Sustainable Engineering and Innovation ISSN 2712-0562
Vol.2, No.2, December 2020, pp.109-118
https://doi.org/10.37868/sei.v2i2.108

Breast cancer identification based on artificial intelligent system

Hassan Khalil Silman[1*], Akbas Ezaldeen Ali[2]
[1]Computer Science Department, University of technology, Iraq-Baghdad
[2]Computer Science Department, University of technology, Iraq-Baghdad

*Corresponding author: hassan1994kh@gmail.com

Abstract

Worldwide, breast cancer causes a high mortality rate. Early diagnosis is important for treatment, but high-density breast tissues are difficult to analyze. Computer-assisted identification systems were introduced to classify by fine needle aspirates FNA with features that better represent the images to be classified as a major challenge. This work is fully automated, and it does not require any manual intervention from user. In this analysis, various texture definitions for the portrayal of breast tissue density on mammograms are examined within addition to contrasting them with other techniques. We have created an algorithm that can be divided into three classes: fatty, fatty-glandular and dense-glandular. The suggested system works in a spatial-related domain and it results with extreme immunity to noise and background area, with a high rate of precision.

Keywords: Breast Cancer; Fine Needle Aspirates; Classification; KNN; NB; J48

1 Introduction

Breast cancer is considered as one prevalent killing disease found in human being (female). It is difficult to safeguard the breast canncer in its early stage, due to the reason that leads to this disease is still unavailable for the scientists. William H. Wolberg, et al, in this work evaluates and diagnoses an interactive computer system based on cytologic features extracted directly from a digital scan of fine-needle aspirate (FNA) slides. The data set consists of a series of 569 patients to develop the system, furthermore, 54 consecutive new patients provided samples to test the system. This work reaches the accuracy of the system predicted with tenfold cross-validation was 97% [1]. Aik Choon tan and David Gilbert condense on three different classification tasks based on several breast cancer available data. They observe that ensemble learning often performs better than a single classification tree [2].

Diana Dumitru explores in 2009 the possible contribution of the Naive Bayesian identification system to distinguish early breast cancer as an effective support in software-aided diagnosis. The well-known dataset for breast cancer in Wisconsin Prognostic was adopted. The analysis showed that the suggested recognizer delivers identical output to other machine learning techniques to low computational effort and high speed [3].

Furthermore, in [4] a breast cancer tumor recognition based on ultrasound images is suggested. The researcher concentrates on the predictive technology of recognizing the state of tumors in the breast tissues. In 2012 breast cancer detection is suggested based on analysis on Confocal Microwave Imaging Algorithm CMI [5].

Studies suggest that data mining techniques are being used to establish predictive models for the recurrence of breast cancer in clinicians who have been followed up for two years [6]. The performance of DT, ANN and SVM out of this process is 0.936, 0.947 and 0.957. Dayong Wang et.al proposed deep learning for identifying metastatic breast cancer. The success rate is 0.995 based on the proposed system [7]. Hiba Asri et al. suggested a comparison of outputs between different machine learning algorithms: SVM, Decision Tree (C4.5), NB and

KNN on the initial Wisconsin Breast Cancer datasets. Experimental results show that SVM with the smallest error rate provides the highest identification (97.13%) [8][5]. In order to use in microarrays of the breast cancer gene expression datasets, an ensemble classifier with correlation-based feature selection with forwarding search is proposed in [9].

1.1 Breast cancer (BC)

The most common type of cancer is BC, the leading cause of cancer murder among women around the world. The illness affects about 10 percent of all women in the Western world at some stage of their lives. This can be diagnosed with careful clinical history analysis, physical examination and mammographic or ultrasonic imaging. Nevertheless, accurate breast mass diagnosis can only be confirmed by biopsy of the fine needle aspiration (FNA), core needle biopsy, or excisional biopsy. FNA is the easiest and fastest form of breast biopsy among these methods and is useful for women with fluid-filled cysts. Research work on the Wisconsin Diagnosis of BC (WDBC) data developed out of the desire to accurately diagnose breast masses based solely on FNA. To improve accuracy and quality of BC detection, a number of scientific initiatives focus on updating methods for Computer Aided Diagnosis (CAD) with FNA BC, including work on image analysis and computational analysis [10].

1.1.1 Types of BC

Two types of benign and malignant tumors are decomposed into BC [11]:
• **Benign** tumors are non-hazardous tumors; their contours are well established. They gradually evolve in the organ they appeared without generating metastatic cases. Benign tumors consist of cells which are identical to normal breast tissue cells.
• **Malignant** tumors are risky tumors, because they can spread to other body organs and cause metastatic cases. Cancer cells of malignant tumors have several abnormalities in shape, size and contours compared to normal cells, where cells lose their original characteristics.

1.1.2 Causes of BC

The first risk factor that can increase the likelihood of BC is the age factor; with age, the risk of BC increases. Some factors that can intervene such as [11]:

- **Family factors, or genetic factors.**
- **Gender:** Women suffer the most from BC.
- **A woman history:** Woman who already had BC in one breast has an increased risk of developing cancer in the other breast.
- **A family history:** When several of the woman's parents are diagnosed with BC, especially at a younger age, the risk of developing BC increases.
- **Genetic factors:** Many genetic mutations cause BC more likely.
- **Obesity:** The risk of BC increases with obesity.
- **Having period in early age**
- **Late menopause:** Woman who started menopause at a later age has a greater chance of developing BC.
- **Having the first child in old age** omen who give birth to their first child after 30 years of age may have increased BC risk.
- **Women who have never been pregnant:** By not having a child, the risk of developing BC is increased.
- **Hormone replacement therapy:** Estrogens and progesterone raises BC risk after 5 years' treatment.
- **Drinking alcohol:** Alcohol consumption raises the risk of BC.

1.2 Fine needle aspiration

Slight drop of viscous fluid collected from the breast through multiple passes with a 23-gage needle as a negative pressure is contributed to an embedded syringe. The aspirated material was expressed onto a silage-coated glass

slide. A specific slide was mounted face down on the amplification and the aspiration extended as the slides were separated horizontally. Arrangements were fixed in 95"/0 ethanol and examined immediately after they were stained with hematoxylin and eosin. Only measurable masses were aspired and only solid masses containing epithelial cells analyzed [11].
The main conclusions of the study may be presented in a short Conclusions section, which may stand alone or form a subsection of a Discussion or Results and Discussion section.

2 Methodology

Each section explains the various stages included in achieving the aim of each research referred to in the literature review. The various stages re consist of pre-processing the input image to remove the characteristics, that are then added to the classifier as an input. The classifier's performance differentiates the ordinary, harmless and malicious cases from the fine needle aspirates applied. After preprocessing operation, several types of features (FNA) were obtained.

2.1 Extraction feature

When the actual image has been pre-processed, the related features listed in the literature can be extracted. The extraction of features may be defined as extracting important fine information from the given input while rejecting all other data [12].
Type of features
WDBC database includes of 569 breast masses with 357 benign and 212malignantcases. In order to evaluate the size, shape, and texture of each cell nuclei, ten characteristics were derived and described as follows [13,14]: **Radius:** (is) measured by averaging the length of segments of the radial line from the mass center of the boundary to each boundary point. **Perimeter:** (is) evaluated as the amount of distances from sequential boundaries. **Area: (**is) computed by calculating the number of pixels inside the boundary and adding half the pixel son perimeter to correct the digitization error. **Compactness:** (is) integrates the diameter and area of the cell to determine its compactness, measured as perimeter2/(area),**Smoothness:** (is) measured by calculating the discrepancy between the length of each circular line and the average length of the two radial lines around it as shown (∑points|ri-(ri+ri+1)/2)/perimeter,
Such that *ri* represents the length of the line across the center of mass boundary to each point.

• *Concavity*
It is obtained by calculating any indentation size at the cell nucleus boundary.
• *Concave points*
It is close to concavity, but only counts the number of boundary points on the concave boundary regions and not the size of those concavities.
• *Symmetry*
Is determined by discovering the proportional length difference between sets of line segments perpendicular to the cell nucleus contour main axis, calculated bysymmetry= =(∑i|lefti-righti)/(∑i(lefti+righti)),
where “lefti” and “righti” represent the lengths of segments perpendicular on the left and right of the major axis.
• *Fractal dimension*
Is accurately measured using the Mandelbrot "coastline approximation," the diameter of the nucleus is calculated using increasingly larger "rulers". As the scale of the ruler rises, the measuring accuracy reduces, and the diameter observed decreases. Plotting these values on a log-log scale and calculating the downward slope means that the relation to the fractal dimension is negative.

2.2 Classification

After extracting the relevant function, the final stage is to identify the obtained data and assign it to a given class. To this end, classifier like Help KNN, NB, J48 [15].

2.3 Proposed system

The overall system consists of two-phase training and testing. The training phase includes 4 stages as shown in Figure 1, first of which is image acquisition, second discrete wavelet transformation for segmentation, while the next stage is fine needle aspirates extraction features, selection of more optimal features. Finally, classification stage is to identify suitable fine needle aspirates class. Features represent the image in a format that focuses on relevant information in particular. For training and testing, the next stage features are selected; this phase is very critical because classification accuracy depends primarily on careful feature choice. In the other step, the fine needle aspirates are classified into normal and malignant class, in order to distinguish fine needle aspirates and identify the BC type.

On the other hand, the testing consists of the same stages that belong to the training phase. The difference between these two phases is that the testing uses the internet in order to send the fine needle aspirates remotely through the website to check the BC type.

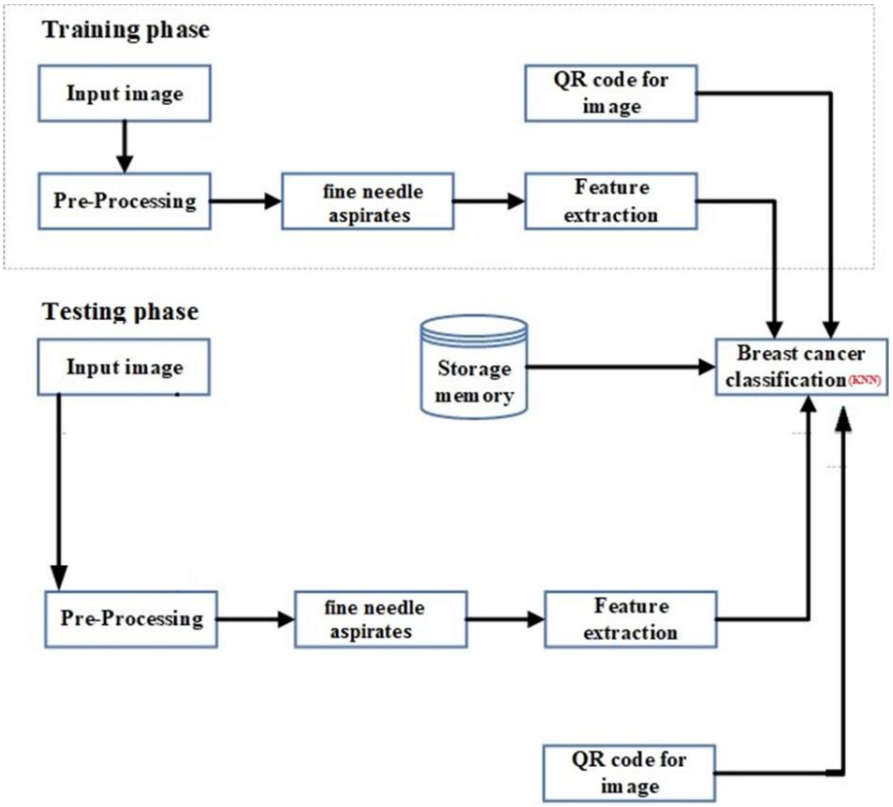

Figure 1. Proposed system

2.4 Classification techniques

Classification is supervised leaning based on known properties that focus on the prediction. A classification function begins with a set of data where class assignments are identified. If there are numerical values for the aim or class mark then a statistical model is used. There are many classification algorithms that mostly use certain methods like Naive Bayes, KNN, and j48. However, breast cancer is considered as one of the diseases that cause a high death toll every year. It is the most common form of all cancers and the leading cause of women's death worldwide. Data mining techniques and classification methods are an effective way of classifying data, particularly in the healthcare sector, where these approaches are commonly used to make diagnostic and analytical decisions [16]. A performance comparison of various machine learning algorithms was made on the Wisconsin BC Datasets (original): Support J48, (k-NN) and Naive Bayes (NB). The main objective was to determine the quality of data classification in terms of the efficacy and efficiency in terms of accuracy of each algorithm, accuracy, sensitivity [11].

3 Results and discussion

A performance comparison is made on the WBCD (Wisconsin Breast Cancer Diagnosis) datasets between different machine learning algorithms: help j48, Naive Bayes (NB) and k Nearest Neighbors (k-NN). The main objective is to determine the quality of the classification of data in terms of the efficiency and effectiveness of each algorithm in terms of accuracy, sensitivity, and specificity. Experimental results show that KNN offers the highest accuracy with the lowest error rate compared to NB classifiers, j48 as shown in Figure 2.

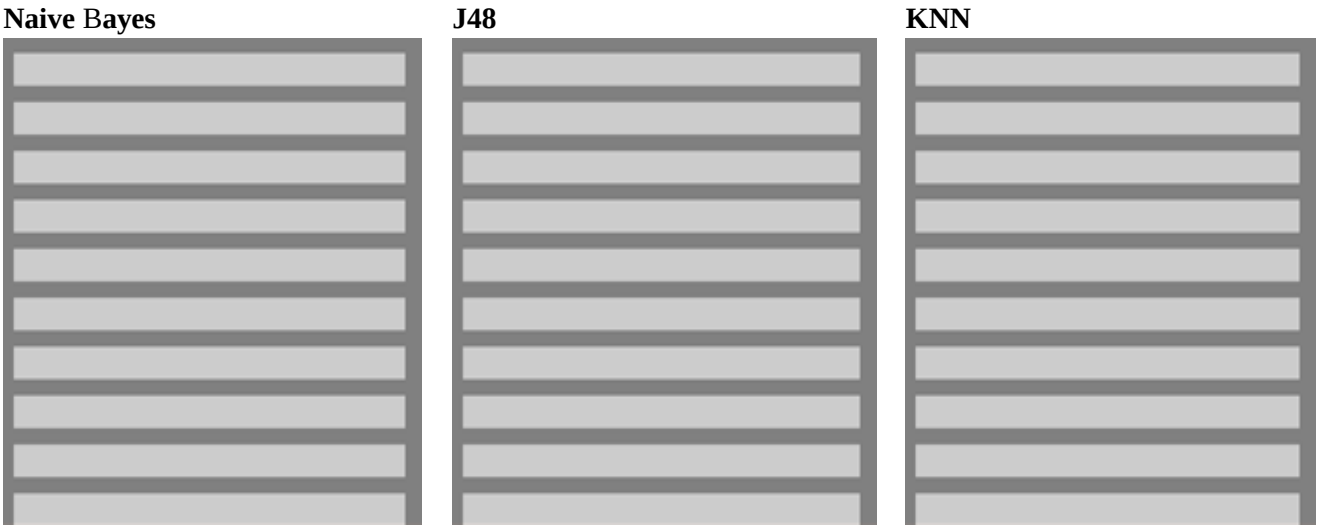

Figure 2. Main interface of the system classification

3.1 Naive Bayes (NB)

After the first NB classification process the outcome of the classification process is observed in this system as shown in Figure 3. It is observed in this system after the first NB classifier process the result of the classification process is as shown in the Figure 3.

Figure 3. NB classifier description result

Table 1. shows that there were 2 classes with 569 features classified using NB algorithm.

Table 1. Detailed Accuracy by Class for NB

NaiveBays Classifier Description Result		
Correctly Classified Instances	529	92.9701 %
Incorrectly Classified Instances	40	7.0299 %
Kappa statistic	0.8491	Success Rate
Mean absolute error	0.07	
Root mean squared error	0.2585	
Relative absolute error	14.9642 %	
Root relative squared error	53.4683 %	
Total Number of Instances	569	

The results of correctly classified features were 529 (92.9701%) while incorrectly classified were 40 (7.0299 %). Mean absolute error and Root mean squared error were 0.07% and 0.2585% for these 2 classes respectively.

3.2 J48

After the second j48 classifier cycle the outcome of the classification cycle is observed in this system as shown in Figure 4.

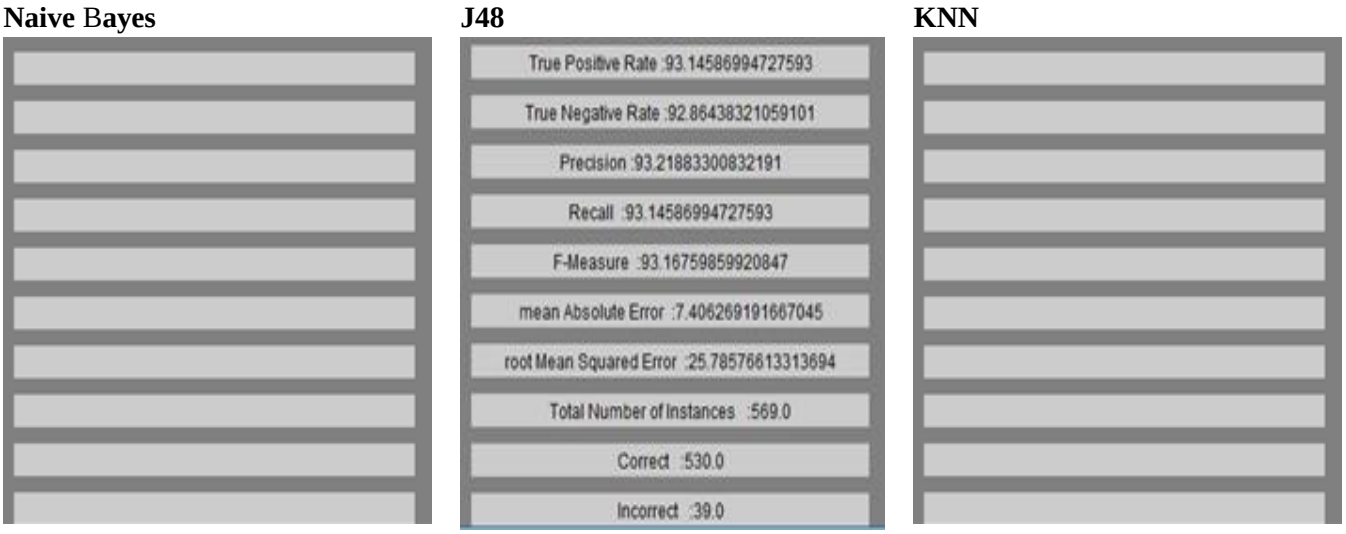

Figure 4. J48 classifier description result

Table 2. shows that there were 2 classes with 569 features after classified u sing J48 algorithm.

Table 2. Detailed Accuracy by Class for j48

J48 Classifier Description Result		
Correctly Classified Instances	530	93.1459 %
Incorrectly Classified Instances	39	6.8541 %
Kappa statistic	0.8544	Success Rate
Mean absolute error	0.0741	
Root mean squared error	0.2579	
Relative absolute error	15.8366 %	
Root relative squared error	53.331 %	
Total Number of Instances	569	

The results of correctly classified features were 530 (93.1459%) while incorrectly classified were 39 (6.8541%). Mean absolute error and Root mean squared error were 0.0741% and 0.2579% for these 2 classes respectively.

3.3 K-nearest neighbor (KNN)

It is observed in this system after the third KNN classifier process the result of the classification process is as shown in the Figure 5.

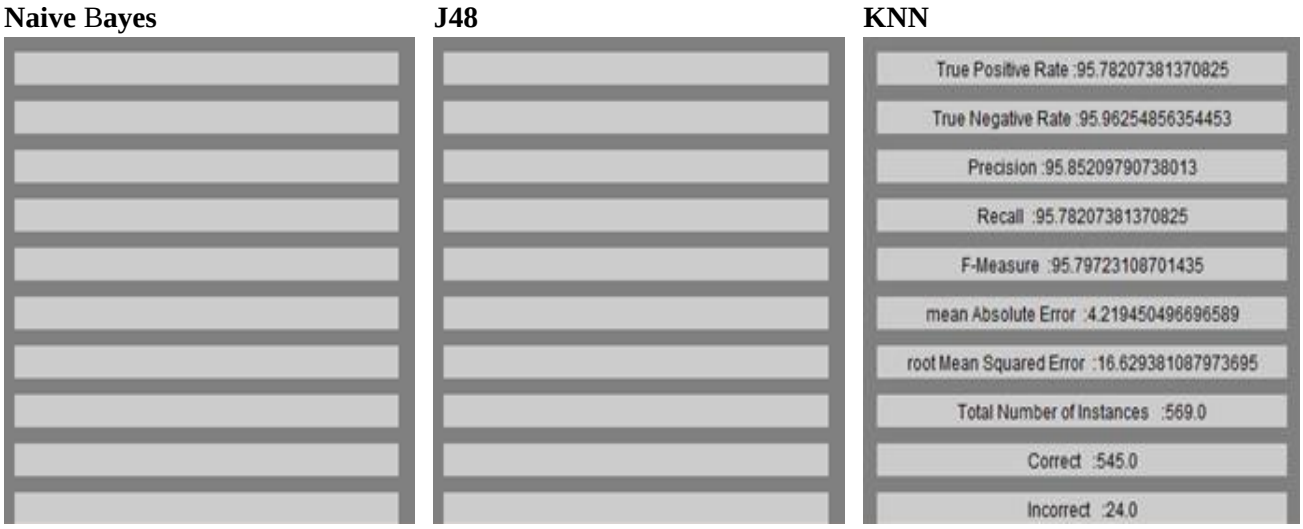

Figure 5. KNN classifier description result

Table 3 shows that there were 2 classes with 569 features after classified using KNN algorithm.

Table 3. Detailed accuracy by class for KNN

KNN Classifier Description Result		
Correctly Classified Instances	545	95.7821 %
Incorrectly Classified Instances	24	4.2179 %
Kappa statistic	0.9105	Success Rate
Mean absolute error	0.0422	
Root mean squared error	0.1663	
Relative absolute error	9.0223 %	
Root relative squared error	34.3934 %	
Total Number of Instances	569	

The results of correctly classified features were 545 (95.7821%) while incorrectly classified were 24 (4.2179%). Mean absolute error and Root mean squared error were 0.0422% and 0.1663% for these 2 classes respectively. The Figure 6 shows that comparison between precision of three machine learning algorithms of 2 classes.

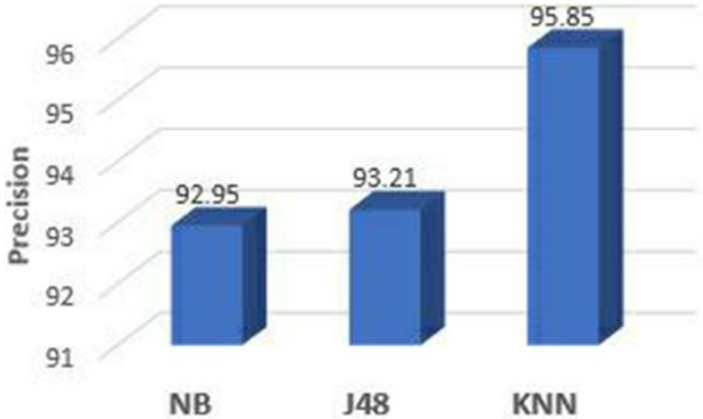

Figure 6. Precision of the three classifiers

3.4 Evaluation of the used classification algorithms in the proposed system

In the proposed method, three classification algorithms were used where the first algorithm was classified as KNN and the second was classified as Bayes navies and the third was classified as j48, the algorithms revealed that the KNN algorithm was much better than the naive Bayes and j48 algorithms. Table 4 displays that the result of KNN, naive Bayes and j48classified of accuracy. The three algorithms have calculated precision; correctly classified instances, incorrectly classified instance and root mean squared error

Table 4. Overall performance for the proposed BC system recognizer

Identification	Correctly Classified Instances	Incorrectly Classified Instances	Root mean squared error	Success Rate
NB	529	40	0.2585	92.9521%
J48	530	39	0.2579	93.2188%
KNN	545	24	0.1663	95.8520%

Experimental results show that KNN provides the highest precision with the lowest error rate as compared to classifiers that use such as NB, j48, Figure 7 show the details of the three classifiers.

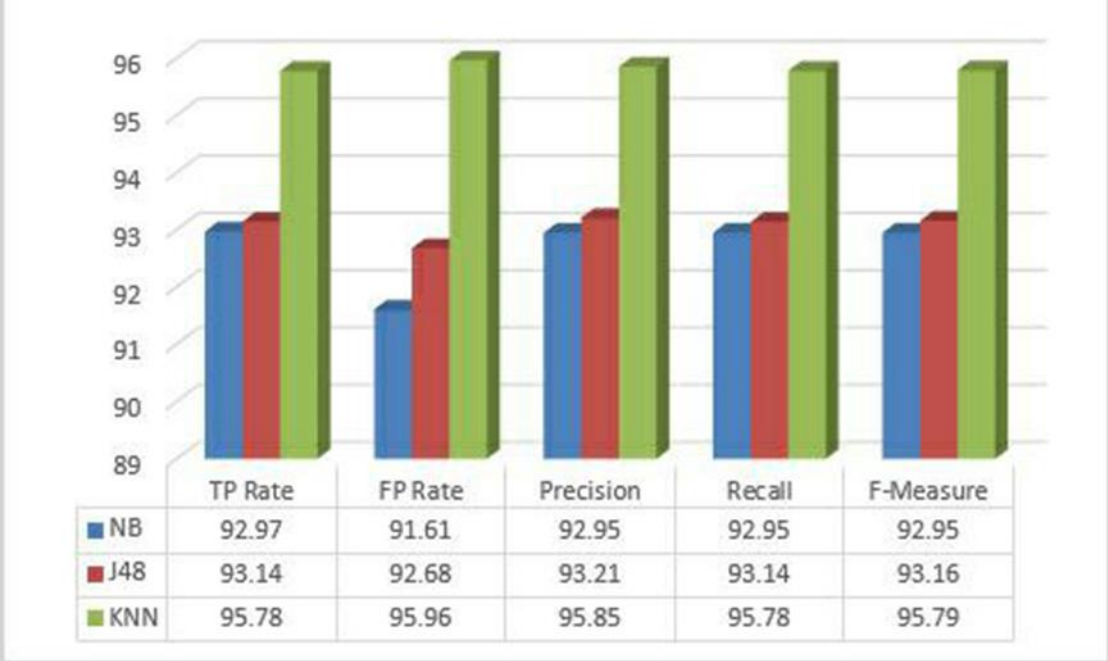

Figure 7. Details of the results of the three classifiers

4 Conclusion

In this article we discussed the use of three machine-learning methods for breast cancer diagnosis. The first KNN algorithm showed good results when dealing with imbalanced data (95,8520 per cent), but it is critical that the dataset should be pre-processed before running the algorithm because it does not deal with missing values and has better output when learning from a dataset with discrete nominal values.

The other algorithm, J48, resulted in a less accurate classifier, with a higher false-negative rate than the first one (93.2188%) Ultimately, the worst success rate for Naves Bayesian Networks was 92.9521% due to reduced filling.

With the KNN Networks, this paper obtained a slightly higher accuracy than those reported in the first papers using this dataset. Nevertheless, several sophisticated machine learning algorithms have been developed and recent papers have achieved accuracy levels of nearly 100% of the cases in this dataset.

References

[1] Wolberg, William H., W. Nick Street, and O. L. Mangasarian. "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates", Cancer letters 77.2-3, pp.163-171, 1994.

[2] Tan, Aik Choon, and David Gilbert. "Ensemble machine learning on gene expression data for cancer classification", 2003.

[3] Dumitru, Diana. "Prediction of recurrent events in breast cancer using the Naive Bayesian classification." Annals of the University of Craiova-Mathematics and Computer Science Series 36.2, pp.92-96. ,2009.

[4] Ishida, Takashi, Gang Niu, and Masashi Sugiyama. "Binary classification from positive-confidence data." Advances in Neural Information Processing Systems. 2018.

[5] Abdul-Sattar, Zubaida. Experimental analysis on effectiveness of confocal algorithm for radar-based breast cancer detection. Diss. Durham University, 2012.

[6] Eshlaghy, A.T. & Pourebrahimi, Alireza & Ebrahimi, Mansour & Razavi, A.R. & Ghasem Ahmad, Leila., "Using three machine learning techniques for predicting breast cancer recurrence", Journal of Health & Medical Informatics. vol.4, no. 2, pp.124-130, 2013.

[7] Wang, Dayong, et al. "Deep learning for identifying metastatic breast cancer." arXiv preprint arXiv:1606.05718, 2016.

[8] Asri, Hiba, et al. "Using machine learning algorithms for breast cancer risk prediction and diagnosis." Procedia Computer Science 83: 1064-1069, 2016.

[9] Al-Quraishi, Tahsien, et al. "Breast Cancer Risk Assessment Prediction Using an Ensemble Classifier." CAINE2017. 2017.

[10] Tingting Mu, Asoke K. Nandi,"Breast cancer diagnosis from fine-needle aspiration using supervised compact hyperspheres and establishment of confidence of malignancy", Lausanne, Switzerland, August 25-29, 2008.

[11] William H. Wolberg*", W. Nick Streetb, O.L. Mangasarianb, "Machine learning techniques to diagnose breast cancer fromimage-processed nuclear features of fine needle aspirates", USA, September 163-171,1994.

[12] R. Guzman-Cabrera, J.R. Guzaman-Supulveda, M. Torres-Cisneros, D.A. May-Arrioja, J. Ruiz-Pinales, O.G. Ibarra-Manzano, G. Avina Cervantes, A. Donzalez Parada, "Digital image processing technique for breast cancer detection", Int J Thermophys, 34, pp.1519-1531, 2013.

[13] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast cytology diagnosis via digital image analysis", Analytical and Quantitative Cytology and Histology, 15(6), pp.396–404,1993.

[14] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Machine learning techniques to diagnose breast cancer fromfine-needleaspirates", CancerLetter, 77, pp.163–171, 1994.

[15] Arnau Oliver, Joan Marti, Robert Marti, Anna Bosch, Jordi Freixenet, "A new Approach to the classification of mammographic masses and normal breast tissue", The 18th International Conference on Pattern Recognition (ICPR'06), pp.1-4, 2006.

[16] S. B. Kotsiantis, et. al., "Supervised machine learning: A review of classification techniques," vol. 160, pp. 3-24, 2007.