PREDICTING AND MAPPING THE GEOGRAPHIC DISTRIBUTION OF GLAUCOMA IN THE UNITED STATES: THE ROLE OF SOCIAL DETERMINANTS USING THE ALL OF US DATASET

By

Ayobami Abolore Alimi

May, 2025

Director of Thesis: Nic Herndon, PhD

Major Department: Computer Science

ABSTRACT

Vision impairment and eye diseases are significant public health concerns in the United States and globally. Glaucoma, a chronic and progressive disease, is one of the leading causes of irreversible blindness worldwide. In the U.S., more than three million individuals are estimated to be affected, with projections indicating a rise as the population ages. While clinical and genetic factors influencing glaucoma onset and progression have been extensively studied, growing evidence suggests that environmental exposures, socioeconomic status, and lifestyle factors also play a crucial role. With disparities in healthcare access and outcomes based on socioeconomic factors, it is crucial to explore how these factors, alongside genetic predispositions, affect glaucoma onset and progression. Addressing these gaps could lead to more targeted interventions, improving outcomes for vulnerable populations. This study aims to bridge this gap by leveraging machine learning techniques to build predictive models for glaucoma risk. By utilizing demographic information and Social Determinant of Health (SDOH) from the *All of Us* dataset, this research develops a comprehensive framework for glaucoma prediction. These models allow for an improved understanding of how SDOH influences glaucoma risk, helping to inform early detection strategies. The optimized Decision Tree model, tuned with GridSearchCV, was the best-performing model for this prediction task, achieving an accuracy of 67.87%. For class 0 (Non-Glaucoma), it yielded a precision

of 0.71, recall of 0.52, and an F1 score of 0.60. For class 1 (Glaucoma), the model achieved a precision of 0.66, recall of 0.81, and an F1 score of 0.73. Feature importance analysis identified age as the most significant predictor, followed by race and the affordability of seeing an eye doctor. In contrast, factors such as affordability of specialist care and copay affordability had minimal impact. The findings from this study have broader implications for enhancing glaucoma risk assessments and healthcare interventions. Additionally, the methodological approach can be applied to other complex diseases, contributing to a more equitable and informed public health approach. By emphasizing social determinants, this research takes a promising step toward reducing the burden of glaucoma and advancing the goals of precision medicine.

PREDICTING AND MAPPING THE GEOGRAPHIC DISTRIBUTION OF GLAUCOMA IN THE UNITED STATES: THE ROLE OF SOCIAL DETERMINANTS USING THE ALL OF US DATASET

A Thesis

Presented to The Faculty of the Department of Computer Science

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Computer Sience

By

Ayobami Abolore Alimi

May, 2025

Director of Thesis: Nic Herndon, PhD

Thesis Committee Members:

Ray Hales Hylock, PhD

David Marvin Hart, PhD

DEDICATION

This work is dedicated to Almighty Allah and my beautiful wife Maryam Alimi.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my parents, whose unwavering support, love, and encouragement have been the foundation of my academic journey. Their sacrifices and belief in my potential have been instrumental in shaping my path.

To my wonderful wife, your patience, understanding, and constant motivation have been my greatest source of strength throughout this process. Your support has made this journey smoother, and I am forever grateful for your love and encouragement.

I extend my heartfelt appreciation to my advisor, Dr. Herndon, for his invaluable guidance, insightful feedback, and continuous encouragement. His expertise and mentorship have played a crucial role in shaping this research and improving my analytical skills.

I am also immensely grateful to my thesis committee members, Dr. Hylock and Dr. Hart, for their time, effort, and constructive criticism. Their perspectives and expertise have significantly enhanced the quality of this work.

Additionally, I would like to thank my supervisor at work, Dr. Cooke-Bailey (Brody School of Medicine), for her support and understanding as I balanced my professional responsibilities with my academic pursuits. Her encouragement and flexibility have been greatly appreciated.

Finally, I extend my gratitude to my family, colleagues, friends, and everyone who has supported me in one way or another during this journey. This achievement would not have been possible without your encouragement and support.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Vision impairment and eye diseases are critical public health concerns in the United States and around the world. Glaucoma, a chronic and progressive disease, is one of the main causes of irreversible blindness worldwide. More than three million people in the United States are estimated to be affected by glaucoma, with projections that this number will increase as the population ages [8]. Primary open-angle glaucoma (POAG), the most common type in the United States, often goes undiagnosed until significant vision loss occurs, underscoring the importance of early detection and prevention. POAG disproportionately affects certain racial and ethnic groups, including African Americans and Hispanics, who tend to experience earlier onset and more severe disease progression.

Although clinical and genetic factors in the onset and progression of glaucoma have been extensively studied [1, 10, 23, 25], growing evidence suggests that environmental factors, such as exposure to pollutants, socioeconomic factors, and lifestyle variables, also play an influential role [2, 5, 13, 19, 22]. However, the exact nature of the interaction between these environmental determinants and the prevalence of glaucoma remains unclear. Recent advances in data collection and analytics, particularly through large datasets such as the *All of Us* (AoU) Research Program [24], present an unprecedented opportunity to analyze these complex relationships. The AoU dataset includes comprehensive demographic, health, genetic, and environmental data from a diverse cohort across the United States, providing a solid foundation for investigating the distribution of glaucoma and other vision diseases and understanding associated environmental factors.

This thesis will leverage the AoU dataset to investigate how environmental factors correlate with the geographic distribution of glaucoma in the United States. By using data science methods to explore this intersection, this study aims to contribute new insights into the role of social and ecological determinants in eye health disparities.

1.1 Research Objectives

The primary objective of this research is to analyze the geographic distribution of glaucoma across the United States and to assess the relationship between this disease and various social factors. This study will focus on several key questions:

1. What is the geographic distribution of glaucoma across the United States?
2. Which environmental / social factors correlate most significantly with the prevalence of glaucoma
3. Are there identifiable patterns of disparity in glaucoma based on socioeconomic or environmental factors?
4. Can glaucoma be predicted with a machine learning model using demographic information and social economic factors?

By addressing these questions, the research will explore whether individuals in certain regions with specific environmental profiles are at greater risk for glaucoma and other vision diseases. Additionally, this study will analyze if disparities observed align with the distribution of these environmental factors, potentially informing targeted interventions for vulnerable communities.

1.2 Significance of the Study

The potential implications of this research are both far-reaching and impactful. Glaucoma has been extensively researched from a genetic and clinical standpoint [1, 10], yet the en-

vironmental factors influencing its development are less understood. Social determinants of health (SDOH) – such as income level, access to healthcare, and community environment – are increasingly recognized as crucial to health outcomes. For example, individuals in areas with high pollution levels or limited healthcare access may be at a greater risk for developing vision problems, potentially due to oxidative stress or chronic health conditions linked to environmental exposures.

Understanding these correlations can aid in identifying vulnerable populations and tailoring preventative healthcare policies that mitigate these risks. Public health interventions informed by this research could address underlying causes, rather than symptoms, by focusing on environmental improvements, education, and increased screening efforts in high-risk areas. By mapping the intersections between environmental variables and vision health, this study hopes to provide insights that can enhance healthcare equity, improve early detection, and ultimately reduce the societal burden of glaucoma and vision impairment.

Furthermore, the data science methodologies applied in this study could serve as a framework for similar analyses of other diseases, demonstrating the potential of large-scale, diverse datasets like AoU to explore complex health-environment interactions.

Chapter 2

Related Work

2.1 Overview of Glaucoma

Glaucoma is a leading cause of irreversible blindness worldwide, characterized by progressive damage to the optic nerve, often associated with elevated intraocular pressure (IOP) [24]. Glaucoma was the cause for blindness in 3.61 million people or 8.4% of the 43.3 million blind people globally in 2020, and glaucoma was the cause for moderate to severe vision impairment (MSVI) in 4.14 million people or 1.4% of the 295 million people visually impaired in 2020 [4]. The most prevalent form, primary open-angle glaucoma (POAG), accounts for the majority of cases in the United States and is particularly concerning due to its asymptomatic onset until advanced stages [20].

The global burden of glaucoma is increasing, with estimates suggesting that by 2040, more than 111 million people will be affected, primarily among populations aged 60 and above [20]. In the U.S., significant racial and ethnic disparities exist, with African Americans being three-four times more likely to develop POAG than White Americans, and Hispanic populations also showing higher prevalence rates [15]. These disparities indicate a strong need to analyze geographic distribution trends and the influence of demographic and SDOH on glaucoma development and progression.

2.2 Genetic and Demographic Risk Factors Influencing Glaucoma

While genetic predisposition plays an essential role in glaucoma susceptibility, demographic factors such as age, sex, and race/ethnicity are equally significant in determining individual risk [10]. Genome-wide association studies (GWAS) have identified multiple glaucoma-associated loci, including MYOC, CAV1/CAV2, TMCO1, MYOF and others which influence disease onset and progression [1]. However, genetic risk alone does not fully explain glaucoma prevalence, necessitating an exploration of demographic predictors such as age, gender, and racial disparities.

- Age: Older adults (60+) face a higher risk of glaucoma due to age-related structural changes in the eye and a decline in neuroprotective mechanisms [11].
- Sex: Some studies suggest that hormonal differences may contribute to glaucoma progression, with postmenopausal women at higher risk due to declining estrogen levels [21].
- Race/Ethnicity: As noted earlier, African Americans and Hispanics have an increased risk of POAG, potentially due to both genetic and environmental factors, warranting further investigation into their predictive role [25].

2.3 Social Determinants of Health and Glaucoma Prediction

The World Health Organization (WHO) defines SDOH as the conditions in which people are born, grow, live, work, and age, which shape health outcomes [14]. Several SDOH factors, listed below, have been linked to glaucoma prevalence and severity, making them key predictors in glaucoma risk modeling.

2.3.1 Healthcare Access and Insurance Coverage

Limited access to ophthalmologic care results in delayed glaucoma diagnosis and poorer outcomes. A study [17] found that uninsured individuals and those without Medicaid were more likely to present with advanced-stage glaucoma, emphasizing the role of healthcare accessibility in disease progression.

2.3.2 Socioeconomic Status (SES) and Income

Lower-income populations are at higher risk of developing glaucoma due to limited healthcare access, poor living conditions, and increased exposure to environmental risk factors [6]. Income disparities influence an individual's ability to afford routine eye exams, contributing to higher rates of undiagnosed or late-stage glaucoma cases.

In addition, factors like housing quality, community safety, and occupational risks can exacerbate health disparities. People in lower-income areas may face greater exposure to environmental hazards, such as air and water pollution, which are linked to glaucoma risk. By examining these social factors, this study aims to investigate the influence of SDOH on glaucoma distribution across the U.S., with a particular focus on high-risk and under-served populations.

2.3.3 Occupational and Environmental Stress

Chronic stress has been linked to increased IOP and optic nerve vulnerability. Work-related stress and long exposure to blue-collar jobs (e.g., exposure to industrial pollutants, strenuous labor) have been suggested as risk factors for glaucoma progression [26, 2].

2.3.4 Urban vs. Rural Disparities

Rural populations face higher rates of blindness due to glaucoma than urban populations, largely due to a lack of specialized eye care services and longer travel times to healthcare facilities [9]. Predictive models incorporating ZIP code-based geographic information can help identify high-risk rural communities needing targeted interventions.

2.4 Predictive Modeling of Glaucoma Using Machine Learning

Given the increasing availability of large-scale health datasets, machine learning (ML) approaches have become an essential tool for predicting glaucoma risk using demographic and SDOH data.

2.4.1 Supervised Learning for Glaucoma Prediction

Supervised learning models such as logistic regression, random forests, and deep learning have been effectively applied to predict glaucoma diagnosis using a combination of demographic, genetic, and SDOH variables [3, 12]. Recent studies have shown that machine learning algorithms, including gradient boosting models, trained on electronic health records (EHRs) and socioeconomic data can significantly enhance glaucoma risk prediction accuracy [12]. These findings highlight the potential of leveraging routinely collected clinical, lifestyle, and demographic data within EHRs to develop scalable and data-driven models for early glaucoma detection.

2.4.2 Geographic Information Systems (GIS) and Spatial Analysis

GIS-based approaches have been employed to map glaucoma distribution and identify environmental risk factors. Integrating machine learning with GIS enables predictive modeling of high-risk geographic zones, improving targeted screening efforts [5].

A notable advantage of *All of Us* is its emphasis on inclusivity, incorporating data from populations historically underrepresented in medical research, including racial minorities and rural communities. This approach not only improves the generalizability of findings but also provides insights into how vision diseases affect different populations. Additionally, with its extensive geographic data, *All of Us* allows for a spatial analysis of glaucoma prevalence, enabling researchers to assess environmental factors such as air and water quality in relation to health outcomes.

2.5 Current Research on Geographic Distribution of Glaucoma

Geographic distribution studies highlight notable patterns in glaucoma prevalence, often correlating with environmental and socioeconomic factors. For instance, rural areas may face higher rates of glaucoma-related blindness due to limited access to specialized eye care, while urban areas with high pollution levels may show an elevated prevalence of glaucoma

and other eye conditions [13].

One significant finding in recent literature is the disparity in vision health outcomes across racial and ethnic lines. African American and Hispanic populations show earlier onset and faster progression of POAG than their White counterparts, which can partially be attributed to environmental and social disparities [7]. This geographic variability, combined with known genetic risk factors, underscores the need for an integrative, data-driven approach to studying glaucoma, one that considers genetics, environment, and social factors within a spatial framework [16].

2.6 The Role of the *All of Us* Dataset in Glaucoma Prediction

The *All of Us* Research Program represents one of the largest, most diverse health databases in the world. By including health information, genetic data, and environmental exposure details from over a million participants, *All of Us* offers a unique opportunity to explore health patterns across a diverse cohort. This dataset allows for the analysis of glaucoma and other vision diseases alongside individual and community-level data, such as environmental exposures and SDOH. The dataset includes:

1. Electronic Health Records (EHRs) – Containing diagnostic codes, prescriptions, and medical histories.
2. Genetic Data – Allowing exploration of polygenic risk scores for glaucoma.
3. Demographic and SDOH Data – Essential for predictive modeling.
4. Geo-spatial Information – Useful for studying environmental exposures and healthcare access disparities.

By leveraging the *All of Us* dataset, this study aims to develop predictive models that integrate demographics, genetic predispositions, environmental factors, and SDOH to improve glaucoma risk assessment and early detection strategies.

2.7 Summary and Research Gaps

While existing studies provide valuable insights into the genetic, environmental, and social factors associated with glaucoma, gaps remain in understanding how these factors interact at the population level. Most studies have been constrained by either a lack of diverse datasets or limited environmental data. The use of the *All of Us* dataset allows this study to fill these gaps, to some extent (since the ZIP codes are truncated), by providing a comprehensive view of glaucoma distribution across the U.S. in relation to environmental and social determinants.

To date, limited research has focused on the cumulative effects of demographic information and social determinants on glaucoma within a geographic context. This research aims to bridge that gap by conducting a spatial analysis of glaucoma in the U.S., linking disease prevalence with geographic, environmental, and social variables. This approach can reveal new insights into the causes and risk factors of glaucoma, potentially leading to targeted interventions that improve vision health outcomes across diverse populations.

Chapter 3

Methodology

3.1 Study Design

This study employs a retrospective observational design using data from the All of Us (AoU) Research Program, a large-scale initiative aimed at advancing precision medicine by collecting diverse health data from over one million participants. It integrates clinical, environmental, and socioeconomic variables to assess their impact on glaucoma prevalence and risk stratification. The research also visualizes the geographical distribution of glaucoma across the United States and develops machine learning models to evaluate risk. All analyses were conducted on the AoU Researcher Workbench, a secure, cloud-based platform that provides approved researchers with access to the program's extensive datasets.

The AoU dataset comprises three primary data types: surveys, physical measurements (PMs), and electronic health records (EHRs). Detailed information about the surveys is available through the Survey Explorer, a tool within the Research Hub designed to assist researchers in navigating the data. The surveys employ branching logic, and all questions are optional, allowing participants to skip any they prefer not to answer. Physical measurements recorded at enrollment include systolic and diastolic blood pressure, height, weight, heart rate, waist and hip circumference, wheelchair use, and current pregnancy status. For participants who consented, EHR data were linked to provide additional clinical context.

All three data types – surveys, PMs, and EHRs – are mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) version 5.2, a stan-

dardized framework maintained by the Observational Health Data Sciences and Informatics (OHDSI) collaborative. This standardization ensures interoperability and facilitates large-scale analyses across diverse datasets.

To safeguard participant privacy, the AoU program applies a series of data transformations. These include:

- Data suppression: Removing codes with a high risk of identification, such as military status.
- Generalization: Aggregating categories for sensitive variables, including age, sex at birth, gender identity, sexual orientation, and race.
- Date shifting: Applying a random offset of less than one year to dates, consistently across each participant's record.

Detailed documentation on privacy measures and the creation of the Curated Data Repository is available in the AoU Registered and Controlled Tier Curated Data Repository Data Dictionary. The Researcher Workbench provides a suite of tools designed to streamline data analysis:

- Cohort Builder: Enables researchers to select groups of participants based on specific criteria.
- Dataset Builder: Facilitates the creation of customized datasets for analysis.
- Workspaces: Offers Jupyter Notebooks for advanced data analysis, supporting both R and Python 3 programming languages. These notebooks allow researchers to work with saved datasets or query the data directly, providing flexibility for complex analyses.

At the time of this thesis report, the AoU Research Program included a total of 413,457 adult participants. To identify individuals with glaucoma, cohort selection was performed using the Systematized Nomenclature of Medicine (SNOMED) code 23986001, which corresponds to the standard concept for glaucoma. It is important to note that there are additional SNOMED codes related to glaucoma, including non-standard codes and those representing specific subtypes of glaucoma. This initial filtering step allowed us to refine the dataset to include only participants with relevant glaucoma-related clinical records. This gave 19,130 individuals with glaucoma related illness.

Figure 3.1 below illustrates the distribution of data across different racial groups in the AoU dataset. The majority of participants are White, accounting for 55.4% of the dataset, followed by African American participants at 19.0%. Other racial groups are represented at lower percentages.

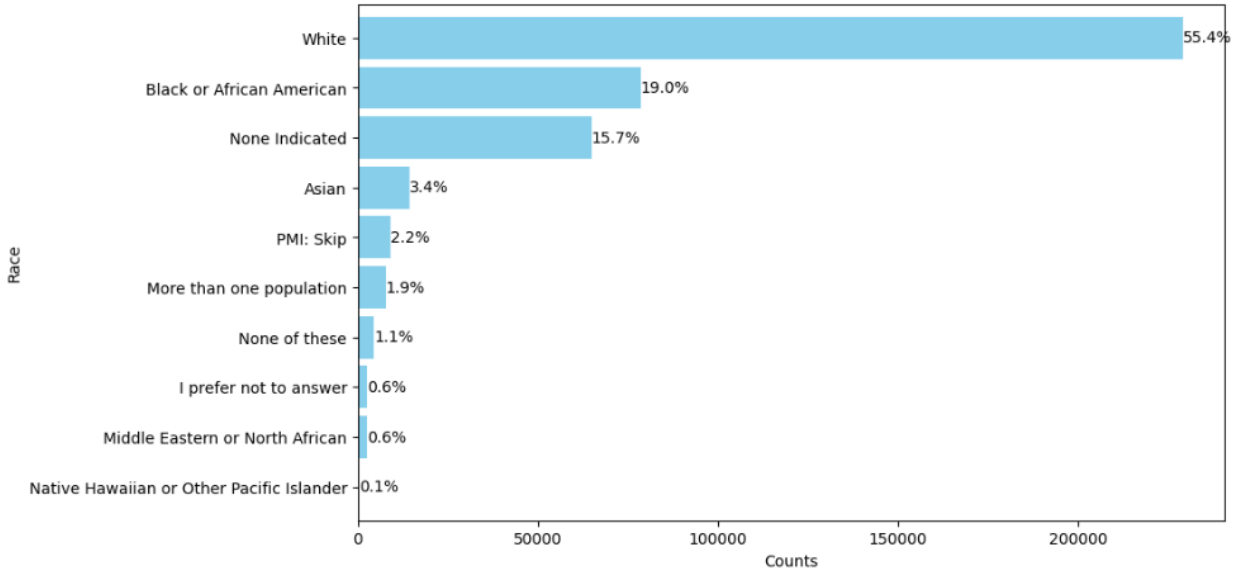

Figure 3.1: *All of Us* data distribution by race

Gender representation is a crucial aspect of the dataset. Figure 3.2 illustrates the distribution of participants across different gender identities. Female participants make up the majority at 59.8%, followed by male participants at 37.3%. The remaining 2.9% either skipped the gender question or selected an alternative response.

Age distribution is another crucial aspect of the dataset. Figure 3.3 illustrates the age distribution in the AoU dataset, showing that all participants are adults, with ages ranging from 20 to 120 years. The highest concentration of participants falls between 58 and 70 years.

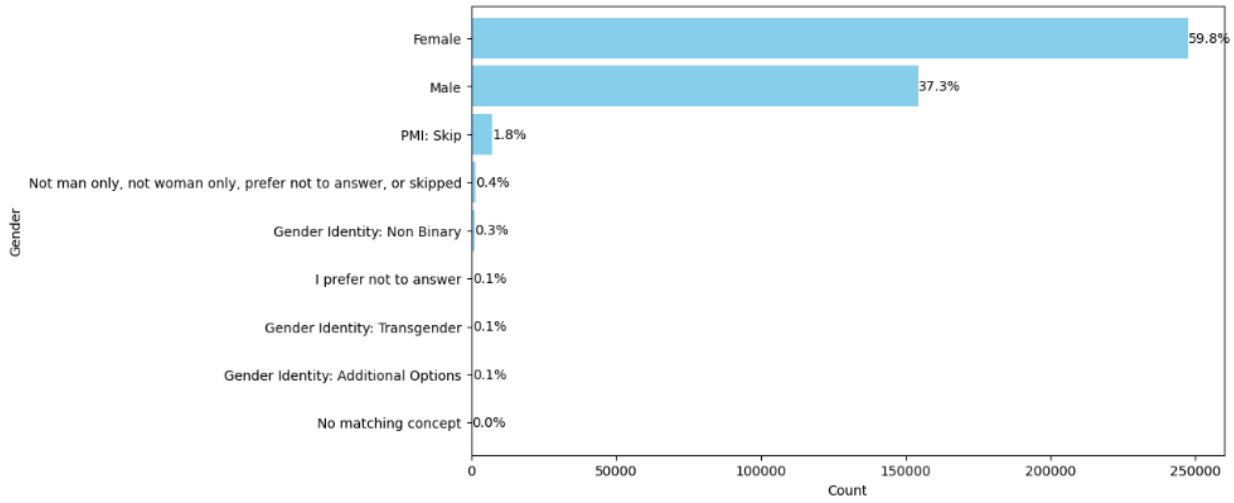

Figure 3.2: *All of Us* data distribution by gender

3.2 Descriptive Analysis

3.2.1 Geo-Spatial Analysis

To incorporate socioeconomic data into the analysis, we extracted zip code-level information for participants in the glaucoma cohort. This dataset includes basic demographic details along with three-digit zip codes formatted as '000**.' The use of three-digit zip codes is a de-identification measure implemented by the AoU Research Program to ensure participant privacy by preventing identification at a more granular geographic level.

To approximate the geographic distribution of individuals in the cohort, we utilized publicly available U.S. zip code data [18], which was downloaded online and uploaded into the AoU Researcher Workbench notebook environment. This dataset contains all U.S. five-digit zip codes, along with corresponding latitude, longitude, city, state, and additional geographic attributes.

Since only three-digit zip codes were available for the glaucoma cohort, we estimated approximate participant locations by calculating centroid coordinates for each three-digit zip region. The centroid was determined by averaging the latitude and longitude values of all five-digit zip codes that fall under the same three-digit zip designation. For example, if a three-digit zip code corresponds to seven different five-digit zip codes, the centroid was

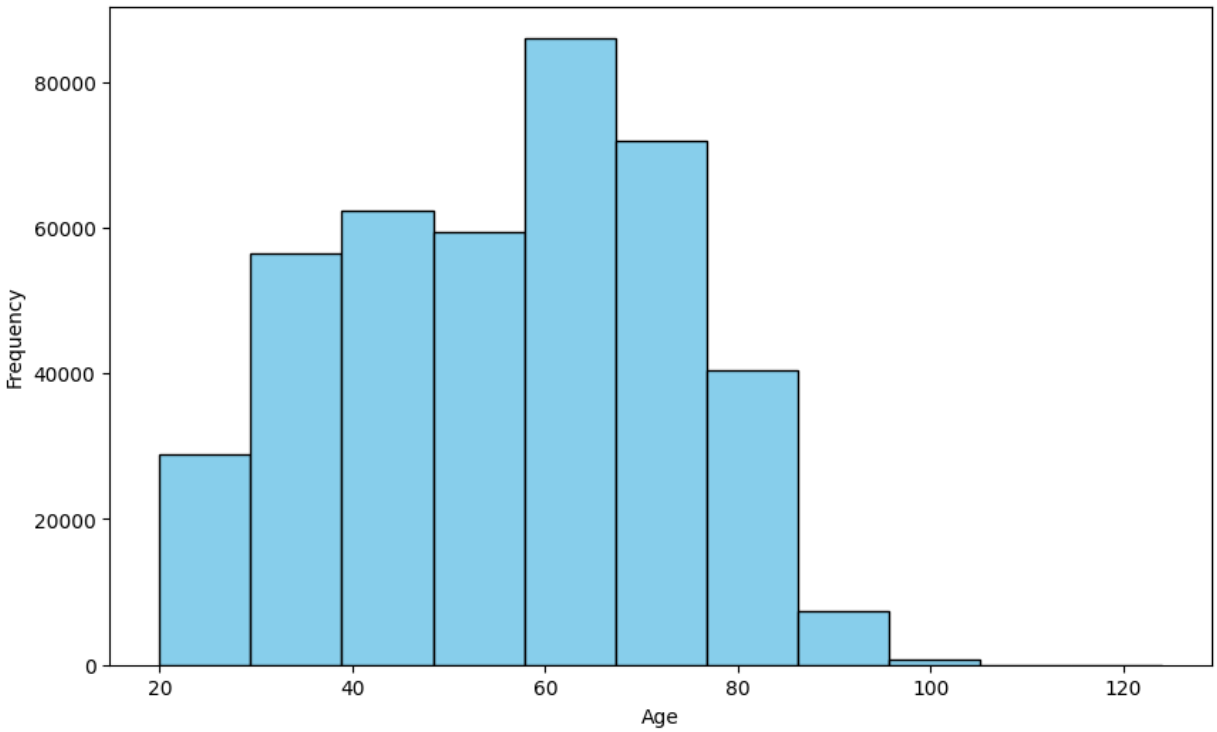

Figure 3.3: *All of Us* data distribution by age

computed by taking the mean latitude and longitude of all seven locations. This calculated centroid was then matched to the closest corresponding five-digit zip code, serving as the approximate geographic location of individuals in the cohort.

The resulting approximate zip code locations were used to generate a geo-spatial visualization in Power BI to illustrate the distribution of glaucoma cases across the United States. However, AoU has strict privacy policies prohibiting the publication of aggregated data with fewer than 20 individuals per group. Therefore, any zip code regions with fewer than 20 participants were excluded from the final visualization to comply with AoU data privacy regulations.

The number of glaucoma recorded per county was plotted and glaucoma prevalence per 1,000,000 population was also visualized. The plot is presented in Figure 4.1 and 4.2 in Chapter 4

3.3 Predictive Analysis

3.3.1 Processing the Cases Group

To analyze the relationship between demographic, socioeconomic, and lifestyle factors with glaucoma prevalence, multiple survey datasets were extracted and processed from the AoU Research Program. The primary focus was on basic lifestyle factors, including:

- Annual income
- Health insurance
- Educational attainment
- Employment status
- Housing

Additionally, a survey on healthcare access was included, which assessed:

- Healthcare affordability
- Prescription affordability
- Affordability of specialist
- Affordability of eye doctor
- Affordability of co-pays

A stress survey was also incorporated to evaluate the self-reported stress levels of each participant. Alongside these variables, demographic data, including age, gender, and race were extracted for individual.

All 19,130 glaucoma cohort participants responded to basic lifestyle surveys with the exception of the health insurance survey with 125 missing values from the cohort. 9,587 participants responded to the health access survey while 6,495 individuals had completed the stress survey. Imputation was used to handle the 125 missing data for health insurance by putting the values to zero, an outer join was performed between the survey datasets and the demographic data using `person_id` as the unique identifier. This integration resulted

in a final dataset of 19,130 with 60% missing values in the stress column and 50% missing values in the health care access surveys columns. Due to the large number of missing values in the variables of interest, we removed the rows with the missing data, leaving us with 5,762 rows of data for model training. Figure 3.4 illustrates the distribution of glaucoma cases across racial groups in the AoU dataset. White participants account for 51.3% of glaucoma cases, while African American participants make up 26%. Figure 3.5 illustrates percentage of people living with glaucoma in each race. The African American has a higher percentage of people with glaucoma despite their lower total participants compared to the white population, a further assertion that glaucoma is more prevalence in African population than other populations.

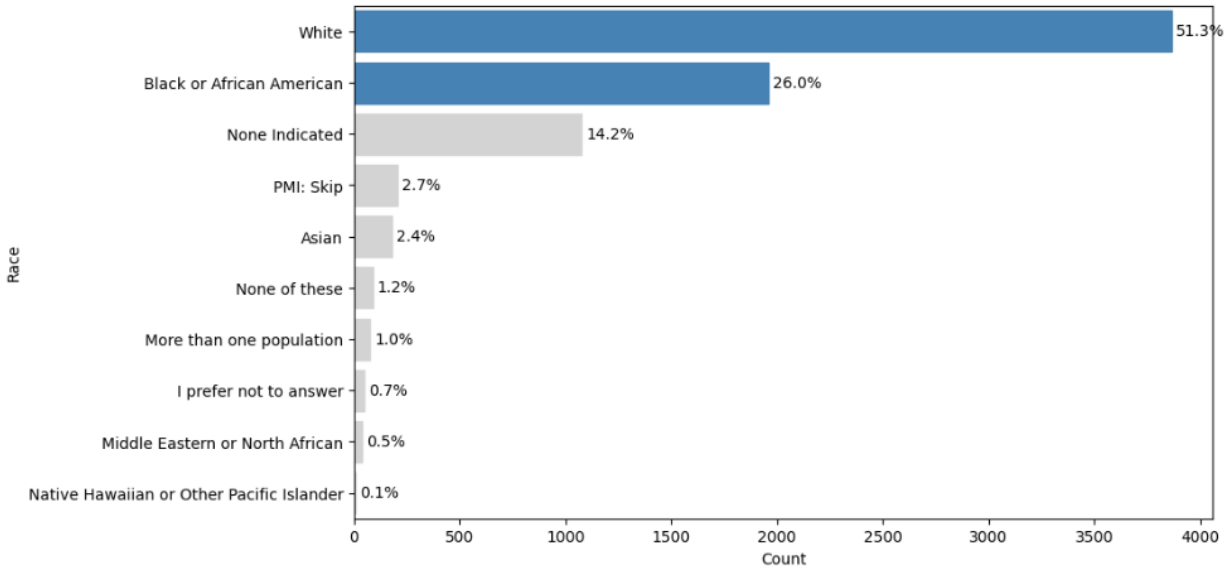

Figure 3.4: *All of Us* dataset glaucoma distribution by race

3.3.2 Processing the Control Group

A similar method was applied to construct the control cohort (individuals without glaucoma). Demographic, lifestyle, healthcare access, and stress survey responses were extracted for 20,000 individuals who met the exclusion criteria for glaucoma. Following the same data integration approach, the final control dataset was reduced to 5,129 participants af-

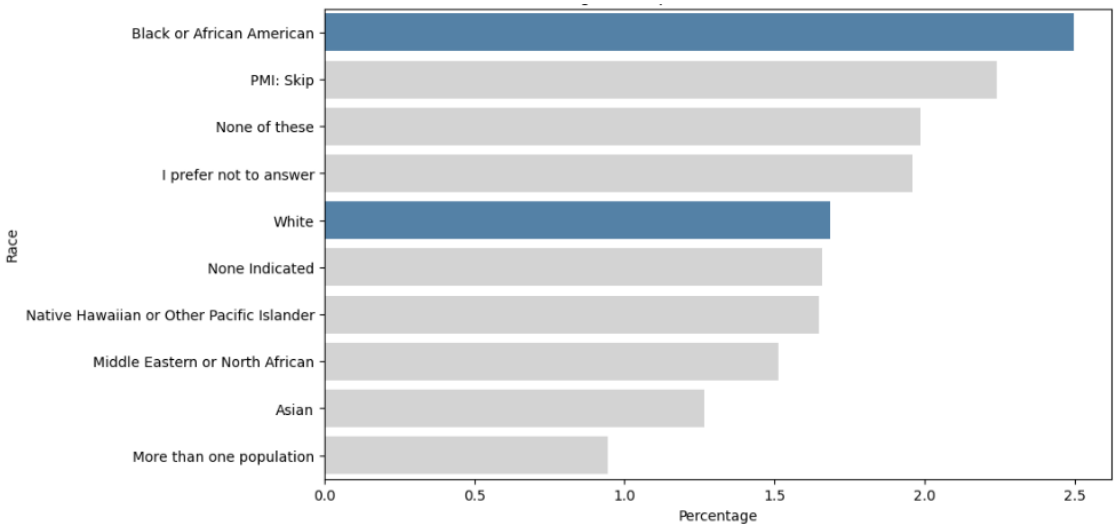

Figure 3.5: *All of Us* dataset Percentage of people with glaucoma in each race

ter merging with survey responses using an outer join and removing the rows with missing values.

To ensure compatibility with machine learning models, all categorical variables were converted into numeric format using mapping method.

3.3.3 Feature Engineering

A total of 14 predictor variables were selected and processed as follows:

1. Age – Continuous numerical variable

2. Gender – Coded as Male and Female (EHR data)

3. Race – Grouped into White, African American, and Others

    - Only the binary category of white and black was used. This categorization was used because the majority of participants in the dataset were White or African American, while other racial groups had significantly fewer participants.
    - Simplifying race to binary form helps reduce dimensionality and sparsity, particularly when minority categories have very low representation.
    - Hispanics are not in race category but ethnicity and this was not included in the variable for the model.

4. Health Insurance – Binary classification:

    - Has insurance (1)
    - No insurance (0)

5. Education Level – Categorized into three groups:

    - Advanced/College degree (2)
    - High school graduate (1)
    - No formal education or incomplete high school (0)

6. Employment Status – Binary classification:

    - Employed (1)
    - Unemployed (0)
    - The survey originally had eight employment categories: retired, out of work (more than a year), out of work (less than a year), homemaker, unable to work, student, employed, and other. These responses were consolidated into a binary employed/unemployed variable.

7. Income Level – Grouped into three categories:

    - Income $\geq$ \$75,000 (2)
    - Income: \$35,001–\$75,000 (1)
    - Income $\leq$ \$35,000 (0)

8. Medication Affordability – Binary classification: Yes (1) / No (0)

    - Originally derived from survey questions regarding prescription affordability.

9. Access to health care provider – Binary classification: Has access (1) / No access (0)

    - This variable was constructed from survey responses about barriers to healthcare access through affordability of the healthcare provider.

10. Ability to Afford Co-Pay – Binary classification: Yes (1) / No (0)

11. Ability to Afford Specialist – Binary classification: Yes (1) / No (0)

12. Spoken to Eye Doctor – Binary classification: Yes (1) / No (0)

13. Stress Level – Binary classification: Stressed (1) / Not stressed (0)

    - Derived from survey responses assessing perceived stress levels among participants.

3.3.4 Final Dataset Construction

The processed case (glaucoma) and control (non-glaucoma) datasets were merged into a single dataset, with glaucoma status as the target variable. The outcome was encoded as a binary classification:

- Individual has glaucoma (1)
- Individual does not have glaucoma (0)

This structured dataset was then used for machine learning model development to predict glaucoma risk based on demographic, socioeconomic, and lifestyle variables.

Figure 3.6 shows the workflow for this study. After data extraction and cleaning, we have 10,891 participants with 5,762 cases and 5,129 controls.

3.4 Machine Learning Model for Glaucoma Prediction

To develop predictive models for glaucoma risk, several supervised machine learning algorithms were implemented. Prior to model development, multicollinearity among the predictor variables was assessed using a correlation matrix. The analysis revealed no significant multicollinearity, indicating that the variables were sufficiently independent for inclusion in the models. Additionally, hyperparameter tuning was conducted for the Decision Tree and Gradient Boosting models using GridSearchCV with 5-fold cross-validation, ensuring optimal model performance and generalizability.

3.4.1 Logistic Regression

A logistic regression model was trained as a baseline classifier due to its interpretability and efficiency in predicting binary outcomes (glaucoma vs. non-glaucoma). The model was

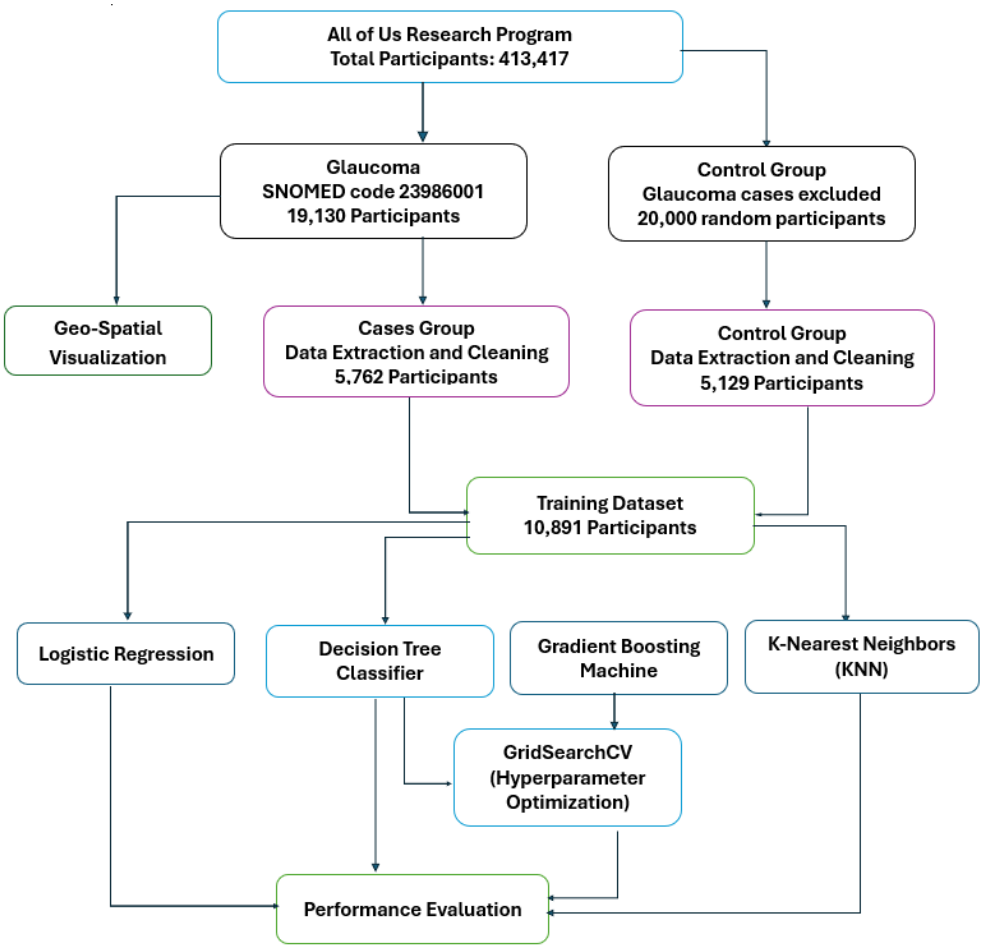

Figure 3.6: Study workflow and cohort definition for evaluating predictive models for participant with glaucoma in the *All of Us* Research Program

optimized using:

- Maximum Iterations: Set to 1000 for better convergence.
- The rest of the parameters were left at their default values, which are suitable for most binary classification tasks.
- Feature Importance Analysis: Used model coefficients to determine the most influential predictors.

3.4.2 Decision Tree Classifier

A Decision Tree Classifier was implemented to capture non-linear relationships between SDOH and glaucoma risk. Key steps included:

- Gini Impurity for Splitting: This was used to determine the best feature splits.
- Other parameters were left at their default values. These include the splitter, which is set to 'best' (the strategy that chooses the best split at each node), maximum depth set to None (allowing the tree to expand until all leaves are pure), minimum samples split set to 2, and minimum samples leaf set to 1.
- Feature Importance Analysis: We evaluated how each variable contributed to model predictions.
- Hyperparameter Optimization: Hyperparameter tuning was conducted using GridSearchCV with 5-fold cross-validation to identify the best combination of parameters for model generalization. The hyperparameters explored included maximum depth {3, 5, 10, None} to control tree depth and prevent overfitting, minimum samples split {2, 5, 10} to define the minimum number of samples required to split a node and minimum samples leaf {1, 2, 4} to specify the minimum number of samples at a leaf node. Other parameters, such as criterion='gini', splitter='best', and maximum features=None, were left at their default setting.

3.4.3 Gradient Boosting Machine (GBM)

The Gradient Boosting Machine (GBM) was employed to enhance predictive performance by sequentially improving weak learners through gradient-based optimization. Key steps included:

- Boosting Framework: Combined multiple weak decision trees to form a strong classifier, minimizing prediction errors iteratively. The model was configured with 100 estimators.
- Learning Rate Tuning: Adjusted the step size of model updates to balance convergence speed and overfitting. A learning rate of 0.1 was used for this model.
- All other parameters, including subsample (1.0), minimum samples split (2), minimum samples leaf (1), and maximum features (None), were left at their default settings. This configuration represents a typical baseline for gradient boosting models prior to hyperparameter tuning.
- Hyperparameter Optimization: The Gradient Boosting Classifier was optimized using GridSearchCV with 5-fold cross-validation, focusing on three hyperparameter: numnber of estimators (100 or 200), learning rate (0.01, 0.1, 0.2), and maximum depth (3, 5, 7). The final model was selected based on the highest AUC-ROC score. All other hyperparameters were retained at their default settings, including subsample of 1.0, minimum samples split of 2, and the criterion was 'friedman mse'. This approach balanced model complexity with performance, helping to avoid overfitting while maximizing predictive accuracy.

- Feature Importance Analysis: Assessed the contribution of each variable to the model's predictive power, identifying key SDOH factors influencing glaucoma risk.

3.4.4 K-Nearest Neighbors (KNN) Classifier

A K-Nearest Neighbors (KNN) model was used to analyze the proximity of glaucoma patients in feature space. The K-Nearest Neighbors (KNN) model was configured with number of neighbors neighbors of 5 to assess glaucoma prediction based on similarity in the feature space. The model used the Euclidean distance (metric='minkowski' with p=2) to compute similarities, and neighbors were uniformly weighted. Other parameters such as algorithm='auto', leaf size=30, and number of jobs=None were left at their default values. This setup allowed the model to classify test instances based on the majority class of the 5 closest neighbors in the transformed feature space.

3.4.5 Model Training and Validation

- Train-Test Split: The dataset was divided into 70% training and 30% testing using stratified sampling to ensure balanced representation of glaucoma cases vs. controls.
- Cross-Validation: Employed 5-fold cross-validation to optimize hyperparameters and prevent overfitting.
- Feature Selection: Identified the most influential predictors based on feature importance scores derived from trained models, ensuring the selection of meaningful variables for prediction.

3.4.6 Model Evaluation Metrics

To assess model performance, we used the following evaluation metrics:

- Accuracy: Measures overall model correctness.
- Precision and Recall: Evaluates the model's ability to identify true glaucoma cases.
- F1 Score: Balances precision and recall for better clinical applicability.
- AUC-ROC Curve: Assesses the discriminatory power of the model.

3.5 Ethical Considerations

Data Privacy and Security: All analyses were conducted within the secure *All of Us* Researcher Workbench to ensure compliance with HIPAA and NIH data-use policies. The study also evaluated model fairness across different racial, socioeconomic, and geographic groups to minimize bias and promote equitable predictions.

Chapter 4

Results and Discussion

4.1 Descriptive Analysis

4.1.1 Glaucoma Cases Distribution

Figure 4.1 Presents a map illustrating the distribution of glaucoma cases across different counties. Larger bubbles indicate areas with a higher number of reported glaucoma cases. The highest concentration appears to be in the vicinity of Cook County, Illinois (1,390 cases), followed by New York County, Manhattan (1,328 cases), and Suffolk County, Massachusetts (1,106 cases). Additionally, Allegheny County, Pennsylvania has 1,075 cases, while other counties report fewer than 1,000 cases. The locations are based on approximate ZIP code-level data and reveal that glaucoma cases are notably concentrated in coastal regions of the Midwest and Northeast. The high number of glaucoma in these areas is not only due to high population density, the prevalence per population is equally higher compared to other areas as shown in Figure 4.2.

4.1.2 Glaucoma Prevalence

The map of glaucoma prevalence in 1,000,000 population is illustrated in Figure 4.1. To better visualize this prevalence, Figure 4.3 illustrates the distribution of glaucoma prevalence per 1,000,000 population across various counties in the United States. The majority of counties (119) fall within the lowest prevalence category of 0–10 cases per 1,000,000 people. This is followed by 49 counties with prevalence between 10–20 cases, 36 counties in the 20–30

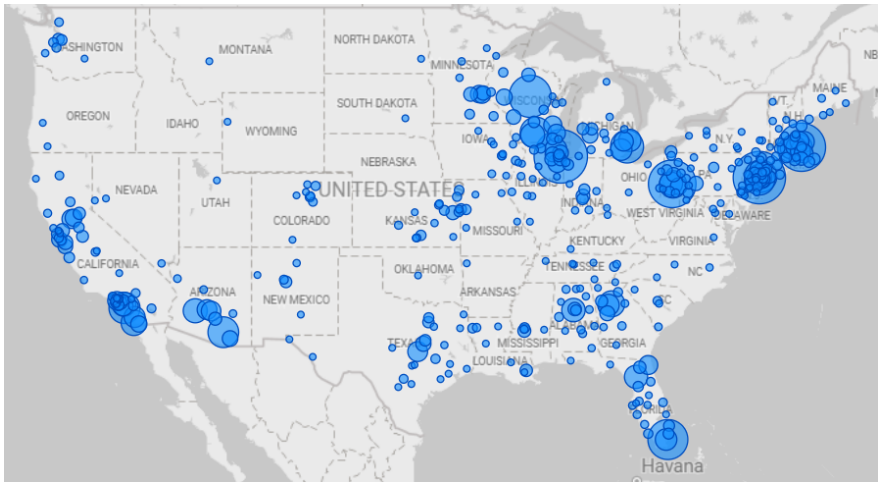

Figure 4.1: Map showing number of reported Glaucoma cases in the United States

range, and 32 counties in the 30–40 range. Notably, 57 counties exhibit prevalence between 100–200 cases per 1,000,000 population, marking a peak in mid-range prevalence.

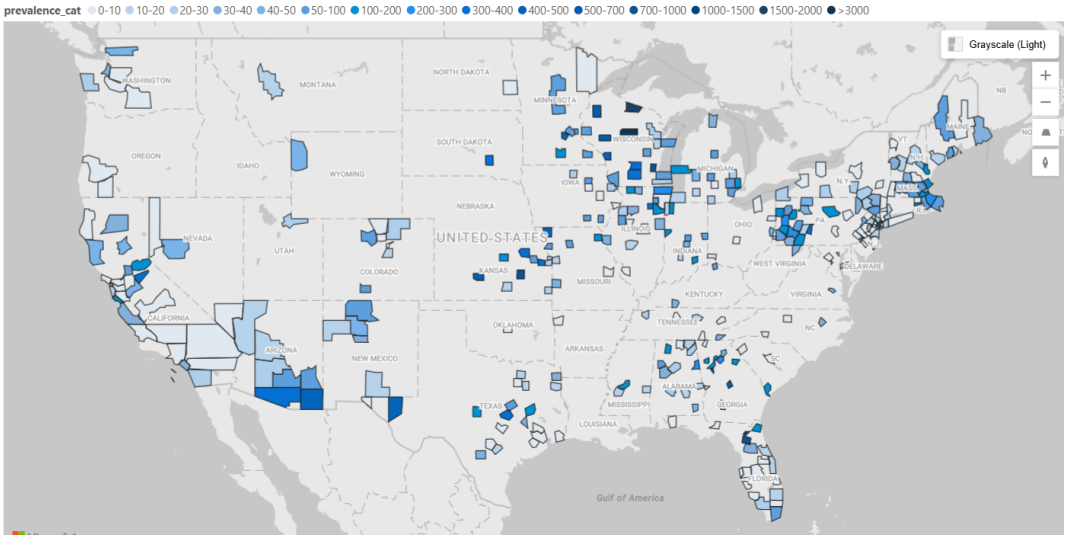

Figure 4.2: Map showing Glaucoma Prevalence per 1,000,000 Population

As prevalence increases, the number of counties in each category generally declines, with only a small number of counties experiencing higher rates. Specifically, only one county, Marathon in Wisconsin, records a prevalence exceeding 3,000 cases per 1,000,000 population.

In the higher prevalence categories, two counties in Wisconsin and two in Kansas fall within the 300–400 prevalence range. Additionally, Arizona, California, Illinois, New York, Pennsylvania, South Dakota, and Texas each have one county in this category. Notably,

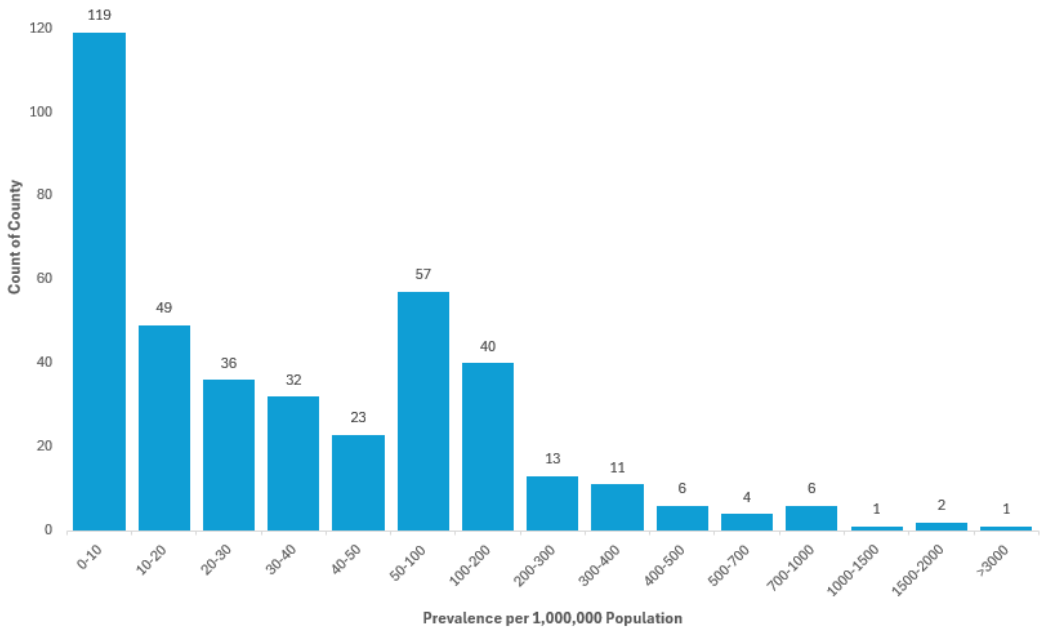

Figure 4.3: County Counts of Glaucoma Prevalence per 1,000,000 Population

Wisconsin has at least one county in most of the higher prevalence category from 300 and above. For the 700–1000 prevalence category, Wisconsin has two counties, while Florida, Kansas, New York, and Pennsylvania each have one county. Suffolk county in Massachusetts is the only county in 1000-1500 prevalence. In the 1500–2000 category, Florida and Wisconsin each have one county. This is illustrated in Table 4.1 below.

State	300-400	400-500	500-700	700-1000	1000-1500	1500-2000	>3000
Arizona	Pima						
California	Calaveras						
Floria				Alachua		Union	
Georgia		Fayette	Twiggs				
Illinois	Kendall						
Kansas	Douglas, Pottawatomie		Pawnee	Chase			
Massachusetts					Suffolk		
Minnesota		Carver					
Missouri		Nodaway					
New York	Bronx			New York			
Pennsylvania	Blair			Allegheny			
South Dakota	Aurora						
Texas	Bell	Culberson					
Wisconsin	Columbia, Dane	La Crosse	Eau Claire, Washburn	Dane, Washington		Vilas	Marathon

Table 4.1: High prevalence categories of counties in different states

4.2 Predictive Analysis

In this section, we present the results of the predictive models used to classify the presence of glaucoma based on demographic and SDOH features. We evaluate and compare the performance of Logistic Regression, Decision Tree, Ensemble Gradient Boosting, optimized Gradient Boosting with GridSearchCV, K-Nearest Neighbors (KNN) models in terms of accuracy, precision, recall, and feature importance.

4.2.1 Logistic Regression Model

The logistic regression model achieved an overall accuracy of 67.2%, indicating a moderate predictive ability with AUC score of 0.73 and ROC curve above the diagonal, indicating the model is better than random guessing. The confusion matrix reveals that the model correctly classified 852 negative cases (no glaucoma) and 1344 positive cases (glaucoma), while misclassifying 661 negative cases and 411 positive cases. The classification report further details the performance across both classes:

- Class 0 (No Glaucoma): Precision = 0.67, Recall = 0.56, F1-score = 0.61
- Class 1 (Glaucoma): Precision = 0.67, Recall = 0.77, F1-score = 0.71
- Overall weighted performance: Precision = 0.67, Recall = 0.67, F1-score = 0.67

The feature importance analysis reveals that affording an eye doctor (0.5629) and having insurance (0.5493) were the most influential predictors of glaucoma classification. Other notable predictors included affording prescriptions (0.1999), gender (0.1132), and stress levels (0.0615). Interestingly, race (-0.3214), income (-0.1316), and education (-0.1021) had negative coefficients, suggesting an inverse relationship with the predicted outcome. This model shows moderate predictive performance with an AUC of 0.73, meaning it is fairly good at distinguishing between glaucoma and non-glaucoma cases. It performs better at detecting glaucoma cases but has lower recall for non-glaucoma cases, meaning some negative cases might be misclassified.

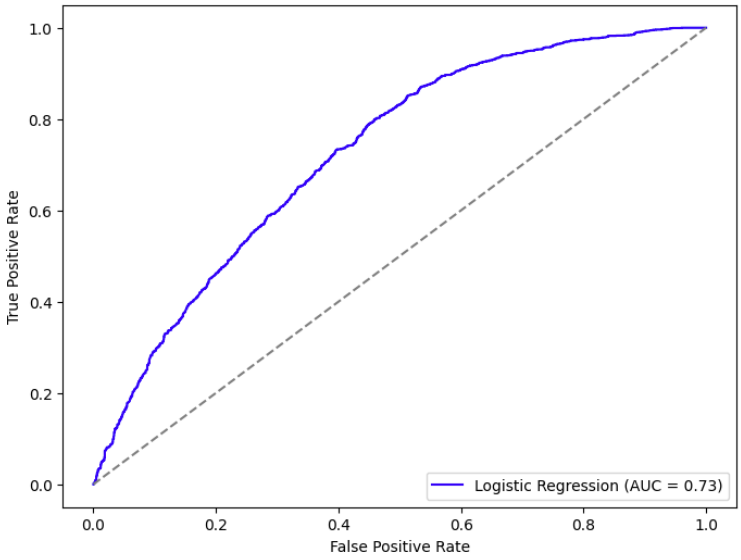

Figure 4.4: AUC-ROC Curve for Logistic Regression

4.2.2 Decision Tree Model

The decision tree model yielded an accuracy of 57.2%, performing notably worse than logistic regression. The confusion matrix showed that the model correctly classified 865 negative cases and 1004 positive cases, while misclassifying 648 negative cases and 751 positive cases. The classification report for this model is as follows:

- Class 0 (No Glaucoma): Precision = 0.54, Recall = 0.57, F1-score = 0.55
- Class 1 (Glaucoma): Precision = 0.61, Recall = 0.57, F1-score = 0.59
- Overall weighted performance: Precision = 0.57, Recall = 0.57, F1-score = 0.57

The feature importance rankings in the decision tree model differed significantly from logistic regression. Here, age (0.4357) was the most important predictor, followed by income (0.1010), education (0.0641), and gender (0.0585). Interestingly, affording an eye doctor (0.0203) and having insurance (0.0159) were less influential in this model, which contrasts with their strong predictive power in logistic regression. The model struggles to differentiate between glaucoma and non-glaucoma cases, as seen in the near-random ROC curve in Figure 4.5.

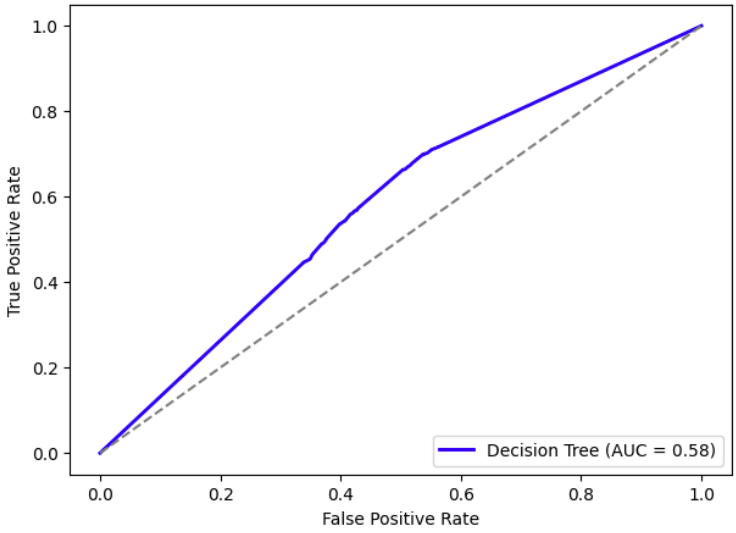

Figure 4.5: AUC-ROC Curve for Decision Tree Model

4.2.3 Ensemble Gradient Boosting Classifier

The Ensemble Gradient Boosting model improved upon previous models with an accuracy of 67.29%. The confusion matrix showed that it correctly classified 842 non-glaucoma cases and 1357 glaucoma cases, while misclassifying 671 non-glaucoma and 398 glaucoma cases. The classification report was as follows:

- Class 0 (Non-Glaucoma): Precision = 0.68, Recall = 0.56, F1-score = 0.61
- Class 1 (Glaucoma): Precision = 0.67, Recall = 0.77, F1-score = 0.72

Feature importance analysis highlighted age (0.7767) as the most significant predictor, followed by race (0.0878) and affordability of an eye doctor (0.0617). Factors such as affordability of specialist care (0.0006) and copay affordability (0.0011) had minimal impact. This model significantly outperforms Decision Tree, achieving higher accuracy (67%) and a stronger AUC-ROC (0.73) with moderate discriminatory power and it does a better job at identifying glaucoma cases (77% recall).

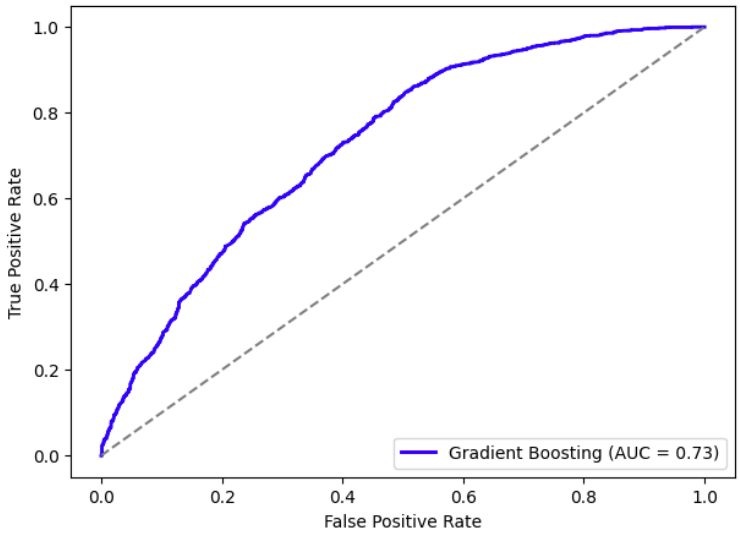

Figure 4.6: AUC-ROC Curve for Gradient Boosting Classifier

4.2.4 Optimized Decision Tree with GridSearchCV

Using hyperparameter tuning, the optimized Decision Tree model improved the accuracy to 67.87%. The best parameters were:

- max depth: 5
- min sample leaf: 1
- min sample split: 2

The confusion matrix showed that 790 non-glaucoma and 1428 glaucoma cases were correctly classified, while 723 non-glaucoma and 327 glaucoma cases were misclassified. The classification report indicated:

- Class 0 (Non-Glaucoma): Precision = 0.71, Recall = 0.52, F1-score = 0.60
- Class 1 (Glaucoma): Precision = 0.66, Recall = 0.81, F1-score = 0.73

The recall for class 1 (glaucoma cases) is the highest so far (81%), making this model the best at detecting glaucoma among the models tested. AUC-ROC of 0.71 shows good discrimination ability, though slightly lower than the previous 0.73. This optimized model performs well, especially in identifying glaucoma cases (high recall). It provides a solid balance between accuracy and recall, making it a strong model for glaucoma prediction.

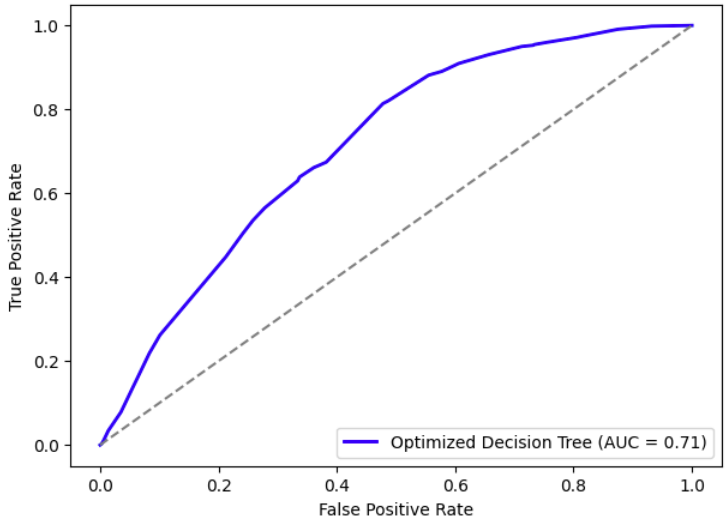

Figure 4.7: AUC-ROC Curve for Optimized Decision Tree

4.2.5 Optimized Gradient Boosting with GridSearchCV

The optimized Gradient Boosting model demonstrated similar predictive performance to the original Gradient Boosting model in identifying glaucoma cases, achieving an accuracy of 67.29% and an AUC-ROC score of 0.73, indicating a moderate discrimination ability between cases and non-cases. The model was fine-tuned with a learning rate of 0.1, max depth of 3, and 100 estimators, optimizing the trade-off between bias and variance.

The confusion matrix revealed that the model correctly identified 1,357 glaucoma cases (true positives) and 842 non-cases (true negatives). However, 398 glaucoma cases were misclassified as non-cases (false negatives), and 671 non-cases were incorrectly classified as glaucoma (false positives).

Despite the hyperparameter tuning, the optimized Gradient Boosting model did not show a significant improvement over the original model in terms of accuracy, recall, or AUC-ROC. While the model maintains strong recall for glaucoma cases, its overall predictive ability remains comparable to the untuned version.

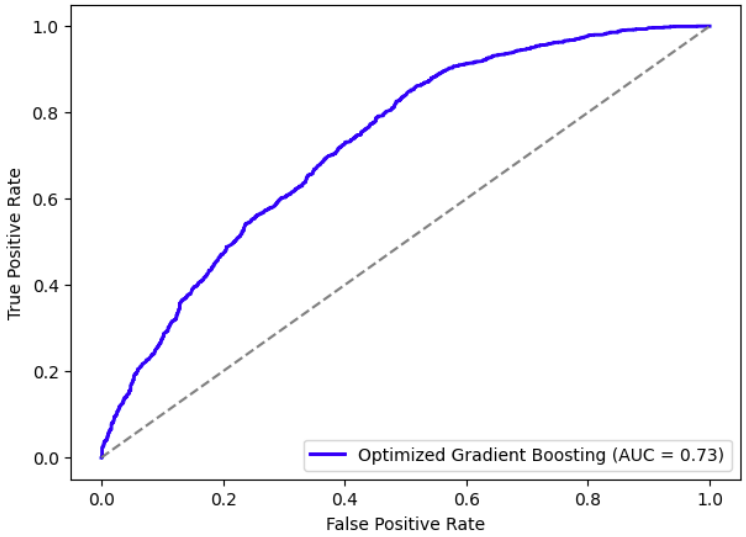

Figure 4.8: AUC-ROC Curve for Optimized Gradient Boosting Classifier

4.2.6 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) model was evaluated for its predictive performance in identifying glaucoma cases. The model achieved an accuracy of 61.57% and an AUC-ROC score of 0.65, indicating moderate discrimination ability between glaucoma and non-glaucoma cases. The confusion matrix revealed that the model correctly classified 1,189 glaucoma cases (true positives) and 823 non-cases (true negatives). However, 566 glaucoma cases were misclassified as non-cases (false negatives), and 690 non-cases were incorrectly classified as glaucoma (false positives). The classification report was:

- Class 0 (Non-Glaucoma): Precision = 0.59, Recall = 0.54, F1-score = 0.57
- Class 1 (Glaucoma): Precision = 0.63, Recall = 0.68, F1-score = 0.65

The AUC-ROC curve Figure 4.9 shows that the model's ability to differentiate between glaucoma and non-glaucoma cases is moderate, with an AUC of 0.65. This indicates that while the model performs better than random guessing (AUC = 0.50), its discriminatory power remains limited.

This model underperformed compared to Gradient Boosting and Logistic Regression, likely due to its sensitivity to feature scaling and data distribution.

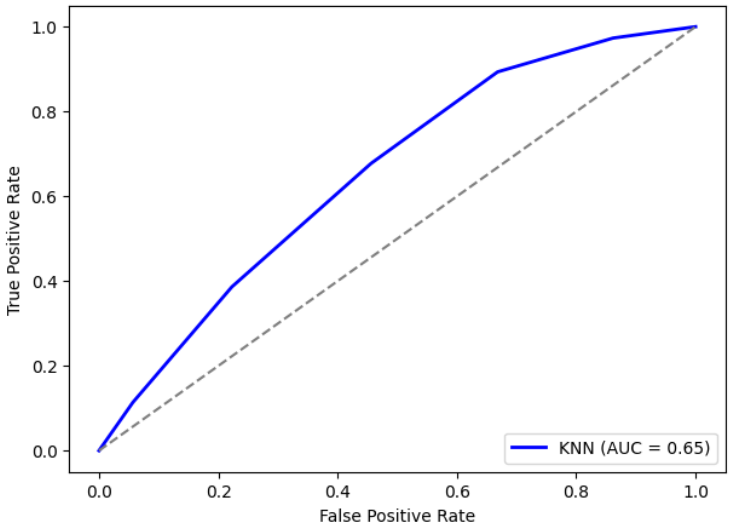

Figure 4.9: AUC-ROC Curve for k-Nearest Neighbors (KNN)

4.3 Model Comparison

Table 4.2 presents the evaluation metrics for all the machine learning models employed in this study. Each model's accuracy, along with class-specific precision, recall, F1 scores, and the AUC-ROC scores, provides a comprehensive comparison at a glance.

Among the models, Logistic Regression and the Ensemble Gradient Boosting classifier performed similarly, each achieving approximately 67% accuracy and an AUC of 0.73. In contrast, the K-Nearest Neighbors (KNN) model had lower performance, with an accuracy of 61.57% and AUC of 0.65, indicating its limited effectiveness for this classification task. The basic Decision Tree model recorded the lowest performance (accuracy: 57.2%, AUC: 0.58), which emphasizes the importance of hyperparameter optimization.

Most models demonstrated better performance on Class 1 (glaucoma cases), with higher recall and F1 scores, suggesting they were more sensitive to detecting positive cases. However, precision for Class 0 (non-glaucoma cases) was generally lower, particularly in the optimized models, indicating a higher rate of false positives.

The Optimized Decision Tree, tuned using GridSearchCV, achieved the highest overall accuracy (67.87%) with optimal hyperparameters: maximum depth of 5, minimum samples

per leaf of 1, and minimum samples required to split of 2. This model's improved performance can be attributed to its ability to capture more complex patterns while avoiding overfitting.

Notably, the optimized decision tree achieved a recall of 81% for glaucoma cases, making it particularly effective for identifying individuals at risk, a critical requirement in medical screening. Additionally, the F1 score of 0.73 for Class 1 (glaucoma) reflects a strong balance between precision and recall. Although its AUC score of 0.71 was slightly lower than that of the Gradient Boosting model, the overall performance highlights its value as a practical and interpretable screening tool for glaucoma detection.

Model	Accuracy (%)	Precision (Class 0) Non-Glaucoma	Recall (Class 0) Non-Glaucoma	F1 Score (Class 0) Non-Glaucoma	Precision (Class 1) Glaucoma	Recall (Class 1) Glaucoma	F1 Score (Class 1) - Glaucoma	AUC
Logistic Regression	67.2	0.67	0.56	0.61	0.67	0.77	0.71	0.73
Decision Tree	57.2	0.54	0.57	0.55	0.61	0.57	0.59	0.58
Ensemble Gradient Boosting	67.29	0.68	0.56	0.61	0.67	0.77	0.72	0.73
Optimized Decision Tree (GridSearchCV)	67.87	0.71	0.52	0.60	0.66	0.81	0.73	0.71
Optimized Gradient Boosting (GridSearchCV)	67.29	0.68	0.56	0.61	0.67	0.77	0.72	0.73
LK-Nearest Neighbors (KNN)	61.57	0.59	0.54	0.57	0.63	0.68	0.65	0.65

Table 4.2: Model Evaluation and Comparison

Chapter 5

Conclusion and Future Work

5.1 Descriptive Analysis

The descriptive analysis revealed a significant concentration of counties with lower glaucoma prevalence, with only a few counties exhibiting markedly higher rates. This uneven distribution suggests potential geographic disparities in disease prevalence, likely influenced by factors such as healthcare access, socioeconomic status, demographic composition, and environmental conditions. These findings underscore the need for targeted public health interventions and improved access to eye care services in high-risk areas.

Furthermore, the observed distribution raises important questions about systemic barriers to diagnosis and treatment. Future studies should examine the interaction between social determinants of health (SDOH), regional healthcare infrastructure, and glaucoma prevalence to better inform policy decisions. Additionally, this approach could be extended to investigate the spatial distribution of other vision-related diseases, such as cataracts, ocular hypertension, and presbyopia, to develop comprehensive eye health strategies.

5.2 Predictive Analysis

The predictive modeling phase provided critical insights into glaucoma risk factors and model performance. Among all tested models, the Optimized Decision Tree demonstrated the highest accuracy (67.87%) and the most balanced trade-off between precision and recall. Its superior performance is likely due to its ability to capture complex patterns in the data while

mitigating overfitting.

Gradient Boosting and Logistic Regression also exhibited strong predictive performance, achieving approximately 67.2% accuracy, reinforcing their reliability in classification tasks. Conversely, models such as K-Nearest Neighbors (KNN) and the baseline Decision Tree displayed lower predictive power, likely due to higher sensitivity to noise and suboptimal handling of feature interactions.

A key takeaway from the feature importance analysis was the consistent identification of age, race, and affordability of eye care as the most influential predictors of glaucoma. This underscores the profound impact of socioeconomic factors on disease risk, reinforcing the need for integrated screening strategies that incorporate both clinical and social risk factors.

5.3 Future Research Directions

While the current study provides valuable insights into glaucoma prediction, several areas warrant further investigation to enhance model robustness and generalizability:

1. Expanding the Dataset: Future research should leverage larger, more diverse datasets to improve model generalization and account for population heterogeneity.

2. Incorporating Genetic and Environmental Factors: Given the known heritability of glaucoma, integrating genetic markers and environmental exposures (e.g., air pollution, water contaminants, and PFAS exposure) could significantly enhance predictive accuracy.

3. Developing Clinical Decision Support Tools: Translating predictive models into practical decision-support tools for ophthalmologists and public health officials could enhance early detection and intervention strategies.

4. Refining Model Interpretability: Advanced machine learning techniques, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Ex-

planations), could provide deeper insights into individual risk factors, making models more useful for personalized risk assessment.

5.4 Final Remark

This study highlights the potential of machine learning-based predictive modeling in understanding glaucoma risk and guiding targeted screening efforts. By integrating social determinants, refining model interpretability, and expanding datasets, future research can contribute to more equitable and effective glaucoma prevention and treatment strategies.

BIBLIOGRAPHY

[1] Aboobakar, I. F., and Wiggs, J. L. The genetics of glaucoma: Disease associations, personalised risk assessment and therapeutic opportunities-a review. *Clinical and Experimental Ophthalmology 50*, 2 (March 2022), 143–162.

[2] Almarzouki, N. Impact of environmental factors on glaucoma progression: A systematic review. *Clinical Ophthalmology 18* (2024), 2705–2720.

[3] Baxter, S. L., Saseendrakumar, B. R., Paul, P., Kim, J., Bonomi, L., Kuo, T.-T., Loperena, R., Ratsimbazafy, F., Boerwinkle, E., Cicek, M., et al. Predictive analytics for glaucoma using data from the all of us research program. *American journal of ophthalmology 227* (2021), 74–86.

[4] Blindness, G. ., Collaborators, V. I., and of the Global Burden of Disease Study, V. L. E. G. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to vision 2020: the right to sight: an analysis for the global burden of disease study. *The Lancet Global Health 9*, 2 (February 2021), e144–e160. Epub 2020 Dec 1. Erratum in: Lancet Glob Health. 2021 Apr;9(4):e408. doi: 10.1016/S2214-109X(21)00050-4.

[5] Chen, K. W., Jiang, A., Kapoor, C., Fine, J. R., Brandt, J. D., and Chen, J. Geographic information system mapping of social risk factors and patient outcomes of pediatric glaucoma. *Ophthalmology Glaucoma 6*, 3 (May-June 2023), 300–307. Epub 2022 Nov 23.

[6] Dada, T., Verma, S., Gagrani, M., Bhartiya, S., Chauhan, N., Satpute, K., and Sharma, N. Ocular and systemic factors associated with glaucoma. *Journal of Current Glaucoma Practice 16*, 3 (September-December 2022), 179–191.

[7] Davuluru, S. S., Jess, A. T., Kim, J. S. B., Yoo, K., Nguyen, V., and Xu, B. Y. Identifying, understanding, and addressing disparities in glaucoma care in the united states. *Translational Vision Science & Technology 12*, 10 (October 2023), 18.

[8] Ehrlich, J. R., Burke-Conte, Z., Wittenborn, J. S., Saaddine, J., Omura, J. D., Friedman, D. S., Flaxman, A. D., and Rein, D. B. Prevalence of glaucoma among us adults in 2022. *JAMA Ophthalmology 142*, 11 (Nov 2024), 1046–1053.

[9] Elam, A. R., Tseng, V. L., Rodriguez, T. M., Mike, E. V., Warren, A. K., Coleman, A. L., and American Academy of Ophthalmology Taskforce on

Disparities in Eye Care. Disparities in vision health and eye care. *Ophthalmology 129*, 10 (2022), e89–e113.

[10] Han, X., Gharahkhani, P., Hamel, A. R., Ong, J.-S., Rentería, M. E., Mehta, P., Dong, X., Pasutto, F., Hammond, C., Young, T. L., Hysi, P., Lotery, A. J., Jorgenson, E., Choquet, H., Hauser, M., Bailey, J. N. C., Nakazawa, T., Akiyama, M., Shiga, Y., Fuller, Z. L., Wang, X., Hewitt, A. W., Craig, J. E., Pasquale, L. R., Mackey, D. A., Wiggs, J. L., Khawaja, A. C., Segrè, A. V., 23andMe Research Team, Consortium, I. G. G., and MacGregor, S. Large-scale multitrait genome-wide association analyses identify hundreds of glaucoma risk loci. *Nature Genetics 55*, 7 (July 2023), 1116–1125. Epub 2023 Jun 29.

[11] Jonas, J. B., Aung, T., Bourne, R. R., Bron, A. M., Ritch, R., and Panda-Jonas, S. Glaucoma. *The Lancet 390*, 10108 (2017), 2183–2193.

[12] Karimi, A., Stanik, A., Kozitza, C., and Chen, A. Integrating deep learning with electronic health records for early glaucoma detection: A multi-dimensional machine learning approach. *Bioengineering 11*, 6 (2024), 577.

[13] Mamidipaka, A., Shi, A., Lee, R., et al. Socioeconomic and environmental factors associated with glaucoma in an african ancestry population: findings from the primary open-angle african american glaucoma genetics (poaagg) study. *Eye* (2024).

[14] Organization, W. H. Social determinants of health and their impact. *WHO Report* (2023).

[15] Quigley, H. A., and Broman, A. T. The number of people with glaucoma worldwide in 2010 and 2020. *British Journal of Ophthalmology 90*, 3 (March 2006), 262–267.

[16] Rahman, M., et al. Integrating social determinants in glaucoma risk prediction. *Public Health Ophthalmology Journal* (2023).

[17] Sekimitsu, S., Elze, T., and Zebardast, N. Impact of the affordable care act on glaucoma severity at first presentation. *Ophthalmic Epidemiology 30*, 3 (June 2023), 326–329. Epub 2022 Jun 20.

[18] Simplemaps. Us zip codes database, 2023. Accessed: [12-20-2024].

[19] Sun, Z., Stuart, K. V., Luben, R. N., Auld, A. L., Strouthidis, N. G., Khaw, P. T., Jayaram, H., Khawaja, A. P., Foster, P. J., on behalf of the UK Biobank Eye, and Consortium, V. Association of ambient air pollution exposure with incident glaucoma: 12-year evidence from the uk biobank cohort. *Investigative Ophthalmology & Visual Science 65*, 12 (2024), 22.

[20] Tham, Y.-C., Li, X., Wong, T. Y., Quigley, H. A., Aung, T., and Cheng, C.-Y. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology 121*, 11 (2014), 2081–2090.

[21] Vajaranant, T. S., Wu, S., Torres, M., and Varma, R. The changing face of primary open-angle glaucoma in the united states: demographic and geographic changes from 2011 to 2050. *American Journal of Ophthalmology 154*, 2 (August 2012), 303–314.e3. Epub 2012 Apr 27.

[22] Wang, L., et al. Geospatial disparities in glaucoma prevalence: A gis-based approach. *Journal of Clinical Ophthalmology* (2023).

[23] Wang, Z., Wiggs, J. L., Aung, T., Khawaja, A. P., and Khor, C. C. The genetic basis for adult onset glaucoma: Recent advances and future directions. *Progress in Retinal and Eye Research 90* (Sep 2022), 101066. Epub 2022 May 17.

[24] Weinreb, R. N., Aung, T., and Medeiros, F. A. The pathophysiology and treatment of glaucoma: A review. *JAMA 311*, 18 (05 2014), 1901–1911.

[25] Wiggs, J. L., and Pasquale, L. R. Genetics of glaucoma. *Human Molecular Genetics 26*, R1 (August 2017), R21–R27.

[26] Yoo, K., Lee, C., Baxter, S. L., and Xu, B. Y. Relationship between glaucoma and chronic stress quantified by allostatic load score in the all of us research program. *American Journal of Ophthalmology 269* (2025), 419–428.