New Analytical Approaches and Practical Implications in Senegalese Clinical Data

Demba Diallo[1], Amadou Dahirou Gueye[2], Mouhamadou Lamine Ba[3]

Faculty of Applied Sciences and Information and Communication Technologies (SATIC), Alioune Diop University of Bambey (UADB), Senegal[1]
Faculty of Advanced Science and Technology (STA), Amadou MahtarMbow University of Diamniadio(UAM), Senegal[2]
Polytechnic Higher School (ESP), Cheikh Anta Diop University of Dakar (UCAD)Senegal[3]
Corresponding author: Demba Diallo, Email: demba.diallo@uadb.edu.sn

This paper investigates the landscape of Senegalese clinical data, emphasizing the diversity and unique challenges inherent in its analysis. The motivation behind this study is rooted in the need to understand health trends and improve healthcare within Senegal. Given the specific demographic profiles and medical practices in the country, this research aims to utilize advanced statistical methods to derive meaningful insights from complex data. The study employs Principal Component Analysis (PCA) to reduce data dimensionality while maintaining essential information, thus clarifying intricate relationships between medical variables. Additionally, machine learning techniques, such as the XGBoost algorithm, are applied to predict post-operative complications and map networks of infectious diseases. These methods provide significant advancements in clinical data analysis, revealing critical insights for public health in Senegal. Through detailed case studies, the practical application of these methods is demonstrated, highlighting their potential to enhance patient care and disease prevention. Ultimately, this research underscores the importance of adopting advanced analytical approaches tailored to the Senegalese context, aiming to foster improvements in healthcare and medical research across the country.

Keywords: Senegalese clinical data, PCA, XGBoost, mapnetworks, diseases.

in: *Advancements in Communication and Systems*. Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, India., 2023, pp. 509–520. DOI: https://doi.org/10.56155/978-81-955020-7-3-45

1. Introduction

At the intersection of medical research and practice, the partnership between Alioune Diop University of Bambey and the Cheikh Ahmadou Bamba Khadim Rassoul Hospital Center of Touba emerges as a driving force in advancing the exploration of clinical data in Senegal. In a context where a thorough understanding of health trends is crucial, clinical data [1]are essential for offering valuable insights into the country's demographic profiles and medical practices. Renowned experts emphasize the crucial importance of this data in improving healthcare and guiding public policies.

To fully exploit this potential, advanced analytical methods are necessary. This article therefore explores how approaches such as Principal Component Analysis (PCA), machine learning (e.g.,XGBoost algorithm) [2][3], and network analysis are applied to Senegalese clinical data. These methods enable the transformation of complex data into meaningful insights, paving the way for a better understanding of health patterns and more effective interventions.

Through concrete case studies, we examine how these analytical approaches are used to explore patient typology, predict post-operative complications, and map networks of infectious diseases. These practical examples demonstrate the tangible impact of advanced clinical data analysis on Senegal's public health, thus opening new perspectives for more precise healthcare and more effective disease prevention.

The structure of this paper is as follows:
Following a critical review of relevant literature, we first present the state of the art by addressing previous research on clinical data analysis in Senegal and the medical landscape of the country. Then, we explore the methodologies used such as Principal Component Analysis, machine learning, and network analysis, with specific case studies. We then present the results and performance of our study. Finally, we conclude.

2. State of the Art

The judicious exploitation of health data allows for the improvement of healthcare quality and decision-making in the Senegalese healthcare system. [4] Health data analysis is essential for evaluating the effectiveness of health programs and guiding future investments towards the most promising interventions. [5] Significant contributions have been made by renowned researchers in this field of medical data analysis.

2.1 Related Work

Existing literature on medical data analysis provides a comprehensive overview of advancements in this field.

Agrawal and Prabakaran [6] addressed the potential impact of Big Data [7] in the healthcare domain, highlighting its benefits and challenges. The authors emphasize that Big Data offers significant opportunities to enhance patient outcomes, particularly by enabling the development of personalized and effective treatments. However, they also note hurdles such as data fragmentation, high costs, and data ownership issues. They explain how these approaches enable the analysis of complex and unstructured datasets, thereby facilitating the discovery of patterns and useful correlations for clinical decision-making.

Using oncology as an example, the authors illustrate how Big Data can directly impactpatient care by enabling a better understanding of the genetic causes of diseases and facilitating the development of more effective treatments. They also highlight challenges related to data ownership, sharing, and privacy, drawing on experiences from the US, UK, and other global healthcare systems. In summary,

their paper highlights the potential of Big Data to transform healthcare while emphasizing the need to address challenges associated with its use, particularly concerning data management and patient privacy.

On the other hand, Cahan et al. [8] underscored the importance of data quality in the effective use of Big Data in medicine. The authors emphasize that while predictive algorithms and machine learning play a crucial role in improving healthcare, biased outcomes can result from biased input data. Thus, they advocate for an inductive approach to Big Data, focusing on data quality and representativeness to ensure fairness and generalization of predictive models. The paper highlights the potential of Big Data to address gaps in preclinical research, accelerate the development of personalized medicine, and enhance healthcare delivery. However, the authors caution that risks associated with Big Data primarily stem from the data itself, including intrinsic biases and gaps in training datasets.

To mitigate these risks, the authors propose several strategies, such as annotating datasets with labeling metadata, redesigning data collection methods to ensure data variety, and using analysis methods such as contrastive principal component analysis to visualize entrenched data biases.

In conclusion, the paper underscores the importance of data transparency, patient health information privacy, and awareness of data gaps to realize the potential of Big Data in personalized medicine while avoiding perpetuating health inequalities.

Additionally, Macias et al. [9] explored the use of big data from electronic medical records (EMRs) in pediatric clinical care. The authors highlighted several potential uses of big data in pediatric clinical care, including improving care quality, collaborative learning, clinical prediction, decision support, and personalized medicine. It utilizes relevant case examples, such as pediatric sepsis diagnosis and management, to illustrate these concepts.

Lastly, Khan et al. [10] provided an in-depth review of big data applications in healthcare, focusing on data analysis for effective care and disease diagnosis. The paper highlights the growing need to use big data in healthcare due to increasing available data, the digital transformation of healthcare systems, and increasing pressure on healthcare providers to deliver effective and personalized care. The primary objective of the analysis is to evaluate existing efforts in big data analysis for healthcare, identify gaps and challenges, and propose solutions to improve the performance of big data-based healthcare applications. The analysis concludes that despite significant progress in the field of big data analysis for healthcare, challenges remain, including reducing treatment costs, minimizing errors, and improving care quality. It is suggested that adopting advanced hybrid models based on machine learning and cloud computing could help overcome these challenges.

The authors conclude by emphasizing the importance of continuing research in the field of big data analysis for healthcare, focusing on developing advanced disease diagnostic models, using machine learning, and improving large-scale health data management.

Overall, while these works have made significant contributions to medical data analysis, their limitation lies in their lack of consideration for the specificities of the Senegalese clinical data landscape. A thorough understanding of these characteristics is essential to develop tailored and effective medical data analysis approaches in this context.

In response to this lack of attention to the specificities of Senegalese medical data, our research has focused on adapting and extending existing methods to better address the needs of this context. The results of our study underscore the importance of contextualizing medical data analysis methods to ensure their relevance and applicability in diverse contexts, including that of Senegal.

2.2 Medical Landscape of Senegal

Let's delve into the healthcare landscape of Senegal, a unique territory where the richness of medical information intertwines with the country's specific challenges. The organization of the Senegalese socio-health sector is pyramid-shaped, aligning with the country's administrative division. The system comprises:

- A central level that includes the Cabinet of the Minister of Health and Social Action, the General Secretariat, the General Directorates, the National Directorates, the Attached Central Services, the National Centers for Social Reintegration, and Level 3 Public Health Facilities.
- A strategic intermediate level that includes medical regions, Regional Hygiene Brigades (BRH), Regional Social Action Services (SRAS), and Level 2 Public Health Facilities.
- An operational peripheral level with Health Districts, Hygiene Sub-Brigades, Departmental Social Action Services, Centers for Promotion and Social Reintegration (CPRS), and Level 1 Public Health Facilities. [11]

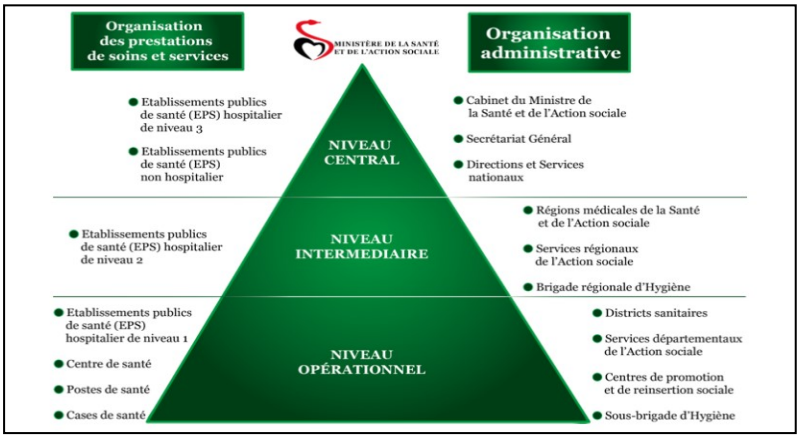

Figure 1. Health and social pyramid: This figure illustrates the pyramid structure of the Senegalese health and social system, highlighting the various levels of health facilities and administrative bodies. This hierarchical organization ensures that healthcare services are efficiently distributed and managed across the country.Source: Ministry of Health and Social Action in its "Directory of Health and Social Statistics 2020".

Clinical data in Senegal encompasses medical information related to patients in healthcare facilities across the country, such as health centers, hospitals, and other institutions. This includes medical records, test results (for example, for tropical diseases), vaccination records, and other information relevant to public health.

Clinical protocols in Senegal are aligned with guidelines issued by the Ministry of Health and Social Action (MSAS). These guidelines define standards for data collection and recording within the framework of primary healthcare, vaccination programs, and other public health initiatives. The underlying objective is to ensure consistent delivery of care nationwide.

Concurrently, data confidentiality and security are major concerns in Senegalese healthcare facilities. Patient data confidentiality, in compliance with national and international standards, is reinforced by cybersecurity measures, such as data encryption, aimed at protecting this sensitive information.

In connection with this, the MSAS has a Health Information System (HIS) with the aim of harmonizing monitoring and evaluation indicators of the healthcare system. For over a decade, the MSAS has been committed to results-based management. It is in this vein that the National Health and Social Information System (NHSIS) has been strengthened to ensure quality throughout the process of production, processing, and utilization of health data. [12]

Furthermore, the adoption of common standards and terminologies in the Senegalese context is a fundamental pillar for promoting system interoperability. Interoperability is defined as the ability of two or more computer platforms to operate and interact jointly. Senegal adopted DHIS2 as the national platform for managing aggregated and individual data through its Tracker module in 2013. [13]

This integration of concepts in the Senegalese context underscores the importance of adapting analytical approaches to the country's specific needs.

Figure 2 illustrates the flow and creation of health data as per the guidelines of the Ministry of Health and Social Action in Senegal.

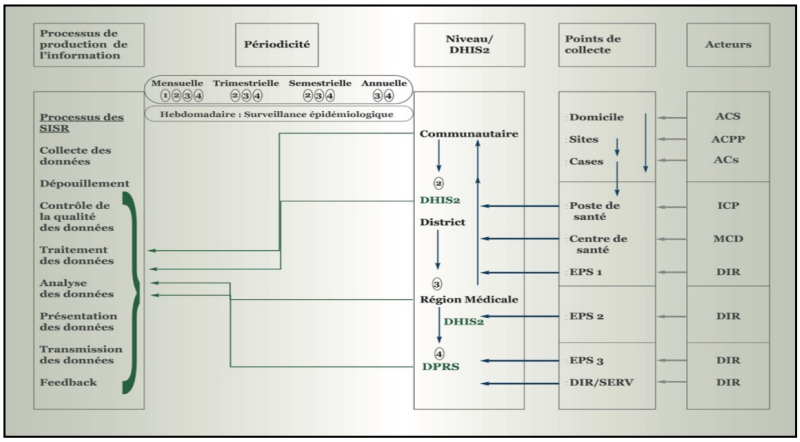

Figure 2. Information circuit and production of health data: The figure illustrates the flow and creation of health data as per the guidelines of the Ministry of Health and Social Action in Senegal. Source: Ministry of Health and Social Action in its "Directory of Health and Social Statistics 2020".

3. Methodologies and Practical Cases

The quest for a deep understanding of Senegalese clinical data now leads us into the captivating world of analytical methods. An essential technique in this approach is Principal Component Analysis (PCA), defined as a method for transforming a set of potentially correlated variables into a set of uncorrelated variables, known as principal components [14]. Applying the eXtreme Gradient Boosting (XGBoost) algorithm [15] to Senegalese clinical data involves rigorous selection of methods based on specific research questions. This classification algorithm is used to predict diagnoses or assess the risk of certain conditions. Here, emphasis is placed on adapting XGBoost to the unique characteristics of Senegalese medical data, incorporating variables such as socio-economic and environmental factors. Network analysis is described as the study of the structural and dynamic properties of networks, and their impact on the behavior of complex systems [16]. Applied to data from the Cheikh Ahmadou Bamba Khadim Rassoul Hospital in Touba, network analysis represents an innovative approach based on the principles of complex network theory. The application of complex network theory to Senegalese

clinical data necessitates a specific methodology. This methodology, encompassing data collection, modeling relationships, and using community detection algorithms, ensures a visual representation of the complex interconnections among clinical elements.

This section thus lays the groundwork for the following sections, which will explore practical applications, results, and the coherent integration of these diverse approaches.

3.1 Principal Component Analysis for Patient Typology

Principal Component Analysis (PCA) emerges as a powerful method for exploring patient typology from clinical data. The main objective is to reduce the dimensionality of the data while retaining essential information, thus allowing for simplified visualization of complex relationships between different medical variables.

The approach began with the collection of diverse clinical data, including parameters such as medical history, diagnostic test results, and patient demographic characteristics. These data were then subjected to in-depth analysis through PCA. The aim was to discover underlying trends, variable clusters, and define principal components representative of the diversity of patient profiles.

3.2 Prediction of Post-Operative Complications

In this case study, the XGBoost machine learning method was deployed to develop a predictive model for post-operative complications in patients. By integrating variables such as medical history, preoperative test results, and demographic characteristics, the model demonstrated significant accuracy in anticipating risks. Figure 3 depicts the structure of the XGBoost algorithm used in the case study for predicting post-operative complications in patients.

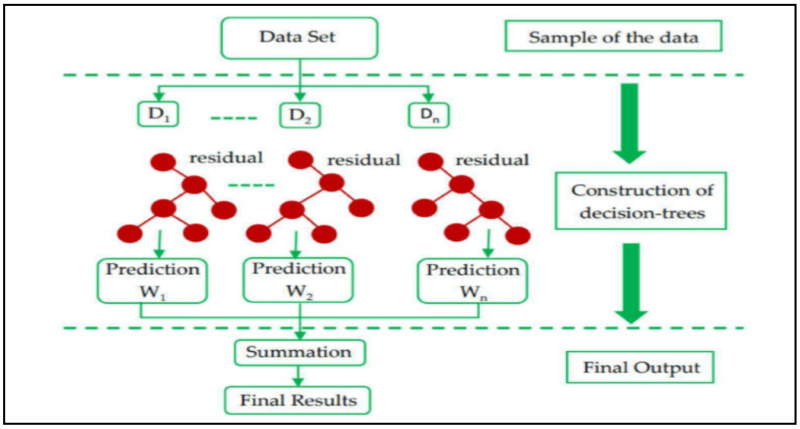

Figure 3. XGBoost Algorithm Structure.[17]

3.3 Mapping Infections Disease Networks

Our third study focuses on the use of a network approach to analyze clinical data related to infectious diseases within the Senegalese population. This mapping of interactions between different diseases has revealed complex patterns of co-occurrence and potential risk factors.

4. Results, Discussion and Performance

The results of the case studies reveal significant insights.
For Case Study 1, Principal Component Analysis (PCA) was used to study patient typology. Figure 4 displays the variance ratio explained by each principal component. It appears that the first two principal components explain a large portion of the total variance in the data, while the other components have a lesser impact. This suggests that most of the important information is captured by the first two principal components. Figure 5 shows the projection of the data onto the first two principal components. This allows visualization of the underlying structure of the data after dimension reduction. One can observe if there are clusters or patterns in the data that might indicate relationships between variables.

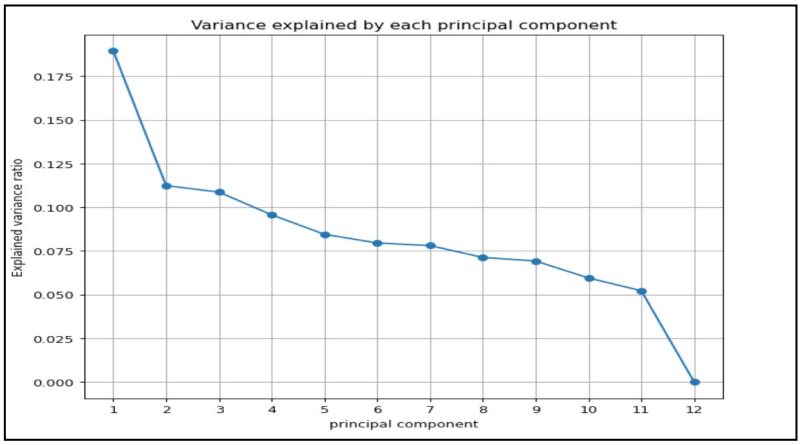

Figure 4. Visualization of the variance explained by each principal component.

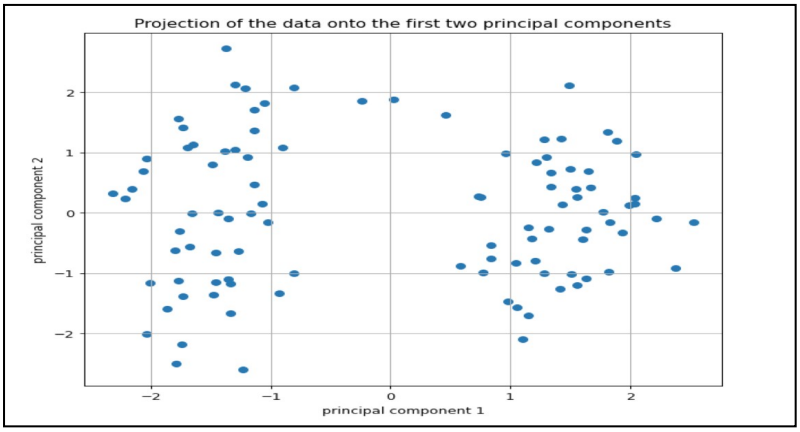

Figure 5. Visualization of the data projection onto the first two principal components.

This study could be significant in several areas of public health and biomedical research in Senegal:

- Detection of patient subgroups: By identifying clusters or patterns in clinical data, this study could help detect subgroups of patients with similar characteristics. This could facilitate a better understanding of risk factors, disease patterns, and treatment responses, potentially improving prevention and treatment strategies.
- Personalization of healthcare: Understanding the structure of clinical data would also allow for better customization of healthcare according to the specific needs of each patient. For instance, by identifying high-risk patient profiles for certain diseases, healthcare professionals could recommend targeted preventive interventions to improve population health.
- Optimization of medical resources: Understanding patterns of medical attendance, hospitalization, and test results, health authorities could better allocate medical resources to meet the population's needs. This could contribute to a more efficient use of medical infrastructure and improve access to healthcare.

In summary, this study could have a positive impact on public health in Senegal by providing valuable insights to guide health policies, medical interventions, and biomedical research.

In Case Study 2, using the XGBoost algorithm, which focused on predicting postoperative complications in patients, the performance of the model was evaluated using several metrics, including accuracy, recall, F1 score, and a confusion matrix (Figure 6).

The confusionmatrix [18] itself is a graphical representation of the model's performance on a classification task. In this case, it shows the number of correctly and incorrectly classified cases of postoperative complications. True positives (TP) represent the number of instances correctly classified by the model as having complications (17 in Figure 6). False positives (FP) represent the number of instances incorrectly classified as having complications (32 in Figure 6). True negatives (TN) represent the number of instances correctly classified as not having complications (108 in Figure 6), and finally, false negatives (FN) represent the number of instances incorrectly classified as not having complications (43 in Figure 6).Given the confusion matrix, we can calculate the performance metrics:[19]

$$Precision = \frac{TP}{TP+} x\ 100\% \quad (1)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+F} x\ 100\% \quad (2)$$

$$Recall = \frac{TP}{TP+FN} x\ 100\% \quad (3)$$

$$F1\ Score = 2\ x\ \frac{Recall\ x\ Precision}{Recall+Precis} x\ 100\% \quad (4)$$

As summarized in Table 1, the XGBoost model achieved an accuracy of 62.50% (equation 2), a recall of 28.33% (equation 3), and an F1 score of 0.31 (equation 4).

While the accuracy of the model appears acceptable (62.50%), the recall (28.33%) indicates that the model may miss a significant number of complication cases. This is further highlighted by the confusion matrix, which shows a substantial number of false negatives (32). This finding suggests that the model needs improvement in its ability to correctly identify complication cases. These limitations have important implications in the context of Senegalese clinical data, where accurate identification of postoperative complications is crucial for patient care.

Several steps can be taken to improve the model's performance. These include reviewing the features used for training, adjusting the model's parameters, and collecting additional data to better represent the diversity of Senegalese patients.

Table 1. XGBoost model performance.

Performance	Value
Accuracy	62.50%
Recall	28.33%
F1 Score	0.31

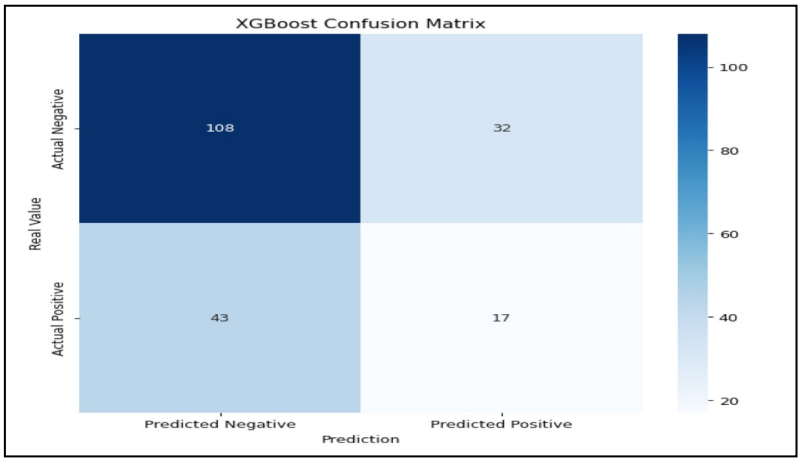

Figure 6. Confusion matrix in graphical form.

Finally, for case study 3 on mapping infectious disease networks, the results obtained provide a thorough analysis of the network based on symptoms. This network graph highlights how certain diseases share common symptoms, which can complicate diagnosis and treatment(Figure 7).

Table 2 provides a detailed statistical analysis of the interconnectedness between different diseases through various measures:

- Interconnected disease clusters: The following diseases are strongly interconnected in the network: Hepatitis A, Dengue, Malaria, Yellow Fever, Typhoid. This suggests that they share several common symptoms or are closely linked in their spread.
- Central diseases: The central diseases in the network, based on degree centrality, are Malaria and Dengue. This means that they have a high number of interactions with other diseases in the network.
- Betweenness centrality: This measure evaluates the degree of control that a disease exerts over the flow of information between other diseases. In this case, Malaria exhibits the highest betweenness centrality among diseases, indicating its crucial role in transmitting symptoms between other diseases.
- Closeness centrality: This metric assesses how quickly a disease can spread information to other diseases within the network. Diseases with high closeness centrality include Malaria, Dengue, Yellow Fever, and Typhoid, indicating their ability to rapidly influence other diseases within the network.

- Communities detected by modularity: The community detection method identified a single group comprising all diseases, suggesting a strong interconnection among them.
- Correlation between symptoms and diseases: This analysis shows which diseases share similar symptoms. For example, Fatigue is associated with several diseases, including Hepatitis A, Malaria, and Dengue.

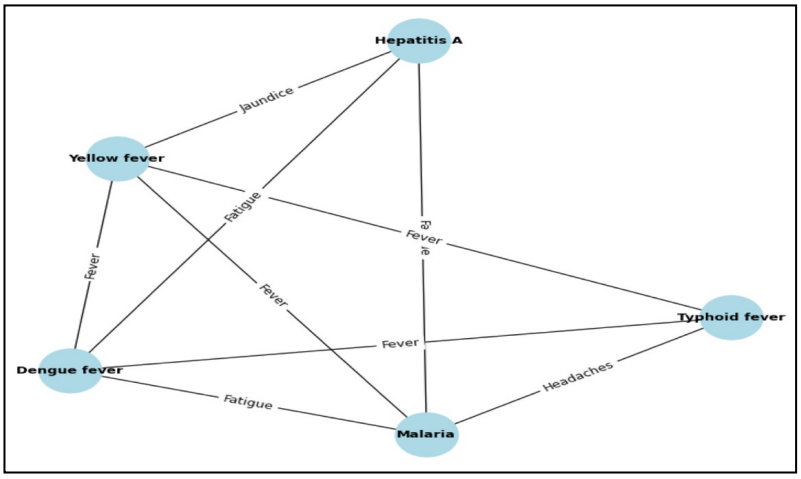

Figure 7. Infectious Disease Network based on Symptoms.

Table 2. Identifying disease clusters and their interconnectedness through specific measures

Statistics	Values
Interconnected disease clusters	['Typhoid fever', 'Malaria', 'Dengue fever', 'Yellow fever', ' Hepatitis A']
Central diseases	['Malaria', 'Dengue fever']
Betweenness centrality	['Hepatitis A': 0.0, 'Malaria': 0.05, 'Dengue fever ': 0.05, 'Yellow fever': 0.05, 'Typhoid fever': 0.0]
Closeness centrality	['Hepatitis A': 0.8, ' Malaria ': 1.0, 'Dengue fever': 1.0, 'Yellow fever': 1.0, 'Typhoid fever': 0.8]
Communities detected by modularity	[frozenset({'Dengue fever', 'Malaria', 'Typhoid fever', 'Hepatitis A', 'Yellow fever'})]
Correlation between symptoms and diseases	['Constipation': {'Typhoid fever'}, 'Abdominal pain': {'Yellow fever'}, 'Muscle pain': {'Dengue fever'}, 'Fatigue': {'Dengue fever', 'Hépatitis A', 'Malaria'}, 'Fever': {'Malaria', 'Dengue fever', 'Typhoid fever', 'Yellow fever'}, 'Shivers': {'Malaria'}, 'Jaundice': {'Hépatitis A', 'Yellow fever'}, 'Headaches': {' Typhoid fever ', 'Malaria'}, 'Nausea': {'Hépatitis A'}, "Loss of appetite": {'Hépatitis A'}, 'Vomiting': {'Yellow fever'}, 'Skin rash': {'Dengue fever'}]

These results could be used to inform strategies for prevention and control of infectious diseases within the Senegalese population. For instance, a targeted awareness campaign focusing on common symptoms of central diseases could be implemented to enhance early detection and management of illnesses. Health authorities could utilize this information to prioritize healthcare resources and surveillance efforts, placing greater emphasis on central diseases and the most significant clusters.

In summary, this analysis provides valuable insights into the structure of interactions among infectious diseases within the Senegalese population, which can guide public health interventions to prevent and control the spread of these diseases.

Definitively, the results from the three case studies offer a comprehensive understanding of various aspects of clinical and infectious data, ranging from patient typology to predicting postoperative complications and mapping networks of infectious diseases. These analyses can offer valuable insights to improve patient care and prevent infectious diseases.

5. Conclusion

In summary, through various case studies, we have explored the diversity of clinical data, health trends, and specific challenges within the Senegalese context. The results of these analyses provide a deep understanding of health patterns, demographic profiles, and complex interactions between diseases, thereby paving the way for more precise and effective interventions. From patient typology to predicting postoperative complications, and mapping networks of infectious diseases, these analytical approaches enable the identification of patient subgroups, personalized healthcare, and optimization of medical resource allocation.

For the future, it is essential to continue these advanced analytical efforts by strengthening collaboration between local practitioners, public health researchers, and data analysis experts. By fully harnessing the potential of clinical data, we can improve patient care, guide health policies, and contribute to disease prevention and control within the Senegalese population. Thus, this article highlights the importance of clinical data analysis for public health in Senegal and opens the door to new perspectives for more effective, precise, and inclusive healthcare.

References

[1] Pan, X., Zhou, X., Song, H.-m., Zhang, R., & Zhang, T. (2012). Enhanced data extraction, transforming and loading processing for Traditional Chinese Medicine clinical data warehouse. In 2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom) (pp. 57-61). doi: 10.1109/Health-Com.2012.6380066

[2] Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer.

[4] Organisation mondiale de la Santé (OMS). (2019). Situation sanitaire au Sénégal. Retrieved from https://www.afro.who.int/fr/countries/senegal.

[5] Organisation mondiale de la Santé (OMS). (2020). Situation sanitaire au Sénégal. Retrieved from https://www.afro.who.int/countries/senegal.

[6] Agrawal, R., &Prabakaran, S. V. (2020, March 5). Big data in digital healthcare: Lessons and recommendations for general practice. Springer.

[7] Cuzzocrea, A. (2021). Big data lakes: Models, frameworks, and techniques. In 2021 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 1-4). doi: 10.1109/Big-Comp51126.2021.00010

[8] Cahan, E. M., Hernandez-Boussard, T., Thadaney-Israni, S., & Rubin, D. L. (2019, August 19). Putting the data before the algorithm in big data addressing personalized healthcare. Springer.

[9] Macias, C. G., Remy, K. E., &Barda, A. J. (2022, November 24). Utilizing big data from electronic health records in pediatric clinical care. Springer.
[10] Khan, S., Khan, H. U., & Nazir, S. (2022). Systematic analysis of healthcare big data analytics for efficient care and disease diagnosing. Springer.
[11] Collé, C. A. (2022, January 22). Santé, Situation économique et sociale du Sénégal 2019 (pp. 74). Agence Nationale de la Statistique et de la Démographie.
[12] Ministère de la Santé et de l'Action Sociale (MSAS). (Oct-Nov-Dec 2022 – No. 54). Bulletin trimestriel du SISS (pp. 01). Retrieved from https://www.sante.gouv.sn/Pr%C3%A9sentation/bulletin-trimestriel-n%C2%B0-1-du-syst%C3%A8me-d%E2%80%99information-sanitaire-et-social
[13] Ministère de la Santé et de l'Action Sociale (MSAS). (Oct-Nov-Dec 2022 – No. 54). Bulletin trimestriel du SISS (pp. 09). Retrieved from https://www.sante.gouv.sn/Pr%C3%A9sentation/bulletin-trimestriel-n%C2%B0-1-du-syst%C3%A8me-d%E2%80%99information-sanitaire-et-social
[14] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. Philosophical Magazine, 2(6), 559-572.
[15] Niu, Y. (2020). Walmart sales forecasting using XGBoost algorithm and feature engineering. In 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE) (pp. 458-461). doi: 10.1109/IC-BASE51474.2020.00103.
[16] Barabási, A.-L. (2016). Network science. Cambridge University Press.
[17] Khan, M. N., &Alabdullah, A. A. (2022). Compressive strength estimation of steel-fiber-reinforced concrete and raw material interactions using advanced algorithms. Polymers, 14(15), 306
[18] Karimi, Z. (2021). Confusion Matrix. Retrieved from https://www.researchgate.net/publication/355096788.
[19] Siddique, M. A. A., Ferdouse, J., Habib, M. T., Mia, M. J., & Uddin, M. S. (2022). Convolutional Neural Network Modeling for Eye Disease Recognition. International Journal of Online and Biomedical Engineering (iJOE), 18(09), 115-122. doi: 10.3991/ijoe.v18i09.29847.