 6	SLN 4	No 9	tamox/letrozol 4	ILC 4
	AXLND 5		tamox/aromysin 5	ILC/LCIS 5
	Unknown 6		Unknown 6	Unknown 6
	SLN/AXLND 7		tamox/decapept 7	IDC/ILC 7
	Padding 8		herceptin 8	LCIS 8
	No 9			no 9
	Darrinage 10			ILC/DCIS 10
	AXLND/SLN/padding 11			
	SLN/padding 12			

In addition, the coding depicted in Table IV is used to numerate the result of test on each feature.

TABLE IV. CODES COMMON

1	yes positive
9	no negative
6	Unknown

In the next section the results of clustering and classification will be discussed.

IV. Clustering and Classification Results

We use 3 clustering for dataset that select some feature in 3 clusters .this cluster show that some feature such as p53 are more than effective to predict breast cancer. In this study, the models were evaluated based on the accuracy measures discussed above (classification accuracy, sensitivity and specificity). The results were achieved using 74 features for each model, and are based on the average results obtained from the test dataset. It has been chosen 983 patients of 1621 as train data and they are clustered into three, and the results are exist in Table V. This classification algorithm uses LVI as a table and predicts cancer more precisely. Using this algorithm, accuracy numbers are listed in Table V. simulation results have been achieved using Rapid Miner.

TABLE V. RESULTS

Clustering	***Table Recurrence***
Result of data train	91.5 %
Result of data test	89.7 %

LVI as a table that shoe is the best feature for predicting effect time.

Classification	***Label: LVI***
Train data97%	Test Data91%

V. Conclusions and Future Directions

Feature selection is one of the most effective methods to enhance data representation and improve performance in terms of specified criteria, e.g., generalization classification accuracy. In the literature, many studies select a subset of salient features using supervised learning rather than unsupervised learning. When the class labels are absent during training, feature selection in unsupervised learning is integral, but its extensible application is rarely studied in the literature. The objective of this study is to select salient features that can be used to identify interesting clusters in the analysis of cancer diagnosis. Specifically, we highlight three qualitative principles that help users to analyze clinical cancer diagnosis using clusters resulting from a subset of salient features. First, the clusters built by a subset of salient features are more practical and interpretable than those built by all of the features, which include noise. Second, the clustering results provide clinical doctors with an understanding of the context of clinical cancer diagnoses. Finally, a search for relevant records based on the clusters obtained when noisy features are ignored is more efficient. These three principles rely on the discovery of natural clusters using salient features and are applicable only to unsupervised learning. To demonstrate the usefulness of these three qualitative principles, we use coincident quantitative measurements to analyze the salient features for discovering clusters. The experiments on the cancer (Diagnostic) and cancer (Original) datasets demonstrate that the selected features are effective for selecting salient features to discover natural clusters. Based on a performance evaluation using well-known validations in statistical model and cluster analysis, our analysis provides an interesting aspect in feature selection for discovering clusters.

References

[1] M. C. Tayade and M. P. M. Karandikar, "Role of Data Mining Techniques in Healthcare sector in India," *Sch. J. Appl. Med. Sci. SJAMS*, vol. 1, no. 3, pp. 158–160, 2013.

[2] S. Chakrabarti, M. Ester, U. Fayyad, J. Gehrke, J. Han, S. Morishita, G. Piatetsky-Shapiro, and W. Wang, "Data mining curriculum: A proposal (Version 0.91)," 2004.

[3] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[4] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[5] T. M. Mitchell, "Machine learning and data mining," *Commun. ACM*, vol. 42, no. 11, pp. 30–36, 1999.

[6] S. B. Kotsiantis, "Supervised machine learning: a review of classification techniques.," *Inform. 03505596*, vol. 31, no. 3, 2007.

[7] S. G. Jacob and R. G. Ramani, "Efficient Classifier for Classification of Prognostic Breast Cancer Data through Data Mining Techniques," in *Proceedings of the World Congress on Engineering and Computer Science*, 2012, vol. 1.

[8] X. Wu and V. Kumar, *The top ten algorithms in data mining*. CRC Press, 2009.

[9] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv. CSUR*, vol. 31, no. 3, pp. 264–323, 1999.

[10] S. Aruna, D. S. Rajagopalan, and L. V. Nandakishore, "Knowledge based analysis of various statistical tools in detecting breast cancer," *Comput. Sci. Inf. Technol. CSIT*, vol. 2, pp. 37–45, 2011.

[11] A. Christobel and Y, Dr.Sivaprakasam, "An Empirical Comparison of Data Mining Classification Methods.," *Int. J. Comput. Inf. Syst.*, vol. 3, No 2011.

[12] D. Lavanya and D. K. U. Rani, "Analysis of feature selection with classification: Breast cancer datasets," *Indian J. Comput. Sci. Eng. IJCSE*, 2011.

[13] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: an application to face detection," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1997, pp. 130–136.

[14] V. N. Chunekar and H. P. Ambulgekar, "Approach of Neural Network to Diagnose Breast Cancer on three different Data Set," in *Advances in Recent Technologies in Communication and Computing, 2009. ARTCom'09. International Conference on*, 2009, pp. 893–895.

[15] D. Lavanya and K. Usha Rani, "ENSEMBLE DECISION TREE CLASSIFIER FOR BREAST CANCER DATA.," *Int. J. Inf. Technol. Converg. Serv.*, vol. 2, no. 1, 2012.

[16] B. Šter and A. Dobnikar, *Neural network in medical diagnosis: comparison with other methods*. 1996.

[17] T. Joachims, "Transductive inference for text classification using support vector machines," in *ICML*, 1999, vol. 99, pp. 200–209.

[18] J. Abonyi and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers," *Pattern Recognit. Lett.*, vol. 24, no. 14, pp. 2195–2207, 2003.

[19] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, and others undefined, "Constrained k-means clustering with background knowledge," in *ICML*, 2001, vol. 1, pp. 577–584.

[20] J. Shi and Z. Luo, "Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples," *Comput. Biol. Med.*, vol. 40, no. 8, pp. 723–732, Aug. 2010.

[21] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[22] G. Krishnasamy, A. J. Kulkarni, and R. Paramesran, "A hybrid approach for data clustering based on modified cohort intelligence and K-means," *Expert Syst. Appl.*, vol. 41, no. 13, pp. 6009–6016, Oct. 2014.

[23] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Syst. Appl.*, vol. 41, no. 4, Part 1, pp. 1476–1482, Mar. 2014.

[24] V. Chaurasia and S. Pal, *Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability*. IJCSMC, 2014.