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. Despite its importance, significant research gaps persist, particularly concerning Native Hawaiian and Pacific Islander (NHPI) populations and Alaska Native and American Indian (AI/AN). These populations experience unique health disparities that are often obscured when aggregated into broader racial categories (Waitzfelder et al., 2023). The distinct genetic, environmental, and social determinants of health (SDHs) affecting these communities necessitate more comprehensive geospatial health research to understand the source of their CVD disparities.

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 unique health disparities faced by these communities (Office of Minority Health, 2023). Many studies underscore the importance of disaggregating social, genetic, and environmental data for these groups to reveal critical health disparities that are otherwise hidden (Espey et al., 2014). This aggregation prevents targeted health interventions and diminishes the efficacy of precision medicine approaches, which consider both biological and social determinants of health essential for addressing the specific needs of NHPI and AI/AN populations (Khoury et al., 2022). Socioeconomic factors 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 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 racial groups. 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 interventions that acknowledge and cater to the unique challenges faced by NHPI and AI/AN communities (National Institutes of Health, 2024).

2) Cardiovascular disease (CVD) remains a significant public health concern. Cardiovascular disease (CVD) remains a significant public health concern, particularly for Native Hawaiian and Pacific Islander (NHPI) and Alaska Native and American Indian (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 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 knowledge in both NHPI and AI/AN populations. Less than 44.4% of NHPI adults possess the recommended knowledge for recognizing and responding to heart attack symptoms, contributing to poorer health outcomes. Similarly, AI/AN populations often have limited awareness of CVD symptoms and risk factors, which may delay seeking medical attention (Felix et al., 2019; Galloway, 2005). Furthermore, NHPI and AI/AN individuals are more likely to die from heart attacks compared to NHWs, emphasizing the critical need for targeted health interventions.

3) Geospatial analysis can fulfill an unmet role in cardiovascular research. Many social determinants of health (SDHs) that impact cardiovascular disease (CVD) vary significantly across 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 techniques

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.

Native Hawaiian and Pacific Islander (NHPI) and Alaska Native and American Indian (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 the well-documented prevalence of CVD risk factors in these communities, there has been limited application of advanced spatial analysis and machine learning techniques to understand how these risks manifest in different geographic contexts. By utilizing SDH and CVD data at the county level and applying sophisticated geospatial machine learning models, our study will fill this critical research gap.

A.2 Background

A.2.1 Cardiovascular Disease

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 the 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, 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 Racial/Ethnic Differences in CVD Outcomes

Cardiovascular disease, and metabolic comorbidities like type II diabetes (T2D), obesity, and dyslipidemia, exhibit significant disparities across different racial and ethnic groups due to a complex interplay of genetic, environmental, and socioeconomic factors. AI/AN populations face disproportionately high rates of cardiometabolic diseases. Studies indicate that AI/AN individuals have higher prevalence rates of T2D, obesity, and hypertension compared to NHWs. These disparities are attributed to factors 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 competent healthcare services exacerbates these disparities (Aluli et al., 2007; Aluli et al., 2010). Understanding these disparities programmatically requires disaggregated data. 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 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 may 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 generated interpretable insights (Ribeiro et al., 2016; Doshi-Velez & Kim, 2017).

A.2.4 Intersection of ML, AI, and Geospatial Health Research

The convergence of machine learning (ML), artificial intelligence (AI), and geospatial health research offers unprecedented opportunities to address cardiovascular disease (CVD) disparities in underserved populations. By incorporating geospatial data—such as socioeconomic status, healthcare access, and environmental exposures—into ML and AI 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 interventions (Fotheringham et al., 2002; Guo & Matisziw, 2021). This spatially weighted approach not only enhances the predictive power of models but also facilitates the creation of tailored public health strategies that address the unique needs of Native Hawaiian and Pacific Islander (NHPI), American Indian and Alaska Native (AI/AN), and multiracial communities, ultimately contributing to the reduction of CVD disparities (Wang, 2020; Krieger, 2012).

B. Innovation

This project introduces innovative approaches to understanding health disparities among Native Hawaiian and Pacific Islander (NHPI) populations through advanced geospatial and machine learning (ML) techniques, with key innovations as follows:

1) Race-Agnostic Analysis for Equitable Health Insights in NHPI & AI/AN

Our models are explicitly designed to be race-agnostic, ensuring that race is not an influencing factor in the identification of key predictors. This approach allows us to focus purely on the impact of SDH variables, reducing racial bias and promoting equitable analyses that identify the true underlying drivers of health disparities. This innovation directly supports health equity and aligns with national efforts to promote fairness in healthcare research.

2) Advanced Geospatial Machine Learning Models: The use of cutting-edge geospatial ML models, such as Geographically Weighted Artificial Neural Networks (GWANN), allows for precise modeling of spatial dependencies in our SDH data and geographically tailored risk predictions, accounting for regional variations in disease burden.

3) Integration of Explainable AI (XAI): By integrating XAI techniques into our analysis, we enhance interpretability in ways that traditional models cannot achieve. Our approach provides clear, actionable insights into the specific SDH factors influencing CVD risk, empowering policymakers and public health officials to develop targeted interventions and drive meaningful change in underserved, high-risk communities.

4) Leveraging Secondary Datasets for Novel Insights. Utilizing an array of publicly available secondary datasets (e.g., American Community Survey, CDC, USDA), we are breaking new ground by extracting deep insights into the interplay between SDH and CVD outcomes. By conducting secondary analysis across thousands of SDH variables, we are transforming these standard datasets into a powerful toolset for understanding and addressing health disparities in historically underrepresented populations.

C. Approach

C.1.1 Data Collection

Our initial step involved the systematic collection of Social Determinants of Health (SDH) data from 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 provide crucial SDH indicators such as household income, health insurance coverage, food security, housing stability, and healthcare access, which are essential for understanding and addressing cardiovascular disease (CVD) disparities in NHPI and AI/AN populations. Specifically, data from the CDC's SVI (2016–2020) will allow us to assess socioeconomic status at the county level, while the USDA Food Environment Atlas will offer 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 will provide critical information on the availability of primary care providers and healthcare infrastructure at a geospatial resolution. The below table provides a comprehensive list of these datasets, as well as the source of the data.

Table 1: Table of collected surveys and their sources.

Survey	Source
ACS 5-Year Estimates	U.S. Census Bureau
Food Environment Atlas	USDA
Health Resources & Services Administration (HRSA)	HRSA
Atlas of Heart Disease and Stroke	CDC
Social Vulnerability Index (SVI)	CDC/ATSDR
Environmental Justice Index (EJI)	CDC
Robert Wood Johnson Foundation's County Health Rankings	RWJF
National Environmental Public Health Tracking	CDC
Walkability Index	EPA
Community Resilience Estimates	U.S. Census Bureau
Medicare Data	CMS
PLACES: SDOH Measures for County ACS	CDC
National Risk Index	CDC/ATSDR
Area Deprivation Index (ADI)	University of Wisconsin
Small Area Income and Poverty Estimates (SAIPE)	U.S. Census Bureau
Environmental Public Health Tracking (Air Pollution Data)	CDC
Farmers' Market SNAP Participation	USDA
Maternal and Child Health Provider Availability	HRSA
National Healthcare Quality and Disparities Report	AHRQ
Spatiotemporal Cardiovascular Disease Mortality Analysis	Multiple Sources
Public School Accessibility	CDC
Community Resilience Estimates for Equity	U.S. Census Bureau

C.1.2 Data Imputation, Standardization, and Cleaning

In order to do geospatially-aware ML and XAI on this data, we must both concatenate the disparate sources into an interoperable dataset, and ensure the dataset is properly cleaned and standardized to generate accurate and powerful models.Given the focus on NHPI and AI/AN groups, we systematically identified and extracted columns pertaining to these demographics by filtering for keywords such as "Asian," "AI/AN," and "NHPI". This targeted selection ensured that our analysis specifically addressed the populations of interest, and allowed us to search the full set of datasets for useful targets. To concentrate our analysis on CVD outcomes, we employed pattern matching techniques to identify columns related to it by searching for the term "CVD" within column names. This step isolated relevant health outcome variables essential for our predictive modeling, allowing us to streamline our dataset and emphasize variables directly linked to CVD outcomes. Expanding beyond CVD-specific keywords, we identified a broader set of VCD-related columns by searching for additional keywords indicative of diseases, conditions, and health outcomes, such as "disease," "condition," "disorder," "illness," "syndrome,", "prevalence," "morbidity," "mortality," "death," and "outcome." This

comprehensive filtering excluded any columns containing "single_race" to avoid redundancy and ensure a focus on multi-racial and diverse population data, resulting in a robust subset for analyzing various health outcomes in relation to SDH. To assess the significance of numerical variables and reduce dimensionality, we conducted a variance analysis on all numerical columns, prioritizing variables with higher variance as they contribute more information to the model. Subsequently, we applied Principal Component Analysis (PCA) to capture at least 95% of the total variance, identifying principal components that summarize the underlying patterns in the data. By examining the PCA loadings, we determined the top contributing variables, aiding in feature selection and interpretation. This dimensionality reduction enhanced the efficiency and performance of our machine learning models by eliminating redundant and less informative variables.Focusing on total population and CVD death rates, we extracted all relevant columns and consolidated them into a separate data frame. To identify potential multicollinearity issues, we computed the correlation matrix for these numerical columns and filtered for pairs exhibiting high correlation (absolute correlation coefficient greater than 0.6 but less than 1.0). High collinearity among predictors can undermine the stability and interpretability of machine learning models, necessitating corrective measures. Identifying these highly correlated pairs allowed us to address redundancy and improve the robustness of our models. Additionally, we employed Factor Analysis to reduce dimensionality while retaining the underlying variance in the data. By transforming collinear variables into uncorrelated factors, we enhanced the robustness of our predictive models. This transformation not only mitigated the issues associated with multicollinearity but also facilitated a more interpretable representation of the data, capturing the essential patterns without the noise of redundant information. Following Factor Analysis, we performed additional correlation checks to ensure that the newly created factor variables did not introduce new collinearity issues, maintaining the integrity and usability of the dataset for subsequent modeling stages.

C.1.3 Geospatially-Aware Machine Learning-Ready Data for Public Health Insights

Once the data has been fully developed and cleaned , it will serve as the foundation for advanced geospatial and machine learning analyses aimed at investigating CVD risks in NHPI/AI/AN populations. The constructed nature will allow the integration of machine learning models that can leverage spatial relationships between SDH factors and health outcomes.

We will also make the cleaned and collected data available to researchers and public health officials, allowing for consistent data driven insights by the scientific community and public at large.

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

Aim 1 focuses on developing and applying geospatial machine learning (ML) models capable of accurately predicting CVD (CVD) risk by integrating complex interactions between social determinants of health (SDH) and geographic factors, particularly for Native Hawaiian and Pacific Islander (NHPI) and Alaska Native and American Indian (AI/AN) populations. This aim involves selecting appropriate ML models, optimizing their performance through spatial weighting and autocorrelation measures, and ensuring robust validation techniques to achieve generalizability and interpretability.

C.2.1 Machine Learning Model Selection and Feature Engineering

Selecting suitable machine learning models is crucial for unraveling the spatial variability and its interplay with SDH factors that contribute to CVD outcomes. Our analysis plan employs a diverse array of linear and nonlinear models to capture both straightforward and intricate relationships within the data, ensuring that we can identify the most effective models for our specific research questions while maintaining high levels of interpretability. We will implement traditional models such as Geographically Weighted Regression (GWR) and its advanced variant, Multiscale GWR (MGWR), alongside sophisticated models like Geographically Weighted Artificial Neural Networks (GWANN), Gradient Boosted Decision Trees (GBDT), Random Forest (RF), and Support Vector Machines (SVM). These models are adept at handling the high-dimensional, complex, and nonlinear relationships inherent our geospatial health data.

- Geographically Weighted Artificial Neural Networks (GWANN) introduce flexibility by allowing neural networks to be trained with spatial weighting functions that account for geographic proximity. This allows the model to adapt its weight matrices according to the spatial distribution of the data, meaning nearby counties exert more influence on the predictions than distant ones. We will implement GWANN using both traditional backpropagation algorithms and newer, more efficient optimization techniques such as Adam and Stochastic Gradient Descent (SGD). GWANN’s architecture will include multiple

hidden layers to capture intricate, nonlinear relationships between SDH factors and CVD risk at a county level.

- Gradient Boosted Decision Trees (GBDT) will be employed due to their robustness in handling complex, non-linear interactions between features. GBDT's ensemble nature—where weak learners (decision trees) are iteratively combined to minimize errors—makes it ideal for datasets with mixed feature types (categorical, continuous) and varying spatial dependencies. In GBDT, we will incorporate geographic coordinates (latitude and longitude) of the counties directly as features, along with our set of other SDHs, to capture the interaction between county and CVD outcomes. We will apply gradient boosting techniques such as XGBoost, LightGBM, and CatBoost.
- Random Forests (RFs) will complement GBDTs by providing insights into feature importance and enhancing our understanding of how different variables contribute to disease risk across counties . The bagging approach inherent in RFs reduces overfitting, making them ideal for predicting outcomes in smaller populations, such as NHPI and AI/AN communities. We will integrate spatial weighting into RFs to ensure that predictions are influenced appropriately by the spatial proximity of counties.
- Support Vector Machines (SVM) enhanced with geographic weighting, will be employed to model the complex boundaries between low and high-risk regions (either counties themselves or groups of counties). Their capability to handle both linear and nonlinear classification tasks will enable us to effectively model CVD risk as a function of spatial and socio-economic disparities. We will experiment with different kernel functions, including linear and radial basis function (RBF) kernels, optimizing them for spatial data to balance interpretability and computational efficiency.
-

Additionally, for use in our ML models, we will perform feature engineering on our data. Feature engineering is a crucial process in machine learning, involving the creation and transformation of raw data into meaningful input variables (features) that improve model performance. Prior to model integration, we will standardize continuous SDH variables (such as poverty rates per 100,000) using z-scores, apply one-hot encoding to categorical variables (such as constructed variables like environmental quality) , and perform log transformations on skewed distributions to enhance interpretability and model performance. Interaction terms will also be engineered to capture complex relationships between features, enabling models to recognize how combinations of SDHs jointly affect health outcomes. Given the high dimensionality and potential sparsity of our dataset, given that we have similar numbers of SDHs and counties, we will iteratively evaluate and refine our feature engineering processes to ensure optimal model performance and interpretability. This iterative process acknowledges the need to experiment with various feature sets and transformations to identify the most informative and interpretable features for our models. Our feature set will encompass spatial, social, and environmental determinants of CVD health, including geographic coordinates of the counties, spatial adjacency, income inequality, healthcare access, education, food security, and environmental factors such as air quality and green space availability.

For each county, we will derive latitude and longitude to capture absolute geographic positioning, which is crucial for evaluating how location affects access to resources and exposure to environmental risks. Spatial adjacency will be incorporated to understand how neighboring counties influence each other's health outcomes, utilizing measures like spatial weights matrices and Moran's I to capture these dependencies. Income inequality will be quantified using metrics such as the Gini coefficient and household income distribution, capturing structural economic challenges across counties. Healthcare access will be assessed through variables related to health insurance coverage and provider density (e.g., number of primary care doctors per 1,000 people), providing insights into the ease or difficulty populations face in accessing healthcare—an essential determinant of health outcomes. Educational attainment and food security will serve as proxies for broader socio-economic stability and the ability to access food, respectively. Environmental factors, including air quality and green space availability, will be generated using data from the Environmental Protection Agency (EPA), focusing on pollutants like Particulate Matter (PM2.5) and Ozone (O3). Additionally, we will develop features for industrial pollution exposure and access to green spaces, reflecting the overall environmental quality of each county.

C.2.2 Spatial Autocorrelation, Weighting, and Geospatial Optimization Techniques

To effectively account for spatial dependencies in CVD risk, we will explore and optimize various spatial weighting schemes, including Gaussian Kernel, Binary Kernel, and Mixed Gaussian-Binary Kernel. Each kernel will be tailored through cross-validation to determine the most appropriate bandwidth for our models. We will

employ both fixed and adaptive bandwidths, where the number of nearest neighbors remains constant while geographic distance varies based on population density and socioeconomic factors. Given the potential sparsity in our graph structures, we will implement strategies such as graph densification by incorporating domain knowledge to add meaningful edges. Additionally, we will utilize graph embedding techniques like GraphSAGE and node2vec to generate dense vector representations of nodes, facilitating better learning from limited relational data. Hierarchical graph structures will also be explored to capture multi-scale spatial patterns, enhancing our models' ability to interpret complex geographic relationships. Leveraging the flexibility of our ML models, particularly GWANNs, we will implement adaptive spatial weighting schemes that dynamically adjust based on local population densities and socio-economic factors. This ensures that the influence of neighboring counties is accurately represented in the model, even in regions with sparse connectivity. Adaptive weighting allows the model to prioritize more relevant neighbors (i.e. close together counties), enhancing its ability to capture essential spatial dependencies without being overwhelmed by irrelevant or noisy connections. To further address sparsity, we will utilize advanced graph embedding techniques that capture the latent structure of the graph in a lower-dimensional space. Techniques such as GraphSAGE and node2vec will be employed to generate dense vector representations of nodes, effectively summarizing the graph's structural and feature-based information. These embeddings facilitate more effective learning by the neural network models, allowing them to leverage the enriched relational information despite the original sparsity of the graph.By integrating these spatial weighting and optimization techniques, we aim to refine our models' ability to accurately predict CVD risks across diverse geographic contexts. This approach ensures that our models can effectively leverage spatial dependencies, even in high-dimensional and sparse datasets, thereby enhancing predictive accuracy and robustness.

C.2.3 Model Validation and Evaluation

We will ensure the validity and robustness of our models through a rigorous validation process that includes both k-fold cross-validation and spatial cross-validation. Spatial cross-validation will be particularly important in our context as it accounts for the geographic structure of the data and ensures that the models do not overfit to specific regions. This is achieved by splitting the data into geographically coherent training and testing sets, ensuring that entire geographic regions are left out during training and then used for validation. To address the risk of overfitting in our complex machine learning models, particularly given the smaller sample sizes of NHPI and AI/An populations, we will incorporate regularization techniques such as L1 (Lasso) and L2 (Ridge) regularization in our regression models. For tree-based models like GBDT and RF, we will implement pruning strategies, set constraints on tree depth, and adjust minimum sample leaf sizes. Cross-validation methods, including nested k-fold and spatial cross-validation, will be used not only for model evaluation but also for hyperparameter tuning. These steps will help ensure that our models maintain high predictive performance while generalizing well to unseen data.

We will evaluate the models using a combination of performance metrics, including:

- Root Mean Squared Error (RMSE): To measure prediction error.
- R^2 and Adjusted R^2: To assess model fit, particularly the proportion of variance in disease risk explained by the model
- Spatial RMSE (sRMSE): An extension of RMSE that accounts for spatial dependencies by adjusting for the spatial structure of the data.

In addition to the entirety of the dataset, these metrics will be applied across different racial and ethnic subgroups to ensure the model generalizes well across populations, particularly for NHPI/AI/AN groups. For advanced models like GWANNs, we will conduct adversarial robustness tests by introducing controlled perturbations to the input data. This will assess the models' resilience to spatial noise and missing data, ensuring reliable performance even under suboptimal data conditions. Also, given the complexity of our data and the multitude of modeling approaches, we will adopt an iterative process of model training, evaluation, and refinement. This approach allows us to continuously enhance model performance and interpretability, ensuring that our final models are both accurate and actionable.

C.2.4 Race-Agnostic Training and Validation

Our models will be trained in a race-agnostic manner, meaning that race will not be used as a feature in the prediction process. This approach ensures that the identification of SDH factors influencing CVD risk is unbiased and equitable across different racial groups. Additionally, we will conduct separate validations to assess the generalizability of models trained on all-race data to specific racial groups, including NHPI and AI/AN populations. By comparing model performance across different racial subsets, we aim to identify potential biases and gaps in generalizability, ensuring that our models provide reliable predictions for

underserved populations without favoring any particular racial group. By integrating these validation and evaluation techniques, we will ensure that our geospatial machine learning models are not only accurate but also robust and generalizable. This comprehensive validation framework will underpin the reliability of our predictive insights into CVD risks among NHPI and AI/AN populations, facilitating the development of targeted and effective public health interventions.

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

Aim 2 emphasizes leveraging Explainable AI (XAI) techniques to interpret and elucidate the influence of SDH factors on CVD outcomes. The objective is to ensure that the models' predictions are understandable to stakeholders, including researchers, policymakers, and public health officials, and to offer actionable insights into the key drivers of CVD disparities in NHPI and AI/AN populations.

C.3.1 SHAP and LIME for Feature Importance and Local Interpretability

To ensure transparency and interpretability in our models, we will leverage SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) as our primary Explainable AI (XAI) techniques.

SHAP assigns an importance score to each feature by quantifying its contribution to the model's prediction, making it particularly effective for complex models like Gradient Boosted Decision Trees (GBDT) and Random Forest (RF), which capture both linear and nonlinear feature interactions. By decomposing model predictions into contributions from individual features, SHAP allows us to identify which factors (e.g., income, healthcare access, geographic location) most significantly impact CVD risk. SHAP analysis will be conducted at both the global level (aggregated across all counties) and the local level (within individual counties or regions), providing a comprehensive view of the factors driving disease disparities.

We will systematically apply SHAP and LIME to our machine learning models to dissect the contributions of various SDH and geographic factors to CVD risk. By focusing on both global and local interpretations, we aim to provide a nuanced understanding of how different variables influence health outcomes across diverse geographic contexts.

Figure 1. 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.

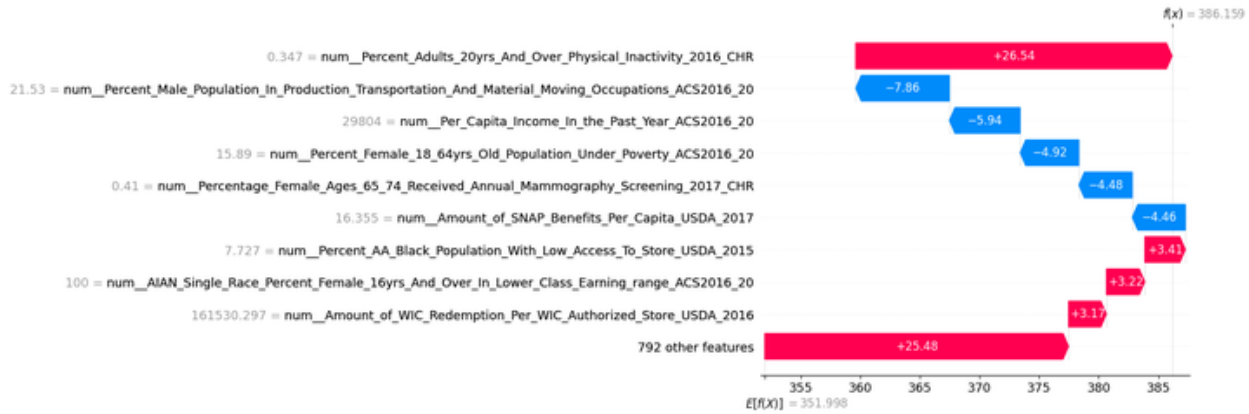

Figure 2. SHAP summary plot for all-cause CVD mortality prediction per 100,000, stratified by feature importance, using an XGBoost model. Red bars indicate features that increase mortality predictions, while blue bars decrease them. Specific feature values, such as a high "Amount of SNAP Benefits" or "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.

Additionally, Local Interpretable Model-agnostic Explanations (LIME) will complement SHAP by providing localized explanations for specific predictions, especially valuable for understanding county-level variations. LIME creates an interpretable, linear approximation of the complex model around a particular prediction, allowing us to explain why the model produced a specific outcome for a given county. This localized view makes it easier for stakeholders to interpret model decisions and enables actionable insights for public health officials at the community level.

C.3.2 Modeling Average SDH Effects with Partial Dependence Plots (PDP)

To understand how specific SDHs influence CVD risk on average across our dataset, we will utilize Partial Dependence Plots (PDPs). By applying PDPs to our machine learning models, we will visualize the marginal effect of one or two SDH features on the predicted CVD outcomes while holding all other features constant. This approach allows us to identify general trends and understand how changes in specific SDH variables impact CVD risk among NHPI and AI/AN populations. We will generate PDPs for key SDH variables identified through our SHAP and LIME analyses, such as "access to healthcare facilities," "income inequality," or "percentage of the population working in transportation jobs." For instance, if "access to healthcare facilities" continues to emerge as a key predictor, we will create PDPs to visualize how varying levels of healthcare access affect predicted CVD risk across different counties in our dataset.

By analyzing these PDPs, we aim to detect specific thresholds where small changes in an SDH variable lead to significant shifts in CVD risk. For example, we might find that CVD risk decreases substantially when the percentage of insured individuals exceeds a certain threshold in a county. Furthermore, we will use two-way PDPs to explore how combinations of SDH variables interact to influence CVD outcomes. For example, we will investigate how "income inequality" and "educational attainment" together affect CVD risk in NHPI and AI/AN populations within our dataset. By focusing on these average effects, PDPs will help us identify general trends and inform broad public health strategies aimed at reducing CVD risk across the U.S. in a geospatially aware way.. The insights gained will also guide us in recommending interventions that address the most influential SDH factors at specific population scales.

C.3.3 Understanding Interaction Effects and Complex Dependencies of CVD Geospatially

To delve deeper into the variability and complex interactions between geographic and SDH variables that may not be captured by PDPs, we will employ Individual Conditional Expectation (ICE) plots and Accumulated Local Effects (ALE) plots. These methodologies will enable us to explore intricate dependencies between features often obscured in machine learning models.

We will apply ICE plots to key SDH features to visualize how the effect of a single SDH feature on CVD risk predictions varies across different counties in our dataset. 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. This will be crucial for developing multifaceted public health strategies aimed at reducing CVD disparities in specific populations.

C.3.4 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 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, 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 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 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 GANTT Chart

We have constructed a five-year plan for the realization of this research project. Given the interoperability of the aims, we anticipate that Aim 2 will be able to be started not long after Aim 1. Regardless of the performance of the models in Aim 1, Aim 2's goals will be conducted and presented to the scientific community.