Vigilance towards the use of artificial intelligence applications for breast cancer screening and early diagnosis

Younes El Ouahabi[11*], My Hachem El yousfi Alaoui[2,3], Benayad Nsiri[2,3], Abdelmajid Soulaymani[1], Abdelrhani Mokhtari[1] Brahim Benaji[3],

[1]Laboratory Health and Biology, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

[2]Research Center STIS, M2CS, National Graduate School of Arts and Crafts of Rabat, Mohammed V University Rabat, Morocco

[3] Groupe of Biomedical Engineering and Pharmaceuticals Sciences - National Graduate School of Arts and Crafts (ENSAM)-Mohammed V University Rabat, Morocco

Abstract: Breast cancer is a real public health problem in Morocco. It is the cause of a significant number of deaths caused by late diagnosis. Mammography plays an essential role in the detection of breast cancer and in the early management of its treatment. Despite the existence of screening programs, there are still high rates of false positives and false negatives. Indeed, women were called back for additional diagnoses based on suspicious results that eventually led to cancer. Artificial intelligence (AI) algorithms represent a promising solution to improve the accuracy of digital mammography offering, on the one hand, the possibility of better cancer detection, and, on the other hand, improved efficiency for radiologists for good decision-making.

In this work, through a review of the literature on the tools used to evaluate the performance of AI systems dedicated to early detection and diagnosis of breast cancer. We set out to answer the following questions: Is the ethics relating to patient data during the development phase of this software is respected? Do these tools take into consideration the specificities of the field? What about the specification, accuracy and limitations of these applications?

At the end, we show through this work recommendations to adapt these evaluation tools of AI applications for breast cancer screening for an optimized and rational consideration of the principle of health vigilance and compliance with the regulatory standards in force governing this field.

Key words: Breast cancer, mammographic image, artificial intelligence, evaluation tools, Heath vigilance

I- Introduction

Breast cancer is the most frequently diagnosed tumor worldwide [1] and represents the first cancer affecting women in Morocco [2] (31.9 per 100,000). It originates in the cells that make up the breast. The cancerous (malignant) tumor is a group of cancerous cells that can invade and destroy nearby tissue. This tumor can also spread to other parts of the body (metastasis).

Breast cells sometimes undergo changes that make their growth pattern or behavior abnormal. These changes can lead to non-cancerous (benign) breast conditions such as atypical hyperplasia and cysts. They can also lead to intraductal papillomas that form in the breast ducts and are usually detectable near the nipple. This type of (benign) breast tumor is a mass that does not spread to the rest of the body (no metastasis) and is usually not life-threatening [3].

A malignant tumor is characterized by a set of events that together result in an abnormal and uncontrolled proliferation of cells.

In this regard, it is worth mentioning that the detection of cancer at an early stage allows for more frequent curative treatment. This helps to increase the chances of cure. Therefore, AI algorithms have become one of the main tools for medical image analysis. Machine learning techniques are solutions to develop tools to help doctors diagnose, predict the risk of diseases and prevent them before it becomes too late. Deep Learning is an emerging area of machine learning that encompasses a wide range of network architectures designed to perform multiple tasks [4]. Nevertheless, the urgency in which we find ourselves should in no way relegate to the background the notion of health vigilance, including the ethical issues and robustness requirements of the AI systems deployed. Those who develop these systems and those who authorize their use must ensure that they comply with certain principles such as respect for patient rights and privacy, security, transparency and fairness [5].

Construction, validation and performance evaluation of an AI model are among the most important phases when designing this model.

*** Correspondance:** My Hachem El yousfi Alaoui **: h.elyousfi@um5r.ac.ma /Younes El Ouahabi: y.elouahabi.gmail@gmail.com,**

In addition to the algorithm used, the size and quality of the data used are also of great importance to justify the results obtained. Furthermore, the splitting of the data and their use during the training and the validation of the models can have a significant impact on the obtained results.

The dataset is the foundation for such work. Its size, content and how it is used implies the quality of the results and the performance of the model.

It was shown in [6] that a model built from data relating to a given population does not give the same results when used with a different population. Also, the size of data set and the way of its exploitation influence the quality of the results [7, 8].

In the context of computer-aided diagnosis and detection of breast cancer, the most popular and successful algorithms are based on deep learning [9] and focus on convolutional neural networks (CNN) [10]

This paper is part of the performance evaluation of AI algorithms. To do so, we are interested in studying some of the most efficient algorithms from the State of the Art. The objective is to be able to classify breast cancer images into two classes (malignant or benign). For this purpose, two algorithms (CNN+ KNN and CNN+ SVM) based on Deep-Learning [11] have been used in order to simplify the task of radiologists who have to process thousands of images every day by offering them a second opinion.

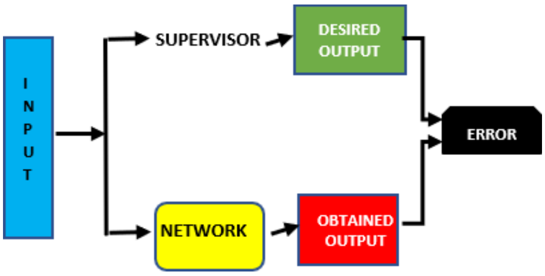


Figure 1: Supervised learning block diagram

The objective of the image classification process is to set up an intelligent system capable of assigning a class to each mammographic image. For this, the algorithm is trained using a learning database containing examples of real cases processed and validated. The objective is to find correlations between the input data (explanatory variables) and the output data (variables to be predicted) in order to facilitate the decision-making.

In supervised learning, we distinguish between classification and regression tasks;

- Classification: when the target variable (to be predicted) is discrete, Y = {1, ...I}. This amounts to assigning a class (or label) to each input. This is the case if we want to predict the trend of a future movement of an asset (high, neutral, low).

- Regression: when the target variable to be predicted is continuous, $Y \subset \mathbb{R}$. Example: predict the future dollar price of the asset in question.

In order to give a faithful and analytical account of this paper and after a general description of breast cancer and the usefulness of its early detection via AI tools, the rest of the work will be divided as follows: Section II is devoted to the materials and methods adopted for the classification of mammographic CNN, KNN, SVM images and their evaluation. In section III, the obtained results will be presented and analyzed in section IV before concluding.

II – Materials and methods

2.1 Convolutional neural networks (CNN)

CNNs are part of AI. They are the foundation of deep learning and the main basis for medical image classification applications [12].

The first part of a CNN is the convolutional part. It works as a feature extractor for images. An image is passed through a succession of filters, or convolution kernels, creating new images called convolution maps. Finally, the convolution maps are flattened and concatenated into a feature vector, called the CNN code. This CNN code at the output of the convolutional part is then connected to the input of a second part, consisting of fully connected layers (multilayer perceptron). The role of this part is to combine the features of the CNN code to classify the image. The output is a last layer with one neuron per category. The numerical values obtained are usually normalized between 0 and 1, with a sum of 1, to produce a probability distributionover the categories.

Basic elements of CNN architecture [13]:

A convolutional neural network architecture is formed by a stack of processing layers:

- The Convolution layer (CONV).
- The Pooling layer (POOL).
- The Correction layer (ReLU).
- The Fully Connected Layer (FC).
- The Loss layer (LOSS).

2.2 Support Vector Machines (SVM)

The Support Vector Machine (SVM) [14] method is a family of machine learning algorithms that can be used to solve classification, regression and anomaly detection problems. SVMs are known for their strong theoretical guarantees, their great flexibility as well as their simplicity of use. Depending on the data, the performance of SVMs is of the same order, or even superior, to that of a neural network or a Gaussian mixture model. These classifiers are based on two key ideas, allowing to deal with non-linear discrimination problems, and to reformulate the classification problem as a quadratic optimization problem. The first key idea is the notion of a maximum margin. The margin is the distance between the separation boundary and the nearest samples. The latter are called support vectors. In SVMs, the separation frontier is chosen as the one that maximizes the margin. The second key idea of SVMs is to transform the representation space of the input data into a higher dimensional space, in which linear separation is likely to exist. Kernel functions are used to transform a scalar product into a high-dimensional space. This technique is known as the kernel trick.

The steps of SVM are:

- Collection of a database;
- Extraction of characteristics;
- Construction of a training database of inputs from these characteristics;
- Classify the inputs;
- We put these outputs in a vector that corresponds to the outputs;
- Construction of the model by the SVM TRAIN command;
- Make a test for signals/images by the SVMCLASSIFY command.

2.3 The k-nearest neighbor method (KNN)

KNN [15] is a pattern recognition algorithm that can be used for both classification and regression. It is one of the nonparametric techniques frequently used in nonlinear financial prediction. This preference is mainly due to two reasons:

First, the algorithmic simplicity of the method compared to other global methods such as neural networks or genetic algorithms.

Second, the KNN method has empirically demonstrated a significant predictive ability.

The idea of the method is to predict the future of a time series by analyzing how it has evolved in a similar situation in the past. Thus, to make a prediction we take the most recent historical data available and we search among these data, the K closest instances also called the K closest vectors.

Deep convolution neural networks are multi-layer architectures designed to extract high-level representations of a given input. They have significantly improved the state of the art in image, video, speech and audio recognition tasks. When trained for supervised classification, CNN layers are ultimately capable of extracting a set of features tailored to the task at hand.

We present a simple and effective technique to account for label noise on deep neural networks. Large amounts of data (required for deep neural networks) usually contain erroneous labels, and the presence of such noise can significantly degrade learning performance.

Feature extraction followed by the application of k-Nearest-Neighbors (kNN) is a deep hybrid learning technique. Some works have also used this strategy, replacing the softmax layer by KNN or by an SVM. They proved that using a CNN plus a KNN [19] or an SVM improves the classification accuracy compared to using the softmax CNN output directly.

The proposed architecture consists of replacing the softmax with a kNN classifier. The CNN and kNN are fully complementary in terms of feature extraction and decision boundaries, in a hybrid system, both algorithms could exploit their potential, and have their drawbacks mitigated by the other. Moreover, CNNs are usually trained with large datasets, and as the training set size approaches.

2.4 The Hybrid Model, CNN+KNN

Our hypothesis is that a hybrid CNN + kNN can outperform in predictive and classification power than a CNN + softmax. To evaluate this, we compare the results of a trained CNN for a classification task with the softmax appris layer of the same CNN, but applying a kNN classifier to the last hidden layer output

a) Direct use of kNN on raw data without any representation learning
b) Using the softmax layer of CNN.
c) Replacing softmax with a kNN classifier.

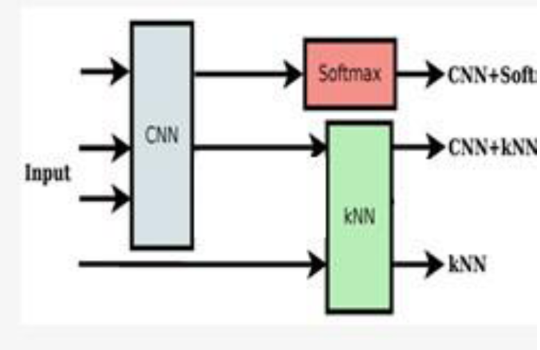


Fig 2: Principle of hybrid CNN+KNN [20]

Certain indulgence in "automatic breast cancer detection algorithms" has raised concerns among researchers, who question AI models for several reasons. On the one hand, they are poorly documented and trained on small or low-quality data sets with a high risk of error. The ultimate danger would be that the premature use of AI technologies would increase diagnostic errors and compromise the quality of care.

Overall, no one can deny the importance of these tools, but only if they are used in close collaboration with medical staff to better define a framework, ensure that the right questions are answered, and provide real added value.

2.5 Evaluation metrics of a classification algorithm

Performances of a binary classification algorithm are calculated from the results obtained by this procedure:

- Deduct the number of samples belonging to class A (e.g., image containing a benign tumor) correctly recognized.
- Deduct the number of samples belonging to class B (for example, image containing a malignant tumor) correctly recognized (true positives and true negatives).
- Then, the number of samples that were incorrectly assigned to the first or second class (false positive, false negative).

These results are used to calculate the "True positive rate" TPR, "False positive rate" FPR and the confusion matrix presented in table 1 are deduced.

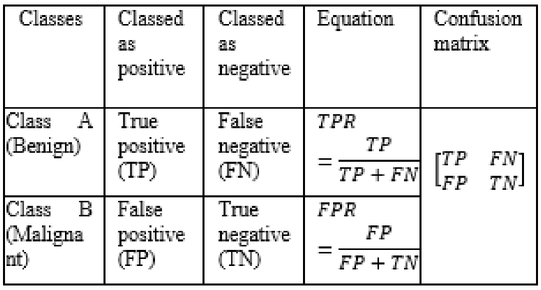
From these elements, we compute the main metrics to evaluate our algorithm. The evaluation focus and the formula for calculating each of the most commonly used metrics are given in Table 2.

To build and validate a CAD-based breast cancer diagnosis/detection algorithm, we need a data set of images processed and classified by specialists. These data are split into training set that is used to build the model and a validation set that allows to evaluate the accuracy of the model [7].

The size of the data and the ratio between training and validation data as well as the validation method used are determining factors for the quality of the metrics obtained.

the size of the data is the deciding factor for the qualities of the generalization performance estimated from the validation set [16, 17].

Table 1: Confusion matrix construction



Classes	Classed as positive	Classed as negative	Equation	Confusion matrix
Class A (Benign)	True positive (TP)	False negative (FN)	$TPR = \frac{TP}{TP + FN}$	$\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$
Class B (Malignant)	False positive (FP)	True negative (TN)	$FPR = \frac{FP}{FP + TN}$	

Table 2: Evaluation focus and formulas of the most commonly used metrics

Measure	Formulas	Evaluation focus
Accuracy:	$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$	Global effectiveness of a classifier
Precision	$Precision = \frac{TP}{TP + FP}$	Classifier effectiveness with respect to the samples of the positive class
Recall (Sensitivity)	$Recall(Sensivity) = \frac{TP}{TP+FN}$	Effectiveness of a classifier to identify positive samples
Fscore	$Fscore = \frac{(\beta^2 + 1)TP}{(\beta^2 + 1)TP + \beta^2 FP + FP}$	Relations between data's positive samples and those given by a classifier
Specificity	$Specificity = \frac{TN}{FP+TN}$	The efficiency with which a classifier identifies negative labels
AUC	$AUC = \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+Fp})$	Ability of the classifier to avoid misclassification

III- Results and discussion

Here we present a supervised algorithm for mammographic images classification. We propose a hybrid approach to improve the performance of our algorithm and adapt it to the images of the database used.

In the Hybrid CNN +KNN (K-Nearest Neighbors) classifier proposed, we exploited the features extracted by our CNN model and used these features as inputs for a KNN classifier. The principle of this hybrid model is presented in Fig. 3.

To build our model, we used the Digital Database for Screening Mammography. It consists more than 2500 case studies with nearly 10,000 images including two images of each breast. Some of these images are normal, others contain benign masses and other malignant ones [20]. Each case is accompanied by a detailed description provided by specialists. It is a publicly available dataset that complies with the Data Protection Regulation (GDPR) [21]. The data contained in this database is no longer personal to identified individuals. It has been de-identified or anonymized. Therefore, the medical code of ethics is respected [22].

https://doi.org/10.1051/e3sconf/202131901099

To implement, train and validate the proposed algorithm, we used Python (with all the necessary libraries) and Google colaboratory.

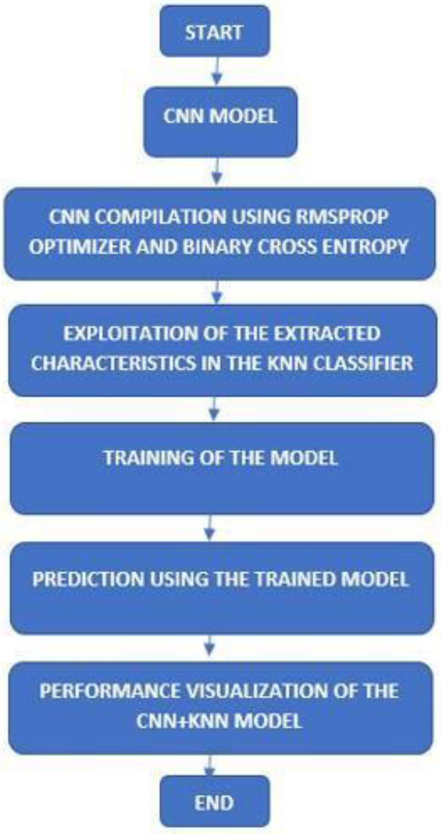

START
CNN MODEL
CNN COMPILATION USING RMSPROP OPTIMIZER AND BINARY CROSS ENTROPY
EXPLOITATION OF THE EXTRACTED CHARACTERISTICS IN THE KNN CLASSIFIER
TRAINING OF THE MODEL
PREDICTION USING THE TRAINED MODEL
PERFORMANCE VISUALIZATION OF THE CNN+KNN MODEL
END

Fig 3: CNN + KNN implementation

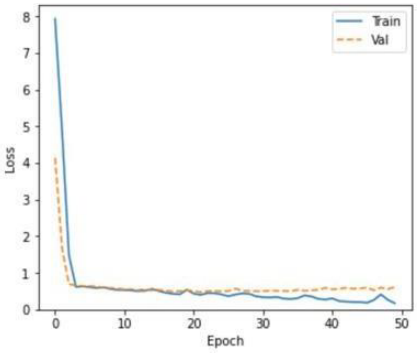


Fig. 4: CNN+KNN loss curve.

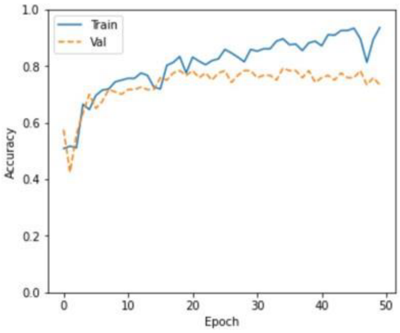
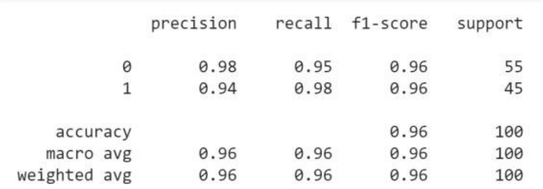


Fig 5: CNN+KNN accuracy curve

Table 3: Evaluation metrics of proposed algorithm



```
              precision    recall  f1-score   support

           0       0.98      0.95      0.96        55
           1       0.94      0.98      0.96        45

    accuracy                           0.96       100
   macro avg       0.96      0.96      0.96       100
weighted avg       0.96      0.96      0.96       100
```

As shown in figures 4 and 5 and table 3, the results of the model are very satisfactory. The confusion matrix for the CNN-only model has 24 false positives and 8 false negatives, whereas the confusion matrix for the combined CNN+KNN model has 12 FPs and 4 FNs. the precision obtained is 98%. In the medical context,unlike in the scientific research context, a goodclassification of samples in the 'Positive' class is much more important than that of the 'Negative' class. Indeed, an error in the classification of samples in the 1st class means a risk of ignoring patients at risk, whereas this is not the case for the classification of samples in the 2nd class. It is therefore important to emphasis the numbers of true positives and false positives in relation to the total numbers of samples in the 'Positive' class. Thismeans that the accuracy of the algorithm should be prioritized over other metrics for evaluating a breast cancer detection/diagnosis algorithm or DAC. The graph in Fig. 6 shows the main indicators of different CNN Architectures using the same data set.

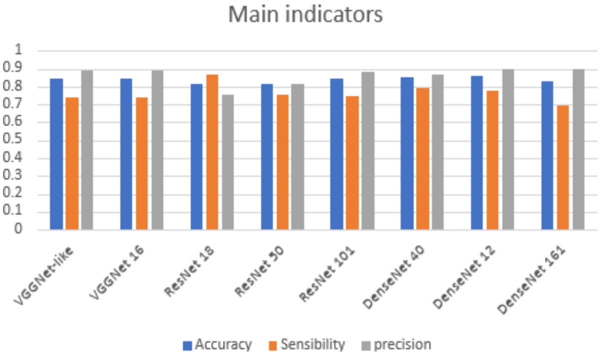


Figure 6: Quantitative indicators used to evaluate different CNN Architectures' performances on the database

For all these models including the one we have proposed, although the results are satisfactory, it should be noted that the training and validation data used to come from the same database. The model test should be done on real-world data and its performance should be verified.

A quick review of the state of the art in this context [9, 13, 24] allows us to conclude that some researchers are only interested in the standard evaluation metrics of the algorithms they propose. They often avoid giving more information about the database used, the evaluation method and the associated risks of error. In terms of ethics and compliance with regulatory standards for the protection of patients' rights, any AI algorithm applied to breast cancer detection/diagnosis must present satisfactory answers to the following questions:

- Does the used dataset comply with the regulatory standards in this area?

- Are the used data appropriate to the target population [23]? And can we use the algorithm in question on any data of the same nature?

- Does the used dataset size used to justify the results reported? And what is the associated error rate?

- What is the error rate associated with the sample segmentation method and the choice of rates used for training and validation?

- What is the cumulative number of positive cases that are likely to be undetected by the algorithm?

IV- Conclusion

Through this work, we presented and implemented an AI algorithm for breast cancer mammography image classification, followed by the parameters of the main hybrid CNN+KNN model with their evaluation metrics. These metrics are very satisfactory (precision = 98%.). Then, we showed that these metrics are not based on real data.

Based on this study and a review of the main algorithms in the literature, we presented the main points of vigilance to adopt in this context.

It should be noted that in addition to the respect of the code of medical ethics relating to the exploitation of patient data, the AI tools must provide real metrics with the margin of error related to the amount of data used and the method of validation. In addition, the precision remains the most important parameter to evaluate such a CAD diagnosis /detection of breast cancer.

References

1. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.

2. Organisation mondiale de la Santé - profils des pays pour le cancer, 2014, Maroc.

3. AL-HAJJ, Muhammad, WICHA, Max S., BENITO-HERNANDEZ, Adalberto, *et al.* Prospective identification of tumorigenic breast cancer cells. *Proceedings of the National Academy of Sciences*, 2003, vol. 100, no 7, p. 3983-3988.

4. ROBERTSON, Stephanie, AZIZPOUR, Hossein, SMITH, Kevin, *et al.* Digital image analysis in breast pathology—from image processing techniques to artificial intelligence. *Translational Research*, 2018, vol. 194, p. 19-35.

5. DIEBOLT, Vincent, AZANCOT, Isaac, BOISSEL, François-Henri, *et al.* «Intelligence artificielle»: quels services, quelles applications, quels résultats et quelle valorisation aujourd'hui en recherche clinique? Quel impact sur la qualité des soins? Quelles recommandations?. *Therapies*, 2019, vol. 74, no 1, p. 141-154.

6. NGO, Vivian, DEGAN, Mona, HO, Eugene, *et al.* Evaluation of the united states preventative services task force screening guidelines for breast cancer in a Hispanic underserved population. *Cureus*, 2020, vol. 12, no 5.

7. VABALAS, Andrius, GOWEN, Emma, POLIAKOFF, Ellen, *et al.* Machine learning algorithm validation with a limited sample size. *PloS one*, 2019, vol. 14, no 11, p. e0224365.

8. MAZUROWSKI, Maciej A., HABAS, Piotr A., ZURADA, Jacek M., *et al.* Training neural network classifiers for medical decision-making: The effects of imbalanced datasets on classification performance. *Neural networks*, 2008, vol. 21, no 2-3, p. 427-436.

9. JIMÉNEZ-GAONA, Yuliana, RODRÍGUEZ-ÁLVAREZ, María José, et LAKSHMINARAYANAN, Vasudevan. Deep-Learning-Based Computer-Aided Systems for Breast Cancer Imaging: A Critical Review. *Applied Sciences*, 2020, vol. 10, no 22, p. 8298.

10. SWIDERSKI, Bartosz, KUREK, Jaroslaw, OSOWSKI, Stanislaw, *et al.* Deep learning and non-negative matrix factorization in recognition of mammograms. In: *Eighth International Conference on Graphic and Image Processing (ICGIP 2016)*. International Society for Optics and Photonics, 2017. p. 102250B.

11. DEBELEE, Taye Girma, SCHWENKER, Friedhelm, IBENTHAL, Achim, *et al.* Survey of deep learning in breast cancer image analysis. *Evolving Systems*, 2020, vol. 11, no 1, p. 143-163.

12. RAWAT, Waseem et WANG, Zenghui. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 2017, vol. 29, no 9, p. 2352-2449.

13. ALKHALEEFAH, Mohammad et WU, Chao-Cheng. A hybrid CNN and RBF-based SVM approach for breast cancer classification in mammograms. In : *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2018. p. 894-899.

14. STATNIKOV, Alexander. *A gentle introduction to support vector machines in biomedicine: Theory and methods*. world scientific, 2011.

15. PETERSON, Leif E. K-nearest neighbor. *Scholarpedia*, 2009, vol. 4, no 2, p. 1883.

16. GALLEGO, Antonio-Javier, CALVO-ZARAGOZA, Jorge, et RICO-JUAN, Juan Ramón. Insights into efficient k-nearest neighbor classification with convolutional neural codes. *IEEE Access*, 2020, vol. 8, p. 99312-99326

17. GALLEGO, Antonio-Javier, PERTUSA, Antonio, et CALVO-ZARAGOZA, Jorge. Improvingconvolutional neural networks' accuracy in noisy environments using k-nearest neighbors. *Applied Sciences*, 2018, vol. 8, no 11, p. 2086

18. XU, Yun et GOODACRE, Royston. On splittingtraining and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, 2018, vol. 2, no3, p. 249-262.

19. SOKOLOVA, Marina et LAPALME, Guy. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 2009, vol. 45, no 4, p. 427-437

21. OLIVEIRA, Júlia EE, GUELD, Mark O., ARAÚJO, Arnaldo de A., *et al.* Toward a standard reference database for computer-aided mammography. In : *Medical imaging 2008: Computer-aided diagnosis*. International Society for Optics and Photonics, 2008. p. 69151Y.

22. RUMBOLD, John Mark Michael et PIERSCIONEK, Barbara. The effect of the general data protection regulation on medical research. *Journal of medical Internet research*, 2017, vol. 19, no 2, p. e47.

23. NGO, Vivian, DEGAN, Mona, HO, Eugene, *et al.* Evaluation of the united states preventative services task force screening guidelines for breast cancer in a Hispanic underserved population. *Cureus*, 2020, vol. 12, no 5.

24. RAMADAN, Saleem Z. Methods used in computer-aided diagnosis for breast cancer detection using mammograms: a review. *Journal of healthcare engineering*, 2020, vol. 2020.