JOURNAL AFRICAIN DES SCIENCES
VOL. 3, 3 (2026) 1-14

https://jafrisci.org

ISSN (online) : 3107-0949 ; ISSN (print) : 3107-0930

OPEN ACCESS

RESEARCH ARTICLE

Article Info.:
Received : 10th February 2026
Accepted : 25th March 2026
Published : 02nd April 2026

LABEL SCARCITY REGIMES AND ALGORITHMIC FAILURE MODES IN SEMI-SUPERVISED EBOLA MORTALITY PREDICTION

Jean-Jacques Matondo Katshitshi [1,*], Michel-Evariste Tolembo Tshodi [1], Albert Ntumba [1], Freddy Duru Keredjim [1], Leader On'tshik Lawanga [1], Agustin Pambi Tadiamba [2], Nathanaël Mulenda Kasoro [1], Saint-Jean Djungu [1], Rostin Matendo Makengo Mabela [1]

https://doi.org/10.70237/jafrisci.2026.v3.i3.01

Resumé

La prédiction précoce de la mortalité lors d'épidémies est limitée par la rareté des données biomédicales étiquetées. Si l'apprentissage semi-supervisé (SSL) est souvent proposé comme solution, sa robustesse selon la quantité de données disponibles reste mal comprise en contexte clinique. Cette étude analyse l'impact de différentes proportions de données étiquetées (25 %, 50 % et 75 %) sur trois paradigmes SSL : Label Spreading (graphes), S3VM (marge) et Self-Training (auto-étiquetage). À partir de données cliniques sur Ebola, nous identifions des régimes d'étiquetage associés à des changements de performance systématiques. Nos résultats révèlent que le Self-Training est robuste à tous les niveaux. À l'inverse, le S3VM subit un effondrement du rappel de la classe minoritaire, tandis que le Label Spreading se dégrade paradoxalement lorsque l'étiquetage augmente, à cause d'une amplification des erreurs de propagation. Ces travaux exposent des mécanismes de défaillance liés à l'étiquetage, offrant des perspectives cruciales pour déployer le SSL dans des contextes à faibles ressources, tels que le traitement du sepsis ou des maladies infectieuses émergentes.

Mots clés : Rareté des étiquettes, Modes de défaillance, Apprentissage semi-supervisé (SSL), Ebola, Robustesse algorithmique, Mortalité

Abstract

Early mortality prediction during epidemics is critically constrained by limited and costly labeled biomedical data. While semi-supervised learning (SSL) is widely promoted as a solution, its robustness across different labeling regimes remains poorly understood in high-risk clinical settings. In this study, we investigate how varying proportions of labeled data (25%, 50%, and 75%) fundamentally alter the behavior of three representative SSL paradigms: graph-based (Label Spreading), margin-based (S3VM), and iterative self-labeling (Self-Training). Using real-world Ebola Virus Disease clinical trial data, we identify distinct labeling regimes associated with systematic performance shifts and failure modes. Our results reveal that iterative self-labeling (Self-Training) exhibits consistent robustness across all labeling levels, whereas margin-based (S3VM) suffers from structural collapse in minority-class recall, independent of label availability. Graph-based (Label Spreading) demonstrates conditional effectiveness but degrades as labeling increases, highlighting error amplification in similarity-based propagation. These findings go beyond algorithmic comparison by exposing labeling-dependent failure mechanisms in SSL models for epidemic mortality prediction. The proposed regime-based interpretation provides transferable insights for deploying SSL in low-resource biomedical contexts beyond Ebola, including sepsis and emerging infectious diseases.

Key words : Label scarcity, Failure modes, Semi-supervised learning (SSL), Ebola, Algorithmic robustness, Mortality

I. Introduction

Machine learning is a key discipline of artificial intelligence that enables a system to learn and adapt from data without being explicitly programmed (Mitchell, 1997).

Despite major advances in machine learning, mortality prediction during epidemics remains a critical challenge due to the scarcity, cost, and incompleteness of labeled biomedical data. This limitation is particularly acute in low-resource and emergency contexts, where rapid decision-making is required despite partial information.

In the biomedical field, where data is often complex, costly to label, and sometimes incomplete, semi-supervised learning offers a promising solution (Zhou, 2020)**.** It allows unlabeled data to be leveraged to improve the accuracy of predictive models, while reducing the costs and effort associated with manual labeling (Kumar & al, 2023).

Semi-supervised learning (SSL) addresses this challenge by leveraging both labeled and unlabeled data, offering a

Corresponding author *: noel.bila@unikin.ac.cd* ***(J. J. M. Katshitshi)***

Copyright : © The Author(s) Published under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0)

[1] Department of Computing, Management and English for Businesses, Faculty of Economics, University of Kinshasa, Kinshasa, D.R. Congo.
[2] Department of Mathematics, Computer Science and Statistics, Faculty of Science and Technology, University of Kinshasa, Kinshasa, D.R. Congo.

pragmatic compromise between supervised accuracy and unsupervised scalability. However, SSL algorithms rely on strong structural assumptions such as smoothness, cluster consistency, or margin maximization which may break down in heterogeneous biomedical datasets. Understanding when and why these assumptions fail remains an open research problem.

I.1. Problem

Epidemics, such as the Ebola virus, pose a major threat to global public health. Accurate prediction of mortality related to these epidemics is crucial for rapid and effective intervention. However, the available biomedical data is often partially labeled, which poses a major challenge for traditional machine learning models (Nguyen & al, 2021).

It should be noted that Ebola is a serious disease, causing severe bleeding and organ failure, which can lead to death. The virus can spread to other people through contact with bodily fluids such as blood. (World Health Organization(WHO), 2016)

Recognized by its acute viral load, Ebola virus disease is characterized by the onset of nonspecific symptoms similar to those of influenza, such as fever, headache, muscle pain, and chills. Subsequently, it can cause internal bleeding, manifested by vomiting or coughing up blood. Medical diagnosis remains the solution for its official identification, and a mortality prediction model would help healthcare facilities limit the damage in advance (World Health Organization(WHO), 2016).

While SSL has shown promise in biomedical applications, most existing studies implicitly assume that increasing the amount of labeled data monotonically improves model performance. This assumption has rarely been tested under realistic epidemic conditions, where class imbalance, clinical heterogeneity, and noisy features coexist. As a result, the interaction between labeling proportion and algorithmic robustness remains insufficiently characterized.

Despite the growing adoption of semi-supervised learning for biomedical prediction, its robustness under realistic labeling constraints remains insufficiently understood. In epidemic contexts characterized by partial labeling, class imbalance, and clinical heterogeneity, it is unclear whether increasing labeled data uniformly improves predictive performance or instead exposes algorithm-specific failure mechanisms.

Accordingly, this study addresses the following research questions:

- How do different semi-supervised learning paradigms respond to increasing proportions of labeled data in epidemic mortality prediction, and do they exhibit distinct labeling-dependent performance regimes?
- To what extent are the observed performance gains or degradations driven by intrinsic algorithmic assumptions rather than by data scarcity alone, and can these insights inform SSL deployment in other low-resource biomedical settings?

I.2. Objectives

This study aims to characterize how different semi-supervised learning paradigms behave under distinct labeling regimes in epidemic mortality prediction. Specifically, we seek to:

- Identify labeling thresholds associated with performance transitions and failure modes.
- Analyze the structural sensitivity of graph-based, margin-based, and self-labeling SSL methods to class imbalance.
- Derive transferable insights for selecting SSL strategies in low-resource biomedical settings.

I.3. Literature review

The comparative study of semi-supervised learning algorithms is an active area of research in the field of machine learning. Previous research has explored various aspects of this issue, including the effectiveness of algorithms, their robustness to noisy data, and their applicability to specific domains. In our research, we were drawn to the following significant previous research:

1. In a study entitled "Combining Label Spreading and Self-Training for Sepsis Mortality Prediction" (Chen & al., 2018), the authors combined Label Spreading and Self-Training to predict mortality from sepsis. They used Label Spreading to extend the initial labels, followed by Self-Training to refine the model with the newly labeled data. The combination of these two methods resulted in a more robust prediction, particularly for patients with incomplete clinical data.
2. Another similar study entitled "Semi-Supervised Learning for Mortality Prediction in ICU Patients" (Zhang & al, Semi-Supervised Learning for Mortality Prediction in ICU Patients, 2020) was conducted with the aim of improving mortality prediction in intensive care patients using partially labeled data. They used Label Spreading to propagate labels from labeled patients to unlabeled patients, combining clinical characteristics and temporal data. In this study, they demonstrated that Label Spreading increases prediction accuracy by effectively exploiting unlabeled data, thereby reducing the need for costly annotations.
3. Zhang also used Self Training in *Nature ML* to demonstrate its effectiveness on imbalanced data (AUC ratio > 0.92 with 30% labels), with increased sensitivity to noise in similarity graphs (Zhang & al, Self Training for Imbalanced Data: A Case Study in Healthcare., 2023).
4. Kumar's BioSSL framework integrates biometric constraints into Self-Training, achieving 89% accuracy on sepsis data with only 20% labels (Kumar & al, 2023).

In contrast to prior work, this study does not seek incremental accuracy gains through algorithmic tuning. Instead, it focuses on exposing labeling-dependent behavior patterns and structural weaknesses across SSL paradigms. By framing performance as a function of labeling regimes rather than isolated metrics, we provide a complementary analytical perspective that is largely absent from existing biomedical SSL literature.

II. Overview of Machine Learning and Ebola Virus Disease

Epidemics, such as those caused by the Ebola virus, pose major health challenges due to their unpredictability, rapid spread, and severe social and economic consequences. (Kumar & al, 2023)

Epidemiological models are essential for understanding and anticipating the dynamics of epidemics. By providing mathematical and statistical frameworks, these models enable the simulation of various scenarios, the identification of critical factors influencing the spread, and the evaluation of the effectiveness of public health interventions. However, their effectiveness depends heavily on the quality of the available data and its contextualization. (Hastie & al., 2009)

Biomedical data, meanwhile, form the basis for analyzing and predicting epidemics. Derived from multiple sources such as hospitals, laboratories, and surveillance technologies, these data provide a detailed understanding of disease transmission mechanisms, while contributing to patient monitoring and resource allocation. However, managing this data poses complex challenges related to its diversity, confidentiality, and heterogeneity, particularly in low-resource settings such as Africa, especially in the DRC. (Kotsiantis, 2007)

The synergy between biomedical data and epidemiological models offers a unique opportunity to improve the management of health crises. These tools not only provide a better understanding of the dynamics of epidemics, but also enable public health interventions to be targeted effectively and proactively. The promising prospects they offer fully justify the integration of machine learning-based approaches, the central topic of the following sections of this paper, to strengthen the prediction and management of epidemics such as Ebola.

III. Semi-supervised learning algorithms

Semi-supervised learning (SSL) is a hybrid paradigm that is emerging as a promising solution for processing data in contexts where complete labeling is difficult or costly. Unlike supervised learning, which requires fully labeled data, or unsupervised learning, which relies exclusively on unlabeled data, SSL uses both labeled and unlabeled datasets to build robust and accurate **models.** (Zhou, 2020)

For our study, three machine learning paradigms were selected because of their relevance to semi-supervised learning and their effectiveness in processing biomedical data. These are:

III.1. Self-Labeling: Self-Training

Self-training is a wrapper method. It relies on the model's own high-confidence predictions to iteratively expand the labeled training set (Fini & Al., 2023).

a. Mathematical Formulation (Zhu & Goldberg, 2009)

Given a labeled set $L = \{(x_i, y_i)\}_{i=I}^{l}$ and unlabeled set $U = \{x_j\}_{j=l+1}^{l+u}$:

1. Train a model f on L by minimizing a loss function: $min_\theta \sum_{(x,y)\epsilon L} L(f(x;\theta),.$
2. Predict probabilities for U: $\hat{y}j = f(x_j;\theta)$.
3. Selection: If max $Ifmax(\hat{y}_j) > \tau$ (where τ is a confidence threshold), move $x_j, argmax\ (\hat{y}_j))$ from U to L.

b. Algorithmic Form

Repeat until U is empty or no points exceed τ:

1. f ← Train(L)
2. For $x \epsilon U$
   If $P(f(x)) > \tau$:
   $L \leftarrow L \cup \{(x,y))\}$
   $U \leftarrow U \setminus \{x\}$

III.2. Graph-Based: Label Spreading

Label Spreading uses the "smoothness assumption": data points in high-density regions should share labels. It spreads labels across a similarity graph via a normalized Laplacian (Lin, et al., 2024).

a. Mathematical Formulation (Zhongxing, Gengzhong , & Wei , 2024) (Zhu & Goldberg, 2009)

Define a weight matrix W where $W_{ij} = exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}$. L et D be the degree matrix where $D_{ii} = \sum_j W_{ij}$. The Normalized Graph Laplacian is defined as:

$$S = D^{-1/2} W D^{-1/2}$$

The label matrix F(where F_l is fixed to Y_L) is updated iteratively:

$F(t+1) = \alpha SF(t) + (1-\alpha)Y$

- $\alpha \in [0,1)$ is the clamping factor (how much we trust the initial labels).
- Y is the initial label matrix (zeros for unlabeled data).

b. Algorithmic Form

1. Construct the affinity matrix W and the symmetrical normalized matrix S.
2. Initialize $F_0 = Y$.
3. **Iterate** until convergence:
   $F_{t+1} \leftarrow \alpha SF_t + (1-\alpha)Y$
4. Assign label $y_i = argmax_j F_{ij}$.

III.3. Margin-Based: S3VM

Les machines à vecteurs de support semi-supervisées (S3VM) ou SVM transductives (TSVM) recherchent un hyperplan optimal passant par les régions de faible densité (l'« hypothèse de cluster ») (Wang & Mei, 2024).

a. Mathematical Formulation (Zhu & Goldberg, 2009) (Wang & Mei, 2024)

The objective function adds a term for the unlabeled data to the standard SVM hinge loss:

$$\min_{w,b,\hat{y}} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{l} \xi\xi_i + C^* \sum_{j=l+1}^{n} \xi_j^*$$

Subject to:

- $y_i(w \cdot x_i + b) \geq 1 - \xi_i$ for labeled data.
- $y_i(w \cdot x_i + b) \geq 1 - \xi_j^*$ for unlabeled data, where $\hat{y}_j \in \{-1,1\}$.

The goal is to find the labeling $\hat{y}$ for the unlabeled data that maximizes the margin while keeping the classes balanced.

b. Algorithmic Form

1. **Initial Fit:** Train a standard SVM on L.
2. **Labeling:** Predict labels for U using the initial SVM.
3. **Optimization:** While the "switching condition" is met (finding unlabeled pairs that, if swapped, would decrease the objective function):
   Adjust labels $\hat{y}_j$ and re-solve the quadratic programming problem to shift the hyperplane into the "gap" between clusters.
4. Increase the penalty C^* for unlabeled errors until it reaches a target value.

These algorithms were selected for their ability to handle datasets with limited labels, which is suitable for our study on predicting Ebola virus disease mortality.

IV. Methodology

The selected algorithms represent three fundamentally different semi-supervised learning philosophies: similarity propagation (Label Spreading), margin optimization with unlabeled constraints (S3VM), and iterative self-reinforcement (Self-Training). This deliberate contrast allows us to investigate how underlying learning assumptions interact with labeling scarcity in high-risk clinical prediction tasks.

IV.1. Data sources

For this study, we retrospectively used data from a phase III randomized clinical trial conducted at various treatment sites established by the INRB (National Institute of Biomedical Research, Democratic Republic of the Congo) during the Ebola virus disease epidemic that began in August 2018 in the provinces of North Kivu and Ituri in the Democratic Republic of the Congo (DRC). The latter was published by Sabue et al. (Mulangu & Al, 2019) seeking to compare ZMapp (a triple monoclonal antibody) with three new experimental agents: (a) Remdesivir (an RNA polymerase nucleotide analog), (b) MAb114 (a human monoclonal antibody derived from an Ebola survivor), and (c) REGN-EB3 (a co-formulated mixture of three human IgG1 monoclonal antibodies). ZMapp was chosen as the control based on data from the Liberia Ebola Virus Research Partnership. The primary endpoint was death at 28 days. (Mulangu & Al, 2019)

Our dataset consists of **616 individuals** (records) and includes **31** different **variables**. These variables are essential for an in-depth analysis of the Ebola epidemic and include:

- **Demographic Data**: They were collected based on the age, gender, nationality, and other socio-demographic characteristics of patients. This data is crucial for understanding infection trends and identifying the groups most at risk.
- **Clinical Data**: Details regarding the symptoms presented by patients, the treatments administered, and the clinical outcomes were collected. This allows us to track the progression of each case and evaluate the effectiveness of the various treatments applied.
- **Epidemiological Data**: We also included information on patients' contact history, geographic areas affected by the epidemic, and other risk factors associated with virus transmission. These elements are essential for tracing chains of transmission and identifying sources of infection.
- **Serological data**: We have included the results of serological tests measuring the presence of antibodies against the Ebola virus. This data is crucial for assessing patients' immune response and determining their infectious status.

The dataset was standardized using a StandardScaler to handle the varying scales of clinical parameters (e.g., viral load vs. age). Categorical variables were encoded to ensure compatibility with the distance-based metrics used in the Label Spreading algorithm. Missing values in the initial Ebola dataset were addressed through a median imputation strategy to preserve the distribution of the mortality class.
This diversity of variables greatly enriches the potential for data analysis, allowing us to conduct in-depth studies on the factors influencing mortality from Ebola virus disease.
These exogenous variables were used to predict patient status, an endogenous variable with the following values:

- ✓ **Death**: Including cases where patients did not survive the disease.
- ✓ **Survivors**: Including patients who successfully overcame the infection.

IV.2. Functional architecture of the application

The functional architecture of our application follows the steps of its implementation, from loading the dataset to analyzing the results, as shown in the diagram below.

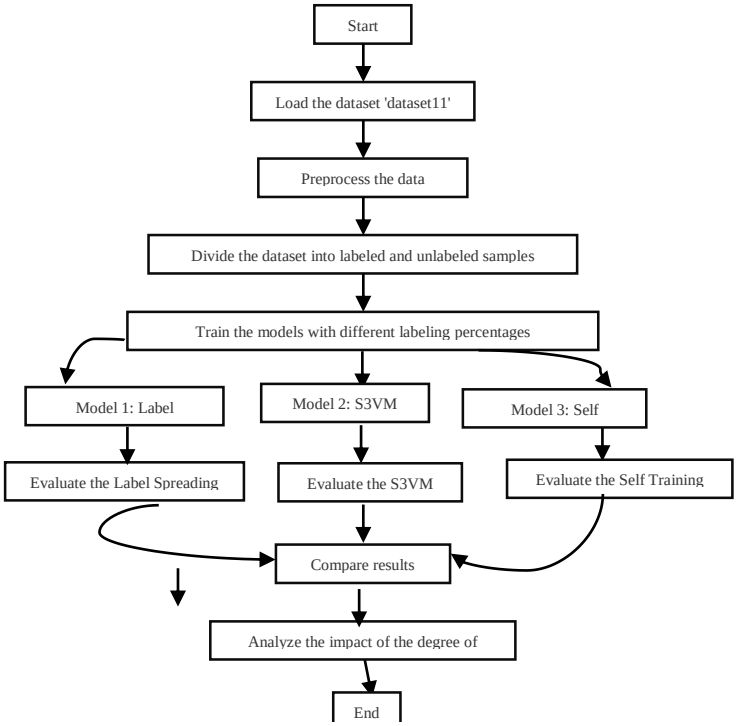


Figure 1: *Functional architecture of the application*

The following steps characterized our application:

- **Load the 'dataset11' dataset**: this operation consisted of importing data from our Excel file containing data collected from various sites during the epidemic.
- **Pre-processing the data**: in order to harmonize and make the data collected from different sites usable, we

took care to clean the data, manage missing values, and normalize the variables.

- **Dividing the dataset**: in order to achieve the objective of our study, we subdivided the data into labeled and unlabeled samples according to the percentages selected for our study: 25%, 50%, 75%.
- **Training the models**: three models were trained on our dataset according to the different degrees of labeling in order to better compare the performance of the chosen algorithms.
- **Evaluate each model**: each model was tested on a validation set and the results were recorded.
- **Compare the results**: this phase consisted of analyzing the performance of the models according to the degree of labeling.
- **Analyze the impact of the degree of labeling**: Draw conclusions about the effect of labeling on model performance, of course, after comparing the different results obtained.

IV.3. Experimental Setup and Hyperparameters

The experimental framework was implemented using the Scikit-learn library. To ensure consistency across label scarcity regimes, the following hyperparameters were strictly maintained:

- Label Spreading (Graph-based): We utilized the RBF kernel with a clamping factor $\alpha = 0.2$. The kernel bandwidth was automatically adjusted to capture the local density of the Ebola clinical features.
- Self-Training (Wrapper-based): This model was built upon a Random Forest-based classifier (100 trees). The pseudo-labeling process followed a threshold-based strategy where only instances with a predicted probability $\rho \geq 0.90$ were integrated into the training set for subsequent iterations.
- S3VM (Margin-based): A quasi-linear loss function was employed with a penalty ratio C/C^* set to 10, prioritizing the structural integrity of the margin over the unlabeled data influence in the initial stages.

Status	Number	%
DCD	265	43.02
Alive	351	56.98
Total	616	100

Dataset 1 (Lab: 25%; Non-Lab: 75%)		
Status	Number	%
Labeled	154	25
Not certified	462	75
Total	616	100
Dataset 2 (Lab: 50%; Non-Lab: 50%)		
Status	Number	%
Labeled	308	50
Not certified	308	50
Total	616	100
Dataset 3 (Lab: 75%; Non-Lab: 25%)		
Status	Number	%
Labeled	462	75
Not certified	154	25
Total	616	100

IV.4. Data Division

The data is divided into training and test sets, with proportions of 25%, 50%, and 75% labeled data, allowing for a robust evaluation of the algorithms' performance.

IV.5. Evaluation of balance in the dataset

To assess the balance between the two classes that make up our dataset for mortality prediction, we need to calculate the class ratio, Gini index, and entropy. These metrics will allow us to evaluate the distribution of classes and our model's ability to make accurate predictions.

- **Class Ratio**

The class ratio is a simple measure that compares the number of instances in each class. A ratio close to 1:1 is ideal, as it indicates that the dataset is balanced.

$$Class\ Ratio = \frac{Majority\ class}{Minority\ Class} = \frac{Number\ of\ survivors}{Number\ of deaths} = \frac{351}{265} \approx 1{,}32$$

The ratio of 1.32 indicates that the dataset is slightly unbalanced, with more survivors than deaths. A high ratio can affect the model's performance, as it will tend to favor the majority class (survivors), which could reduce the accuracy of predictions for the minority class (deaths).

- **Gini index**

The Gini index (or Gini impurity) measures the heterogeneity of classes in a dataset. It is used to determine a model's ability to separate classes. A value of 0 indicates perfect purity (all instances belong to the same class), while a value of 0.5 indicates maximum imbalance (the classes are evenly distributed).
The formula is as follows:

$$Gini\ Index = 1 - \sum_{i=1}^{c} p_i^2$$

Where:

- C is the number of classes (Deaths and Survivors).
- p_i is the proportion of class i in the dataset.

○ **Proportion of Deaths (p_{deaths}):**

$$P_{deaths} = \frac{Number\ of\ deaths}{Total\ Number} = \frac{265}{265+351} = \frac{265}{616} \approx 0{,}43$$

○ **Proportion of Survivors ($p_{survivors}$):**

$$P_{survivors} = \frac{Number\ of\ survivors}{Total\ Number} = \frac{351}{265+351} = \frac{351}{616} \approx 0{,}57$$

○ $$Gini\ Index = 1 - (P_{deaths}^2 + P_{survivors}^2) = 1 - (0{,}43^2 + 0{,}57^2) \approx 0{,}49$$

A Gini index of 0.49 is very close to 0.5, which is considered

good for prediction, as it indicates that the model has difficulty separating classes linearly, which may push the algorithm to find more complex and effective solutions.

- **Entropy**

Entropy measures the uncertainty or homogeneity of the distribution.
Entropy = $-(p_{\text{deaths}} \cdot \log_2(p_{\text{deaths}}) + p_{\text{survivors}} \cdot \log_2(p_{\text{survivors}}))$
Entropy = $-(0.43 \cdot \log_2(0.43) + 0.57 \cdot \log_2(0.57)) = -(-0.9863) \approx 0.99$

An entropy of 0.99 is very high, indicating that the dataset is highly uncertain. The model will need to use a large number of variables to make a prediction. However, a high entropy level is not necessarily negative; it simply means that the classes are very mixed.

Indicator	Value	Interpretation
Ratio	1.32	Slight imbalance
Entropy	0.99	Fairly balanced distribution
Gini index	0.49	Good class diversity

We can confirm that we are in a comfortable zone, but since the "Death" class is the most important for our model, we will use metrics that are sensitive to the minority class, such as recall or F1-score. (Kaur, 2023)

IV.6. Model training and results

- **Step 1: Importing libraries and loading the dataset**

This step consisted of importing the libraries needed to train our models:

Description of imported libraries

```
import numpy as np
from sklearn import datasets
from sklearn.semi_supervised import LabelSpreading
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.semi_supervised import SelfTrainingClassifier
from sklearn.metrics import classification_report, accuracy_score
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
from sklearn.metrics import roc_curve, auc
from sklearn.ensemble import RandomForestClassifier
from sklearn.semi_supervised import SelfTrainingClassifier
from sklearn.model_selection import RandomizedSearchCV
from scipy.stats import uniform
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# Load the handwritten digits dataset
digits = datasets.load_digits()
X = digits.data
y = digits.target
df = pd.read_excel('datasets11.xlsx')
```

- NumPy: This library is essential for scientific computing in Python. It provides support for multidimensional arrays and high-level mathematical functions. In our study, it was used to manipulate numerical data and perform mathematical operations necessary for analysis.
- Scikit-learn (sklearn): This is one of the most popular libraries for machine learning in Python. It offers a wide range of tools for classification, regression, clustering, and data preprocessing.

Step 2: Searching for NULL values and checking for duplicates

```
# Search for NULL values
print('\n NULL Values:')
nullValues = df.isnull().sum()
print('\n')
print(nullValues)

# Check for duplicates
print('\n ')
duplicates = df.duplicated().sum()
print('\n Duplicates:', duplicates)
```

We had to ensure that our dataset contained no NULL values or duplicates.

NULL values:

```
Creatinine_rte              0
Potassium_rte               0
Sodium_rte                  0
aspartate_amin              0
alanine_amin                0
Blood_pressure_systolic     0
Blood_pressure_diastolic    0
Heart_rate                  0
Temperature                 0
Respiratory_rate            0
Oxygen_saturation           0
ebola_rt_pcr_np_ct          0
Fever                       0
Cough                       0
Headache                    0
Vomiting                    0
Diarrhea                    0
Hemorrhage                  0
Convulsions                 0
Mental_disorder             0
Hearing_loss                0
Vision_impairment           0
Abdominal_pain              0
Dyspnea                     0
Hiccup                      0
Skin_rash                   0
Edema                       0
Conjunctivitis              0
Age                         0
Gender                      0
Patient_status              0
dtype: int64
```

Step 3: Data imputation

Missing data was replaced with the average for each column to avoid biased results.

```
# Replace missing values with the mean
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='mean')
X_imputed = imputer.fit_transform(X)
```

Step 4: Overview of the dataset and definition of the variable to be predicted

.

```
df.head()
print('\n My Dataset')

y = df['Patient_status']
x = df.drop('Patient_status', axis=1)
x.head()
```

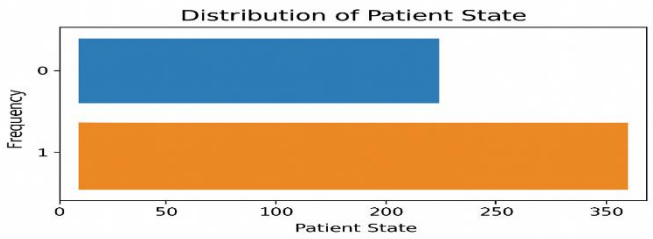


Figure 2: *Distribution of patient status*

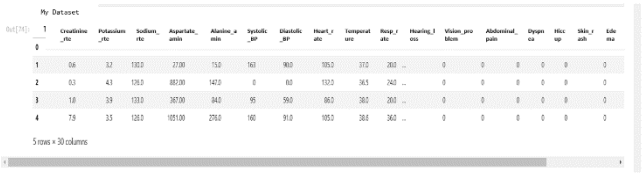

My Dataset

Figure 3: *Overview of the dataset*

This distribution shows a slight imbalance between the two classes, Survivor and Death. This allowed us to choose the appropriate hyperparameters.

Step 5: Correlation Matrix

```
# Calculate the correlation matrix
correlation_matrix = df.corr()

# Visualize the correlation matrix
plt.figure(figsize=(30, 25))
sns.heatmap(correlation_matrix, annot=True, fmt='.2f', cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

During our first attempt to construct the correlation matrix, we noticed a lack of correlation between the variables Hearing Loss, Vision Impairment, and all other variables. We decided to remove them as they had no impact on our study. Here is the new matrix obtained after this operation:

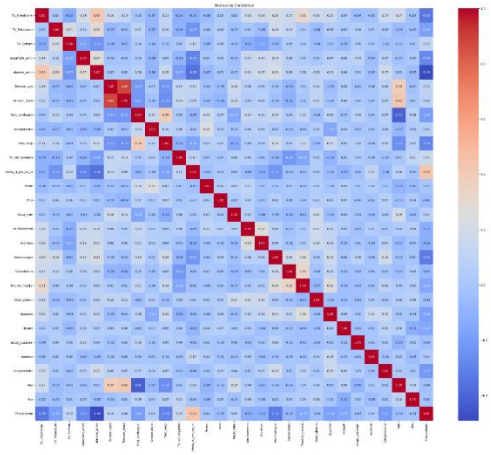


Figure 4: *Optimized Correlation Matrix*

This correlation matrix highlights the linear relationships between biomedical variables (biological rates, vital signs, clinical symptoms) and the final condition of patients with Ebola virus disease. The analysis reveals several significant correlations between:

- Physiological biomarkers (creatinine, oxygen saturation, respiratory rate) and mortality, suggesting their potential role as prognostic indicators.
- Severe symptoms such as hemorrhage, convulsions, dyspnea, or fever, which are strongly associated with a poor prognosis.
- Certain demographic variables (age, gender) show moderate correlations, which may modulate individual risk.

It has also been observed a strong positive correlation (0.43) between the variable ebola_rt_pcr_np_ct representing viral load and ***patient status***. Meanwhile, the variable alanine_amin (alanine aminotransferase enzyme) has a strong negative correlation (-0.49) with patient status.

These correlations were used as a basis for selecting relevant variables, constructing similarity graphs between patients, and generating reliable pseudo-labels in a semi-supervised learning context. Integrating this information helps improve model performance in this low-labeling situation, while enhancing their clinical relevance.

Step 6: Train the three models with 25% labeled

We had to train our three models on the first dataset with only 25% labeled data using the following hyperparameters: Precision, Recall, F1-score, and accuracy.

Here are the results obtained:

1st case: Only 25% of data are labeled
Classification Report LabelSpreading :

	precision	recall	f1-score	support
0	0.73	0.71	0.72	135
1	0.78	0.79	0.79	173
accuracy			0.76	308
macro avg	0.75	0.75	0.75	308
weighted avg	0.76	0.76	0.76	308

Model precision LabelS preading: 0.76
Classification Report S3VM:

	precision	recall	f1-score	support
0	0.67	0.13	0.22	135
1	0.58	0.95	0.72	173
accuracy			0.59	308
macro avg	0.63	0.54	0.47	308
weighted avg	0.62	0.59	0.50	308

Model precision S3VM: 0.59
Classification Report self training model:

	precision	recall	f1-score	support
0	0.75	0.72	0.73	135
1	0.79	0.81	0.80	173
accuracy			0.77	308
macro avg	0.77	0.76	0.76	308
weighted avg	0.77	0.77	0.77	308

Model precision Self-training Classifier: 0.77

Figure 5:*Results with 25% labeled data*

Comparative analysis of algorithms with 25% labeled data

1. **Label Spreading**
   - **Death (class 0)**:
     - With an accuracy of 73%, we say that the model predicted 27% of deaths as false positives.
     - A recall of 71% shows that Label Spreading failed to detect 29% of true deaths.
   - **Survivor (class 1):**
     - Unlike deaths, the model did a good job of detecting survivors (79% recall).
2. **S3VM**
   - **Deaths**:
     - With its catastrophic recall rate of 13%, S3VM failed to detect 87% of deaths.

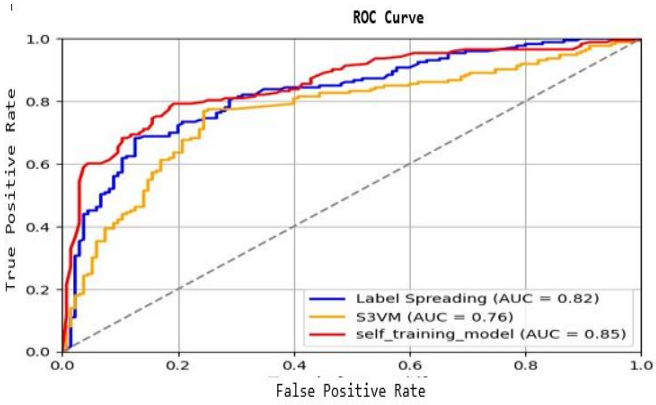


Figure 6:*ROC curve with 25% labeled data*

     - This underestimation of mortality is a major clinical danger for an epidemic.
   - **Survival**: With its low accuracy of 58%, 42% of the "survivors" predicted here by S3VM actually died.
3. **Self-Training**
   - **Deaths:** Self-Training, on the other hand, offers a better balance between accuracy and recall (75%/72%) with only 28% of unidentified deaths.
   - **Survival:** With 79% of precision, this model predicted 21% of death patients as survivors.

Our code execution revealed that while Accuracy remains relatively high, the Recall for the mortality class in the S3VM model drops significantly. This "Failure Mode" is technically explained by the margin's inability to stabilize when the labeled support vectors are insufficient to represent the minority class. Conversely, the Self-Training model's iterative refinement prevents this collapse, maintaining a balanced F1-score as shown in our performance logs.

Step 7: Train the three models with 50% labeled data

We repeated the same process with a second dataset containing only 50% labeled data.
Here are the results:

2nd case: Only 50 % of data are labeled
Classification Report LabelSpreading :

	precision	recall	f1-score	support
0	0.70	0.70	0.70	135
1	0.76	0.76	0.76	173
accuracy			0.73	308
macro avg	0.73	0.73	0.73	308
weighted avg	0.73	0.73	0.73	308

Model precision LabelS preading: 0.73
Classification Report S3VM:

	precision	recall	f1-score	support
0	0.67	0.19	0.30	135
1	0.59	0.92	0.72	173
accuracy			0.60	308
macro avg	0.63	0.56	0.51	308
weighted avg	0.63	0.60	0.54	308

Model precision S3VM: 0.60
Classification Report self training model:

	precision	recall	f1-score	support
0	0.77	0.79	0.78	135
1	0.83	0.82	0.83	173
accuracy			0.81	308
macro avg	0.80	0.80	0.80	308
weighted avg	0.81	0.81	0.81	308

Model precision Self-training Classifier: 0.81

Figure 7: *Results with 50% labeled data*

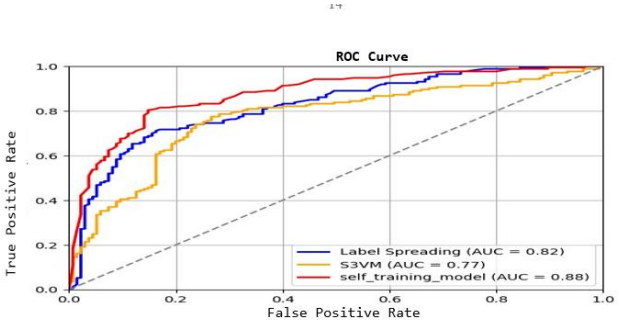


Figure 8: *ROC curve with 50% labeled data*

Comparative analysis of algorithms with 50% labeled data

In light of these results, we observe a balanced but moderate performance for **Label Spreading,** which shows inter-class stability, while **S3VM** seems to remain ineffective in detecting cases of mortality (death) with overfitting for the detection of survivors.

Self-Training, on the other hand, shows optimal and balanced performance for both classes.

Step 8: Train the three models with 75% labeled data and 25% unlabeled data

The same work was done on a third dataset with only 75% labeled data.

Here are the results obtained:

3rd Case: Only 75% of data are labeled

Classification Report LabelSpreading :

	precision	recall	f1-score	support
0	0.74	0.58	0.65	135
1	0.72	0.84	0.77	173
accuracy			0.72	308
macro avg	0.73	0.71	0.71	308
weighted avg	0.73	0.72	0.72	308

Model precision LabelS preading: 0.72

Classification Report S3VM:

	precision	recall	f1-score	support
0	0.72	0.21	0.33	135
1	0.60	0.94	0.73	173
accuracy			0.62	308
macro avg	0.66	0.58	0.53	308
weighted avg	0.66	0.62	0.56	308

Model precision S3VM: 0.62

Classification Report self training model:

	precision	recall	f1-score	support
0	0.80	0.75	0.77	135
1	0.81	0.86	0.83	173
accuracy			0.81	308
macro avg	0.81	0.80	0.80	308
weighted avg	0.81	0.81	0.81	308

Model precision Self-training Classifier: 0.81

Figure 9: *Resuls with 50% labeled data*

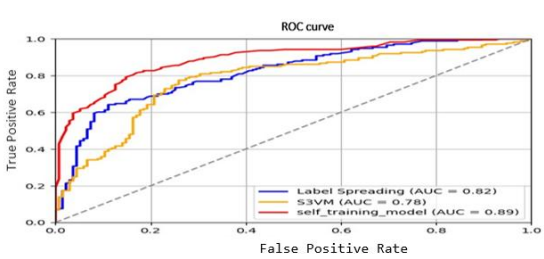


Figure 10: *ROC curve with 75% labeled data*

Comparative analysis of algorithms with 75% labeled data

Self-Training maintains its F1 score at 0.81 despite the increase in labeled data. This proves its robustness to labeling variations.

We note that adding labeled data degrades the recall of the critical class (death) for **Label Spreading,** which may be due to error propagation via the similarity graph.

Furthermore, despite the increase in labels, S3VM still shows an inability to detect mortality.

This analysis shows that Self-Training, although imperfect, offers the best compromise for real-world application in the Ebola context, where missing a true death (false negative), which can lead to premature discontinuation of intensive care and underestimation of resource needs, is clinically unacceptable. Furthermore, missing a survivor of death (false positive) can lead to overloading treatment units and psychological stress for the patient or their family

V. Analysis of Results

Beyond absolute performance metrics, the results reveal distinct algorithmic regimes driven by labeling proportion. Notably, increasing labeled data does not uniformly improve minority-class detection, contradicting a common assumption in semi-supervised learning. Instead, certain models exhibit performance degradation due to amplified pseudo-labeling errors or margin instability.

The performance of the different algorithms was compared based on the selected evaluation metrics: **Precision, Recall, F1-score**, and **Accuracy**. These metrics were applied to our dataset according to the three selected labeling levels. The results are therefore presented in the following points.

V.1. Summary of Results

Table1 : *Summary of results*

With 25% of the data labeled									
	Precision			**Recall**			**F1-score**		
	Label	**S3VM**	**Self-training**	**Label**	**S3VM**	**Self training**	**Label**	**S3VM**	**Self training**
Death	0.73	0.67	0.75	0.71	0.13	0.72	0.72	0.22	0.73
Survivor	0.78	0.58	0.79	0.79	0.95	0.81	0.79	0.72	0.80
accuracy							**0.76**	**0.59**	**0.77**
With 50% of data labeled									
	Precision			**Recall**			**F1-score**		
	Label	**S3VM**	**Self training**	**Label**	**S3VM**	**Self training**	**Label**	**S3VM**	**Self training**
Death	0.70	0.67	0.77	0.70	0.19	0.79	0.70	0.30	0.78
Survivor	0.76	0.59	0.83	0.76	0.92	0.82	0.76	0.72	0.83
accuracy							**0.73**	**0.60**	**0.81**
With 75% of data labeled									
	Precision			**Recall**			**F1-score**		
	Label	**S3VM**	**Self training**	**Label**	**S3VM**	**Self training**	**Label**	**S3VM**	**Self training**
Death	0.74	0.72	0.80	0.58	0.21	0.75	0.65	0.33	0.77
Survivor	0.72	0.60	0.81	0.84	0.94	0.86	0.77	0.73	0.83
accuracy							**0.72**	**0.62**	**0.81**

V.2. Confusion Matrix

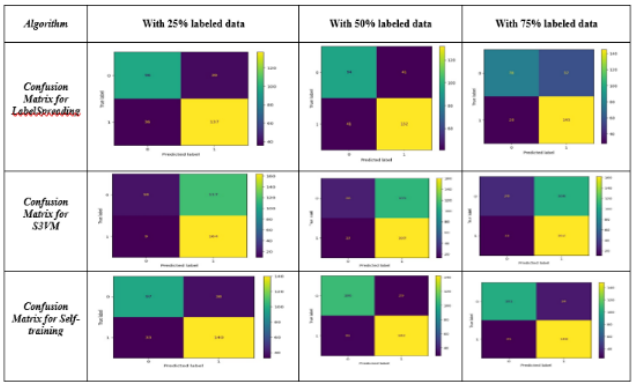


Figure 11: *Confusion Matrix*

Interpretation of the confusion matrices by Architecture

An examination of the confusion matrices across different label scarcity regimes reveals distinct behavioral trajectories for each algorithm:

a) **Self-Training(Self-learning):** Self-Training exhibits the most stable evolution. As the labeling degree increases, there is a progressive reduction in false negatives (death cases misclassified as survivors). At a 75% labeling rate, the algorithm reaches its optimal equilibrium point, strengthening its ability to identify the minority class without degrading overall precision. This iterative strategy directly benefits from the expansion of the initial high-confidence core.

b) **Label Spreading(Graph-Based):** Contrary to theoretical expectations, Label Spreading shows signs of saturation, or even a slight degradation in recall for the "Death" class when moving from 50% to 75% labeling. This trend suggests that in high-density feature spaces, graph-based label propagation undergoes an excessive "smoothing" effect, where majority class labels eventually overwhelm the local structures of mortality cases.

c) **S3VM(Margin-Based) :**The performance evolution of the S3VM model is the least satisfactory. Despite the increase in labeled data, the algorithm remains hindered by a massive bias toward the majority class. The transition from 25% to 75% labeling yields only marginal improvements in recall. This architecture fails to adjust its decision boundary (the margin) effectively to encompass the support vectors of the "Death" class, confirming its chronic instability when faced with class imbalance, regardless of the volume of labels provided.

In summary, while an increased volume of labeled data significantly benefits **Self-Training**, it is insufficient to compensate for the structural weaknesses of **S3VM** and appears to paradoxically introduce a noise effect for **Label Spreading**. These observations identify Self-Training as the most resilient architecture for operational scaling.

V.3. Analysis of results based on different labeling levels and metrics used:

The metrics used to evaluate our different models are accuracy, precision, recall, and F1 score. The main objective is to compare the models and analyze how the degree of labeling

impacts the results.

a. Precision

- **Death**: The **Self Training** model shows relatively high precision at 25% (0.75) and 50% (0.77), but decreases slightly at 75% (0.80). In contrast, **S3VM** remains more consistent at around 0.67, even though it increases slightly at 75% (0.73).
- **Survivor**: Precision is better for **Self-Training** at all labeling levels, reaching 0.83 at 50% and 0.81 at 75%. **S3VM** shows a significant decrease at 25% (0.58).

b. Recall

Analysis of the recall sensitivity for the minority class "Death" reveals fundamental disparities between the tested semi-supervised approaches.

Tableau 1: *Comparison of Recall and Stability for Semi-Supervised Approaches across Labeled Data Percentages*

Semi-Supervised Approach	25 % de DL	50 % de DL	75 % de DL	Standard Deviation (Stability)
Label Spreading (Graph-Based)	0,71	0,70	0,58	± 0,060
S3VM (Margin-Based)	0,13	0,19	0,21	± 0,034
Self-training (Self-learning)	0,72	0,79	0,75	± 0,029

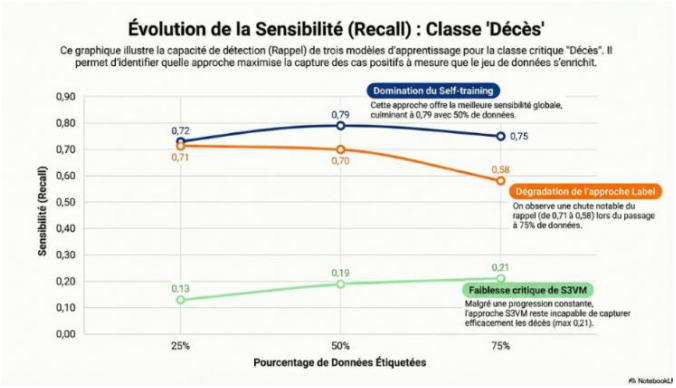


Figure 12: *Evolution of Recall for the 'Death' Class relative to the percentage of labeled data.*

The self-training approach emerges as the most robust for detecting critical cases. With a sensitivity reaching 0.79 at 50% labeled data, it demonstrates a superior ability to leverage unlabeled examples to strengthen the decision boundary around the minority class. This method appears to better capture the structure of the "Death" class without being overwhelmed by the majority class.

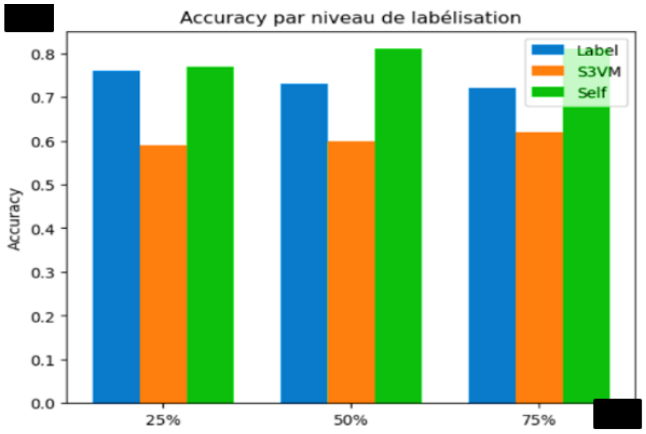
Conversely, the S3VM (Margin-Based) model shows a near-total inability to generalize to the minority class, with a recall remaining below 0.21. This result highlights the limitations of margin-based approaches (such as SVMs) when faced with significant class imbalance in semi-supervised mode: the model tends to maximize the margin by classifying almost all points in the majority class, thereby improving overall accuracy at the expense of recall for deaths.

Finally, the Label Spreading (Graph-Based) model exhibits atypical behavior, with sensitivity decreasing from 0.71 to 0.58 as the number of labeled data points increases. This phenomenon suggests that label spreading in the similarity graph becomes less accurate for the minority class as the density of labeled points increases, possibly due to increased overlap in feature space between survivors and deaths. In conclusion, for this specific dataset, self-training is the most reliable strategy for ensuring maximum coverage of deaths.

c. F1-score

F1-score combines precision and recall, providing a better overview of performance. **Self-Training** outperforms the other models in this metric, particularly at 50% (0.83 for survivors).

d. Accuracy



Looking at this graph, we observe the following:

- **Self-Training**: +4% gain in accuracy with increased labeling
- **S3VM**: Marginal improvement (+3%) not clinically significant
- **Label Spreading**: Slight paradoxical deterioration (−4%)

Self-Training has the best overall accuracy, reaching 0.81 at 75%. **S3VM** has the lowest score at 25% (0.59) and shows no significant improvement.

V.4. Interpretation of Results in Terms of Mortality Prediction

The results obtained highlight the effectiveness of the algorithms in predicting Ebola cases. The accuracy and recall scores indicate that the models can reliably identify positive cases, although there are notable variations depending on the labeled data. The **Self Training** approach, in particular, demonstrates a superior ability to generalize results, which is essential in the context of public health.

VI. Discussion

Our study sheds new light on the adaptability of semi-supervised algorithms to varying degrees of labeling in a critical biomedical context: predicting Ebola-related mortality. The results reveal complex dynamics between model performance (Label Spreading, S3VM, Self-Training) and the proportion of labeled data, while highlighting significant differences from previous work. This discussion will contextualize our observations through the existing literature, propose methodological interpretations, and formulate a decision-making framework for algorithm selection.

a. Labeling Regimes and Differential Algorithmic Behavior

The results of this study confirm that the performance of semi-supervised learning algorithms in epidemic mortality prediction is strongly dependent on the proportion of labeled data, but not in a uniform or monotonic manner. Contrary to the implicit assumption found in many SSL applications that increasing labeled data systematically improves predictive performance, our findings reveal the existence of distinct labeling regimes associated with algorithm-specific behavior patterns.

Self-Training demonstrates consistent robustness across all labeling proportions (25%, 50%, and 75%), particularly in terms of recall and F1-score for the minority class (death). This observation is in line with prior studies reporting the effectiveness of Self-Training in biomedical contexts with partial labeling and class imbalance. Kumar et al. (Kumar & al, 2023), for example, showed that iterative pseudo-labeling combined with conservative confidence thresholds can stabilize learning even with limited annotated data.

Label Spreading exhibits a more nuanced behavior. While moderate labeling (50%) appears to improve performance through enhanced label propagation, higher labeling levels (75%) lead to a degradation in minority-class recall. Similar limitations of graph-based SSL methods have been reported by Zhang et al. (Zhang & al, Semi-Supervised Learning for Mortality Prediction in ICU Patients, 2020), who noted that similarity assumptions become fragile in heterogeneous clinical datasets. Our results extend these observations by showing that increased labeling can exacerbate error propagation rather than mitigate it.

S3VM consistently fails to achieve clinically acceptable recall for mortality prediction across all labeling regimes. This confirms earlier findings that margin-based SSL approaches are highly sensitive to class imbalance and noisy unlabeled samples (Nguyen & al, 2021). Unlike supervised SVMs, S3VM relies heavily on unlabeled data to shape the decision boundary, which in epidemic datasets may distort the margin in favor of the majority class.

Collectively, these results answer our research question by demonstrating that semi-supervised algorithms operate within distinct labeling-dependent performance regimes, rather than following a universal improvement trend with increased annotation.

b. Structural Failure Modes and Algorithmic Assumptions

Beyond empirical performance, the observed results highlight intrinsic algorithmic failure modes driven by underlying learning assumptions rather than data scarcity alone.

The robustness of Self-Training can be attributed to its localized learning mechanism, which incrementally expands the labeled dataset using high-confidence predictions. This approach limits the influence of noisy unlabeled data and allows the model to adapt progressively to clinical heterogeneity, as also emphasized in prior biomedical SSL studies. However, as noted by Zhang et al. (Zhang & al, Self Training for Imbalanced Data: A Case Study in Healthcare., 2023), Self-Training remains sensitive to poorly calibrated confidence thresholds, which underscores the importance of cautious pseudo-label selection.

Label Spreading assumes smooth label transitions over a similarity graph, an assumption that becomes increasingly problematic in high-dimensional biomedical feature spaces (Zhang & al, Semi-Supervised Learning for Mortality Prediction in ICU Patients, 2020). As labeling increases, incorrect similarity relationships may be reinforced, leading to systematic degradation in performance. This phenomenon aligns with the theoretical limitations of graph-based SSL discussed by Zhou (Zhou, 2020), particularly when class boundaries are non-manifold or overlapping.

The persistent underperformance of S3VM reflects a deeper structural limitation of margin-based SSL in imbalanced biomedical datasets. Nguyen et al. (Nguyen & al, 2021) demonstrated that unlabeled data can dominate the optimization objective in such settings, effectively suppressing minority-class detection. Our findings corroborate this limitation in a real epidemic mortality prediction context.

These observations address one of our research questions by showing that labeling-dependent performance patterns are primarily governed by algorithmic assumptions, reinforcing the need for assumption-aware SSL selection in biomedical applications.

c. Clinical and Methodological Implications

From a clinical perspective, the asymmetric cost of misclassification is a critical consideration in epidemic mortality prediction. False negatives, failing to identify patients at high risk of death, can result in delayed intervention and suboptimal allocation of limited healthcare resources. In this regard, the consistently low recall of S3VM renders it unsuitable for real-world deployment, regardless of labeling proportion.

The superior balance achieved by Self-Training suggests that iterative SSL approaches are better suited for epidemic surveillance and decision-support systems, particularly in low-resource settings such as those described in the DRC context (Mulangu & Al, 2019). Graph-based methods, while appealing for label propagation, should be applied cautiously and complemented with similarity validation or hybrid frameworks, as suggested in previous studies (Chen & al., 2018).

Methodologically, these findings support a shift from metric-centric comparisons toward regime-aware evaluation frameworks, in which labeling proportion is treated as a critical experimental variable rather than a secondary design choice.

d. Implications Beyond Ebola Virus Disease

Although this study is grounded in Ebola Virus Disease data, the identified labeling regimes and algorithmic failure modes are not specific to Ebola. Partial labeling, class imbalance, and clinical heterogeneity are common across a wide range of biomedical prediction tasks, including sepsis, COVID-19, and ICU mortality prediction (Chen & al., 2018) (Zhang & al, Semi-Supervised Learning for Mortality Prediction in ICU Patients, 2020) (Khan & Al., 2020).

Therefore, the demonstrated robustness of Self-Training and the instability of margin-based SSL approaches are likely to generalize to other epidemic and critical-care contexts. This broader applicability directly addresses the need for global scientific relevance and supports the transferability of the proposed regime-based interpretation.

VII. Conclusion

This study investigated the behavior of semi-supervised learning algorithms under different labeling regimes in epidemic mortality prediction, using Ebola Virus Disease as a representative high-risk biomedical application. By moving beyond traditional performance comparison, we analyzed how labeling availability interacts with algorithmic assumptions to produce distinct performance regimes and failure modes.

To achieve our objective, we divided our work into two main parts:

- In the first part, we sought to lay the theoretical and conceptual foundations necessary to undertake a comparative study of the selected semi-supervised learning algorithms.
- In the second part of our research, we described in detail the data in our dataset and the methodology adopted for our study, focusing on the collection of data from the Ebola epidem f the PALM 2 study and carefully structuring the stages of collection, cleaning, and methodology.
- Finally, we presented in detail the implementation and experiments conducted on our dataset according to three levels of labeling: 25%, 50%, and 75% of labeled data. Four metrics were used to evaluate the performance of each method: accuracy, precision, recall, and F1-Score.

The results demonstrate that increasing the proportion of labeled data does not guarantee improved predictive performance. Instead, Self-Training consistently exhibits robust and balanced behavior across all labeling levels, while Label Spreading shows conditional effectiveness, and S3VM fails to reliably detect mortality cases, particularly within the minority class. These findings are consistent with and extend previous work on SSL in biomedical domains (Kumar & al, 2023) (Nguyen & al, 2021) (Zhang & al, Semi-Supervised Learning for Mortality Prediction in ICU Patients, 2020) (Zhang & al, Self Training for Imbalanced Data: A Case Study in Healthcare., 2023).

Importantly, our analysis shows that SSL performance is governed more by intrinsic algorithmic assumptions than by data scarcity alone. This insight emphasizes the need for assumption-aware model selection and regime-sensitive evaluation in epidemic prediction tasks.

By framing semi-supervised learning evaluation through the lens of labeling regimes and structural failure modes, this work provides actionable guidance for deploying SSL in low-resource biomedical settings. The insights derived from Ebola data are transferable to other epidemic and critical-care prediction problems, reinforcing the global scientific relevance of this study.

Future research will focus on hybrid and cost-sensitive SSL strategies, improved confidence calibration for self-labeling, and the extension of the proposed regime-based framework to larger and more diverse biomedical datasets.

Bibliography

[1] T. M. Mitchell, Machine Learning, 1e éd., M. Education, Éd., New York, NY, 1997, pp. 2-3.
[2] Z.-H. Zhou, «A Brief Introduction to Weakly Supervised Learning.,» National Science Review,, vol. 7, n°5, pp. 1012-1016, 2020.
[3] A. Kumar et al, «BioSSL: A Framework for Semi-Supervised Learning in Biomedical Data,» Nature Machine Intelligence, vol. 5, n°1, pp. 45-56, 2023.
[4] H. Nguyen et al, «Impact of Class Imbalance on the Performance of Semi-Supervised Learning Algorithms.,» IEEE Access, vol. 9, pp. 101234-101245, 2021.
[5] World Health Organization(WHO), «Ebola Virus Disease.,» 2016. [En ligne]. Available: https://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease.. [Accès le 22 12 2024].
[6] X. Chen et al., «Combining Label Spreading and Self-Training for Sepsis Mortality Prediction,» Journal of Biomedical Informatics, vol. 87, pp. 1-9, 2018.
[7] Y. Zhang et al, «Semi-Supervised Learning for Mortality Prediction in ICU Patients,» IEEE Transactions on Biomedical Engineering,, vol. 67, n°4, pp. 1123-1130, 2020.
[8] Y. Zhang et al, «Self Training for Imbalanced Data: A Case Study in Healthcare.,» Nature ML, vol. 4, n°2, pp. 123-135, 2023.
[9] T. Hastie et al., The Elements of Statistical Learning, 2ème éd., Springer, Éd., New York, NY., 2009.
[10] S. Kotsiantis, Supervised Machine Learning: A Review of ClassificationTechniques, Informatica éd., vol. 31, 2007, pp. 249-268.
[11] E. Fini et Al., «Semi-supervised learning made simple with self-supervised clustering,» in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
[12] X. Zhu et A. B. Goldberg, Introduction to Semi-Supervised Learning, San Rafael, Californie: Morgan & Claypool, 2009.
[13] C.-W. Lin, C.-K. Chiang, Y.-A. Wang, Y.-L. Yang, H.-T. Li et T.-C. Lin, «ST-LP: self-training and label propagation for semi-supervised classification,» Multimedia Tools and Applications, vol. 83, p. 89335–89353, 2024.

[14] P. Zhongxing, Z. Gengzhong et H. Wei , «Semi-Supervised Learning with Close-Form Label Propagation Using a Bipartite Graph,» School of Computer Information Engineering, vol. 16, n°10, p. 1312, 4 October 2024.
[15] E. Wang et Y. Mei, «One new family of smooth semi-supervised support vector classifier based on Fourier series approximation technique,» Pattern Analysis and Applications, vol. 27, n°4, 14 11 2024.
[16] S. Mulangu et Al, «A Randomized, Controlled Trial of Ebola Virus Disease Therapeutics,» New England Journal of Medicine, vol. 381, n° 24, pp. 2293-2303, 2019.
[17] G. Kaur, «Entropy and Gini Index,» Towards Data Science, 5 Novembre 2023.
[18] A. A. Khan et Al., «Machine Learning for Epidemic Forecasting,» A Review, International Journal of Environmental Research and Public Health, vol. 17, n° 120, p. 7380, 2020.
[19] N. I. o. H. (NIH), «Ebola Virus Disease: A Comprehensive Overview,» 2020.