doi:10.3233/SHTI210834

# Deep Learning, a Not so Magical Problem Solver: A Case Study with Predicting the Complexity of Breast Cancer Cases

My-Anh LE THIEN[a,1], Akram REDJDAL[a], Jacques BOUAUD[b,a] and Brigitte SEROUSSI[a,c,d]
[a] *Sorbonne Université, Université Sorbonne Paris Nord, Inserm, UMR S_1142, LIMICS, Paris, France*
[b] *AP-HP, DRCI, Paris, France*
[c] *AP-HP, Hôpital Tenon, Paris, France*
[d] *APREC, Paris, France*

**Abstract.** Using guideline-based clinical decision support systems (CDSSs) has improved clinical practice, especially during multidisciplinary tumour boards (MTBs) in cancer patient management. However, MTBs have been reported to be overcrowded, with limited time to discuss all cases. Complex breast cancer cases that need further MTB discussions should have priority in the organization of MTBs. In order to optimize MTB workflow, we attempted to predict complex cases defined as non-compliant cases despite the use of the decision support system OncoDoc. After previously obtaining insufficient performance with machine learning algorithms, we tested Multi Layer Perceptron for classification, compared various samplers to compensate data imbalance combined with cross-validation, and optimized all models with hyperparameter tuning and feature selection with no improvement and lacklustre results (F1-score: 31.4%).

**Keywords.** Clinical decision support systems, Deep learning, Breast cancer

## 1. Introduction

Patient-specific treatment plans as recommended by clinical practice guidelines (CPGs) have improved patient outcomes and clinicians are highly encouraged to implement them [1]. In many countries, the management of cancer patients is discussed in multidisciplinary tumour boards (MTBs) to allow for the best collective decision-making. In addition, clinical decision support systems (CDSSs) have shown to improve the compliance of MTB decisions with CPG recommendations. However, MTBs have been reported to be overloaded with limited time to discuss properly each clinical case.

OncoDoc is a CDSS developed to provide patient-specific recommendations and promote the implementation of breast cancer CPGs [2]. The system has been routinely used in MTBs at the Tenon hospital (Paris, France) in a three-year period. MTB decision compliance rate with CPGs for invasive breast cancers reached 91.7%. In the remaining 8.3%, patients presented specific medical circumstances not formally

[1] Corresponding author, My-Anh Le Thien, Email: myanh.lethien@gmail.com

covered by the CPGs in OncoDoc knowledge base, which explain that clinicians might in some cases disagree with OncoDoc treatment plans and choose not to follow them.

We wish to pre-emptively identify patients whose profiles are too complex to be properly handled by CPGs and optimize patient triage ahead of MTBs. In this way, non-complex cases might be treated more rapidly, and additional time and focus could be allocated to cases identified as complex. As OncoDoc is directly built from CPGs, we hypothesize that non-compliance with OncoDoc recommendations is a marker of case complexity and that predicting cases leading to non-compliant MTB decisions might help identify complex cases.

In a previous study, we used different machine learning algorithms (RandomForest, DecisionTree, XGBoost) to classify OncoDoc's complex cases based on available routine data and tested numerous sampling methods to compensate data imbalance, with unsatisfying results (F1-score did not exceed 40%) [3]. As deep learning is known to offer possible improvement on imbalanced samples when machine learning lacks sufficient performance, we implemented a Multi Layer Perceptron (MLP) on OncoDoc data in order to improve complex cases classification and identification.

## 2. Material and Methods

### *2.1. Dataset*

Data was collected from the existing OncoDoc database and included MTB decisions for adult women treated for breast cancer from February 2007 to September 2009 at the Tenon hospital (AP-HP, Paris, France). Data consisted of 1,887 MTB decision instances (1,054 patients) with 127 collected variables. A sizable number of variables was incomplete due to OncoDoc's architecture as a decision tree: data not relevant to the case is not asked, and therefore not entered.

We applied supervised deep learning with labelled training datasets and all values predicted from the test datasets verified against the actual class. According to our hypothesis, cases where clinicians did not comply with OncoDoc recommendations (8.3%) were labelled as "complex" whereas others cases were labelled as "non-complex". Whenever possible, missing values were imputed so as to remain logically sound, e.g., when a tumour was non-invasive, all tumour-invasive-related variables were filled as "not applicable". Continuous variables were converted to categorical variables (e.g. patient age) and all variables with labels were encoded as integers. All missing or "not applicable" values were considered as integers, e.g., excision margins outlying invasive tumour originally coded as invaded (1) or not (0) were encoded as missing (0), not applicable to non-invasive tumour (1), non-invaded (2) and invaded (3). Additional variables were built to reflect known factors of clinical complexity, e.g., triple negative breast cancer patients (hormonal receptors = negative AND Her2 receptors = negative). The final dataset consisted of 1,887 decisions and 70 variables.

### *2.2. Multi Layer Perceptron*

MLP is a subtype of artificial neural network (ANN) which uses back-propagation, and presents at least three neuron layers for data classification: an input layer, a hidden layer, and an output layer. In MLP, data is handled one way and passes all layers once, as opposed to the way data is handled in recurrent neural networks. Back-propagation

allows estimating what constitutes erroneous values in the hidden layer (which has no reference in the initial data to compare itself to) from the final classification labels [4].

### *2.3. Splitting and sampling dataset*

The following standard procedures were applied for this study:

- With stratification on case complexity: stratification keeps the class ratio between the original dataset and the training and testing datasets.
- With or without k-fold cross-validation: in k-fold cross-validation, the training set is split into k smaller sets (here, k=5). Each of the k "folds" are used in turn as testing set against the rest and cross-validation then averages performance measures to give a more accurate estimate of model performance.

Since data was severely imbalanced, as there were few complex cases, we systematically tested and compared the following samplers on each training datasets generated for cross-validation, in order to offset data imbalance [5]:

- Random Under Sampling (RUS)
- Random Over Sampling (ROS)
- Adaptive Synthetic (ADASYN)
- Synthetic Minority Oversampling Technique (SMOTE)
- SMOTE and Edited Nearest Neighbours (SMOTEEN)

### *2.4. Hyperparameter tuning and feature selection*

We searched the best values for the model's hyperparameters using Random Search and Grid Search. Random Search tested random combinations of hyperparameters in a given range and Grid Search was applied to narrow down the best values of hyperparameters. We optimized model variables with feature selection: all variables were first included in the analyses, then progressively excluded from the model from lowest to highest feature importance, e.g. from least to most useful variable as identified by the trained model. The optimized number of variables to include was then retained.

### *2.5. Evaluation indices*

MLP was evaluated using precision, recall, and F1-score. Accuracy was available but considered to be unreliable as the testing dataset was unbalanced (if 90% of data belongs to class A, a model might simply choose to systematically class data as class A to obtain a 90% accuracy). We compared the mean cross-validation indices for each sampling methods and model. The overall process is illustrated in Figure 1.

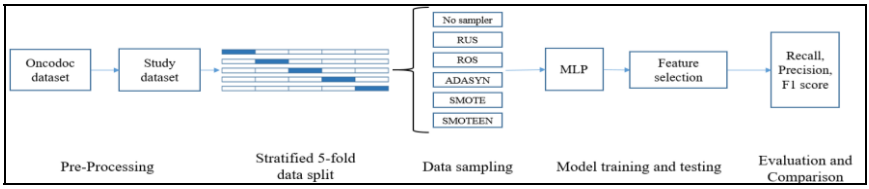


**Figure 1.** Analytic plan.

## 3. Results

Before feature selection, MLP presented its best F1 score without any sampler (recall=25.9%, precision=40.4%, F1-score=31.4%). Applying samplers improved recall at the cost of precision, especially for RUS (recall=65.5%, precision=16.9%, F1-score=26.7%) and SMOTEEN (recall=52.3%, precision=20.2%, F1-score=28.6%), with decreased F1-score.
Feature selection did not improve MLP when no sampler was applied but showed slight improvement for ROS (before feature selection: recall=39.1%, precision=19.3%, F1-score=25.6% *versus* after feature selection: recall=47.6%, precision=23.5%, F1-score=30.9%) and ADASYN (before feature selection: recall=33.3%, precision=21.3%, F1-score=25.8% *versus* after feature selection: recall=37.3%, precision=25.0%, F1-score=29.2%). Almost no improvement was observed for SMOTE, and performance stayed the same for SMOTEEN. (cf. Table 1).

**Table 1**. MLP model scores by sampling technique before and after feature selection (%)

| Model | Score | No sampling | RUS | ROS | ADASYN | SMOTE | SMOTEEN |
|---|---|---|---|---|---|---|---|
| Before feature selection | Recall | 25.9 | 65.5 | 39.1 | 33.3 | 33.3 | 52.3 |
| | Precision | 40.4 | 16.9 | 19.3 | 21.3 | 26.2 | 20.2 |
| | F1 | 31.4 | 26.7 | 25.6 | 25.8 | 28.6 | 29.0 |
| After feature selection | Recall | 25.9 | 63.2 | 47.6 | 37.3 | 33.9 | 50.0 |
| | Precision | 40.4 | 17.0 | 23.5 | 25.0 | 27.5 | 20.4 |
| | F1 | 31.4 | 26.9 | 30.9 | 29.2 | 29.5 | 28.7 |

## 4. Discussion

Multiple sampling methods were used to compensate data imbalance with MLP. Improvement was observed for recall in all samplers but showed deteriorated values for precision, with unsatisfactory F1-score results. In other words, our models could not easily identify complex cases and could only improve recall performance by indiscriminately selecting cases in the hope to identify complex cases, decreasing precision performance. Calculated from recall and precision, F1-score could not increase when one score improved at the expense of the other, and showed the overall performance plateau. Hyperparameters tuning improved all models through Random Search and Grid Search but remains insufficient. Throughout hyperparameters tuning, all models had the tendency to sacrifice precision for recall. Feature selection granted slight improvement by excluding variables presenting no added value to the model. Few variables were consistently more “important” than others e.g., ‘Profile frequency’ or ‘Age_35_75’ are easily explained as known factors of clinical complexity.

As a pre-constructed model in Python, MLP generated results similar or worse than those obtained with machine learning algorithms [3]. However, hyperparameter tuning with MLP was a non-compressible time-consuming task that did not allow for extensive testing and a larger range of parameters could possibly yield better results. On the other hand, we attempted to construct deep learning models “from scratch” without using a pre-existing Python library such as MLP, but all of them choose to classify all data as “non-complex” with no attempt to identify “complex” cases. Such behaviour could not be explained as a way to exploit the data imbalance (when 90% of

data was 'non-complex') since similar results were obtained after compensating with data sampling. As it is possible that further testing could lead to a functional model "from scratch", further research and understanding of deep learning models and neural networks might improve our current results.

Errors in data were encountered during pre-processing, faulty recording of patients' characteristics might have occurred. Likewise, data structure and variables were significantly modified during pre-processing (various variables were outright excluded from analysis, or heavily modified to simplify or avoid missing data) which may have removed important information for data classification. Additional data might also benefit model training and it is plausible that OncoDoc available data simply does not include the information needed to accurately predict complex cases.

Lastly, we assumed non-conformity with OncoDoc as the sole explicit marker of complexity and acknowledge the possible limits associated: complex cases which eventually remained conform to OncoDoc after a lengthy discussion are not identified in our study, likewise, some treatments might be dismissed without posing difficulty to MTB clinicians (such as patient's preference). Further reviews of MTB decisions and analyses of causes for non-compliance might give us a more accurate indication of complexity for further studies.

## 5. Conclusion

We collected data from OncoDoc routine data to identify complex patient cases, with deep learning methods, assuming complex cases could not be systematically managed by a CPG-based CDSS. The dataset was prepared for analysis and underwent various sampling techniques to overcome its class imbalance. Several improvement plans were implemented but results did not improve our previous performance with machine learning models, hinting at a possible need for different models, additional data and/or additional data structuring for further improvement.

## References

[1] Kreienberg R, Wöckel A, Wischnewsky M. Highly significant improvement in guideline adherence, relapse-free and overall survival in breast cancer patients when treated at certified breast cancer centres: An evaluation of 8323 patients. The breast. 2018 Aug 1;40:54-9.

[2] Séroussi B, Bouaud J, Gligorov J, Uzan S. Supporting multidisciplinary staff meetings for guideline-based breast cancer management: a study with OncoDoc2. InAMIA Annual Symposium Proceedings 2007 (Vol. 2007, p. 656). American Medical Informatics Association.

[3] Le Thien M-A, Redjdal A, Bouaud J, et al. Using Machine Learning on Imbalanced Guideline Compliance Data to Optimize Multidisciplinary Tumour Board Decision Making for the Management of Breast Cancer Patients Studies in Health Technology and Informatics. Stud Health Technol Inform. 2021, October 2-4, To appear

[4] Moghaddasi H, Ahmadzadeh B, Rabiei R, et al. (2017 [cited 2021 Sep 14]) Study on the Efficiency of a Multi-layer Perceptron Neural Network Based on the Number of Hidden Layers and Nodes for Diagnosing Coronary- Artery Disease Jentashapir J Cell Mol Biol. 2017; 8(3):e63032.

[5] Sun Y, Wong AKC, Kamel MS (2009) Classification of imbalanced data: a review Int J Patt Recogn Artif Intell. 23:687–719.