Received September 8, 2021, accepted October 6, 2021, date of publication October 26, 2021,
date of current version November 18, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3122948

Machine Learning for the Analysis of Conductivity From Mono Frequency Electrical Impedance Mammography as a Breast Cancer Risk Factor

ROSARIO LISSIET ROMERO CORIPUNA[1,2], DELIA IRAZÚ HERNÁNDEZ FARÍAS[1], BLANCA OLIVIA MURILLO ORTIZ[3,4], LUIS CARLOS PADIERNA[1], AND TEODORO CÓRDOVA FRAGA[1]

[1]División de Ciencias e Ingenierías Campus León, Universidad de Guanajuato Lomas del Bosque No. 103, Lomas del Campestre; León, Guanajuato 37672, México
[2]Escuela profesional de Física, Facultad de Ciencias Naturales y Formales, Universidad Nacional de San Agustín, Arequipa 04000, Perú
[3]Unidad de Investigación en Epidemiología Clínica, Unidad Médica de Alta Especialidad No. 1 Bajío, Instituto Mexicano del Seguro Social, León, Guanajuato 37320, México
[4]OOAD Guanajuato, Instituto Mexicano del Seguro Social, León, Guanajuato 37320, México

Corresponding author: Delia Irazú Hernández Farías (di.hernandez@ugto.mx)

This work was supported in part by the Grant from the Health Research Fund under Grant FIS7IMSS7PROT7G17–2/1715, and in part by the Medical Research Council of the Mexican Social Security Institute (IMSS), México. The work of Rosario Lissiet Romero Coripuna was supported by the Consejo Nacional de Ciencia y Tecnología (CONACYT) Scholarship for Postgraduate Students under Grant CVU 737873.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the National Commission for Scientific Research of the Mexican Institute of Social Security under Application No. R-2017-785-108.

ABSTRACT Computational approaches have been used for analyzing risk factors together with conventional mammograms for breast cancer detection. Currently, other screening methods like electro-impedance mammography are available. Notwithstanding, as far as we know there is not related work evaluating the role of electrical-conductivity index of the mammary gland as a quantitative factor for early detection of breast cancer. This paper aims to demonstrate the importance of including breast conductivity index as a quantitative local risk-factor by analyzing a dataset of Mexican patients from a machine learning perspective. There are 12 attributes distributed into two groups: *electrical-conductivity* (3) and *medical records* (9). According to the obtained results with unsupervised methods, the performance in terms of accuracy of using only electrical-conductivity (43%) is better than using all available features (38%) and the medical records (33%). On the other hand, we identified that SVM achieves higher results in comparison with other algorithms when only the electrical-features are used. The obtained results demonstrate the important role of conductivity index as a quantitative local risk-factor for being considered in screening processes. Besides, it emerges as an important aspect to be included in the development of automatic tools for experts to perform breast cancer diagnosis.

INDEX TERMS Electro-impedance, conductivity, machine learning, mammography MEIK, risk factor.

I. INTRODUCTION

Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020 [1], Breast Cancer (henceforth BC) being the most common type of cancer with 2.26 million cases, causing 685,000 deaths, placing it in the fifth place among different types of cancer.

People suffering from cancer have higher chances of survival when they have an early detection together with the opportunity to access proper medical treatment. The Sustainable Development program from the United Nations contemplates reducing the rate of premature deaths due to non-transmissible diseases (including cancer) by 2030. Such a challenging task could result in saving more than 40 million lives. However, it requires a multi-sectoral effort for expanding the available resources as well as establishing political commitments to offer an effective global response to cancer [2]. Attempting to improve the control of the BC, the World Health Organization promotes its early detection as well as the use of screening tools such as clinical breast exam,

The associate editor coordinating the review of this manuscript and approving it for publication was Nazar Zaki.

breast self-exam, and X-Ray mammography. The latter being the most widely used tool around the world.

An alternative to such methods is the Electro-Impedance Tomography (henceforth denoted as *EIT*) which has emerged as a novel procedure. It is less invasive and it does not use any ionizing radiation, which is one of the biggest drawbacks of traditional approaches [3]. EIT has its basis on the potential difference stored in the normal and pathologically altered tissues. According to the literature, the electrical properties are different among normal and malignant breast tissues, setting the stage for cancer detection by means of measuring electrical properties [4]. Malignant tumors show capacity and conductance values increased, which results in a decrease in impedance [5].

Research concerning the assessment of electro-impedance in physiological systems can be found in [6], [7]. Brown and Barber [8] laid the foundations of the EIT by developing systems for retrieving and reconstructing *in vivo* images. They also studied the electrical resistivity measurements for a wide range of tissues [9] and analyzed the potential clinical applications of this kind of tomography [10]. The first EIT taken from an upper arm was introduced in [11]. Surowiec *et al.* [12] studied the dielectric properties of the carcinoma and surrounding tissue from the breast considering a frequency range of 20 kHz up to 100 MHz. Exploiting the algorithm described in [13], some images of conductivity and electric permissiveness were reconstructed by using phantoms. Kejariwal *et al.* [14] published the first results on breast cancer detection throughout EIT. A medical device for obtaining 3D images of the conductivity from under the skin regions was developed by [5]. Wexler and Murugan [15] proposed a method for detecting breast cancer using high definition EIT. An algorithm for estimating both the location and size of abnormalities in an electrically conductive medium based on the EIT technique was developed in [16], [17]. A mammogram scheme of electrical impedance which describes information regarding standards and pathology on this topic was proposed by Karpov *et al.* [18].

With the help of electrical impedance mammograms, it is possible to study the anatomical structure of the mammary gland of women of diverse ages at different physiological periods, even in the one with maximum functional activity, i.e., the lactation period. Understanding the anatomy of the mammary gland makes it easier to understand the physiological and pathological processes that occur. EIT can be considered as a functional diagnostic technique that provides data from the mammary gland, its physiological state, and that it is sensitive to changes in the electrical conductivity of the tissues. Because of its high sensitivity in the primary diagnosis of benign tumors, EIT helps to give a dynamic follow-up of patients with diffuse and nodular types of mastopathy [19]. In this paper, we will refer to the use of EIT medical images for breast cancer detection as Electrical Impedance Mammography (henceforth denoted as EIM).

As already mentioned, BC represents a big main concern in developing countries. Therefore, strategies for promoting its early detection are crucial. In this paper, we focused our attention on analyzing a dataset of Mexican patients. According to the Mexican's National Institute of Geography and Statistics (*Instituto Nacional de Estadística y Geografía* - INEGI), in 2014 such a disease had a wide impact on the population with an average of 28.75 new cases per 100,000 women over 20 years old. Our main aim is to assess the role of the *conductivity index* as a potential risk factor to be considered for developing automatic systems that could serve as an aid tool during breast cancer diagnosis. In addition, we are interested in investigating whether or not a combination of well-known risk factors (namely anthropometric, gynecological-obstetric (OB/GYN), inherited-family, and environmental) with conductivity index could help in such a prediction. We are addressing automatic breast cancer diagnosis by casting it as a machine learning task where the objective is to determine the corresponding Breast Imaging Electrical Impedance Classification (denoted as BI-EIM) label as defined by medical specialists. BI-EIM is an annotation schema used by physicians when interpreting an EIM mammogram. It is parallel to the BI-RADS (Breast Imaging Reporting and Data System) used by radiologists to categorize X-Ray mammograms. It is important to highlight that we are not dealing with the task of assigning a positive or negative label according to whether or not a patient is suffering from breast cancer, but the obtained result of determining a suspicious malignancy can be a crucial aid insight for the final diagnosis.

According to [20], only a few research works in the literature have used machine learning techniques for dealing with EIT data. The progress so far achieved has been focused on two main tasks: i) To distinguish between benign and malign tumors [21]; and ii) To classify breast tissues according to the following classes: *Carcinoma*, *Fibro-adenoma*, *Mastopathy*, *Glandular*, *Connective*, and *Adipose* [22]–[26]. In all these papers, the authors took advantage of a dataset retrieved by [27], which is publicly available.[1] A similar task was performed by [28], where the aim was to discriminate between tissues extracted from different body parts (among them there is breast tissue) by using electrical impedance information. An aspect in common in the literature on this topic is that for experimental purposes only electrical conductivity information is used.

On the other hand, more research on automatic breast cancer detection has been done by exploiting other types of data. Aiming to distinguish between benign and malignant tumors from histopathological data, in [29] the authors used different machine learning techniques to achieve classification rates over 95% in precision terms. More information on the use of this data for BC can be found in [30]. Ultrasound images have been also widely used for performing this task, in [31] the authors present a comprehensive overview of the techniques and methods applied with ultrasound images. X-Ray mammograms have been, without doubt, the most widely

[1] https://archive.ics.uci.edu/ml/datasets/Breast+Tissue

exploited for performing automatic breast cancer detection. Recently, novel methods like deep learning have been applied reaching classification rates above 95% [32]. It is important to mention that, there are some works in the literature where the BI-RADS labels have been used as features during classification [33]–[35].

In this paper, we are proposing to use well-known risk factors in addition to electrical impedance information to determine the BI-EIM value that a medical report could have during screening inspired by what is done by physicians. Similar approaches have been addressed by using textual mammography reports [36]–[38] and also X-Ray mammographies [39]–[41] but not with the use of EIT data.

The major contributions of this paper are:

i) The results of this paper could be considered as a baseline for researching on classification of medical reports in terms of BI-EIM by taking advantage of electrical conductivity properties. To the best of our knowledge, there is no previous literature exploiting such kind of information for determining a BI-EIM value.
ii) We validate the usefulness of considering electrical conductivity data as a risk factor to be considered for breast cancer screening through a wide set of experiments using different sets of features.
iii) According to the obtained results, we identified which machine learning methods are most precise for identifying a label in terms of any suspicious malignancy when electrical conductivity data is used.

The rest of the paper is organized as follows. Section II describes the data and the methodology we used for experimental purposes. Sections III and IV present the obtained results from unsupervised and supervised machine learning approaches for evaluating the role of breast conductivity indexes as risk factors, respectively. Finally, in Section V we point out some conclusions and directions for future work.

II. MATERIALS AND METHODS

Computational sciences advances have derived in the development of software for *Computer Aided Diagnosis* (henceforth CAD). Such applications have been used as an aid tool for specialists for detecting different illnesses [42], [43]. CAD systems have been mainly exploited for detecting breast cancer in X-Ray mammograms [44]. As already mentioned, the main aim of this paper is to apply machine learning techniques over a dataset comprising recognized risk factors as well as breast-conductivity indexes obtained from the use of EIM as a potential risk factor. The idea is to establish the foundations for the development of an automatic auxiliary tool for identifying BC when breast conductivity indexes are obtained from the use of EIM as a screening tool.

A. DATA

The dataset we use for experimental purposes is composed by a total of 12 attributes: electrical conductivity changes in the mammary gland tissues [45], denoted as conductivity index of the left breast IC_{left},[2] conductivity index of the right breast IC_{right}[2] and discrepancy of the distribution between the left and right glands D_{dist}[2]. Besides, some well-known risk factors associated with reproductive, environmental, and lifestyle aspects are also considered: *Patients' age*[2], *Body Mass Index* (BMI)[2] [46], *Parity*[3] [47], [48], *Age of menarche*[3] [49], [50], *Menopause*[3], *Family history of breast cancer*[3], *Hormone therapy*[3] [51], *Alcohol consumption*[3], and *smoking*[3] [52].

Risk factors and breast-conductivity data were retrieved in the framework of the research project *Monofrequency Electrical Impedance Mammography (EIM) Diagnostic System in Breast Cancer Screening* [53], [54]. Such a project considered a total of 1200 female patients with a strong chance of having breast cancer. Information concerning the advantages and disadvantages of performing studies for breast cancer screening was provided to them. Afterward, they signed a written informed consent. The patients have not previously received surgical or pharmacological treatment for breast carcinoma; they were subjected to a physical examination before a pre- or postmenopausal screening mammogram performed by well-trained personnel according to standard protocols. The study population includes women with a wide range of characteristics: pregnant and breastfeeding, with breast prosthesis after a mastectomy, or with cosmetic breast surgery. This procedure was carried out in High Specialty Medical centers: *Unidad Médica de Alta Especialidad No.* 1 *Bajío* and *Unidad Médica de Alta Especialidad No. 48* located in Leon, Guanajuato, Mexico. Tables 1 and 2 show some general statistics about some of the aforementioned attributes.

TABLE 1. Average (avg), maximum (max), and minimum (min) values of the continuous quantitative features.

	Age	D_{dist}	IC_{left}	IC_{right}	BMI	$Menarche$
avg	48	10.16	0.46	0.46	28.87	13
max	80	55.53	0.84	0.90	50.66	21
min	20	1.56	0.09	0.11	10.8	8

TABLE 2. Distribution of the qualitative nominal features. Each of them was recorded as the answer to a dichotomous question (YES/NO).

	Parity	*Menopause*	*Hormone Therapy*	*Family history of breast cancer*	*Alcohol consumption*	*Smoking*
Yes	1101	562	450	195*	126	143
No	184	633	745	1000	1069	1052

*A total of 104 cases reported having such antecedent from her mother, grandmother or sister, while the remaining mentioned another familiar member.

An EIM was done for each patient, allowing us to obtain the three values related to the breast conductivity indexes. Besides, an additional X-Ray mammography or ECO Doppler was also performed for those patients older and younger than forty years old, respectively. When there is any suspicious of malignancy in the younger patients, a supplementary X-Ray mammography was done. The EIM was performed by using a computerized electro-impedance mammography equipment denoted as

[2]The attribute value type is continuous quantitative.
[3]The attribute value type is qualitative nominal.

MEIKv.5.6 (0.5 mA, 50 kHz) developed and manufactured by PKF SIM - Technika.

MEIKv.5.6 provides visualization of the conductivity of subsoil areas and it is suitable for conducting clinical research. This system uses three-dimensional measurement and reconstruction of conductivity distribution in biological tissues for clinical diagnosis. It consists of a compact array of 256 electrodes arranged in a square matrix with sides of 12 cm collocated in a rigid plane. Besides, the system includes an output multiplexer that serves as the connection with the electrodes array for providing an alternating current. The electrodes' array is connected to a potential difference measuring unit through an input multiplexer. Remote electrodes are attached to the extremities of the patient. Further details on the configuration of the device can be found in [5]. During the screening, the electrodes array is pressed against the breast. This allows for an increasing of the number of electrodes in contact and a decreasing of the thickness of the tissue to be measured. Figure 1 shows how the device is used for reconstructing and visualization of the resulting conductivity distribution as a stack of tomographic images. For what concerns the data we used for experimental purposes, details on the screening process can be found in [53].

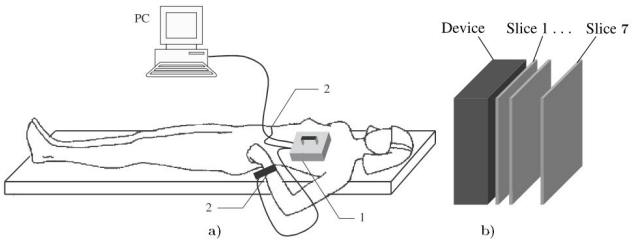

FIGURE 1. a) Physical configuration of the system and measuring procedure: 1—plane with 256 electrodes, 2—remote electrodes. b) 3D imaging planes [5].

The conductivity index (IC) is a quantitative feature of the breast structure received during an electric impedance scanning. It is measured in terms of siemens per meter (S/m). The ion concentration varies depending on the composition of the cellular elements in the breast. Acinar-ductal type of breast structure shows a low IC since it contains a large number of cellular elements with a high ions' concentration. Conversely, a high IC and low ions' concentration is observed in a breast with a large number of fat lobules and connective tissues (amorphous structure) [55].

Regarding the electrical impedance information, the dataset previously described was captured in terms of *mono-frequency electrical impedance mammography* [53] while in the state-of-the-art other settings have been used for data collection: *electrical impedance spectroscopy* [22], [23], *tetrapolar impedance measurement* [21], and *multi-frequency electrical impedance* [24], [28].

The dataset was manually annotated by medical specialists with a correspondent BI-EIM[4] [56] value assigned after interpreting each EIM. Table 3 shows the EIM scale and its correspondent BI-EIM category used for labeling the dataset.

TABLE 3. EIM categories used for dataset labeling.

Scale EIM	BI-EIM categories
0 - 1	$BI\text{-}EIM_1$ lesion is not defined
2	$BI\text{-}EIM_2$ benign tumors–routine mammography
3	$BI\text{-}EIM_3$ probably benign findings
4	$BI\text{-}EIM_4$ suspicious abnormality–biopsy
5	$BI\text{-}EIM_5$ highly suggestive of malignancy–treatment/biopsy

The distribution of BI-EIM values among the dataset is shown in Figure 2. It is important to mention that a subset of 5 cases were excluded from the 1200 patients due to the lack of or incomplete data. Therefore, in the following sections, we are considering only a total of 1195 cases.

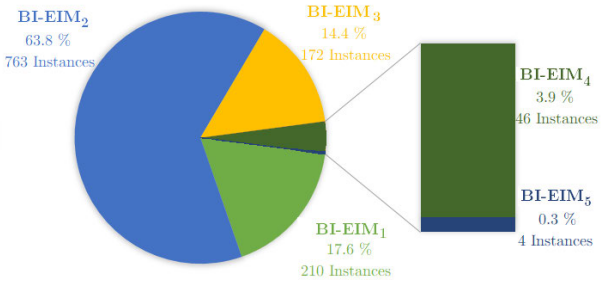

FIGURE 2. BI-EIM distribution in the dataset.

Figure 3 shows an histogram regarding the conductivity index values from the dataset. The IC_{left} shows a mean of (0.460 ± 0.132) with 0.84 and 0.09 as maximum and minimum values. For the IC_{right}, the values are (0.461 ± 0.133), 0.9 and 0.11, for the mean, maximum, and minimum values, respectively.

B. PROPOSED METHOD

We are interested in analyzing the performance of automatic methods in terms of how many instances could be correctly identified against the labels determined by a specialist. In order to analyze the dataset previously described, we applied both unsupervised and supervised machine learning algorithms. For assessing the performance of such methods, standard metrics from classification tasks were used.

Under an *unsupervised approach*, the aim is to study how the patients are clustered according to the different groups of features used. We hypothesize that the similarities between the patients being diagnosed with the same BI-EIM value could help for grouping them. On the other hand, the supervised learning approach has been used for developing models able to predict the corresponding BI-EIM value that a given patient has, considering different sets of features for training purposes. Experiments were carried out using different classification approaches by taking advantage of a set of well-known classifiers. For both approaches, we used Scikit-learn [57] implementation for all the machine

[4]http://www.onkocet.eu/download/MEIKUserManual.pdf

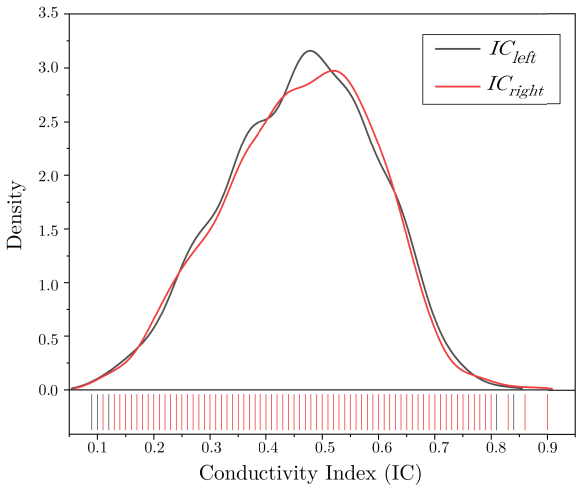

FIGURE 3. Conductivity index histogram for left and right breast in the dataset.

learning algorithms. Different experimental settings were assessed. In all the experiments, we first use a subset including only data coming from electrical impedance, i.e., IC_{left}, IC_{right}, and d_{dist} (henceforth the experiments involving only these features are denoted as *ElecFts*); then, the remaining nine attributes regarding risk factors are exploited (denoted as *RiskFts*); and finally, both groups of features are merged in to a single one (from now on it such experiments are denoted as *AllFts*). In the following sections, we describe each experiment carried out in more detail as well as the obtained results.

III. UNSUPERVISED ANALYSIS

For experimental purposes, two types of settings were evaluated first considering the data with its original values and also after applying normalization data techniques (namely StandarScaler (removing the mean and scaling to unit variance), MinMax (scaling in a range between zero and one), and MaxAbsScaler (scaling each feature by its maximum absolute value)) aiming to standardize the attributes we have, taking into account that, as they come from different aspects their values have diverse scales. Two different centroids initialization methods were used: *k-means++* and *k-random*. The number of clusters (*n_clusters*) to generate was set to 4. Given the fact that there are only a few instances in BI-EIM$_5$, we decided to only consider four clusters to be done by merging together BI-EIM$_4$ and BI-EIM$_5$. Default parameters settings were used. Table 4 shows the obtained results over the original data. The outcomes after applying the normalization data techniques with both sets of features show no significant differences or improvements with respect to the original data, therefore, we decided not to include them in the manuscript.

The distribution of the clusters is practically equal for both algorithms with each feature set. Considering the *ElecFts*, the cluster number 2 has the highest amount of instances. However, more than 20% of the cases with BI-EIM$_2$ were not included in this group by both algorithms. Instead, it appears

TABLE 4. Obtained clusters distribution applying *k-means++* and *k-random* using electrical-conductivity features, all available features and only risk features.

BI-EIM	Original	*ElecFts*		*AllFts*		*RiskFts*	
		K-m	*K-r*	*K-m*	*K-r*	*K-m*	*K-r*
1	210	427	427	322	323	298	301
	17.57%	35.73%	35.73%	26.95%	27.03%	24.94%	25.19%
2	763	504	504	447	463	415	414
	63.85%	42.18%	42.18%	37.41%	38.74%	34.73%	34.64%
3	172	214	213	231	230	245	243
	14.29%	17.91%	17.82%	19.33%	19.25%	20.50%	20.33%
4	46	50	51	195	179	237	237
	3.85%	4.18%	4.27%	16.32%	14.98%	19.83%	19.83%
5	4						
	0.33%						

that these instances were included in cluster number 1. Finally, in cluster 3 there are around 3% of instances placed in the wrong group. For what concerns to *RiskFts* and *AllFts*, both configurations seem to have a similar performance concerning clusters 1 and 3, but not for clusters 2 and 4. With *AllFts*, while the *k-means++* implementation assigns more instances to cluster 4, the *k-random* does to the cluster 1. It is interesting to note that, unlike using only electrical conductivity information where there is a notable skewed distribution towards cluster 2, in these cases the samples are a little bit more disseminated among the clusters, being the cluster 2 the one having the highest number of instances.

Figure 4 shows a schematic representation of the instances according to its BI-EIM value with each group of features as well as how they were distributed in the clusters. As it can be observed, the instances are highly overlapped making the clustering very challenging. In the *ElectFts* and *AllFts* it is possible to note some trends, like most of the instances in BI-EIM$_4$ have high values of D_{dist} (being the x-axis in the *ElectFts* and y-axis in *AllFts*); in this regard, there is a salient instance of BI-EIM$_5$ having the highest value of the same attribute. Therefore, the attribute providing such separation is in the *ElectFts* set. Similarly, it is also possible to note a group composed by instances in BI-EIM$_3$. For what concerns to the BI-EIM$_1$ and BI-EIM$_2$, the instances are very close between them. Considering the *RiskFts* is even harder to distinguish groups of instances belonging to the same class. With this group of features all the instances are more dispersed.

1) COMPARING CLUSTERING RESULTS

We further analyze how many instances were correctly assigned to its corresponding BI-EIM value. Table 5 shows the obtained results when the three subsets of features were used for generating the clusters. The first two columns show the Original distribution of instances according to the golden labels. Then, the samples' distribution when using each version of the *k-means* are shown for each set of features. The only electrical-based features showed a higher accuracy than using the other subsets. A total of 516 instances were correctly clustered, which represents the 43% of the total

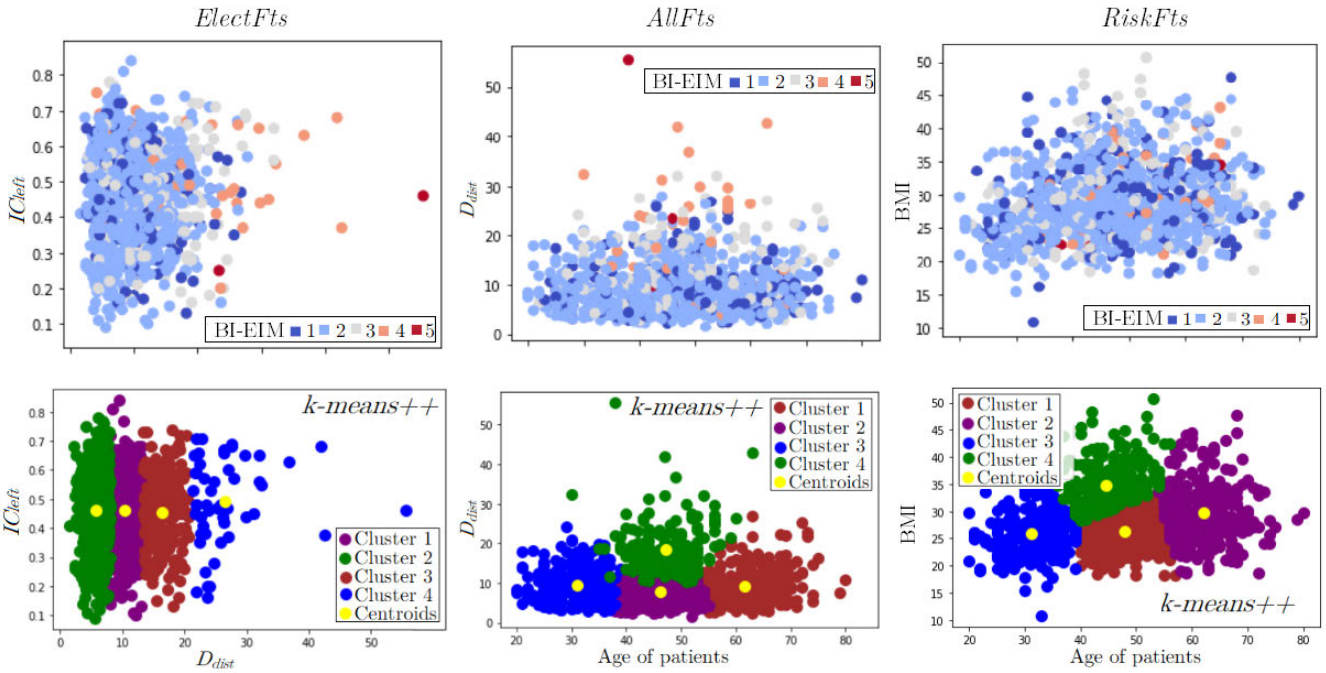

FIGURE 4. Schematic representation of the generated clusters with each group of features. In the upper side the instances with its corresponding BI-EIM value are plotted.

amount of data. The *ElecFts* have a better performance for the BI-EIM$_1$, BI-EIM$_2$, and BI-EIM$_3$, while on the contrary, by using *AllFts* there are a greater amount of correctly assigned instances in the BI-EIM$_4$ and BI-EIM$_5$. As it can be observed, when only the *RiskFts* are used, the amount of instances correctly clustered is lower than with the others groups in particular for BI-EIM$_4$. According to the obtained results, with the *ElecFts* almost half of the instances are correctly clustered, which can be an insight on the usefulness of using only these attributes for assigning a BI-EIM value automatically. These results can be also related to the plots in Figure 4, where the distribution of the instances due to the feature set used allows us to observe differences among them that are inline with the outcomes in terms of how many instances from the same class were grouped together.

IV. SUPERVISED ANALYSIS

Taking advantage of annotated data, we also decided to assess the possibility of assigning a given BI-EIM value as a classification task. A set of classifiers[5] composed by: *Naïve Bayes* (NB), *Decision Tree* (DT), *Random Forest* (RF), *k-Nearest Neighbors* (with three values of k namely: $3NN$, $5NN$, and $7NN$), *Logistic Regression* (LR), and *Support Vector Machine* (SVM) was used. Some of these algorithms have been already exploited for addressing classification with electrical conductivity features: SVM [24], [28] and KNN [21]; apart from that

[5]Default parameters of the methods were used, except for the SVM and LR where a GridSearch strategy was used in each case of obtaining optimal parameter values over the original data distribution.

TABLE 5. Comparison of the obtained clusters with different subsets of features. In this table, experiments carried out with *K-means++* are on the *K-m* column, and the ones with *K-random* on the *K-r*.

BI-EIM	Original	*ElecFts*		*AllFts*		*RiskFts*	
		K-m	*K-r*	*K-m*	*K-r*	*K-m*	*K-r*
1	210	82	82	69	69	62	62
2	763	352	352	318	329	286	286
3	172	64	64	26	26	46	45
4	46	16	16	23	22	5	5
5	4						
4 & 5	50	18	18	25	28	6	6
	1195	516	516	438	452	399	398
	100%	43.18%	43.18%	36.65%	37.82%	33.39%	33.31%

neural network classifiers [22] and Linear Discriminant Analysis [23] have been also exploited. The task was addressed by casting it as a *multi-class* classification problem, where the aim is to generate a model for identifying the BI-EIM value of a given instance.

Similar to the previous Section, we experimented only with electrical conductivity features, with only the risks factors and then, using also all the available information. As golden labels, the BI-EIM values assigned by the experts were used. For evaluation purposes, we used the *Accuracy* and the *F-score*. All the experiments were performed following a *Stratified k-fold strategy*, having values of $k = 4$. Such a value was fixed according to the number of available instances of BI-EIM$_5$.

Figure 5 shows the obtained results in Accuracy terms (defined as in Eq. 1) for each of the classifiers and sets of features. The highest accuracy was obtained by NB when

using the *ElecFts* reaching a rate of 0.641. SVM achieves 0.638 and 0.63 for *AllFts* and *RiskFts*, respectively; in both cases, these are the highest outcomes for these features' sets. In most of the cases, when using the *RiskFts* the lowest results are achieved. The classifiers performed more poorly are LR and DT. In the literature, some works addressing the task of assigning a BI-EIM label report accuracy rates of 0.85 and 0.83, when using textual content (features like a predefined set of terms in the mammography reports domain were exploited) [38] and X-Ray mammograms (by using information regarding Region of interest and masses manually identifies were used) [40], respectively. Even when the highest accuracy obtained by us is lower than the outcomes in related tasks, it is still competitive considering that in the aforementioned approaches there is more available information for experimenting with supervised learning unlike in our task, where only three features were exploited.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where:

$TP =$ True Positives, $TN =$ True Negatives

$FP =$ False Positives, $FN =$ False Negatives

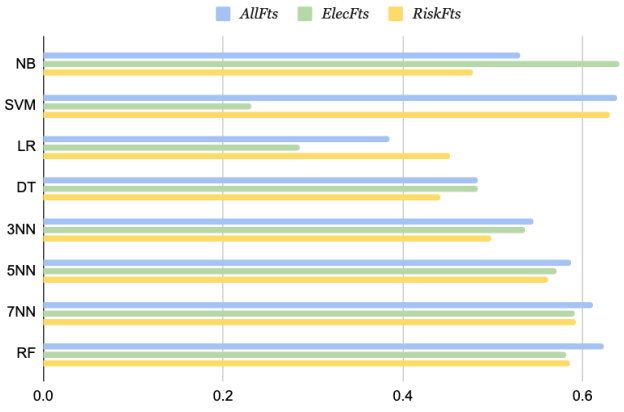

FIGURE 5. Obtained results in terms of accuracy with the different sets of features.

It is important to mention that, even when Accuracy is a widely used evaluation metric in related tasks, it has some drawbacks when the dataset in hand is not balanced, as in our case. Given the fact that the classifiers' performance can be undermined due to the imbalance degree among the classes, and that in the dataset in hand there is a skewed imbalanced distribution towards the BI-EIM$_5$ class, the problem we are addressing can be considered as a *class imbalance problem*. Let us to mention that, for each instance in BI-EIM$_5$, there are 52, 191, 43, and 12 instances from BI-EIM$_1$, BI-EIM$_2$, BI-EIM$_3$, and BI-EIM$_4$, respectively. According to [58], a dataset is imbalanced when the number of examples representing one class is much lower than the ones of the other classes. Imbalanced datasets are pervasive in real life, especially in the medical field. In fact, imbalanced distribution has been recognized as an important challenge for medical real-world data, particularly for breast cancer detection [59]. In our study case, as in most of the imbalanced problems, the underrepresented class is the one of interest, we are aimed to identify patients having a high BI-EIM value, which could be a potential alert signal of the presence of any anomaly in the mammary gland.

Most of the classification algorithms, when generating a criteria for classification, tend to be biased by the number of samples belonging to a given class (the one with most samples) provoking a misclassification of the instances in the minority class. Attempting to deal with these kinds of problems, research has been done devoted to develop strategies for addressing such a problem [60]. Broadly speaking, there are two main techniques: *a) Algorithm level approaches*, which involve to adapt learning algorithms for dealing with class imbalance by the use of different criteria to optimize classification rates for minority and majority classes; and, *b) Data level approaches* that work at pre-processing stage, which aim is to modify imbalanced data using different procedures to provide a balanced or more adequate data distribution by the use of sampling methods [58]; they are independent of the learning algorithm used. Such methods randomly discard majority class instances (commonly denoted as *Random Under-Sampling* - RUS) and, on the contrary, instances from the minority class are replicated (denoted as *Random Over-Sampling* - ROS). However, there are some drawbacks of such alternatives. After discarding some instances by RUS, useful information for building the model can be excluded. While applying ROS, the probabilities of over-fitting the learning algorithm increase, then, alternative approaches have been proposed. In this sense, there is SMOTE (Synthetic Minority Oversampling Technique) [61] which creates new instances of the minority class by interpolating them by the use of different techniques.

Aiming to assess the usefulness of *Data level* approaches in the task at hand, we decided to generate a set of corpora with different imbalanced degrees for experimental purposes. First, an ideal scenario where the classes are fully balanced, i.e., the same amount instances per class was generated (denoted as *1:1*). In the second one, a configuration where the instances from classes BI-EIM$_1$, BI-EIM$_2$, and BI-EIM$_3$ were under-sampled and the remaining two were over-sampled (denoted as *Conf1*). Finally, in the *Conf2* the instances in BI-EIM$_1$, BI-EIM$_2$, and BI-EIM$_3$ were left untouched while the remaining two were also over-sampled. We exploited the Imbalanced-Learn Python implementation [62] of RUS and SMOTE (for over-sampling) with the k-neighbors parameter fixed to 2. Table 6 shows the distribution per class after applying *data-level pre-processing* methods.

Figure 6 shows a schematic representation of the data distribution we used for experimental purposes considering all subsets of features. It is important to highlight that, five different subsets were generated for each configuration. Then, we manually selected one of each trying to choose the one which ensures that data from minority classes were more

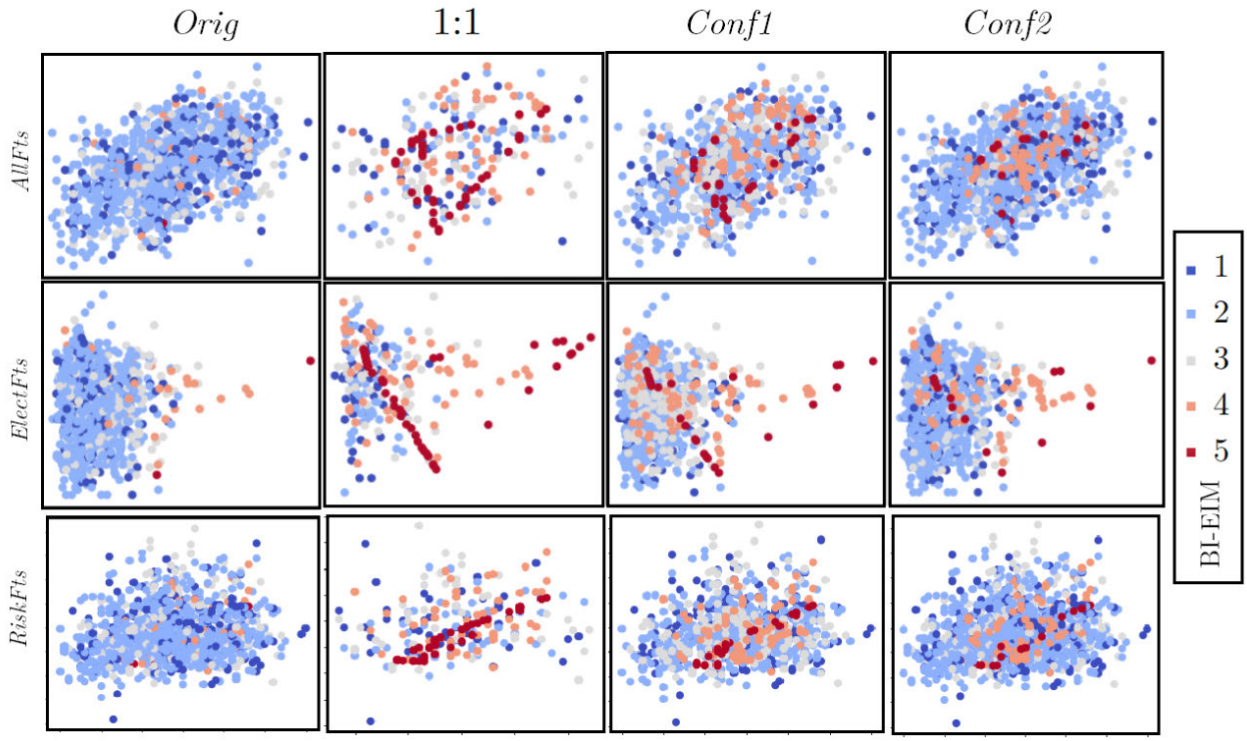

FIGURE 6. Data distribution of the whole dataset after applying *data-level pre-processing* with both sets of features.

TABLE 6. Distribution of each class after applying data-level pre-processing techniques. RUS is highlighted with dark-gray color box, while ROS with light-gray. The original distribution is denoted as *Orig*.

		All Data			
		Orig	1:1	*Conf1*	*Conf2*
BI-EIM	1	210	50	200	210
	2	763	50	450	763
	3	172	50	150	172
	4	46	50	75	92
	5	4	50	25	16

sparse in the feature space. As it can be observed, in the original distribution, it is very hard to identify the instances belonging to BI-EIM$_5$, on the contrary, with the proposed configurations it is possible to distinguish some of them.

In a similar fashion to the Unsupervised Approach, we applied the same normalization data techniques over the original data distribution, once again we observed very similar results in classification rates terms. Then, only experiments with the original data together with *data-level* techniques for compensating class imbalance were carried out. As mentioned before, a set of classifiers were used, however, for the sake of the readability, we filtered out only the three best performing classifiers considering the results obtained with the *ElecFts* subset in F-score terms. In this second set of experiments, we decided to use F-score (as defined in Eq. 2) aiming to evaluate the classifiers' performance for each class on its own.

$$\text{F-score} = 2 \times \frac{precision \times recall}{precision + recall} \tag{2}$$

where:

$$precision = \frac{TP}{TP + FP}, \quad recall = \frac{TP}{TP + FN}$$

Table 7 shows a summary of the obtained results with the different settings when the NB, SVM, and RF classifiers were exploited. The results where the classification rate for *ElecFts* are higher than the rest are highlighted in bold. It is important to mention that, in order to assess the variability of the classifiers, each experiment was performed 30 times, the results in Table 7 correspond to the obtained average and standard deviation in each case.

A. DISCUSSION

Considering the *Orig* distribution with all classifiers, in 8 out of 15 of the cases the *ElecFts* reached the highest classification rate. BI-EIM$_2$ is the class with the highest classification rates considering all classifiers. It is important to note that, in the case of BI-EIM$_4$ in the three classifiers a better performance is observed with the *ElecFts*. In the case of BI-EIM$_5$, the performance is remarkably low in all experiments, this can be due to the low amount of instances belonging to this class. In the balanced scenario (i.e., 1:1), a drop in the results in BI-EIM$_2$ is observed in all cases, probably because of the loss of information during undersampling; however, in the experiments with the three sets of features, the results with *ElecFts* the highest. For BI-EIM$_3$, the outcomes of using electrical features with SVM are the highest; besides, it is interesting to note that, in this class the results are very similar also among the different distribution settings. A quite comparable

TABLE 7. Obtained results in F-score terms. Each result is the average of the 30 experiments carried out, below in small font size the standard deviation is included. As it can be observed, the results across the experiments were very similar in all cases. In this Table, we used the acronyms: ***AF***, ***EF***, and ***RF*** for ***AllFts***, ***ElectFts***, and ***RiskFts***, respectively.

		NB				SVM				RF			
		Orig	1:1	*Conf1*	*Conf2*	*Orig*	1:1	*Conf1*	*Conf2*	*Orig*	1:1	*Conf1*	*Conf2*
BI-EIM$_1$	*AF*	0.0291 ± 0.0124	0.2699 ± 0.0727	0.1325 ± 0.0193	0.0263 ± 0.0131	0	0.0769 ± 0.0698	0.3641 ± 0.0003	0	0.0379 ± 0.0154	0.3827 ± 0.0497	0.1411 ± 0.0207	0.0479 ± 0.0184
	EF	0	**0.4098** ± 0.0217	0	0	**0.2775** ± 0.0332	**0.4010** ± 0.0214	0.3504 ± 0.0411	**0.2744** ± 0.0105	**0.1843** ± 0.0201	0.3561 ± 0.0348	**0.3035** ± 0.0201	**0.1452** ± 0.0234
	RF	0.0162 ±0.0089	0.2774 ±0.0470	0.0942 ±0.0227	0.0177 ±0.0103	0.0	0.0582 ±0.0558	0.3695 ±0.0013	0.0	0.0696 ±0.0200	0.4118 ±0.0541	0.1771 ±0.0156	0.0724 ±0.0183
BI-EIM$_2$	*AF*	0.7189 ± 0.0061	0.2060 ± 0.047	0.6268 ± 0.0101	0.7208 ± 0.0087	0.7794 ± 0.0000	0.1541 ± 0.0381	0	0.7594 ± 0.0005	0.7737 ± 0.0034	0.2622 ± 0.0420	0.67 ± 0.0078	0.7684 ± 0.0037
	EF	**0.7866** ±0.0014	**0.2409** ± 0.0269	**0.6875** ± 0.0020	**0.7731** ± 0.0014	0.2234 ± 0.0756	**0.1823** ± 0.0415	**0.2137** ± 0.0664	0.4656 ± 0.0171	0.7345 ± 0.0059	**0.4393** ± 0.0425	0.616 ± 0.0125	0.7461 ± 0.0058
	RF	0.6696 ±0.009	0.2245 ±0.0564	0.5874 ±0.0247	0.6902 ±0.0245	0.7744 ±0.0012	0.1557 ±0.0359	0.0272 ±0.0070	0.7564 ±0.0012	0.7454 ±0.0048	0.2593 ±0.0489	0.6153 ±0.0072	0.7346 ±0.0060
BI-EIM$_3$	*AF*	0.2179 ± 0.0192	0.4370 ± 0.0422	0.2736 ± 0.0197	0.2139 ± 0.0182	0	0.1765 ± 0.0207	0	0	0.1636 ± 0.0264	0.3362 ± 0.0463	0.2713 ± 0.0255	0.155 ± 0.0203
	EF	0.1958 ± 0.0142	0.2819 ± 0.0321	0.2019 ± 0.0182	0.1844 ± 0.0163	**0.2493** ± 0.0155	**0.2189** ± 0.0307	**0.2606** ± 0.0208	**0.2548** ± 0.0193	**0.1900** ± 0.0192	0.2871 ± 0.0550	0.2163 ± 0.0272	**0.1953** ± 0.0193
	RF	0.1013 ±0.0137	0.4409 ±0.0353	0.1187 ±0.0172	0.1025 ±0.0148	0	0.1756 ±0.0126	0	0	0.0941 ±0.0167	0.3401 ±0.0389	0.1375 ±0.0222	0.0810 ±0.0196
BI-EIM$_4$	*AF*	0.205 ± 0.0376	0.3853 ± 0.0521	0.3064 ± 0.0208	0.3578 ± 0.0328	0	0.044 ± 0.0364	0	0.1275 ± 0.0267	0.1366 ± 0.0577	0.5474 ± 0.0432	0.5589 ± 0.0382	0.6226 ± 0.0279
	EF	**0.2682** ± 0.0309	0.2092 ± 0.0343	0.2763 ± 0.0155	**0.3934** ± 0.0149	**0.1757** ± 0.0298	**0.1969** ± 0.0323	**0.1496** ± 0.0208	**0.2117** ± 0.0450	**0.1599** ± 0.0403	0.4864 ± 0.0426	0.3941 ± 0.0374	0.4767 ± 0.0356
	RF	0.0339 ±0.0217	0.3779 ±0.0451	0.1889 ±0.0330	0.2626 ±0.0367	0.0	0.0443 ±0.0353	0.0405 ±0.0202	0.1711 ±0.0360	0.0515 ±0.0326	0.5372 ±0.0446	0.3349 ±0.0356	0.4224 ±0.0314
BI-EIM$_5$	*AF*	0	0.7172 ± 0.0259	0.4484 ± 0.0396	0.2395 ± 0.0541	0	0.5357 ± 0.0413	0.0911 ± 0.0645	0	0	0.9204 ± 0.0160	0.8776 ± 0.0493	0.5579 ± 0.0768
	EF	0	0.4843 ± 0.0242	0.3117 ± 0.0531	0.109 ± 0.0733	0	0.5009 ± 0.0236	**0.3193** ± 0.0346	0.0522 ± 0.0353	0	0.705 ± 0.0476	0.3799 ± 0.0592	0.2724 ± 0.0902
	RF	0	0.7150 ±0.0333	0.4034 ±0.0496	0.215 ±0.0391	0	0.5557 ±0.0501	0.2824 ±0.1001	0.1400 ±0.0917	0	0.9194 ±0.0172	0.8539 ±0.0357	0.5254 ±0.1055

scenario is found for BI-EIM$_4$. Overall, there is a positive impact on the classification rate for the rest of the classes, particularly, we observed the highest differences in BI-EIM$_5$. In this distribution, when SVM is used as a classifier all but the BI-EIM$_5$ have better performance with *ElecFts*.

For what concerns to *Conf1*, *ElecFts* with SVM show better results than using *AllFts* and *RiskFts* in all cases except in BI-EIM$_1$, where the performance is very similar among the three groups of features. However, it is important to highlight that the parameters used during the experiments are the same as in the original distribution, then it seems that such setting has no impact on the classifier when the *ElecFts* are used, and the contrary occurs with the whole set of features. Finally, in *Conf2*, where a more realistic distribution was considered, the positive effect of applying *data-level* techniques for class imbalance is particularly remarkable in the three classifiers on the BI-EIM$_5$, where, the performance of *ElecFts* (despite being lower than with *AllFts*) is still competitive. Besides, the classifiers' performance in this configuration concerning those classes that were left untouched (BI-EIM$_1$, BI-EIM$_2$, and BI-EIM$_3$) is very similar to the one in the original distribution, there are even cases where there is an improvement in the obtained results when the *ElecFts* are used.

Overall, the classification rates are very low for all classes except for BI-EIM$_2$, the one with more available instances for building the classification models. However, given the fact that data with conductivity indexes for addressing breast cancer detection is scarce, the obtained results could be considered as a starting point for further research on this topic. Furthermore, the outcomes serve to validate that, comparing the results of the features settings, the performance of *ElecFts* is still competitive taking into account that such subset is composed of only three features.

V. CONCLUSION AND FUTURE WORK

The main conclusion is that the electrical-conductivity of mammary gland was proved to be an effective index to classify medical records in terms of BI-EIM. Experimental results summarized in Table 7 shows that this index equals or surpasses the classification accuracy of the three machine learning techniques in the four considered configurations; therefore, it can be considered an alternative to the classification based on medical records, with the advantage of reducing the number of attributes from nine to three. The BI-EIM supervised classification problem was hard to solve for any of the considered classifiers since it is strong unbalanced. Similarly, the results reported in Table 5 indicates that the electrical-conductivity features improve the accuracy indexes from 33 to 43% for both unsupervised clustering techniques (k-means and k-random). These results leave a wide room for improvements and position our results as a base line of forthcoming machine learning techniques. The related work found in literature indicates that the approach is different

to both the electro-impedance and the clinical record-based screening methods. To the best of our knowledge this is the first time the electrical conductivity is evaluated as an index for the BI-EIM classification problem in a real scenario with Mexican population. As future work, the EIM data acquisition will be analyzed to study differences among age cohorts.

ACKNOWLEDGMENT

Non-financial competing interests. Informed consent was obtained from all patients in writing.

REFERENCES

[1] J. M. Ferlay, F. Ervik, M. Lam, L. Colombet, M. Mery, and M. Piñeros. (2020). *Observatorio Global del Cancer: Cancer Today. Lyon: Agencia Internacional de Investigación Sobre el Cancer*. Accessed: Apr. 22, 2020. [Online]. Available: https://gco.iarc.fr/today

[2] IAEA. (2020). *Programme of Action for Cancer Therapy (PACT)*. Accessed: Jul. 19, 2020. [Online]. Available: https://www.iaea.org/services/key-programmes/programme-of-action-for-ca%ncer-therapy-pact

[3] E. Y. K. Ng, S. V. Sree, K. H. Ng, and G. Kaw, "The use of tissue electrical characteristics for breast cancer detection: A perspective review," *Technol. Cancer Res. Treatment*, vol. 7, no. 4, pp. 295–308, Aug. 2008.

[4] T. A. Khan and S. H. Ling, "Review on electrical impedance tomography: Artificial intelligence methods and its applications," *Algorithms*, vol. 12, no. 5, p. 88, 2019.

[5] V. Cherepenin, A. Karpov, A. Korjenevsky, V. Kornienko, A. Mazaletskaya, D. Mazourov, and D. Meister, "A 3D electrical impedance tomography (EIT) system for breast cancer detection," *Physiolog. Meas.*, vol. 22, no. 1, pp. 9–18, Feb. 2001.

[6] R. Plonsey and R. Collin, "Electrode guarding in electrical impedance measurements of physiological systems—A critique," *Med. Biol. Eng. Comput.*, vol. 15, no. 5, pp. 519–527, Sep. 1977.

[7] R. D. Stoy, K. R. Foster, and H. P. Schwan, "Dielectric properties of mammalian tissues from 0.1 to 100 MHz; a summary of recent data," *Phys. Med. Biol.*, vol. 27, no. 4, p. 501, 1982.

[8] D. C. Barber and B. H. Brown, "Applied potential tomography," *J. Phys. E, Sci. Instrum.*, vol. 17, no. 9, p. 723, 1984.

[9] D. C. Barber, B. H. Brown, and I. L. Freeston, "Imaging spatial distributions of resistivity using applied potential tomography—APT," in *Information Processing in Medical Imaging*. Springer, 1984, pp. 446–462.

[10] B. Brown, D. Barber, and A. Seagar, "Applied potential tomography: Possible clinical applications," *Clin. Phys. Physiolog. Meas.*, vol. 6, no. 2, p. 109, 1985.

[11] B. Brown, T. Karatzas, R. Nakielny, and R. Clarke, "Determination of upper arm muscle and fat areas using electrical impedance measurements," *Clin. Phys. Physiolog. Meas.*, vol. 9, no. 1, p. 47, 1988.

[12] A. J. Surowiec, S. S. Stuchly, J. R. Barr, and A. Swarup, "Dielectric properties of breast carcinoma and the surrounding tissues," *IEEE Trans. Biomed. Eng.*, vol. BME-35, no. 4, pp. 257–263, Apr. 1988.

[13] H. Griffiths, "A phantom for electrical impedance tomography," *Clin. Phys. Physiolog. Meas.*, vol. 9, no. 4A, pp. 15–20, Nov. 1988.

[14] M. Kejariwal, K. Kaster, J. Jurist, and J. Pakanati, "Breast cancer detection using electrical impedance tomography: Spice simulation," in *Proc. 15th Annu. Int. Conf. IEEE Eng. Med. Biol. Societ*, Oct. 1993, pp. 64–65.

[15] A. Wexler and R. Murugan, "High definition electrical impedance tomography methods for the detection and diagnosis of early stages of breast cancer," U.S. Patent 09 991 993, Aug. 8 2002.

[16] O. Kwon, J. R. Yoon, J. K. Seo, E. J. Woo, and Y. G. Cho, "Estimation of anomaly location and size using electrical impedance tomography," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 1, pp. 89–96, Jan. 2003.

[17] N. K. Soni, A. Hartov, C. Kogel, S. P. Poplack, and K. D. Paulsen, "Multi-frequency electrical impedance tomography of the breast: New clinical results," *Physiolog. Meas.*, vol. 25, no. 1, p. 301, 2004.

[18] A. Karpov, A. Kolobanov, and M. Korotkova, *An Electrical Impedance Mammographic Scheme—Norms and Pathology*. IntechOpen, 2015, ch. 1.

[19] A. Karpov, A. Kolobanov, and M. Korotkova, "Diagnostic system in electrical impedance mammography: Background," in *Breast Imaging*, A. M. Malik, Ed. Rijeka, Croatia: IntechOpen, 2017, Ch. 8.

[20] J. Zuluaga-Gomez, N. Zerhouni, Z. Al Masry, C. Devalland, and C. Varnier, "A survey of breast cancer screening techniques: Thermography and electrical impedance tomography," *J. Med. Eng. Technol.*, vol. 43, no. 5, pp. 305–322, Jul. 2019.

[21] A. Al Amin, S. Parvin, M. Kadir, T. Tahmid, S. K. Alam, and K. S.-E. Rabbani, "Classification of breast tumour using electrical impedance and machine learning techniques," *Physiol. Meas.*, vol. 35, no. 6, p. 965, 2014.

[22] A. Helwan, J. B. Idoko, and H. Rahib Abiyev, "Machine learning techniques for classification of breast tissue," *Proc. Comput. Sci.*, vol. 120, pp. 402–410, Aug. 2017.

[23] J. E. da Silva, J. P. M. de Sá, and J. Jossinet, "Classification of breast tissue by electrical impedance spectroscopy," *Med. Biol. Eng. Comput.*, vol. 38, no. 1, pp. 26–30, Jan. 2000.

[24] P. C. Shetiye, "Detection of breast cancer using electrical impedance and RBF neural network," *Int. J. Inf. Electron. Eng.*, vol. 5, no. 5, p. 356, 2015.

[25] P. Verma, S. Ramasamy, and D. D. D. P. Selvam, "Classification of breast cancer from electrical impedance measurements dataset in samples of freshly excised breast tissues," *Platform, A J. Sci. Technol.*, vol. 4, no. 1, pp. 107–116, 2021.

[26] N. Chumuang, P. Pramkeaw, and A. Farooq, "Electrical impedance of Breast's tissue classification by using bootstrap aggregating," in *Proc. 15th Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, Nov. 2019, pp. 551–556.

[27] J. Jossinet, "Variability of impedivity in normal and pathological breast tissue," *Med. Biol. Eng. Comput.*, vol. 34, no. 5, pp. 346–350, Sep. 1996.

[28] P. K. Grewal and F. Golnaraghi, "Pilot study: Electrical impedance based tissue classification using support vector machine classifier," *IET Sci., Meas. Technol.*, vol. 8, no. 6, pp. 579–587, 2014.

[29] A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," in *Proc. 5th Int. Symp. Health Informat. Bioinf.*, Apr. 2010, pp. 114–120.

[30] M. Narendra, "Breast cancer detection using histology images: A survey," *J. Adv. Res. Dyn. Control Syst.*, vol. 12, no. SP7, pp. 561–565, Jul. 2020.

[31] H. D. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern Recognit.*, vol. 43, no. 1, pp. 299–317, Jan. 2010.

[32] M. A. Cifci and Z. Aslan, "Deep learning algorithms for diagnosis of breast cancer with maximum likelihood estimation," in *Proc. Int. Conf. Comput. Sci. Appl.* Springer, 2020, pp. 486–502.

[33] A.-A. Nahid and Y. Kong, "Involvement of machine learning for breast cancer image classification: A survey," *Comput. Math. Methods Med.*, vol. 2017, pp. 1–29, Dec. 2017.

[34] M. Benndorf, E. Kotter, M. Langer, C. Herda, Y. Wu, and E. S. Burnside, "Development of an online, publicly accessible naive Bayesian decision support tool for mammographic mass lesions based on the American college of radiology (ACR) BI-RADS lexicon," *Eur. Radiol.*, vol. 25, no. 6, pp. 1768–1775, Jun. 2015.

[35] Y. Chen, L. Ling, and Q. Huang, "Classification of breast tumors in ultrasound using biclustering mining and neural network," in *Proc. 9th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2016, pp. 1787–1791.

[36] S. M. Castro, E. Tseytlin, O. Medvedeva, K. Mitchell, S. Visweswaran, T. Bekhuis, and R. S. Jacobson, "Automated annotation and classification of BI-RADS assessment from radiology reports," *J. Biomed. Informat.*, vol. 69, pp. 177–187, May 2017.

[37] B. Percha, H. Nassif, J. Lipson, E. Burnside, and D. Rubin, "Automatic classification of mammography reports by BI-RADS breast tissue composition class," *J. Amer. Med. Inform. Assoc.*, vol. 19, no. 5, pp. 913–916, Sep. 2012.

[38] M. Boroumandzadeh and E. Parvinnia, "Automated classification of BI-RADS in textual mammography reports," *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 29, no. 2, pp. 632–647, Mar. 2021.

[39] A. Oliver, J. Freixenet, and R. Zwiggelaar, "Automatic classification of breast density," in *Proc. IEEE Int. Conf. Image Process.*, vol. 2, Sep. 2005, p. 1258.

[40] F. Chokri and M. H. Farida, "Mammographic mass classification according to Bi-RADS lexicon," *IET Comput. Vis.*, vol. 11, no. 3, pp. 189–198, Apr. 2017.

[41] N. Khan, K. Wang, A. Chan, and R. Highnam, "Automatic BI-RADS classification of mammograms," in *Image and Video Technology*, T. Bräunl, B. McCane, M. Rivera, and X. Yu, Eds. Springer, 2016, pp. 475–487.

[42] M. Yasmin, M. Sharif, and S. Mohsin, "Survey paper on diagnosis of breast cancer using image processing techniques," *Res. J. Recent Sci.*, vol. 2, pp. 88–98, 2013.
[43] R. Ramani, S. Suthanthiravanitha, and S. Valarmathy, "A survey of current image segmentation techniques for detection of breast cancer," *Int. J. Eng. Res. Appl.*, vol. 2, no. 5, pp. 1124–1129, 2012.
[44] M. Korotkova, A. Karpov, M. Machin, Y. Tsofin, V. Tsyplyonkov, and A. Tchayev, "Electric impedance imaging of the mammary gland in circumstances of skin abnormality or damage," in *World Congress on Medical Physics and Biomedical Engineering*. Springer: Munich, Germany, Sep. 2009, pp. 284–287.
[45] O. Trokhanova, A. Karpov, V. Cherepenin, A. Korjenevsky, V. Kornienko, Y. Kultiasov, and V. Marushkov, "Electro-impedance mammography testing at some physiological woman's periods," in *Proc. 11th Int. Conf. Electr. Bio-Impedance (Oslo)*, Jun. 2001, pp. 461–465.
[46] S. P. Helmrich, S. Shapiro, L. Rosenberg, D. W. Kaufman, D. Slone, C. Bain, O. S. Miettinen, P. D. Stolley, N. B. Rosenshein, R. C. Knapp, and T. H. Leavitt, Jr., "Risk factors for breast cancer," *Amer. J. Epidemiol.*, vol. 117, no. 1, pp. 35–45, 1983.
[47] Z. Momenimovahed and H. Salehiniya, "Epidemiological characteristics of and risk factors for breast cancer in the world," *Breast Cancer, Targets Therapy*, vol. 11, p. 151, Apr. 2019.
[48] J. L. Kelsey, M. D. Gammon, and E. M. John, "Reproductive factors and breast cancer," *Epidemiol. Rev.*, vol. 15, no. 1, p. 36, 1993.
[49] X. R. Yang, J. Chang-Claude, E. L. Goode, F. J. Couch, H. Nevanlinna, R. L. Milne, M. Gaudet, M. K. Schmidt, A. Broeks, A. Cox, and P. A. Fasching, "Associations of breast cancer risk factors with tumor subtypes: A pooled analysis from the breast cancer association consortium studies," *J. Nat. Cancer Inst.*, vol. 103, no. 3, pp. 250–263, 2011.
[50] L. Bernstein, "Epidemiology of endocrine-related risk factors for breast cancer," *J. Mammary Gland Biol. Neoplasia*, vol. 7, no. 1, pp. 3–15, 2002.
[51] M. C. Pike, B. E. Henderson, J. T. Casagrande, I. Rosario, and G. E. Gray, "Oral contraceptive use and early abortion as risk factors for breast cancer in young women," *Brit. J. Cancer*, vol. 43, no. 1, pp. 72–76, Jan. 1981.
[52] S. E. Singletary, "Rating the risk factors for breast cancer," *Ann. Surg.*, vol. 237, no. 4, p. 474, 2003.
[53] B. Murillo-Ortiz, A. Hernández-Ramírez, T. Rivera-Villanueva, D. Suárez-García, M. Murguía-Pérez, S. Martínez-Garza, A. Rodríguez-Penin, R. Romero-Coripuna, and X. M. López-Partida, "Monofrequency electrical impedance mammography (EIM) diagnostic system in breast cancer screening," *BMC Cancer*, vol. 20, no. 1, pp. 1–10, Dec. 2020.
[54] B. Murillo-Ortiz, A. Rodríguez-Penin, A. Hernández-Ramírez, T. Rivera-Villanueva, A. E. Moran-Gonzalez, S. Martínez-Garza, D. Suárez-García, M. Pérez-Murguía, and R. Romero-Coripuna, "Diagnóstico de Cáncer de mama mediante Mamografía por electroimpedancia computarizada MEIK," *Mastología*, vol. 9, no. 1, pp. 20–28, 2019.
[55] A. Karpov, M. Korotkova, G. Shiferson, and E. Kotomina, "Electrical impedance mammography: Screening and basic principles," in *Breast Cancer Breast Reconstruction*, L. Tejedor, S. G. Modet, L. Manchev, and A. A. Parikesit, Eds. Rijeka, Croatia: IntechOpen, 2020, Ch. 1.
[56] N. M. Zain and K. K. Chelliah, "Breast imaging using electrical impedance tomography: Correlation of quantitative assessment with visual interpretation," *Asian Pacific J. Cancer Prevention*, vol. 15, no. 3, pp. 1327–1331, Feb. 2014.
[57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[58] A. F. Hilario, S. G. López, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning From Imbalanced Data Sets*. Springer, 2018.
[59] S. Fotouhi, S. Asadi, and M. W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *J. Biomed. Informat.*, vol. 90, Feb. 2019, Art. no. 103089.
[60] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[61] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[62] G. Lemaître, F. Nogueira, and K. C. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.

ROSARIO LISSIET ROMERO CORIPUNA was born in Cabanaconde, Caylloma, Arequipa, Perú, in 1993. She received the B.Sc. degree from the Universidad Nacional de San Agustin, Arequipa, in 2014, and the M.Sc. degree in physics from the Universidad de Guanajuato, León, Mexico, in 2017, where she is currently pursuing the Ph.D. degree in physics. Her main research interests include medical-physics, medical images, data analysis, machine learning, simulation, and numerical modeling aimed at achieving results in the medical area.

DELIA IRAZÚ HERNÁNDEZ FARÍAS was born in León, Guanajuato, Mexico, in 1987. She received the B.Sc. and M.Sc. degrees in computer science from the Leon Institute of Technology and the Ph.D. degree in computer sciences from the Universitat Politècnica de València, Spain, and the University of Turin, Italia, in 2017. She is currently a part-time Professor at the División de Ciencias e Ingenierías Campus León, Universidad de Guanajuato, Mexico. Her main research interests include machine learning, natural language processing related tasks, pattern recognition, medical image processing, natural language processing, and machine learning algorithms.

BLANCA OLIVIA MURILLO ORTIZ was born in León, Guanajuato, Mexico, in 1968. She received the B.Med. degree in medicine from the University of Guanajuato, Mexico, in 1998, and the M.Sc. and Ph.D. degrees in medical sciences from the Medical Science Institute, University of Guanajuato, in 2000 and 2004, respectively. She has authored research works on telomeres and telomerase in breast cancer. She is particularly interested in non-invasive biomedical techniques for early diagnostic the breast cancer and novel electroimpedance techniques and their applications.

LUIS CARLOS PADIERNA was born in Dolores Hidalgo, Guanajuato, Mexico, in 1985. He received the B.Eng. degree in computer systems engineering from the Celaya Institute of Technology, in 2009, the M.Sc. degree in computer science from the León Institute of Technology, in 2011, and the Ph.D. degree in computer science from the Tijuana Institute of Technology, in 2018. He is currently an Associate Professor at the University of Guanajuato. He has authored research works on the designing of support vector machines. His main research interests include machine learning, pattern recognition, evolutionary algorithms, and deep learning.

TEODORO CÓRDOVA FRAGA was born in La Purísima, Pinos, Zacatecas, Mexico, in 1967. He received the B.Sc. degree from the Universidad Autónoma de San Luis Potosi, in 1997, and the M.Sc. and Ph.D. degrees in physics from the University of Guanajuato, in 2001 and 2003, respectively. He did his Ph.D. degree with a research stage of a year at Sao Paulo University, Brazil, from September 2002 to August 2003. He spent a year (from August 2004 to September 2005) as a Research Fellow at Vanderbilt University, Nashville, TN, USA. He is currently a Titular Professor at the University of Guanajuato, Campus León. He has authored research works on MRI, US, signal processing, digital image processing, biomagnetism, Raman spectroscopy, and biomedical instrumentation. He is the owner of five patents. His main research interests include bioinstrumentation, data analysis, medical images, and bioelectromagnetism registers.