Supervised deep learning embeddings for the prediction of cervical cancer diagnosis

Kelwin Fernandes[1,2], **Davide Chicco**[3], **Jaime S. Cardoso**[1,2], and **Jessica Fernandes**[4]

[1]**Instituto de Engenharia de Sistemas e Computadores Tecnologia e Ciência**
(INESC TEC), Porto, Portugal
[2]**Universidade do Porto, Porto, Portugal**
[3]**Princess Margaret Cancer Centre, Toronto, Ontario, Canada**
[4]**Universidad Central de Venezuela, Caracas, Venezuela**

Corresponding author:
Kelwin Fernandes[1,2]

Email address: kafc@inesctec.pt

ABSTRACT

Cervical cancer remains a significant cause of mortality all around the world, even though it can be prevented and cured by surgical resections in early stages. Providing universal and efficient access to cervical screening programs is a challenge that requires identifying vulnerable individuals in the population, among other steps. In this work, we present a computationally automated strategy for predicting the outcome of the patient biopsy given risk patterns from individual medical records. We propose a machine learning technique that allows a joint and fully supervised optimization of dimensionality reduction and classification models. We also build a model able to highlight relevant properties in the low dimensional space, to ease the classification of patients. We instantiated the proposed approach with deep learning architectures, and achieved accurate prediction results (top area under the curve $AUC = 0.6875$) which outperform previously developed methods, such as denoising autoencoders. Additionally, we explored some clinical findings from the embedding spaces, and we validated them through the medical literature, making them reliable for physicians and biomedical researchers.

INTRODUCTION

Despite the possibility of prevention with regular cytological screening, cervical cancer remains a significant cause of mortality in low-income countries [22]. The cervical tumor is the cause of more than 500,000 cases per year, and kills more than 250,000 patients in the same period, on world basis [16]. However, cervical cancer can be prevented by means of the human papillomavirus infection (HPV) vaccine, and regular low-cost screening programs [6]. The two most widespread techniques in screening programs are conventional or liquid cytology and colposcopy [16, 37, 17, 49]. Furthermore, this cancer can be cured by removing the affected tissues when identified in early stages [16, 6], in most cases.

The development of cervical cancer is usually slow and preceded by abnormalities in the cervix (dysplasia). However, the absence of early stage symptoms might incur in carelessness prevention. Additionally, in developing countries, resources lack and patients usually have poor adherence to routine screening due to low problem awareness.

While improving the resection of lesions in the first visits has a direct impact on patients that attend screening programs, the most vulnerable populations have poor or even non-existent adherence to treatment programs. Scarce awareness of the problem and patients' discomfort with the medical procedure might be the main causes of this problem. Furthermore, in low-income countries, this issue can be due to lack of access to vulnerable populations with difficult access to information and medical centers. Consequently, the computational prediction of individual patient's risk has a key role in this context. Identifying patients with the highest risk of developing cervical cancer can improve the targeting efficacy of cervical cancer screening programs: our software performs this operation computationally in few minutes by producing accurate prediction scores.

⁴⁷ Fernandes and colleagues performed a preliminary attempt to tackle the problem of predicting the
⁴⁸ patient's risk to develop cervical cancer through machine learning software [17]. In that project, the
⁴⁹ authors employed transfer learning strategies for the prediction of the individual patient's risk on a dataset
⁵⁰ of cervical patients' medical tests. They focused on transferring knowledge between linear classifiers on
⁵¹ similar tasks, to predict the patient's risk [17].

⁵² Given the high sparsity of the associated risk factors in the population, dimensionality reduction
⁵³ techniques can improve the robustness of the machine learning predictive models. However, many projects
⁵⁴ that take advantage of dimensionality reduction and classification use suboptimal approaches, where each
⁵⁵ component is learned separately [28, 4, 25] .

⁵⁶ In this work, we propose a joint strategy to learn the low-dimensional space and the classifier itself in
⁵⁷ a fully supervised way. Our strategy is able to reduce class overlap by concentrating observations from
⁵⁸ the healthy patients class into a single point of the space, while retaining as much information as possible
⁵⁹ from the patients with high risk of developing cervical cancer.

⁶⁰ We based our prediction algorithm on artificial neural networks (ANNs), which are machine learning
⁶¹ methods able to discover non-linear patterns by means of aggregation of functions with non-linear
⁶² activations. A recent trend in this field is deep learning [26], which involves large neural network
⁶³ architectures with successive applications of such functions. Deep learning, in fact, has been able to
⁶⁴ provide accurate predictions of patients diagnosis in multiple medical domains [49, 9, 15, 5, 2]. We
⁶⁵ applied our learning scheme to deep variational autoencoders and feedforward neural networks. Finally,
⁶⁶ we explored visualization techniques to understand and validate the medical concepts captured by the
⁶⁷ embeddings.

⁶⁸ We organize the rest of the paper as follows. After this Introduction, we describe the proposed method
⁶⁹ and the dataset analyzed in Methods and Dataset. Afterwards, we describe the computational prediction
⁷⁰ results in Results, and the model outcome interpretation in Discussion, and we conclude the manuscript
⁷¹ outlining some conclusion and future development.

METHODS

⁷³ High dimensional data can lead to several problems: in addition to high computational costs (in memory
⁷⁴ and time), it often leads to overfitting [47, 7, 34]. Dimensionality reduction can limit these problems and,
⁷⁵ additionally, can help for the visualization and interpretation of the dataset, because it allows researchers
⁷⁶ to focus on a reduced number of features. For these reasons, we decided to map the original dataset
⁷⁷ features into a reduced dimensionality before performing the classification task.

⁷⁸ Generally, to tackle high-dimensional classification problems, machine learning traditional approaches
⁷⁹ attempt to reduce the high-dimensional feature space to a low-dimensional one, to facilitate the posterior
⁸⁰ fitting of a predictive model. In many cases, researchers perform these two steps separately, deriving
⁸¹ suboptimal combined models [28, 4, 25]. Moreover, since dimensionality reduction techniques are often
⁸² learned in an unsupervised fashion, they are unable to preserve and exploit the separability between
⁸³ observations from different classes.

⁸⁴ In dimensionality reduction, researchers use two categories of objective functions: one for maximizing
⁸⁵ the model capability of recovering the original feature space from the compressed low dimensional one,
⁸⁶ and another one for maximizing the consistency of pairwise similarities in both high and low dimensional
⁸⁷ spaces.

⁸⁸ Since defining a similarity metric in a high-dimensional space might be difficult, we limit the scope
⁸⁹ of this work to minimizing the reconstruction loss. In this sense, given a set of labeled input vectors
⁹⁰ $X = \{x_1, x_2, \ldots, x_n\}$, where $x_i \in \mathbb{R}^d, \forall i \in 1, \ldots, n$ and Y is a vector with the labels associated to each
⁹¹ observation, we want to obtain two functions $C : \mathbb{R}^d \rightarrow \mathbb{R}^m$ and $D : \mathbb{R}^m \rightarrow \mathbb{R}^d$ such that $m < d$ and that
⁹² minimizes the following loss:

$$L_r(C, D, X) = \frac{1}{|X|} \sum_{x \in X} ((D \circ C)(x)) - x)^2 \qquad (1)$$

⁹³ Namely, the composition ($\circ$) of the compressing (C) and decompressing (D) functions approximate
⁹⁴ the identity function.

⁹⁵ In the following sections, we describe the proposed dimensionality reduction technique and its
⁹⁶ instantiation to deep learning architectures.

Joint dimensionality reduction and classification

Since our final goal is to classify the data instances (observations), we need to achieve a good low-dimensional mapping and build the classifier independently. Thereby, we propose a joint loss function that minimizes the trade-off between data reconstruction and classification performance:

$$
\begin{aligned}
L(M,C,D,X,Y) &= L_c((M \circ C)(X),Y) + \\
&\quad \lambda L_r(C,D,X)
\end{aligned}
\tag{2}
$$

where M is a classifier that receives as input the vectors in the low dimensional space $(C(X))$, L_c is a classification loss function such as categorical cross-entropy, and $\lambda \geq 0$. In this case, we focus on the classification performance while using Equation 1 as a regularization factor of the models of interest. Hereafter, we will denote this method as semi-supervised dimensionality reduction.

Fully supervised embeddings

The previously proposed loss function consists of two components: a supervised component given by the classification task, and an unsupervised component given by the low-dimensional mapping. However, the scientific community aims at understanding the properties captured in the embeddings, especially on visual and text embeddings [23, 27]. Moreover, inducing properties in the low-dimensional space can improve the class separability. To apply this enhancement, we introduce partial supervision in the L_r loss.

We can explore these properties by learning the dimensionality reduction process in an supervised way. Namely, learning a bottleneck supervised mapping function $((D \circ C)(x) \approx M(x,y))$ instead of the traditional identity function $((D \circ C)(x) \approx x)$ used in reconstruction-based dimensionality reduction techniques. The reconstruction loss $L_r(C,D,X)$ becomes:

$$
L_M(C,D,X,Y) = \frac{1}{|X|} \sum_{\langle x,y \rangle \in X,Y} ((D \circ C)(x)) - M(x,y))^2
\tag{3}
$$

where $M(x)$ is the desired supervised mapping.

For instance, to facilitate the classification task, removing the overlap between both classes should be captured in low-dimensional spaces. Without loss of generality, we assume that the feature space is non-negative. Thereby we favor models with high linear separability between observations by using the mapping function Equation 4 in Equation 3.

$$
Sym(x,y) = \begin{cases} x & \text{, if } y \\ -x & \text{, if } \neg y \end{cases}
\tag{4}
$$

In our application, if all the features are non-negative, the optimal patient's behavior associates to the zero vector with total lack of risk patterns. On the other hand, a patient with high feature values is prone to have cancer. Within the context of cervical cancer screening, we propose the mapping given by Equation 5, where the decoded version of the healthy patients is the zero vector. This idea resembles the fact that their risk conduct has not contributed to the disease occurrence. On the other hand, we mapped ill patients to their original feature space, for promoting the low-dimensional vectors to explain the original risk patterns that originated the disease.

$$
Zero(x,y) = \mathbb{1}(y) \cdot x
\tag{5}
$$

While the definition of the properties of interest to be captured by the low-dimensional space is application-dependent, the strategy to promote such behavior can be adapted to other contexts.

Deep supervised autoencoders

Autoencoders are special cases of deep neural networks for dimensionality reduction [9, 48]. They can be seen as general feedforward neural networks with two main sub-components: the first part of the neural network is known as the *encoder*, and its main purpose is to compress the feature space. The neural

network achieves this step by using hidden layers with less units than the input features, or by enforcing
sparsity in the hidden representation. The second part of the neural network, also known as the *decoder*,
behaves in the opposite way, and tries to approximate the inverse encoding function. While these two
components correspond to the *C* and *D* functions in Equation 1, respectively, they can be broadly seen as a
single artificial neural network that learns the identity function through a bottleneck, either a low number
of units, or through sparse activations. Autoencoders are usually learned in an unsupervised fashion by
minimizing the quadratic error (Equation 1).
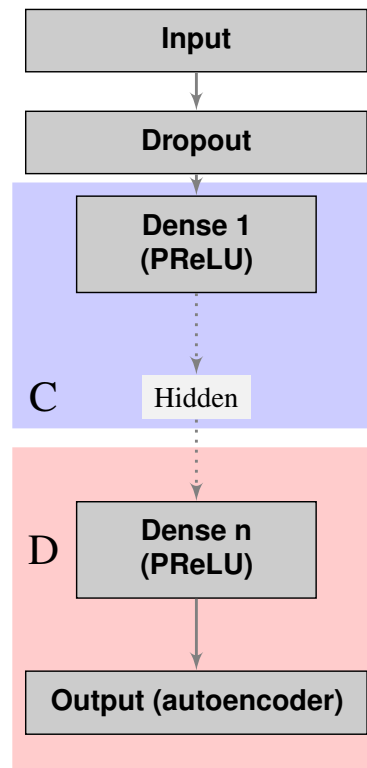
Figure 1. Deep denoising autoencoder. The blocks in blue and red represent the
encoding (*C*) and decoding (*D*) components of the network respectively.

Denoising autoencoders (DA) represent a special case of deep autoencoders that attempt to reconstruct
the input vector given a corrupted version of it [48]. Denoising autoencoders can learn valuable repre-
sentations even on presence of noise. Scientists can experiment this task by adding an artificial source
of noise in the input vectors. In the neural network architecture (Figure 1), we also included a dropout
layer after the input layer that randomly turns off at most one feature per patient [43]. Thereby, we aim to
build stable classifiers that produce similar outcomes for patients with small differences in their historical
records. Furthermore, we aim at producing stable decisions when patients lie on a subset of the answers
to the doctors' questions during the medical visit, by indicating absence of a given risk behavior (for
example, high number of sexual partners, drug consumption, and others). We use a Parametric Rectifier
Linear Unit (PReLU) [20] as activation function in the hidden layers of our architectures (Figure 1).
PReLU is a generalization of standard rectifier activation units, which can improve model fitting with low
additional computational cost [20].

The loss functions (Equation 2 and Equation 3) can learn a joint classification and encoding-decoding
network in a multitask fashion (Figure 2). Additionally, to allow the neural network to use either the learned
or the original representation, we include a *bypass layer* that concatenates the hidden representation with
the corrupted input. In the past, researchers have used this technique in biomedical image segmentation
with U-net architectures [39], to recover possible losses in the compression process, and to reduce the
problem of vanishing gradients. We use this *bypass layer* with cross-validation.

In a nutshell, our contribution can be summarized as follows: (i) we formalized a loss function to
handle dimensionality reduction and classification in a joint fashion, leading to an global optimal pipeline;
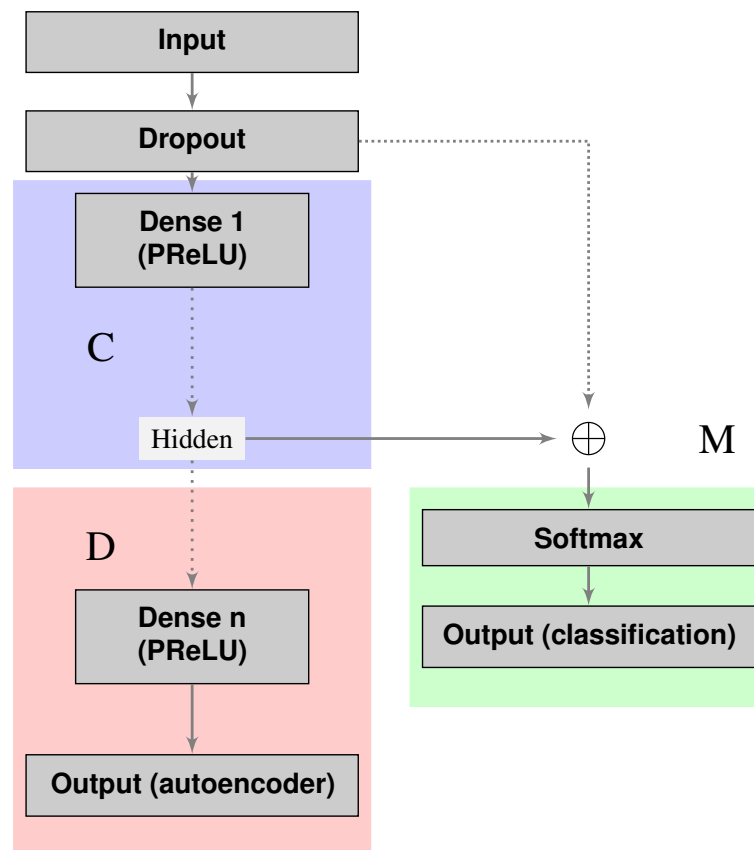
Figure 2. Supervised Deep Embedding Architecture. The blocks in blue, red and green represent the encoding (*C*), decoding (*D*) and classification (*M*) components of the network respectively.

160 (ii) in order to induce desired properties on the compressed space, we proposed a loss that measures
161 the model's capability to recreate a mapping with the desired property instead of the identity function
162 usually applied in in dimensionality reduction; (iii) we showed that multitask autoencoders based on
163 neural networks can be used as a specific instance to solve this problem, and we instantiated this idea to
164 model an individual patient's risk of having cervical cancer.

DATASET

166 The dataset we analyze contains medical records of 858 patients, and covers a random sampling of patients
167 between 2012 and 2013 who attended the gynecology service at Hospital Universitario de Caracas in
168 Caracas, Venezuela. Most of the patients belong to the lowest socioeconomic status (Graffar classification:
169 IV-V [18]) with low income and educational level, being the population with the highest risk. The age
170 of the patients spans between 13 and 84 years old (27 years old on average). All patients are sexually
171 active and most of them (98%) have been pregnant at least once. The screening process covers traditional
172 cytology, the colposcopic assessment with acetic acid and the Schiller test (Lugol's iodine solution) [17].
173 The medical records include the age of the patient, sexual activity (number of sexual partners and age of
174 first sexual intercourse), number of pregnancies, smoking behavior, use of contraceptives (hormonal and
175 intrauterine devices) and historical records of sexually transmitted diseases (STDs) (Table 1). Hence, we
176 encoded the features denoted by bool $\times\, T$, $T \in \{\text{bool}, \text{int}\}$ as two independent values: whether or not the
177 patient answered the question and, if she did, the answered value. In some cases, the patients decided not
178 to answer some questions for privacy concerns. This behavior is often associated to risk behaviors being a
179 relevant feature to explore when modeling risk patterns. Therefore, we added a flag feature that allows the
180 model to identify if the question was answered or not after missing value imputation. We encoded the
181 categorical features using the one-of-K scheme. The hospital anonymized all the records before releasing

Feature	Type	Feature	Type
Age	int	IUD (years)	int
Number of sexual partners	bool × int	Sexually transmitted diseases (STDs) (yes/no)	bool × bool
Age of first sexual intercourse	bool × int	Number of STDs	int
Number of of pregnancies	bool × int	Diagnosed STDs	categorical
Smokes (yes/no)	bool × bool	STDs (years since first diagnosis)	int
Smokes (years & packs)	int × int	STDs (years last diagnosis)	int
Hormonal Contraceptives (yes/no)	bool	Previous cervical diagnosis (yes/no)	bool
Hormonal Contraceptives (years)	int	Previous cervical diagnosis (years)	int
Intrauterine device (IUD) (yes/no)	bool	Previous cervical diagnosis	categorical

Table 1. Features names and data type acquired in the risk factors dataset [17]. int: integer. bool: boolean.

182 the dataset. The dataset is now publically available on the Machine Learning Repository website of the
183 University of California Irvine (UCI ML) [45], which also contains a description of the features [46].
184 To avoid problems of the algorithm behavior related to different value ranges of each feature, we
185 scaled all the features in our experiments using $[0, 1]$ normalization, and we inputed missing data using
186 the average value [7]. While more complex pre-processing schemes could be introduced, such as inferring
187 the missing value with a k-nearest neighbor model [40], we decided to use this methodology to avoid
188 additional complexity that would make it difficult to fairly compare the explored techniques. In most
189 cases, the features positively correlate to the cancer variable, with 0 representing the lack of that risk
190 pattern and 1 representing the maximum risk.

RESULTS

192 We measured the performance of the proposed methods with the area under the *Precision-Recall* (PR)
193 curves [12, 7] and the *logistic loss* (also known as *cross-entropy loss*) function.
194 As baseline, we use a deep feed-forward neural network with a softmax activation in the output layer.
195 The remaining parameters (such as the initial dropout layer, depth and optimization algorithm) conform to
196 the ones used in the proposed methodologies (Table 2). The main hyper-parameters related to the network
197 topology are the depth and width, which define the number of layers in the architecture and the size of the
198 low-dimensional representation.
199 We used a stratified 10-fold cross-validation in the assessment of the proposed methods. We optimized
200 the neural networks by using the RMSProp optimization strategy [44] for a maximum number of 500
201 epochs, with early stopping after 100 iterations without improvement and a batch size of 32. We validated
202 these parameters empirically, being enough to ensure model convergence in all cases. We also validated the
203 performance of other optimization strategies such as *Adam* and stochastic gradient descent. However, we
204 did not observe any gain in terms of predictive performance or convergence. We use sparse autoencoders
205 by adding an L_1 penalization term, to ensure that each unit combines a small subset of risk factors, as
206 would be done by a human expert.
207 We fine-tuned all the hyper-parameters using a grid search strategy with nested stratified 3-fold
208 cross-validation. In this sense, we validated the performance of each network configuration on three
209 training-validation partitions and choose the one that maximizes the area under the Precision-Recall curve.
210 Then, for the best configuration, we re-trained the model using the entire training set. We chose the size
211 of the low-dimensional space as part of this nested cross-validation procedure, and chose empirically the
212 parameters related to the optimization algorithm (that are strategy, number of epochs, early stopping).

Parameter	Values
depth	$\{1, \ldots, 6\}$
width	$\{10, 20\}$
regularization	$\{0.01, 0.1\}$
bypass usage	$\{false, true\}$

Table 2. Set of possible options for fine-tuning each parameter.

To recreate the decisions made by the physician at different configurations of the screening process, we consider the observability of all possible subsets of screening outcomes when predicting the biopsy results. Thereby, we cover scenarios where only behavioral and demographic information is observable (first line of each table with empty subset) up to settings where cytology and colposcopy (Hinselmann and Schiller) results are available.

Diagnosis prediction results

Our proposed architectures with embedding regularization achieved the best diagnosis prediction results in most cases (Table 3 and Table 4) when compared with other neural network approaches. Furthermore, the fully supervised embeddings improved the performance of the semi-supervised approach (Equation 2), through both the strategies (symmetric and zero mapping). The relative gains in terms of area under the Precision-Recall curve depend on the subset of observable modalities, ranging from 30.7% when only medical records are observed to 3.3% when the outcome of all the screening procedures is known.

Using a paired difference Student's t-test [33] with a 95% confidence level, zero-mapping methodology achieved better results than the baseline and semi-supervised learning schemes. We found no statistical differences between the symmetry and zero mappings.

We validated the performance of traditional machine learning models such as support vector machines (SVM) with radial basis function (RBF) kernel [41], k-nearest neighbors [35], and decision trees [38]. In general, the proposed models surpassed the performance of the classical methodologies in terms of area under the Precision-Recall curve. The SVM model achieved better logarithmic loss given the post-processing of its scores using the Logistic Regression model that directly optimize this metric. Further improvements could be observed by post-processing the outcome of the other strategies.

Subset	Baseline	Semi	Sym	Zero	SVM	k-NN	DecTree
	0.1334	0.1424	0.1534	0.1744	0.0877	0.0345	**0.1941**
C	0.1998	0.1853	0.2115	0.2174	0.1550	**0.3033**	0.2560
H	0.4536	0.4459	0.4407	**0.4625**	0.4192	0.3885	0.3616
S	0.6416	0.6411	0.6335	**0.6668**	0.5905	0.5681	0.6242
CH	0.4752	0.4684	**0.4754**	0.4609	0.4423	0.4095	0.4023
CS	0.6265	**0.6424**	0.6388	0.5985	0.6205	0.5379	0.6089
HS	0.6200	**0.6356**	0.6277	0.5864	0.6199	0.6335	0.5956
CHS	0.6665	0.6351	**0.6875**	0.6404	0.6374	0.6653	0.5542
best	0	2	2	2	0	1	1

Table 3. Performance of the proposed architectures in terms of Area Under the Precision-Recall curve. The subset of observable screening strategies include: Cytology (C), Hinselmann (H) and Schiller (S). Baseline: deep feed-forward neural network. Semi: semi-supervised dimensionality reduction (Equation 2). Sym: symmetry mapping dimensionality reduction (Equation 4). Zero: zero mapping dimensionality reduction (Equation 5). SVM: support vector machine. k-NN: k-nearest neighbors. DecTree: decision tree.

The gains achieved by the mapping-based supervised embeddings happen because the proposed fully-supervised strategies aim to reduce the overlap between observations from both classes. In the past, researchers showed that class overlap has higher correlation with the model performance than the imbalance ratio in highly unbalanced datasets [11]. The visualization of the embeddings through the t-distributed stochastic neighbor embedding (t-SNE) [31] confirms this aspect, because in t-SNE fully supervised embeddings achieve better separability and less overlapping clusters (Figure 3, Figure 4, Figure 5, and Figure 6).

For visualization purposes, we are using t-SNE based upon neighborhood similarities, since learning a valuable representation in a 2-dimensional space raises difficulties. Moreover, because of the high dimensionality of our embeddings, their reduction capabilities rely on their sparsity.

Subset	Baseline	Semi	Sym	Zero	SVM	*k*-NN	DecTree
	0.3004	0.2708	0.2657	0.2716	**0.2421**	4.3670	4.1889
C	0.2829	0.2757	0.2868	**0.2609**	0.2614	2.6884	3.5001
H	0.2169	0.2274	0.2422	0.2031	0.1984	0.7178	3.2175
S	0.1710	0.1475	0.1489	0.1359	**0.1273**	0.9366	1.6893
CH	0.2210	**0.2054**	0.2286	0.2123	0.2196	1.0477	2.8509
CS	0.1594	0.1469	**0.1240**	0.1464	0.1248	0.4036	1.7687
HS	0.1632	0.1786	0.1615	0.1622	**0.1225**	0.3238	1.8098
CHS	0.1563	0.1577	0.1494	0.1514	**0.1099**	0.4037	1.8906
best	0	1	1	1	4	0	0

Table 4. Performance of the proposed architectures in terms of logarithmic loss. The subset of observable screening strategies include: Cytology (C), Hinselmann (H) and Schiller (S). Area Under the Precision-Recall curve. Baseline: deep feed-forward neural network. Semi: semi-supervised dimensionality reduction (Equation 2). Sym: symmetry mapping dimensionality reduction (Equation 4). Zero: zero mapping dimensionality reduction (Equation 5). SVM: support vector machine. *k*-NN: *k*-nearest neighbors. DecTree: decision tree.

Results in other applications

To observe the impact of our method, we validated the performance of the aforementioned model architectures on several biomedical datasets available on the UC Irvine Machine Learning Repository. Thus, we assessed the models performance on nine datasets. The machine learning models we proposed achieved high prediction results, being the zero-mapping approach the best model in most cases (Table 5 and Table 6). This outcome suggests that mapping the majority class to a unique point in the space might improve the learning effectiveness in unbalanced settings. This idea draws a link between binary and one-class classification, and we plan to explore it more in the future.

Dataset		Baseline	Semi	Sym	Zero
Breast Cancer	[32]	0.9795	**0.9864**	0.9835	0.9856
Mammography	[14]	**0.8551**	0.8539	0.8533	0.8489
Parkinson	[29]	0.9517	0.9526	0.9573	**0.9604**
Pima Diabetes	[42]	0.7328	0.7262	0.7095	**0.7331**
Lung Cancer	[21]	0.7083	0.6042	0.6927	**0.8021**
Cardiotocography	[3]	0.9948	0.9948	0.9925	**0.9958**
SPECTF Heart	[24]	0.9470	**0.9492**	0.9462	0.9463
Arcene	[19]	0.8108	0.8433	**0.8900**	0.8455
Colposcopy QA	[17]	0.7760	0.8122	0.7961	**0.8470**
best		1	2	1	5

Table 5. Performance of the proposed architectures on other datasets downloaded from UC Irvine Machine Learning Repository [45], measured through the Area Under the Precision-Recall curve

DISCUSSION

As shown in the Results section, our deep learning algorithm can predict cervical cancer diagnosis with high accuracy. To further understand the clinical interpretability of our prediction model, we investigated which dataset risk features have the highest impact in the cervical cancer diagnosis for the patients.

In fact, pre-invasive intra-epithelial lesions of the cervix and cervical cancer relate to HPV infection of
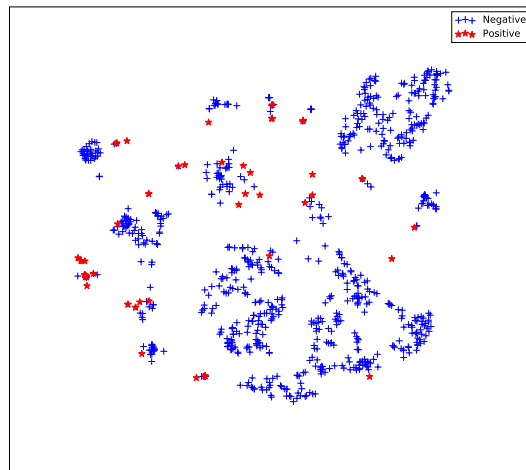
Figure 3. Two dimensional projection of the unsupervised embedding using t-distributed stochastic neighbor embedding (t-SNE) [31].
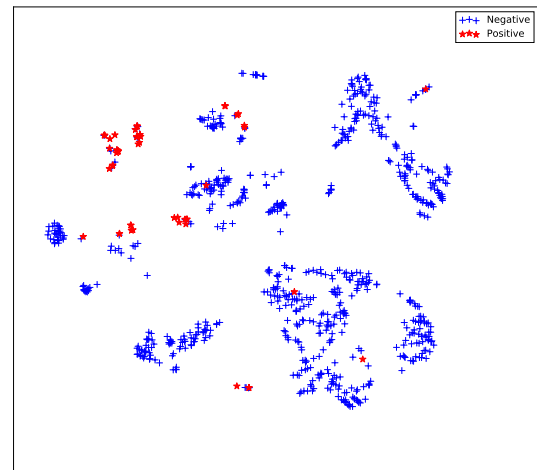
Figure 4. Two dimensional projection of the semi-supervised embedding using t-distributed stochastic neighbor embedding (t-SNE) [31].
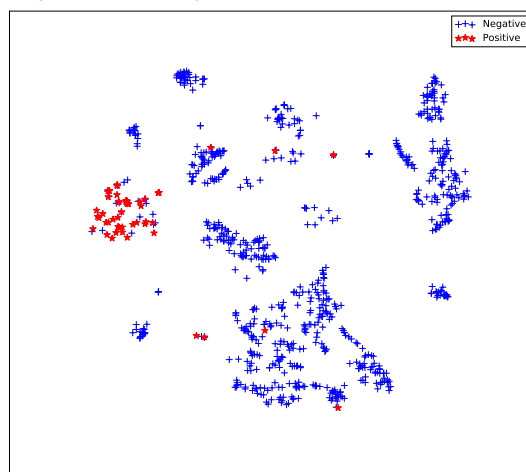
Figure 5. Two dimensional projection of the semi-supervised embedding with symmetry mapping using t-distributed stochastic neighbor embedding (t-SNE) [31].
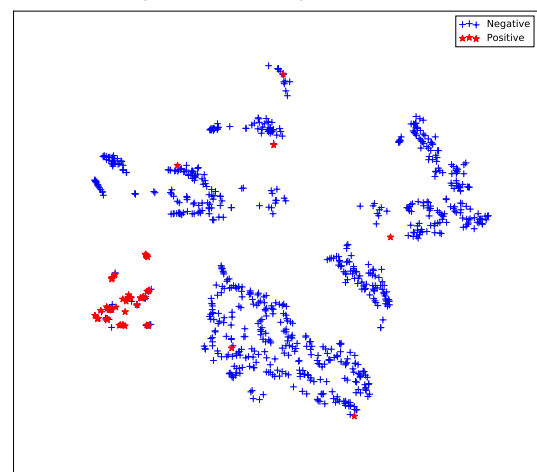
Figure 6. Two dimensional projection of the supervised embedding with zero mapping using t-distributed stochastic neighbor embedding (t-SNE) [31].

257 oncological serotypes that progress to oncological lesions, and multiple factors contribute to this progress
258 without a definite cause-dependent relation. The patterns that have highest acceptance in the literature
259 regard presence of human immunodeficiency virus (HIV) and smoking, followed by sexual risk behaviors
260 such as early sexual initiation, promiscuity, multiple pregnancies, and a history of sexually transmitted
261 infections. Another factor involved is the use of oral contraceptives.
262 From a technical point of view, while black-box machine learning techniques have achieved state-of-
263 the-art results in several applications, the lack of interpretability of the induced models can limit their
264 general acceptance by the medical community. Thus, we tried to understand the relations by our prediction
265 model to corroborate if they are supported by the medical literature.
266 In this context, we studied the impact of the original features on the embedding space to find
267 correlations in the decision process. To determine this impact, we perturbed each feature using all the
268 other values from the feature's domain, and then we computed the maximum impact of the features in the
269 embedded space. Finally, we applied an agglomerative clustering technique to aggregate features with
270 similar impact in the embedding features. From a medical point of view, we validated several properties
271 of interest (Figure 7).
272 For instance, risky sexual patterns such as an early sexual initiation and presence (and lifespan) of

9/12

Dataset		Baseline	Semi	Sym	Zero
Breast Cancer	[32]	0.0984	**0.0888**	0.0966	0.0930
Mammographic	[14]	0.5122	0.5051	0.4973	**0.4822**
Parkinson	[29]	0.3945	0.4042	**0.3883**	0.4323
Pima Diabetes	[42]	0.5269	**0.5229**	0.5250	0.5472
Lung Cancer	[21]	1.1083	0.8017	**0.6050**	0.8328
Cardiotocography	[3]	0.0113	0.0118	0.0116	**0.0110**
SPECTF Heart	[24]	**0.4107**	0.4205	0.4121	0.4196
Arcene	[19]	1.3516	**0.8855**	1.0230	1.1518
Colposcopy QA	[17]	0.5429	0.5406	0.5195	**0.4850**
best		1	3	2	3

Table 6. Performance of the proposed architectures on other datasets downloaded from UC Irvine Machine Learning Repository [45], measured through logarithmic loss.
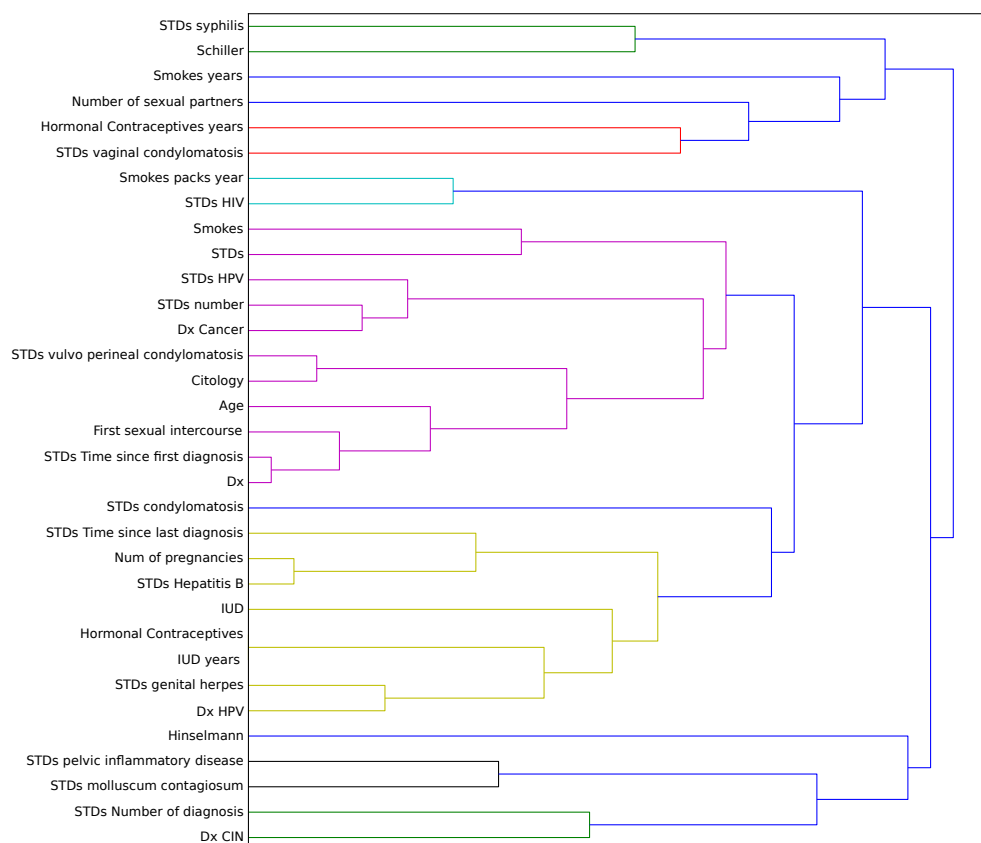
Figure 7. Agglomerative clustering of features by impact on the embedding space.

sexually transmitted diseases (with a special focus on HPV) have the most similar impact in the predictive outcome of the model. Also, smoking habits are associated by the model as having a similar effect as these sexual patterns. These relations were already studied in the medical literature [30, 13].

The similarity between the use of hormonal contraceptives with condylomatosis and the use of intrauterine devices with STDs shows another interesting pattern that has not been quantified yet to the best of our knowledge. These patterns might can be evidence of sexual patterns with high risk.

CONCLUSION

Cervical cancer is still a widespread disease nowadays, and its diagnosis often requires long times and multiple time-consuming clinical exams. In this context, machine learning can provide effective tools to speed up the diagnosis process, by processing high-scale patients' datasets in a few minutes.

In this manuscript, we presented a computational system for the prediction of cervical patient diagnosis, and for the interpretation of its results. Our system consists of a loss function that allows joint optimization of dimensionality reduction, and classification techniques able to promote relevant properties in the embedded spaces. Our deep learning methods predicted the diagnosis of the patients with high accuracy, and their application to other datasets showed that their robustness and effectiveness is not bounded to cervical cancer. Our methods can be used to analyze profiles of patients where the biopsy and potentially other screening results are missing, and is able to predict confidently if they have cervical cancer.

In the future, we plan to employing alternative approaches for data missing imputation, such as oversampling through k-nearest neighbors [40] or latent semantic indexing similarity [8]. We also plan to try alternative prediction models, like probabilistic latent semantic analysis [36]. Finally, we plan to extend our computational system by adding a feature selection step, able to state the most relevant features among the dataset.

SOFTWARE AND DATA AVAILABILITY

The cervical cancer dataset is publicly available on the University of California Irvine Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+(Risk+Factors)

The software code of the methods is available at: https://github.com/kelwinfc/cervical-cancer-screening

We implemented the software in Python 2.7 using the Keras [10] and TensorFlow [1] frameworks, and tested it on a computer running the Linux Ubuntu 16.04 operating system.

ACKNOWLEDGMENTS

The authors thank the Gynecology Service of the Hospital Universitario de Caracas, and Francis Nguyen (Princess Margaret Cancer Centre) for the English proof-reading of this manuscript.

REFERENCES

[1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., and Google Brain (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2016)*, volume 16, pages 265–283.

[2] Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838.

[3] Ayres-de Campos, D., Bernardes, J., Garrido, A., Marques-de Sa, J., and Pereira-Leite, L. (2000). Sisporto 2.0: a program for automated analysis of cardiotocograms. *Journal of Maternal-Fetal Medicine*, 9(5):311–318.

[4] Bessa, S., Domingues, I., Cardoso, J. S., Passarinho, P., Cardoso, P., Rodrigues, V., and Lage, F. (2014). Normal breast identification in screening mammography: A study on 18 000 images. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 325–330. IEEE.

[5] Cangelosi, D., Pelassa, S., Morini, M., Conte, M., Bosco, M. C., Eva, A., Sementa, A. R., and Varesio, L. (2016). Artificial neural network classifier predicts neuroblastoma patients' outcome. *BMC Bioinformatics*, 17(12):83.

[6] Centers for Disease Control and Prevention (CDC) (2013). Cervical cancer screening among women aged 18-30 years-United States, 2000-2010. *Morbidity and Mortality Weekly Report*, 61(51-52):1038.

[7] Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(1):35.

[8] Chicco, D. and Masseroli, M. (2015). Software suite for gene and protein annotation prediction and similarity search. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(4):837–843.

[9] Chicco, D., Sadowski, P., and Baldi, P. (2014). Deep autoencoder neural networks for Gene Ontology annotation predictions. In *Proceedings of ACM BCB 2014*, pages 533–540. ACM.

[10] Chollet, F. (2015). Keras. https://github.com/keras-team/keras.

[11] Cruz, R., Fernandes, K., Cardoso, J. S., and Costa, J. F. P. (2016). Tackling class imbalance with ranking. In *The 2016 International Joint Conference on Neural Networks*, pages 2182–2187. IEEE.

[12] Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240. ACM.

[13] Deacon, J. M., Evans, C. D., Yule, R., Desai, M., Binns, W., Taylor, C., and Peto, J. (2000). Sexual behaviour and smoking as determinants of cervical HPV infection and of CIN3 among those infected: a case–control study nested within the Manchester cohort. *British Journal of Cancer*, 83(11):1565.

[14] Elter, M., Schulz-Wendtland, R., and Wittenberg, T. (2007). The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Medical Physics*, 34(11):4164–4172.

[15] Fernandes, K., Cardoso, J. S., and Astrup, B. S. (2017a). Automated detection and categorization of genital injuries using digital colposcopy. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 251–258. Springer.

[16] Fernandes, K., Cardoso, J. S., and Fernandes, J. (2015). Temporal segmentation of digital colposcopies. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 262–271. Springer.

[17] Fernandes, K., Cardoso, J. S., and Fernandes, J. (2017b). Transfer learning with partial observability applied to cervical cancer screening. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 243–250. Springer.

[18] Graffar, M. (1956). Une méthode de classification sociale d'échantillons de population. *Courrier*, 6(8):455–459.

[19] Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. (2005). Result analysis of the nips 2003 feature selection challenge. In *Advances in neural information processing systems*, pages 545–552.

[20] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of IEEE ICCV 2015*, pages 1026–1034.

[21] Hong, Z.-Q. and Yang, J.-Y. (1991). Optimal discriminant plane for a small number of samples and design method of classifier on the plane. *Pattern Recognition*, 24(4):317–324.

[22] Kauffman, R. P., Griffin, S. J., Lund, J. D., and Tullar, P. E. (2013). Current recommendations for cervical cancer screening: do they render the annual pelvic examination obsolete? *Medical Principles and Practice*, 22(4):313–322.

[23] Kiros, R., Salakhutdinov, R., and Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

[24] Kurgan, L. A., Cios, K. J., Tadeusiewicz, R., Ogiela, M., and Goodenday, L. S. (2001). Knowledge discovery approach to automated cardiac spect diagnosis. *Artificial intelligence in medicine*, 23(2):149–169.

[25] Lacoste-Julien, S., Sha, F., and Jordan, M. I. (2009). Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems*, pages 897–904.

[26] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[27] Levy, O., Goldberg, Y., and Ramat-Gan, I. (2014). Linguistic regularities in sparse and explicit word representations. In *CoNLL*, pages 171–180.

[28] Li, W., Prasad, S., Fowler, J. E., and Bruce, L. M. (2012). Locality-preserving dimensionality reduction and classification for hyperspectral image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 50(4):1185–1198.

[29] Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A., and Moroz, I. M. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Biomedical Engineering Online*, 6(1):23.

[30] Louie, K. S., De Sanjose, S., Diaz, M., Castellsague, X., Herrero, R., Meijer, C. J., Shah, K., Franceschi, S., Munoz, N., and Bosch, F. X. (2009). Early age at first sexual intercourse and early pregnancy are risk factors for cervical cancer in developing countries. *British Journal of Cancer*, 100(7):1191.

[31] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

[32] Mangasarian, O. L., Street, W. N., and Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577.

[33] Menke, J. and Martinez, T. R. (2004). Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons. In *2004 IEEE International Joint Conference on Neural Networks, 2004. Proceedings*, volume 2, pages 1331–1335. IEEE.

[34] Moore, J. H. (2004). Computational analysis of gene-gene interactions using multifactor dimensionality reduction. *Expert review of molecular diagnostics*, 4(6):795–803.

[35] Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2):1883.

[36] Pinoli, P., Chicco, D., and Masseroli, M. (2015). Computational algorithms to predict Gene Ontology annotations. *BMC Bioinformatics*, 16(Suppl 6):S4.

[37] Plissiti, M. E. and Nikou, C. (2013). A review of automated techniques for cervical cell image analysis and classification. In *Biomedical Imaging and Computational Modeling in Biomechanics*, pages 1–18. Springer.

[38] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

[39] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer.

[40] Santos, M. S., Abreu, P. H., García-Laencina, P. J., Simão, A., and Carvalho, A. (2015). A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. *Journal of biomedical informatics*, 58:49–59.

[41] Scholkopf, B., Sung, K.-K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, 45(11):2758–2765.

[42] Smith, J. W., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association.

[43] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

[44] Tieleman, T. and Hinton, G. (2012). Lecture 6.5 – Rmsprop: divide the gradient by a running average of its recent magnitude. *Coursera: Neural Networks for Machine Learning*, 4(2):26–31.

[45] University of California Irvine (1987, accessed on 2017-08-10). *Machine Learning Repository*.

[46] University of California Irvine Machine Learning Repository (2017, on accessed 2018-02-01). *Cervical cancer (risk factors) Data Set*.

[47] Van Der Maaten, L., Postma, E., and Van den Herik, J. (2009). Dimensionality reduction: a comparative. *Journal of Machine Learning Research*, 10:66–71.

[48] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of ICML 2008*, pages 1096–1103. ACM.

[49] Xu, T., Zhang, H., Huang, X., Zhang, S., and Metaxas, D. N. (2016). Multimodal deep learning for cervical dysplasia diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 115–123. Springer.