The substantial influence of negative sampling and prevalence when presenting classification results: case study with TOF-MRA

**Introduction**

Recent advancements in deep learning, in particular, have resulted in improved solutions for the detection and characterization of intracranial vascular anomalies. From a technical point of view, we refer to major papers reporting the procedures followed by different groups in the development of these applications. While the focus of these contributions may differ, most datasets share the same characteristics, i.e., they are gathered in distinct stages and prepared by a limited number of individuals. All patterns obtained from the available capture steps, e.g., from the raw data to the images, are processed identically. While this similarity is a benefit to avoid overfitting, it requires to some extent the application of techniques to improve classification results. One such technique is to use negative samples extracted from pathology-free areas of the same volume. While this approach is well established, it has not been implemented consistently. The effect of prevalence, none of them corrected classification metrics accordingly. In this regard, this work explores how classification performances of a typical case study are influenced by different sampling strategies in the context of true aneurysm detection. We show that the classification performance can vary drastically with the sampling strategy as well as the prevalence of positive cases.

**Methods**

We performed an extensive study to evaluate the effect of sampling strategies on deep learning-based classification results. We used three different datasets (Hounsfield units, intensity-matched, vesselness-based) to evaluate the effect of sampling strategy on classification results, along with the oftentimes-forgotten real-world prevalence adjustment for the study population. Though similar works have been performed before, the results have not been generalized or applied to real-world clinical cases. Our study explores whether the prevalence of positive cases can influence classification results and how this can be compensated for.

**Results**

Our study demonstrates how classification results that at first glance might seem promising, should instead always be contextualized with respect to the real-world prevalence of positives. We observed that classification results can change drastically between different sampling strategies, even when the prevalence of positives is kept constant. This effect is more pronounced when the prevalence of positives is low. Our study also highlights the importance of real-world prevalence adjustment for classification results.

**Discussion**

This work demonstrates how classification results that at first glance might seem promising, should instead always be contextualized with respect to the real-world prevalence of positives. We observed that classification results can change drastically between different sampling strategies, even when the prevalence of positives is kept constant. This effect is more pronounced when the prevalence of positives is low. Our study also highlights the importance of real-world prevalence adjustment for classification results.

**Acknowledgements**

No acknowledgements found.

**References**


**Table 1: MR acquisition parameters of the study population.**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field strength</td>
<td>3.0 T</td>
</tr>
<tr>
<td>Echo time</td>
<td>6.19 ms</td>
</tr>
<tr>
<td>Repetition time</td>
<td>4.5 ms</td>
</tr>
<tr>
<td>Flip angle</td>
<td>8°</td>
</tr>
<tr>
<td>Resolution</td>
<td>0.625 x 0.625 x 0.625 mm³</td>
</tr>
<tr>
<td>Matrix size</td>
<td>224 x 224 x 128</td>
</tr>
</tbody>
</table>

**Figure 1:** Example of test T2FLAIR sequence on an 80-year-old patient with microscopic Hemorrhagic Stroke - MBI_2019-07-24_02_21_23. The red vessel mask was updated with Vascular Modelling Tools - VMTK from the local aneurysm source.

**Figure 2:** Comparison of positive and negative patch extraction from (a) D1 (random sampling), (b) D2 (intensity-matched sampling) and (c) D3 (vesselness-based sampling). These are computed specifying an explicit prevalence: 

- **Acc = accuracy**, **Sens = sensitivity**, **Spec = specificity**, **AUC = Area Under ROC Curve**, **RW-PPV = real-world positive predictive value**, **RW-NPV = real-world negative predictive value**.