For cited: *Ahmad, A. A., Polat, H., Olegovich, T. M., Ahmad, A. A. & Porfirevich, D. V. (2022). Breast Cancer Predication Using Data Mining Techniques. Journal of Current Research on Engineering, Science and Technology, 8 (1), 69-80.*

Journal of Current Research on Engineering, Science and Technology (JoCREST)
ISSN: 2651-2521

Year: 2022
Volume: 8
Issue: 1

Crossref doi: 10.26579/jocrest.8.1.7

Research Article

Breast Cancer Predication Using Data Mining Techniques

Ahmad Ayid AHMAD[1], Hüseyin POLAT[2], Tanygin Maxim OLEGOVICH[3], Ali Ayid AHMAD[4] & Dobritsa Vyacheslav PORFIREVICH[5]

Keywords
Breast cancer; Decision tree; Naïve Bayes; attribute reduction; classification, PCA, LDA.

Abstract
As an addendum to the reviewed recent studies on the detection of breast cancer and came to the conclusion that proposed a model to aid in the resolution of the problem of evaluating the severity of the disease's risk, and to learn about the best practices, to reduce time and cost with the aim of improving well-being. We compared three automated methods for blood diseases detection using different method for attribute reduction: PCA (Principle Component Analysis), LDA (Linear Discriminate Analysis), ICA (Independent Component Analysis) ,original data set and six algorithms, which are Naive Bayes, K-Nearest Neighbor, Decision Tree, Logistic Regression, ANN(Artificial Neural Network), and SVM (Support Vector Machine) for classification. On data set which we have get from kaggle. The outcomes show that in the wake of applying the PCA procedure, the exactness is: 0.909, 0. 87 ,0.91, 0.72, 0.904 and 0.90 for Naive Bayes, Decision Tree, Logistic Regression, SVM, ANN and KNN individually and in the wake of applying the ICA strategy, the exactness is : 0.92, 0.89, 0.93, 0.92, 0.92 and 0.90 for Naive Bayes, Decision Tree, Logistic Regression, SVM, ANN and KNN separately , the precision in the wake of applying the LDA procedure is: 0.92, 0. 90, 0. 90 , 0.925, 0.92 and 0.91 for Naive Bayes, Decision Tree, Logistic Regression, SVM, ANN and KNN separately, while the exactness on the first informational index is: 0.91, 0. 87, 0. 91 , 0.72, 0.91 and 0.91 for Naive Bayes, Decision Tree, Logistic Regression, SVM, ANN and KNN separately. All in all the outcomes gave the idea that when the credulous bayes calculation give the most noteworthy precision subsequent to applying the ICA and LDA strategy which is 92%.

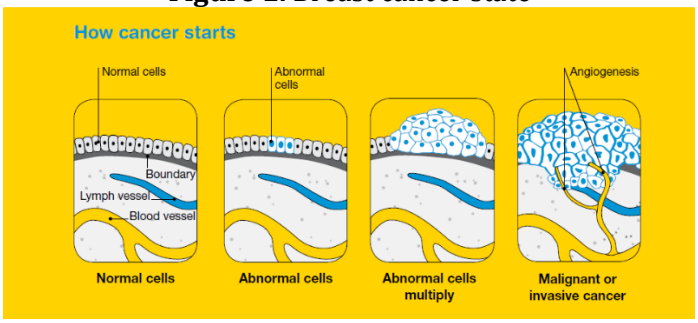
Article History
Received
9 Mar, 2022
Accepted
14 May, 2022

[1] Corresponding Author. ORCID: 0000-0002-6031-9414. Department of Computer Engineering, Gazi University, Ankara, Turkey, ahmadayid@yahoo.com
[2] ORCID: 0000-0001-8921-140X. Department of Computer Engineering, Gazi University, Ankara, Turkey, polath@gazi.edu.tr
[3] ORCID: 0000-0002-4099-1414. Southwest State University Kursk, Russian Federation, tanygin@yandex.ru
[4] ORCID: 0000-0002-6031-9414. Kirkuk University, Kirkuk, Iraq, aliayid2013@gmail.com
[5] ORCID: 0000-0001-7236-6552. Southwest State University Kursk, Russian Federation, dobritsa@mail.ru

1. Introduction

Cancer is a genetic disease that is characterized with cell growth. The tumor has the ability to destroy nearby tissue by invasion and spread to other parts of the body by metastasis. Despite significant improvements in early detection of cancer and new treatment strategies, more than sixty percent of the people diagnosed with cancer worldwide still have died. Breast cancer is one of the most commonly seen cancer types, which is unique to women. Figure 1 shows how to start breast cancer.

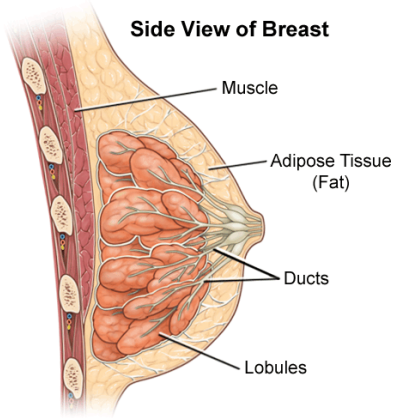
Figure 1. Breast cancer state



In this figure, first image on the left are the normal cells while the second left demonstrates the abnormality of the breast cells. After some period, the abnormal cells appear and the abnormal section will increase and this cancer gets multiple cells, which can be seen on the right two images. The final image illustrates how much the cancer cell took over. This is the most dangerous era of the cancer.

Breast cancer is a common problem that occurs in women more than 1 million cases per year worldwide. Therefore, early diagnosis of breast cancer in women's life plays a very important role. The goal is to detect cancers before symptoms start. A breast cancer type and propagation velocity are the most important factor for successful treatment. Thus, early detection tests for breast cancer can save thousands of lives every year. Breast consists of glands and greasy tissue among the skin, chest wall, blood vessels and lymph vessels. Each node also called lobules, and many lobes, a lobe form. Here are 15 to 20 parts in each breast. The Anatomy of the breast tissue is shown in figure 2.

Figure 2. Anatomy of the breast tissue



The lymph nodes of the breast have a clear liquid waste. Lymph nodes filter lymph and little tissue to clean; there are pea-sized pieces. Breast drainage is often called the axillary lymph nodes under the arm.

Cancer is a name given to a group of diseases that affect several organs in the human body. Cancer begins in the body when cells begin to spread and divide randomly (Institute, N. C, 2015). Cancer begins to appear anywhere in the human body and may arise in any of the billions of cells in the body. Usually, the cells of the body divide and new cells appear in the order and according to the need for them. When the cells age or become damaged, new cells replace the damaged cells (Longo et. al, 2012). When cancer appears, this system completely collapses, with new cells that are not needed, while old or damaged cells survive. New cells that are not needed by the body continue to grow, develop and spread in the body, in what is known as tumors (Duke et. al, 1996). In other words, cancer occurs when cells in one portion of the body divide uncontrollably and cause harm to other cells. Currently, cancer has become one of the main causes of death all over the world. Several factors affect the creation or spreading cancers including: gender, age, genetics, marital status, quality of life, living location etc. (Duke et. al, 1996). Data mining in a medical database takes advantage of a vast volume of data and is considered a methodology for changing and managing information to relevant data as well as extracting hidden data from preprocessed information. Data mining in the medical archive takes advantage of the vast volume of data which is a methodology for changing and managing information to relevant data and extracting hidden data from preprocessed information. To predict the probability of heart attack, various data mining strategies such as Nave Bayes, KNN algorithm, Decision tree, and Neural Network are used. The KNN algorithm finds the values of cancer factors by using the K user defined value. The classification report for cancer is generated using the decision tree algorithm. The Naïve Bayes method is used to predict cancer through probability. The Neural Network gives the limited mistake of the forecast of disease. Many quality decrease methods like Principal Component Analysis (PCA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA). Head segment investigation (PCA) is an integral asset for property decrease as proposed by Turk and Pentland. The principle benefit of PCA is that it can diminish the element of the information without losing a lot of data. Break down segregate direct (LDA) otherwise called Fisher's Discriminate Analysis, is another dimensionality decrease strategy, it decides a subspace where the between-class disperse (additional individual inconstancy) is just about as extensive as could really be expected, while the inside class dissipate (intrapersonal changeability) is kept steady. Autonomous part examination (ICA) produce premise vectors that are measurably free, it is option of PCA, which give an all the more remarkable information portrayal, and it's a separate investigation rule which can be utilized improve PCA (Bouzalmat et. al,2014). On account of AI calculations and figuring science, specialists can anticipate bosom malignant growth at a previous stage. This paper gives a knowledge about information mining procedure used to anticipate bosom malignancy.

2. Related Studies

Numerous investigates utilized information mining procedures to acquire new information , find obscure examples in the clinical field, and examine cautiously the information they need to discover the reasons for sicknesses uncommonly malignancy, diabetes , thalassemia, heart infections, AIDS… etc.in the following line, we present a portion of these explores. Ashraf Y. A. Maghari, Asem H. Shurrab used medical data set and try to find the relationship between CBC and Tumor, using three of data mining techniques, which are Decision tree, Naïve Bayes and Rule Induction(Maghari& Shurrab,2017). Clare, D., and Avanija have expressed Hybrid diseases predict system (HDPS) with high-level accuracy using Neural Network, Naïve Bays', Decision tree, Linear Regression and EM Algorithm (Clare& Avanija,2016). Lokanayaki and Malathi have discussed a well-balanced dataset is very important for creating a good predictive model. Medical datasets are often not balanced in their class labels(Lokanayaki &Malathi,2013). Shubpreet kaur and Bawa Review on Data Mining Techniques in Healthcare using Neural Network, Naïve Bays', Decision tree and KNN (kaur & Bawa,2017).

3. Materials And Methods

3.1. Dataset And Preprocessing

Business understanding: understanding the problem in breast cancer, thinking how to use data mining to predict the breast cancer. Collecting data: understand patient's data after collecting it from kaggle. Structuring: organizing the data, which is necessary because raw data comes in many different shapes and sizes. Missing values: dealing with missing values, by deleting the rows that containing missing values. We divided the data set into two parts, the first one is represent 60% iof the data and it represent the training set, while the other is 40% of the data and represent the test set.

3.2. Experiment

In this stage we lead probes our information utilizing three characteristic size decrease methods, which are Principal Component Analysis (PCA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA), while the fourth trial is on the first informational index. Subsequent to applying all of these techniques, we apply sixe grouping calculations, which are Naive Bayes, K-Nearest Neighbor, Decision Tree, Logistic Regression, Artificial Neural Network, and Support Vector Machine. At the end of the day we apply the test multiple times on a similar informational collection. The test is applied by python language, numerous instruments are utilized which are, dominate 2016 and boa constrictor.

3.3. Experimental Results and Findings

To evaluate a classifier, you have to apply it in a number of cases where one has knowledge of the "true" class of the respective objects, at least in retrospect. An example of such a case is a medical laboratory test, which seeks to determine whether a person has a particular disease. Later it is determined by more complex tests, whether the person actually suffers from this disease. The test represents a classifier, which "sick" and "healthy" classifies the persons in the categories. Since this is a yes / no question, we also say that the test is positive (classification "sick")

or negative (classification "healthy") from. In order to assess how well suited the laboratory test for the diagnosis of the disease, its actual state of health is now compared with the result of the tests in each patient. There are four possible cases may occur. True positive: The patient is sick, and the test has the display correctly. False negative: The patient is sick, but the test has wrongly classified it as healthy. False positive: The patient is healthy, but the test has wrongly classified it as ill. True negative: The patient is healthy, and the test has the display correctly. The parameters that used for evaluation results is shown in table 1.

Table 1. The parameters that used for evaluation results

Parameter	Abbreviation	Mathematic expression
Sensitivity	Recall, hit rate, True Positive Rate (TPR)	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$
Specificity	Selectivity, True Negative Rate (TNR)	$\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1 - \mathrm{FPR}$
Precision	Positive Predictive Value (PPV)	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = 1 - \mathrm{FDR}$
Accuracy	(ACC)	$\frac{\mathrm{TP}+\mathrm{TN}}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})}$
Balanced Accuracy	(BA)	$\frac{\mathrm{TPR}+\mathrm{TNR}}{2}$
F1 Score	-	$\frac{2*(\mathrm{PPV}*\mathrm{TPR})}{(\mathrm{PPV}*\mathrm{TPR})} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$

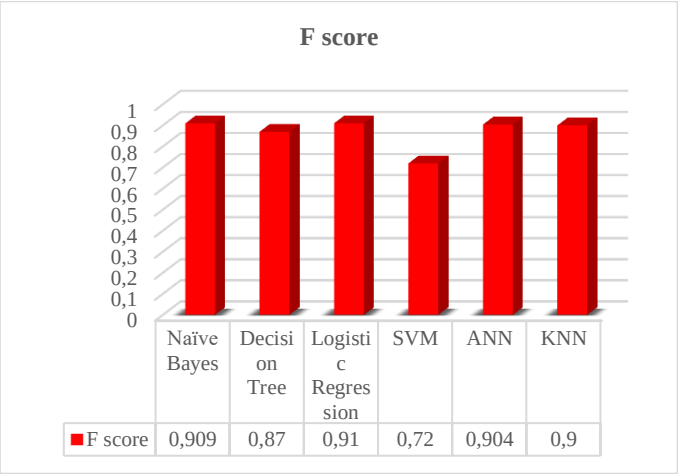
As referenced above, we applied the examination multiple times, five calculations are utilized without fail. The main test is applied by utilizing Principal Component Analysis (PCA) to lessen the quality size, Naive Bayes, K-Nearest Neighbors, Decision Tree, Logistic Regression, Artificial Neural Network, and Support Vector Machine are applied in the subsequent advance.

The results of the first experiment is as appear in the table 1 and figure 3.

Table 1. The results of classification algorithm after applying PCA method

Algorithm	Naïve Bayes	Decision Tree	Logistic Regression	SVM	ANN	KNN
F score	0.909	0.87	.91	0.72	0.904	0.90

Figure 3. Classifiers algorithms results after applying PCA method

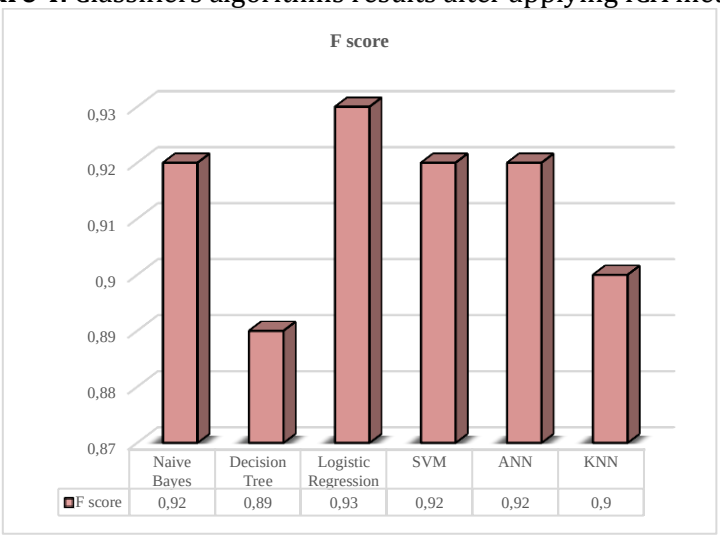


The subsequent investigation is applied by utilizing Independent Component Analysis (ICA), to lessen the characteristic size .similar calculations which are applied in the main examination are applied in the subsequent test. The aftereffects of the subsequent examination is as show up in the table 2 and figure 4.

Table 2. The results of classification algorithm after applying ICA method

Algorithm	Naive Bayes	Decision Tree	Logistic Regression	SVM	ANN	KNN
F score	0.92	0.89	0.93	0.92	0.92	0.90

Figure 4. Classifiers algorithms results after applying ICA method

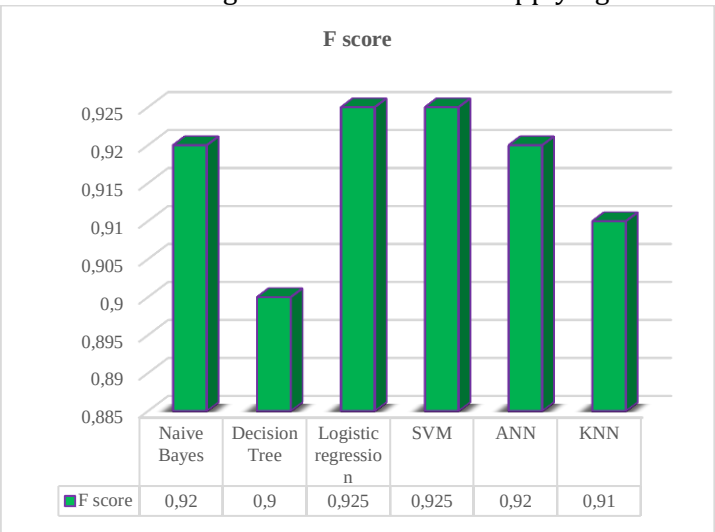


The third examination is applied by Linear Discriminant Analysis (LDA), to lessen the property size .similar calculations which are applied in the last trials are applied in the subsequent trial. The consequences of the third investigation is as show up in the table 3 and figure 5.

Table 3. The results of classification algorithm after applying LDA method

Algorithm	Naive Bayes	Decision Tree	Logistic regression	SVM	ANN	KNN
F score	0.92	0.90	0.925	0.925	0.92	0.91

Figure 5. classifiers algorithms results after applying LDA method

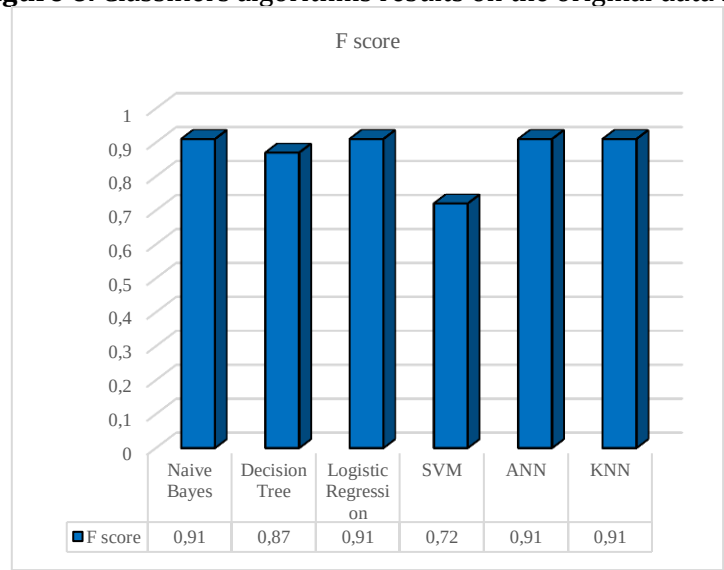


The last experiment is applied on the original data set. The results of the last experiment is as appear in the table 4 and figure 6.

Table 4. The results of classification algorithm on the original data set

Algorithm	Naive Bayes	Decision Tree	Logistic Regression	SVM	ANN	KNN
F score	0.91	0.87	0.91	0.72	0.91	0.91

Figure 6. Classifiers algorithms results on the original data set



When compare the same algorithms after applying the attribute reduction, the results appears as shown in figures 7, 8, 9, 10, 11 and 12.

Figure 7. Naive Bayes accuracy values

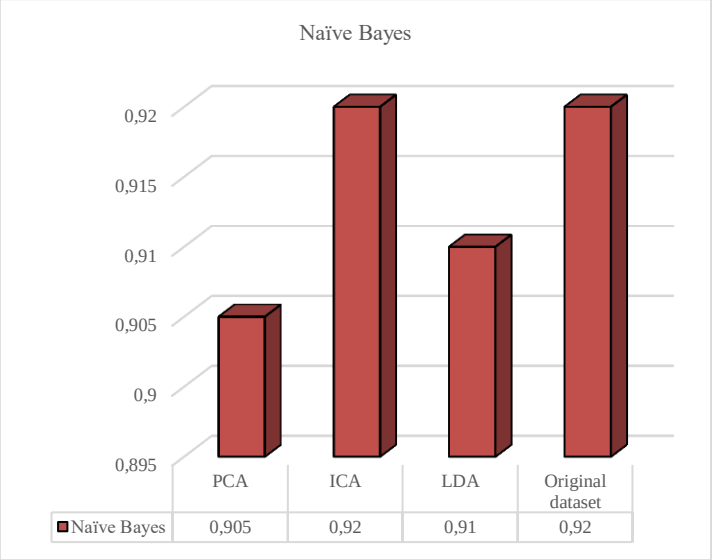


Figure 8. Decision Tree accuracy values

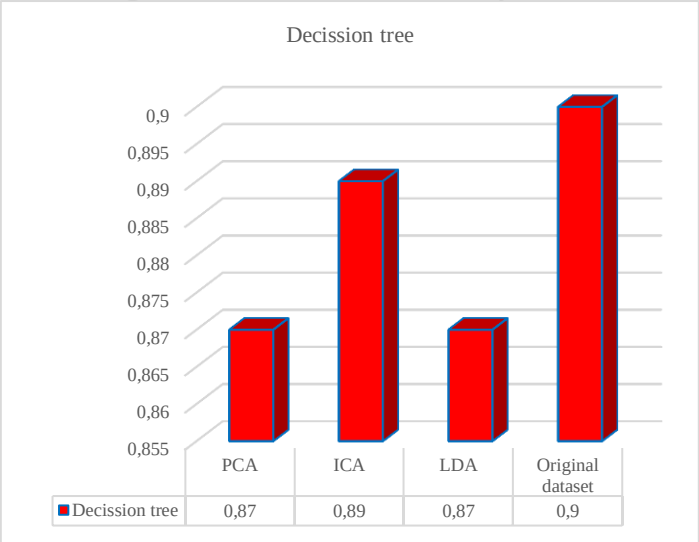


Figure 9. Logistic Regression accuracy values

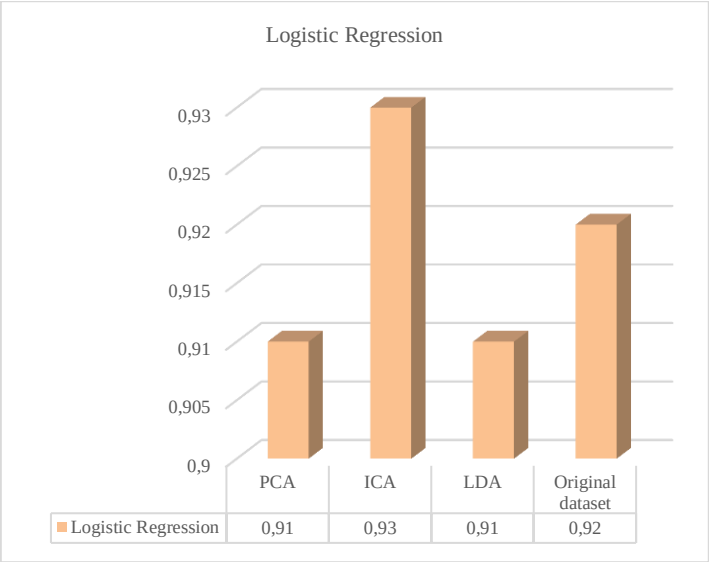


Figure 10. SVM accuracy values

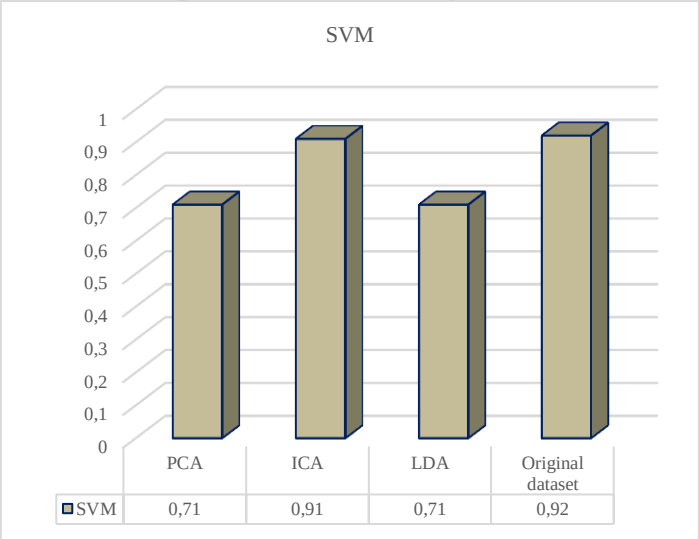


Figure 11. ANN accuracy values

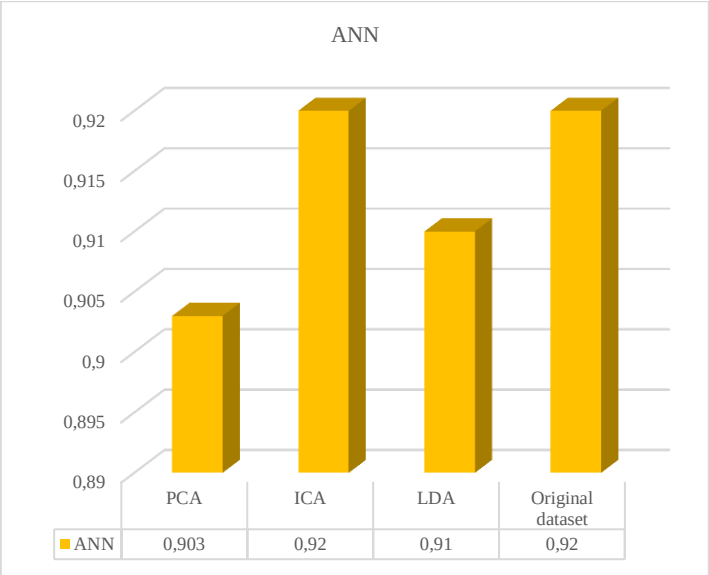
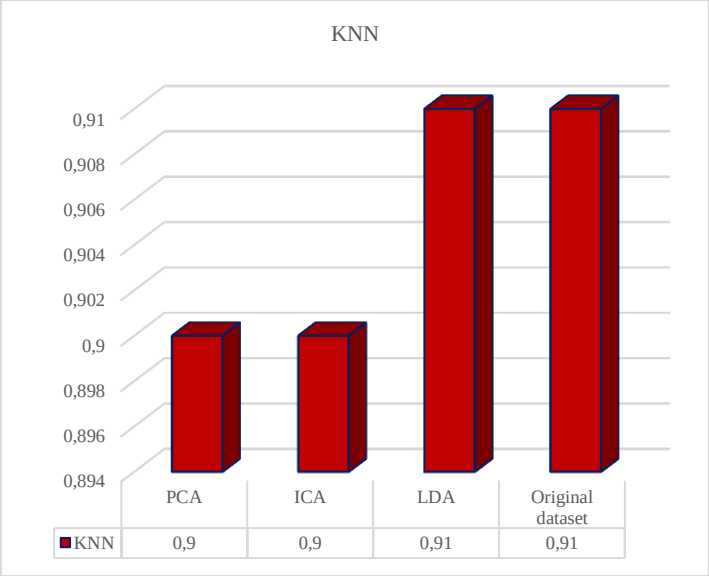


Figure 12. KNN accuracy values



4. Results and Recommendations

In the world of medicine, data mining has become a basic technique for computer applications. Data mining is a technique for extracting useful information from raw data sets by investigating and compressing them while taking into account multiple points of view. Medical data mining is a series of techniques for extracting useful and innovative knowledge from human resources records in order to assist physicians in making the best decision possible. Cancer and diabetes have been the leading causes of death in the recent years around the world. As a result, data mining is the most commonly used component, as it involves extensive and time-consuming experiments. In this research compared three automated methods for blood diseases detection using different method for attribute reduction: PCA (Principle Component Analysis), LDA (Linear Discriminate Analysis), ICA

(Independent Component Analysis) ,original data set and six algorithms, which are Naive Bayes, K-Nearest Neighbor, Decision Tree, Logistic Regression, ANN(Artificial Neural Network), and SVM (Support Vector Machine) for classification. On data set which we have get from kaggle. The results show that after applying the PCA technique, the accuracy is: 0.909, 0. 87 ,0.91, 0.72, 0.904 and 0.90 for Naive Bayes, Decision Tree, Logistic Regression, SVM, ANN and KNN respectively and after applying the ICA technique, the accuracy is : 0.92, 0.89, 0.93, 0.92, 0.92 and 0.90 for Naive Bayes, Decision Tree, Logistic Regression, SVM, ANN and KNN respectively , the accuracy after applying the LDA technique is: 0.92, 0. 90, 0. 90 , 0.925, 0.92 and 0.91 for Naive Bayes, Decision Tree, Logistic Regression, SVM, ANN and KNN respectively, while the accuracy on the original data set is: 0.91, 0. 87, 0. 91 , 0.72, 0.91 and 0.91 for Naive Bayes, Decision Tree, Logistic Regression, SVM, ANN and KNN respectively.

In other words the results appeared that when the naïve bayes algorithm give the highest accuracy after applying the ICA and LDA method which is 92%.

References

Ashraf Y. A. Maghari, Asem H. Shurrab," Blood Diseases Detection Using Data Mining Techniques ", 2017 8th International Conference on Information Technology (ICIT).

Bouzalmat, A., Kharroubi, J., & Zarghili, A. (2014). Comparative study of PCA, ICA, LDA using SVM classifier. *Journal of Emerging Technologies in Web Intelligence*, *6*(1), 64–68. https://doi.org/10.4304/jetwi.6.1.64-68

Clare, D., and Avanija, J., "Contextual Learning Approach to Improve Diagnostic Accuracy for Hybrid (Lung, HIV and Heat) Diseases", IJSTE International Journal of Science Technology & Engineering, May 2016, Volume. 2, Issue. 11, pp: 585-588.

Duke, R. C., Ojcius, D. M., & Young, J. D.-E. (1996). Cell suicide in health and disease. Scientific American, 275(6), 80-87.
[4] Duke, R. C., Ojcius, D. M., & Young, J. D.-E. (1996). Cell suicide in health and disease. Scientific American, 275(6), 80-87.

Institute, N. C. (February 9, 2015). What Is Cancer? Retrieved: June 1, 2017 from https://www.cancer.gov/about-cancer/understanding/what-iscancer,

Lokanayaki and Malathi, "Data Preprocessing for Liver Dataset Using SMOTE," International Journal of Advanced Research in Computer Science and Software Engineering, Volume. 3, No. 11, Nov. 2013.

Longo, D. L., Fauci, A. S., Kasper, D. L., Hauser, S. L., Jameson, J. L., & Loscalzo, J. (2012). Harrison's Principles of Internal Medicine 18E Vol 2 EB: McGraw Hill Professional.

Shubpreet kaur and Bawa" Review on Data Mining Techniques in Healthcare", Proceedings of the Second International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud, 2017.