

A Generative AI-Enabled Contextual Data Curation Framework for Emerging Demographic Health and Malaria Indicator Survey Data

¹*Udo, Ifiok J., **James, Idara I., **Udoeka, Ifreke. J., ***Dan, Emmanuel A., ****Usip, Usoro I.,
***Asuquo, Suzanlindsy I. & ***Usip, Patience U.

*Department of Information Systems, University of Uyo. Nigeria

**Department of Computer Science, Akwa Ibom State University, Ikot Akpaden

***Department of Computer Science, University of Uyo, Nigeria

****Department of Cybersecurity, University of Uyo, Uyo.

E-mail: ¹udojames@uniuyo.edu.ng

ABSTRACT

The increasing complexity and volume of demographic health and malaria indicators surveys data from diverse sources present significant challenges for effective data curation and contextual interpretation pipeline, especially in machine learning applications reliant on curated data to improve its predictive performance. While Generative AI has advanced language understanding, its integration into data curation frameworks remains underexplored. This paper proposes an intelligent framework that embeds Generative AI algorithms into NLP model to enhance semantic representation of data during real-time curation, focusing on analysis that transforms textual contents to equivalent fact embeddings representations required for machine learning. By embedding Generative AI algorithms over conventional data preprocessing approach, the framework improves both the efficiency and accuracy of contextual data interpretation, enabling more robust machine learning task for evidence-based decision-making. Experimental results with selected tree-based algorithms demonstrate that our approach outperforms traditional methods in terms of average accuracy (F1-Score) of approximately 5%, with data obtained from Nigeria demographic health and Malaria indicator surveys. This signifies the imperative contextually representing curated data for machine learning task, as it enhances the quality of outputs and supports evidence-based decision-making.

Keywords: AI, Algorithms, Demography, Machine Learning, Tree-Based Algorithms, Malaria Burden, Evidence-Based Decision-Making

CISDI Journal Reference Format

Udo, Ifiok J., James, Idara I., Udoeka, Ifreke. J., Dan, Emmanuel A., Usip, Usoro I., Asuquo, Suzanlindsy I. & Usip, Patience U. (2026): A Generative AI-Enabled Contextual Data Curation Framework for Emerging Demographic Health and Malaria Indicator Survey Data. Computing, Information Systems, Development Informatics & Allied Research Journal. Vol 17 No 2, Pp 51-60. Available online at www.isteams.net/cisdijournal. dx.doi.org/10.22624/AIMS/CISDI/V17N2P5

1. INTRODUCTION

The surge in unstructured demographic health survey (DHS) and Malaria indicator Survey (MIS) data (Zegeye et al., 2025) driven by the availability of modern data collection, storage and computing infrastructure (Awoyelu et al., 2014) has presented significant challenges for accurate contextual representation of data for machine learning task (Methuku, 2025).

Traditional NLP-based classification models frequently encounter difficulties with contextual ambiguity, semantic variation, multilingual expressions (Lalrinawma and Nunsanga, 2025), and the dynamic nature household's language, often resulting in suboptimal contextual encoding of data for machine learning. While Generative AI has demonstrated advanced capabilities in language understanding, its application within data curation frameworks remains largely unexplored (James and Osubor, 2023; Bhardwaj et al. 2024). Efficient management and interpretation of demographic health data are critical for informed public health decision-making. However, the proliferation of data sources, coupled with the increasing volume and heterogeneity of collected data, poses substantial challenges for data curation, especially in low-resource settings.

Traditional approaches to data preprocessing (Udo and Afolabi, 2011) and curation often struggle to capture the nuanced context and semantic relationships embedded within textual survey responses, limiting the effectiveness of downstream machine learning models. Recent advancements in Generative Artificial Intelligence (AI) (Udo et al., 2026), particularly in natural language processing (NLP) (Chowdhary, 2020), have enabled unprecedented capabilities in understanding and representing complex language constructs. Despite these advances, the application of Generative AI within data curation frameworks for demographic and health datasets remains largely unexplored. Harnessing these technologies offers the potential to address existing gaps in semantic representation and contextual interpretation, ultimately improving the quality and utility of curated data in intelligent applications across online and mobile applications ecosystems (James et al., 2016).

This research presents an intelligent framework that integrates Generative AI into natural language processing (NLP) models for real-time curation of textual data from demographic health and Malaria indicator surveys. The study addresses two primary research questions: how generative AI can enhance contextual and semantic representations of textual data for machine learning tasks, and how transforming textual data into numerical representations improves machine learning outcomes. By leveraging Generative AI for semantic enrichment during data curation, the proposed approach aims to increase the efficiency and accuracy of converting contextual textual information into numerical formats suitable for machine learning algorithms. Specifically, the methodology involves extracting survey data, integrating and cleaning it using NLP techniques, generating numerical equivalents via a Generative AI preprocessing pipeline, and subsequently training and testing with select tree-based machine learning algorithms. Experimental results demonstrate that the Generative AI-enhanced curation framework outperforms traditional preprocessing methods, leading to improved predictive performance and supporting more robust, evidence-based decision-making in public health.

The remaining section of the paper is subdivided into relate work in section 2.0, proposed framework is addressed in section 3.0, section 4.0 describes the experiment, results and discussion.

2. RELATED WORKS

James and Osubor (2023) highlighted the critical role of data curation in enhancing the performance of supervised machine learning models within malaria research. Studies emphasize that integrating datasets from sources such as Nigeria's Malaria Indicator Survey (MIS) and Demographic and Health Survey (DHS) through comprehensive preprocessing techniques, including data cleaning, integration, feature selection, handling missing values, and transformation can significantly improve data quality.

These processes reduce redundancy, inconsistencies, and noise, thereby creating unified datasets that are more suitable for predictive modeling. Improved data curation not only facilitates more accurate and robust machine learning outcomes but also supports evidence-based decision-making in public health interventions targeting malaria. Samuel et al. (2024) exemplify a comprehensive data preprocessing pipeline by addressing missing data and class imbalance and selecting relevant features for analysis of the 2018 Demographic and Health Survey (DHS) data. Specifically, the authors employed imputation methods for missing values, SMOTE for class imbalance, and information ranking for feature selection. Their approach, however, primarily focused on the numeric attributes of the DHS dataset. Similarly, Faremi et al. (2024) tackled class imbalance in the 2021 Malaria Indicator Survey (MIS) data using SMOTE, further demonstrating the value of such preprocessing techniques. Mariki et al. (2022) adopted a different approach, handling missing clinical symptom and patient data by deleting incomplete records and encoding numeric features to facilitate machine learning model implementation. Nkiruka et al. (2021) engineered features to pinpoint climatic factors contributing to malaria incidence, while Khan et al. (2020) aggregated and reduced the feature space to streamline the detection process of malaria parasites in red blood cells, thereby increasing model transparency and efficiency.

The existing data preprocessing methods, while useful for traditional data analysis, are insufficient for modern machine learning workflows. These conventional techniques often overlook the importance of semantic and contextual understanding, which are essential for enabling machine learning algorithms to achieve higher predictive accuracy. In contrast, generative AI approaches particularly those utilizing transformer-based architectures have demonstrated remarkable capabilities in various domains, especially in capturing semantic and contextual nuances to encode and process complex, large-scale datasets. For example, Shaikat et al. (2026) highlighted several gaps in existing research and introduced an Explainable Machine Learning Framework that incorporates BERT (Koroteev, 2021) for classifying AI-generated text from multiple sources. The approach combines contextual semantic representation learning with explainable AI methods, utilizing fine-tuned BERT embeddings to extract robust contextual features.

This enables the classification of text generated by various large language models, as opposed to relying on outputs from a single data source. However, while this framework advances the field in terms of semantic and contextual representation for text classification, it does not address the unique challenges posed by datasets containing mixed data types, such as numerical, categorical, and textual information which are common in data harvested demographic health and malaria indicator surveys. Usip et al. (2021) developed a PeNLP parser to extract and visualize geolocation information related to maternal, neonatal, and child healthcare, focusing on social determinants of health. The approach used indicators of healthcare at the household and community levels. Although the study applied NLP techniques, the contextual representation of the information was constrained by the limitations of the available datasets.

In light of these limitations, there is a compelling need for generative AI-enabled contextual data curation tailored specifically to emerging demographic health and malaria indicator surveys. Unlike traditional preprocessing methods, generative AI models can integrate semantic, contextual, and structural information from heterogeneous data sources, including numerical, categorical, and textual fields. This capability is critical for enabling more accurate and actionable insights in public health informatics research, where the complexity and diversity of survey data often hinder conventional analysis.

Leveraging transformer-based architectures, generative AI can bridge existing gaps by providing a unified, context-aware representation of complex datasets, thus empowering machine learning workflows to achieve robust predictive performance and deeper interpretability. As the scale and intricacy of health data continue to grow, generative AI-driven curation is essential for unlocking the full potential of data-driven decision-making in public health interventions, particularly for diseases like malaria.

3. A GENERATIVE AI-ENABLED CONTEXTUAL DATA CURATION FRAMEWORK

The proposed Generative AI-Enabled Contextual Data Curation Framework overcomes several limitations of traditional health data management systems, such as weak data encoding mechanisms. These limitations often introduce bias in machine learning outcomes, as the contextual meaning of health information is rarely captured by the purely numerical representations needed for machine learning. Instead of serving as a data storage system, this framework provides an intelligent evidence-generation platform that integrates data, applies natural language processing (NLP), enriches semantics contextually, and leverages machine learning. By embedding Generative AI algorithms for the curation process, machine learning models can extract more meaningful insights from Demographic Health Surveys (DHS) and Malaria Indicator Surveys (MIS). This leads to improved data interoperability, supports explainable analytics, and ultimately enables more informed public health decision-making for malaria prevention and broader healthcare interventions. The workflow of the generative AI-enabled contextual data curation framework is illustrated in Figure 2.

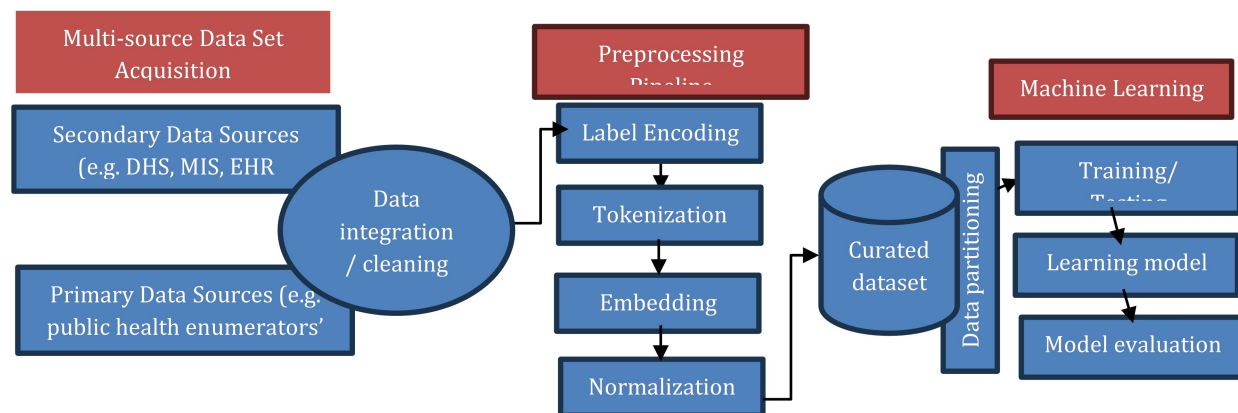


Figure 2: Workflow of the Generative AI-Enabled Contextual Data Curation Framework

This framework operates through a structured, sequential workflow designed to generate insights for malaria control and public health planning. The major components are: data acquisition layer, preprocessing pipeline using NLP, and machine learning pipeline.

Data Acquisition

The first step involves registration and request for the dataset from the DHS online repository (<https://dhsprogram.com/>) also collection of MIS data from a variety of primary sources including households in select households in Akwa Ibom State, Nigeria. This ensures a comprehensive dataset that captures diverse perspectives and information relevant to malaria research and intervention.

Data Integration and Cleaning

The collected datasets, which may be heterogeneous in nature, are integrated and cleaned. This process enhances the quality and consistency of the data, making it suitable for subsequent analysis.

NLP Preprocessing

The framework applies natural language processing (NLP) preprocessing techniques. These include label encoding, tokenization, contextual embedding, and normalization, all of which help standardize and structure the data for advanced analysis. After data integration, the unified dataset undergoes a comprehensive Natural Language Processing (NLP) preprocessing pipeline, especially targeting textual variables, metadata, and survey descriptions to prepare the data for advanced machine learning and analytical tasks.

Label Encoding

Categorical variables such as malaria status, geographical regions, education level, and intervention categories are converted into numerical representations through label encoding. This transformation is crucial for making the data compatible with machine learning algorithms that require numeric input.

Tokenization

Textual survey responses, questionnaire items, and metadata are split into meaningful linguistic units, known as tokens.

Embedding Generation

The tokenized text is further processed to generate contextual vector representations using transformer-based language models like BERT. These embeddings capture semantic relationships among survey variables, enabling the analytical framework to understand context and meaning beyond isolated keywords.

Normalization

Numerical feature values and embedding vectors are standardized through normalization. This step ensures consistent feature scaling across the dataset and improves the convergence of machine learning models during training. Collectively, these preprocessing steps transform heterogeneous survey information into structured, machine-readable representations, laying the foundation for robust data analysis and modeling.

Contextual Curation

Once processed, the data is curated into a semantically enriched repository using Generative AI techniques. This step organizes the information in a way that enables deeper understanding and knowledge extraction for related context tokens and types.

Machine Learning Analysis

The curated data undergoes machine learning and Generative AI analysis. This phase generates contextual evidence, predictive insights, and robust decision support, all of which contribute to precise evidence-based malaria burden classification and public health planning.

4. EXPERIMENT

Data collection

A combination of textual and numeric data from the MIS dataset is summarized in Table 1. This excerpt presents the summary of features obtained from both the MIS and DHS datasets, which were curated from selected households across the six geopolitical zones of urban and rural Nigeria. In total, 70,428 household records with 40 features were extracted from these repositories for implementing the proposed learning model. Additionally, the MIS and DHS feature categories related to malaria burden considered in this study are those comprehensively analyzed and documented by James and Osubor (2025b).

Table 1: Features associated with MIS and DHS

Selected Input Features
Wealth index
number of household members
Number of children 5 and under
Main floor material
Wall material
Number of mosquito bed nets
Number of children under mosquito bed net previous night
Type of mosquito bed net(s) slept under last night
Children under 5 slept under mosquito bed net last night (household questionnaire)
Result of the malaria test
Relationship to household head
Ideal number of children (grouped)
Husband/partner worked in last 7 days/12 months
Respondent currently working
Cost of treatment of fever
Household has: electricity
Literacy
Type of cooking fuel
Number of days after fever began sought advice or treatment
Currently pregnant
Duration of pregnancy (months)
Proportion of children under 5 and below over the total no of household
Proportion of ideal children under 5 and below over the no. of ideal children in the household
Levels of burden children under 5 who slept under mosquito net
Level of burden associated with proportion of mosquito net per no. of household.
Level of burden associated with proportion of children under mosquito net previous night over ideal children.

Contextual Curation using Generative AI

In the implementations leveraging Generative AI, BERT WordPiece (Shaikat et al., 2025) tokenization is used to preserve the contextual semantics of the text, ensuring that subtle meanings are retained.

The pseudocode for BERTH algorithm is listed in Algorithm 1.

Algorithm 1: BERT for Malaria Indicator Survey Data Classification

Input:

Curated Dataset D
Pre-trained BERT
Epochs E
Learning Rate α

Output:

BERT Model

Begin

Initialize BERT parameters

For epoch = 1 to E do

Shuffle training dataset

For each mini-batch B do

Tokenize text

Encode using BERT

Obtain CLS embeddings

Predict class labels

Compute CrossEntropy Loss

Compute gradients

Update parameters using AdamW optimizer

End For

Validate model

End For

Save BERT model

Return trained BERT model

End

Machine Learning module

A tree-based boosting ensemble algorithm combined with a majority voting mechanism to accurately classify malaria burden at the household level was implemented using the curated data. This approach from the strength ensemble learning, where boosting algorithms iteratively reduce bias and enhance prediction accuracy. Also, the transparency associated with tree-based ensemble learning supports interpretability in machine learning outcomes, especially in handling complex, heterogeneous data such as emerging DHS and MIS data.

4. RESULTS AND DISCUSSION

Tables 2, 3 and 4, tabulate the results of select tree-based algorithms using conventional data preprocessing versus contextual data curation framework as well as the comparative performances, respectively.

Table 2: Summary of Results of machine learning model without conventional preprocessing technique

Algorithms	Precision	Recall	AUC
XGboost	0.9473	0.8396	0.9501
Gradient Boosting	0.7431	0.8357	0.9461
Catboost	0.6725	0.8816	0.9412
Adaboost	0.6812	0.8532	0.9377
LightGBM	0.7295	0.8169	0.8875

Table 3: Summary of results of machine learning with contextual data curation framework

Algorithm	Precision	Recall	AUC
XGboost	0.9658	0.8816	0.9717
Gradient Boosting	0.7800	0.8762	0.9650
Catboost	0.7061	0.9211	0.9589
Adaboost	0.7153	0.8854	0.9548
LightGBM	0.7661	0.8577	0.9218

Table 4: Summary of comparative performances between the results in Tables 2 and 3

Algorithm	F1-Score (Table 2)	F1-Score (Table 3)	Performance improvement
XGboost	0.7902	0.8327	0.0537
Gradient Boosting	0.7861	0.8246	0.0490
Catboost	0.7590	0.7976	0.0500
Adaboost	0.7549	0.7926	0.0500
LightGBM	0.7699	0.8084	0.0500

The tree-based model implemented within the contextual data curation framework demonstrate that XGBoost consistently outperforms other models, achieving the highest Precision (0.9658), F1-score (0.8327), and AUC (0.9717), which highlights its superior discriminative power in classifying malaria burden (James and Osubor, 2025). Gradient Boosting secures the second-best performance, with a balanced Precision (0.7800) and Recall (0.8762), resulting in an F1-score of 0.8246. CatBoost excels in Recall (0.9211), reflecting strong sensitivity for detecting classes of malaria burden cases, though its Precision is relatively lower. Both AdaBoost and LightGBM show consistent improvements across all metrics, indicating greater predictive stability following enhancements in contextual data curation and feature representation.

5. CONCLUSION

In conclusion, while DHS and MIS surveys are key resources for tracking population health and guiding malaria policy, their growing complexity presents data management challenges that traditional methods struggle to address. Emerging Generative AI technologies offer promising solutions for integrating, interpreting, and synthesizing data, but their application to the curation of demographic and malaria survey data is still limited. However, the findings presented in this paper align with expectations for a Generative AI-enabled contextual data curation framework, where higher data quality, improved semantic embeddings, and optimized preprocessing yield measurable gains in classifier performance. The observed approximate 5% improvement across evaluation metrics validates the impact of enhanced data engineering and model optimization, while preserving the relative ranking of all evaluated algorithms.

REFERENCES

1. Awoyelu, I., Omodunbi, T., & Udo, J. (2014). Bridging the gap in modern computing infrastructures: Issues and challenges of data warehousing and cloud computing. *Computer and Information Science*, 7(1), 33.
2. Bhardwaj, E., Gujral, H., Wu, S., Zogheib, C., Maharaj, T., & Becker, C. (2024, June). Machine learning data practices through a data curation lens: An evaluation framework. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1055-1067).
3. Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603-649.
4. Faremi, A. S., Akinnuwesi, B., Mbunge, E., Mashwama, P., Fashoto, S. G., Ncube, P. Z., ... & Metfula, A. (2024, March). Machine Learning Models for Identifying Factors Influencing and Predicting Malaria Among Children Under Five Years in Nigeria. In *2024 Conference on Information Communications Technology and Society (ICTAS)* (pp. 88-94). IEEE. South Africa. DOI: 10.1109/ICTAS59620.2024.10507142
5. James, I., Ansa, G., & Udoeka, I. (2016). Conceptual framework on overcoming the challenges of multiple vehicle registration in nigeria: A mobile application approach. *IOSR Journal of Mobile Computing and Application*, 3(5), 1-3.
6. James, I and Osubor, V. (2023). Curated big data for supervised machine learning: A case of malaria indicator and demographic health surveys data. [Online]. Available: <https://library.ncs.org.ng/download/curated-big-data-for-supervised-machine-learning-a-case-of-malaria-indicator-and-demographic-health-surveys-data/>
7. James, I., & Osubor, V. (2025). Machine learning evidence towards eradication of malaria burden: A scoping review. *Applied Computer Science*, 21(1), 44-69
8. Khan, A., Gupta, K. D., Venugopal, D., & Kumar, N. (2020, July). Cidmp: Completely interpretable detection of malaria parasite in red blood cells using lower-dimensional feature space. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.
9. Koroteev, M. V. (2021). BERT: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.
10. Lalrinawma, C., & Nunsanga, M. V. (2025, February). Sentiment Analysis Review: Methods and Challenges. In *2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)* (pp. 831-838). IEEE.

11. Mariki, M., Mkoba, E., & Mduma, N. (2022). Combining clinical symptoms and patient features for malaria diagnosis: machine learning approach. *Applied Artificial Intelligence*, 36(1), 2031826.
12. Methuku, V. (2025). NLP and AI for public health intelligence: Automating disease surveillance from unstructured data. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 2(1), 43-56.
13. Nkiruka, O., Prasad, R., & Clement, O. (2021). Prediction of malaria incidence using climate variability and machine learning. *Informatics in medicine Unlocked*, 22, 100508.
14. Samuel, O., Zewotir, T., and North, D. (2024). Application of machine learning methods for predicting under five mortality: analysis of Nigerian demographic health survey 2018 dataset. *BMC medical informatics and decision making*. 24(1): 86
15. Shaikat, M. T. A., Chowdhury, S. H., Mamun, M., Shaha, P., Hossain, M. M., & Iqbal, S. (2026). An Explainable Machine Learning Framework with BERT for Multi-Source AI-Generated Text Classification. *Telematics and Informatics Reports*, 100350.
16. Udo, I. J., & Afolabi, B. (2011). Hybrid data reduction technique for classification of transaction data. *Journal of Computer Science and Engineering*, 6(2), 12-16.
17. Udo, I. J., James, I. I., Dan, E. A., & Ofon, U. (2026). Integrated GenAI Data Observatory Framework for Evidence-based Analytics of Malaria Burden. *International Journal of Innovative Computing*, 16(1-2), 271-278.
18. Usip, P. U., Ekpenyong, M. E., Ijebu, F. F., Usang, K. J., & Udo, I. J. (2021). PeNLP parser: an extraction and visualization tool for precise maternal, neonatal and child healthcare geo-locations from unstructured data. In *Deep Learning in Biomedical and Health Informatics* (pp. 157-181). CRC Press.
19. Zegeye, A. F., Mekonen, E. G., Gebrehana, D. A., Tekeba, B., & Tamir, T. T. (2025). Spatial variation and multilevel determinants of malaria infection among pregnant women in Sub-Saharan Africa: using malaria indicator surveys. *BMC infectious diseases*, 25(1), 654.