Global Journal of Sciences
Volume 3, Issue 1, 2026, Pages 91-106
DOI: https://doi.org/10.48165/gjs.2026.3109; ISSN: 3049-0456
https://acspublisher.com/journals/index.php/gjs

Research Article

K-means Clustering in Identifying Epidemiological Patterns of Measles in United States of America Between 2009-2024

Basil B. Duwa[1,2], Murat Sayan[3,4], James Gambu[5], Huzaifa Umar [1], Celestine Iwendi[1,6]

[1]*Operational Research Center in Healthcare, Near East University, Nicosia, TRNC, Mersin 10, Türkiye.*
[2]*Department of Biomedical Engineering Near East University, Nicosia, TRNC, Mersin 10, Türkiye.*
[3]*Educational and Research Hospital, PCR Unit, Kocaeli University, Kocaeli, Türkiye.*
[4]*DESAM Research Institute, Near East University, Nicosia, TRNC Türkiye.*
[5]*Department of Medical Microbiology and Clinical Microbiology, Health Science Institute, Near East University, Nicosia, TRNC Türkiye.*
[6]*Centre of Intelligence of Things, University of Bolton, United Kingdom.*

ARTICLE INFO

Article History
Received 17 March 2026
Revised 26 April 2026
Accepted 05 May 2026
Published 14 May 2026

Keywords
Epidemiological surveillance
K-means clustering
Measles
Public health analytics
Vaccination coverage
United States

ABSTRACT

Tracking measles vaccination coverage is an essential factor in preventing subsequent outbreaks and building herd immunity. Thus, recognizing trends in immunization and surveillance assist in tailoring public health efforts in different locations. The study employs K-means clustering to detect distinctive epidemiological trends among U.S.A. states using measles, mumps, and rubella (MMR) vaccine coverage data obtained across states from kindergarteners from 2009 to 2024 using the CDC dataset. States sourced dataset on population size, percentage surveyed, and coverage of vaccination groups were preprocessed and standardized to exclude outliers. Using the elbow approach and silhouette analysis, we calculated the ideal number of clusters. K-means clustering was then utilized to categorize states based on similarities in surveillance features. In our results, three major clusters appeared: Cluster 1 featured states with substantial populations and wide variation in survey coverage. Cluster 2 included smaller states with regular high survey rates and MMR coverage. Cluster 3 showed modest population sizes and coverage rates. Visualizations indicated major differences in survey approach and immunization criteria across clusters. K-means clustering successfully identified hidden patterns in MMR vaccine monitoring trends across USA. These results provide useful insights for modifying public health efforts and enhancing vaccination reporting systems across the country.

1. INTRODUCTION

Measles is an extremely infectious viral disease transmitted by the Measles morbillivirus, a *Paramyxoviridae* member. It spreads mostly by respiratory droplets and is capable of staying airborne for a period of two hours after an infected person coughs or sneezes [1]. This disease is characterized by conjunctivitis, a distinctive maculopapular rash, fever, and cough. Those diagnosed also exhibit prevalent symptoms such as blindness, encephalitis, pneumonia, and death, particularly in kids under

*Corresponding author. Email: basil.barthduw@neu.edu.tr/huzaifa.umar@neu.edu.tr

age five, who are immunocompromised and malnourished. Measles requires roughly 95% vaccine coverage to obtain community immunity due to its high basic rate of reproduction (R), which ranges from 12 to 18. In spite of the measles-mumps-rubella (MMR) vaccination, measles continues [2] to pose a serious threat to public health around the world. Over ten million incidents and 136,000 deaths from measles occurred globally in 2023, according to the World Health Organization (WHO) [3]. The COVID-19 pandemic-related disruptions in immunizations are responsible for the 20% rise in incidence compared to 2022. Although the disease burden is highest in low-income nations, vaccination hesitancy, disinformation, and under-vaccinated areas have also resulted in occurrences in high-income nations, such as the US. In 2025, the Centres for Disease Control and Prevention (CDC) recorded the greatest epidemic of measles in 30 years, with over 1,000 cases documented in 31 states. Interestingly, 96% of these instances included people whose vaccination status was uncertain or who were not immunized [4].

In the etiology of measles, the virus replicates first in the nasopharynx and local lymph nodes before spreading throughout the body through the circulation. Due to the virus's ability to inhibit host immune responses, subsequent infections may become more likely to occur. Although laboratory confirmation via PCR testing or serologic testing for measles-specific IgM antibodies is usual in the United States, the diagnosis is mostly clinical in settings with inadequate resources [5]. There is no particular antiviral cure for measles; instead, supportive care, vitamin A supplementation, hydration, and, if required, treatment of subsequent bacterial infections are the mainstays of treatment [6].

Strong surveillance data is essential for healthcare systems in order to identify at-risk groups and create focused interventions in light of growing concerns about vaccine reluctance and declining coverage. However, it can be challenging to identify basic trends in vaccination coverage employing classic epidemiological approaches alone since states differ in how data are gathered and presented [7]. Approaches from artificial intelligence (AI) and machine learning (ML) provide strong substitutes in this situation. Specifically, without the need for pre-established groups, algorithms for unsupervised learning like K-means clustering can group regions according to common characteristics like vaccination rates, size of population, or survey coverage. These methods have been effective in areas including syndromic surveillance, illness mapping, and medical allocation of resources [8].

Despite the increasing use of ML in healthcare, few researchers have used unsupervised clustering techniques to investigate trends over time in MMR vaccination coverage in US states. A significant portion of the current literature focuses on prediction models of outbreaks or vaccination uptake instead of discovering structural trends in surveillance data. This is a crucial study gap, especially in determining how population patterns and reporting behaviors influence observed prevalence rates and possible vulnerability to epidemics. To ensure uniformity, significance, and statistical validity, the subsequent inclusion criteria were used during choosing the data and preparation for the study.

- To discover long-term developments and trends, the data must be incorporated from 2009 to 2024, a 15-year time range.
- The clustering approach involved only US states and authorities with continuous yearly MMR vaccination data.
- Jurisdictions that lacked over twenty percent of data points during the study timeframe were eliminated to ensure dataset integrity.
- The study concentrated on childhood, as they are an important indicator population for monitoring adherence to school-entry vaccination regulations.

2. LITERATURE REVIEW

Recently released statistics from the CDC revealed a disturbing rise in measles in the US. As of May 15, 2025, 1,024 confirmed measles cases had been reported in 31 regions, including big states like California, Texas, Florida, and New York. This is the biggest epidemic in many years, with Fourteen multi-case breakouts identified and 92% of cases connected to outbreak clusters. Comparatively, in 2024, there were only 285 confirmed cases, with 69% related with epidemics, indicating a significant rise both in case number and outbreak intensity [9]. While the 2025 data revealed significant risks, with 96% of infected people being unvaccinated or having an uncertain vaccination status, these cases were particularly focused in areas with a high vaccine hesitancy or refusal, indicating a strong link between vaccine denial and measles incidence.

Children under the age of five made up 30% of incidents and recorded the highest hospitalization rate (23%), and three verified measles-related deaths highlighted the disease's continued lethality [10]. In addition to highlighting the effects of decreased MMR vaccination coverage, under-reporting, and inconsistent implementation of immunization regulations, these results show a significant immunity gap within the population. Additionally, they reveal systemic irregularities in vaccination monitoring and imply that present predictive techniques and outbreak response procedures are unable to stop the spread of disease at a rapid pace [11].

Recent research examined exploring the implementation of ML to enhance monitoring and predicting of the outbreak of measles in the US. A clear example is study by [8] “Predicting Measles Outbreaks in the United States: Evaluation of Machine Learning Approaches" which provides significant insights into county-level epidemic prediction applying supervised machine learning techniques such as XGBoost and logistic regression. Their findings showed that XGBoost exceeded standard logistic regression in the detection of counties at the highest risk of measles incidents, with AUROC values greater than 0.92. They also created a hybrid approach that integrated attributes from unsupervised clustering techniques (HDBSCAN and unsupervised random forest), which enhanced model performance significantly on imbalanced datasets [8]. While their study advances considerably to the field by focusing on outbreak prediction, it is fundamentally event-driven and retrospective, with a concentration on single outbreak incidences instead of future structural trends in vaccination monitoring or prevalence trends. Furthermore, the clustering methodologies used were supplemental and not utilized to comprehend bigger demographic patterns in the United States throughout time.

Furthermore, new developments in surveillance studies have explored atypical epidemiological approaches for improving early measles detection. Chen and Bibby (2025) explored the possibility of adopting wastewater-based epidemiology (WBE) as an innovative approach for early measles outbreak detection. Their model-based research measured measles RNA dispersion and established the spatial and temporal criteria required for wastewater identification [12]. Though WBE is an outstanding surveillance technique, its application is essentially reactive, applicable only in high-incidence contexts, and presently limited by technology constraints. Furthermore, WBE fails to identify trends in vaccination behaviour or vaccination efficiency, nor does it classify people based on fundamental coverage stages, which could indicate possible weaknesses.

Similarly, in order to improve epidemic prevention, Kujawski et al. (2024) created an ML approach to forecast national measles incidence in the US. The predictive model was developed on historical data from 2014 and 2018 and verified with 2019 measles incident records, which included 17 predictor factors. Their system outperformed expectations, with 95% specificity for areas with no incidents and 72% sensitivity for areas that experienced measles outbreaks. Incredibly, the algorithm detected 73% of the top 30 vulnerable countries and 94% of all instances of measles recorded in 2019 [13]. While this technique emphasizes the effectiveness of supervised learning for epidemic prediction, it is

strongly reliant on historical occurrence data and focuses on immediate threat estimation. Furthermore, the study prioritizes prediction accuracy over interpretability or the detection of long-term structural trends in vaccination coverage or monitoring quality. As such, it provides a reactive tool for detecting outbreaks but does not explain why particular states or counties have consistently low vaccination effectiveness over time. Therefore, our study contributes to this existing predictive framework by using unsupervised K-means clustering on a 15-year collection (2009-2024) of MMR vaccination coverage amongst US kindergarteners. Rather than predicting future outbreaks, we want to uncover hidden epidemiological trends in vaccination behaviour as well as reporting procedures among states. This structural understanding is crucial for developing preventive measures such as focused outreach, improved public health coverage, and region-specific vaccination campaigns.

Figure 1 provides the visualization of the major keywords used in Scopus streamlined for serching the relevant studies. These keywords are measles, the United States and prediction. Precisely forty-two current and relevant documents were found, which were subsequently applied in the bibliometric network visualization using the approach as adopted from Van and Waltman's VOS viewer analysis [14].

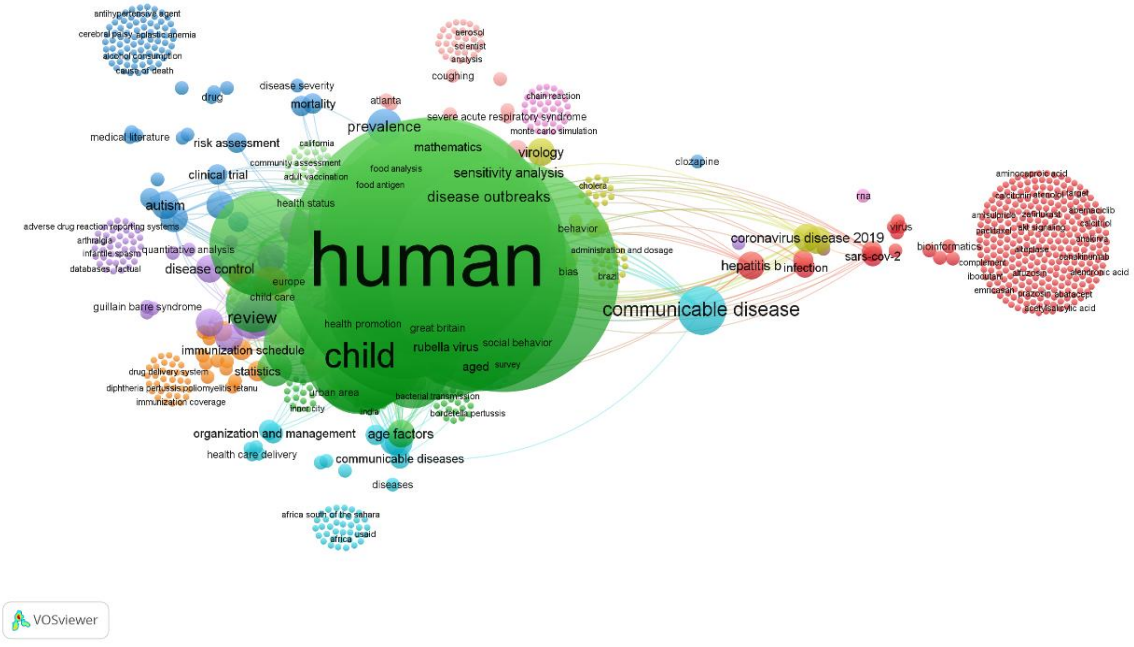


Figure 1: Bibliometric network visualization

Additionally, during active epidemics, recent studies have highlighted the importance of promptly assessing vaccination coverage and immunity gaps. A unique statistical approach, VaxEstim, was established by Martoma et al. (2024) in their descriptive investigation of the 2022 Central Ohio measles outbreak to determine vaccination coverage and population vulnerability in real time. With a predicted resistance gap of 42% and based on restored daily impact and serial interval presumptions, their model calculated that the target population's early-phase vaccination coverage was just 53% (95% credible interval: 21%–77%), considerably lower than the 93% critical vaccination threshold [15]. This study

demonstrates how decades of U.S. attempts to eradicate measles can be jeopardized by unreported vaccination gaps, particularly in areas with poor vaccination databases. It also highlights how crucial localized, real-time modelling is for directing public health actions during epidemics. Although VaxEstim works well in outbreak situations, it is based on high-incidence, retrospective data and ignores systemic under-vaccination over time or between states, as well as long-term patterns of surveillance variability.

3. METHODOLOGY

3.1 Study workflow

The study employs a systematic approach that comprises data collection, preprocessing, building [16] our unsupervised model, K-means clustering, cluster assessment, and interpretation. The research uses 15-year datasets (2009-2024) to find hidden trends in MMR vaccination coverage across states in the United States as presented on Figure 2.

This study's methodology for discovering epidemiological trends includes a systematic series of analytical phases. It starts with data collection, in which MMR vaccination data is obtained from the CDC. This is then followed by data exploration, which includes analyzing feature distributions and relationships to identify underlying patterns. The next phase, data preprocessing, involves cleaning, dividing, and correcting class imbalances to assure model ready. The model architecture is then developed using the K-means algorithm, and important features like % surveyed and population size are chosen. The model is cluster validated using visual approaches such as pair plots and boxplots to verify meaningful classifications. Lastly, the results are analyzed in the interpretation phase to generate findings that can be used to guide public health policies and decision-making.

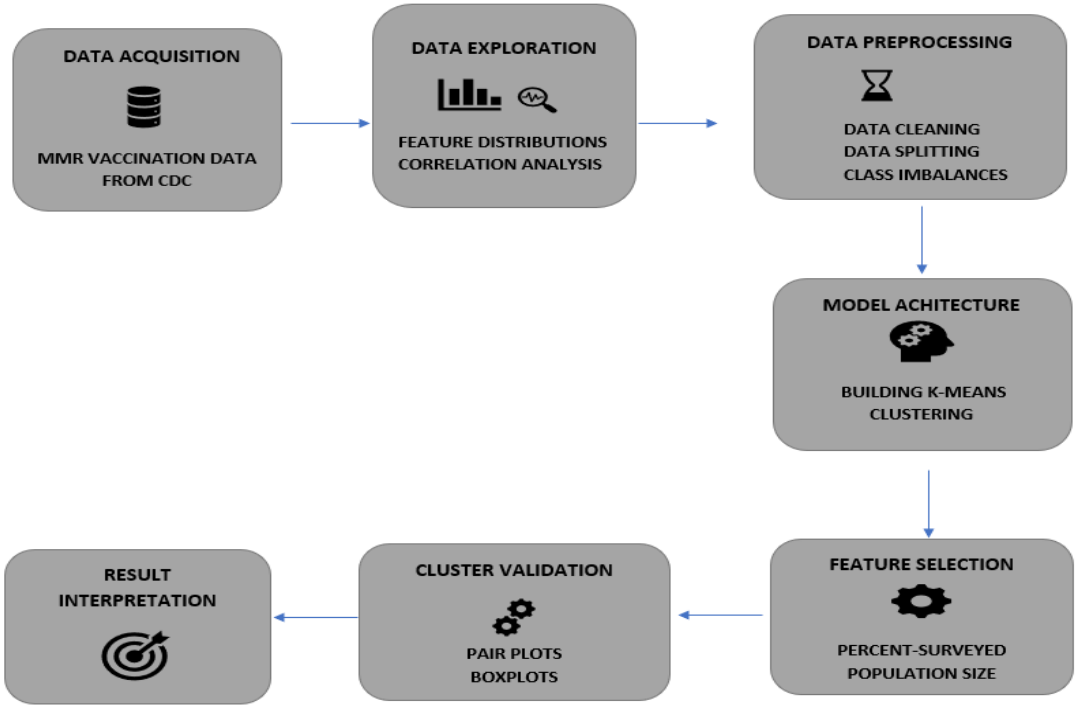


Figure 2: Workflow of the proposed study

3.2 Dataset Description

The dataset includes yearly state MMR vaccination history for kindergarten pupils across the United States from 2009 to 2024. The data, which comes from the CDC and other validated immunization registries, contains important demographical and epidemiological factors. With characteristics like the population size, survey of the number of kindergarteners, percent surveyed method of analysis, percentage of MMR vaccination coverage, and an average coverage group based on CDC thresholds, which each row indicates a distinct data point for a U.S. state or jurisdiction in a particular year. Initially, there were 769 occurrences in the dataset. The total number of suitable samples were 765, however, because states with more than 20% missing data over the course of the 15-year period were eliminated to guarantee data consistency and quality.

3.3 Data Exploration

Figure 3 shows the trend in mean MMR vaccination coverage amongst kindergarteners in the United States from 2009-10 to 2023-24. Hence, between 2009 and 2020, the trend-maintained stability, consistently above 93.5%, showing continued high levels of immunization. Notably, the highest mean vaccination rate was achieved during the 2019-20 school year, shortly prior to the COVID-19 pandemic.

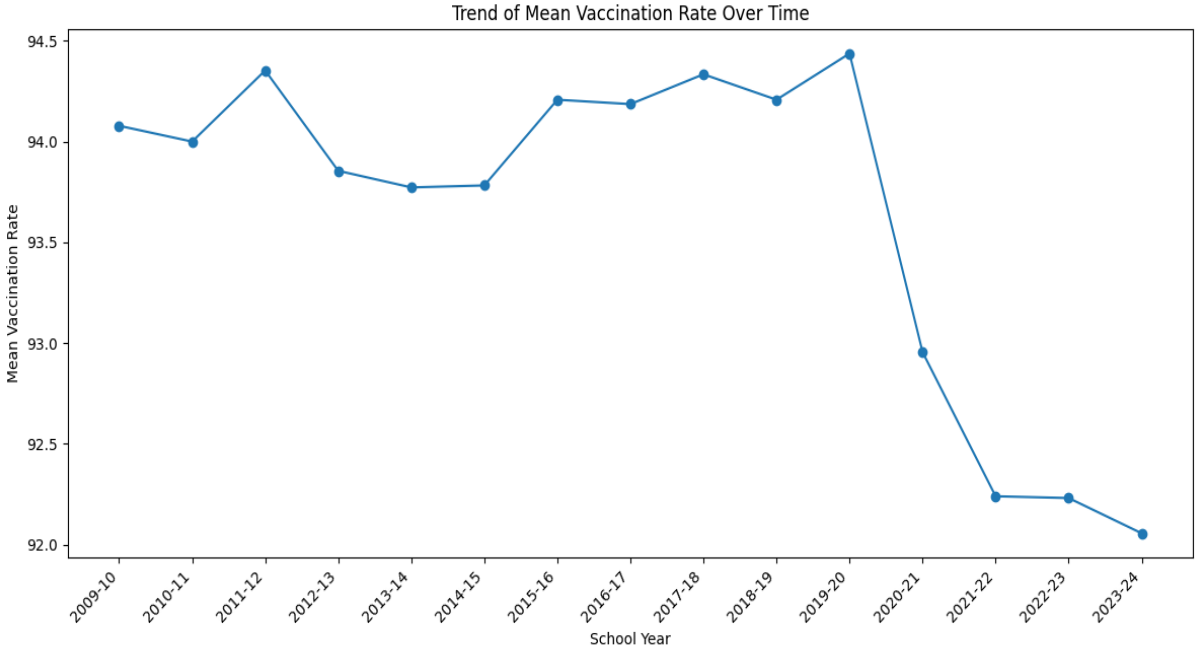


Figure 3: Trends of mean vaccination over time

However, beginning in 2020-21, there was a substantial fall in mean vaccination rates. This large reduction, which will dip below 93% and finally approach 92% by the 2023-24 school year, correlates with the interruption of healthcare and educational activities caused by the Covid-19 pandemic. The pandemic most certainly contributed to decreased access to routine immunizations, delayed medical appointments, and increased vaccine reluctance.

3.4 Data Preprocessing

To guarantee the dependability and efficiency of the ML pipeline, data preparation is an essential stage [17]. Preprocessing in this study included a number of steps designed to get the MMR vaccination dataset ready for clustering analysis. Initially, data cleaning was done to fix typos, missing numbers, and inconsistencies. To maintain the dataset's integrity, states having more than 20% missing data during the 15-year period were completely eliminated. Depending on the kind of parameter, relevant imputation methods, such as mean replacement or forward-fill, were used to handle missing values for the remaining data. Subsequent data transformation procedures were carried out, such as normalizing continuous variables (such as vaccination coverage and population size) to place them on a similar scale, which is necessary for distance-based methods like K-means. To make sure that the framework could read categorical variables like survey technique and coverage category correctly, one-hot encoding was used to encode them. In order to find any skewed distribution that would compromise clustering quality, the dataset was further analyzed for class imbalances, particularly in the coverage group feature. In order to maintain chronological consistency, the data was finally divided into different time periods, allowing for both retrospective and prospective trend analysis. Additionally, outliers were evaluated, and if they were shown to substantially alter feature distributions, they were either capped or eliminated. These pretreatment procedures made that the dataset was clean, well-organized, and ready for reliable clustering and trend identification in the study's later stages.

3.4.1 Feature selection

To determine which variables were most pertinent for clustering analysis, feature selection was performed after preprocessing [18]. The goal was to preserve characteristics that accurately depicted the differences in immunization surveillance methods and epidemiological trends between states. Based on their importance and accessibility across all observations, three essential qualities were chosen. These included: population size, which represents the number of kindergarteners surveyed in each state each year; vaccination coverage category, a discretized measure that groups states based on CDC-defined levels for herd immunity and percent surveyed, which represents the percentage of the eligible kindergarten population included in the immunization data. The K-means algorithm's significant proximity calculation was feasible by numerically encoding the vaccination coverage group, which was initially ordinal. This is because they incorporate variations in state immunization performance, data completeness, and demographic scale, all crucial for identifying unique epidemiological clusters, these features were selected. Since dimensional clarity and minimizing redundancy are essential for interpretability in unsupervised learning designs, additional factors, such as the survey approach, were examined but left out of the final clustering framework.

3.5 Model Architecture

The K-means clustering approach, an unsupervised ML method, was used in this study to find trends and categories in the collected data that reflect correlations in vaccination monitoring and coverage parameters. K-means was chosen for its ease of use, efficiency, and efficacy in dealing with structured numerical datasets. The model's goal is to classify the dataset into, non-overlapping clusters, containing each data point assigned to the cluster with the corresponding mean significance, known as the centroid. To guarantee that the algorithm operated effectively, all selected features, including population size, % surveyed, and vaccine coverage group, were standardized with StandardScaler. This change guaranteed that each attribute contributes equally to the distance computation performed by K-

means, which is dependent to feature size. The number of clusters (K) was empirically determined using two commonly used approaches: the elbow method, which shows the within-cluster sum of squares (WCSS) to identify the area with low returns, and silhouette analysis, which gauges the similarity each point is to its own cluster in comparison with other clusters.

3.5.1 Implementation Tools

The K-means clustering methodology was implemented on a powerful computing system to ensure that the dataset was processed smoothly and efficiently. The system was setup with Windows 10 Pro (64-bit), operating on an 11th Gen Intel® Core™ i7-11700KF processor @ 3.60 GHz, equipped with 64 GB RAM and 1 TB storage. An NVIDIA GeForce RTX 3070 GPU was employed for more advanced computing and visual functions similar to the study by [19]. The primary programming platform used for model building and analysis was Python, including important libraries such as scikit-learn for ML, pandas and NumPy for data manipulation, and Matplotlib and Seaborn for visualization. This robust system allowed for smooth preprocessing, modelling, and interpretation of the enormous multi-year immunization dataset. Table 1 presents the implementation tools required in the study design and system.

Table 1: Implementation tools and frameworks

s/n	Implementation tools	Settings
1	Windows	Windows 10 pro
	Frameworks/libraries	Matplotlib, seaborn, NumPy, pandas, sklearn
2	Operating system	64-bit operating system* 64-based processor
3	Processor	11th Gen Intel ® Core ™ 17-11 700 KF@ 3.60 G Hz
4	Programming language	python
5	Graphics processing unit (GPU)	Nvidia GeForce RT*3070
6	Random access memory (RAM)	64.0 GB
7	Storage space	1 TB*1

4. RESULTS AND DISCUSSION

To assess the clustering model's efficacy and ability to disclose important epidemiological trends, we performed a comparative study of cluster characteristics using key criteria such as population size and percentage surveyed. These measures shed light on both the demographic setting and the dependability of MMR vaccination data across states in the US.

Figure 4 demonstrates the Elbow technique, which was used to estimate the ideal number of clusters (k) for K-means clustering in MMR vaccination coverage trends across USA from 2009 to 2024. The x-axis portrays various values of number of clusters, whereas the y-axis indicates the inertia, or within-cluster total of squares, which measures the closely the clusters are categorized. Clusters decrease and more suited to the data as the number of clusters rises, hence the inertia falls. More clusters, nevertheless, barely help the situation after a certain point. The "elbow" of the curve, which is shown

at k = 3, indicates that three clusters is the ideal ratio of overfitting to underfitting because it indicates the point at which the rate of decline abruptly levels off.

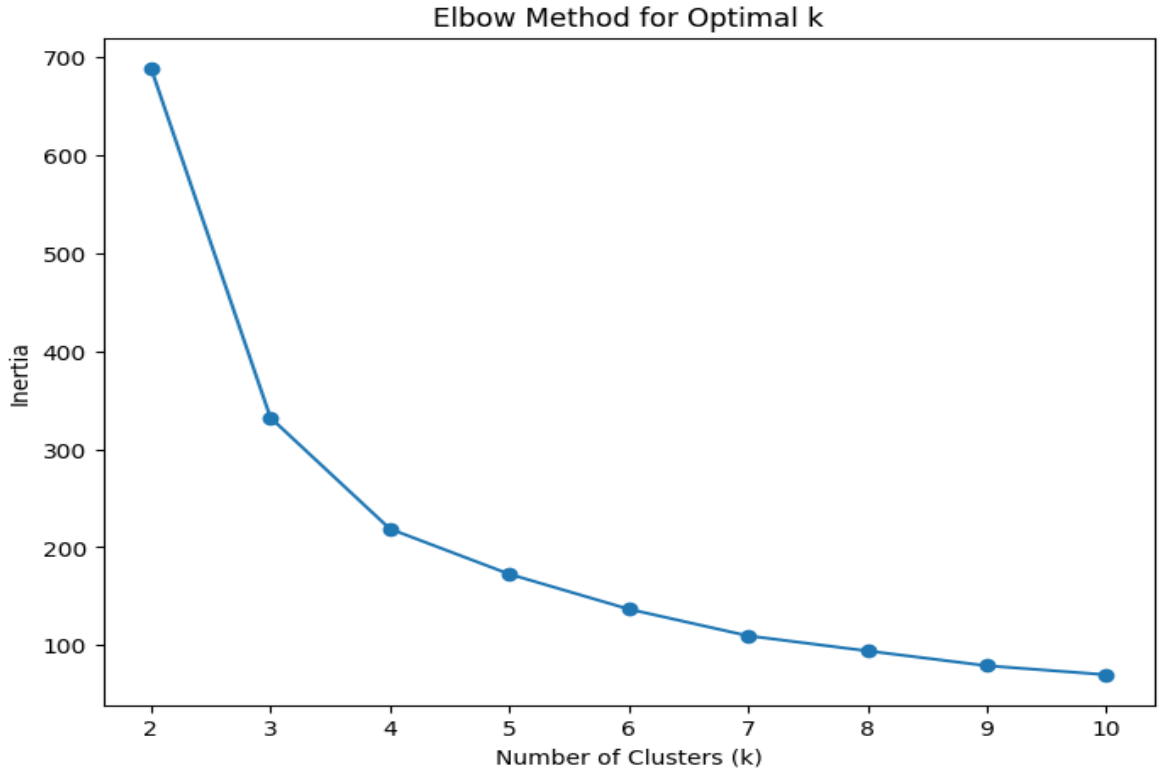


Figure 4: Elbow method for Optimal k

This affirms the study's methodological technique for splitting states into three epidemiological clusters, which effectively capture diverse behavioural trends in vaccination effectiveness and population surveillance. This selection facilitates clearer understanding and policy targeting, emphasizing the value of an unsupervised machine learning approach in health-related decision-making.

Figure 5 above displays the Silhouette Analysis utilized to estimate the ideal number of clusters for the K-means clustering method applied to MMR vaccine coverage information across USA. The x-axis reveals the k, while the y-axis indicates the silhouette outcome, which determines the similarity of a thing to its own cluster in comparison to other clusters. An increased silhouette score suggests better-defined and better clusters. As shown in the graph, the silhouette rating increases dramatically at k = 3, indicating that clusters are well separated and cohere at this value.

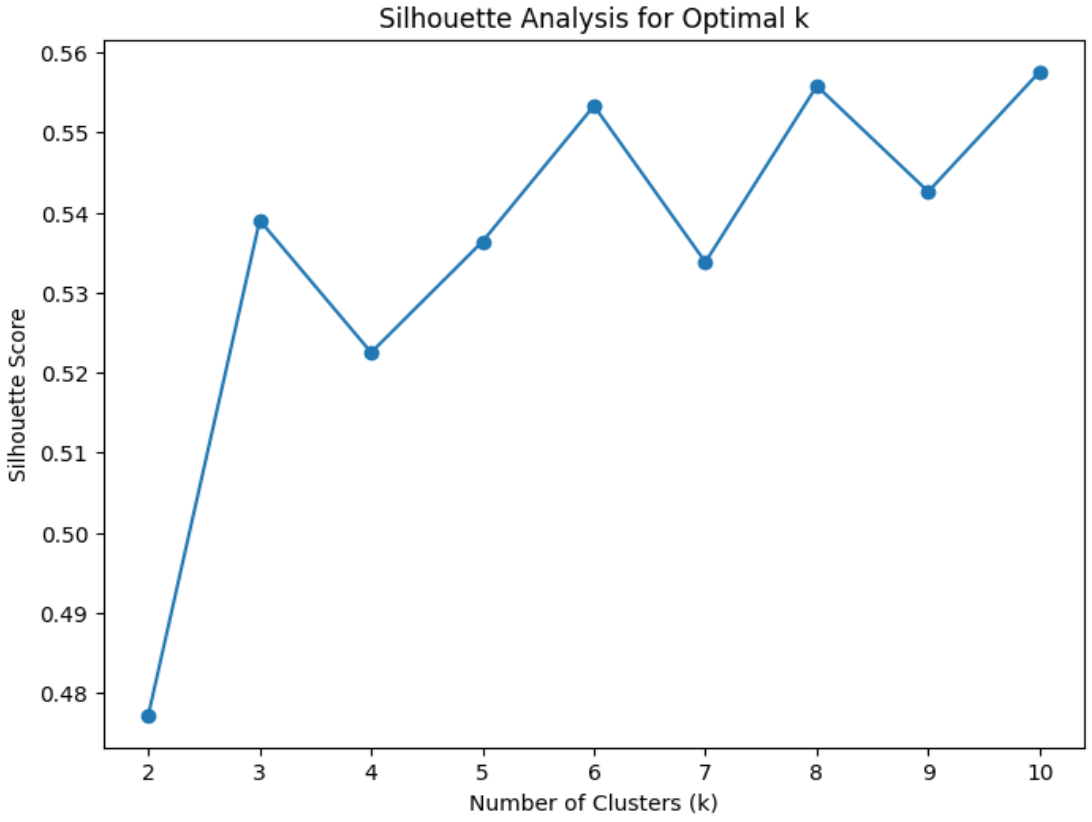


Figure 5: Silhouette Analysis for Optimal K

While the rating continues to grow slightly above k = 3, the enhancement is minor and may result in unneeded complexity. As a result, k = 3 strikes a suitable balance in model simplicity and cluster stability. This result supports the Elbow Method's results and supports the study's use of three clusters, allowing for significant categorization of states according to vaccination data features and boosting interpretability for health policy.

The violin plots in Figure 6 show the pattern of population size and percentage surveyed among the three selected clusters. Cluster 0 is mostly made up of states with lower populations and great survey coverage, often near or at 100%. Cluster 1 has the states with the highest demographics and regularly high percentage surveyed numbers, signifying adequate regions with strong monitoring infrastructure. Cluster 2, on the other hand, has a wider range of population sizes but is distinguished by a substantially lower percentage polled, exposing potential flaws in data collection or public health participation. These visual insights support the cluster disparities reported by the K-means model, emphasizing disparities in both demographic size and monitoring efficacy among USA.

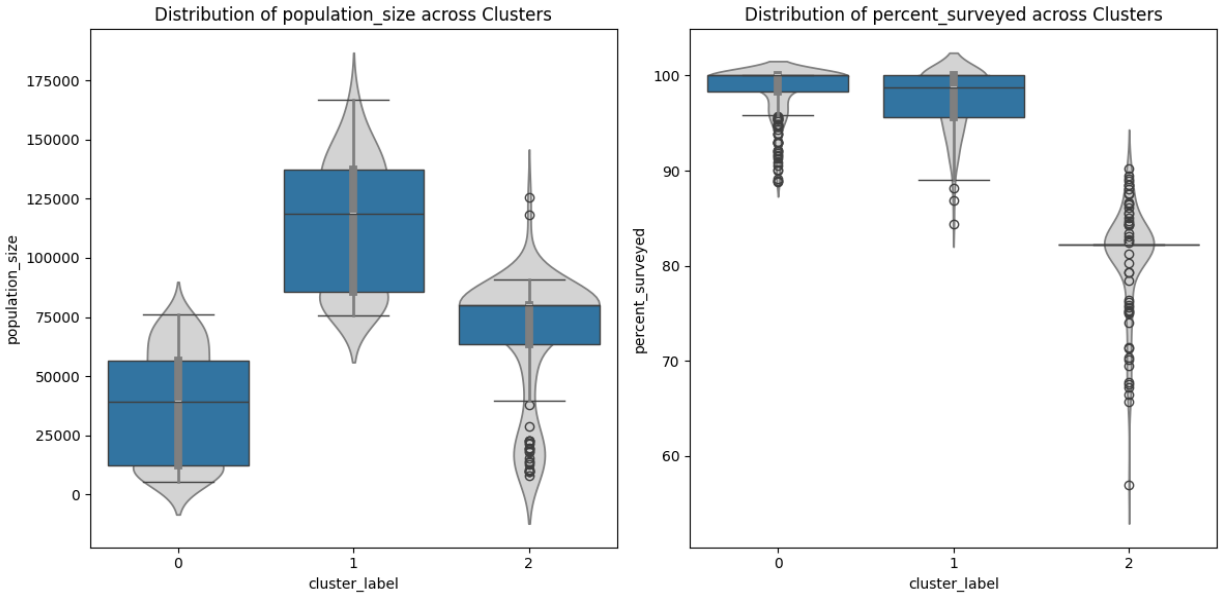


Figure 6: Distribution of population size and percent surveyed

The pair plot depicts the correlations connecting population size and percent surveyed in the three K-means clusters, emphasizing unique epidemiological groups. Cluster 0 (light pink) has a concentrated distribution, with consistently high survey coverage and moderate population sizes. Cluster 1 (purple) includes states with larger populations and correspondingly high percent surveyed rates, indicating a robust public health system. Cluster 2 (dark violet) shows more dispersed population sizes and significantly lower survey coverage, indicating possible underreporting or resource restrictions. The density charts along the diagonal emphasize these disparities in variable distribution, demonstrating the clustering structure's durability as in Figure 7.

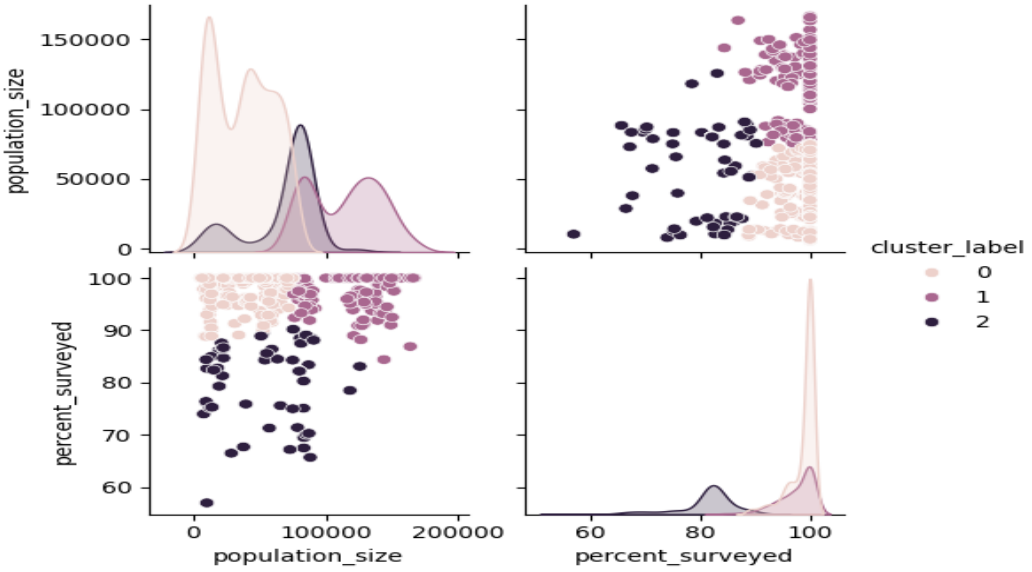


Figure 7: Correlations of population size and percent surveyed

The present study employed K-means clustering to discern latent trends in MMR vaccination coverage among kindergarteners across the United States from 2009 to 2024. This unsupervised learning technique yielded three distinct clusters, each characterized by a unique epidemiological profile. Cluster 0 encompassed states with modest population sizes yet consistently demonstrated exemplary survey coverage. Cluster 1 comprised high-population states exhibiting comparably robust coverage, indicative of well-established public health reporting mechanisms. In contrast, Cluster 2 presented a broader spectrum of population sizes coupled with diminished survey percentages, signifying systemic deficiencies in immunization data collection or access to healthcare facilities.

Statistical analysis elucidated a persistent decline in mean vaccination rates commencing around the 2020-2021 academic year, likely attributable to the COVID-19 pandemic's disruption of routine immunization initiatives ass reported by [20]. This trend aligns with CDC reports indicating a surge in measles outbreaks and a concomitant decrease in vaccination coverage post-2020. Recent investigations, including those conducted by Kujawski et al. (2024), have underscored the spatial and structural variances in measles risk prediction, thereby reinforcing the clustering findings. In a separate study, Ru et al. (2023) employed both supervised machine learning methodologies (XGBoost, logistic regression) and unsupervised clustering techniques (HDBSCAN, uRF) to forecast county-level outbreaks. Although their model yielded commendable performance metrics (AUROC = 0.926), its primary focus remained on prediction rather than the analysis of systemic coverage deficiencies.

Our approach builds upon prior research by facilitating a more nuanced epidemiological classification of states, which can inform targeted upstream intervention strategies. Moreover, our study represents one of the earliest applications of K-means clustering utilizing an unsupervised machine learning methodology in relation to MMR vaccination data over an extended temporal framework. This innovative technique yields epidemiological categories without necessitating incident labels, thereby rendering it applicable for real-time vaccination policy formulation. This analysis, encompassing the years 2009 to 2024, integrates fluctuations from both pre- and post-pandemic periods, enabling policymakers to evaluate systemic disruptions and discern long-term ramifications within survey and vaccination infrastructure. In contrast to antecedent models that employed supervised learning to project epidemic probabilities, our study adopts a data-driven, exploratory methodology to delineate state segments and pinpoint those necessitating specific interventions.

5. Study Limitation

Despite its commendable attributes, the study presents several notable shortcomings. Consequently, the dataset primarily focused on kindergarteners, thereby neglecting other vulnerable demographics such as toddlers and adolescents, which could provide a more holistic understanding of epidemiological trends. Moreover, the clustering methodology considered only two factors: population size and the percentage surveyed, which, while significant, fail to encompass relevant sociodemographic and policy variables. Additionally, considerable data discrepancies or underreporting in certain states may compromise the precision and interpretation of cluster classifications. Lastly, the framework does not explicitly integrate temporal dynamics within the clustering process, thereby constraining insights into the evolving state-level patterns over time.

6. Conclusion

This study employed K-means clustering to elucidate patterns in MMR vaccination coverage across US states and regions over a 15-year period (2009-2024). Utilizing publicly accessible data from the CDC, the framework unveiled three distinct clusters that encapsulated significant transformations in

population size, survey participation rates, and overall data integrity. The findings underscore regional and systemic disparities in immunization infrastructures, often obscured by national averages. Among the notable discoveries from this investigation is the pronounced decline in average MMR vaccine coverage attributed to the onset of the COVID-19 pandemic, suggesting that routine pediatric immunization initiatives may have been compromised due to pandemic-induced disruptions, similar to the report by (*COVID-19 Pandemic Leads to Major Backsliding on Childhood Vaccinations, New WHO, UNICEF Data Shows*, n.d.). The cluster-based classification further enabled the identification of states that are particularly susceptible to outbreaks due to inadequate coverage and survey participation, which serve as critical precursors.

The methodological strengths of the study arise from its clarity, interpretability, and reliance on readily accessible components, rendering it exceptionally suitable for prompt implementation by both local and national health authorities. The integration of various validation techniques, including elbow and silhouette analysis, bolsters the credibility of the formed clusters, ensuring that the findings are robust and actionable. In light of the CDC's 2025 forecast of a significant increase in measles cases, with over one thousand reported instances across 31 countries, our results are both timely and imperative. They furnish a framework for proactive risk assessment, which, when coupled with predictive models and targeted surveillance, can substantially enhance the United States' capacity to maintain its measles eradication status.

Finally, this work elucidates the significance of unsupervised machine learning as an invaluable adjunct in public health research. By uncovering structural discrepancies in vaccine coverage, it facilitates the implementation of targeted immunization initiatives, enhances data reporting methodologies, and promotes equitable medical policies. Future inquiries may build upon this foundational work by integrating socio-demographic characteristics, spatial forecasting, and cross-validations with outbreak data to augment the precision and applicability of the clustering methodology.

Declarations

Author contributions

Conceptualization, B.B.D; Methodology, B.B.D; software, B.B.D and H.U; Analysis, B.B.D and C.I.; data curation, B.B.D.; writing—original draft preparation, B.B.D and M.S.; writing—review and editing, B.B.D and M.S.; visualization, B.B.D;C.I; M.S; and H.U; supervision, M.S. and H.U.; project administration, B.B.D and M.S. The authors have read and agreed to the published version of the manuscript.

Funding

This project received no funding

Competing interests

There are no competing interests.

Data availability statement:

The dataset used in this study is publicly available online https://www.cdc.gov/measles/data-research/index.html

List of Acronyms

1. **AI** – Artificial Intelligence
2. **ANN** – Artificial Neural Network
3. **AUROC** – Area Under the Receiver Operating Characteristic Curve

4. **ANFIS** – Adaptive Neuro-Fuzzy Inference System
5. **CDC** – Centers for Disease Control and Prevention
6. **GPU** – Graphics Processing Unit
7. **MLR** – Multiple Linear Regression
8. **MMR** – Measles, Mumps, and Rubella
9. **ML** – Machine Learning
10. **PCR** – Polymerase Chain Reaction
11. **RAM** – Random Access Memory
12. **TB** – Terabyte (used in the context of storage)
13. **WBE** – Wastewater-Based Epidemiology
14. **WCSS** – Within-Cluster Sum of Squares

References

[1] Emmelot, M. E., Bodewes, R., Maissan, C., Vos, M., de Swart, R. L., van Els, C. A. C. M., & Kaaijk, P. (2025). Impact of genotypic variability of measles virus T-cell epitopes on vaccine-induced T-cell immunity. *NPJ Vaccines*, 10(1), 36.

[2] Martell, R., Reade, M., Boesch, L., Kaur, D. P., Kumar, S., McArthur, M., & Maar, M. A. (2025). The role of narratives in promoting vaccine confidence among Indigenous peoples in Canada, the United States, Australia, and New Zealand: a scoping review. *International Journal for Equity in Health*, 24(1), 63.

[3] Global measles threat continues to grow as another year passes with millions of children unvaccinated | CDC Online Newsroom | CDC.

[4] MEASLES CASES SURGE WORLDWIDE, INFECTING 10.3 MILLION PEOPLE IN 2023. (2025). *Neurosciences (Riyadh, Saudi Arabia)*, 30(1), 73–74.

[5] Moss, W. J., Griffin, D. E., & Feinstone, W. H. (2025). Measles. *Vaccines for Biodefense and Emerging and Neglected Diseases*, 551–565.

[6] Khan, L. (2018). Measles in children. *Pediatric Annals*, 47(9), e340–e344.

[7] Naureckas Li, C., Kaplan, S. L., Edwards, K. M., Marshall, G. S., Parker, S., & Mary Healy, C. (2025). What's Old Is New Again: Measles. *Pediatrics*.

[8] Ru, B., Kujawski, S., Afanador, N. L., Baumgartner, R., Pawaskar, M., & Das, A. (2023). Predicting Measles Outbreaks in the United States: Evaluation of Machine Learning Approaches. *JMIR Formative Research*, 7, e42832.

[9] Desai, A. N., & Majumder, M. S. (2020). What Is Herd Immunity? *JAMA*, 324(20), 2113.

[10] Tanne, J. H. (2025). Measles: Second child dies in Texas as RFK Jr finally recommends vaccination. *BMJ*, 389, r726.

[11] Holzmann, H., Hengel, H., Tenbusch, M., & Doerr, H. W. (2016). Eradication of measles: remaining challenges. *Medical Microbiology and Immunology*, 205(3), 201.

[12] Chen, W., & Bibby, K. (2025). Temporal, spatial, and methodological considerations in evaluating the viability of measles wastewater surveillance. *Science of The Total Environment*, 959, 178141.

[13] Kujawski, S. A., Ru, B., Afanador, N. L., Conway, J. H., Baumgartner, R., & Pawaskar, M. (2024). Prediction of measles cases in US counties: A machine learning approach. *Vaccine*, 42(26).

[14] van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538.

[15] Martoma, R. A., Washam, M., Martoma, J. C., Cori, A., & Majumder, M. S. (2023). Modeling vaccination coverage during the 2022 central Ohio measles outbreak: a cross-sectional study. *The Lancet Regional Health - Americas*, 23.

[16] Uzun Ozsahin, D., Duwa, B. B., Ozsahin, I., & Uzun, B. (2024). Quantitative Forecasting of Malaria Parasite Using Machine Learning Models. *Diagnostics*, 14(4), 385.

[17] Duwa, B. B., Onakpojeruo, E. P., Uzun, B., Hussain, A. J., Ozsahin, I., David, L. R., & Ozsahin, D. U. (2024). Ensemble Predictive Modeling for Dementia Diagnosis. *Proceedings - International Conference on Developments in ESystems Engineering*, 352–357.

[18] Duwa, B. B., Usanase, N., & Uzun, B. Enhanced Drug Classification for Cancers of the Liver with Multi-Criteria Decision-Making Method-PROMETHEE.

[19] Mustapha, M. T., & Ozsahin, D. U. (2025). Morphological Analysis and Subtype Detection of Acute Myeloid Leukemia in High-Resolution Blood Smears Using ConvNeXT. *AI*, 6(3), 45.

[20] COVID-19 pandemic leads to major backsliding on childhood vaccinations, new WHO, UNICEF data shows. UNICEF/WHO report.