O^{2}PF: Oversampling via Optimum-Path Forest for Breast Cancer Detection

Leandro A. Passos*, Danilo S. Jodas*, Luiz C. F. Ribeiro*, Thierry Pinheiro, João P. Papa
Department of Computing
São Paulo State University
Bauru, Brazil
{leandro.passos, luiz.felix, joao.papa}@unesp.br, {danilojodas, thierrypin}@gmail.com

***Abstract*—Breast cancer is among the most deadly diseases, distressing mostly women worldwide. Although traditional methods for detection have presented themselves as valid for the task, they still commonly present low accuracies and demand considerable time and effort from professionals. Therefore, a computer-aided diagnosis (CAD) system capable of providing early detection becomes hugely desirable. In the last decade, machine learning-based techniques have been of paramount importance in this context, since they are capable of extracting essential information from data and reasoning about it. However, such approaches still suffer from imbalanced data, specifically on medical issues, where the number of healthy people samples is, in general, considerably higher than the number of patients. Therefore this paper proposes the O^{2}PF, a data oversampling method based on the unsupervised Optimum-Path Forest Algorithm. Experiments conducted over the full oversampling scenario state the robustness of the model, which is compared against three well-established oversampling methods considering three breast cancer and three general-purpose tasks for medical issues datasets.**

***Index Terms*—Data imbalance, Oversampling, Optimum-Path Forest.**

I. Introduction

Computer-Aided Diagnosis (CAD) systems aim at helping physicians to quickly report better diagnosis to patients, thus representing an essential step towards dangerous disease accurate diagnosis. Besides, intelligent-based CAD systems have successfully employed machine learning techniques, a promising subfield of artificial intelligence, to tackle complicated problems that demand knowledge and reasoning about the subject. Regarding the latter, several works were developed in the last few years to aid detecting atherosclerosis [1], Parkinson's disease [2], and breast cancer [3], to name a few.

Breast cancer is a dangerous illness, affecting millions of women, as well as few men, around the world. According to the World Health Organization, such a disease was responsible for the death of approximately $627,000$ women in 2018 worldwide [4]. Therefore, early diagnosis is crucial to effectively prevent its progress, consequently making it possible to elaborate more efficient treatment plans. In general, standard breast cancer diagnosis consists of a visual analysis performed by professionals aiming to identify possible abnormalities (e.g., nodules), which may indicate cancer risk signs. Once identified, it is possible to extract relevant measures from such nodules, assisting physicians to judge the presence or absence of cancerous tissue.

Many image processing-based algorithms have also been proposed to help in such tasks [5]–[7], as well as machine learning techniques, which have been commonly employed for both nodule segmentation and classification. Agarap [8], for instance, studied the performance of several machine learning techniques to label a nodule as benign or malignant. Further, Passos et al. [9] developed a neural network to tell benign from malignant nodules, and to label the latter according to the most likely cancer type.

Moreover, once the nodules can be seen as abnormalities, their identification can also be tackled in an unsupervised fashion. In such a direction, Ribeiro et al. [10] used the unsupervised Optimum-Path Forest (OPF) [11] algorithm to distinct malignant nodes from benign, obtaining significant results. The unsupervised Optimum-Path Forest algorithm possesses properties commonly employed to improve other machine learning techniques, such as the Brainstorm Optimization [12], as well as to create new techniques, such as the OPF-based approach for anomaly detection [13] and the Fuzzy OPF classifier [14].

Although the above-mentioned techniques obtained promising results in the context of breast cancer detection, most of the classification algorithms usually suffer from imbalanced dataset problems, which arises when the number of samples among classes differs significantly. In this scenario, the trained classifier is more likely to label a new sample as belonging to the most common (majority) class, degrading its performance for the smallest (minority) class. Moreover, such a situation may degenerate when the dataset presents outliers. Numerous studies have also been proposed to cope with such a problem in medical datasets [15]–[17]. Therefore, the generation of synthetic samples for the minority class, which is usually referred to as oversampling, is recognized as a prominent contribution to rebalancing the dataset for classifier training purposes, consequently improving their robustness to label minority class samples correctly. In this context, several

*Authors contributed equally.

The authors would like to thank FAPESP grants #2013/07375-0, #2014/12236-1, #2019/18287-0, and #2019/07665-4, as well as CNPq grants #427968/2018-6 and #307066/2017-7.

powerful oversampling strategies have been proposed in the literature to tackle such a problem [18]–[20], most of them still presenting difficulties while enforcing diversity among new synthetic samples, which denotes a problem that worth to be addressed, since it can improve classifier generalization properties.

To such an extent, this work proposes an approach to perform oversampling via the Optimum-Path Forest, hereinafter named O^{2}PF. The method employs the unsupervised OPF algorithm to capture features intrinsic to the minority class into different clusters. Further, new training examples are generated by sampling from a Gaussian distribution parametrized by the cluster characteristics. Therefore, the main contribution of this paper are twofold: (i) extending the OPF algorithm capabilities by introducing a novel oversampling mechanism that enforces synthetic intra-class variability; and (ii) studying how O^{2}PF can benefit the development of CAD systems by extensively evaluating it on five tumor-classification and one retinopathy identification tasks. Finally, the full-balanced datasets are employed to train the OPF classifier, and results, i.e., the accuracy, recall, and F1-measure, are compared against the ones obtained by the standard training dataset without oversampling. Additionally, results are compared against three baselines: SMOTE [18], the Borderline SMOTE [19], and ADASYN [20].

In the remainder of this paper, Section II presents the theoretical background regarding the unsupervised and supervised OPF algorithm variants, whereas Section III introduces the proposed approach. Further, Section IV outlines the experimental setup and Section V discusses experimental results. Finally, Section VI presents conclusions and future works.

II. Background

This section presents the theoretical background regarding the unsupervised and supervised variants of the OPF algorithm.

A. Unsupervised Optimum-Path Forest

Let $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ be a dataset such that $\boldsymbol{x}_i \in \mathbb{R}^m$ represents the features extracted from the i-th sample. Further, let $\mathcal{G} = (\mathcal{D}, \mathcal{A}_{k^\star})$ be a graph where each node corresponds to a different feature vector connected to its $k^\star$-nearest neighbors, as defined in the adjacency relationship set $\mathcal{A}_{k^\star}$.

The unsupervised OPF algorithm consists in partitioning the graph through a competitive process, in which a few samples are marked as "prototypes" and compete among themselves to conquer the remaining nodes. Such a procedure partitions the $\mathcal{G}$ into optimum-path trees (OPTs) rooted at a prototype, each corresponding to a cluster, where a sample is more similar to the elements of its tree than any other tree. Overall, the process can be divided into three steps: (i) computing a proper neighborhood size $k^\star$ and the adjacency relationship $\mathcal{A}_{k^\star}$, (ii) electing the prototype nodes, and (iii) performing the competition process to partition the dataset into OPTs.

Regarding the first step, different approaches may be considered. Among others, Rocha et al. [11] proposed to find $k^\star$ by minimizing the normalized graph cut function, as it takes into account the dissimilarities between clusters as well as the similarity degree among samples of each cluster.

Concerning the second step, the algorithm must select the prototypes to form the root of each OPT (which will ultimately form clusters) to rule the competition process and conquer the remaining samples in the graph. The supervised OPF proposed by Papa et al. [21], [22] selects as prototypes the nearest samples from different classes, found by computing the graph Minimum Spanning Tree (MST). However, in the unsupervised variant, since labels are usually unavailable, Rocha et al. [11] proposed to select the prototypes as the samples located in the center of each cluster. To such an extent, all samples are assigned a density score $\rho(\boldsymbol{x}_i), \quad \forall \boldsymbol{x}_i \in \mathcal{D}$, computed through a Gaussian probability density function (pdf), defined as follows:

$$\rho(\boldsymbol{x}_i) = \frac{1}{k^\star \sqrt{2\pi\sigma^2}} \sum_{\forall \boldsymbol{x}_j \in \mathcal{A}_{k^\star}(\boldsymbol{x}_i)} \exp\left(\frac{-d(\boldsymbol{x}_i, \boldsymbol{x}_j)}{2\sigma^2}\right), \quad (1)$$

where $i \neq j$, $\sigma = d_{\max}/3$ and $d_{\max}$ stands for the maximum arc-weight in $\mathcal{G}$. This formulation considers all adjacent nodes for density computations, as the Gaussian distribution covers 99.7% of the samples with distance $d(\boldsymbol{x}_i, \boldsymbol{x}_j) \in [0, 3\sigma]$.

After evaluating Equation 1 for each node in the graph, the density values are used to populate a priority queue in a way that the unsupervised OPF maximizes the cost of each sample, thus partitioning the graph into OPTs. Such a cost is defined in terms of paths on $\mathcal{G}$, which is an acyclic sequence of adjacent samples in $\mathcal{A}_{k^\star}$.

Let $\pi_{\boldsymbol{x}_i}$ be a path with terminus at sample $\boldsymbol{x}_i$ and starting from some root $\mathcal{R}(\boldsymbol{x}_j)$, being the latter the set of all prototype samples. Further, let $\pi_{\boldsymbol{x}_i} = \langle \boldsymbol{x}_i \rangle$ be a trivial path (i.e., a path containing only one sample), whereas $\pi_{\boldsymbol{x}_i} \cdot \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$ denotes the concatenation of a path $\pi_{\boldsymbol{x}_i}$ and the arc $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ such that $i \neq j$.

In the third step, the algorithm assigns to each path $\pi_{\boldsymbol{x}_i}$ a value $f_{\min}(\pi_{\boldsymbol{x}_i})$ given by a smooth connectivity function $f_{\min} : \mathcal{D} \rightarrow \mathbb{R}^+$, which must satisfy some constraints to ensure the algorithm theoretic correctness [23], [24]. A path $\pi_{\boldsymbol{x}_i}$ is considered optimal if $f_{\min}(\pi_{\boldsymbol{x}_i}) \geq f_{\min}(\tau_{\boldsymbol{x}_i})$ for any other path $\tau_{\boldsymbol{x}_i}$. Among the proposed path-cost functions in the literature, the unsupervised OPF relies on the following formulation:

$$\begin{aligned} f_{\min}(\langle \boldsymbol{x}_i \rangle) &= \begin{cases} \rho(\boldsymbol{x}_i) & \text{if } \boldsymbol{x}_i \in \mathcal{R} \\ \rho(\boldsymbol{x}_i) - \delta & \text{otherwise.} \end{cases} \\ f_{\min}(\pi_{\boldsymbol{x}_i} \cdot \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle) &= \min\{f_{\min}(\pi_{\boldsymbol{x}_i}), \rho(\boldsymbol{x}_j)\}, \end{aligned} \quad (2)$$

where $\delta = \min_{\forall (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{A}_{k^\star} | \rho(t) \neq \rho(s)} |\rho(t) - \rho(s)|$ consists in the smallest quantity to avoid plateaus and over-segmentation in regions near prototypes (i.e., areas with the highest density).

Among all possible paths $\pi_{\boldsymbol{x}_i}$ that originate in some local maximum, i.e., some prototype, the OPF algorithm assigns to each sample a final path whose minimum density value along

it is maximum. Such final path value is represented by a cost map $\mathcal{C}$, as follows:

$$\mathcal{C}(\boldsymbol{x}_i) = \max_{\forall \pi_{\boldsymbol{x}_j} \in \mathcal{D},\, i \neq j} \{f(\pi_{\boldsymbol{x}_j} \cdot \langle \boldsymbol{x}_j, \boldsymbol{x}_i \rangle)\}. \tag{3}$$

The OPF algorithm maximizes $\mathcal{C}(\boldsymbol{x}_i)\ \forall \boldsymbol{x}_i \in \mathcal{D}$ by computing an optimum-path forest for each sample in descending order of cost. Each forest is encoded as an acyclic predecessor map $\mathcal{P}$ which assigns to each sample $\boldsymbol{x}_i \notin \mathcal{R}$ its predecessor $\mathcal{P}(\boldsymbol{x}_i)$ in the optimum path from $\mathcal{R}$, or a marker nil when $\boldsymbol{x}_i \in \mathcal{R}$.

It is important to remark that the unsupervised OPF algorithm determines the number of clusters (OPTs) automatically, hence such information is not required beforehand, differently from other algorithms. Furthermore, the only hyperparameter that must be set is the search interval upper bound $k_{\max}$ for the proper neighborhood size $k^\star \in [1, k_{\max}]$.

B. Supervised Optimum-Path Forest

Differently from its unsupervised version, the supervised variant uses a fully-connected graph $\mathcal{G}' = (\mathcal{D}, \mathcal{B})$ instead. Moreover, the closest samples from different classes are marked as prototypes, as aforementioned. Regarding the competition process, instead of using Equation 2, the following smooth function is employed:

$$\begin{aligned} f_{\max}(\langle \boldsymbol{x}_i \rangle) &= \begin{cases} 0 & \text{if } \boldsymbol{x}_i \in \mathcal{R} \\ \infty & \text{otherwise.} \end{cases} \\ f_{\max}(\pi_{\boldsymbol{x}_i} \cdot \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle) &= \max\left\{f_{\max}(\pi_{\boldsymbol{x}_i}), d(\boldsymbol{x}_i, \boldsymbol{x}_j)\right\}. \end{aligned} \tag{4}$$

Further, the following cost map is used to partition the graph:

$$\mathcal{S}(\boldsymbol{x}_i) = \min_{\boldsymbol{x}_i \in \mathcal{D}} \left\{\max\left\{\mathcal{S}(\boldsymbol{x}_i), d(\boldsymbol{x}_i, \boldsymbol{x}_j)\right\}\right\}. \tag{5}$$

Such costs are used to initialize the algorithm before evaluating Equation 5, which is performed for every node in an ascending order of costs. After partitioning the graph, each prototype propagates its ground truth label to all samples in its OPT. Afterwards, prediction is performed by solving the same equation for new samples individually.

III. Proposed Approach

Although the unsupervised OPF was conceived for clustering purposes, such clusters possess a set of features that can also be employed to synthesize new samples, thus being suitable for oversampling. This section describes the procedure for a binary classification problem, which intends to oversample the class composed of the smallest number of features. Notwithstanding, the same approach can be easily applied to multiclass problems by individually repeating the procedure for each class to be oversampled.

The process of synthetic samples generation contemplates two main steps: (i) creating plausible samples, i.e., synthetic elements with characteristics that are coherent with the selected class; and (ii) introducing sample variability, avoiding making the classifier biased towards a subset of characteristics present in that class. To tackle such issues, the O^{2}PF first performs the clustering from minority class samples, which turns possible the extraction of common patterns intrinsic to the class, i.e., the samples' average position and variance. Further, the algorithm assumes that all features from a class follow a normal distribution. Thus, a new sample $\boldsymbol{z} \in \mathbb{R}^m$ can be generated by sampling such a distribution from some of the q clusters found. Notice the number of synthetic samples generated by each cluster is proportional to the number of original samples compounding it, i.e., we are always doubling the number of samples from the minority class, although the user can set that percentage. The distribution is performed as follows:

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}_q, \Sigma_q), \tag{6}$$

where $\boldsymbol{\mu}_q \in \mathbb{R}^m$ stands for the distribution mean, defined as the average feature vector from all samples within the q-th cluster. Moreover, $\Sigma_q \in \mathbb{R}^{m \times m}$ is the covariance matrix, computed as follows:

$$\Sigma_q = \frac{1}{n_q - 1}(X_q - \boldsymbol{\mu}_q)(X_q - \boldsymbol{\mu}_q)^T, \tag{7}$$

given that $X_q \in \mathbb{R}^{m \times n_q}$ is a matrix formed by concatenating all n_q cluster feature vectors. Figure 1 illustrates the proposed approach behavior.

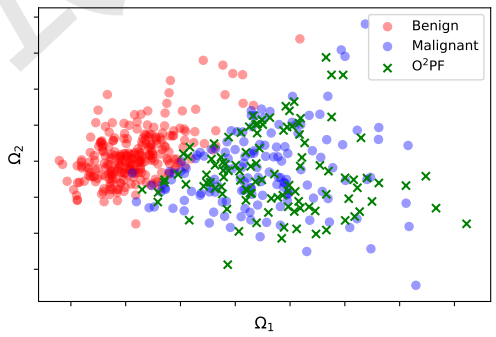

Fig. 1: Oversampling using O^{2}PF. The image comprises WBCD Diagnostic II dataset samples and synthetic points from the minority class generated using O^{2}PF. Data were projected onto a 2-dimensional space using the Principal Component Analysis algorithm for visualization purposes.

IV. Methodology

This section describes the datasets used in the experiments. Further, the experimental setup is outlined.

A. Datasets

Experiments were conducted over two sets of three databases each. The first set comprises three datasets for breast cancer detection, i.e., the Wisconsin Breast Cancer Database, which is composed of the datasets Prognostic, Diagnostic I,

and Diagnostic II. The second set is composed of general-purpose datasets for medical issues. All of them are unbalanced, binary, and were obtained from the UCI repository [25]. A brief description of each one follows below:

- **Wisconsin Breast Cancer Database (WBCD) Prognostic**[1] [26]**:** Regards predicting a sample as recurrent or non-recurrent type of cancer based on 32 features. There are 198 samples, being 151 (76.3%) non-recurrent and 47 (23.7%) recurrent;
- **WBCD Diagnostic I:** Consists of classifying a tumor as benignant or malignant based on 32 features as well. There are 569 instances, from which 357 (63.7%) are benign and 212 (37.3%) are malignant;
- **WBCD Diagnostic II:** Corresponds to labelling each of the 699 samples as benign or malignant tumor. Each sample comprises 9 features and each class contains 458 (65.5%) and 241 (34.5%) examples, respectively;
- **Diabetic Retinopathy Debrecen (DRD)**[2] [27]**:** Regards predicting whether an image contains signs of diabetic retinopathy or not based on 19 variables. The dataset contains 1,151 samples, from which 611 (53.1%) are positive and 540 (46.9%) are negative;
- **Cervical Cancer (CC)**[3] [28]**:** For this task we predict the binary *biopsy* variable based on 32 features for 858 samples. Differently from other datasets, all variables in this scenario are either integer or binary. Further, the dataset is significantly skewed, with 55 (6.4%) positive and 803 (93.6%) negative samples;
- **Mammographic Mass (MM)**[4] [29]**:** Concerns predicting if a mammographic mass is benign or malignant based on six features. The dataset contains 516 (53.7%) benign and 445 (46.3%) malignant samples, comprising 961 examples.

B. Experimental setup

The experiments conducted in this paper considered preprocessing the data such that missing features were replaced by their corresponding mean in the training partition. All the features were normalized to have zero mean and unitary standard deviation. Further, the datasets were randomly divided into training, validation and testing sets, each containing 70%, 15%, and 15% of the data, respectively[5].

The validation set was employed to fine-tune the oversampling method hyperparameters, i.e., finding the $k_{\max}$ for the unsupervised OPF and κ for the other methods that maximize the minority class recall. Afterward, the augmented dataset is used to train the OPF classifier for further computing the results over the testing sets. Notice that such a procedure is repeated 20 times for statistical analysis, and the results are compared through the Wilcoxon signed-rank test [30] with 0.05 significance concerning recall values. Figure 2 depicts such a pipeline. Implementation-wise, we rely on the supervised and unsupervised implementations provided by Opfython [31]. Additionally, the source code was implemented using Python and is available on GitHub.[6]

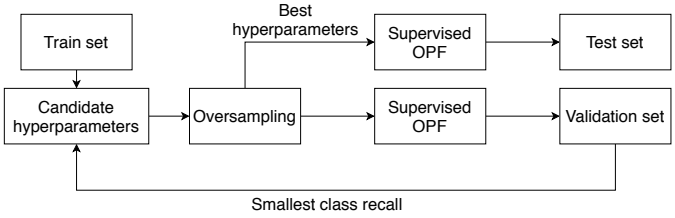

Fig. 2: Experimental pipeline for each dataset partition.

V. Experimental Results

This section is divided into four main steps: (i) datasets augmented using the O^{2}PF are compared against the standard version, (ii) the proposed approach is compared against three baselines for oversampling considering three distinct versions of the Wisconsin Breast Cancer Database, i.e., Prognostic, Diagnostic I, and Diagnostic II, for the task of breast cancer detection. In step (iii), a similar experiment is conducted over three general-purpose medical issues datasets, and (iv) it provides a brief discussion concerning the optimization of the proposed method hyperparameter, i.e., the $k_{\max}$.

A. O^{2}PF Data Augmentation Versus Standard Datasets

This section presents the results obtained by the Optimum-Path Forest classifiers considering the datasets balanced through O^{2}PF oversampling, i.e., minority classes are augmented such that both classes present a similar number of samples, against the standard version of the datasets. Table I presents the recall considering each dataset. Values in bold denote the best results according to the Wilcoxon signed-rank test with 5% of significance.

TABLE I: Evaluation of the proposed O^{2}PF against the standard datasets versions.

Ds. Version	Recall	Prognostic	Diagnostic I	Diagnostic II	DRD	CC	MM
ORIGINAL	Avg.	0.4945	**0.9184**	0.8903	0.4922	**0.6320**	**0.6180**
	Std.	±0.1872	±0.0346	±0.0550	±0.2017	±0.1002	±0.0480
O^{2}PF	Avg.	**0.5739**	**0.9311**	**0.9104**	**0.5834**	**0.6292**	0.6086
	Std.	±0.1368	±0.0331	±0.0452	±0.2308	±0.0924	±0.0434

In this context, one can observe O^{2}PF obtained the best results in five out of six datasets according to the Wilcoxon signed-rank test, obtaining the best results alone in three of them.

[1]All considered versions are available at https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic).

[2]Available at https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set.

[3]Available at https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29.

[4]Available at https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass.

[5]Such percentages were obtained empirically.

[6]The source code is available online at https://github.com/Leandropassosjr/o2pf.

B. Results concerning the Breast Cancer datasets

Table II presents the average recall, accuracy, F1-measure and best $k_{\max}$ considering O^{2}PF or κ considering the other techniques, as well as their standard deviation. The results comprise a minority oversampling, thus providing balanced datasets. The proposed approach is compared against three baseline techniques, i.e., SMOTE, Borderline SMOTE, and ADASYN.

TABLE II: Results considering WBCD Breast Cancer datasets.

Dataset	Statistics	O^{2}PF	SMOTE	Borderline SMOTE	ADASYN
Prognostic	Recall	$\mathbf{0.5739 \pm 0.1368}$	$\mathbf{0.5803 \pm 0.1253}$	$\mathbf{0.5802 \pm 0.1343}$	$\mathbf{0.6311 \pm 0.1283}$
	Accuracy	0.6317 ± 0.0654	0.6183 ± 0.0532	0.6317 ± 0.0619	0.6300 ± 0.0547
	F1	0.4143 ± 0.1310	0.4112 ± 0.1330	0.4175 ± 0.1266	0.4366 ± 0.1413
	Best k	37.2500 ± 27.0867	6.4000 ± 1.5297	6.0500 ± 1.4654	6.1500 ± 1.4586
Diagnostic I	Recall	$\mathbf{0.9311 \pm 0.0331}$	$\mathbf{0.9368 \pm 0.0275}$	$\mathbf{0.9347 \pm 0.0322}$	$\mathbf{0.9422 \pm 0.0336}$
	Accuracy	0.9471 ± 0.0170	0.9500 ± 0.0138	0.9424 ± 0.0178	0.9448 ± 0.0191
	F1	0.9285 ± 0.0257	0.9330 ± 0.0198	0.9231 ± 0.0278	0.9264 ± 0.0285
	Best k	12.7500 ± 12.6960	5.4000 ± 0.8602	5.6500 ± 1.2359	6.0000 ± 1.6432
Diagnostic II	Recall	$\mathbf{0.9104 \pm 0.0452}$	$\mathbf{0.9075 \pm 0.0541}$	0.8947 ± 0.0586	$\mathbf{0.9108 \pm 0.0520}$
	Accuracy	0.9490 ± 0.0176	0.9476 ± 0.0192	0.9419 ± 0.0217	0.9490 ± 0.0191
	F1	0.9195 ± 0.0294	0.9170 ± 0.0331	0.9080 ± 0.0353	0.9196 ± 0.0315
	Best k	23.5000 ± 29.6268	6.4500 ± 1.4992	5.4000 ± 0.8602	6.8500 ± 1.5580

Results observed over WBCD Prognostic dataset show all techniques obtained similar statistical results considering the recall. Regarding the accuracy, one can observe the proposed approach obtained the highest average value, together with the borderline SMOTE. Such a result suggests samples generated by O^{2}PF fit better the class distribution, being less prone to false negatives. A similar behavior is observed over the Diagnostic I dataset. Concerning the Diagnostic II dataset, the proposed approach obtained the best results, together with ADASYN and SMOTE, outperforming Borderline SMOTE. Considering the accuracy, O^{2}PF obtained the highest averages.

Despite the good performance achieved by the oversampling approach using O^{2}PF, the baseline algorithms have shown statistical similarity in most of the cases. Sparse clusters with low density may be responsible for introducing new samples that are distant from the original distribution of the minority class, a situation that would be likely to produce outliers in the new resample dataset. Such behavior may also influence in a moderate recall, as observed in Tables II.

C. General Purpose Medical Datasets Results

Table III presents the results obtained over the three general purpose medical datasets, i.e., Diabetic Retinopathy Debrecen, Cervical Cancer and Mammographic Mass.

Considering Diabetic Retinopathy Debrecen, the proposed approach outperformed the average recall overall techniques, although SMOTE and Borderline SMOTE achieved similar results. Similar behavior is observed over the F1 metric. On the other hand, O^{2}PF, SMOTE and ADASYN obtained the best results regarding the Cervical Cancer dataset. Since both Borderline SMOTE and ADASYN are variants of SMOTE, they are expected to perform differently over different scenarios. However, as observed in most of the experiments, they generally are outperformed by SMOTE technique itself,

TABLE III: Results considering General Purpose Datasets.

Dataset	Statistics	O^{2}PF	SMOTE	Borderline SMOTE	ADASYN
DRD	Recall	$\mathbf{0.6086 \pm 0.0434}$	$\mathbf{0.5831 \pm 0.0476}$	$\mathbf{0.5855 \pm 0.0498}$	0.5716 ± 0.0487
	Accuracy	0.5934 ± 0.0327	0.5983 ± 0.0353	0.5974 ± 0.0343	0.5957 ± 0.0354
	F1	0.6092 ± 0.0302	0.5799 ± 0.0465	0.5803 ± 0.0460	0.5734 ± 0.0469
	Best k	26.2500 ± 26.9664	6.8500 ± 1.6515	7.5000 ± 1.7748	5.0000 ± 0.0000
CC	Recall	$\mathbf{0.5834 \pm 0.2308}$	$\mathbf{0.6287 \pm 0.1505}$	0.5791 ± 0.1811	$\mathbf{0.6237 \pm 0.1560}$
	Accuracy	0.9388 ± 0.0212	0.9403 ± 0.0157	0.9403 ± 0.0165	0.9415 ± 0.0138
	F1	0.5330 ± 0.1548	0.5673 ± 0.0900	0.5414 ± 0.1109	0.5675 ± 0.0905
	Best k	13.2500 ± 10.8714	6.5500 ± 1.4654	6.5500 ± 1.7168	6.0000 ± 1.5166
MM	Recall	$\mathbf{0.6292 \pm 0.0924}$	$\mathbf{0.6757 \pm 0.0739}$	$\mathbf{0.6876 \pm 0.0608}$	$\mathbf{0.6614 \pm 0.0706}$
	Accuracy	0.6769 ± 0.0531	0.6738 ± 0.0503	0.6852 ± 0.0426	0.6876 ± 0.0396
	F1	0.6439 ± 0.0639	0.6591 ± 0.0545	0.6710 ± 0.0479	0.6631 ± 0.0567
	Best k	48.0000 ± 34.2199	7.1000 ± 1.4799	7.2500 ± 1.5772	6.1500 ± 1.5898

considering the average values. Regarding Mammographic Mass datasets, all techniques performed in a very much alike fashion, obtaining similar results.

One can notice that all three datasets present a challenging task since no technique reached a 0.7 recall. Unusual behavior is observed over the Cervical Cancer dataset, whose all techniques obtained an approximate accuracy of 94%, despite the recall below 0.63. Such behavior may suggest samples from the minority class are distributed among a subcluster from majority class, therefore providing high accuracy despite the low recall.

D. O^{2}PF Hyperparameter Selection

O^{2}PF requires a proper selection of a single hyperparameter, the k_{max}, which is employed in the clustering process. Such a hyperparameter, however, is way less sensitive when compared to a proper selection of the best k, performed by the other techniques. Figure 3 depicts a grid search considering a proper selection of those hyperparameters for each technique. Notice the central line describes the average value over the validation dataset, while the broader area describes the standard deviation. Notice even though O^{2}PF considers a very wider interval, i.e., ranging from $[5, 100]$, most of the time, the results outperform the other techniques, which assume a shorter interval between $[5, 10]$.

VI. Conclusion

This paper presented an oversampling approach based on the unsupervised Optimum-Path Forest Algorithm. The proposed O^{2}PF showed to be capable of handling the class imbalance problem through a simple and effective procedure that generates new synthetic samples based on the normal distribution of the feature vectors inside each cluster. The experiments performed in three datasets of breast cancer, apart from three complementary medical issue datasets, showed that the O^{2}PF approach demonstrated similar or superior results when compared to the baseline methods already proposed in the literature.

Notwithstanding, the effectiveness of the proposed approach may still suffer in synthesizing new samples based on low-density clusters, a situation that may introduce noise samples in the training set and, consequently, affect creating of the

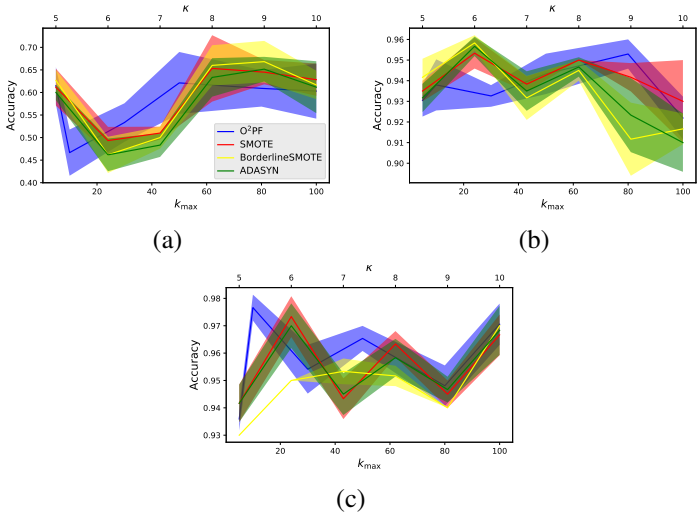

Fig. 3: Grid searching for a proper selection of the techniques' hyperparameter over the validation for WBCD (a) Prognostic, (b) Diagnostic I, and (c) Diagnostic II datasets.

prediction model. Future studies will be conducted to overcome the influence of sparse clusters with low density in the process of synthesizing new outliers. Moreover, experiments with a multiclass problem will also be performed in subsequent investigations.

References

[1] D. S. Jodas, A. S. Pereira, and J. M. R. Tavares, "A review of computational methods applied for identification and quantification of atherosclerotic plaques in images," *Expert Systems with Applications*, vol. 46, pp. 1–14, 2016.

[2] L. C. Ribeiro, L. C. Afonso, and J. P. Papa, "Bag of samplings for computer-assisted parkinson's disease diagnosis based on recurrent neural networks," *Computers in biology and medicine*, vol. 115, p. 103477, 2019.

[3] N. I. Yassin, S. Omran, E. M. E. Houby, and H. Allam, "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review," *Computer Methods and Programs in Biomedicine*, vol. 156, pp. 25–45, 2018.

[4] W. H. Organization. Breast cancer. [Online]. Available: https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

[5] M. Xian, Y. Zhang, H. Cheng, F. Xu, B. Zhang, and J. Ding, "Automatic breast ultrasound image segmentation: A survey," *Pattern Recognition*, vol. 79, pp. 340 – 355, 2018.

[6] W. Wieclawek, M. Rudzki, A. Wijata, and M. Galinska, "Preliminary development of an automatic breast tumour segmentation algorithm from ultrasound volumetric images," in *Information Technology in Biomedicine*, E. Pietka, P. Badura, J. Kawa, and W. Wieclawek, Eds. Cham: Springer International Publishing, 2019, pp. 77–88.

[7] L. Liu, K. Li, W. Qin, T. Wen, L. Li, J. Wu, and J. Gu, "Automated breast tumor detection and segmentation with a novel computational framework of whole ultrasound images," *Medical & Biological Engineering & Computing*, vol. 56, pp. 183–199, 2017.

[8] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset," in *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*, 2018, pp. 5–9.

[9] L. A. Passos, C. Santos, C. R. Pereira, L. C. S. Afonso, and J. P. Papa, "A hybrid approach for breast mass categorization," in *ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing*. Springer, 2019, pp. 159–168.

[10] P. B. Ribeiro, L. A. Passos, L. A. da Silva, K. A. da Costa, J. P. Papa, and R. A. Romero, "Unsupervised breast masses classification through optimum-path forest," in *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*. IEEE, 2015, pp. 238–243.

[11] L. M. Rocha, F. A. M. Cappabianco, and A. X. Falcão, "Data clustering as an optimum-path forest problem with applications in image analysis," *International Journal of Imaging Systems and Technology*, vol. 19, no. 2, pp. 50–68, 2009.

[12] L. C. S. Afonso, L. A. Passos, and J. P. Papa, "Enhancing brain storm optimization through optimum-path forest," in *2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE, 2018, pp. 000 183–000 188.

[13] R. R. Guimarães, L. A. Passos, R. Holanda Filho, V. H. C. de Albuquerque, J. J. Rodrigues, M. M. Komarov, and J. P. Papa, "Intelligent network security monitoring based on optimum-path forest clustering," *IEEE Network*, vol. 33, no. 2, pp. 126–131, 2018.

[14] R. W. R. Souza, J. V. C. Oliveira, L. A. Passos, W. Ding, J. P. Papa, and V. H. Albuquerque, "A novel approach for optimum-path forest classification using fuzzy logic," *IEEE Transactions on Fuzzy Systems*, 2019.

[15] M. Zhu, J. Xia, X. Jin, M. Yan, G. Cai, J. Yan, and G. Ning, "Class weights random forest algorithm for processing class imbalanced medical data," *IEEE Access*, vol. 6, pp. 4641–4652, 2018.

[16] T. Razzaghi, I. Safro, J. Ewing, E. Sadrfaridpour, and J. D. Scott, "Predictive models for bariatric surgery risks with imbalanced medical datasets," *Annals of Operations Research*, vol. 280, no. 1-2, pp. 1–18, 2019.

[17] M. Rezaei, H. Yang, and C. Meinel, "Recurrent generative adversarial network for learning imbalanced medical image semantic segmentation," *Multimedia Tools and Applications*, pp. 1–20, 2019.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[19] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing*. Springer, 2005, pp. 878–887.

[20] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 2008, pp. 1322–1328.

[21] J. P. Papa, A. X. Falcão, and C. T. N. Suzuki, "Supervised pattern classification based on optimum-path forest," *International Journal of Imaging Systems and Technology*, vol. 19, no. 2, pp. 120–131, 2009.

[22] J. P. Papa, A. X. Falcão, V. H. C. De Albuquerque, and J. M. R. Tavares, "Efficient supervised optimum-path forest classification for large datasets," *Pattern Recognition*, vol. 45, no. 1, pp. 512–520, 2012.

[23] A. X. Falcão, J. Stolfi, and R. A. Lotufo, "The image foresting transform: theory, algorithms, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 19–29, 2004.

[24] K. C. Ciesielski, A. X. Falcão, and P. A. V. Miranda, "Path-value functions for which dijkstra's algorithm returns optimal mapping," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 7, pp. 1025–1036, Sep 2018.

[25] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[26] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 1990.

[27] B. Antal and A. Hajdu, "An ensemble-based system for automatic screening of diabetic retinopathy," *Knowledge-based systems*, vol. 60, pp. 20–27, 2014.

[28] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," in *Iberian conference on pattern recognition and image analysis*. Springer, 2017, pp. 243–250.

[29] M. Elter, R. Schulz-Wendtland, and T. Wittenberg, "The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process," *Medical physics*, vol. 34, no. 11, pp. 4164–4172, 2007.

[30] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[31] G. H. de Rosa, J. P. Papa, and A. X. Falcão, "Opfython: A python-inspired optimum-path forest classifier," 2020.