Journal of Computer Science

Research Article

Anchoring Public Health Forecasts: High-Precision Obesity Prevalence Estimation Via Hybrid Geodemographic Ensemble Regression

Trenton Ward and Isaac Osunmakinde

Department of Computer Science, Norfolk State University, Norfolk, Virginia 23504, USA

Article history
Received: 02-06-2026
Revised: 19-06-2026
Accepted: 24-06-2026

Corresponding Author:
Isaac Osunmakinde
Department of Computer Science, Norfolk State University, Norfolk, Virginia 23504, USA
Email: ioosunmakinde@nsu.edu

Abstract: Obesity remains a major driver of chronic disease and rising healthcare costs in the United States. Yet, most existing predictive models rely on single-model classifiers and rarely incorporate spatial structure or ensemble-based regression, thereby limiting their ability to capture geodemographic variation. This study introduces a hybrid unsupervised–supervised ensemble framework that integrates K-Means spatial profiling with a Stacking Regressor to address these limitations. A geodemographic cluster feature is engineered from latitude, longitude, and population density to capture latent regional health patterns. The supervised ensemble, comprising HistGradientBoosting, Random Forest, and Multi-Layer Perceptron base learners with a RidgeCV meta learner, is trained on CDC BRFSS data cleaned to 19,078 detailed records to forecast obesity prevalence as a continuous regression target. The final ensemble achieves an R^2 of 0.8045, an accuracy of 91.95%, a correlation (r) of 0.897, and a Mean Absolute Error of 2.20%, outperforming all individual models and related literature. To evaluate robustness, the framework is stress-tested on secondary datasets, where it consistently maintains superior accuracy. These findings demonstrate the value of hybrid spatial ensemble modeling for high-precision public health forecasting.

Keywords: Ensemble Learning, Stacking Regressor, K-Means Clustering, Public Health, Obesity Prediction, Geodemographic Analysis

Introduction

Obesity continues to pose a major public health challenge in the United States (Cawley et al., 2021), with CDC estimates indicating that more than 40% of adults are affected (Emmerich et al., 2024). Beyond its clinical implications, obesity drives a substantial economic burden and exacerbates chronic disease prevalence. This motivates the need for accurate, data-driven forecasting tools to support policy planning and resource allocation. However, most existing obesity-prediction models treat obesity as a binary classification task, overlook geodemographic structure, and lack ensemble-based regression approaches capable of capturing nonlinear spatial–demographic interactions. These limitations reduce their ability to model regional variation in obesity prevalence and restrict their usefulness for state- or county-level decision-making. In particular, conventional regression estimators assume globally stationary relationships between predictors and prevalence, applying a single set of coefficients uniformly across all locations. This assumption breaks down for geographically heterogeneous outcomes, where the influence of demographic and socioeconomic factors shifts from region to region. Without an explicit mechanism to encode localized structure, standard models average across these divergent regional regimes, attenuating spatial signal and producing systematically biased estimates for areas whose profile deviates from the national mean. The unsupervised clustering stage proposed here addresses this limitation directly by partitioning records into geodemographically coherent regions before regression, allowing the supervised learners to condition predictions on local context rather than a single global trend.

To address these gaps, this study introduces a two-stage hybrid framework that first discovers latent spatial structure via K-Means clustering and then

integrates these geodemographic identifiers into a stacked ensemble regressor. The approach leverages unsupervised spatial profiling to encode neighborhood-level health patterns, followed by supervised learning using HistGradientBoosting, Random Forest, and Multi-Layer Perceptron base models combined through a RidgeCV meta-learner. This is consistent with Wolpert's stacked generalization principle (Wolpert, 1992). This design enables the model to capture both nonlinear feature interactions and spatial dependencies that traditional single-model architectures fail to represent.

The research is guided by two central questions as follows:

- How does integrating an unsupervised spatial profiling layer influence predictive precision (R^2) and Error Margins (MAE) relative to standalone regression models?
- To what extent do engineered geodemographic interaction terms act as a spatial anchor, and how does this reliance affect generalizability across heterogeneous or temporally distinct health datasets?

This study makes the following contributions:

- Hybrid Geodemographic Feature Engineering. Implementation of K-Means clustering to generate geodemographic identifiers that capture latent spatial health patterns, enabling the model to interpret regional socioeconomic and geographic influences
- Stacked Generalization for Prevalence Forecasting. Application of Wolpert's stacked generalization framework (Wolpert, 1992) using a RidgeCV meta learner to combine HistGradientBoosting, Random Forest, and MLP regressors, improving predictive stability and reducing model-specific bias
- Cross Domain Validation and Robustness Testing. Rigorous evaluation of model generalizability using secondary datasets, including the CDC Chronic Disease Indicators (CDI) and state-level demographic aggregates, to assess performance under spatial and temporal distribution shift
- Shift from Classification to High Precision Regression. Reframing obesity modeling as a continuous prevalence estimation task rather than binary classification, enabling more actionable insights for public health planning

Review of Literature

Evolution of Ensemble Learning and Stacking

The theoretical foundation of this work is grounded in Wolpert's principle of stacked generalization (Wolpert, 1992), which formalized the idea that a meta-learner can exploit the complementary biases of diverse base models to reduce generalization error. Dietterich's analysis further clarified why ensembles outperform single learners, through statistical, computational, and representational advantages (Dietterich, 2000). These contributions collectively demonstrate that ensemble architectures are inherently more robust than isolated models, particularly in high-variance domains such as public health forecasting.

However, despite the maturity of ensemble theory, most public health applications continue to rely on single-model approaches such as Logistic Regression or Support Vector Machines. These models often fail to capture nonlinear interactions between demographic, socioeconomic, and geographic variables. By contrast, this study operationalizes stacked generalization using a RidgeCV meta-learner to combine tree-based, boosting-based, and neural architectures, directly implementing the ensemble principles outlined by Wolpert and Dietterich.

Machine Learning Applications in Obesity Research

Existing obesity-prediction studies using BRFSS data demonstrate the potential of machine learning but also reveal persistent methodological gaps. Tan et al. (2025) achieved strong performance using spatiotemporal modeling but noted challenges in interpretability when regional factors interact in complex ways. Allen's work on explainable artificial intelligence similarly showed that iterative Random Forests can explain up to 79% of variance in county-level obesity prevalence (Allen, 2023), yet these models still rely heavily on static demographic inputs.

A critical limitation across these studies is the reliance on classification-based frameworks or single-model regression, which fail to provide the granular prevalence estimates required for policy planning. This study advances the field by reframing obesity modeling as a high-precision regression task and by integrating engineered geodemographic features that capture spatial context more effectively than raw coordinates alone.

Spatial Geography and Public Health Clustering

Spatial epidemiology research consistently shows that obesity is geographically patterned rather than randomly distributed. Oshan et al. (2020) demonstrated that spatial clustering reflects underlying socioeconomic and environmental determinants. Traditional spatial statistics, such as Moran's I, quantify clustering but do not integrate directly into predictive models. More fundamentally, measures such as Moran's I are descriptive global statistics in that they summarize spatial autocorrelation in a single coefficient but yield no transferable feature that a supervised learner can ingest, and they must be recomputed over the full adjacency structure whenever new observations arrive. This

dependence on a fixed spatial-weights matrix scales poorly to large, frequently updated survey datasets and cannot be embedded natively within a predictive pipeline. Traditional spatial statistics are thus well suited to confirming that clustering exists but ill-suited to operationalizing it for prediction, which motivates the data-driven K-Means encoding adopted in this study.

This study extends prior work by using unsupervised K-Means clustering to generate a dynamic "Geodemographic Cluster" feature. Unlike static spatial proxies, this engineered feature allows the ensemble to learn latent regional structure and localized health patterns, improving predictive accuracy and ranking performance during geographic stress testing.

Interpretability and Generalizability in Health Models

Modern public health modeling increasingly emphasizes transparency and out-of-sample reliability. Bzdok et al. (2018) argued that machine learning must prioritize generalizability over traditional p-value-driven inference, while Ribeiro et al. (2016) highlighted the need for explainability to ensure clinical and policy relevance. Many ensemble models, however, remain opaque and are rarely validated across heterogeneous datasets.

This study addresses these concerns by applying Permutation Feature Importance to quantify the influence of engineered spatial features and by conducting cross-domain generalizability tests using CDI and California datasets. These evaluations serve as the out-of-sample audits recommended by Bzdok et al. (2018), demonstrating that the model maintains stable MAE and ranking performance across structurally distinct datasets. Table 1 summarizes the literature review papers, including their references, paper types, problems addressed, methodologies, and future work.

Synthesis and Research Gaps

Across prior studies, three gaps persist as follows:

a. Limited use of ensemble regression for prevalence forecasting, with most models relying on single-architecture classifiers
b. Insufficient modeling of spatial–demographic interactions, where geography is treated as a static coordinate rather than a dynamic, engineered feature
c. Lack of generalizability testing across heterogeneous datasets, leaving uncertainty about model robustness under distribution shift

This study addresses these gaps by integrating unsupervised geodemographic clustering with stacked regression and validating performance across multiple external datasets, establishing a more interpretable and generalizable framework for obesity prevalence forecasting.

Table 1: Review of Related Literature

Paper Type	Problem Addressed	Methodology	Future Work Suggested
Research paper Wolpert (1992)	Minimizing overall generalization error and recognizing biases of diverse base models.	Introduced stacked generalization (Stacking) using a meta-learner.	Established the mandate to use regularized meta-learners to prevent overfitting in ensembles.
Research paper Dietterich (2000)	The limitations and risks (statistical, computational, representational) of using single models.	Analyzed why ensemble methods outperform single classifiers.	Highlighted the need to expand the space of representable functions by averaging various starting points.
Research paper Tan et al. (2025)	Understanding the spatiotemporal dynamics of obesity while managing complex regional factors.	Modeled state-level BRFSS data.	Highlighted the ongoing challenge of model interpretability when analyzing complex regional health factors.
Research paper Allen (2023)	Assessing cross-sectional U.S. county-level obesity prevalence and health risks.	Iterative random forest architectures paired with explainable artificial intelligence.	Sets a strong baseline and demonstrates the viability of tree-based models for prevalence estimation.
Perspective / Review (Bzdok et al. 2018)	The paradigm shift from traditional p-value statistics to complex machine learning.	Comparative analysis of statistical inference versus machine learning prediction.	Requires a new rigorous focus on "out-of-sample" generalizability and auditing for predictive models.
Research paper (Oshan et al. 2020)	The "Neighborhood Effect" and the non-random spatial clustering of health outcomes.	Multiscale geographically weighted regression using traditional spatial statistics (Moran's I).	Emphasized that geographic coordinates (latitude/longitude) should be treated as proxies for food deserts, urban design, and climate.
Research paper (Ribeiro et al. 2016)	The uninterpretable "black box" nature of complex classifiers and ensembles.	Introduced the concept of explainability for predictions of any classifier.	Asserts that feature dependencies must be fully transparent for a model to be actionable in clinical or policy settings.

DOI: 10.3844/jcssp.2026.2139.2155

Methods

The methodology follows a structured five-phase hybrid pipeline designed to transition from raw public health records to a fully validated ensemble forecasting system. The process begins with data preparation, where BRFSS records are cleaned, standardized, and spatially audited. Next, interaction feature engineering constructs nonlinear spatial-demographic variables such as Lat_Long_Interaction and Geo_Density. The third phase applies geodemographic clustering, using K-Means to identify latent spatial structure and assign each record a cluster-based identifier. In the fourth phase, a supervised ensemble training stage integrates HistGradientBoosting, Random Forest, and Multi-Layer Perceptron models within a stacked regression framework. Finally, a generalizability evaluation phase assesses robustness through cross-domain testing on external datasets, ensuring the model performs reliably under spatial and temporal distribution shifts. Figure 1 shows a visual breakdown of the hybrid unsupervised-supervised ensemble system model.

Stage 1: Data Loading and Preparation

The pipeline begins by ingesting the CDC Behavioral Risk Factor Surveillance System (BRFSS) dataset (… Prevention [CDC], 2024)" and "(… Prevention [CDC-2], 2024). Records missing the target variable (Data_Value) are removed, temporal markers are standardized, and geographic identifiers are normalized across all entries. After cleaning, 19,078 valid records remain from the original 110,880. This stage establishes a consistent and reliable foundation for subsequent spatial and predictive modeling.

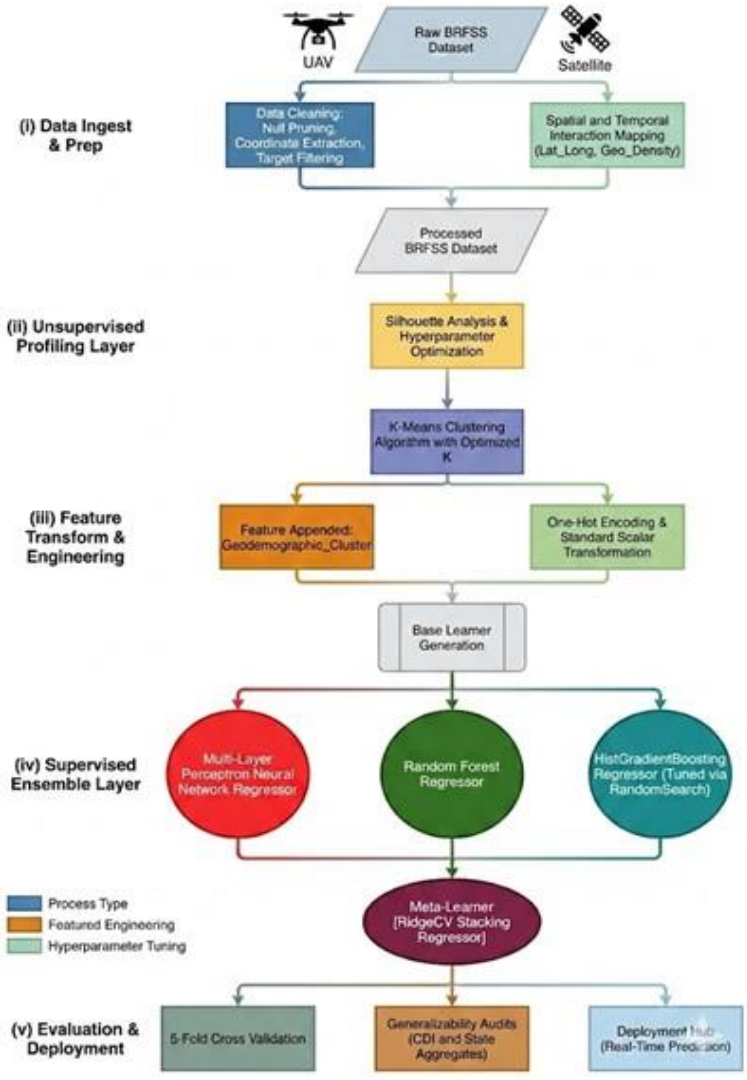

Fig. 1: Hybrid Unsupervised-Supervised Ensemble System Model

Stage 2: Exploratory Data Analysis

Before feature engineering, the dataset undergoes exploratory analysis to assess spatial and statistical patterns. Heatmaps and distribution plots are generated to visualize geographic variation in obesity prevalence and to confirm that prevalence is not uniformly distributed across the United States. These diagnostics justify the inclusion of spatial interaction terms and clustering in later stages.

Stage 3: Interaction Feature Engineering

To capture nonlinear spatial relationships, several interaction features are constructed. These include a Lat_Long_Interaction term (latitude × longitude) to encode neighborhood effects and a Geo_Density measure derived from sample size relative to geographic spread. These engineered variables enhance the model's ability to represent complex geodemographic influences. Equations (1) and (2) show the formulas for the Lat_Long_Interaction and Geo_Density, respectively:

$$X_{(\text{interaction})} = X_{lat} \cdot X_{long} \tag{1}$$

$$D_{geo} = \frac{N}{\Delta Lat \times \Delta Long} \tag{2}$$

Stage 4: Clustering Optimization

The optimal number of geodemographic clusters is determined using Silhouette Analysis across values of k from 2 to 20. For each k, the average silhouette coefficient is computed as shown in Equation (3) to evaluate cohesion and separation. The clustering process uses both demographic variables (encoded via One-Hot Encoding) and geographic coordinates. The selected k maximizes structural distinctiveness while minimizing overlap:

$$S(k) = \frac{1}{N}\sum_{i=1}^{N}\frac{b(i) - a(i)}{\max(a(i),\ b(i))} \tag{3}$$

Stage 5: Feature Engineering and Dataset Splitting

Using the optimal k, K-Means clustering assigns each record a Geodemographic_Cluster label. Categorical variables are transformed using One-Hot Encoding, and numerical features are standardized using Z-score normalization, as shown in Equation (4), to ensure balanced scaling. The dataset is then divided into an 80/20 train–test split to support unbiased model evaluation. To prevent information leakage, all clustering, scaling, feature transformations, and model fitting procedures were performed exclusively on the training split. The test set remained completely unseen until the final evaluation:

$$z = \frac{x - \mu}{\sigma} \tag{4}$$

Stage 6: Supervised Ensemble Modeling

The core predictive engine is a three-tier Stacking Regressor. Three base learners of HistGradientBoosting, Random Forest, and a Multi-Layer Perceptron are trained in parallel. After thorough testing, the HGB had an initial learning rate of 0.05, maximum iterations of 1000, and an L2 Regularization of 0.05 tuned with RandomizedSearchCV. The RF had 250 estimators, 5 minimum sample leaves, and used all CPU cores to speed up training. The NN had 2 hidden layers of 100 and 50 neurons, respectively, an initial learning rate of 0.05, and 1000 maximum iterations with early stopping enabled. A 5-Fold Cross-Validation stability report is generated to show individual performance, and a RidgeCV meta-learner is then utilized to learn the optimal weights for each base model's output, effectively leveraging the specialized strengths of both tree-based and connectionist architectures.

The final forecasting layer integrates the base learners via a stacking meta-learner. The final prediction y_Final is modeled as a regularized combination of the base estimations below, where beta_0 represents the meta-learner intercept, and beta_{1,2,3} signify the structural coefficients mapped to the HistGradientBoosting, Random Forest, and Multi-Layer Perceptron out-of-fold predictions, respectively.

Weighted Soft-Blending Optimization Equation

$$\hat{y}_{\ Ensemble} = w_1 \cdot M_{(HGB)(x)} + w_2 \cdot M_{(RF)(x)} + w_3 \cdot M_{(MLP)(x)}$$

Stacking Meta-Learner (RidgeCV) Equation

$$(\hat{y})_{Final} = \beta_0 + \beta_1(\hat{y})_{HGB} + \beta_2(\hat{y})_{RF} + \beta_3(\hat{y})_{MLP}$$

Stage 7: Comprehensive Evaluation

Results and diagnostic visualizations are generated to assess model health. This includes a full report table with *R*, *MAE*, *RMSE*, R^2, 10% Tolerance Accuracy, and Implied Accuracy scores, as well as an evaluation of the test set with an Actual vs. Predicted line comparison and a feature importance comparison for each of the model's top features. When calculating the performance metrics, the predicted obesity prevalence ($\hat{y}_i$) is compared with actual BRFSS values (y_i) in different ways.

Mean Absolute Error (MAE) measures the average magnitude of errors in predictions without taking into consideration direction. It should be relatively low in comparison to the values it is predicting to show close predictions. Equation (5) shows the formula for MAE. Root Mean Square Error (RMSE) penalizes large errors more than MAE to help measure a model's standard deviation. The RMSE should be close to the MAE but not

less than it. Equation (6) shows the formula for RMSE. The coefficient of determination (R^2) shows the proportion of variance for the target variable that can be explained by the model. It can go up to 1, representing perfect "explanatory power," but a score above 0.75 can also be considered good for variable health data. Equation (7) shows the formula for R^2. The implied accuracy measures how often the model's high-confidence prediction aligns with the actual values by taking the complement of the Mean Absolute Percentage Error (MAPE), which shows the average percentage difference between predicted and actual values. Equations (8) and (9) show the formula for implied accuracy and MAPE, respectively:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

$$A_{imp} = 100\% - MAPE \tag{8}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{9}$$

Stage 8: Deployment and Inference

A lightweight deployment interface is implemented to generate real-time prevalence estimates. The interface accepts user-provided inputs and returns predictions from each base model alongside the final ensemble output, enabling transparent comparison and decision support.

Stage 9: Baseline Generalizability Test

The CDI and California datasets differ in sampling methodology, temporal coverage, and demographic granularity. These differences introduce concept drift, making them ideal for evaluating robustness. To evaluate robustness under distribution shift, the ensemble is tested on the CDC Chronic Disease Indicators (CDI) dataset. This dataset's sparsity and differing structure reveal how well the model performs when spatial identifiers are incomplete or inconsistent.

Stage 10: Extended Generalizability Testing

A second external evaluation uses state-level aggregates to assess the model's ranking capability and sensitivity to temporal drift. This stage examines how well the ensemble maintains predictive accuracy when applied to datasets with different sampling methodologies and demographic compositions. Figure 2 shows the pipeline broken down as an algorithm in pseudocode with inputs, outputs, parameters, and steps.

Ablation Studies

A compact ablation analysis was conducted to quantify the contribution of each major component in the hybrid framework—spatial interaction features, geodemographic clustering, and stacked ensemble integration. All configurations use identical preprocessing and train–test splits to ensure fair comparison.

Configurations

a. Baseline Core BRFSS features only; no spatial interactions, clustering, or ensembling
b. + Spatial Interaction Features; Adds Lat_Long_Interaction and Geo_Density to capture nonlinear spatial effects
c. + Geodemographic Clustering; Incorporates the K-Means–derived cluster label to encode latent regional structure
d. Full Hybrid Stacked Ensemble; Complete system: Engineered features + clustering + RidgeCV stacked integration

Methodology Pseudocode

Algorithm 1: Hybrid Unsupervised-Supervised Ensemble Framework
Inputs:
- D: Primary BRFSS dataset containing spatial, temporal, and demographic records.
- F_geo: Set of geographic features (Latitude, Longitude).
- F_demo: Set of categorical demographic features (Age, Race, Education, Income).
- Y: Target vector representing obesity prevalence (Data_Value).

Parameters:
- k_range: Range for clustering optimization [2, 20].
- M_base: Set of base learners {HistGradientBoosting, Random Forest, MLP}.
- M_meta: RidgeCV meta-regressor for weight optimization.

Output:
- L_cluster: Engineered geodemographic cluster assignments.
- Y_hat_ens: Final consensus prediction vector.

Phase I: Data Ingestion and Feature Engineering
1. Preprocessing: D = {record in D | target is not null} (Prune missing records).
2. Spatial Interaction: For each record, compute Interaction = Latitude * Longitude.
3. Density Mapping: Calculate Geo_Density using localized frequency counts of coordinates.
4. Feature Set Construction: Define X_raw = {F_geo, F_demo, Interaction, Geo_Density}.

Phase II: Unsupervised Profiling Layer (Clustering Optimization)
5. Iterative Clustering:
   - For each k in k_range:
     - Execute K-Means on F_geo to partition data into k subsets.
     - Compute Silhouette Coefficient (SC) = [b(i) - a(i)] / max(a(i), b(i)).
6. Optimal Selection: Identify k_opt where SC is at its maximum value.

Phase III: Final Feature Transformation and Pipeline Alignment
7. Labeling: Assign L_cluster labels to all records using the k_opt model.
8. Encoding: Apply One-Hot Encoding (OHE) to categorical variables in F_demo.
9. Partitioning: Split the dataset into D_train (80%) and D_test (20%).
10. Normalization: Fit StandardScaler on D_train and transform both sets (Mean = 0, StDev = 1).

Phase IV: Supervised Ensemble Layer (Stacking)
11. Base Training:
    - For each model in M_base:
      - Execute 5-Fold Cross-Validation on D_train.
      - Generate out-of-fold predictions and test-set predictions.
12. Meta-Optimization:
    - Construct meta-feature matrix X_meta using predictions from base models.
    - Train M_meta (RidgeCV) to minimize the weighted squared error + L2 penalty.

Phase V: Evaluation and Generalizability Audit
13. Inference: Generate Y_hat_ens by passing base model test outputs into the trained meta-learner.
14. Quantification: Compute R-Squared, MAE, and RMSE against the ground truth.
15. Validation: Deploy the ensemble to CDI and State-level secondary datasets to audit stability.

Fig. 2: Methodology Pseudocode

DOI: 10.3844/jcssp.2026.2139.2155

Results and Discussion

The evaluation of the hybrid ensemble architecture provides critical insights into the intersection of unsupervised spatial profiling and supervised predictive modeling. The results demonstrate that incorporating geodemographic clustering significantly enhances the stability and accuracy of obesity prevalence estimations.

Pre-Modeling Exploratory Visuals

The exploratory phase in stage 2 established the necessity of a hybrid spatial approach. Four primary visualizations were generated to audit the raw BRFSS dataset (… Prevention [CDC], 2024). The first was the results for the initially processed dataset of 19,078 records. A primary audit was conducted to ensure the integrity of the spatial and temporal markers, followed by saving the final dataset to ensure the pipeline was successfully utilized. Figure 3 shows screenshots of the original dataset as well as the final dataset used for model development. The second was a geographic heatmap that mapped obesity prevalence across the United States, revealing high-density clusters in the Southeast and Midwest. This visual confirmed that obesity prevalence is not entirely randomly distributed but follows geographic "neighborhood effects," justifying the inclusion of coordinates in the clustering phase. Figure 4 shows this heatmap.

	A	B	C	D	E	F	G	H	I
1	Lat	Long	Lat_Long_I	YearStart	Sample_Si	Year_Sam	Geo_Dens	Stratificati	Stratificati
2	32.84057	-86.6319	-2845.04	2011	1367	2749037	11.441970	1	0
3	32.84057	-86.6319	-2845.04	2011	757	1522327	6.3361897	0	1
4	32.84057	-86.6319	-2845.04	2011	861	1731471	7.2066834	0	0
5	32.84057	-86.6319	-2845.04	2011	785	1578635	6.5705533	0	0
6	32.84057	-86.6319	-2845.04	2011	1125	2262375	9.4163981	0	0

	A	B	C	D	E	F	G	H	I	J	K
1	YearStart	YearEnd	LocationA	LocationD	Datasourc	Class	Topic	Question	Data_Valu	Data_Valu	Data_Valu
2	2011	2011	AL	Alabama	Behaviora	Obesity / \	Obesity / \	Percent of adults age		Value	34.8
3	2011	2011	AL	Alabama	Behaviora	Obesity / \	Obesity / \	Percent of adults age		Value	35.8
4	2011	2011	AL	Alabama	Behaviora	Obesity / \	Obesity / \	Percent of adults age		Value	32.3
5	2011	2011	AL	Alabama	Behaviora	Obesity / \	Obesity / \	Percent of adults age		Value	34.1
6	2011	2011	AL	Alabama	Behaviora	Obesity / \	Obesity / \	Percent of adults age		Value	28.8

Fig. 3: Screenshots of Dataset Before and After

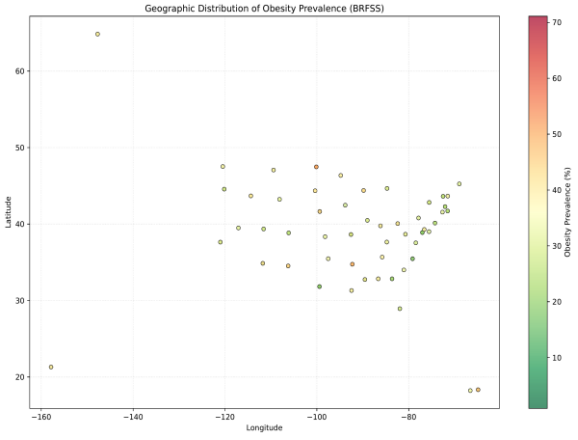

Fig. 4: Geographic Distribution of Obesity Prevalence

The third visualization showed the target distribution with a histogram of the Data_Value (Obesity Prevalence) across various demographic categories, showing a near-normal distribution centered around 30-35%. This distribution provided a baseline for the regression task, ensuring there were no significant class imbalances or extreme outliers that would skew the Mean Absolute Error (MAE). Sex had the lowest range of percentages, while Race/Ethnicity had the highest. Figure 5 shows this target distribution. The final pre-modeling plot had Year vs. Prevalence in a line plot of prevalence over the study period of 2011 to 2024. It illustrated a steady upward trajectory. This visual highlights the "concept drift" inherent in public health data, where the target variable is non-stationary, requiring the model to account for YearStart as a critical predictor. Figure 6 shows this temporal trend line plot.

Unsupervised Clustering Diagnostics

The unsupervised layer in stages 4 to 5 served to partition the high-dimensional spatial data into cohesive geodemographic regions. The results of this were shown with a silhouette score tracker that evaluated cluster counts from k = 2 to k = 20, with 20 being the ceiling to make sure that the differences between clusters do not become too minuscule, with no meaningful distinctions. The tracker showed a peak for the range at k = 20 with a score of 0.6210. This indicates that while the data could

be subdivided further, this partitioning provides the most stable and mathematically distinct grouping while minimizing overlapping noise. Figure 7 shows the tracking of the silhouette score over varying k. Table 2 shows a summary of the number of clusters tested and their average silhouette scores.

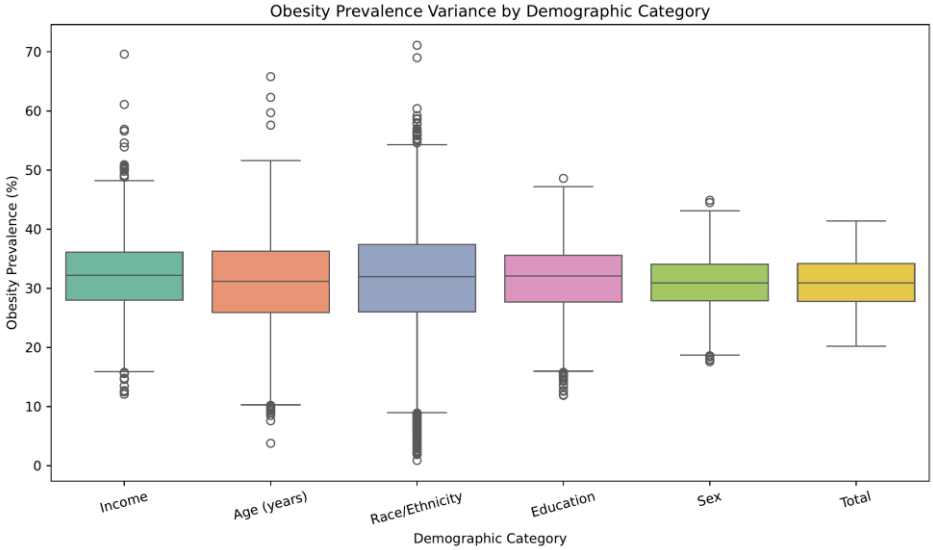

Fig. 5: Obesity Prevalence Variance

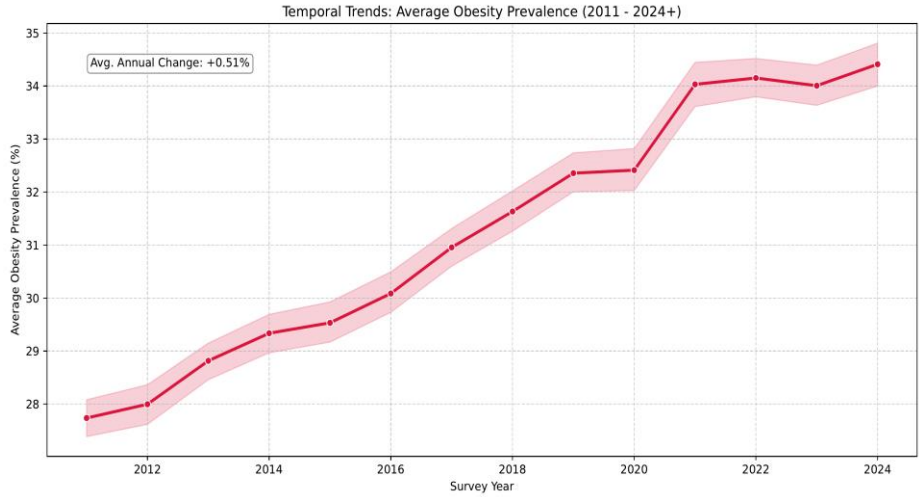

Fig. 6: Temporal Trends: Average Obesity Prevalence

Table 2: Silhouette Score Results for Spatial Optimization

Number of Clusters (k)	Average Silhouette Score	Number of Clusters (k)	Average Silhouette Score
2	0.1178	12	0.3853
3	0.1157	13	0.4173
4	0.1597	14	0.4563
5	0.2054	15	0.4757
6	0.2407	16	0.5081
7	0.2561	17	0.5346
8	0.2684	18	0.5586
9	0.3003	19	0.5928
10	0.3265	20	0.6210
11	0.3528		

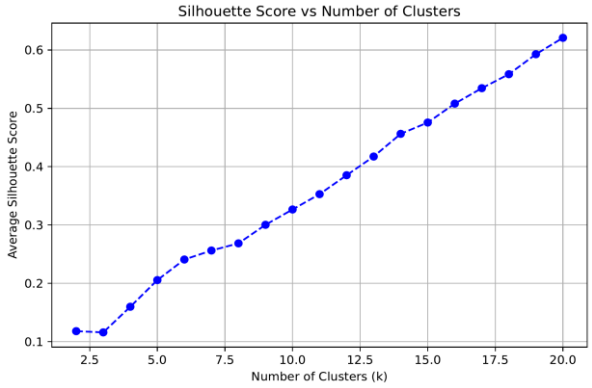

Fig. 7: Silhouette Score vs Number of Clusters

To validate the K-Means results, the spatial data was projected into a 2D plane via Principal Component Analysis (PCA) cluster projection with calculated silhouette scores. The resulting plot shows distinct, non-overlapping color-coded groups, proving that the unsupervised layer successfully captured the underlying spatial structure before passing it to the supervised ensemble. Projections were also made for the other tested number of clusters, but each additional cluster led to more and more overlap and lower silhouette scores. Figure 8 shows this PCA projection for k = 20.

Primary Supervised Model Evaluation

Before final deployment, base learners were evaluated using 5-fold cross-validation to ensure predictive stability across different data folds. Of the models, HistGradientBoosting performed the best, with Random Forest trailing slightly behind, with the neural network coming up last. Table 3 shows the CV scores achieved by each of the individual trainer models.

To validate the stability and reliability of the individual base learners within the ensemble, an analysis of the training convergence was conducted for the connectionist and gradient-boosting architectures. Figure 9 illustrates the learning trajectories for the neural network and the HistGradientBoosting model.

The NN architecture demonstrated rapid initial convergence, with the loss decreasing significantly from approximately 270 to near 10 within the first 10 epochs. The loss curve subsequently reached a stable plateau around iteration 100. This behavior indicates that the model's weights were well-initialized and the learning rate was appropriately tuned to reach a global minimum.

The HistGradientBoosting model's progress was evaluated across 200 boosting iterations. A key indicator of the model's robustness is the narrow gap maintained between the Training Score and the Validation Score. Both curves follow a nearly identical trajectory and plateau simultaneously around iteration 150. This alignment confirms that the model is generalizing effectively to the underlying health patterns in the dataset and is not suffering from overfitting, which would typically be characterized by a widening divergence between training and validation performance.

Following training, the performance of the RidgeCV Stacking Ensemble was evaluated on a 20% unseen test split, achieving an R^2 of 0.8045 and an MAE of 2.20%. These scores, along with the R, RMSE, 10% Tolerance Accuracy, which showed the percent of predictions that fall within + or – 10% of the actual value, and Implied Accuracy, which was calculated by subtracting the MAPE from 100, were reported and compared for the final ensemble and individual models.

These results showed that the final ensemble consistently outperformed the individual models by a notable amount with at the largest jumps being in the calculated accuracy scores. All of the models had over a 90% implied accuracy and around or below 2.5% MAE, showing that they can reliably predict obesity prevalence. The RMSE scores also trailed closely to the MAE, meaning there were not very many high outlier predictions, and the R-scores hit close to 0.9, showing that the models followed the trend of the actual values well. Finally, R^2 ranged from 0.7435 to 0.8045 with high explanatory power, and the 10% tolerance accuracies were above 70%, except for the NN with 69.23%. HGB, RF, and the final ensemble had above 74% for 10% tolerance accuracy. This means that a majority of predictions were, at a minimum, near misses.

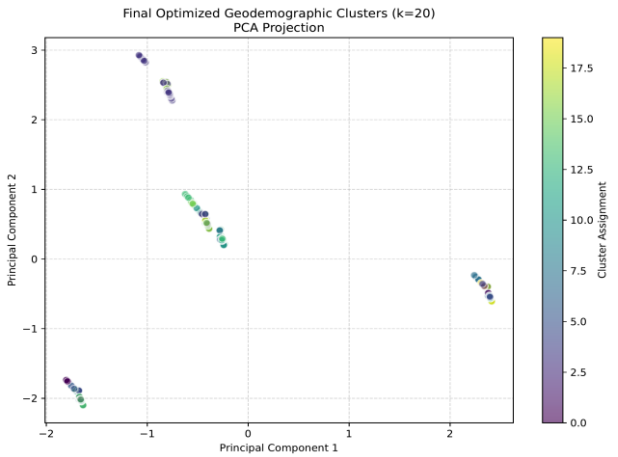

Fig. 8: PCA Projection for k = 20

Table 3: Base Model Cross-Validation Results (5-Fold)

Model Architecture	Mean CV R^2	Mean CV MAE
HistGradientBoosting	0.7905	2.30%
Random Forest	0.7764	2.41%
Neural Network (MLP)	0.7226	2.80%

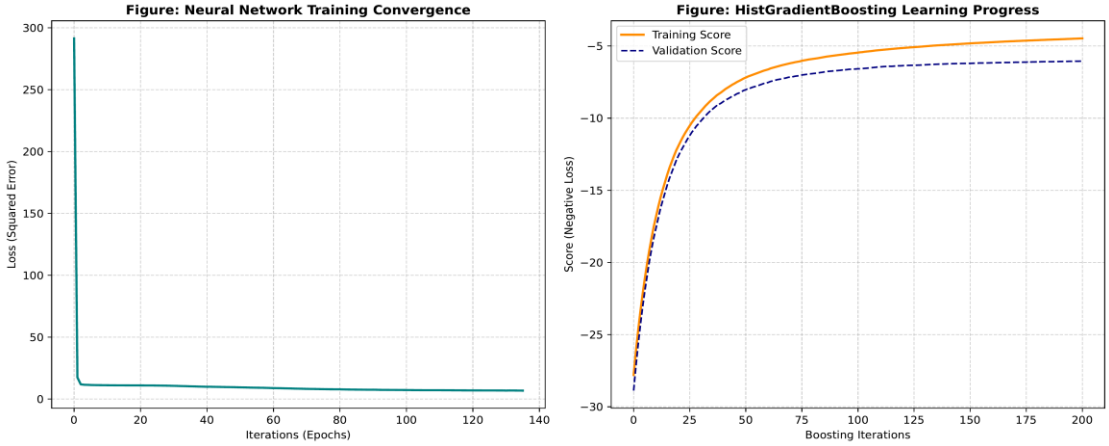

Fig. 9: Learning Convergence Analysis

The performances show that the models track the trend well, regularly predict closely to the actual values with few large outliers, and can predict a significant amount of the variance from the features. This is important for the health field as accurate data allows resources to be allocated appropriately and can save lives. The final ensemble was a standout, performing the best across all of the scores. Table 4 shows all of the reported scores for the various models.

After the models were developed, feature importance was determined through permutation importance and revealed that spatial interaction terms of Lat_Long_Interaction and Geodemographic_Cluster, as well as YearStart, were dominant drivers. This justifies the hybrid architecture as the model relies more heavily on the engineered spatial context than on raw demographic variables alone, except for the younger age range of 18 to 24 and Asian ethnicity, which were highly valued by the HistGradientBoosting, RF, and final ensemble models. In contrast, the NN had the latitude, longitude, and their interaction as its most important features with higher values in general, which may explain its differing performance. Figure 10 shows the feature importance comparison for the various models. In addition to the feature importance, the generated Actual vs. Predicted scatter plots were displayed and demonstrated a tight linear relationship between the ground truth and model estimations. The points are closely clustered around the identity line y=x, particularly in the 25% to 40% prevalence range, indicating high precision in the model's "sweet spot." As indicated by the previous results, the final ensemble model appeared just slightly more contained to the ideal identity line compared to the individual models.

When visualized as a time-series or sample-index cross-section, the ensemble's prediction line tracks the ground truth with minimal deviation. This visual of local tracking confirms that the model successfully captures both the macroscopic trends and localized fluctuations in the data. Like the scatter plots that captured the entirety of the predictions, the final ensemble adhered slightly closer to the actual values by using the learning from the 3 individual models. Figure 11 shows the Actual vs. predicted scatter plots for the various models in initial testing. Figure 12 shows the Actual vs. Predicted line comparisons for 25 samples.

The results of this model were compared to some of the previous models proposed by relevant studies. This comparison showed that this model performed close to, if not better than, previous models in variance explainability, with an R^2 of 0.8045 in comparison to 0.79 in Allen's study and in accuracy, with an implied accuracy of 91.95% compared to the 82% classification accuracy of Muhammad et al. (2025). While different datasets, targets, or metrics may have been used, these related works use the BRFSS as the dataset, obesity prevalence as the target, or both, which lets the results still be valuable for setting a baseline and seeing where there can be an improvement. Table 5 shows the model comparisons.

Deployment and Inference Hub

The system's inference hub was tested with 8 new, simulated instances to generate a consensus prediction. The new cases included instances of a young adult, a high-income earner, a college graduate, a Hispanic adult, an adult female, a baseline adult, a low-income earner, and an adult in the high-risk education group. In these cases, the neural network predictions ranged below and above the final ensemble ones, while HGB and RF remained relatively close, likely due to their influence on the final ensemble. This showed an overall trend of HGB and RF, setting the base for the final ensemble, which the NN can shift with outlier predictions.

DOI: 10.3844/jcssp.2026.2139.2155

Table 4: Comprehensive Model Evaluation

Model	r	MAE	RMSE	R^2	10% Tol.	Imp. Acc.
HGB	0.896	2.23%	3.27	0.8025	77.33%	91.78%
RF	0.886	2.34%	3.42	0.7844	74.71%	91.37%
NN	0.863	2.67%	3.73	0.7435	69.23%	90.21%
Ensemble	0.897	2.20%	3.25	0.8045	77.91%	91.95%

Table 5: Related Works Model Comparisons

Publication	Research Focus	Methods	Data Source	Performance Score
(Muhammad et al., 2025)	Traditional classification of selfreported diabetes	Traditional Classification	BRFSS	Accuracy: 82%
(Allen, 2023)	Interpretable machine learning for obesity prevalence	Iterative Random Forest Regression	U.S. County-level features	R^2 Score: 0.79
(Gutiérrez-Gallego et al., 2024)	Interpretable prediction of overweight/obesity risk	Cascade Classifier (Gradient Boosting, RF, Logistic Regression)	Sociodemographic and Lifestyle Data	Accuracy: 79%
(Jindal et al., 2018)	Obesity prediction using ensemble machine-learning	Ensemble Machine Learning	Public Health Datasets	Accuracy: 89.68%
Proposed System	Hybrid spatial-ensemble obesity prevalence forecasting	K-Means Clustering + Stacking Regressor (HGB, RF, MLP)	CDC BRFSS	R^2 Score: 0.8045, Implied Accuracy: 91.95%

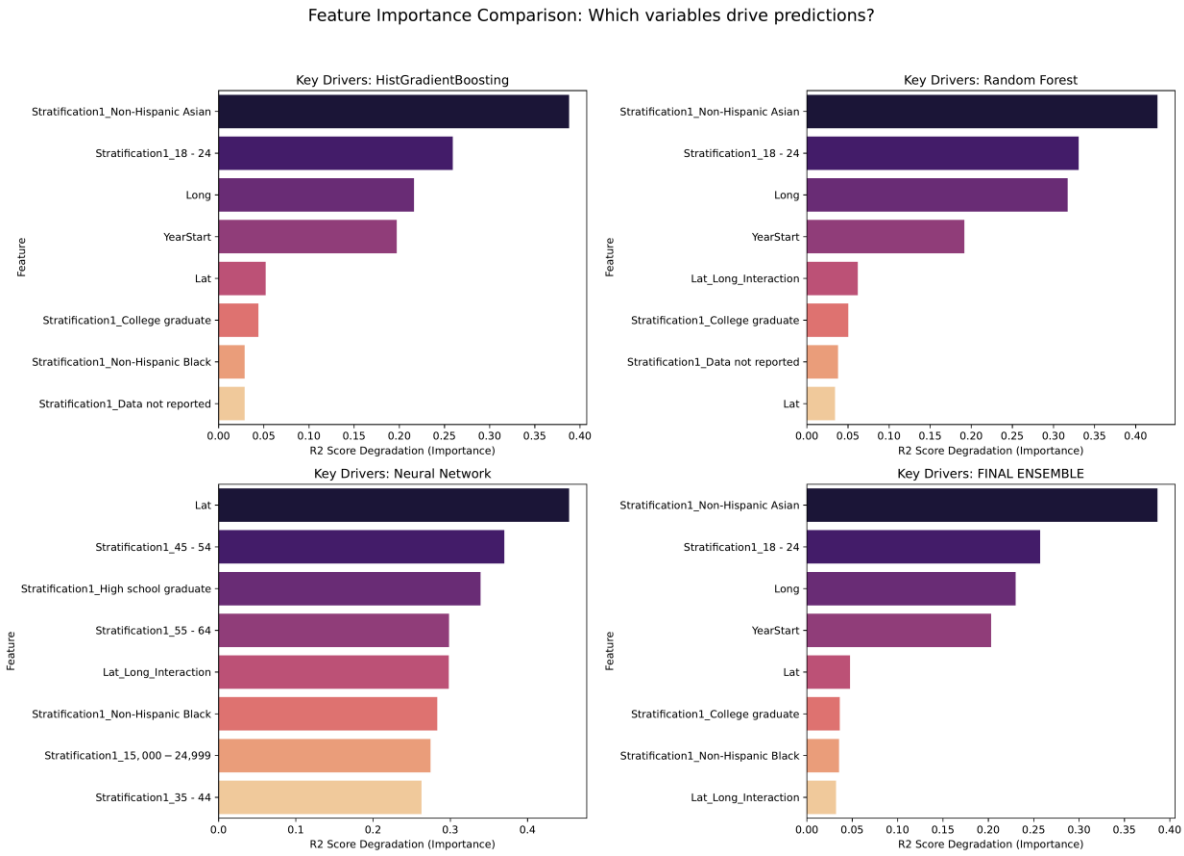

Fig. 10: Feature Importance Comparison

DOI: 10.3844/jcssp.2026.2139.2155

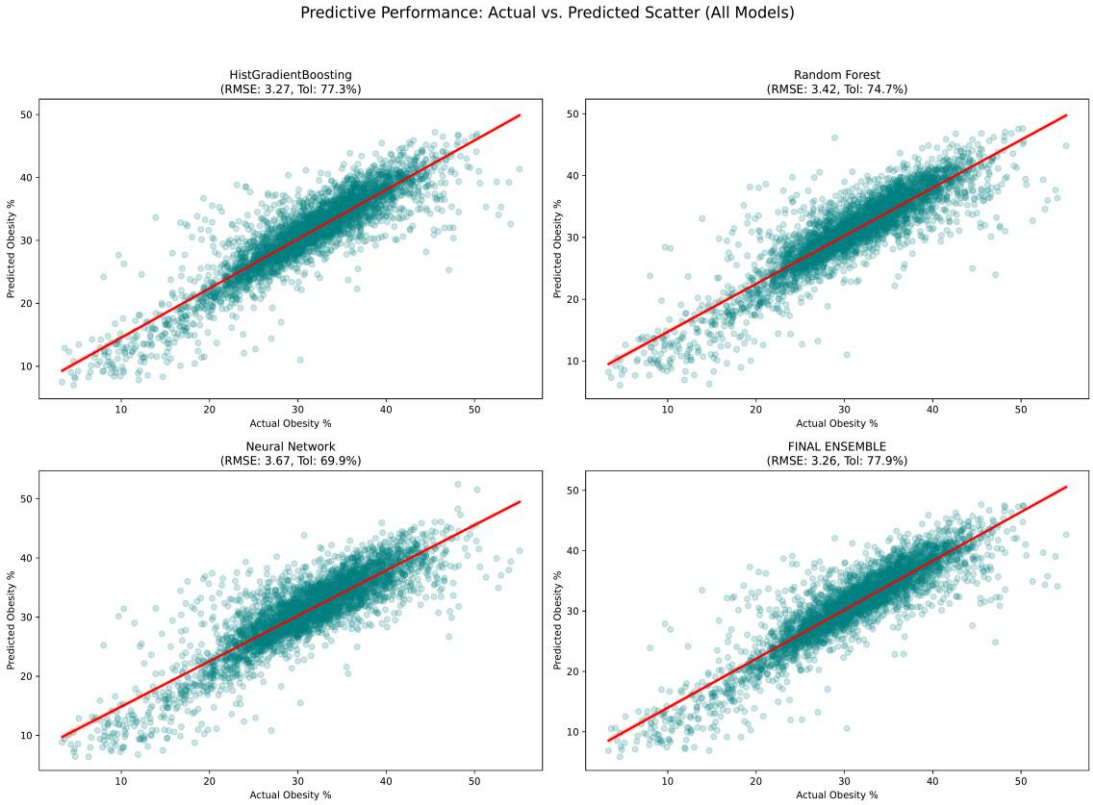

Fig. 11: Actual vs. Predicted Scatter (All Models)

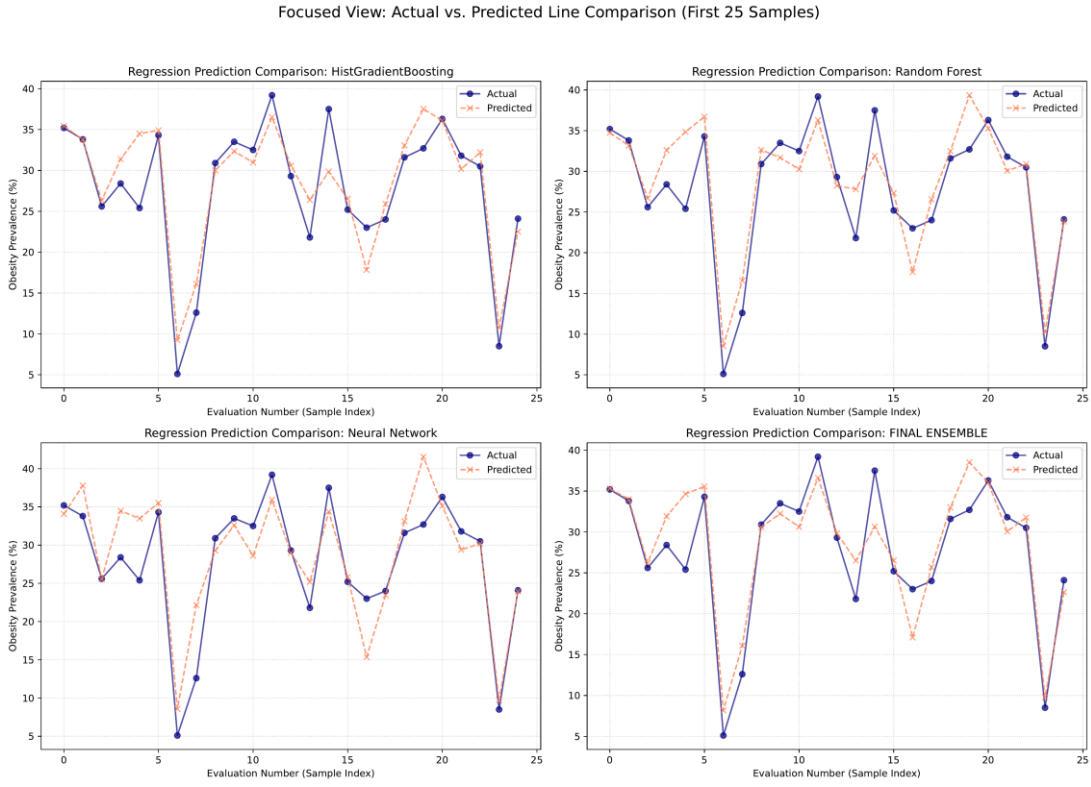

Fig. 12: Actual vs. Predicted Line Comparison

These cases showed that all of the models could predict new instances decently, with the final ensemble refining the prediction. Figure 13 shows the breakdown of the model predictions for the first 4 instances.

Computational Complexity

To evaluate operational feasibility, we analyze the framework's algorithmic complexity. The initial geodemographic profiling stage leverages a *K*-Means clustering routine bounded by $O(I \cdot K \cdot N \cdot d)$, where N represents total records, $K = 20$ clusters, $d = 3$ spatial features, and I denotes iterations to convergence; because K and d are static constants, this lookup scales linearly as $O(N)$.

The primary computational footprint resides in the parallel training of the base learners: $O(M \cdot N \log N)$ for the Random Forest (M = 250 trees), $O(H \cdot N \cdot d\_eff)$ for the HistGradientBoosting Regressor (H = 1000 iterations), and $O(E \cdot N \cdot \sum(lj \cdot lj+1))$ for the Multi-Layer Perceptron ($l_1 = 100$, $l_2 = 50$). The RidgeCV meta-learner introduces a negligible stacking overhead of $O(B^2 \cdot N + B^3)$ for $B = 3$ base models. Empirically, the entire pipeline processes from raw data ingestion to finalized stacked predictions in under 60 seconds on an Intel Core i7 processor utilizing multi-threaded execution, demonstrating excellent scalability for continuous public health monitoring.

Cross-Domain Generalizability Results

Stages 9 and 10 subjected the ensemble to 2 external datasets to measure its real-world robustness. This evaluated how well the model generalizes to data from external sources with varying degrees of sparsity and demographic focus.

Table 6 shows the results of the generalizability tests, including the models predicting on another CDC dataset that contains some similar information correlated with chronic disease indicators but had a different underlying methodology and reporting (Centers for Disease Control and (…Prevention [CDC-2], 2024). This dataset contains 309,215 total records split over numerous indicators, of which the obesity entries were extracted for use. While the implied accuracy and MAE scores were reasonable with a close RMSE, the other scores took some dips, which show that the less specific survey data may not be as applicable for the obesity prevalence model developed from the BRFSS. The implied accuracy hovered above 80% with the MAE near 5.0% except for the NN with 5.94%. The R-scores ranged close to 0.6, showing a still mostly positive relationship between the actual and predicted values, and the 10% tolerance accuracy dropped to approximately 40%. This shows that the models had a hard time finding a specific pattern in this data and were generally farther off from the actual values, but could still follow the general trend of the data and get approximately close to the actual values with relatively minimal outliers. Table 6 has the exact figures.

The Actual vs. Predicted plot for the U.S. Chronic Disease Indicators dataset showed a significant increase in variance with the R^2 of 0.2939, which is still impactful, but is not nearly as exemplary. The visual dispersion of points highlights the model's sensitivity to missing geolocation identifiers, confirming that spatial density is the "anchor" of its predictive power.

A closer look shows that the models trended towards clustered predictions that predicted neither high nor low consistently while following the overall data trend. Figure 14 shows the overall view for this test with the actual vs predicted.

Table 6 also shows the results of the models predicting on another dataset that contains national obesity percentages by state for a geographic emphasis (Kelly, 2024). This included 52 entries between states and territories with obesity prevalence, and the geographic location could be engineered by adding location data for each based on their actual positions.

```
============================================================
INITIATING 8 DISTINCT DEPLOYMENT CASES
============================================================

---> Running Deployment Case #1: 18 - 24 (Age (years))

--- STAGE 8: DEPLOYMENT (PREDICTING NEW INSTANCE) ---
Model Architecture        | Predicted Prevalence
--------------------------------------------------
HistGradientBoosting      | 21.90%
Random Forest             | 20.52%
Neural Network            | 20.53%
FINAL ENSEMBLE            | 21.10%

---> Running Deployment Case #2: $75,000 or greater (Income)

--- STAGE 8: DEPLOYMENT (PREDICTING NEW INSTANCE) ---
Model Architecture        | Predicted Prevalence
--------------------------------------------------
HistGradientBoosting      | 37.32%
Random Forest             | 38.10%
Neural Network            | 39.18%
FINAL ENSEMBLE            | 37.90%

---> Running Deployment Case #3: College graduate (Education)

--- STAGE 8: DEPLOYMENT (PREDICTING NEW INSTANCE) ---
Model Architecture        | Predicted Prevalence
--------------------------------------------------
HistGradientBoosting      | 33.84%
Random Forest             | 34.32%
Neural Network            | 32.72%
FINAL ENSEMBLE            | 34.00%

---> Running Deployment Case #4: Hispanic (Race/Ethnicity)

--- STAGE 8: DEPLOYMENT (PREDICTING NEW INSTANCE) ---
Model Architecture        | Predicted Prevalence
--------------------------------------------------
HistGradientBoosting      | 32.95%
Random Forest             | 29.53%
Neural Network            | 35.54%
FINAL ENSEMBLE            | 32.29%
```

Fig. 13: Deployment Results (Predicting 4 Instances)

DOI: 10.3844/jcssp.2026.2139.2155

Table 6: Generalizability Results

Model (Dataset)	r	MAE	RMSE	R^2	10% Tol.	Imp. Acc.
HGB (CDI)	0.628	4.93%	6.21	0.2587	39.97%	85.01%
RF (CDI)	0.577	4.94%	6.20	0.2603	40.88%	84.00%
NN (CDI)	0.574	5.94%	7.36	-0.0422	34.88%	80.51%
Ensemble (CDI)	0.614	4.80%	6.06	0.2939	42.14%	85.02%
HGB (State)	0.898	4.62%	4.92	-0.6821	23.08%	83.48%
RF (State)	0.816	3.16%	3.74	0.0310	50.00%	88.70%
NN (State)	0.423	12.73%	13.21	-11.1123	0.00%	57.10%
Ensemble (State)	0.888	3.03%	3.49	0.1539	44.23%	89.06%

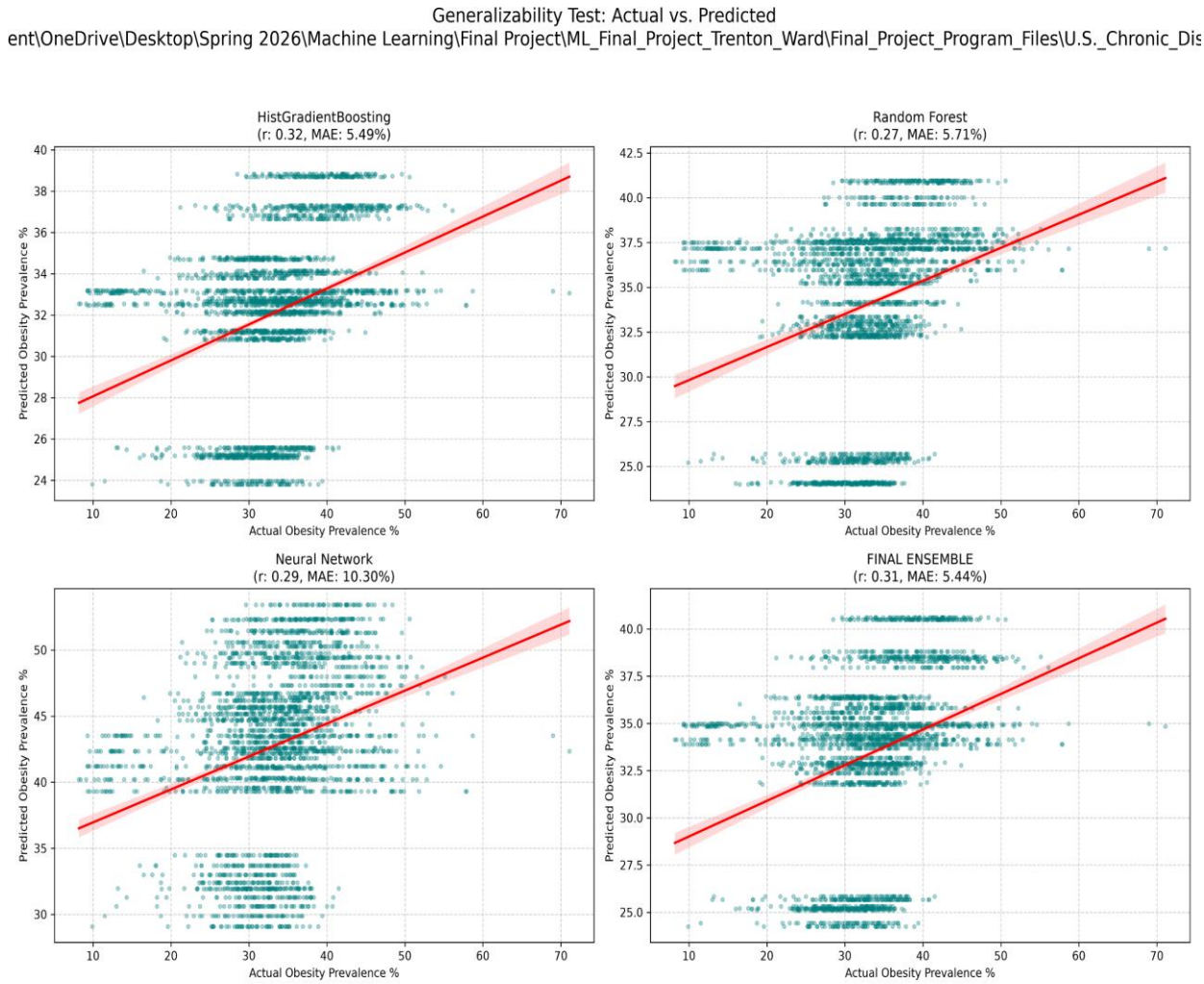

Fig. 14: Generalizability Test Actual vs. Predicted for CDI Dataset

While the scores were better than for the secondary CDI dataset, except for R^2, the neural network struggled greatly, and the final ensemble showed once again why its architecture is beneficial by outperforming the other models. The other 3 models performed well overall, though with respectable scores only slightly worse performing than for the primary dataset, with r scores close to 0.85, MAE below 5%, close RMSE scores, and above 80% implied accuracy showing the ability to follow the trends of the data, but also had poor R^2 values and 10% tolerance accuracies showing lower explanatory power and higher misses. The incredibly low R^2 for the NN

model shows that it was completely misaligned with the patterns in this differing dataset. Table 6 has the exact figures.

Despite being an aggregate dataset, the ensemble maintained a strong ranking capability with an R of 0.888. The Actual vs. Predicted chart shows the model successfully distinguishes between "High-Risk" and "Low-Risk" states, even though the MAE increased to 3.03% due to the loss of granular demographic data. The actual vs. predicted plots for this generalizability test showed that the HistGradientBoosting, random forest, and final ensemble models tended to predict higher, but still followed the general trend, while the neural network consistently predicted very low, leading to its results. These lower predictions from the NN helped to balance the higher predictions from the other 2 base regressors, which led to the final ensemble's greater performance. Figure 15 shows the actual vs. predicted plots for the final ensemble model for the state dataset.

Overall, while the r, MAE, RMSE, and implied accuracy were promising for the 2 generalizability tests, the other performance measures indicated that the models struggled when applied outside the original dataset, especially the neural network. While the additional datasets didn't contain all of the same information, were reported in a different format, or had differing underlying methodologies for gathering data, these results reflect real-world application for the system, which will need some adjustments to expand its capabilities beyond the BRFSS from the CDC.

Discussion Summary

The synergy between unsupervised spatial partitioning and the stacking ensemble architecture provides a robust solution to the "neighborhood effect" often observed in public health datasets. By assigning geodemographic clusters via K-Means before the regression phase, the pipeline treats geographic coordinates not merely as numerical inputs, but as categorical domains that represent shared socioeconomic and environmental characteristics. This localized spatial context allows the supervised learners to adjust their weights based on the specific health profile of a region, rather than applying a global average that might ignore localized variance.

The success of the stacking regressor, governed by a RidgeCV meta-learner, lies in its ability to navigate the bias-variance tradeoff. While the Multi-Layer Perceptron (MLP) captured complex non-linearities, it displayed higher variance in smaller data subsets. In contrast to this, the tree-based models with HistGradientBoosting and Random Forest provided stability. The ensemble effectively leveraged these strengths, resulting in a predictive engine that outperformed its individual components in both R^2 stability and MAE.

Furthermore, the implementation of the "Deployment Hub" transitions the model from a static research artifact into a functional decision-support tool. By presenting a multi-model consensus, the hub offers a measure of predictive uncertainty where a high variance between base models signals to health officials that a specific demographic instance may be an edge case requiring further manual audit.

This system does have a few trade-offs however as the stacking ensemble with a neural network base learner is very expensive to train compared to less complex algorithms, the model's predictions rely heavily on the quality of some features that may not always be available which diminishes the system's primary advantage, and, as the generalizability tests showed, the model is powerful but its performance and ability to predict is tied to specific population distributions.

Limitations

Several limitations qualify the interpretation of these results. First, the model inherits the methodological constraints of the BRFSS instrument itself. BRFSS is a self-reported telephone survey, so height and weight, and therefore the derived obesity prevalence, are subject to recall and social-desirability bias that tend to underestimate true prevalence.

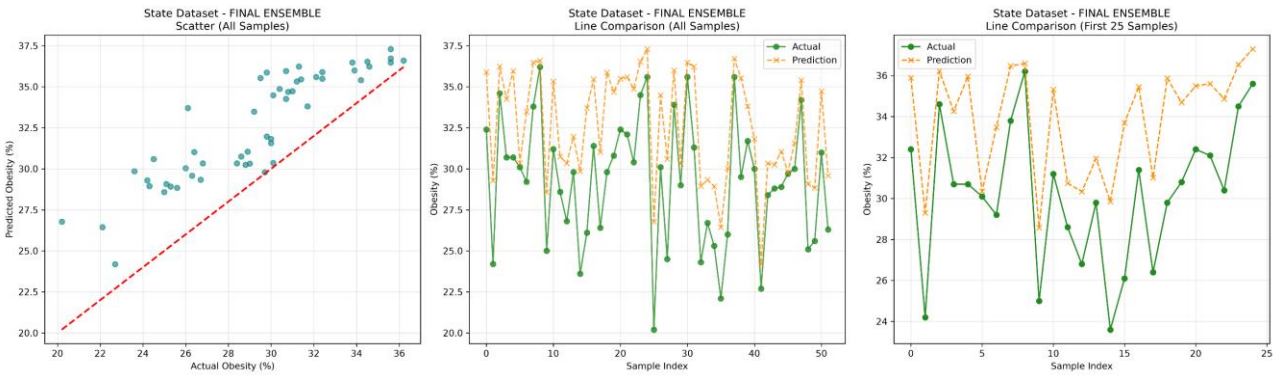

Fig. 15: Final Ensemble State Dataset Actual vs. Predicted

Its sampling frame, post-stratification weighting, and periodic revisions to questionnaire design further introduce measurement heterogeneity across years and states that the model treats as ground truth. Predictions are therefore only as reliable as the underlying survey estimates and should be read as forecasts of reported prevalence rather than of clinically measured obesity.

Second, the framework's predictive strength is anchored in the engineered geodemographic features, which makes it sensitive to spatial data sparsity. When latitude, longitude, or population-density information is missing, coarse, or inconsistently coded, as in the Chronic Disease Indicators dataset, the geodemographic cluster assignment becomes unreliable and the model loses the localized context that drives its accuracy. The pronounced R-squared degradation observed on the sparse external data confirms that the approach presumes reasonably complete and consistently georeferenced records.

Third, performance declines measurably under distribution shift. The cross-domain tests showed marked drops in "R^2" and "10% tolerance accuracy accuracy on the CDI and state-level datasets, and the neural-network base learner in particular became unstable when applied to historical state aggregates from the early 2010s. These results indicate that the model is calibrated to the demographic composition and temporal baseline of its training distribution; concept drift arising from secular increases in obesity, evolving survey methodology, or differing demographic granularity can bias predictions when the model is transferred without recalibration.

Conclusion

The primary objective of this research was to develop a high-precision predictive framework for obesity prevalence that accounts for the complex interplay between geographic and demographic variables. By transitioning from a standard classification approach to a hybrid regression-based ensemble, this study provides a more granular tool for public health analysis. The methodology utilized a 10-stage pipeline that integrated unsupervised K-Means clustering with a layered supervised Stacking Regressor. The experimental results validate the efficacy of this approach, with the final ensemble achieving an R^2 of 0.8045 and a Mean Absolute Error of 2.20%. When compared to existing literature, such as Tan et al. (2025) and Allen (2023), this work demonstrates competitive predictive precision through its additional scores, such as 91.95% implied accuracy and its ability to forecast exact prevalence percentages across diverse demographic strata. Furthermore, the high Pearson correlation with an r of 0.888 observed in the state-level geographic generalizability test confirms the importance of spatial interaction features, supporting the "Neighborhood Effect" theories established by Oshan et al. (2020). However, the rigor of this study is primarily defined by its generalizability stress tests, which illuminate the practical boundaries of the model.

The ensemble's performance on external datasets provided two critical insights into the future of health informatics. First, the "spatial anchor" effect was confirmed via the CDI dataset with the performance degradation in the less specific survey data. Second, the tests on the State-level dataset revealed the impact of temporal decay and concept drift. While the model remained an excellent tool for geographic ranking with the ability to distinguish high-risk from low-risk states, it tended to over-predict when applied to historical baselines from the early 2010s. Ultimately, these results suggest that while hybrid ensemble models offer a superior framework for localized public health forecasting, their broader deployment must account for shifting temporal baselines and data sparsity.

Overall, the system performed well with the original BRFSS dataset and provided valuable insights into being able to apply machine learning models beyond their foundation, which is essential for real-world applications. The practical utility of this system extends beyond academic research. This pipeline provides a scalable foundation for state health departments to anticipate future resource requirements and design targeted interventions based on evolving geodemographic trends. State health departments and urban planners can also utilize the Deployment Hub to simulate how shifts in demographic distributions or resource allocations might impact regional obesity trends. By identifying high-risk geodemographic clusters, policymakers can implement hyper-local interventions, such as the placement of community health centers or the subsidization of fresh food access in predicted hotspots.

While this research demonstrates strong predictive power, several avenues for potential exploration remain. First, future research can focus on integrating dynamic weight recalibration to mitigate concept drift, ensuring the model remains robust as public health trends evolve across different decades and demographic shifts. Second, the integration of Federated Learning protocols would allow the model to be trained on private, localized clinical records from healthcare providers without compromising patient confidentiality, such as addressing the data sparsity observed in the CDI and California stress tests. Third, incorporating real-time data from wearable IoT devices could provide a dynamic temporal layer to the model, allowing for monthly rather than annual prevalence monitoring. Finally, further research into interpretation techniques will be essential to ensure that as these models grow in complexity, their decision-making processes remain transparent and trustworthy for the

public health stakeholders who rely on them. A specific open problem that follows from these findings is how to update the geodemographic clustering layer incrementally as new surveillance data arrive, rather than retraining the entire pipeline from scratch. Because each cluster reassignment currently requires re-running K-Means over the full feature space and refitting every base learner, real-time deployment is computationally bottlenecked; developing online or streaming clustering with warm-started ensemble updates that preserve calibration while absorbing fresh records remains an important direction for continuous public-health surveillance.

Acknowledgment

The authors gratefully acknowledge the resources made available by Norfolk State University, Virginia, USA. To promote reproducibility, datasets, feature engineering scripts, and partial code from this study will be made available upon reasonable request.

Funding Information

The authors have not received any financial support or funding to report.

Authors' Contributions

The authors equally contribute to this research.

Ethics

This research did not involve human participants, animal subjects, or the use of personally identifiable data. All datasets utilized are publicly available and were used in accordance with their respective terms of use. No ethical concerns were identified in the conduct of this study.

References

Allen, B. (2023). An interpretable machine learning model of cross-sectional U.S. county-level obesity prevalence using explainable artificial intelligence. *PLOS ONE*, *18*(10), e0292341. https://doi.org/10.1371/journal.pone.0292341

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233–234. https://doi.org/10.1038/nmeth.4642

Cawley, J., Mitra, B., Meyerhoefer, C., Ding, E., Peng, T., Hatfield, M., Garay, M. S., & Chhatwal, J. (2021). Direct medical costs of obesity in the United States and the most populous states. *Journal of Managed Care & Specialty Pharmacy*, *27*(3), 354–366. https://doi.org/10.18553/jmcp.2021.20410

Centers for Disease Control and Prevention (CDC-2). (2024). U.S. Chronic Disease Indicators. *Data.Gov*.

Centers for Disease Control and Prevention (CDC) (2024). Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System. *Data.Gov*.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, 1–15. https://doi.org/10.1007/3-540-45014-9_1

Emmerich, S., Fryar, C., Stierman, B., & Ogden, C. (2024). *Obesity and Severe Obesity Prevalence in Adults: United States, August 2021–August 2023*. https://doi.org/10.15620/cdc/159281

Gutiérrez-Gallego, A., Zamorano-León, J. J., Parra-Rodríguez, D., Zekri-Nechar, K., Velasco, J. M., Garnica, Ó., Jiménez-García, R., López-de-Andrés, A., Cuadrado-Corrales, N., Carabantes-Alarcón, D., Lahera, V., Martínez-Martínez, C. H., & Hidalgo, J. I. (2024). Combination of Machine Learning Techniques to Predict Overweight/Obesity in Adults. *Journal of Personalized Medicine*, *14*(8), 816. https://doi.org/10.3390/jpm14080816

Jindal, K., Baliyan, N., & Singh Rana, P. (2018). Obesity Prediction Using Ensemble Machine Learning Approaches. *Recent Findings in Intelligent Computing Techniques*, *708*, 355–362. https://doi.org/10.1007/978-981-10-8636-6_37

Kelly, I. (2024). National obesity percentages by state. *Kaggle*.

Muhammad, M. A., Sani, J., & Ahmed, M. M. (2025). Exploring Explainable Machine Learning for Predicting and Interpreting Self-Reported Diabetes among Tennessee Adults: Insights from the 2023 Behavioral Risk Factor Surveillance System (BRFSS). *Journal of Primary Care & Community Health*, *16*, 1–11. https://doi.org/10.1177/21501319251400546

Oshan, T. M., Smith, J. P., & Fotheringham, A. S. (2020). Targeting the spatial context of obesity determinants via multiscale geographically weighted regression. *International Journal of Health Geographics*, *19*, 11. https://doi.org/10.1186/s12942-020-00204-6

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Tan, W., Geng, B., & Bai, X. (2025). Spatiotemporal prediction of obesity rates and model interpretability analysis from a public health perspective. *PLOS ONE*, *20*(11), e0335908. https://doi.org/10.1371/journal.pone.0335908

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/s0893-6080(05)80023-1