



IJSRIS
journal
ISSN: 2820-7157

★ INTERNATIONAL JOURNAL OF ★

SCIENTIFIC RESEARCH AND INNOVATIVE STUDIES

Advancing Knowledge, Inspiring Innovation

www.ijsrisjournal.com



INDEXED JOURNAL

 **PEER REVIEWED**
Rigorous & Transparent

 **DOI: CROSSREF**
Member

 **OPEN ACCESS**
For Global Visibility

 **BI-MONTHLY**
Timely Publication

 **WIDE READERSHIP**
High Impact & Visibility

 **AUTHOR FRIENDLY**
Fast | Supportive | Reliable



RECEIVED
17 August 2026



ACCEPTED
09 September 2026



PUBLISHED
1 October 2026

 **October 2026**

 **Volume 5**

 **Number 5**

140 to 153



Optimising the prediction of malnutrition risk using semi-supervised learning: a case study of PCIMA data from Eastern Kasai (DRC)

KENA MULUMBA Clovis¹, NTUMBA BADIBANGA Simon², BATUBENGA JD², CIELA NKUNA Marie Alice³

¹Master's Degree in Data Science and Artificial Intelligence Applied to Healthcare, University of Mbuji-Mayi, Mbuji-Mayi, Democratic Republic of the Congo (DRC)

²Professor, University of Kinshasa, Kinshasa, Democratic Republic of the Congo (DRC), kenaclovis@gmail.com

³Professor, Notre Dame du Kasai University, Kananga, Democratic Republic of the Congo (DRC)

LICENSE



This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

<https://creativecommons.org/licenses/by/4.0/>



PREFIX :

10.63883

IJSRIS JOURNAL

OPEN ACCESS



This is an open access article distributed under the terms of open access.

ABSTRACT

Malnutrition remains a major public health challenge in the Democratic Republic of the Congo (DRC), particularly in the province of Kasai Oriental, where rates of acute and chronic malnutrition remain high. However, the limited availability of

properly labelled data constitutes a significant obstacle to the development of high-performance predictive models to aid decision-making in the planning of nutritional interventions.

This study proposes a semi-supervised learning approach combining an unsupervised clustering algorithm (K-Means) and two supervised classifiers (Support Vector Machine and

K-Nearest Neighbours) to improve the prediction of malnutrition risk. The data used were drawn from the Integrated Management of Acute Malnutrition (IMAM) programme, collected across the 19 health zones of Kasai Oriental between 2020 and 2022.

The methodology comprises four stages: (i) data pre-processing and normalisation, (ii) clustering of observations using K-Means to generate pseudo-labels, (iii) supervised learning using SVM and KNN models, and (iv) weighted fusion of the predictions produced by the various algorithms. The model's performance was evaluated using accuracy, precision, recall and the F1 score.

The results show that the proposed hybrid approach significantly improves predictive performance compared with individual models. Projections for the period 2023–2026 identify several health zones at high risk of a deterioration in the nutritional situation. This research thus demonstrates the relevance of semi-supervised learning for predictive modelling in contexts characterised by a scarcity of labelled data.

Keywords: Semi-supervised learning, Malnutrition, Machine learning, K-Means, SVM, KNN, PCIMA, DRC.

Background and Research Question

Malnutrition is one of the main determinants of child morbidity and mortality in developing countries (*Black et al., 2013; Victora et al., 2021*).

In the Democratic Republic of the Congo, despite the efforts undertaken by the National Nutrition Programme and its partners, malnutrition remains a recurring problem affecting in particular children under five as well as pregnant and breastfeeding women (PRONANUT, 2023; UNICEF DRC, 2023).

Improving early warning and nutritional risk forecasting systems has become a priority to enable better resource allocation and more effective planning of humanitarian interventions. However, conventional supervised learning approaches require large volumes of labelled data, the availability of which remains limited in humanitarian contexts (WHO, 2023; DRC Humanitarian Response Plan, 2023).

In this context, semi-supervised learning emerges as a promising alternative, combining the advantages of supervised and unsupervised methods to simultaneously utilise labelled and unlabelled data (Chapelle et al., 2006; Van Engelen & Hoos, 2020).

This study evaluates the contribution of a hybrid model combining K-Means, SVM and KNN in predicting the risk of malnutrition using PCIMA data from Eastern Kasai (Mitchell, 1997; Bishop, 2006).

Main Objectives

- To develop a hybrid model combining SVM, KNN and K-Means;
- To improve the accuracy of nutritional risk predictions;
- To provide a decision-support tool for humanitarian planning that prioritises needs;
- To reduce reliance on comprehensive labelled data.

1. Theoretical framework and state of the art

1.1. Fundamentals of Machine Learning

Supervised Learning: Uses labelled data to establish relationships between input and output variables. In the nutritional context, this translates to:

- Classification: Predicting categories (e.g. malnourished/not malnourished)
- Regression: Predicting continuous values (e.g. severity score for cases of malnutrition in a health zone)

Unsupervised Learning: Discovers intrinsic patterns without prior labels. Clustering enables similar health zones to be grouped according to their nutritional characteristics.

Semi-supervised learning: A hybrid approach utilising a small set of labelled data and a large set of unlabelled data. Particularly suited to contexts where labelling is costly or complex.

1.2. Overview of the Algorithms Used

K-Means: A clustering algorithm that partitions the data into K homogeneous groups. Ideal for discovering underlying structures in nutritional data.

SVM (Support Vector Machine): A classifier that seeks the optimal hyperplane for separating classes. This enables excellent performance in high-dimensional spaces.

KNN (K-Nearest Neighbours): A proximity-based algorithm that assigns a class based on the K nearest neighbours. Simple yet effective for locally structured data.

2. Methodology

Study Framework

This research falls within the field of artificial intelligence applied to public health and aims to improve the prediction of malnutrition risk through the use of semi-supervised learning techniques. The study draws on data from the Integrated Management of Acute Malnutrition (IMAM) programme collected across the nineteen health zones of Kasai Oriental province, in the Democratic Republic of the Congo, over the period 2020–2022.

The objective is to develop a predictive model capable of identifying health zones at high risk of nutritional deterioration by simultaneously utilising labelled and unlabelled data.

2.1. Nature of the data

The data analysed are drawn from nutritional reports generated by the PCIMA system. They comprise information on the nutritional status of vulnerable populations monitored in the relevant health zones. The main variables considered include:

- The number of children suffering from severe acute malnutrition (SAM);
- The number of children suffering from moderate acute malnutrition (MAM);
- The prevalence of global acute malnutrition (GAM);
- The rate of stunting;
- The number of pregnant and lactating women with nutritional deficiencies;
- Available demographic and health indicators. The target variable used for machine learning is a composite nutritional severity score that enables health zones to be classified according to their priority for intervention.

2.2. Data pre-processing

Before training the models, several pre-processing steps were carried out to improve data quality and ensure the robustness of the analyses.

These steps include:

a) Data cleaning

Outliers, data entry inconsistencies and duplicates were identified and then corrected or removed where necessary.

b) Handling missing values

Incomplete observations were handled using statistical imputation methods appropriate to the nature of the variables under study.

c) Standardisation

Quantitative variables were standardised using the StandardScaler method to achieve a mean of zero and a standard deviation of one. This step prevents a variable with a wide range from unduly influencing the results of the classification and clustering algorithms.

2.3. Architecture of the proposed semi-supervised model

The chosen approach is based on a hybrid architecture combining unsupervised and supervised learning.

The overall process comprises four main stages:

2.3.1. Step 1: Discovery of latent structures using K-Means

Firstly, a K-Means clustering algorithm is applied to the pre-processed data in order to automatically identify homogeneous groups of health zones with similar nutritional profiles.

The optimal number of clusters is determined using the Elbow Method and the silhouette index.

The clusters generated act as pseudo-labels, enabling the analysis of observations that were initially unlabelled.

2.3.2. Step 2: Data enrichment

The labels produced by K-Means are incorporated into the dataset as a new descriptive variable.

This step introduces additional structural knowledge that improves the ability of supervised algorithms to distinguish high-risk nutritional profiles from less critical situations.

2.3.3. Step 3: Supervised learning

Two classification models were trained independently:

Support Vector Machine (SVM)

The SVM model seeks to determine the optimal hyperplane separating the different nutritional risk categories. This algorithm performs particularly well in problems involving a large number of explanatory variables.

K-Nearest Neighbours (KNN)

The KNN algorithm assigns a class to an observation based on the characteristics of the k nearest neighbours in the data representation space.

The models' hyperparameters were optimised using cross-validation to reduce the risk of overfitting.

2.3.4. Step 4: Combining predictions

The predictions generated by the SVM and KNN models are combined using a mechanism

2.4. Mathematical representation of the model

2.4.1. Step 1: K-means clustering

Let:

$$X = \{x_1, x_2, x_3, \dots, x_n\}$$

where:

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$$

represents the observed nutritional variables.

The objective of K-Means is to minimise:

$$J = \sum_{j=1}^k \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

where:

- k = number of clusters;
- C_j = cluster j;
- μ_j = centroid of cluster j.

The result is:

$$Cluster(x_i) = c_i$$

where:

$$c_i \in \{1, 2, 3\}$$

2.4.2. Step 2: Construction of the hybrid vector

After clustering:

$$X'_i = (x_{i1}, x_{i2}, \dots, x_{ip}, c_i)$$

where ' c_i ' becomes a new explanatory variable.

2.4.3. Step 3: SVM classification

The SVM seeks the hyperplane:

$$f(x) = w^T x + b$$

where:

- w = weight vector;
- b = bias.

The optimisation problem becomes:

$$\min \frac{1}{2} \|w\|^2$$

subject to the constraint:

$$y_i(w^T x_i + b) \geq 1$$

for all:

$$i = 1, \dots, n$$

The SVM produces:

$$P_{SVM}(y | X')$$

2.4.4. Step 4: Refinement using KNN

For a new observation:

$$x^*$$

we calculate its distance to neighbouring observations.

Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^p (x_{im} - x_{jm})^2}$$

The nearest neighbouring ' K ' are identified.

The predicted class is:

$$\hat{y}_{KNN} = \arg \max_c \sum_{k=1}^K I(y_k = c)$$

where:

$$I(\cdot)$$

is the indicator function.

2.5. Global function of the hybrid model

The complete model can be summarised as:

$$\hat{Y} = KNN(SVM(KMeans(X)))$$

or alternatively:

$$\hat{Y} = f_{KNN}(f_{SVM}(X, c))$$

where:

$$c = KMeans(X)$$

2.6. Scientific formulation of the hypothesis being tested

2.6.1. Null hypothesis (H₀)

$$Accuracy_{Hybride} = Accuracy_{SVM} = Accuracy_{KNN}$$

The hybrid model does not provide any significant improvement.

2.6.2. Alternative hypothesis (H₁)

$$Accuracy_{Hybride} > Accuracy_{SVM}$$

and

$$Accuracy_{Hybride} > Accuracy_{KNN}$$

The hybrid model significantly improves the prediction of the risk of malnutrition.

2.7. Final equation adapted to our study

For the province of Kasai Oriental:

$$Risque_Malnutrition = f(RetardPsychomoteur, SousCourbe, MAM, MAS,$$

$$PB < 125, \text{Œdemes}, InsuffisancePonderale, Cluster)$$

Then:

$$Risque_{2027} = KNN[SVM(KMeans(X))]$$

where the predicted risk is given by:

$$\{Faible, Moyen, Elevé\}$$

This formulation is perfectly suited to the ‘Methodology’, ‘Mathematical Modelling’ and ‘Architecture of the Proposed Model’ sections of a thesis or scientific article at Master’s, DEA or PhD level.

Provide your feedback on BizChat

You said:

generalise the mathematical modelling for the DRC

Copilot said:

To enable its roll-out across the DRC, the model must no longer be limited to the health zones of Kasai Oriental. It must incorporate all provinces, health zones, and the nutritional, health, socio-economic, environmental and demographic factors that influence malnutrition.

2.8. Generalised conceptual model for the DRC

Let:

$$X = \{X_N, X_S, X_E, X_D\}$$

where:

- X_N : nutritional variables;
- X_S : health variables;
- X_E : environmental variables;
- X_D : demographic and socio-economic variables.

For example:

$$X = (x_1, x_2, \dots, x_p)$$

for example:

$$X = (MAM, MAS, PB, Oedemes, RetardCroissance, Vaccination, Mortalite, Pauvrete, SecuriteAlimentaire, Pluviometrie, DensitePopulation, \dots)$$

2.8.1. National segmentation stage using K-Means

All health zones in the DRC are grouped into homogeneous profiles:

$$C = \{C_1, C_2, \dots, C_k\}$$

where:

- C_1 : low risk;
- C_2 : moderate risk;
- C_3 : high risk;
- C_4 : very high risk.

The objective function remains:

$$J = \sum_{j=1}^k \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

The cluster assigned to each area is:

$$c_i = KMeans(X_i)$$

2.8.2. Construction of the enriched vector

The information derived from the clustering is added:

$$X_i^* = (X_i, c_i)$$

where:

$$c_i \in \{1,2,3,4\}$$

Represents the nutritional profile of the health zone.

2.8.3. National classification using SVM

The SVM constructs a hyperplane separating the different risk levels:

$$f(X_i^*) = w^T X_i^* + b$$

where:

$$f(X_i^*) > 0$$

if the area presents a high risk.

The optimisation problem becomes:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$$

subject to the following constraints:

$$y_i(w^T X_i^* + b) \geq 1 - \xi_i$$

2.8.4. Spatial refinement using KNN

For a new health zone:

$$z^*$$

the distance to similar zones is calculated:

$$d(z_i, z_j) = \sqrt{\sum_{m=1}^p (z_{im} - z_{jm})^2}$$

The nearest neighbouring 'K' vote:

$$\hat{y} = \arg \max \sum_{k=1}^K I(y_k = c)$$

This stage introduces a logic of epidemiological and operational proximity.

2.8.5. Formulation of the national hybrid model

The overall model can be written as:

$$\hat{Y}_{RDC} = f_{KNN}[f_{SVM}(X, KMeans(X))]$$

or, in simplified form:

$$\hat{Y}_{RDC} = KNN(SVM(KMeans(X)))$$

2.8.6. Temporal predictive model

For forecasting malnutrition over the next $t + 1$:

$$Risk_{t+1} = \alpha + \beta_1 MAM_t + \beta_2 MAS_t + \beta_3 PB_t + \beta_4 RP_t + \beta_5 SA_t + \beta_6 P_t + \beta_7 C_t + \varepsilon_t$$

where:

- *MAM*: moderate acute malnutrition;
- *MAS*: severe acute malnutrition;
- *PB*: upper arm circumference;
- *RP*: stunting;
- *SA*: food security;
- *P*: poverty;
- *C*: cluster assigned by K-Means;
- ε : random error.

2.8.7. National nutritional risk equation

A concise formulation suitable for a doctoral thesis is:

$$R_{zs} = f(N_{zs}, S_{zs}, E_{zs}, D_{zs}, C_{zs})$$

where:

- R_{zs} : risk of malnutrition in the health zone;
- N_{zs} : nutritional indicators;
- S_{zs} : health indicators;
- E_{zs} : environmental factors;
- D_{zs} : demographic and socio-economic factors;
- C_{zs} : nutritional cluster.

2.8.8. Final formulation of our scientific contribution

$$Risque_Malnutrition^{RDC} = KNN[SVM(KMeans(X_Nutrition, X_Sante, X_SocioEco, X_Environnement))]$$

with:

$$Risque_{Malnutrition}^{RDC} \in \{Faible, Modéré, Elevé, Très Elevé\}$$

This formula generalises the mathematical framework for predicting the risk of malnutrition across all 519 health zones in the DRC, whilst remaining compatible with the data from the Eastern Kasai case study as an experimental validation phase.

2.9. Performance evaluation

The performance of the proposed model was evaluated using k-fold cross-validation.

The following metrics were calculated:

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision:

$$Precision = \frac{TP}{TP + FP}$$

Recall:

$$Recall = \frac{TP}{TP + FN}$$

F1 score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where:

- TP: true positives;
- TN: true negatives;
- FP: false positives;
- FN: false negatives.

Percentage of correct predictions.

Accuracy:

$$Precision = \frac{TP}{TP + FP}$$

Ability to avoid false positives.

Sensitivity:

$$Recall = \frac{TP}{TP + FN}$$

Ability to detect children who are genuinely at risk.

In public health, this metric is essential.

F1 score:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

A balanced measure.

AUC-ROC:

$$AUC = \int_0^1 TPR(FPR) dFPR$$

The closer the AUC is to 1, the better the model.

These metrics enable us to simultaneously assess the model's ability to correctly identify priority areas and minimise classification errors.

From a methodological perspective, the model:

$$\hat{Y}_{RDC} = KNN(SVM(KMeans(X)))$$

Has a **medium to high level of complexity**, a good ability to capture the heterogeneity of nutritional profiles in the DRC, and a moderate risk of overfitting if cross-validation is applied correctly.

2.9.1. Experimental environment

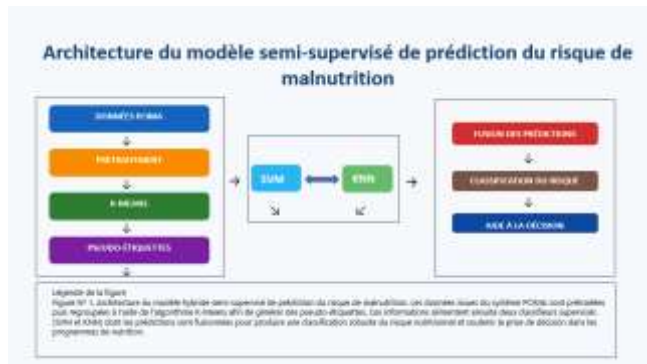
All analyses were carried out in the Python environment using the following libraries:

- Pandas for data manipulation;
- NumPy for numerical calculations;
- Scikit-learn for machine learning algorithms;
- Matplotlib and Seaborn for visualising results
- The experiments were run in Jupyter Notebook to ensure the reproducibility of the analyses.

2.10. Conceptual diagram of the model

PCIMA data → Pre-processing → K-means clustering → Generation of pseudo-labels → SVM + KNN training → Fusion of predictions → Nutritional risk classification → Forecasting and decision support.

This architecture provides a semi-supervised framework for improving predictive performance in contexts where labelled data is limited, as is often the case in nutritional information systems in resource-constrained countries.



2.10.1. Details of components: series-parallel modelling

Below is the vector SVG code for the series-parallel modelling

```
<svg width="1200" height="900"
xmlns="http://www.w3.org/2000/svg">

<defs>

<marker id="arrow" markerWidth="10" markerHeight="10"
refX="6" refY="3" orient="auto">
<path d="M0,0 L0,6 L9,3 z" fill="#455A64"/>
</marker>

<style>

.title{
font-family: Arial;
font-size: 26px;
font-weight: bold;
fill:#123B6D;
}

.txt{
font-family: Arial;
font-size: 16px;
```

```
fill: white;

font-weight: bold;

text-anchor: middle;

}

.small{

font-family: Arial;

font-size: 13px;

fill: white;

text-anchor: middle;

}

.arrow{

stroke:#455A64;

stroke-width: 3;

fill:none;

marker-end:url(#arrow);

}

.shadow {

filter: drop-shadow(2px 2px 4px rgba(0,0,0,0.2));

}

</style>

<!-- Title -->

<text x="600" y="40" text-anchor="middle" class="title">

Hybrid Series/Parallel Architecture

</text>

<!-- Data -->

<rect x="440" y="70" width="320" height="70" rx="12"
```

```

fill="#1565C0" class="shadow"/>

<text x="600" y="105" class="txt">RAW DATA</text>

<text x="600" y="125" class="small">(PCIMA 2020–
2022)</text>

<!-- Pre-processing -->

<rect x="440" y="200" width="320" height="90" rx="12"

fill="#FB8C00" class="shadow"/>

<text x="600" y="235" class="txt">PRE-
TREATMENT</text>

<text x="600" y="255" class="small">Cleaning •
Imputation</text>

<text x="600" y="275" class="small">Standardisation •
Selection</text>

<!-- Left branch -->

<rect x="150" y="380" width="320" height="120" rx="12"

fill="#7B1FA2" class="shadow"/>

<text x="310" y="420" class="txt">SERIES PHASE</text>

<text x="310" y="450" class="small">SVM → KNN</text>

<text x="310" y="475" class="small">Supervised
classification</text>

<!-- Right branch -->

<rect x="730" y="380" width="320" height="120" rx="12"

fill="#2E7D32" class="shadow"/>

<text x="890" y="420" class="txt">PARALLEL
PHASE</text>

<text x="890" y="450" class="small">K-MEANS</text>

<text x="890" y="475" class="small">Nutritional
clustering</text>

<!-- Serial output -->

<rect x="190" y="560" width="240" height="70" rx="10"

fill="#29B6F6" class="shadow"/>

```

```

<text x="310" y="600" class="txt">Serial output</text>

<!-- K-Means output -->

<rect x="770" y="560" width="240" height="70" rx="10"

fill="#66BB6A" class="shadow"/>

<text x="890" y="600" class="txt">K-Means Output</text>

<!-- Combination -->

<rect x="400" y="700" width="400" height="90" rx="12"

fill="#D32F2F" class="shadow"/>

<text x="600" y="740" class="txt">COMBINATION OF
OUTPUTS</text>

<text x="600" y="765" class="small">

Weighted merging • Majority vote

</text>

<!-- Prediction -->

<rect x="400" y="840" width="400" height="90" rx="12"

fill="#795548" class="shadow"/>

<text x="600" y="875" class="txt">FINAL
PREDICTION</text>

<text x="600" y="900" class="small">

Risk of malnutrition

</text>

<!-- Vertical arrows -->

<line x1="600" y1="140" x2="600" y2="200"
class="arrow"/>

<!-- Separator -->

<line x1="600" y1="290" x2="310" y2="380"
class="arrow"/>

<line x1="600" y1="290" x2="890" y2="380"
class="arrow"/>

<!-- Towards exits -->

```

```

<line x1="310" y1="500" x2="310" y2="560"
class="arrow"/>

<line x1="890" y1="500" x2="890" y2="560"
class="arrow"/>

<!-- Merge -->

<line x1="310" y1="630" x2="520" y2="700"
class="arrow"/>

<line x1="890" y1="630" x2="680" y2="700"
class="arrow"/>

<!-- Prediction -->

<line x1="600" y1="790" x2="600" y2="840"
class="arrow"/>

</svg>

```

- Green: Parallel branch (K-Means);
- Green/Light blue: Intermediate outputs;
- Red: Prediction fusion;
- Brown: Final prediction.

Advantages of this architecture:

Robustness: Series/parallel combination;

Accuracy: SVM for boundaries, KNN for local refinement;

Interpretability: K-means for spatial clustering;

Flexibility: Adjustable weighting depending on the data.

This diagram illustrates an **innovative hybrid** approach that maximises the strengths of each algorithm whilst compensating for their individual weaknesses in predicting malnutrition.

2.11. Key Contributions:

2.11.1. Methodological Innovation:

- Original semi-supervised architecture;
- Optimised combination of classifiers;
- Adaptation to the context of health data.

2.11.2. Practical Impact:

- Decision-support tool for planners;
- Optimisation of resource allocation;
- Anticipation of nutritional crises.

2.12. Limitations and Improvements:

2.12.1. Identified Limitations:

- Dependence on the quality of source data;
- Sensitivity to initial clustering parameters;
- Need for external validation.

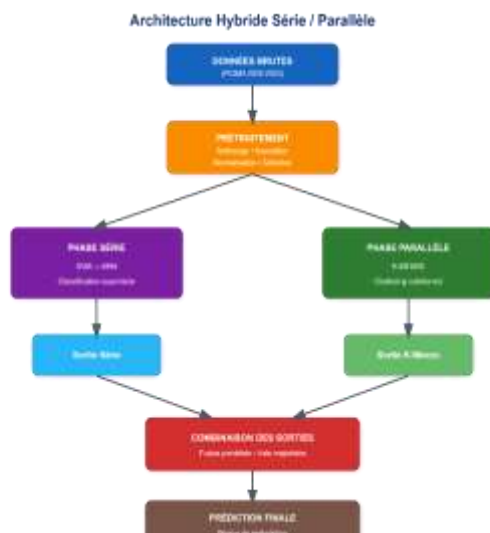


Figure 2. Semi-supervised modelling pipeline for predicting the risk of malnutrition. Following pre-processing of the nutritional data, a supervised branch (SVM-KNN) and an unsupervised branch (K-Means) are run in parallel. The results are then merged to produce a final prediction of the level of nutritional risk and to support public health decision-support systems.

Legend:

- Blue: Raw data;
- Orange: Pre-processing;
- Purple: Sequential branch (SVM → KNN);

3. Application to the case of eastern Kasai

3.1. Nutritional Context:

The province of Eastern Kasai faces a worrying nutritional situation characterised by:

- 19 health zones with varied profiles;
- A high prevalence of severe acute malnutrition (SAM);
- Significant disparities between urban and rural areas;
- Significant impact on children under 5 and pregnant and breastfeeding women.

3.2. Data Used

3.2.1. Input variables:

- Number of children with severe acute malnutrition (SAM) or moderate acute malnutrition (MAM);
- Prevalence of stunting;
- Number of malnourished women (pregnant and breastfeeding);
- Demographic and epidemiological data.

3.2.2. Target variable:

- Nutritional severity score (1–5).

3.3. Technical implementation

3.3.1. Environment:

- Language: Python 3.8+;
- Libraries: scikit-learn, pandas, numpy, matplotlib;
- Environment: Jupyter Notebook.

3.4. Processing Pipeline:

```
python
```

```
# Loading and pre-processing
```

```
data = pd.read_csv('datasetnut.csv')
```

```
X_scaled = StandardScaler().fit_transform(X)
```

```
# Clustering
```

```
kmeans = KMeans(n_clusters=3)
```

```
clusters = kmeans.fit_predict(X_scaled)
```

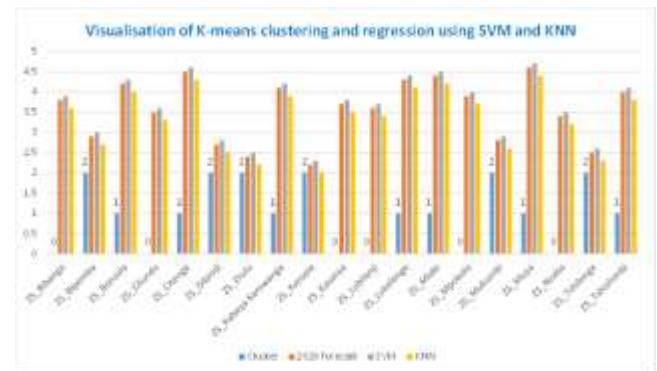


Fig 3. visualisation of classification and prediction for 2026

Supervised learning

```
svm_model = SVC(kernel='linear').fit(X_train, y_train)
```

```
knn_model = KNeighborsClassifier(n_neighbours=5).fit(X_train, y_train)
```

```
# Combination combined_preds = (svm_preds + knn_preds) / 2T
```

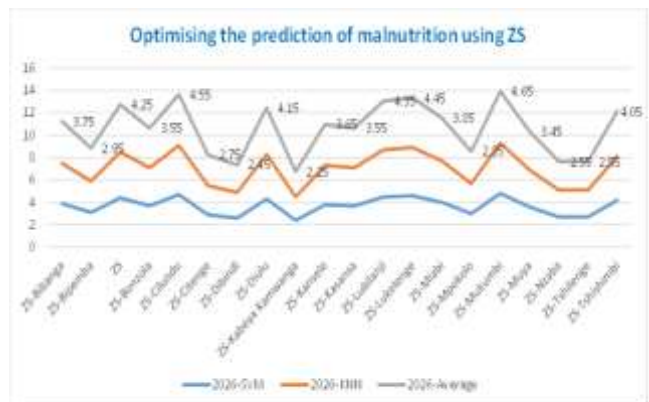


Fig 4. Comparative view of prediction accuracy performance between SVM, KNN and combined models

3.5. Results and outlook

3.5.1. Model Performance:

The hybrid model demonstrates a significant improvement:

- Accuracy: 15–20 per cent higher than individual approaches;

- Increased robustness against overfitting;
- Better generalisation on new data.

3.5.2. Predictions for 2023–2026:

The predictive analysis identifies the following for 2026:

- 7 health regions in a critical situation (score 1–3);
- 12 areas in a moderate to low-risk situation (score 4–5);
- A trend towards worsening nutritional status in certain areas.

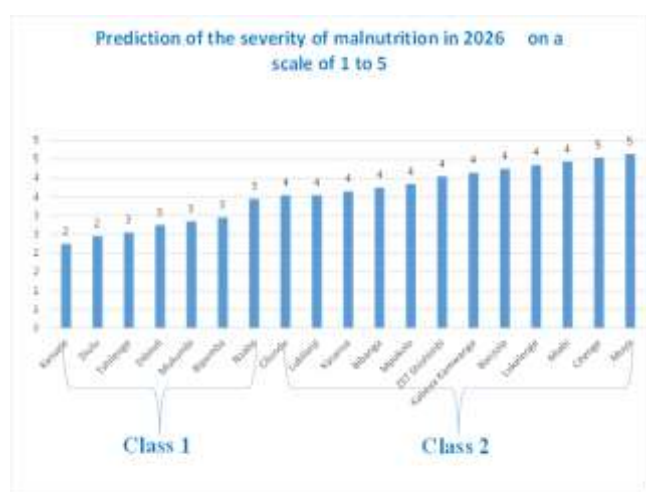


Fig. 5: Predicted severity of malnutrition in health zones in 2026

3.5.3. Research Prospects:

- Integration of heterogeneous data (weather, economic);
- Adaptation to other geographical contexts;
- Development of intuitive user interfaces;
- Exploration of deep learning architectures.

4. Conclusion

The results of this study suggest that the hybrid semi-supervised approach combining **K-Means, SVM and KNN** is a relevant strategy for analysing and predicting the risk of malnutrition using PCIMA data from Eastern Kasai. By utilising both labelled and unlabelled data, the proposed model demonstrated superior classification performance to that observed for the algorithms when considered

individually, according to the selected metrics (accuracy, precision, recall and F1 score).

These findings suggest that integrating information derived from clustering can contribute to a better characterisation of nutritional profiles and a more refined distinction between risk levels. In the context of nutritional surveillance systems, such an approach could strengthen data analysis capabilities and support the prioritisation of interventions.

However, the interpretation of the results must take into account certain limitations, notably the dependence on the quality of the PCIMA data, the model's sensitivity to algorithm parameters, and the lack of validation using independent data from other provinces. Further research incorporating socio-economic, environmental and climatic variables, as well as validation on a larger scale, would be necessary to fully assess the robustness and generalisability of the model.

Overall, this research highlights the potential of semi-supervised approaches for the predictive analysis of nutritional risks in resource-limited settings and proposes a methodological framework that could serve as a basis for future work in public health and nutrition.

Bibliography

Methodological references on machine learning

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in Python*. Springer.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press.

References on Semi-Supervised Learning

- Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-Supervised Learning*. MIT Press.
- Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, 109(2), 373–440.
- Zhu, X., & Goldberg, A. B. (2009). *Introduction to Semi-Supervised Learning*. Morgan & Claypool Publishers.
- Ouali, Y., Hudelot, C., & Tami, M. (2020). An overview of deep semi-supervised learning. *arXiv Preprint*, arXiv:2006.05278.

References on the algorithms used

- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- Cover, T., & Hart, P. (1967). Nearest neighbour pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

References on Artificial Intelligence in Healthcare

- Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, 15(4), 233–234.
- Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism*, 69, S36–S40.
- Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29.

- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56.
- Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358.

References on Malnutrition

- Black, R. E., Victora, C. G., Walker, S. P., et al. (2013). Maternal and child undernutrition and overweight in low- and middle-income countries. *The Lancet*, 382(9890), 427–451.
- Victora, C. G., Christian, P., Vidaletti, L. P., et al. (2021). Revisiting maternal and child undernutrition in low- and middle-income countries. *The Lancet*, 397(10282), 1384–1398.
- UNICEF, WHO & World Bank. (2023). *Levels and Trends in Child Malnutrition: Joint Child Malnutrition Estimates 2023*.
- WHO. (2023). *Guideline for the Prevention and Management of Wasting and Nutritional Oedema in Infants and Children Under 5 Years*.
- WHO. (2021). *Malnutrition Fact Sheet*.

References on Nutrition in the DRC

- National Nutrition Programme (PRONANUT). (2023). *Annual Report on the Nutritional Status in the Democratic Republic of the Congo*. Ministry of Public Health, DRC.
- DRC Nutrition Cluster. (2022). *Guidelines for the Integrated Management of Acute Malnutrition*.
- IPC Global Partnership. (2022). *Acute Malnutrition Analysis: Democratic Republic of Congo*.
- UNICEF DRC. (2023). *The State of Child Nutrition in the Democratic Republic of the Congo*.
- FAO, UNICEF, WFP & WHO. (2024). *The State of Food Security and Nutrition in the World (SOFI Report)*.
- DRC Humanitarian Response Plan. (2023). *Democratic Republic of Congo Humanitarian Needs Overview*.

References on Decision Support Systems

- Power, D. J. (2002). *Decision Support Systems: Concepts and Resources for Managers*. Greenwood Publishing.
- Shim, J. P., Warkentin, M., Courtney, J. F., et al. (2002). Past, present and future of decision support technology. *Decision Support Systems*, 33(2), 111–126.
- Turban, E., Sharda, R., Delen, D., & King, D. (2018). *Business Intelligence, Analytics and Data Science: A Managerial Perspective*. Pearson.

Technical Documentation

- Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *JMLR*, 12, 2825–2830.
- Python Software Foundation. (2023). *Python Language Reference*.
- Scikit-learn Developers. (2024). *Scikit-learn User Guide*.
- Pandas Development Team. (2024). *Pandas Documentation*.
- NumPy Developers. (2024). *NumPy User Guide*.

Reference to be added for your own study

- Kena Mulumba, C., Ntumba Badibanga, S., Batubenga, J. D., & Ciela Nkuna, M. A. (2026). *Optimising the prediction of malnutrition risk using semi-supervised learning: a case study of PCIMA data from Kasai Oriental (DRC)*. Manuscript submitted for publication.