IJCRT.ORG ISSN : 2320-2882
INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

HEALTHSCAPE: A GIS EXPLORATION OF SOCIAL DETERMINANTS

Roshan Aadithya
Dept. of CSE(Data Science)
RNS Institute of Technology
Bangalore, Karnataka, India

Shreya H S
Dept. of CSE(Data Science)
RNS Institute of Technology
Bangalore, Karnataka, India

Shreya L Gowda
Dept. of CSE(Data Science)
RNS Institute of Technology
Bangalore, Karnataka, India

Spoorthi M
Dept. of CSE(Data Science)
RNS Institute of Technology
Bangalore, Karnataka, India

Mrs. Smitha BA
Dept. of CSE(Data Science)
RNS Institute of Technology
Bangalore, Karnataka, India

***Abstract*—The health inequities across Indian districts have roots in Social Determinants of Health inclusive of maternal care, child nutrition, sanitation, education, and access to healthcare[4]. While datasets like NFHS-5 [1] are rich in information on these indicators, the high dimensionality coupled with inconsistent administrative formats causes serious challenge in systematic vulnerability assessment[14]. This paper presents HealthScape, a data-driven unified analytics framework that integrates PCA-based dimensionality reduction[14], [15], machine learning-based classification[6], [8], and geospatial visualization[7], [10] for generating district-level SDOH vulnerability maps. Standardized data cleaning and feature engineering, dimensionality reduction based on PCA, vulnerability scoring, and classification by XGBoost form the pipeline. District boundaries are spatially joined using GeoJSON files; hence, interactive choropleth maps can be generated to highlight high-risk clusters and regional disparities[15]. Experimental results confirm strong predictive performance and accurate spatial representation by identifying the top vulnerable districts and subsequently yielding interpretable insights for policymakers. Complex survey data will now be transformed into actionable intelligence, and the model proposed is scalable for population-health monitoring across India.**

***Index Terms*—SDOH, NFHS-5, PCA, Machine Learning, GIS, Vulnerability Index, Spatial Analysis**

I. INTRODUCTION

Health inequality persists in the country, and it is caused not only by medical factors but also by a wide spectrum of social, economic, and environmental conditions commonly referred to as the Social Determinants of Health[4].. These different determinants, ranging from access to maternal healthcare and sanitation to education levels and nutritional status, affect how various population groups experience health risks differently. While national programs have improved several indicators, significant variation still exists at the district level. Understanding these variations requires analytical tools that can move beyond isolated statistics and reveal the underlying patterns embedded within large-scale public health datasets.

The NFHS-5[1] is one of India's most comprehensive sources of district-level health and demographic information. However, this depth also poses a challenge because interpreting hundreds of indicators, inconsistent naming of districts, missing values, and multidimensional patterns directly from the raw data is difficult. Traditional reporting formats summarize indicators separately, offering limited insight into how multiple SDOH factors interact to influence overall vulnerability. Consequently, policymakers often lack a consolidated, comparative picture of district-level disparities and are thus limited in their ability to design targeted interventions.

Recent advances in data science and machine learning[6], combined with geospatial analytics [7], [15], have provided powerful tools for tackling these challenges. Methods like PCA, clustering, and machine learning models are able to extract meaningful relationships from high-dimensional datasets. In turn, GIS visualizes complex analytical results on intuitive maps, allowing hotspots, regional trends, and structural inequalities to be more easily comprehended.Combining these provides a means of converting complex survey data sets into actionable insight.

HealthScape was designed with this objective in mind. It integrates statistical modeling, machine learning, and GIS visualization into a single end-to-end analytic framework for evaluating district-level health vulnerability throughout India. From cleaning and standardizing NFHS-5 data, extracting core vulnerability dimensions through PCA, predicting high-risk districts using machine learning, and then translating those outputs into spatial maps, the system produces a unified, interpretable representation of SDOH patterns. Such a system enables researchers, policymakers, and health administrators to understand where vulnerabilities are concentrated and to understand better the factors that create vulnerability.

II. Problem Statement

Health outcomes across districts in India vary significantly due to a wide range of disparities in maternal health care, nutrition, sanitation, education, and access to healthcare services, which also go by the name of Social Determinants of Health [4]. While datasets like NFHS-5[1] include extensive district-level information, their high dimensionality, inconsistency in district naming, and lack of an integrated analytical framework render objective assessment of vulnerability quite challenging for policymakers. Existing systems provide only fragmented or state-level insights and do not offer a unified data-driven vulnerability index. As such, decision-makers do not have a clear method to identify high-risk districts and prioritize interventions

III. Literature Review

SDOH-based analytics using NFHS-5 [1] and administrative health data [2] have become central to identifying district-level disparities in India's health outcomes. Prior work highlights persistent inequalities[6] in maternal health, spatial clustering [7]of service gaps, child malnutrition hotspots[8] , and digital-health access [13], emphasizing the need for integrated, multi-dimensional vulnerability assessment frameworks. Studies applying explainable ML to public-health datasets and GIS-based spatial epidemiology further demonstrate that combining statistical modelling, machine learning, and geospatial tools[6], [15] improves precision and interpretability in population-level risk assessment. These findings motivate unified systems such as HealthScape, which leverage ML and GIS to convert complex SDOH indicators into actionable district-level vulnerability insights.

A. Explainable Machine Learning for Maternal Mortality Analysis

Saragadam et al. (2025)[6] applied explainable machine learning techniques to district-level Health Management Information System (HMIS)[2] data to identify the determinants of Maternal Mortality Ratio (MMR) [5]in India. Using gradient boosting models combined with SHAP-based interpretability, the study demonstrated that antenatal care coverage, institutional deliveries, female literacy, and referral system strength were major contributors to MMR variation across districts. The work highlights the usefulness of administrative datasets for SDG monitoring and emphasizes the need for transparent analytical methods that support targeted decision-making.

B. Spatial Clustering of Maternal Health Indicators

Sharma, Kumar, and Singh (2024)[7] studied geographic clustering of maternal health outcomes by applying NFHS-5[1] district-level indicators. Using global Moran's I and local spatial autocorrelation LISA, high-risk clusters were identified with statistical significance, in the EAG states and parts of Northeast India. Subsequently, using spatial regression, the study found that female literacy, socio-economic status, and accessibility to maternal health services are the significant predictors. The findings identify the role played by spatial epidemiology in guiding region-specific maternal health interventions.

C. Machine Learning and Geospatial Modelling for Child Malnutrition

Agarwal et al. (2023)[8] combined NFHS-5 microdata with environmental and socio-economic covariates to predict district- and cluster-level malnutrition hotspots in India. Using ensemble machine learning models and spatial cross-validation, the study produced high-resolution predictive maps of stunting, wasting, and underweight prevalence. This work presents the potential benefits of combining ML and GIS tools for precision nutrition planning.

D. Intra-Urban Health Inequalities in Megacities

Singh et al. (2023)[9] investigated the intra-urban differentials in maternal and child health in Indian megacities by integrating NFHS-5 cluster data with municipal administrative boundaries. The study showed significant inequalities in health between slum and non-slum settlements; poorer urban neighborhoods consistently recorded lower utilization of maternal care services and higher malnutrition rates. Using spatial statistics and multilevel models, the study established that urban health outcomes are significantly affected by neighborhood-level deprivation and service availability.

E. Climate–Health Vulnerability Assessment Using NFHS-5

Rao et al. (2023)[10]constructed a composite climate-health vulnerability index by integrating NFHS-5[1] indicators with high-resolution climatic exposure datasets. The study mapped climatic health vulnerability for the entire nation using standardized sensitivity, exposure, and adaptive-capacity metrics. High-risk zones were identified within the Indo-Gangetic plains and the drought-prone central regions. Spatial autocorrelation techniques were used to validate these clusters.

F. Geographic Determinants of COVID-19 Vaccination Coverage

Patel et al. (2024)[11] examined district-level vaccination patterns for COVID-19 in relation to measures of social determinants captured in NFHS-5[1]. Spatial clustering analyses highlighted significant geographic inequities, with low-coverage districts concentrated in socioeconomically disadvantaged regions. Using spatial regression, it was possible to identify digital access, female literacy, media exposure, and wealth as the key predictors of vaccination uptake, highlighting the impact of digital and social inequity on the vaccination rollout across India.

G. Digital Health Equity and Telemedicine Adoption

Verma, Jain, and Srivastava (2024) [13]investigated the uptake of telemedicine services across districts in India, using NFHS-5[1] socio-demographic indicators coupled with administrative utilization data. Strong spatial patterns of digital health access were found, with low adoption in areas with poor mobile ownership, limited internet connectivity, and overall socio-economic vulnerability. Moran's I and LISA analyses

confirmed that digital exclusion exhibited significant clustering, underlining structural barriers to the equitable deployment of digital health.

H. Comparison of existing approaches

approaches	features	strengths	challenges
Explainable ML on Administrative Health Data	Uses HMIS indicators, gradient boosting models, SHAP interpretability	Provides transparent insights into health-system drivers; supports SDG monitoring	Data quality issues, under-reporting, ecological bias
Spatial Epidemiology Using NFHS-5	Moran's I, LISA hotspot detection,spatial regression	Identifies geographic clusters; highlights regional disparities; effective for targeted planning	Cross-sectional data limits causal inference; unstable estimates in low-sample districts
ML–GIS Hybrid Models for Nutrition /Vulnerability Mapping	ML prediction (RF, XGBoost), geospatial layers, environmental covariates	Captures complex interactions; produces high-resolution risk maps; supports precision intervention	Model interpretability varies; temporal mismatch in datasets; spatial autocorrelation issues

I. Open research directions and gaps

Despite progress in applying machine learning and geospatial methods to SDOH-driven vulnerability[10] assessment, several gaps remain. First, most studies rely on cross-sectional datasets such as NFHS-5[1], limiting the ability to capture temporal changes or emerging vulnerabilities. Developing longitudinal or real-time SDOH monitoring systems remains an open research need. Second, existing models often treat environmental, digital, and socio-economic determinants separately; integrating these multi-layered exposures into unified, causal frameworks is still understudied. Third, current ML-based vulnerability tools face interpretability challenges, particularly when incorporating complex, high-dimensional features or ensemble models.

Another critical gap is the limited availability of fine-grained spatial data. District-level aggregation masks within-district heterogeneity[7], especially in urban slums, tribal areas, and remote regions. Future research must explore small-area estimation and satellite-derived proxies to overcome data scarcity. Finally, while many studies generate spatial risk maps, there is limited evidence of their operational uptake in government planning processes. Designing deployable, user-centric decision-support systems that integrate ML models with health management platforms represents a key translational opportunity.

IV. System Architecture

The propoosed HealthScape system architecture and its principal components design emphasizes modularity, reproducibility, and scalability so that NFHS-5[1] and related datasets can be transformed into robust, interpretable vulnerability measures and spatial outputs. Subsections A–G present the overall design followed by detailed descriptions of each architectural layer.

A. Overall Design

The architecture follows a structured flow consisting of data ingestion, cleaning, feature engineering, modeling, and GIS-based visualization. Each layer is independent, enabling flexible updates when new datasets or indicators become available.

1. Data Acquisition Layer
2. Data Cleaning and Harmonization
3. Feature Engineering and Feature Store
4. PCA and Machine Learning Modeling
5. Spatial Visualization and Deployment

The design prioritizes Reproducibility, Interpretability, Scalability, Actionability.

B. Data Acquisition Layer

This layer collects all input data required for the framework, including

1. NFHS-5 indicators (district-level).
2. Optional HMIS administrative metrics.
3. GeoJSON files containing district boundaries.

The system validates schemas, checks completeness, and stores raw inputs with metadata for future traceability.

C. Data Cleaning and Harmonization

To ensure consistency across sources, this layer performs

1. Standardization of district names and formats.
2. Fuzzy-matching for mismatched district entries.
3. Missing-value imputation using rule-based methods.
4. Conversion of numeric, categorical, and percentage fields into consistent formats.

The output is a clean, unified dataset ready for analytical processing.

D. Feature Engineering and Feature Store

This layer transforms raw indicators into structured SDOH features

1. Grouping variables into key domains (maternal health, nutrition, WASH, socio-economic, digital access).
2. Applying normalization and scaling.
3. Creating composite indicators where relevant.

A versioned feature store preserves engineered features for reproducible modeling and comparisons over time.

E. PCA and Machine Learning Modeling

The analytical engine combines dimensionality-reduction and predictive modeling.

1. PCA - extracts underlying vulnerability patterns across SDOH domains. PC1 is converted into the Vulnerability Index.
2. ML Classification - Logistic Regression classify districts into high- or low-vulnerability categories.
3. Evaluation - accuracy, ROC-AUC, calibration curves, and Brier scores ensure reliable predictions.
4. Explainability:feature-importance and SHAP-based insights support transparency in decision-making.

F. Spatial Visualization and Deployment

This final layer links analytical results with spatial boundaries to generate actionable outputs. 1. Interactive GIS dashboards using GeoJSON and Plotly/Leaflet.

2. Static choropleth maps for reporting and presentations

3. District-level summaries highlighting top-risk and low-risk regions

4. Deployment options include notebook-based analysis, containerized execution, or cloud-hosted map services.

V. METHODOLOGY AND IMPLEMENTATION

HealthScape represents an integrated analytical workflow developed through the integration of statistical transformation, machine learning, and geospatial intelligence to assess health vulnerability at district levels. This section describes the end-to-end pipeline from raw data acquisition, processing, and generation of vulnerability scores and spatial maps. The methodological choices aim at transparency, reproducibility, interpretability, and operational suitability for public-health planning..

A. Data Preparation

NFHS-5 indicators and auxiliary data are harmonized into a uniform analytical structure in a rigorous process of data preparation 1. Source consolidation: different NFHS-5 district-level tables have been combined into one dataset on maternal health, nutrition, sanitation, demographics, and use of digital channels, using the district names as keys

2. Standardization of data: Districts are normalized by lowercasing, punctuation removal, and fuzzy matching to ensure cross-dataset alignment with GeoJSON boundaries.

3. Missing-value handling: Median imputation of the numerical variables and mode-based filling for categorical indicators; missingness patterns also record for sensitivity checks.

4. Range validation: Indicators that include measures of percentages are validated against the 0–100 limit, extreme or implausible values are identified and corrected.

5. Consistency enforcement: Units of measurement, data type, and column format standardization ensure a coherent and analysis-ready dataset.

This structured preprocessing improves the reliability of downstream modeling and ensures compatibility of the data with spatial datasets.

B. Feature Engineering

Feature engineering transforms this cleaned dataset into analytically expressive variables, capturing multi-dimensional SDOH characteristics.

1. Domain classification: Indicators are grouped into coherent SDOH domains, including maternal care utilization, child nutritional status, WASH conditions, socio-economic context, and digital access.

2. Scaling and normalization: Application of the Z-score standardization for the variables that feed PCA in ML models, Min–Max normalization is preferred for interpretability.

3. Composite indicators: Summary measures for domains (such as the Maternal Care Index, Sanitation Index) are obtained by weighted or unweighted aggregation.

4. Correlation filtering: Highly collinear features are filtered out to reduce redundancy and stabilize PCA and ML output.

5. Feature documentation: A structured ”data dictionary” is developed that traces data transformations, variable definitions, and domain assignments. Here is a list of indicators that were used for analysis

TABLE I
LIST OF INDICATORS USED FOR ANALYSIS

Indicator Name	Category	Description	Source
ANC 4+ Visits	Maternal Health	Percentage of mothers who received atleast four antenatal care visits	NFHS-5
Institutional Delivery	Maternal Health	Births delivered in health facilities	NFHS-5
Immunization Coverage	Health Services	Facility-reported vaccination completion	HMIS
Stunting	Child Nutrition	Children under 5 whose height-for-age is below WHO standard	NFHS-5
Wasting	Child Nutrition	Children whose weight-for-height is below WHO standard	NFHS-5
Full Immunization	Child Health	Children (12–23 months) receiving all basic vaccinations	NFHS-5
Anaemia in Women	Maternal Health	Prevalence of anaemia among women aged 15–49	NFHS-5

This step enhances not only the performance of the model but also interpretability, organizing a complex indicator space into meaningful analytical components.

C. Dimensionality Reduction (PCA)

Using PCA to develop a compact representation of the district vulnerabilities across different SDOH dimensions.

1. Input Formation: The PCA input is a matrix comprising standardized SDOH indicators (districts × features).

2. Component Extraction: PCA identifies orthogonal components capturing the maximum variance in the dataset. PC1 is the most interpretable dimension, reflecting broad structural disadvantage.

3. Index Construction: PC1 is min–max scaled to yield the Vulnerability Index ranging from 0 (least vulnerable) to 1 (most vulnerable).

4. Loading Interpretation: Through PCA loadings, one can ascertain which social determinants, such as sanitation deficits, nutritional weaknesses, or poor matern‘al health, are more contributing factors to vulnerability.

5. Dimesionality Justification: PCA eliminates multicollinearity and removes noise, helping ML models focus on dominant patterns rather than redundant indicators. The resulting index is a strong, data-driven indicator of district-level disadvantage.

D. Machine Learning Modeling

To complement the vulnerability index with predictive capability, multiple classification models—including Logistic

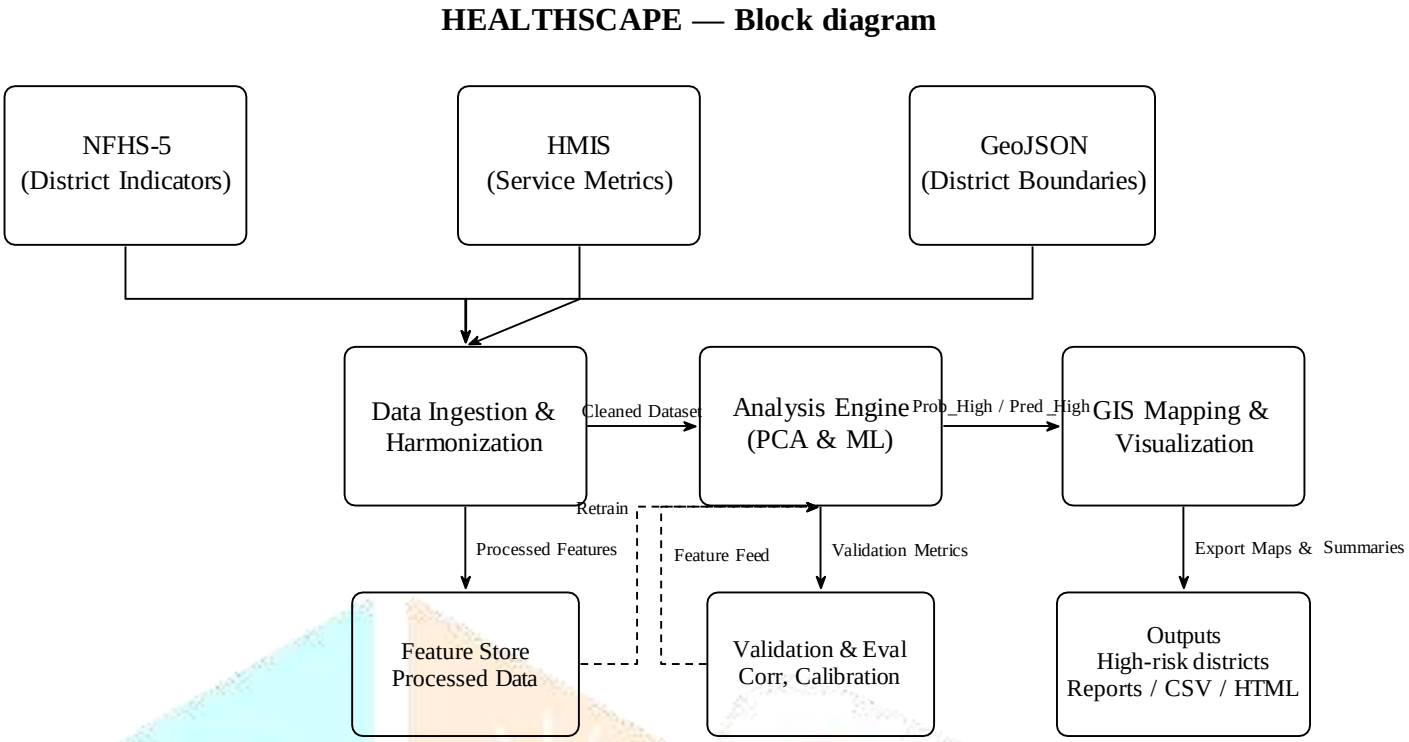

Fig. 1. Block diagram of the HealthScape analytical pipeline after adjusting box sizes to fit page width.

Regression, Random Forest, and SVM—are tested. XGBoost is selected due to its superior performance, robustness with tabular health data, and ability to model nonlinear interactions. The modelling pipeline includes stratified train-test splitting, applying SMOTE to balance high-risk vs non-high-risk districts, and performing calibration to ensure reliable probability outputs. Evaluation uses metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Below is a Performance comparison of ML models

TABLE II
PERFORMANCE COMPARISON OF ML MODELS

Model	Accuracy	F1	ROC-AUC
XGBoost	0.880282	0.849558	0.942414
Random Forest	0.866197	0.837607	0.927711
Naive Bayes	0.830986	0.800000	0.888503
Logistic Regression	0.823944	0.800000	0.929140
SVM (RBF)	0.823944	0.778761	0.905861
KNN	0.802817	0.777778	0.897999

E. Spatial Integration and GIS Mapping

Geospatial visualization turns the results of analytics into interpretable maps that support regional planning.

1. Spatial Joins : The dataset is combined with district-level GeoJSON boundaries based on harmonized district identifiers.

2. Map generation: Vulnerability Index, probability scores, and PCA components are visualized as Choropleth layers to enable comparison of patterns across regions.

3. Cluster identification: Regional clusters of high vulnerability visually emerge, reflecting the previously documented disparities in maternal health and socio-economic outcomes.

4. Interactive dashboards: HTML-based maps include hover tooltips, legend controls, and district-level summaries that facilitate real-time exploration.

5. Reporting outputs: High-resolution PNG maps are produced for printed reports, presentations, and government briefings.

This step links statistical complexity with intuitive insights such that immediate comprehension of spatial inequity is possible.

F. Implementation Environment

The complete pipeline was implemented using modern data-science tools in a reproducible analytical environment.

1. Software stack in use: Python (Pandas, NumPy), Scikit-learn, GeoPandas, Plotly, and SHAP.

2. Versioning: All intermediate artifacts are versioned-cleaned datasets, feature matrices, PCA loadings, ML models-for reproducibility.

3. Deployment configurations:

• Notebook workflows for exploration and prototyping

• Containerized environments for reproducible execution

• Optional deployment on cloud for scalable data processing and map hosting

4. Documentation: The pipeline is completely logged to provide traceability for all transformations, updates of the models, and visual outputs.

This systematic environment supports long-term maintainability and adaptability to new data sources.

VI. Results

The HealthScape framework produces an integrated set of analytical and visual outputs that collectively capture district-level health vulnerability across India by combining PCA-based dimensionality reduction, supervised machine learning, and geospatial visualization. Figure 1 shows the HealthScape dashboard highlighting the district-level vulnerability[10] map derived from the PCA-based composite index. The Vulnerability Index, scaled between 0 and 1, consolidates over one hundred correlated NFHS-5[1] indicators related to maternal health, child nutrition, sanitation, education, and healthcare access into a single interpretable measure. The resulting spatial distribution reveals strong geographic clustering of vulnerability, with consistently higher scores observed across districts in eastern, north-central, and north-eastern India, while southern and western regions exhibit comparatively lower vulnerability levels. The same HealthScape dashboard also integrates complementary analytical views, including feature-importance plots, state-level vulnerability comparisons, and model performance metrics, enabling multi-level interpretation of results within a single interface.

Figure 2 presents the chatbot component embedded within the HealthScape dashboard, which enables interactive querying of model outputs and vulnerability rankings. Using the PCA-derived index, the chatbot identifies the most vulnerable districts, with several districts from Nagaland—such as Tuensang, Mon, Kiphire, Zunheboto, and Longleng—appearing among the highest-ranked. This result is consistent with the spatial patterns observed in the vulnerability map and reflects structural challenges in healthcare accessibility, terrain, and service coverage in the region. The chatbot enhances interpretability by translating complex analytical results into natural-language responses, allowing users to explore district-level vulnerability, understand contributing factors, and validate findings without requiring direct interaction with raw data or code. This human-centric layer bridges the gap between advanced analytics and policy-oriented decision-making.

Figure 3 further examines the relationship between the model-generated probability of high vulnerability and a key maternal health indicator—the percentage of mothers receiving at least four antenatal care visits. The scatter plot demonstrates a clear inverse relationship, with districts exhibiting higher ProbHigh values generally showing lower ANC coverage. This negative association confirms that the machine learning model effectively captures meaningful public-health signals rather than statistical artifacts. While some dispersion is observed at higher probability levels, indicating heterogeneity among vulnerable districts, the overall trend validates the model's sensitivity to critical maternal health determinants and supports the use of ProbHigh as a reliable risk indicator.

These patterns were intuitive in GIS-based visualizations, where one can visualize district-wise vulnerability gradients and thematic variations across maternal health, sanitation, and socio-economic factors. Altogether, the results strengthen the case for HealthScape as a scalable and interpretable framework for vulnerability assessment at the district level.

Fig. 2. District-level PCA-based vulnerability map

Fig. 3. HealthScape chatbot

These patterns were intuitive in GIS-based visualizations, where one can visualize district-wise vulnerability gradients and thematic variations across maternal health, sanitation, and socio-economic factors. Altogether, the results strengthen the case for HealthScape as a scalable and interpretable framework for vulnerability assessment at the district level. These results highlight the system's potential to support evidence-based prioritization of districts, guide targeted interventions, and assist policymakers in addressing health inequities driven by social determinants.

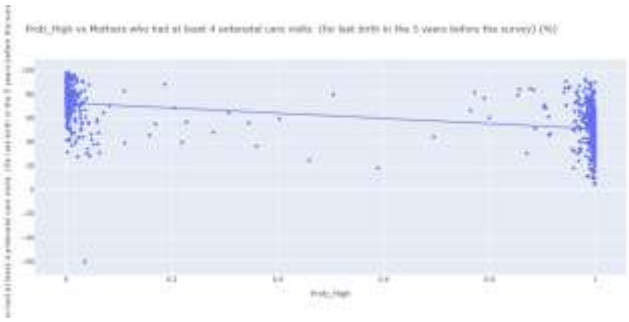

Fig. 4. relationship between predicted high-risk probability and ANC

REFERENCES

[1] National Family Health Survey (NFHS-5), Ministry of Health and Family Welfare (MoHFW), Government of India. International Institute for Population Sciences (IIPS), 2019–21.

[2] Health Management Information System (HMIS), Ministry of Health and Family Welfare, Government of India, 2019-21.

[3] Census of India , Office of the Registrar General and Census Commissioner, India, 2011.

[4] World Health Organization (WHO), A Conceptual Framework for Action on the Social Determinants of Health, 2021.

[5] UNICEF India, Maternal and Child Health Indicators Report, 2020.

[6] Saragadam et al., Explainable Machine Learning on Health Management Information System Data to Unveil Health Factors, 2025.

[7] Sharma, Kumar and Singh, Spatial clustering of maternal health outcomes in India: Evidence from NFHS-5, 2024.

[8] Agarwal, Joshi and Mehta, Predicting child malnutrition hotspots in India using machine learning and geospatial analysis of NFHS-5 data, 2023.

[9] Singh, Kumar and Pandey, Intra-urban health inequalities in Indian megacities: A spatial analysis using NFHS-5 and urban administrative data, 2023.

[10] Rao, Chandra and Gupta, Climate-Health Vulnerability Mapping in India: Integrating NFHS-5 Data with Environmental Indicators for Adaptive Planning, 2023.

[11] Patel, Gupta and Reddy, Geographic Disparities in COVID-19 Vaccination Coverage and Social Determinants of Health in India: A Spatial Analysis Using NFHS-5 Data, 2024.

[12] Mishra, Saxena and Tiwari, Health System Resilience and Social Determinants: Lessons from COVID-19 Response Across Indian States Using NFHS-5 Baseline Data, 2023.

[13] Verma, Jain and Srivastava, Digital Health Equity in India: A Geospatial Analysis of Telemedicine Adoption Using NFHS-5 and Administrative Data, 2024.

[14] Acharya, R., and Porwal, A, A vulnerability index for maternal health in India: A social, economic and demographic assessment, 2020.

[15] Khan, A. A., and Mohanty, S,GIS-based mapping and assessment of health vulnerability in India,2020.