

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,700

Open access books available

182,000

International authors and editors

195M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



Chapter

# Automatic BI-RADS Classification of Breast Magnetic Resonance Medical Records Using Transformer-Based Models for Brazilian Portuguese

*Ricardo de Oliveira, Bruno Menezes, Júnia Ortiz  
and Erick Nascimento*

## Abstract

This chapter aims to present a classification model for categorizing textual clinical records of breast magnetic resonance imaging, based on lexical, syntactic and semantic analysis of clinical reports according to the Breast Imaging-Reporting and Data System (BI-RADS) classification, using Deep Learning and Natural Language Processing (NLP). The model was developed from transfer learning based on the pre-trained BERTimbau model, BERT model (Bidirectional Encoder Representations from Transformers) trained in Brazilian Portuguese. The dataset is composed of medical reports in Brazilian Portuguese classified into six categories: Inconclusive; Normal or Negative; Certainly Benign Findings; Probably Benign Findings; Suspicious Findings; High Risk of Cancer; Previously Known Malignant Injury. The following models were implemented and compared: Random Forest, SVM, Naïve Bayes, BERTimbau with and without finetuning. The BERTimbau model presented better results, with better performance after finetuning.

**Keywords:** BI-RADS, deep learning, transformers, BERTimbau, Portuguese NLP

## 1. Introduction

The healthcare sector is characterized as a large data catalyst, contained in medical records, reports, test results and so on. In the case of textual medical records, the correct classification of unstructured parts of the texts incorporated into medical documents can support healthcare professionals for managing relevant data effectively and efficiently, organizing the data related to patients and their findings in diagnostic tests.

This work presents a classification system for categorization of BI-RADS [1], based on lexical, syntactic, and semantic analysis of documents, derived from textual

clinical records, using Deep Learning and NLP (Natural Language Processing). The main goal is to verify the performance of BERTimbau model [2] in BI-RADS categories classification from breast magnetic resonance imaging clinical records. Machine learning models were also used to classify BI-RADS, in order to establish a baseline. The following models were implemented and compared: Random Forest, SVM, Naïve Bayes, BERTimbau with and without finetuning.

After submitting a dataset containing 8813 records of medical texts to deep learning training, a new expert model based on existing rules was created for automated BI-RADS category classification in breast MRI reports, using a supervised machine learning approach. In addition to being able to classify medical texts related to breast MRIs to their corresponding BI-RADS, the model will be able to inform about the quality of the medical record, in relation to pre-existing statistics.

## 2. Materials and method

### 2.1 BI-RADS

BI-RADS [1], is an acronym for Breast Imaging-Reporting and Data System, a quality assurance tool originally designed for using in mammography. The system is a collaborative effort of many health care groups but is published and copyrighted by the American College of Radiology (ACR).<sup>1</sup> The system was designed to standardize clinical reporting and is used by medical professionals to communicate a patient's risk of developing breast cancer, particularly for patients with dense breast tissue. The document focuses on patient reports used by medical professionals. The six classification categories of the American College of Radiology are described below.

#### 2.1.1 BI-RADS category 0 - inconclusive

When the radiologist classifies the result as BI-RADS 0 [1], it means that the examination was considered inconclusive or incomplete. Causes for a category 0 include technical factors, such as poor image quality, which may be due to improper breast positioning or patient movement during the exam. Category 0 can also be assigned when there is doubt about the existence or not of an injury, requiring another imaging exam to take the test.

#### 2.1.2 BI-RADS category 1: Normal or negative

When the radiologist classifies the result as BI-RADS 1 [1], it means that no alteration was presented. The exam is completely normal. The breasts are symmetrical and do not present masses, architectural distortions or suspicious calcifications. The risk of malignant lesion in an exam classified as category 1 is 0%.

#### 2.1.3 BI-RADS category 2: Certainly benign findings

When the radiologist classifies the result as BI-RADS 2 [1], it means that some alteration was found in the images, but that the characteristics of the lesion allow us to state that it is benign. To be classified as category 2, the physician needs to be

---

<sup>1</sup> [www.acr.org](http://www.acr.org).

confident in stating that the lesion is of benign origin. If the physician is in doubt, the result cannot be classified as BI-RADS 2, but as BI-RADS 3. Therefore, in practice, a BI-RADS 2 result has the same clinical value as a BI-RADS 1. The risk of malignant lesion is 0%.

#### *2.1.4 BI-RADS category 3 - probably benign findings*

When the radiologist classifies the result as BI-RADS [1], it means that some alteration was found in the images, which is probably benign, but which is not 100% safe. As much as the doctor is almost sure that the lesion is benign, if he has the slightest doubt, the classification should be category 3. Therefore, a result in category 3 indicates a lesion with very low risk of malignancy, which does not need to be biopsied initially, but which, as a precaution, should be followed closely over the next 2 years. The risk of malignant lesions in BI-RADS 3 is only 2%, that is, 98% of cases are actually benign lesions.

#### *2.1.5 BI-RADS category 4 - suspicious findings*

When the radiologist classifies the result as BI-RADS 4 [1], it means that some alteration was found in the images, which may be cancer, but which is not necessarily cancer. All patients with a BI-RADS 4 result should undergo biopsy of the lesion so that the correct diagnosis can be established. Category 4 is usually divided into 3 subcategories according to cancer risk:

- BI-RADS 4A – Lesion with low suspicion of malignancy – 2 to 10% risk of cancer.
- BI-RADS 4B – Lesion with moderate suspicion of malignancy – 11 to 50% risk of cancer.
- BI-RADS 4C – Lesion with high suspicion of malignancy – 51 to 95% risk of cancer.

Regardless of the BI-RADS 4 subcategory, all cases should undergo biopsy. The difference is that in the patient with BI-RADS 4A, the biopsy is expected to confirm a benign lesion, while in the BI-RADS 4C, the biopsy is expected to confirm the diagnosis of cancer.

#### *2.1.6 BI-RADS category 5 - high cancer risk*

When the radiologist classifies the result as BI-RADS 5 [1], it means that some alteration was found in the images, which almost certainly is derived from breast cancer. Breast lesions with typical features of cancer include dense, spiculated nodules, pleomorphic calcifications, lesions with skin retraction or distortions of breast architecture, or fine linear calcifications arranged in a segment of the breast. Thus, all category 5 lesions should be biopsied and the risk of malignancy in a BI-RADS 5 classification is greater than 95%.

#### *2.1.7 BI-RADS category 6: Previously known malignant lesion*

The BI-RADS 6 classification [1] is only used in patients who already have a diagnosis of breast cancer established and end up undergoing a diagnostic imaging

exam to monitor the disease, for example, after the onset of chemotherapy. This classification serves only to confirm to the physician that the malignant lesion identified in the mammogram is the same previously known.

## 2.2 Dataset

For this study, 8813 instances of reports issued by a radiology service, fully anonymized, for breast MRIs, comprised between April 2016 and December 2021 were collected. For bilateral breast MRIs, 7360 instances, representing 83.51% of the total number of instances; for resonances of left breasts, 750 instances, representing 8.51%; for MRIs of right breasts, 657 instances, representing 7.45%; and for breast MRIs using the mamotomy technique, 46 instances, representing 0.52%.

The medical record with the highest number of words had a value of 484; the smallest, 130. The average number of words found was 202.

Breast MRI scans grade BI-RADS if indicated. With that, due to the standardization existing in the instances, there is a BI-RADS classification at the conclusion of medical reports. To extract this information and the population of a specific variable, the `loc` method was used, combined with the `str.contains` function to extract keywords related to the BI-RADS categories contained in the medical records.

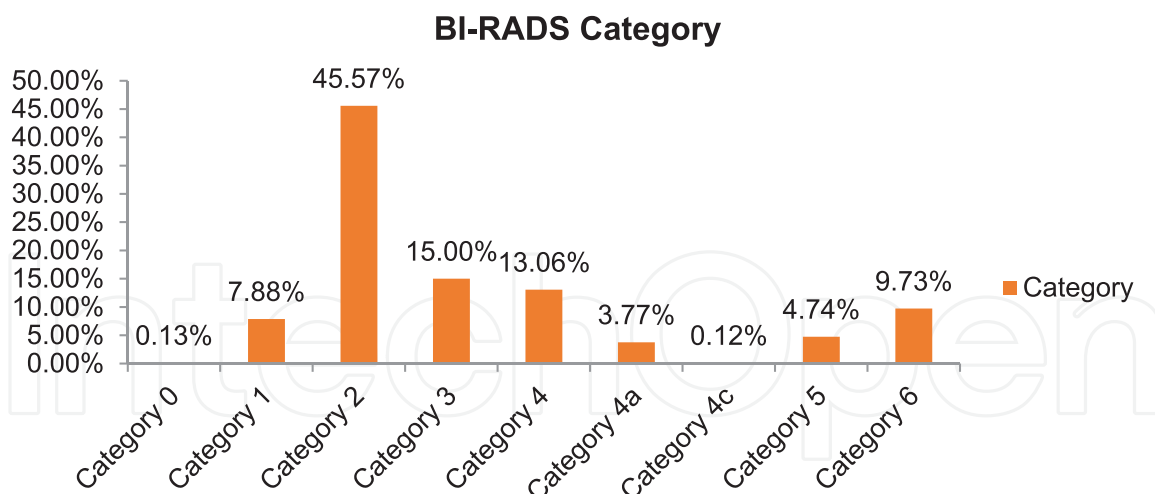
For the algorithms to work more efficiently, a new variable was created, containing the mapped information, but numerically. The variable is then represented like this:

BI-RADS by Category (numeric variable (Category\_Code)):

- 2, 3387 instances;
- 3, 1115 instances;
- 4, 971 instances;
- 6, 723 instances;
- 1586;
- 5, 352 instances;
- 4A, 280 instances;
- 0, 10 instances;
- 4C, 9 instances.

The data are naturally “unbalanced”, as this reflects what is actually found in a population that undergoes this type of technique for diagnosing breast cancer, which is an expected behavior for such a set and according to the classes observed.

Some of the most common indications for magnetic resonance imaging of the breasts are clarification of inconclusive findings on mammography and/or ultrasound, as well as tracking high-risk patients, not being indicated for initial investigation, as are, for example, diagnostic exams via mammography (**Figure 1**).



**Figure 1.**  
Percentage of records by class.

## 2.3 Models

The Transformer architecture (Vaswani et al.) is a natural language processing neural network architecture developed by Google in 2017. It was introduced in the paper “Attention Is All You Need” [3] and has revolutionized the way neural networks are trained to handle language processing tasks such as automatic translation and text generation. The main innovation of the Transformer architecture is the use of attention, which allows the network to consider all input words simultaneously when producing an output. This helps to deal with the variable length dependency problem present in many natural language processing tasks. In addition, the Transformer architecture uses multi-header layers of attention, which helps extend the network’s modeling capability.

BERT (Bidirectional Encoder Representations from Transformers) [4] is a language pre-training technique developed by Google in 2018. It uses the Transformer architecture to learn bidirectional representations of each word in a text corpus. This means that, unlike other pre-training techniques that only consider the left or right context of each word, BERT considers the left and right context of each word simultaneously. This allows the model to learn richer and more accurate representations of the words. BERT was trained on a large amount of text from the internet and can be easily adapted to various natural language processing tasks such as text classification, entity extraction and question-answering. It has shown excellent results in many natural language processing tasks and has become a basis for many other language models.

Language model pre-training has been shown to be effective in improving many tasks related to natural language processing [5]. This includes sentence-level tasks such as natural language inference [6], which aim to predict relationships between sentences by analyzing them holistically [7], as well as token-level tasks, such as named entity recognition and answering queries, where models are needed to produce token-level output [8].

In this study, data were submitted to a neural network algorithm called BERTimbau [2], for natural language processing (NLP) in Portuguese, a variation of the BERT algorithm [9]. Random Forest, Support Vector Machine (SVM) and Naïve Bayes machine learning algorithms were also used in order to create a baseline.

Machine learning [10] is a sub-area of artificial intelligence that has shown enormous growth in recent decades. These are mathematical, statistical and computational algorithms that are capable of carrying out an inference process through example-based learning.

Random Forest [11] is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for classification and regression problems in machine learning. It is based on the concept of ensemble learning, which is a process of combining several classifiers to solve a complex problem and improve model performance.

SVM [12] is one of the most popular supervised learning algorithms used for classification and regression problems. However, it is primarily used for classification problems in machine learning.

The Naïve Bayes algorithm [13] is a supervised learning algorithm based on Bayes' theorem and used to solve classification problems. It is primarily used in classifying text that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simplest and most effective classification algorithms that helps in building fast machine learning models that can make fast predictions.

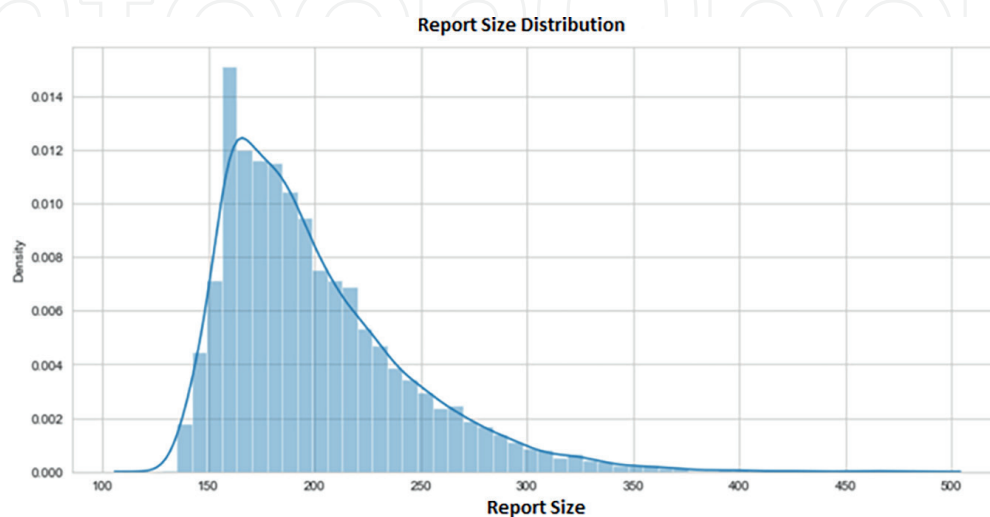
### 3. Results

#### 3.1 Exploratory analysis

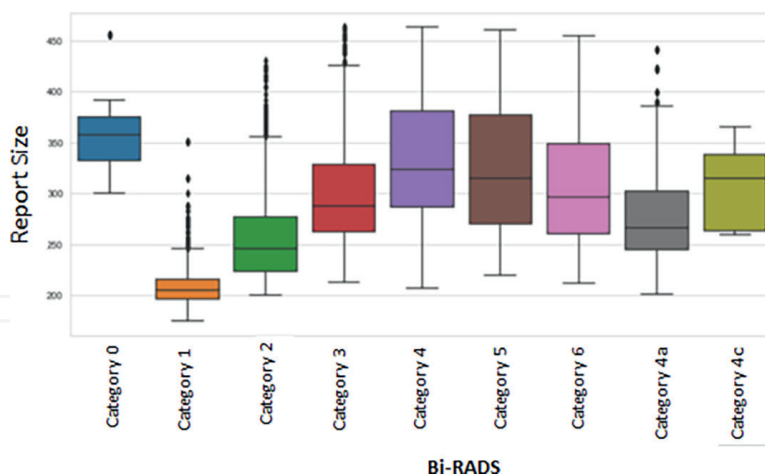
With the dataset still having its original characteristics, in terms of the variable that stores the medical records, the size distribution (number of words per medical record) – see **Figure 2**.

The distribution of the number of existing words per document for each category in the original dataset is presented in **Figure 3**.

In order to clean the data and to improve computational performance, pre-processing techniques were applied such as lowercase, besides of \r, \n, punctuation and stopwords (for Portuguese) removal. The experiments in this work were performed using the dataset in its original characteristic.



**Figure 2.**  
*Report size distribution. Original dataset.*



**Figure 3.**  
Boxplot of words in reports. Original dataset.

### 3.2 Classification models

The original dataset was submitted to three machine learning algorithms (Random Forest, SVM and Naïve Bayes) and a deep learning algorithm (BERTimbau).

The metrics applied to verify the performance of the models were:

*Precision:* The ability of a classification model to identify only relevant data points. Mathematically, precision is the number of true positives (VP) divided by the number of true positives (VP) plus the number of false positives (FN):  $VP / (VP + FN)$ ; *Recall:* The ability of a model to find all relevant cases in a dataset. Mathematically, recall is defined as the number of true positives (VP) divided by the number of true positives (VP) plus the number of false negatives (FN):  $VP / (VP + FN)$ ; *F1-score:* is defined as the harmonic mean of precision (P) and recall (S). The harmonic mean is an alternative metric to the more common arithmetic mean. It is often useful when calculating an average rate:  $2 \times (P \times S) / (P + S)$ ; *Accuracy:* is the number of data points correctly predicted from all data points. More formally, it is defined as the number of true positives (VP) and true negatives (VN) divided by the number of true positives (VP), true negatives (VN), false positives (FP) and false negatives (FN):  $(VP + VN) / N$ .

It is important to note that when submitting the dataset to an attribute selection technique, for the machine learning models, categories 0 and 43 were excluded, as their number of instances were inexpressive for the performance of the models. For the Random Forest algorithm, Randomized Search Cross Validation and Grid Search Cross Validation techniques were applied.

The best hyperparameters found with Random Search were:

Bootstrap = False. Method for sampling data points (with or without replacement); max\_depth = 30. The max\_depth of a tree in Random Forest is defined as the longest path between the root node and the leaf node; max\_features = sqrt. This is similar to the maximum number of resources given to each tree in a random forest; min\_samples\_leaf = 1. Specifies the minimum number of samples that must be present in the leaf node after splitting a node; min\_samples\_split = 5. Parameter that tells the decision tree in a random forest the minimum number of observations needed at any node to split it; n\_estimators = 800. Number of trees in the forest.

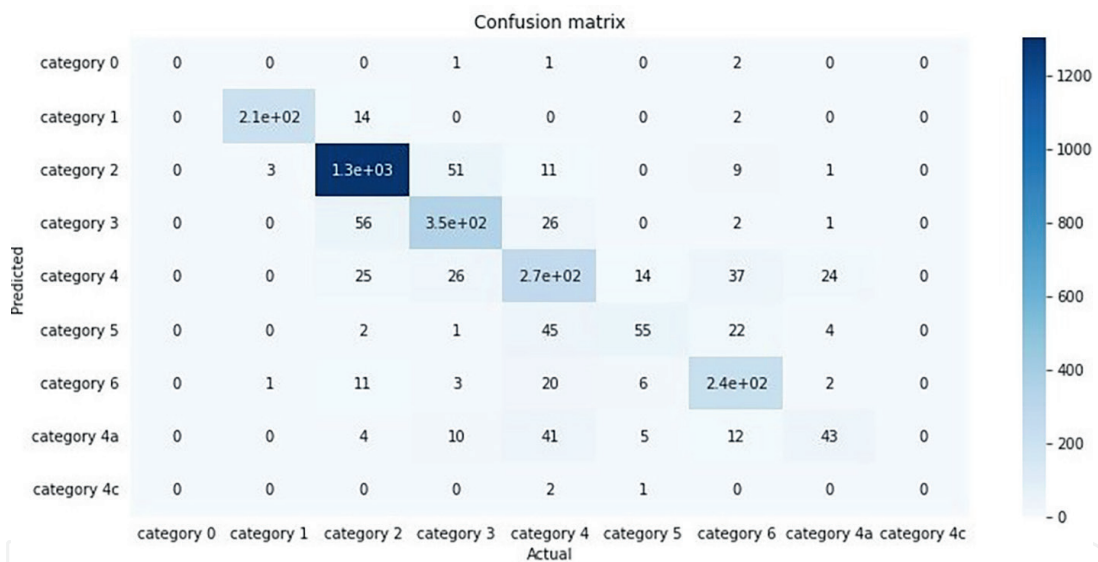
**Table 1** presents the results found for Random Forest after training.

The confusion matrix, which shows the classification frequencies for each class in the model, for the Random Forest model results, is presented in **Figure 4**.



Random Forest				
Class	Precision	Recall	F1-Score	Support
1	0.96	0.92	0.94	84
2	0.92	0.96	0.94	533
3	0,87	0.8	0.83	157
4	0.79	0.82	0.81	145
5	0.59	0,22	0.32	46
6	0.77	0.9	0.83	115
41	0.85	0.83	0.84	35
Accuracy			0.88	1115
Macro AVG	0.82	0.78	0.79	1115
Wighted AVG	0.87	0.88	0.87	1115

**Table 1.**  
Random Forest results.



**Figure 4.**  
Random Forest confusion matrix.

For the SVM algorithm, the Randomized Search Cross Validation technique was applied. The best hyperparameters found with Random Search were:

Probability = True, enable probability estimates; Kernel = poly, specifying the kernel type (in this case, polynomial) to be used in the algorithm; Gamma = 10, kernel coefficient for what was specified in hyperparameter Kernel = poly; Degree = 4, Degree of polynomial kernel function (poly); C = 0.01, being the regularization parameter. The strength of the regularization is inversely proportional to C. It must be strictly positive. The penalty is a l2 squared penalty.

**Table 2** shows the results found for the SVM after training.

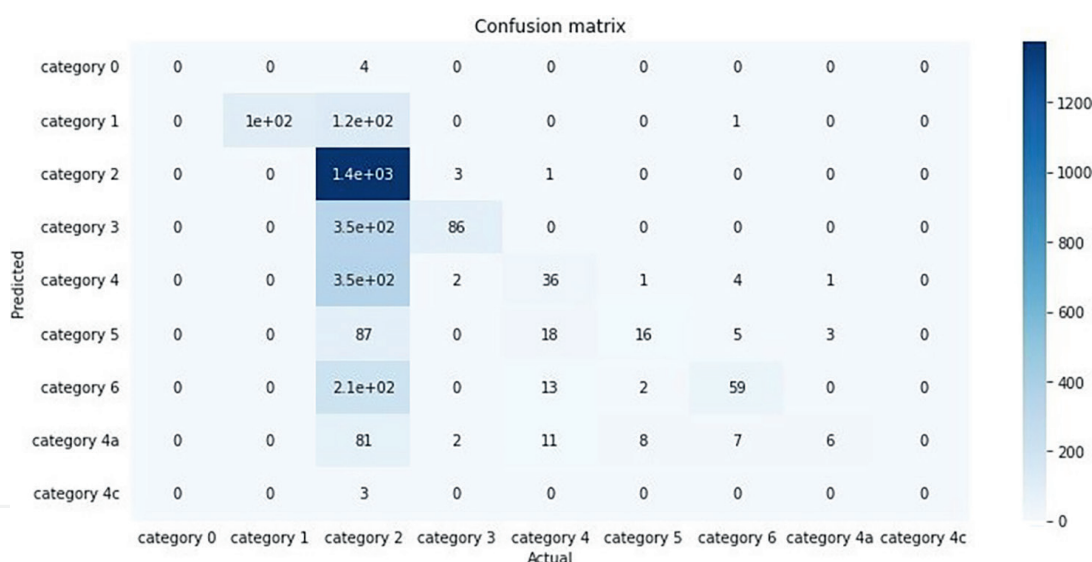
The confusion matrix for the SVM-based model is presented in **Figure 5**.

The values found after training the Naïve Bayes model are presented in **Table 3**.

**Figure 6** presents the confusion matrix for Naive Bayes-based model results.

SVM				
Class	Precision	Recall	F1-Score	Support
1	1	0.6	0.75	84
2	0.64	0.99	0.78	533
3	0.95	0.39	0.56	157
4	0.56	0.35	0.43	145
5	0.46	0.13	0.2	46
6	0.84	0.37	0.52	115
41	0.67	0.29	0.4	35
Accuracy			0.67	1115
Macro AVG	0.73	0.45	0.52	1115
Wighted AVG	0.71	0.67	0.64	1115

**Table 2.**  
SVM results.



**Figure 5.**  
SVM confusion matrix.

**Table 4** presents a summary of the machine learning models results for comparison, which shows that Random Forest algorithm presented the best result.

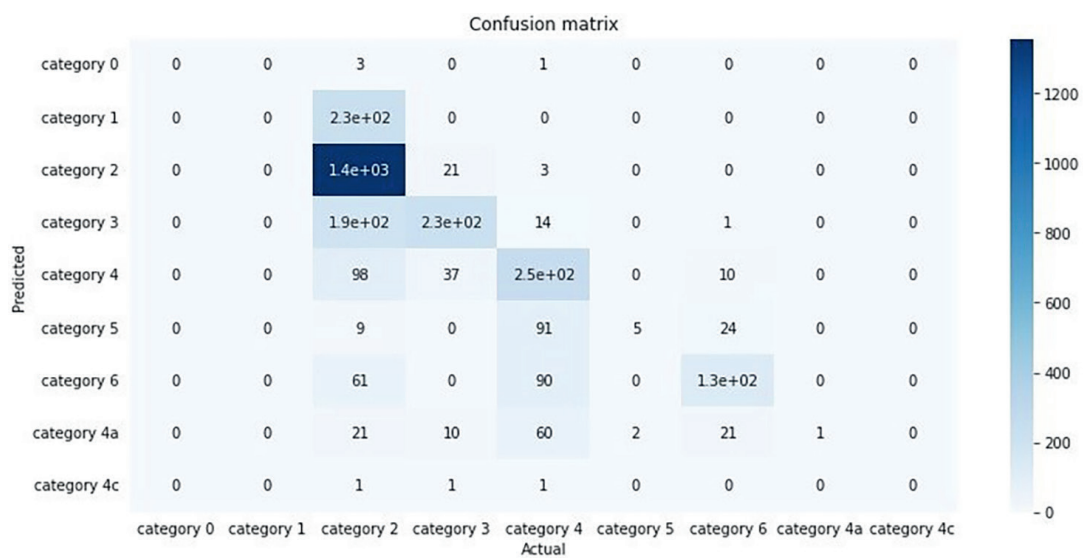
To submit the dataset to the BERTimbau algorithm, the One-Hot Encoding technique was adopted, transforming the categorical variables into binary ones, with a one-hot encoding being a representation of categorical variables as binary vectors. Specific test steps were applied, with and without finetuning. **Table 5** presents values found after submitting the dataset to four epochs training.

The optimizer used was AdamW with the following parameters: optimizer = AdamW(optimizer\_grouped\_parameters, lr = 2e-5, correct\_bias = True).

The custom optimization parameters were 'params' with the following rule [p for n, p in param\_optimizer if not any(nd in n for nd in no\_decay)] being the value for no\_decay equal to ['bias', 'gamma', 'beta'], which is an iterable thing of parameters to optimize

Naive Bayes				
Class	Precision	Recall	F1-Score	Support
1	0.88	0.27	0.42	84
2	0.8	0.98	0.88	533
3	0.81	0.61	0.69	157
4	0.56	0.78	0.65	145
5	0.5	0.04	0.08	46
6	0.75	0.65	0.7	115
41	0.75	0.34	0.47	35
Accuracy			0.75	1115
Macro AVG	0.72	0.52	0.56	1115
Wighted AVG	0.76	0.75	0.72	1115

**Table 3.**  
Naive Bayes results.



**Figure 6.**  
Naive Bayes confusion matrix.

Test set scores		
Model	F1	Accuracy
Random Forest	0.787	0.876
Naive Bayes	0.556	0.753
SVM	0.519	0,674

**Table 4.**  
Summary of machine learning models results.

or dictionaries that define groups of parameters; 'weight\_decay\_rate' with value 0.01 which is the decoupled weight decay to apply or 'params' with the rule [p for n, p in param\_optimizer if any(nd in n for nd in no\_decay)] and 'weight\_decay\_rate': 0.0.

BERTimbau				
Class	Precision	Recall	F1-Score	Support
birads0	0	0	0	10
birads1	1	0.97	0.98	586
birads2	0.99	0.98	0.99	3387
birads3	0.95	0.97	0.96	1115
birads4	0.86	0.95	0.9	971
birads4a	1	0.7	0.83	280
birads4c	0	0	0	9
birads5	0.51	0.79	0.62	352
birads6	0.93	0.55	0.7	723
Micro AVG	0.93	0.91	0.92	7433
Macro AVG	0.69	0.66	0.66	7433
Weighted AVG	0.94	0.91	0.92	7433
Samples AVG	0.91	0.91	0.91	7433
Test F1 Accuracy	0.92			
Test Flat Accuracy	0.91			

**Table 5.**  
*BERTimbau results.*

Applying a finetuning, with the aim of enriching the vocabulary of BERTimbau and thus creating both a new specialist model in the area in question and also a specific tokenizer, 1819 new tokens were added. After training in four epochs, the new model was created, expressing a perplexity at a value of 2.17. Perplexity is a measure of how well a probability distribution or probability model predicts a sample. Can be used to compare probability models. A low perplexity indicates that the probability distribution is good at predicting the sample.

The values found using the created expert model, are presented in **Table 6**.

In general, BERTimbau model presented better results compared to machine learning algorithms. **Figure 7** presents the comparative values between BERTimbau model stages.

By observing the values shown in the table above, it is clearly seen that in the vast majority of situations in which the classes were present, the performance of the adjusted model was better than all previously tested models.

## 4. Conclusions

The Transformer architecture has become the dominant architecture for natural language processing, frequently outperforming models such as convolutional neural networks and recurrent networks in different tasks [14]. Pre-trained models are able to be trained on generic or specialist sets and, consequently, they are easily adapted to tasks with excellent performance. The architecture is particularly conducive to large corpora pre-training, providing accuracy increase in later tasks, such as text classification, language comprehension, and more.

BERTimbau				
Class	Precision	Recall	F1-Score	Support
birads0	0	0	0	10
birads1	1	0.98	0.99	586
birads2	1	0.99	0.99	3387
birads3	0.98	0.99	0.98	1115
birads4	0.95	0.98	0.97	971
birads4a	0.95	0.96	0.96	280
birads4c	0	0	0	9
birads5	0.95	0.8	0.87	352
birads6	0.93	0.96	0.95	723
Micro AVG	0.98	0.97	0.98	7433
Macro AVG	0.75	0.74	0.75	7433
Weighted AVG	0.97	0.97	0.97	7433
Samples AVG	0.97	0.97	0.97	7433
Test F1 Accuracy	0.98			
Test Flat Accuracy	0.97			

**Table 6.**  
Fine-tuned BERTimbau results.

	1 - ORIGINAL BERTIMBAU MODEL					2 - POST FINE TUNING MODEL					FINE TUNING		
	BERTimbau Tokenizer and Model					BIRADS Tokenizer and Model					1 versus 2		
	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score
birads0	0	0	0	10	birads0	0	0	0	10	birads0	EQUAL	EQUAL	EQUAL
birads1	1	0.97	0.98	586	birads1	1	0.98	0.99	586	birads1	EQUAL	BEST	BEST
birads2	0.99	0.98	0.99	3387	birads2	1	0.99	0.99	3387	birads2	BEST	BEST	EQUAL
birads3	0.95	0.97	0.96	1115	birads3	0.98	0.99	0.98	1115	birads3	BEST	BEST	BEST
birads4	0.86	0.95	0.9	971	birads4	0.95	0.98	0.97	971	birads4	BEST	BEST	BEST
birads4a	1	0.7	0.83	280	birads4a	0.95	0.96	0.96	280	birads4a	WORSE	BEST	BEST
birads4c	0	0	0	9	birads4c	0	0	0	9	birads4c	EQUAL	EQUAL	EQUAL
birads5	0.51	0.79	0.62	352	birads5	0.95	0.8	0.87	352	birads5	BEST	BEST	BEST
birads6	0.93	0.55	0.7	723	birads6	0.93	0.96	0.95	723	birads6	EQUAL	BEST	BEST
micro avg	0.93	0.91	0.92	7433	micro avg	0.98	0.97	0.98	7433	micro avg	BEST	BEST	BEST
macro avg	0.69	0.66	0.66	7433	macro avg	0.75	0.74	0.75	7433	macro avg	BEST	BEST	BEST
weighted avg	0.94	0.91	0.92	7433	weighted avg	0.97	0.97	0.97	7433	weighted avg	BEST	BEST	BEST
samples avg	0.91	0.91	0.91	7433	samples avg	0.97	0.97	0.97	7433	samples avg	BEST	BEST	BEST

**Figure 7.**  
Comparison between BERTimbau and fine-tuned BERTimbau.

The idea of using machine learning to classify texts in this work with supervised approach is to develop a classification model based on an initial set of labeled texts, using the reached values as baseline for the project.

BERT is undoubtedly a breakthrough in using deep learning for natural language processing. The progress is very significant when it comes to the Portuguese language. The accessibility and fast fine-tuning provide a wide range of practical applications, including using the generated model itself as a basis for creating specialist models in

health area, for example. In the case of this study, fine-tuned BERTimbau managed to capture specific information for a generalist area, increasing its vocabulary and becoming a good model for classifying medical records data, structuring data which is normally unstructured.

### **Conflict of interest**

The authors declare no conflict of interest.

### **Author details**

Ricardo de Oliveira<sup>1</sup>, Bruno Menezes<sup>1\*</sup>, Júnia Ortiz<sup>1</sup> and Erick Nascimento<sup>2</sup>


1 Senai Cimatec, Salvador, Brazil

2 University of Surrey, Surrey, UK

\*Address all correspondence to: [bruno.menezes@fieb.org.br](mailto:bruno.menezes@fieb.org.br)

### **IntechOpen**

---

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Castro S, M, Tseytlin E, Medvedeva O, Mitchell K, Visweswaran S, Bekhuis T, et al. Automated annotation and classification of BI-RADS assessment from radiology reports. *Journal of Biomedical Informatics*. May 2017;**69**:177-187
- [2] Souza F, Nogueira R, Lotufo R. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems. Rio Grande do Sul, Brazil, October 20-23 (to appear). [S.l.: s.n.]: BRACIS; 2020
- [3] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Available from: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
- [4] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. 2019. Available from: <https://arxiv.org/abs/1810.04805>
- [5] Dai Andrew M, Le Quoc V. Semi-Supervised Sequence Learning. Available from: <https://arxiv.org/abs/1511.01432>. 2015
- [6] Bowman SR, Angeli G, Potts C, Manning CD. A Large Annotated Corpus for Learning Natural Language Inference. Available from: <https://arxiv.org/abs/1508.05326>
- [7] Dolan W, B, Brockett C. Automatically constructing a corpus of sentential paraphrases. In: Proceedings of the Third International Workshop on Paraphrasing (IWP2005). 2005. p. 13
- [8] Tjong EF, Tjong S, Sang M, De F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Disponível em: <https://arxiv.org/abs/cs/0306050>
- [9] Deep Learning Book. disponível em: <https://www.deeplearningbook.com.br/o-que-e-bert-bidirectional-encoder-representations-from-transformers/>
- [10] Rudin C, Wagstaff KL. Machine learning for science and society. *Machine Learning*. 2014;**95**:1-9
- [11] Tin Kam HO. Random Decision Forests (PDF). In: Proceedings of the 3rd International 9 Conference on Document Analysis and Recognition; Montreal, QC; 14-16 August 1995. 1995. pp. 278-282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016
- [12] Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995;**20**(3):273-297. DOI: 10.1007/BF00994018
- [13] Andrew Mccallum. Graphical Models, Lecture2: Bayesian Network Representation (PDF). pp. 4-6. [Accessed: 22 October 2019]
- [14] Rothman D. Transformers for Natural Language Processing: Build Innovative Deep Neural Network Architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and More. Packt Publishing Ltd; 2021. p. 2