Artificial intelligence applied to support the breast cancer diagnosis

Woldson Leonne Pereira Gomes
Federal University of Pará - UFPA
Control and Systems Laboratory - LACOS
Belém, Brazil
woldson.gomes@itec.ufpa.br

Antonio da Silva Silveira
Federal University of Pará - UFPA
Control and Systems Laboratory - LACOS
Belém, Brazil
asilveira@ufpa.br

***Abstract*— Breast cancer is a more common neoplasm among women (not considering non-melanoma skin cancer). The estimate for the coming years is still growing and poses a threat to human health. Currently, the methods used in the diagnosis of breast cancer are performed through analysis of mammography images. Allowed, an analysis made by two specialists, which are subject to errors due to factors such as fatigue and lack of capacity. Not only the factor of human errors in diagnoses, certainly the long periods of time until the final diagnosis is another factor to be taken into account, because cancer is a progressive disease over time. In this sense, the present work applied a solution through the automatic classification of mammography images, in order to determine as normal or cancer. In addition, for simulations, two machine learning techniques were added independently, as they can eventually serve as a support in the diagnosis of breast cancer, that is, a CAD system, which means "computer-aided diagnosis". As machine learning techniques applied for classification referenced as convolutional neural networks and support vector machines. Subsequently, the construction of the classification algorithms, they were subjected to the testing phase, which was found to be more than 85% accurate in the classification of mammography images.**

Keywords—cancer prediction, mammographic CAD, machine learning, convolutional neural networks, support vector machine.

I. INTRODUCTION

Currently, cancer is one of the most complex public health problems facing the Brazilian health system, given its epidemiological, social and economic magnitude [1]. Among the most common types of cancer, breast cancer stands out, which develops in the breasts. According to [2], all cancer is characterized by a fast and disordered growth of cells.

According to [3], for the general population, the lifetime risk of having breast cancer in Brazil is around 8%, that is, one in every twelve women will develop breast cancer throughout their lives. For Brazil, the estimate for each year of the 2020-2022 triennium indicates that there will be 625,000 new cases of cancer, the most common cancer among women will be breast cancer, except for non-melanoma skin cancer [4].

According to [5], the main method used for diagnosing breast cancer occurs through the analysis of mammography images. Screening by mammography is the safest way to detect the presence of breast lesions, especially at an early stage.

The mammography exam uses a device called a mammograph, which is an X-ray device with a more restricted field and greater sensitivity for detecting smaller structures [6].

However, according to [7], the time to diagnosis or even a misdiagnosis are crucial factors for the patient's longevity. The delay between diagnosis and initiation of treatment aggravates breast cancer, making it progressive and irreversible.

Furthermore, according to [8], it is indicated that between 10 and 15% of mammograms are wrongly evaluated, which may result in unnecessary biopsies, or even delay the treatment of cancer.

In the medical field, image classification is used to aid in the diagnosis of various types of disease, including breast cancer [9]. Thus, the machine becomes an ally in cancer diagnosis, optimizing time and not subject to human errors such as fatigue or lack of well-trained professionals. These types of systems are called CAD, computer-aided diagnosis.

Aiming to reduce the rate of errors in the identification of breast cancer, a way that radiology clinics and hospitals use, is the double reading. That is, the exam is viewed and diagnosed by two specialists. It is proven that double reading increases the detection of breast cancer by up to 8.5% [10].

However, not all health care facilities have two specialists to assess the mammography image. In addition, the option for dual diagnosis by specialists takes time, which can result in a delay in the delivery of test results.

Thus, in the search for a solution in the identification of breast cancer through mammography images, this work aims to experiment with two types of machine learning methods, in order to classify mammography images from two databases between two classes, normal and cancer, and a posteriori to evaluate the performance of both classifiers. The proposed classifiers use the Convolutional Neural Networks (CNN) and Support Vector Machine (SVM) methods.

II. METHODS AND PROCEDURES

For the construction of classification algorithms, the programming language Python was used, due to the great potential of the language with the use of machine learning. The first classifier proposed is a convolutional artificial neural network and the second a support vector machine. The fundamental idea is to keep the same training and testing data to be submitted in both classifiers, that is, both the SVM and CNN methods are trained and tested with the same data set. Thus, the parameters for the final evaluation of the models can be compared using statistical methods without discrepancies in relation to the database, as well as an analysis of the similarity of errors between the classifiers.

A. Database

First, it was necessary to use an extensive database that contained mammography images and their associated medical

report. The medical report of interest to the project consists of defining which images are characterized as breast cancer and which are classified as normal.

In other words, the sample collection phase consisted of obtaining two types of images, thus forming two groups. In the project, it was agreed to call each image diagnosis as a label, and the images as samples. Thus, each sample has a label attached, which defines which class the image belongs to.

When using a database for the development of a diagnostic support system, it must be safe, varied, and have a large amount of samples. Several tests reported in the state of the art prove that the performance of a CAD system is heavily influenced by the quality and quantity of images used.

For the elaboration of the project, datasets from two different databases were partially used, namely the MIAS [11] and the DDSM [12]. The fundamental choice for using the aforementioned databases is due to the nature of the public domain and the wide variety of samples with labels. The sample sets have dimensions averaging 1024x1024 and varying storage sizes averaging 500KB.

Altogether, the dataset consisted of 519 mammography images, divided into two classes. With 459 images for training and 60 images for tests. The choice of division was based on the book by [13], which describes that, for a small dataset, it is feasible to use 90% of the data for training, and 10% for testing. Therefore, for the present work approximately 88% of the data were used for training and 12% for tests.

The test images were divided into two parts, one part for an extended test, consisting of 50 images, 25 cancer class and another 25 normal class images; the second part was intended for a reduced test, with 10 images in total, 5 images from the cancer class and another 5 from the normal class. The idea of the reduced test is to visually insert the classification result and the real label linked to the display of the respective image. For training, 241 images from the cancer class were used, and another 218 images from the normal class, as best illustrated in Fig. 1.

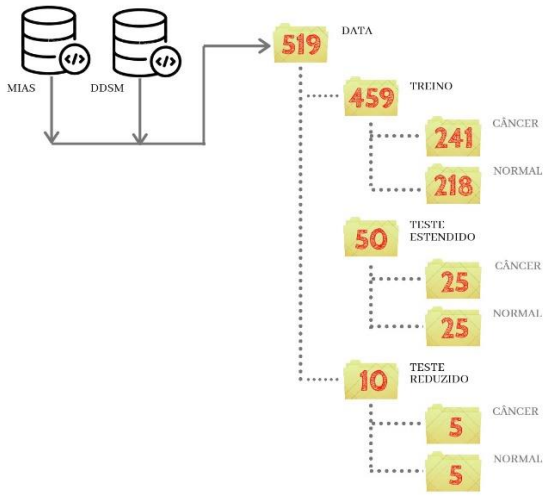

Fig. 1. Division of the dataset into training and testing.

B. Classifier with convolutional neural networks

Convolutional neural networks (CNN) belong to a category of algorithms based on artificial neural networks. The great advantage of using CNN is its great efficiency in recognizing images and videos, in addition to other types of classifications. Thus, with the labels and images composed in the dataset, the development of a mammography image classifier is proposed.

For the construction of CNN, two reputable libraries in the area of machine learning were used, the libs Keras and TensorFlow. Keras is a high-level TensorFlow API, that is, it serves as a bridge to connect applications in a way that makes using TensorFlow easier and more intuitive [14]. On the other hand, TensorFlow is one of the most renowned libraries nowadays, being used in the most varied machine learning applications, and it is an open source platform with the most modern machine learning currently available [15].

The initial step in building the CNN is to read the dataset. Thus, as arguments used, it must have the path that the database is in, that is, the physical directory in which the images of the MIAS and DDSM databases are saved.

Reading all data is only possible if all files are in the same extension configured by the algorithm, in this case, the extension “jpg”. For other sets of images with different extensions, import is possible after including the extension for reading in the algorithm.

After reading the dataset from the disk, the transformations to be applied to the images are defined, as well as the percentage of data for training and validation. The initial transformation applied to the image consists of a normalization of the data input. Furthermore, the number of images read at a time consists of 150, that is, the algorithm is capable of reading and processing 150 images at once during CNN training.

After data entry, it is necessary to build the CNN architecture. The parameters were defined empirically using the sequential method and based on existing models for image applications, present in the TensorFlow documentation and in the works of [16] and [17]. Thus, the architecture is arranged with two convolutional layers of two dimensions (2D), which have 64 filters and another with 128, each filter, or convolution kernel, has a 2x2 dimension that runs through the entire image, and an activation function of type ReLU.

The convolutional layer is responsible for extracting resources from an image, contributing with the help of blurring, sharpening, edge detection, noise reduction and other operations that help in the machine learning process. In the convolution process, the dot product of the pixel values in the current filter window with the weights defined in the filter is calculated.

Furthermore, the layer input must be of the three-dimensional (3D) type, in RGB format, that is, with three color channels. Despite 3D input, the convolutional layer is considered 2D because the kernel traverses the image in two dimensions, running three times, once for each color channel.

The architecture has a Max Pooling layer of size 2x2, which will perform a downsampling of the samples by a factor of 2, that is, it will reduce the dimensions of the dataset and will help to avoid overfitting. It is noteworthy that the reduction of image dimensionality does not necessarily promote the loss of resources or important standards used for image characterization.

As it is known, overfitting is an unwanted phenomenon in the construction of the model, that is why the proposed CNN has three dropouts, responsible for deactivating a percentage

of neurons. The defined percentage to inactivate the neurons was 30% for the first two and 50% for the last dropout.

Neurons ignored in the dropout are randomly chosen, and it is necessary to avoid over-adjustment, as a layer occupies many parameters, thus, neurons develop codependent with each other during training, which restricts the individual power of each neuron, leading to overfitting.

After the convolutional layer it is necessary to transform the output from 2 dimensions to one dimension, that is, an array. The need for transformation is due to the standardization of data entry in the fully connected layer. Thus, the architecture has a flatten operation, which is responsible for the desired spatial transformation.

Finally, the last structures of the CNN are two dense layers, which make up the fully connected layer, with a total of 256 neurons and a ReLU activation function; and the output layer, with only two neurons, which will define the classification between cancer and normal classes, with a sigmoid activation function, which establishes an output prediction probability with values between 0 and 1.

In order to facilitate the understanding of how the CNN model is structured, Figure 2 illustrates the connections and description of the layers. It is noteworthy that in the classification algorithm developed, there is a linked function that briefly shows how the convolutional neural network model is structured, detailing each layer as well as in Fig. 2.

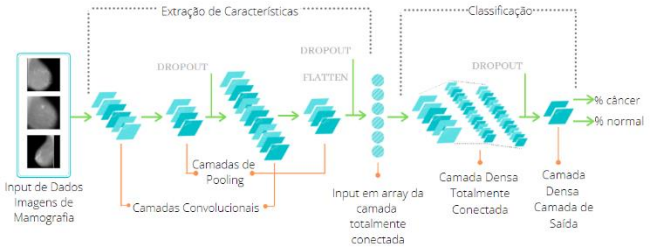

Fig. 2. Proposed architecture for the convolutional artificial neural network.

Other parameters defined in the CNN classification algorithm refer to compilation, as it is necessary to define how the CNN will learn. In this sense, there are three vital compilation parameters: the loss function, that is, the loss function, which was defined as binary cross entropy; the optimizer, which used the TensorFlow pattern, the Adam optimizer; and accuracy, which is a simple metric for evaluating the model during training [17].

Binary cross entropy is a metric used to measure the performance of the binary classification model, ie cancer or normal. The output of this variable is a probability value, so it varies between 0 and 1. The loss of cross entropy increases as the probability of the prediction deviates from the true label, and tends to zero as the network improves computing the desired output for the training inputs. In other words, the crossed entropy will be responsible for describing the loss between the probability distributions of the cancer and normal classes.

The Adam optimizer is an improved version of the descending gradient, and incorporates an adaptive learning rate and momentum [17]. The Adam optimizer was chosen because it is an efficient first-order algorithm for gradient-based optimization of stochastic functions, based on an adapted estimate of low-order moments.

The accuracy parameter represents a metric of model performance, comparing the true labels to those ranked by the algorithm. The model is fine if it converges on one. The reciprocal is true, that is, close to zero means that the model did not perform well in the prediction.

After the training phase, it is necessary to test the finished classifier. For this, the training algorithm has a specific function to save the trained model to disk, in addition to being configured to save the model with the lowest loss, that is, the one with the best performance, avoiding overfitting in the classifier. So, after saving it to disk, a new algorithm responsible for the tests was designed. It is important to note that the machine needs enough space to save the model, as it occupies a space of 1.45 GB, in .h5 format.

In CNN's test algorithm, the model is loaded from disk and subjected to two types of tests, with images never seen by CNN. The reduced test, with a total of 10 samples, and an extended test with 50 samples. The reduced and extended tests were used to test both the CNN model and the SVM, and serve as a basis for evaluating the model, in the generalizability of the classifier.

The need for a reduced test arose from the importance of bringing the operator closer to reality, not only showing graphs to evaluate the model, thus, the algorithms display the 10 images of the reduced test, together with the true labels and the labels of the classification performed, visualizing more intuitively how the classifier works in practice.

C. Classifier with support vector machine

The second classification method uses support vector machine (SVM) techniques. The SMV technique was chosen because it is recommended algorithms for non-linear classification, which is characteristic of the data in the data set. The great advantage of the algorithm lies in its robustness, precision and efficiency even using a small set of samples for training; moreover, the SVM is essentially a binary classifier, which fits perfectly into the type of classification proposed in the work.

To construct the SVM classification algorithm, the Scikit-Learn library was used, which has several integrated modules, including the classification via SVM [18].

The initial steps, as well as in the construction of the CNN, become similar in the use of the SVM. A priori, it was necessary to load the images from the dataset present in physical folders on the disk to memory, that is, store them in a variable. After loading, it is important to perform a count of the number of samples, to know if all the images contained in the database were successfully imported.

Based on the study of SVM, it is possible to infer the use of an RBF kernel, which is capable of handling non-linear data. SVM does not work with probability directly, they are calculated using cross validation, it is the SVC method of the reference library. SVC is the reference name for the class that implements the support vector machine classification [19].

The constructor of the estimator takes model parameters as arguments. Thus, the problem data and labels must be passed as arguments to the functions, and the classifier parameters must be defined as the γ and the C parameter.

Parameter γ is used for hyperplane with RBF kernel, the bigger it is, the more it is influenced by the support vectors, that is, it tries to fit exactly to the training dataset. Therefore, the gamma parameter defines how far the influence of a single training example reaches, with low values meaning “far” and high values meaning “near”.

The “C” parameter refers to an error term penalty, responsible for controlling the compromise between the soft decision limit and the correct classification of training points. The higher the value of C, the more likely the classifier is to be overfitting. Thus, the C parameter trades the correct classification of training examples against maximizing the decision function margin.

The parameters of C and γ were tested with the help of algorithms capable of estimating the best values for them. The value of γ was defined as the inverse of the number of features, that is, sampling characteristics. And the value of C was tested in a loop of repetition until saving the value that presents the best accuracy for the model.

III. Results and Discussion

A. *Database Analysis*

Before training the classification algorithms, it is necessary to view the images contained in the database, to identify whether they were correctly loaded, as shown in Fig. 3.

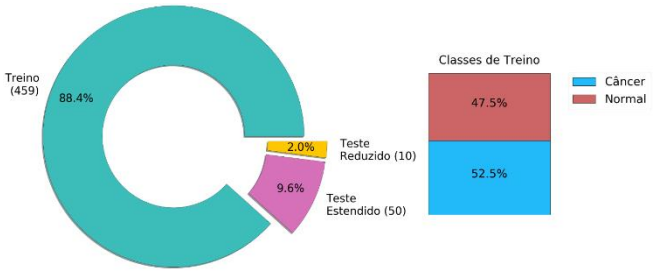

Fig. 3. Data distribution after loading the mammography image samples contained in the folder.

B. *Classifier with convolutional neural networks*

After creating the CNN model, according to the methodology applied for each layer, with a model visualization function, it is possible to verify the number of parameters and layers present.

Thus, it can be confirmed that the model proposed in the methodology was built accordingly, thus, the model can be trained with the dataset already imported. After viewing the built model, the next step consists of training, which was configured for 50 epochs.

In Figure 4, the learning curves are displayed, on the right the loss function graph is displayed, and on the left the accuracy graph after each epoch, both for validation and training. For validation, an average accuracy of 81% was obtained.

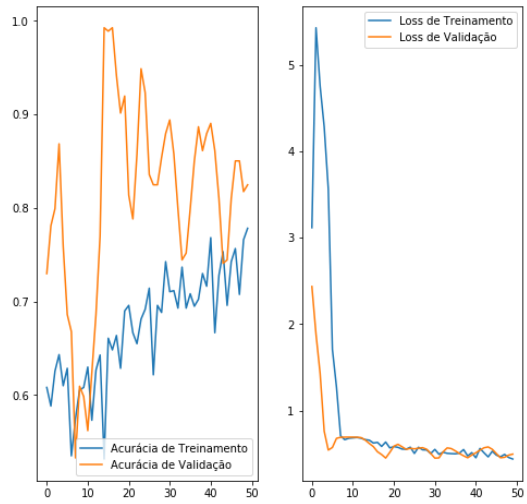

Fig. 4. Accuracy and loss of training with CNN.

It is observed that the learning curves behave as expected, that is, the loss function decreasing at each epoch, and the accuracy of the model increasing. An important detail lies in the fact that validation accuracy has high values compared to training accuracy, and this strongly demonstrates that CNN is able to generalize, potentially able to correctly classify samples never seen in the training phase.

With CNN trained, the extended and reduced tests are applied. For the extended test, the network is submitted to 50 never-seen samples, and the classifier is evaluated using the confusion matrix, illustrated in Fig. 5.

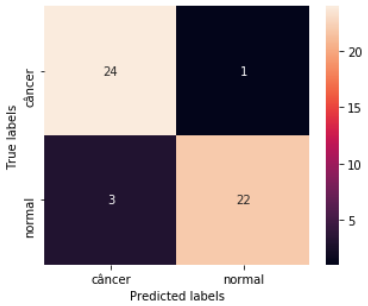

Fig. 5. Confusion Matrix for Extended Testing with CNN.

Another parameter evaluated, and of fundamental importance, is the area of the ROC curve, reaching a percentage of 92%, which is satisfactory due to the nature of the data. Note, in Fig. 5, that the classifier was only 4x wrong, with 3 normal images classified as cancer, and 1 image with cancer classified as normal by the network.

As reported in the methodology, a reduced test was separated in order to better visualize the prediction result together with the corresponding image, as illustrated in Fig. 6 and 7 the prediction result of 10 test samples, and the confusion matrix, respectively.

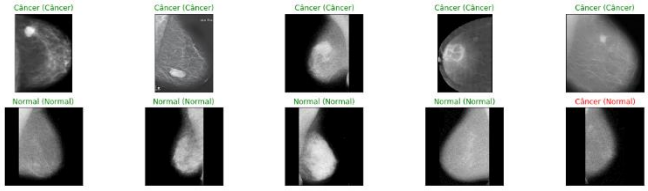

Fig. 6. Visual illustration of test rating with CNN.

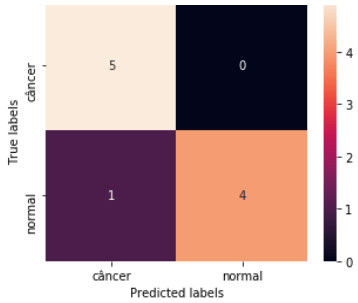

Fig. 7. Confusion matrix for reduced testing with CNN.

Thus, as well as in the extended test, for the reduced test the ROC metric is evaluated. Thus, the average of the area under the ROC curve obtained a value of 90%.

C. *Classifier with support vector machine*

For the second proposed classifier, the SVM technique was used and, after importing the data, the model was built. With the model built, the training phase can be started, and the loss function of the training phase can be seen in Fig. 8.

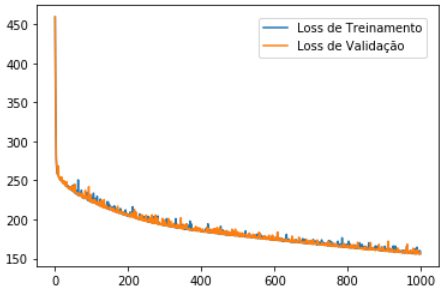

Fig. 8. Log-loss of the SVM classifier.

Furthermore, as it was applied to the CNN classifier, another fundamentally evaluated metric refers to the accuracy of the model, illustrated in Fig. 9.

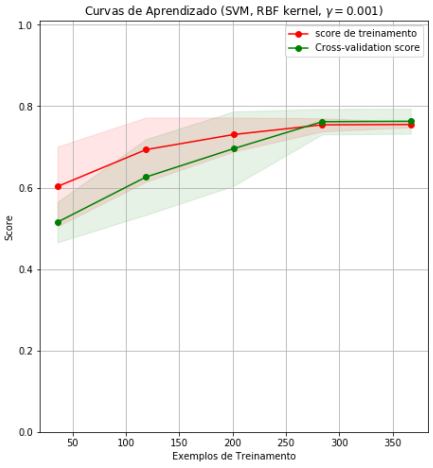

Fig. 9. Training accuracy of the SVM classifier.

For the cross-validation using the k-fold method, with a value of k=10, the following values were obtained, as shown in Fig. 10, of which the average was 76%.

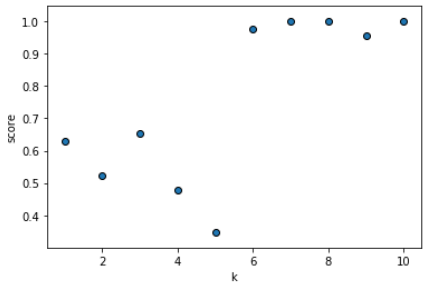

Fig. 10. Cross validation by the k-fold method.

After training, tests were performed, as well as for CNN. Therefore, the same test samples were subjected to extended and reduced tests in the SVM classification algorithm. For the extended test, the confusion matrix in Fig. 11 was obtained and the area under the ROC curve was 88%.

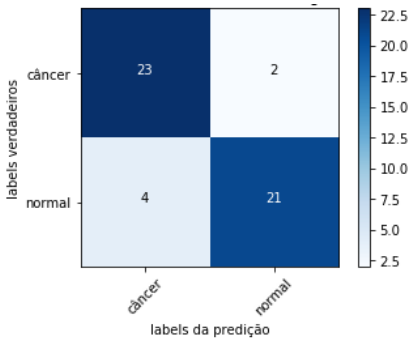

Fig. 11. Confusion matrix for extended testing with SVM.

Furthermore, the reduced test applied the same reasoning used in the classification via CNN, that is, the images, true and prediction labels were displayed, as illustrated in Fig. 12, and the confusion matrix can be seen in Fig. 13.

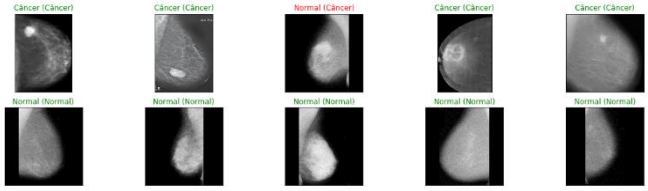

Fig. 12. Visual illustration of test classification with SVM.

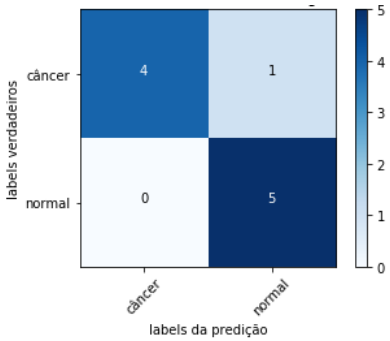

Fig. 13. Confusion matrix for reduced testing with SVM.

D. *Classifier performance evaluation.*

In order to objectively compare the classifiers, the metrics of the area under the ROC curve were taken into account for the testing phase, and the percentage of accuracy achieved for the training phase, specifically the validation of the model, which demonstrates how the classifier is able to generalize and perform well. The results of the comparison between models and the comparison algorithm can be seen in Fig. 14.

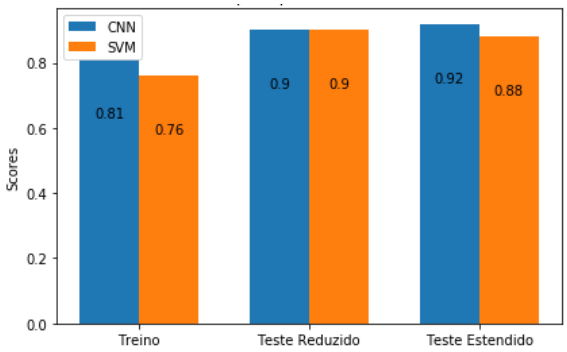

Fig. 14. Comparative performance evaluation of models.

As shown in Fig. 14, it can be seen that CNN's performance indices obtained better results than the support vector machine. Despite this, the difference was not so out of line, with a difference of 5% for training and 4% for the extended test.

Other parameters at the level of comparison evaluated refer to the time spent to train each model. With the help of the CProfile library, an execution log was generated for each classifier, which shows the time spent in the training phases. Thus, 435 seconds were obtained during the SVM training phase, which is equivalent to approximately 7 minutes and 25 seconds, and the training algorithm using CNN required 2102 seconds, that is, approximately 35 minutes.

After obtaining the classification data of both methods used, it was possible to carry out a more succinct analysis of which samples each classifier accepted or failed in the prediction. Thus, an algorithm capable of identifying the indices of each erroneous prediction of both classifiers was implemented. After analysis, it was observed that, between the techniques of SVM and CNN, for both the extended and reduced tests, there were no prediction errors in common. That is, all test samples that CNN missed, SVM got right, and all that SVM missed, CNN sorted correctly.

IV. Conclusions

In this working paper, the importance of a method capable of intervening positively on cancer was addressed, in the proposed case, the use of machine learning to aid in the diagnosis of breast cancer through the analysis of mammography images, serving as a second opinion for decision making.

The proposed classifiers were developed according to the applied methodology. And, the classifier method using convolutional neural network, obtained better results compared to the support vector machine method. The performance difference between them was relatively small and is due to the convolutional layers of the CNN, which has a high efficiency in the extraction of features.

The good performance of the classifier is proven by making a comparison between similar works observed in the literature. Thus, it is worth highlighting the test accuracy of the classifiers addressed in the works of [20], [21], and [22]. In the works mentioned, the test accuracies were, respectively, 72%, 81.73% and 97%. It is noteworthy that these values refer to the maximum accuracy obtained in each work, while, for the present study, the maximum accuracy obtained was 92%.

With the similarity analysis of the errors between the classifiers carried out, it can be inferred that, for a better classification of the test set, it is recommended to use both classifiers. In this way, it promotes greater security in the classification of images.

Therefore, the classifiers have a high capacity to support the diagnosis of breast cancer, and can be further improved with interfaces and data integration in a network. Thus, contributing to the promotion of health and excellence in diagnostic imaging.

References

[1] Instituto Nacional do Câncer, ABC do câncer: abordagens básicas para o controle do câncer, vol. 1. Rio de Janeiro: CEDC, 2011.

[2] T. Santos e M. Gonzaga, “Revista Saúde em Foco”, Fisiopatologia do câncer de mama e os fatores relacionados, vol. 1, no 10, p. 359–366, 2018.

[3] S. Vieira, C. Reis, D. Silva, R. Junior, R. Valença, e J. Mendes, Câncer de mama: Consenso da Sociedade Brasileira de Mastologia, 1o ed, vol. 1. Teresina: EDUFPI, 2017.

[4] Instituto Nacional do Câncer, Estimativa 2020: incidência de câncer no Brasil. Rio de Janeiro: Serviço de Educação e Informação Técnico-Científica, 2019.

[5] L. Souto, “Mineração de imagens para a classificação de tumores de mama”, Monografia, Universidade Federal Rural do Semi-Árido, Mossoró, 2014.

[6] B. Matheus, “BancoWeb: base de imagens mamográficas para auxílio em avaliações de esquemas CAD”, Universidade de São Paulo, São Carlos, 2010.

[7] C. Souza, S. Fustinoni, M. Amorim, E. Zandonade, J. Matos, e J. Schirmer, “Ciência e saúde coletiva”, Estudo do tempo entre o diagnóstico e início do tratamento do câncer de mama em idosas de um hospital de referência em São Paulo, Brasil, vol. 20, no 12, p. 3805–3016, 2015.

[8] J. Peixoto, E. Canella, e A. Azevedo, Mamografia: da prática ao controle. Rio de Janeiro: Gráfica Esdeva, 2007.

[9] A. Xavier, J. Sato, G. Giraldi, e C. Thomaz, “VII Workshop de Visão Computacional”, WVC, p. 67–72, 2011.

[10] Benveniste, A. Ferreira, e V. Aguillar, “Radiologia Brasileira”, Dupla leitura no rastreamento mamográfico, vol. 39, no 2, p. 85–89, 2006.

[11] J. Suckling, “The mini-MIAS database of mammograms”, PEIPA, the Pilot European Image Processing Archive, 2003. http://peipa.essex.ac.uk/info/mias.html (acessado fev. 15, 2020).

[12] M. Heath, “The Digital Database for Screening Mammography”, Proceedings of the Fifth International Workshop on Digital Mammography, p. 212–218, 2011.

[13] A. Bari, M. Chaouchi, e T. Jung, Análise Preditiva para leigos, 2o ed. Rio de Janeiro: Alta Books, 2019.

[14] Keras, “Keras: The Python Deep Learning library”, Keras: The Python Deep Learning library, 2015. https://keras.io/ (acessado jan. 07, 2020).

[15] Google Brain, “An end-to-end open source machine learning platform”, An end-to-end open source machine learning platform, 2015. https://www.tensorflow.org/ (acessado jan. 12, 2020).

[16] F. Chollet, Deep learning with Python. Shelter Island, New York: Manning Publications Co, 2018.

[17] A. Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems, 2o ed. Sebastopol, CA, EUA: O’Reilly Media, Inc., 2019.

[18] D. Cournapeau, “scikit-learn: Machine Learning in Python”, scikit-learn: Machine Learning in Python, 2007. https://scikit-learn.org/ (acessado fev. 09, 2020).

[19] Pedregosa, “JMLR”, Scikit-learn: Machine Learning em Python, p. 2825–2830, 2011.

[20] M. Reyes, “Clasificación de mamografías mediante Redes Neuronales Convolucionales”, Dissertação de Mestrado, Universidad Autónoma de Nuevo León, San Nicolás de los Garza, 2019.

[21] R. Silva e A. Carvalho, “Automatic Classification of Breast Lesions Using Transfer Learning”, IEEE Latin America Transactions, vol. 17, no 12, p. 1964–1969, 2019.

[22] F. Oliveira e A. Lins, “LUISA: Uma proposta de ferramenta para auxílio ao diagnóstico do câncer de mama a partir de imagens de mamografias digitalizadas.”, Revista de Engenharia e Pesquisa Aplicada, vol. 5, no 12, p. 73–83, 2020.