Research Strategy

A. Significance and Background

A.1 Significance

Place-based health research has emerged as a crucial method for understanding the interplay between geographic, social, and biological factors influencing cardiovascular disease (CVD) outcomes. Significant research gaps persist despite its importance, particularly among Native Hawaiian and Pacific Islander (NHPI) populations and Alaska Native and American Indian (AI/AN) individuals. These underrepresented, understudied, underserved (U3) populations experience unique health disparities often obscured when aggregated into broader racial categories (Waitzfelder et al., 2023). The integration of advanced methodologies, including longitudinal analyses, combinatorial frameworks for social determinants of health (SDHs), and 4D geospatial approaches, offers an unprecedented opportunity to address these gaps and better understand the etiology of CVD in these communities.

1) Research has historically failed to generalize to NHPI/AI/AN populations, aggregating them into wider racial groups. Research has historically failed to generalize to NHPI and AI/AN populations, often aggregating them into broader racial groups like "Asian" or "Other," which obscures true identity and the unique health disparities faced by these U3 communities (Office of Minority Health, 2023). Past and present studies have continued highlighting the importance of disaggregating social, genetic, and environmental components in these groups to reveal critical health disparities that are otherwise hidden (Espey et al., 2014; Lee et al., 2024). This aggregation diminishes the ability to uncover intricate combinations of SDH synergized to amplify these health risks and requires combinatorial approaches to understand the multidimensional relationships between these factors. Social and economic parameters such as limited access to healthcare, lower educational attainment, unemployment, and housing instability intensify health disparities among these groups (Centers for Disease Control and Prevention, 2014). For instance, NHPI and AI/AN populations experience higher rates of poverty compared to non-Hispanic Whites, affecting their ability to access quality healthcare and nutritious food (U.S. Census Bureau, 2020). Educational disparities persist, with lower high school and college graduation rates impacting employment opportunities and income levels (National Center for Education Statistics, 2019). Additionally, housing instability and overcrowding are more prevalent in NHPI and AI/AN communities, leading to increased psychosocial stress and exposure to environmental hazards (Department of Housing and Urban Development, 2021). These challenges suggest that NHPI and AI/AN populations face greater obstacles in accessing quality healthcare, education, employment opportunities, and safe housing compared to other demographics nationwide. Such disparities contribute to higher rates of chronic diseases, mental health issues, and lower life expectancy (Indian Health Service, 2021). Addressing these disparities requires targeted research and community interventions that acknowledge and cater to the unique challenges faced by the NHPI and AI/AN populace (National Institutes of Health, 2024).

2) Cardiovascular disease (CVD) remains a significant public health concern. CVD remains a significant public health concern, particularly for NHPI and AI/AN populations. The estimated prevalence of CVD is significantly higher among NHPI and AI/AN populations compared to non-Hispanic White (NHW) populations. For instance, a study on NHPI populations in Hawaii and California found that the estimated prevalence of CVD ranged from 5.27% among single-ethnic NHPI groups to 8.53% among Pacific Islander-White multiracial groups, compared to just 1.81% among NHWs (Waitzfelder et al., 2019). Similarly, AI/AN individuals experience higher rates of heart disease and related mortality than NHWs (CDC, 2019). The prevalence of ischemic stroke is also more than twice as high in NHPI and AI/AN populations compared to NHWs, with these individuals experiencing an earlier onset of stroke by an average of 10 years (Nakagawa et al., 2013; Howard et al., 2014). This earlier onset could be attributed to the higher prevalence of metabolic syndrome, diabetes, obesity, and other CVD risk factors such as dyslipidemia within NHPI and AI/AN communities (Barnes et al., 2010; Rodriguez et al., 2014). These disparities are further compounded by lower levels of heart attack education or awareness of early warning signs in both NHPI and AI/AN populations. Less than 44.4% of NHPI adults possess the recommended skills for recognizing and appropriately responding to heart attack symptoms, contributing to poorer health outcomes. Similarly, AI/AN populations often have limited awareness of heart attack or stroke symptoms and CVD risk factors in general, which may delay seeking medical attention (Felix et al., 2019; Galloway, 2005). Much uncertainty persists about how environmental, genetic, and behavioral factors dynamically interact over time and space to drive these disparities.

3) Geospatial analysis can fulfill an unmet role in cardiovascular research. Many social determinants of health (SDHs) impacting disparities in CVD prevalence and management vary significantly across geographic

regions. Traditional epidemiological techniques often fail to account for this spatial variability, potentially overlooking critical factors influencing health outcomes. Geographic Information Systems (GIS) and geospatial analysis approaches enable researchers to model these spatial variations, offering deeper insights into how specific socioeconomic and environmental factors affect health across different areas. Previous studies, such as Hadley et al. (2022), have utilized geospatial methods to highlight the importance of place-based health research, particularly in addressing health disparities among marginalized populations. NHPI and AI/AN populations are especially vulnerable to place-based health inequities due to factors such as residential segregation, historical trauma, and limited access to healthcare resources (Mamun et al., 2024). Despite considerable evidence of CVD risk factors in these U3 populations, few studies employ advanced geospatial or machine learning techniques to capture how these risks unfold across longitudinal timeframes. By integrating SDH and CVD data in a more robust, 4D (spatial + temporal) framework, our work aims to fill a crucial gap in current biomedical research using state-of-the-art geospatial ML models.

A.2 Background

A.2.1 Cardiovascular Disease Pathophysiology

Cardiovascular disease (CVD) encompasses a range of conditions affecting the heart and blood vessels, including coronary artery disease (CAD), heart failure (HF), hypertension, and stroke. CVD is the leading cause of mortality globally, accounting for approximately 17.9 million deaths each year (World Health Organization, 2023). The burden of CVD is not only measured in terms of mortality but also in morbidity, as it significantly impacts quality of life and imposes substantial economic costs on healthcare systems. CVD is influenced by a combination of modifiable and non-modifiable risk factors. Non-modifiable factors include age, sex, and genetic predisposition. Modifiable risk factors encompass lifestyle choices such as smoking, physical inactivity, poor diet, and excessive alcohol consumption. Additionally, conditions like hypertension, diabetes, dyslipidemia, and obesity are critical contributors to the development and progression of CVD (Benjamin et al., 2019). The underlying pathophysiology of CVD involves the buildup of atherosclerotic plaques within arterial walls, leading to narrowed and hardened arteries. This process restricts blood flow, reducing oxygen and nutrient delivery to vital organs. Inflammation and oxidative stress play pivotal roles in plaque formation and destabilization, increasing the risk of acute cardiovascular events like myocardial infarction (MI) and stroke (Libby et al., 2002).

While advancements in medical treatments and preventive strategies have reduced CVD mortality in some populations, disparities persist. Certain racial and ethnic groups, including NHPI, AI/AN, and multiracial populations within these demographics, experience higher prevalence and mortality rates from CVD compared to NHWs. These disparities are exacerbated by socioeconomic factors, access to healthcare, and environmental exposures unique to these communities (Younus et al., 2016; Weir et al., 2016; Heidenreich et al., 2013; Khomtchouk et al., 2020).

A.2.2 CVD Outcomes in Diverse Populations

Persons with diverse racial and ethnic backgrounds are unduly impacted by CVD, including metabolic comorbidities such as type II diabetes (T2D), obesity, and dyslipidemia. Racial and ethnic differences in the epidemiology of CVD in the United States result in significant disparities due to a complex interplay of genetic, environmental, and socioeconomic factors. For example, AI/AN populations face disproportionately high rates of cardiometabolic diseases. Studies indicate that AI/AN individuals have significantly higher prevalence rates of T2D, obesity, and hypertension compared to NHWs. These disparities are attributed to social and economic parameters such as limited access to healthcare, historical trauma, socioeconomic challenges, and lifestyle changes resulting from acculturation and environmental shifts (Devi et al., 2018; Singh et al., 2020). Multiracial individuals often encounter compounded health disparities due to intersecting identities and the lack of targeted research. The aggregation of multiracial data into broad categories can obscure specific health risks and protective factors unique to these groups. Research indicates that multiracial individuals may experience higher stress levels, discrimination, and socioeconomic disadvantages, which contribute to increased CVD risk (Adams et al., 2015; Ju et al., 2017). As previously highlighted, NHPI populations exhibit significantly higher rates of CVD and its risk factors. The intersection of genetic factors, dietary patterns, socioeconomic status, and limited access to culturally appropriate and concordant care exacerbates these disparities further (Aluli et al., 2007; Aluli et al., 2010). Understanding these systemic issues programmatically requires access to disaggregated datasets. Therefore, precision medicine approaches that account for genetic diversity, environmental exposures, and social determinants of health are essential for developing effective prevention and treatment strategies for these U3 populations (Taparra et al., 2022).

A.2.3 Role of Machine Learning and Artificial Intelligence in Cardiovascular Research

Machine Learning (ML) and Artificial Intelligence (AI) have revolutionized cardiovascular research by enabling the analysis of large, complex datasets to uncover patterns and insights that traditional statistical methods often overlook. These technologies facilitate the integration of diverse data sources, including electronic health records (EHRs), genetic information, and geospatial data, enhancing the accuracy of CVD risk prediction and diagnosis (Wiemken & Kelley, 2019; Wiens & Shenoy, 2017). For example, deep learning algorithms such as convolutional neural networks (CNNs) are employed to analyze medical imaging data for early detection of CVD, while ML models integrate genomic data to identify novel genetic markers associated with disease susceptibility and progression (Litjens et al., 2017; Kourou et al., 2015). Furthermore, explainable AI (XAI) techniques, including Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), enhance model transparency and trust by elucidating the factors driving predictions, thereby enabling clinicians to make informed decisions based on the interpretable insights that are generated from it (Ribeiro et al., 2016; Doshi-Velez & Kim, 2017).

A.2.4 Geospatial Health Research Interface with AI/ML Modeling

The convergence of AI/machine learning and geospatial health research offers unprecedented opportunities to address cardiovascular disease (CVD) disparities in underserved populations across the United States. By incorporating geospatial data—such as socioeconomic status, healthcare access, and environmental exposures—as features trained in large-scale AI/ML models, researchers can develop more nuanced and location-specific CVD risk assessments. Geospatial machine learning models, including Geographically Weighted Regression (GWR) and Geographically Weighted Artificial Neural Networks (GWANN), account for regional variability and spatial dependencies, enabling the identification of localized risk factors and the development of targeted community interventions (Fotheringham et al., 2002; Guo & Matisziw, 2021). This spatially-weighted approach not only enhances the predictive analytics power of these data models but also facilitates the creation of tailored public health strategies that can address the unique needs of Native Hawaiian and Pacific Islander (NHPI), American Indian, and Alaska Native (AI/AN), and associated multiracial communities, ultimately contributing to the reduction of CVD disparities (Wang, 2020; Krieger, 2012).

B. Innovation

This project introduces novel epidemiological approaches for investigating health disparities among Native Hawaiian and Pacific Islander (NHPI) and American Indian and Alaska Native (AI/AN) populations, leveraging 4D (spatiotemporal) insights and XAI approaches within advanced geospatial ML frameworks to define the complex, multidimensional social determinants of health (SDH) drivers of cardiovascular disease (CVD) risk which provide actionable, interpretable insights for underserved populations. Three major innovations distinguish this research:

1. **Race-Agnostic Analysis in a 4D Context**
   By omitting race as a predictive input, our approach mitigates bias while focusing on how SDHs evolve across not only space but time. This 4D, race-agnostic perspective enables a more equitable lens on CVD disparities and allows for more robust modeling of these predictors.
2. **Combinatorial Geospatial Machine Learning Models**
   We deploy cutting-edge geospatial ML techniques such as Geographically Weighted Artificial Neural Networks (GWANN) to capture combinatorial interactions among environmental factors, healthcare access, and socioeconomic variables. By modeling these intricate, multi-layered relationships at the micro level, we can uncover nuanced “hot spots” (both geospatially and temporally) of CVD risk that conventional epidemiological methods often overlook.
3. **XAI for Interpretable CVD Risk Factors**
   We integrate XAI methods to clarify which SDH factors most influence CVD outcomes for these underserved populations. This transparency ensures that public health officials and community leaders can develop highly effective community interventions guided by our data-driven, 4D insights (from interpretable white-box models versus opaque black-box ones), ultimately bridging the gap between our research findings and the real-world impact we hope to achieve with this project.

C. Approach

C.1.1 Data Collection

To investigate cardiovascular disease (CVD) disparities among Native Hawaiian and Pacific Islander (NHPI) and American Indian and Alaska Native (AI/AN) populations, we systematically curated Social Determinants of Health (SDH) data from a diverse set of governmental sources, including the American Community Survey (ACS), CDC Social Vulnerability Index (SVI), USDA Food Environment Atlas, and HRSA's Maternal and Child Health Bureau. These datasets encompass vital SDH indicators such as income levels, food security, healthcare access, and housing stability. For example, data from the CDC's SVI (2016–2020) allows us to assess socioeconomic status at the geographic county level, while the USDA Food Environment Atlas offers insights into food availability and access to healthy food, helping identify regions where food insecurity may exacerbate CVD risks. Additionally, healthcare access data from HRSA provides critical information on the availability of primary care providers and healthcare infrastructure systems at a geospatial level. The below table provides a comprehensive list of the datasets we use in our study, as well as the source of each data point.

Survey	Source	Number of Features
ACS 5-Year Estimates	U.S. Census Bureau	18
Food Environment Atlas	USDA	281
Health Resources & Services Administration	HRSA	2
Atlas of Heart Disease and Stroke	CDC	200
Social Vulnerability Index (SVI)	CDC/ATSDR	157
Environmental Justice Index (EJI)	CDC	36
Robert Wood Johnson Foundation's County Rankings	RWJF	28
National Environmental Public Health Tracking	CDC	11
Walkability Index	EPA	129
Community Resilience Estimates (2022)	U.S. Census Bureau	17
Medicare Data	CMS	100
PLACES: SDOH Measures for County ACS	CDC	18
National Risk Index	CDC/ATSDR	11
Area Deprivation Index (ADI)	University of Wisconsin	17
Small Area Income and Poverty Estimates (SAIPE)	U.S. Census Bureau	31
Environmental Public Health Tracking (Air)	CDC	14
Farmers' Market SNAP Participation	USDA	10
Maternal and Child Health Provider Availability	HRSA	55
National Healthcare Quality and Disparities	AHRQ	250
Public School Accessibility	CDC	15
Community Resilience Estimates for Equity	U.S. Census Bureau	344

Table 1: Table of collected surveys and their respective sources.

C.1.2 Data Imputation, Standardization, and Cleaning

In order to do geospatially-aware machine learning on our collected dataset, we first had to concentrate the constituent datasets into a single, larger dataset, and perform proper data standardization and feature engineering to maintain the quality and biological relevance (i.e. to not drop biologically relevant variables when feature engineering) while making it tenable for machine learning.

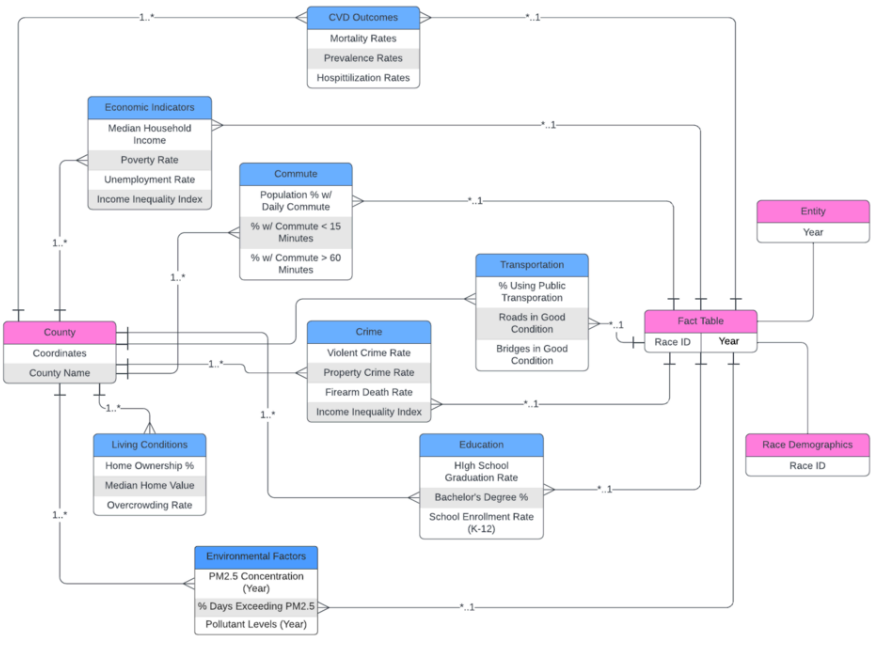
This is especially important given that some columns exhibited a high proportion of missing values with invalid input rates reaching up to 99%, as some variables were only collected for a few counties. To maintain data integrity, U.S. territories, including Puerto Rico, the U.S. Virgin Islands, and other non-state territories, were excluded from the dataset. Each row in the dataset corresponds to a county or county equivalent, and after this process, the final row count was reduced to 3,143, aligning with the official count of U.S. counties in 2022. After the initial data collection phase that constructed the main dataset, as addressed earlier, we identified two other useful smaller datasets—one containing important geological data such as latitude and longitude, and the other containing valuable immigration data to help contextualize immigration patterns when explaining the social determinants of CVD. This second dataset is rich in NHPI population data. Both datasets required only minor cleaning. To resolve inconsistencies in county naming conventions across these datasets (e.g., "Acadia Parish" vs. "Acadia"), we utilized FAISS (Facebook AI Similarity Search) to identify and match similar entries. The resulting unified dataset contained 3,143 rows and 2,066 columns.

The preparation of the dataset involved achieving a base level of cleanliness and standardization suitable for training ML models while maintaining flexibility for model-specific adaptations. Different ML models impose unique requirements on data structures, and we carefully preprocessed the data to ensure compatibility. We began by identifying and removing exact duplicate columns, retaining only one instance of each. To mitigate issues related to multicollinearity, which can distort feature importance and model interpretability, we analyzed pairwise correlations across features. Columns with correlations above 0.9 were consolidated by retaining only one representative feature, except in cases where highly correlated columns captured similar data across closely related time periods (air_quality_2019 vs. air_quality_2020, for example), which were retained for temporal context. To address missing data, we set a threshold of 30% invalid inputs per column. Features exceeding this threshold were removed to ensure that imputation did not introduce excessive bias. For the

remaining columns, missing values were addressed using various imputation techniques, including Bayesian imputation, K-Nearest Neighbors (KNN), and bootstrapping, chosen based on the specific nature and distribution of the missing data within each column. These preprocessing steps reduced the dataset to approximately 1,300 columns. This base-level dataset can now be further tailored to meet the specific requirements of geospatially aware machine learning models aimed at predicting cardiovascular disease (CVD) outcomes. For example, geospatial models like geographically weighted regression (GWR) or spatially explicit neural networks require latitude and longitude features to account for spatial heterogeneity in predictors. These models may necessitate additional transformations, such as creating distance-based features or spatial weight matrices, to capture regional patterns in CVD prevalence. In contrast, non-spatial models, such as random forests, focus on leveraging socio-demographic and environmental predictors without explicit geospatial inputs, relying instead on feature importance ranking to refine the predictors relevant to CVD, and thus do not require those additional transformations.

Additionally, in order to address concerns about data quality in many remote/underserved counties–particularly those with large NHPI and AI/AN communities, which often have incomplete SDH data, we will construct an adjacency graph with each node representing a county and each edge capturing the spatial proximity of the counties (or a domain specific connection, such as counties that share tribal health resources/fall under the same tribal governance system). We will use the GraphSAGE algorithm to generate node embedding that encapsulate these connections, which will then allow us to augment each county's baseline features (i.e. its poverty rate, healthcare utilization rate, etc) with the embedding, so we can borrow strength from better sampled neighbors or similar counties to improve the accuracy of our models.

To identify and refine the target outcomes for our geospatially aware machine learning models, we implemented a systematic approach to extract relevant variables from the extensive dataset. Using advanced pattern-matching techniques integrated with FAISS, we efficiently parsed the columns in the dataset to isolate those relevant to cardiovascular disease (CVD) outcomes. This process ensured that our models were trained on a solid foundation of well-defined and biologically meaningful targets while maintaining a focus on populations of interest, and to ensure there were no important targets that may be left out with a manual search. We searched for columns containing keywords associated with disease and health outcomes, such as "CVD," "disease," "condition," "disorder," "illness," "syndrome," "prevalence," "morbidity," "mortality," "death," and "outcome." This approach enabled the extraction of a robust subset of CVD outcome variables for use in our models.



Finally, in order to programmatically capture the structure of the cleaned dataset, we have constructed an entity-relationship diagram (ERD) that depicts the variable relationships.

Figure 1. Entity-Relationship Diagram (ERD) Plot of Dataset Schema. At the core of the diagram is the *County* entity, which serves as the primary unit of analysis, and connects the constituent SDH data. Each of these entities is linked to relevant attributes such as median household income, PM2.5 pollutant levels, or high school graduation rates, providing a comprehensive view of the social determinants of health (SDHs). The CVD Outcomes entity, containing variables like mortality and hospitalization rates, is directly tied to these SDH factors through

the Fact Table, which consolidates data across counties, years, and race demographics.
This relational structure allows us to perform both granular analyses of individual factors and combinatorial investigations into how these variables interact. The ERD framework supports further data cleaning and imputation workflows by clearly delineating the dependencies of the data and enabling the identification of missing or inconsistent entries across linked entities.

C.2 Developing and Applying Advanced Geospatial Machine Learning Models for CVD Risk Prediction

Aim 1 focuses on designing and applying robust, high-dimensional geospatial ML models that draw on both SDH and geographic information of U.S. counties to predict cardiovascular disease (CVD) risk, with particular emphasis on Native Hawaiian, Pacific Islander (NHPI), and Alaska Native and American Indian (AI/AN) populations. These groups often reside in regions with sparse coverage or and are often aggregated/not included in this type of data. Our plan thus features a systematic comparison of multiple ML approaches, advanced feature engineering, and careful use of spatial autocorrelation and weighting techniques. We will also incorporate combinatorial and temporal (4D) data analysis methods to capture subtle interdependencies that might be overlooked by simpler, static models.

C.2.0 Initial Approach

To explore cardiovascular disease (CVD) disparities, particularly in Native Hawaiian and Pacific Islander (NHPI) populations, we trained an Explainable Boosting Machine (EBM) using the cleaned data from C.1. The model underwent hyperparameter tuning using a 5-fold cross-validation approach, achieving an R^2 of 0.69, indicating strong predictive capability.

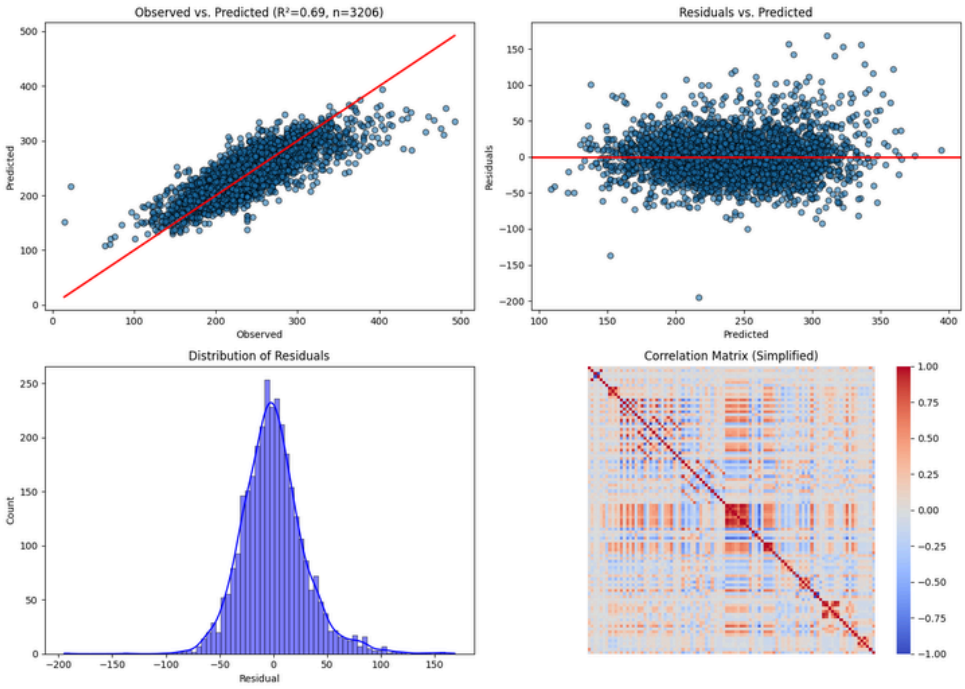


Figure 2. Visualizing the performance of the Explainable Boosting Machine (EBM) model trained on the dataset. (Top Left) Observed vs. Predicted plot showing an R^2 value of 0.69, indicating the model explains 69% of the variance in the target variable. (Top Right) Residuals vs. Predicted plot highlighting the distribution of residual errors. (Bottom Left) Distribution of residuals confirming their approximate normality, crucial for model validation. (Bottom Right) Simplified correlation matrix of the input features used for training.

With that being said, several issues emerge upon closer examination of the residuals and outlier analysis. While the residuals vs. predicted plot suggests an absence of systematic bias, the variance in residual magnitudes increases with higher predictions, revealing challenges with heteroscedasticity. Moreover, the residual distribution includes extreme outliers, highlighting the model's difficulty in generalizing to certain subsets of the data, particularly in low-data counties. For example, counties like Mellette in South Dakota and Ketchikan in Alaska show extreme prediction errors, with residuals of -194.60 and -137.30, respectively. This underscores the model's poor performance in regions with sparse or low-quality data, which disproportionately affects underserved and rural areas. Additionally, while multicollinearity is not a direct issue for EBMs due to their ability to handle correlated features through pairwise interaction terms and additive models, the correlation matrix reveals significant interdependencies among predictors. These interdependencies suggest the need for further feature engineering to better disentangle overlapping effects and improve the interpretability of the model's outputs. Furthermore, despite the model's interpretability strengths, the current framework does not account for geospatial dependencies, which are particularly relevant for social determinants of health (SDHs) influencing cardiovascular disease (CVD) outcomes.

C.2.1 Machine Learning Model Selection

Selecting suitable machine learning models is crucial for unraveling how the interaction of SDHs, geography, and other environmental factors converge to shape cardiovascular disease risk, with a specific emphasis on modeling for subpopulations such as NHPI and AI/AN. To do this we will combine traditional models such as Geographically Weighted Regression (GWR) and its multiscale variant, Multiscale GWR (MGWR), alongside nonlinear learners that can incorporate further spatial weighting—Random Forests (RF), Gradient Boosted Decision Trees (GBDT), ensemble methods like XGBoost (and related gradient boosting frameworks such as LightGBM and CatBoost), Support Vector Machines (SVM), and Geographically Weighted Artificial Neural Networks (GWANN). We will systematically compare these learners to ensure the robustness of our predictions and to ensure the highest accuracy on the data whale ensuring the results remain interpretable.

In our initial approach, our key priority was balancing predictive performance with interpretability and computational efficiency. The Explainable Boosting Machine (EBM) model excels in this regard, as demonstrated by its cumulative feature importance curve (Figure 5). This figure highlights that the top 25 features explain over 80% of the variance in cardiovascular disease (CVD) mortality predictions, underscoring the EBM model's ability to focus on the most impactful predictors while reducing reliance on less relevant features.

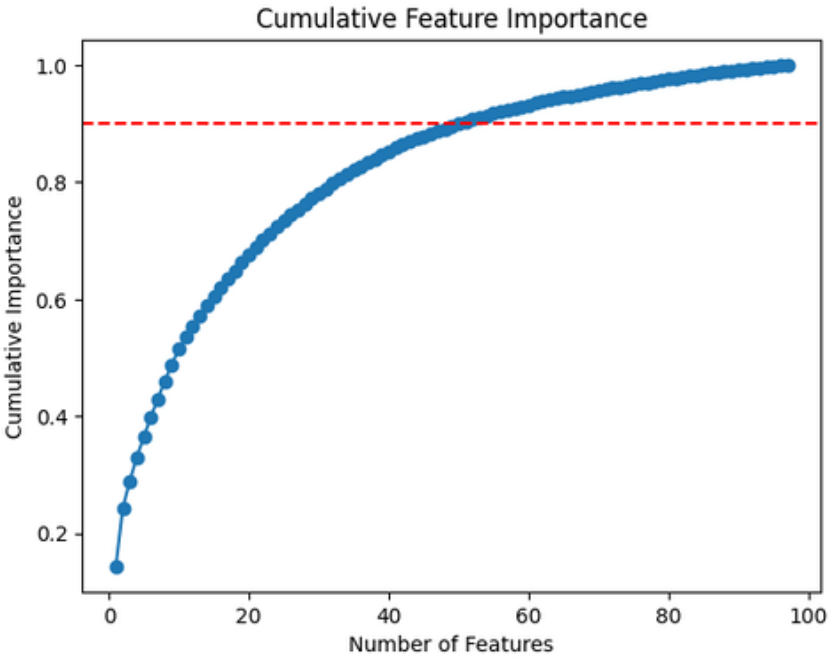


Figure 3. Cumulative Feature Importance Curve

The cumulative feature importance curve for the EBM model. This plot shows that the top 25 features contribute to over 80% of the model's predictive power, illustrating the efficiency of feature selection and the significance of key predictors in explaining variance in CVD mortality rates.

Delineating the purpose of these specific model choices, Geographically Weighted Regression (GWR) and its extension in Multiscale GWR (MGWR), capture how associations between SDHs and CVD risk vary across different locations. By estimating locally adaptive coefficients, these methods generate spatial heatmaps showing where certain factors—such as poverty or insurance coverage—exert stronger effects, making their outputs particularly straightforward to interpret for public health decision-makers. Geographically Weighted Artificial Neural Networks (GWANN) add another layer of flexibility by integrating spatial kernels (e.g., Gaussian or binary) into backpropagation, so that nearby counties carry greater weight during training. This approach not only learns intricate, nonlinear relationships but also reveals which locations are most influential for predicting CVD outcomes, thereby highlighting potential "hotspots.". Meanwhile, Gradient Boosted Decision Trees (GBDT) such as XGBoost refine multiple weak learners in an iterative ensemble and can directly incorporate geocoordinates as features, producing feature-importance scores and partial dependence plots that clarify how location and other SDHs jointly shape CVD risk. Random Forests (RF) complement GBDT by averaging many decision trees, thereby limiting overfitting and offering robust performance in smaller populations—especially for NHPI and AI/AN groups where sample sizes may be limited. Spatial weighting can further sharpen their local predictions and yield clear rankings of influential variables (e.g., healthcare access or pollution). Support Vector Machines (SVM), modified with geographic kernels, capture abrupt boundaries between low- and high-risk regions, enabling us to pinpoint where modest changes in socioeconomic thresholds produce large increases in disease burden. By combining GWR/MGWR, GWANN, GBDT, RF, and SVM, we achieve strong predictive accuracy along with interpretability: local coefficient maps (GWR/MGWR), feature-importance metrics (GBDT,

RF), and region-specific decision boundaries (SVM) together reveal why CVD risk clusters in particular counties and how best to address it.

C.2.2 Model Validation and Evaluation

To most rigorously assess the performance of our models, we will validate them through both k-fold cross-validation and with spatial cross-validation strategies. The purpose of the latter is to partition the data into geographically coherent units, such as contiguous sets of counties, to ensure the models are not inadvertently using neighboring locations in both training and testing. This ensures that spatial autocorrelation does not inflate model accuracy. To further reduce overfitting risks, we will employ nested cross-validation, to isolate hyperparameter tuning from final model evaluation. In order to do this, we will perform systematic Bayesian optimization. For each model from C2.2, we will define a tailored parameter space (covering items like learning rate, number of trees, kernel type, bandwidth, and regularization strength) and start with an initial set of random hyperparameter configurations. After each round of model training and validation, we will use a probabilistic surrogate function in the form of a Tree-structured Parzen-estimator (TPE) to model the likelihood of obtaining high v.s. low scores in different regions of hyperparameter space. After each run, this model will be updated to reflect observed performance, and we can propose a new set of hyperparameters that we predict will yield better results to quickly converge on model configurations that offer the best performance.

We will measure performance through a combination of conventional and spatially aware metrics. We will use Root Mean Squared Error (RMSE) and the coefficient of determination (R^2) to gauge overall predictive precision and variance explained. We will also use Adjusted R^2 controls for the number of predictors, ensuring we do not overfit the models simply by adding more features (given the high dimensionality of our data). In addition, we will use Spatial RMSE (sRMSE), which explicitly accounts for autocorrelation, to highlight geographic hotspots where the model being assessed may perform poorly. We will also dichotomize CVD mortality as high vs. low and compute AUC and ROC. Furthermore, beyond standard spatial cross-validation, we will employ a customized approach that withholds entire connected subgraphs from the node embeddings from C1.2 (such as the counties on Hawaii/Alaska or isolated pockets of counties with high data quality) to ensure that the learned embedding and the ML pipeline as a whole can generalize to new low density counties without requiring direct data overlap. We will compute the same metric as above for these clusters, and compare their performance with and without the node embedding to ensure that our models perform well in well sampled and sparse counties alike. Residual analysis across different states provides valuable insights into the geographic robustness of the EBM model. Figures 4 and 5 compare the residual distributions for California and Alaska, two states with distinct social determinants of health (SDHs) and demographic profiles. To be more specific, Residual analysis across different states provides valuable insights into the geographic robustness of the EBM model.

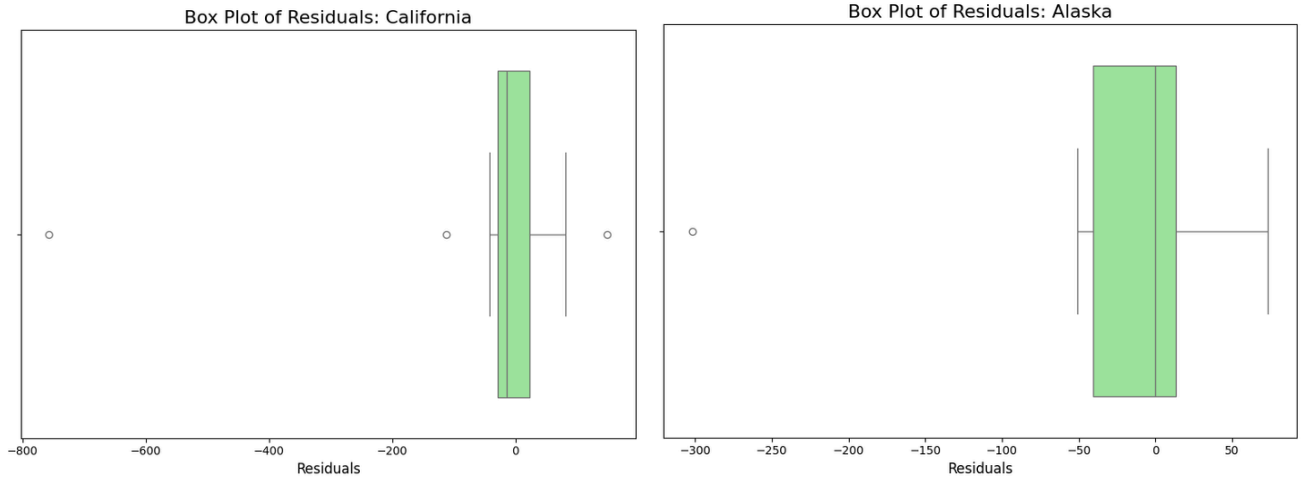


Figure 4 and 5. Box Plot of Residuals for California and Alaska

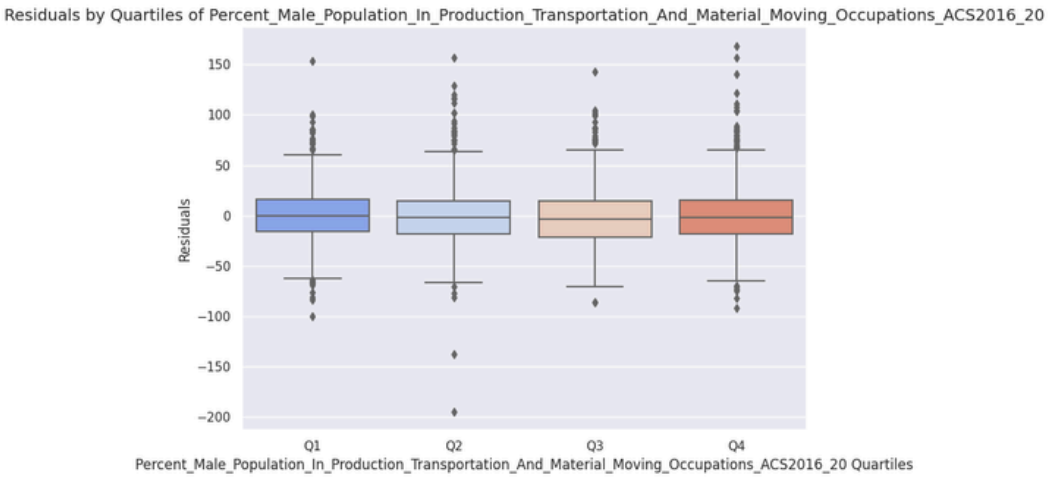
This box plot shows the distribution of residuals for the EBM model predictions of CVD mortality rates in California. While most residuals are clustered near zero, extreme outliers with large negative residuals (e.g., -800) indicate significant errors in certain counties, suggesting challenges in modeling unique regional factors.This box plot shows the residuals for Alaska, with a tighter range of residuals compared to California.

In California (Figure 4), while most residuals cluster near zero, several extreme outliers reveal significant prediction errors for specific counties. These large negative residuals, such as -800, suggest the model struggles to fully capture the diversity and complexity of SDHs in California. This aligns with earlier findings that the model performs poorly in areas with sparse or unique data characteristics. Such large deviations highlight the need for additional feature engineering or regional adjustments to improve model accuracy in this state.

In contrast, Alaska (Figure 5) exhibits a tighter distribution of residuals, with a smaller overall spread and fewer outliers. This suggests that the model performs better in predicting CVD mortality in Alaska, likely due to less variation in certain SDHs or more homogeneous data. However, one outlier with a residual of approximately -300 indicates some difficulty in specific counties, possibly due to limited data availability or unmodeled local factors.

Additionally, an essential aspect of robust model validation is evaluating residual behavior across key features to identify patterns and biases that may affect predictive accuracy. To this end, we analyzed residuals for the EBM model across quartiles of Percent Male Population in Production, Transportation, and Material Moving Occupations (ACS 2016–20), a significant socioeconomic predictor of cardiovascular disease (CVD) mortality. The results, shown in Figure 2, reveal that while the median residuals across quartiles are generally close to zero, the variability and the presence of extreme outliers increase in higher quartiles (e.g., Q4). This suggests that the model struggles to accurately capture the effects of this socioeconomic factor, particularly in regions or subpopulations with higher proportions of males in these occupations.

Figure 6. Residuals by quartiles of Percent Male Population in Production, Transportation, and Material Moving Occupations (ACS2016–20). The box plot illustrates the distribution of residuals for the Explainable Boosting Machine (EBM) model across quartiles of this feature. While the median residual remains close to zero, there is noticeable variance and extreme outliers, particularly in the highest quartile (Q4). This highlights the need for improved handling of socioeconomic predictors to reduce error magnitudes for specific subgroups.



C.2.3 Implementation, Scalability, and Reproducibility

To ensure that we can successfully deploy our strategy for building/training our geospatial machine learning models, we will use Indiana University's BigRed200 and Quartz Supercomputing Clusters, as well as Google Cloud Platform (GCP), for use as high-performance computing (HPC) environments. These provide us with the necessary space/computing resources to store the above data and train the models in an efficient manner. To manage our data processing and model training workflow, we will use Snakemake, which will allow us to easily run jobs across multiple computing nodes and as a whole optimize our use of HPC resources. Additionally, to maintain consistency between the HPC environments, as well as with local computing environments, we will containerize our application using Docker, to help ensure a standard set of dependencies and software versions across the environment, and to enable collaboration of different researchers working on this project effectively. Also for version control and collaborative development, we will employ Git, maintaining all code, scripts, and documentation within a centralized Git repository on Github.. By integrating Git with our workflow manager, we ensure that every iteration of our models is documented and reproducible, allowing for easy rollback and auditing of changes.

C.2.4 Race-Agnostic Training and Validation

Our modeling strategy consciously avoids using race or ethnicity as predictive variables to prevent perpetuating biases against NHPI, AI/AN, or other underserved communities. Instead, to ensure that the identification of SDH factors influencing CVD risk is unbiased and equitable across different racial groups, we base predictions on core structural factors such as socioeconomic status, environmental exposures, and healthcare accessibility. Nonetheless, we will conduct subgroup analyses after training each race-agnostic model, calculating the same RMSE, sRMSE, R^2, AUC, and ROC metrics for CVD mortality data for these groups, specifically focusing on NHPI and AI/AN. If we observe systematic underperformance for these subgroups, we will revisit kernel bandwidths, refine the choice of geospatial or socioeconomic predictors, or incorporate fairness-centric regularization techniques to address any disparities.

C.3 Explainable AI (XAI) for Interpreting CVD Risk Factors

The objective of Aim 2 is to leverage Explainable AI (XAI) techniques to interpret and elucidate the influence of SDH factors on CVD mortality at a U.S. county level. While Aim 1 emphasizes design robust, geospatially-aware ML models to model this relationship, Aim 2 independently aims to analyze the predictions from the models, as well as the dataset itself, to uncover patterns/relationships in and between the SDHs and CVD mortality, and create an interpretable framework for understanding the SDH factors and their implications for CVD, especially for underserved populations such as as NHPI and AI/AN.

C.3.1 SHAP and LIME for SDH Importance

To interpret the models generated within Aim 1, we will mainly employ Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). SHAP will provide a global view of feature importance, which will allow us to decompose the predictions of the models into contributions from individual SDH factors, allowing us to understand the important features for further analysis.

Figure 7. SHAP violin plot for all-cause CVD mortality prediction per 100,000, illustrating the distribution of feature impacts on model output. The width of each "violin" represents the density of SHAP values for different levels of each feature, with high values in pink and low values in blue. Features like "Percent Adults 20yrs And Over Physical Inactivity" and "Amount of SNAP Benefits Per Capita" have substantial influence on the prediction, with their specific values pushing mortality rates higher or lower. This plot highlights both the strength and direction of each feature's impact on predicted mortality.

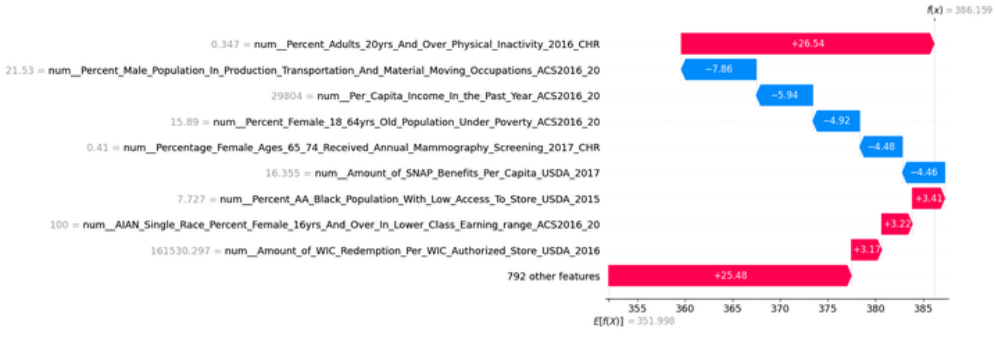


Figure 8. SHAP summary plot for all-cause CVD mortality prediction per 100,000, stratified by feature importance, using an XGBoost model. Red bars indicate features that increase mortality predictions, while blue bars decrease them. Specific feature values, such as a high "Amount of SNAP Benefits" or "Percentage of AA/Black Population with Low Access to Store," show substantial influence on the prediction. The bottom axis represents the model's base prediction $E[f(x)]$, with SHAP values adjusting this base to reach the final predicted mortality rate.

Beyond SHAP, we will also seek to understand feature importance more locally, within specific counties or clusters/regions. Furthermore, in our explicitly geographically weighted models, we will apply geoSHAP, to incorporate spatial autocorrelation and heterogeneity between counties into the interpretation of the most important features for the models. This will allow us to understand how the impact of different SDHs differs across domain-specific groups, such as states, the contiguous U.S. versus states like Hawai'i and Alaska, and other similar relationships. To complement the use of geoSHAP, we will compute Local Interpretable Model-agnostic Explanations (LIME) to explain why specific counties or regions are at higher or lower CVD risk. LIME excels at explaining individual predictions by approximating the model locally with a simpler, interpretable model. Suppose a model predicts high CVD risk for a rural NHPI-dense county. LIME might identify "lack of healthcare facilities" and "high unemployment rates" as the key contributors for this specific prediction. geoSHAP could then show that "lack of healthcare facilities" is a critical factor across a cluster of rural counties, indicating a regional issue rather than an isolated problem. Together, these insights provide a complete picture: LIME identifies the immediate drivers for the specific county, while geoSHAP contextualizes these findings within the broader geographic landscape. As a whole, we will be able to curate a list of the most important SDH variables that drive CVD mortality across dif we ferent models and regions.

Finally, to move beyond prediction-focused interpretation, we will apply interpretability tools like SHAP directly to the raw dataset, integrating them with unsupervised clustering and exploratory models. This approach allows us to uncover latent patterns and relationships in the data without relying on predictive model outputs. For example, SHAP values can be calculated on clusters identified through k-means or hierarchical clustering, revealing the most influential SDH variables driving the formation of these clusters. Combining this with methods like random forest feature importance, we can better understand which SDH factors differentiate clusters, such as regions with high CVD risk versus low risk.

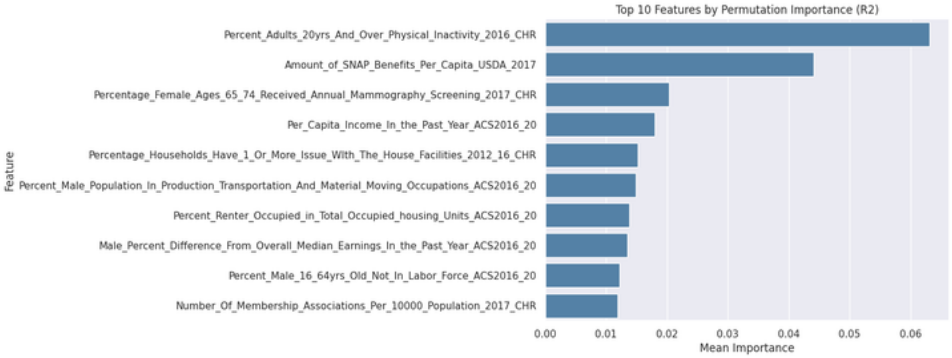
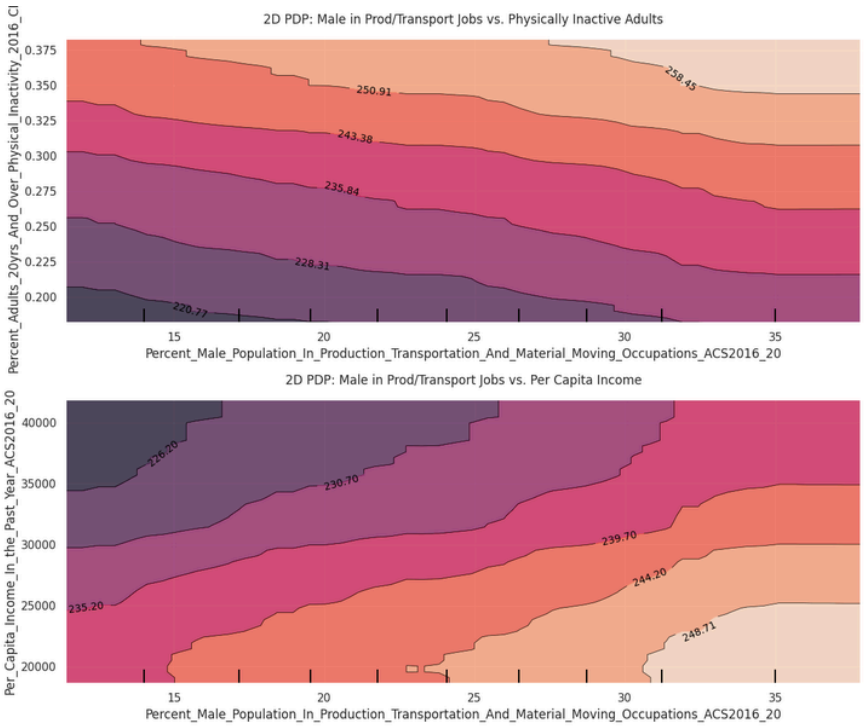


Figure 9. Top 10 Features by Permutation Importance (R^2)
Bar plot ranking the top 10 most important features for the EBM model based on permutation importance (R^2). Features like the percentage of physically inactive adults (20 years and older, CHR2016) and per capita income (ACS2016–20) significantly influence predictions of CVD mortality rates, emphasizing the role of both health behaviors and socioeconomic conditions.



C.3.2 Understanding Interaction Effects and Complex Dependencies of CVD Geospatially

Building on the important features identified in C.3.1, we will use Partial Dependence Plots (PDPs), Individual Conditional Expectation (ICE) plots, and Accumulated Local Effects (ALE) plots to explore their marginal effects, interactions, and thresholds. PDPs will illustrate how average changes in critical SDH factors, such as healthcare access, impact CVD risk. ICE plots will add granularity by capturing variability in these effects across counties, highlighting geographic nuances. ALE plots will disentangle correlated SDH variables, such as average income and

average educational attainment, to uncover their joint contributions to health disparities. The two-dimensional partial dependence plots (PDPs) in Figure 3, derived from the Explainable Boosting Machine (EBM) model, provide a clear visualization of how key predictors interact. For example, the top plot illustrates the combined effects of the percentage of males in production, transportation, and material moving occupations and the percentage of physically inactive adults (20 years and older). The contours of predicted CVD mortality rates reveal that physical inactivity amplifies the effects of high occupational exposure, suggesting a compounding risk in regions with high prevalence of both factors.

Figure 10: Two-dimensional Partial Dependence Plots (PDPs):

(Top) PDP showing the interaction between the percentage of males in production, transportation, and material moving occupations (ACS2016–20) and physically inactive adults (20 years and older, CHR2016). The contour lines depict the predicted cardiovascular disease (CVD) mortality rates for different combinations of these two features, with darker shades indicating lower mortality rates.

(Bottom) PDP illustrating the interaction between the same percentage of males in production-related jobs and per capita income (ACS2016–20). This plot highlights how socioeconomic disparities interact with occupational data to influence predicted CVD mortality rates.

ALE can further clarify how healthcare access interacts with other factors like income inequality. Also, for instance, we will analyze how access to healthy food options influences CVD risk differently in urban versus rural counties that have significant NHPI and AI/AN populations. ICE plots will allow us to capture geographic variability in feature importance, highlighting how the same SDH factor may have different impacts on CVD risk in various spatial contexts. This method is particularly useful for identifying regions where certain SDH factors have a stronger or weaker influence on health outcomes.

In addition, we will employ ALE plots to analyze complex feature interactions by accounting for interactions between multiple SDH features, especially when features are correlated. For example, we will examine how the combination of air pollution levels and access to healthcare services jointly affects CVD risk across different regions. ALE plots will help us uncover subtle interactions between SDH factors that might not be immediately apparent. We may identify compounded effects, such as how environmental hazards and limited healthcare access together disproportionately increase CVD risk in economically disadvantaged areas.

By understanding these complex dependencies, we will model how combinations of SDH factors produce unique risk profiles for NHPI and AI/AN communities in our dataset. The insights from ICE and ALE plots will enable us to suggest interventions tailored to the unique combinations of risk factors present in different regions.

C.3.3 Dynamic Temporal and Subgroup-Specific Analysis of SDH and CVD Risk in 4D

To capture the full complexity of the relationship between the SDHs and CVD risk, especially pursuant to underserved communities and groups, we will conduct dynamic temporal and subgroup-specific analyses using a four-dimensional (4D) framework. This approach considers the spatial, temporal, demographic, and health outcome dimensions simultaneously.

While not all variables in our dataset are time-weighted, some are split into distinct time windows, such as 2016–2018 and 2018–2020. For example, healthcare facility density or insurance enrollment data may show significant shifts between these periods. To explore temporal dynamics, we will employ dynamic time warping (DTW) to align trends in variables across counties, even when temporal patterns vary. This method is particularly useful for comparing SDH trajectories in counties with similar initial conditions but diverging CVD outcomes.

Additionally, we will use state-space modeling to track transitions between high- and low-risk states over time, stratified by demographic subgroups. Using subgroup specific factors, such as NHPI housing ownership rates, for NHPI and AI/AN populations will allow us to understand distinct patterns of risk transitions. This will allow us to identify thresholds, both for more broad variables, and for targeted variables like the above, necessary to shaft counties from high to low CVD risk.

Also, we will use policy scenario projection graphs to model the potential outcomes of implementing specific interventions, such as increased educational funding or group-specific opportunities by way of distributed lag models (DLMs) and system dynamic modeling. These simulations will account for temporal delays in policy impacts, such as insurance enrollment in 2016 leading to reduced CVD risk by 2020. For each scenario, we will estimate the magnitude of CVD risk reduction over time, stratified by counties and subgroups. More specifically, we will compare scenarios where healthcare funding is increased broadly versus scenarios targeting NHPI- or AI/AN-dense regions to determine where interventions yield the greatest impact.

Additionally, interaction effects between policies, such as healthcare access improvements combined with income inequality reductions, will be modeled to uncover synergistic effects.
We will also incorporate temporal dispersion heatmaps to visualize variability in SDH variables over time, highlighting regions or subgroups experiencing unstable or rapidly changing conditions. For each SDH variable, we will calculate metrics such as standard deviation, coefficient of variation, and range within time windows to identify high-dispersion regions. For example, counties with significant fluctuations in healthcare access may exhibit inconsistent CVD risk reductions, signaling systemic instability. By correlating dispersion metrics with CVD outcomes, we can identify regions requiring stabilization efforts.

C.3.3 Visualizing Geospatial Predictions and Risk Factors

A pivotal component of our analysis plan is translating model predictions and XAI outputs into geospatial visualizations that are easily interpretable by public health officials and stakeholders. These visualizations will facilitate data-driven decision-making and the implementation of targeted interventions. We will develop geospatial risk maps highlighting areas with the highest CVD risks for NHPI and AI/AN populations. These maps will be overlaid with heatmaps of significant SDH variables identified through SHAP and LIME analyses, providing a clear depiction of geographic disparities. Using popular GIS tools and spatial analysis libraries such as QGIS and ArcGIS, we will create layered maps that integrate CVD risk predictions with SDH variables. Additionally, interactive maps will be developed iteratively, allowing users to explore how different SDH factors influence disease risk across various counties and states. These interactive features will empower stakeholders to pinpoint areas where community interventions—such as improving healthcare access or mitigating environmental risk factors—can have the most significant impact on reducing CVD burden, particularly among NHPI and AI/AN populations. In addition to traditional geospatial maps, we will create graph-based visualizations that illustrate the influence of neighboring counties and key SDH factors on each county's CVD risk. These visualizations will highlight clusters of high-risk areas and the interdependencies contributing to these patterns, providing a comprehensive view of spatial health disparities. Ultimately, these visual tools will assist decision-makers and public health/policy officials in identifying and prioritizing areas for intervention, ensuring that resources are allocated effectively to address the most pressing health disparities among NHPI and AI/AN populations.

C.4 Limitations and Alternative Approaches

Despite the comprehensive nature of our study, we acknowledge several limitations of the study, and propose alternative approaches to address them. One critical limitation is the quality and completeness of data, particularly for counties with large NHPI and AI/AN populations where missingness is present. To address this, we, as aforementioned, have chosen to use the GraphSAGE algorithm to learn embeddings that strengthen the data for low-data counties near ones with better data, as well as imputation techniques within the data pre-training. Another limitation of this study is the aggregation of data at the county level, which may obscure intra-country variation and neighborhood level differences. In another study, utilizing more granular spatial units, such as census tracts or block groups, could provide more granular insights. In spite of these limitations,, given the scale of data we have collected as well as the utility of assigning risk factors at constituent units like counties, (making them more addressable) both make this study worthwhile. The dynamic nature of SDHs further complicates our analyses due to their evolving nature. To capture this, we have employed dynamic lag models and temporal dispersion heatmaps, which allows us to discover regions where conditions are rapidly changing, and directly incorporate these into our modeling.

C.5 GANTT Chart

Figure 3 presents a five-year roadmap aligning our project's two overarching aims to time-specific milestones and deliverables. In Year 1, we will start Aim 1 centers by data cleaning, feature selection, and establishing the foundational dataset needed for modeling. These efforts (shown in the beige/orange cells) ensure the data's integrity and will provide a strong basis for developing geospatially aware ML models. During Year 2 and the first half of Year 3, we train and validate these ML models.

Concurrently, Aim 2 begins at the end of Year 1, focusing on explainable AI (XAI) analyses of social determinants of health (SDHs). Work in Aim 2 (highlighted in blue) extends through Years 3 and 4, as we apply advanced XAI methods to identify influential SDH factors driving cardiovascular disease (CVD) outcomes at the county level. In the later stages of Year 4 and Year 5, the team conducts deeper investigations of region-specific risk interactions—such as assessing how various SDH variables synergize to affect local CVD burdens—and develops user-friendly visualizations and dashboards for policymakers and stakeholders.

Overall, this structured timeline ensures synergy between dataset preparation, geospatial modeling, and the interpretability components required to translate research findings into actionable public health insights.

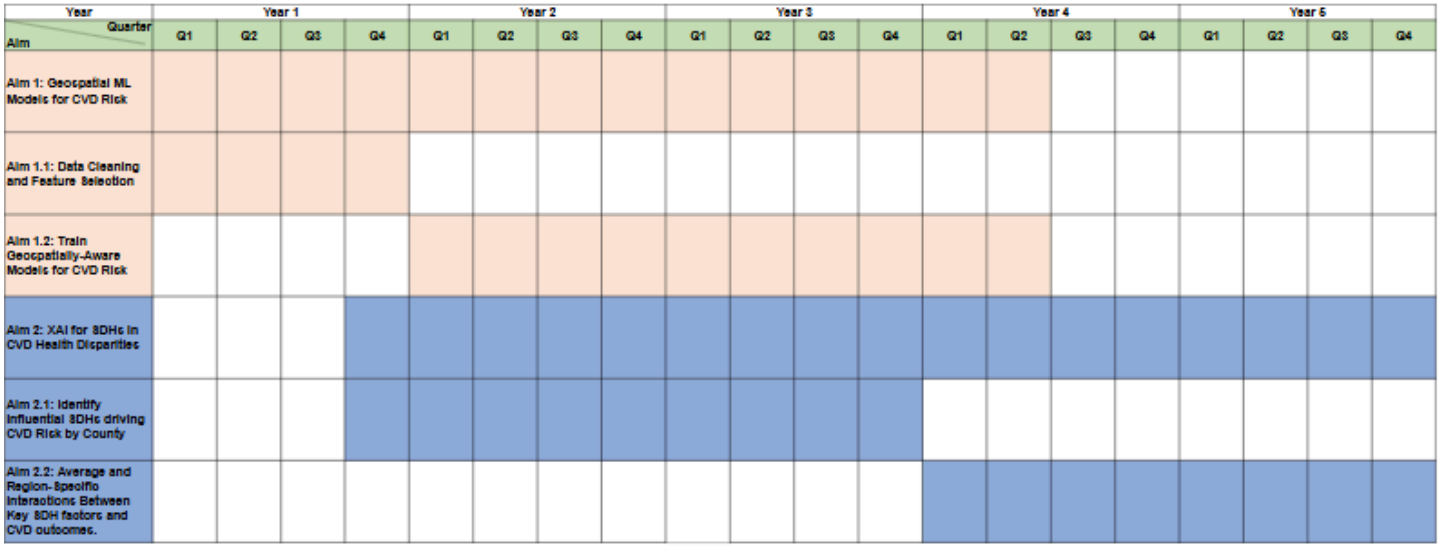


Figure 11. GANTT chart of the Specific Aims.