

# PREDICTING HYPERTENSION AMONGST BLACK WOMEN USING MACHINE LEARNING MODELS

Ahmed Rauf Klasra  
*Georgia Institute of Technology*  
ahmedraufk@gatech.edu

## Undergraduate Thesis

Approved by:

Dr. Andrea Parker, Advisor  
School of Interactive Computing  
*Georgia Institute of Technology*

Date Approved: 12/4/2024 | 11:40 AM EST

Signed by:  
*Andrea Parker*

David Peeler,  
Research Scientist  
Institute for People and Technology  
*Georgia Institute of Technology*

Date Approved: 12/2/2024 | 8:33 PM EST

DocuSigned by:  
*David Peeler*

## TABLE OF CONTENTS

<b>CHAPTER 1. Introduction</b>	<b>1</b>
1.1 Introduction	1
1.2 Literature Review	3
1.3 Research Questions and Study Aims	5
<b>CHAPTER 2. Methodology</b>	<b>6</b>
2.1 Data Collection	7
2.2 Data Pre-processing	9
2.3 Model Selection and Evaluation Metrics	10
2.4 Identifying Predictors of Hypertension	13
<b>CHAPTER 3. Results</b>	<b>15</b>
3.1 A Majority of EHRs Consisted of an Understudied Population of Low-Income Black Women Without a College Degree	15
3.2 GBM Was the Best ML Model to Predict Hypertension Amongst Black Women	17
3.3 Identification of Established and Novel Features Associated with Hypertension Amongst Black Women	18
<b>CHAPTER 4. Discussion</b>	<b>20</b>
<b>CHAPTER 5. Conclusion</b>	<b>23</b>

# CHAPTER 1. INTRODUCTION

## 1.1 Introduction

Black women in the United States are disproportionately affected by adverse maternal and general health outcomes related to high blood pressure and hypertension. They have the highest prevalence of hypertension among any race or ethnicity group [1], being twice as likely as white women to have the condition [2]. Moreover, They're 60% more likely to suffer from preeclampsia compared to white women [3, 4], adding to pregnancy complications as hypertension is seen to cause an increased risk of adverse cardiovascular outcomes for both the mother and the child [2]. The detrimental outcomes associated with hypertension highlight the critical importance of early detection and prediction of this condition. Prompt identification allows for the timely implementation of treatment strategies and lifestyle modifications that have the potential to prevent the occurrence of detrimental health consequences. Recent advancements in machine learning (ML) methodologies have established these models as effective alternatives to traditional statistical methods for predicting health conditions like hypertension. Their benefits include improved predictive accuracy and integration of demographic factors, enhancing the overall understanding of hypertension prevalence [5, 6]. While the accuracy of machine learning models that aid in clinical decisions is impressive, many are trained on broader populations, using un-balanced electronic health records (EHR) and patient data. This is a concern as EHRs often fail to account for nonwhite populations, leading to issues of bias amongst such models [7, 8].

To address this concern, we pre-processed black women's electronic health records (EHRs) from the All of Us research dataset to train multiple interpretable ML models. Our selected models, extreme gradient boosting(XGBoost), gradient boosting machine(GBM), and Random Forest(RF), performed with moderate accuracy (>67%) but would require more data and medical expertise to improve accuracy for clinical use. We found the gradient boosting machine to be the best model for this task and ran a SHAP analysis on it to identify which features from our dataset were the best predictors of hypertension amongst black women. Our study identified well-established features such as age, body mass index (BMI), weight, and smoking as the best predictors. Moreover, we determined through the SHAP analysis that widowed black women are at higher risk of developing hypertension compared to black women with a different marital status and that black women born outside of the United States were at lower risk of developing hypertension. Finally, we discovered that the self-reported general health scores of black women have the potential to be reliable predictors of hypertension prevalence amongst our focal population. Our research presents several noteworthy contributions to the understanding and identification of hypertension among black women. First, we demonstrate the feasibility of classic machine-learning algorithms as practical tools in clinical settings for identifying hypertension, addressing a significant gap in current literature, which often overlooks this population. Second, we employ SHAP (SHapley Additive exPlanations) analysis to enhance the interpretability of our model, ensuring that both clinicians and the target demographic can comprehend its implications. Third, our study contributes to the ongoing effort to elucidate the prevalence of hypertension among black women by identifying both established and novel predictors associated with this condition.

## 1.2 Literature Review

EHR data can be seen as a treasure trove for machine learning developers. Many studies have used EHRs to produce models that can successfully predict hypertension, such as in the case of Islam et al.[6] where a large dataset (N=8,18,603) consisting of three South Asian countries was used to train standard ML-based classifiers such as extreme gradient boosting(XGBoost), random forest(RF), and gradient boosting machine (GBM). Islam et al. employed a multifaceted approach to predict hypertension by integrating various features. Their methodology included a set of standard clinical measurements, namely age, body mass index (BMI), weight, and height. In addition to these clinical indicators, the study also considered various sociodemographic and economic factors, thereby enhancing the comprehensiveness of their predictive model. This holistic methodology shows the complex interplay between physiological and socio-economic determinants in assessing hypertension risk.

These models performed great on standard ML metrics such as F1-score, where XGBoost and GBM scored 95% overall, highlighting these models as some of the best choices for hypertension prediction. In a similar study, Shrivastava et al. [5] used a smaller dataset (N = 50,000) using similar predictive features such as age, weight, and height and additional ones such as smoking, alcohol use, and physical activity. The best-performing model in this study (RF) performed with mixed results, as it seemed to do well in its efforts to predict cases of non-hypertensive patients (F1-Score > 0.7). Unfortunately, the model performed poorly when attempting to predict cases that do have hypertension (F1-Score < 0.3). The lower amount of training data available for the model could explain this result, emphasizing the need for an extensive dataset.

Along with their impressive accuracy, ML models are effective analyzers of predictors of clinical conditions. Huang et al. [9] show that Shapely additive values (SHAP) can successfully aid in identifying predictors of hypertension. Huang et al.'s evaluation of their hypertension model using SHAP identified common predictors such as age to be strong predictors, alongside additional demographic predictors such as poverty and being black being classified as strong predictors. This shows the potential for this approach to identify not only common predictors for hypertension but perhaps even new predictors of the condition that are specific or more relevant to black women. In a study with similar goals to our own, Huang, Yongchao, et al. [8] used Machine learning to predict prenatal depression while attempting to reduce model bias towards black and Latina women. Huang, Yongchao, et al. were able to use SHAP to identify several new and unexplored markers that predicted prenatal depression. Despite the efforts to enhance the dataset by incorporating a higher representation of black and Latina women, the model continued to exhibit bias against black women. Specifically, its AUROC (area under the receiver operating characteristic curve) for prenatal depression remained inferior when compared to the predictions made for white women. The study implied that the observed results could have stemmed from poor quality data, particularly due to the possibility that black women may underreport their cases of depression because of social stigma and medical mistrust.

To mitigate the impact of poor-quality health data, it is essential to consider a range of factors.—such as socioeconomic status, access to healthcare, and social support networks—in order to better understand and mitigate the issues surrounding hypertension among these populations. These factors play a critical role in managing hypertension [10], and literature has highlighted that lifestyle elements, including diet and physical activity, also contribute to the racial disparities seen

in hypertension prevalence [11]. This underscores the necessity of incorporating both psychosocial and lifestyle variables to more accurately predict hypertension prevalence among black women.

## **1.3 Research Questions and Study Aims**

### 1.3.1 First Research Question

*How accurately can a machine learning model trained on the EHRs of black women predict hypertension among this population?*

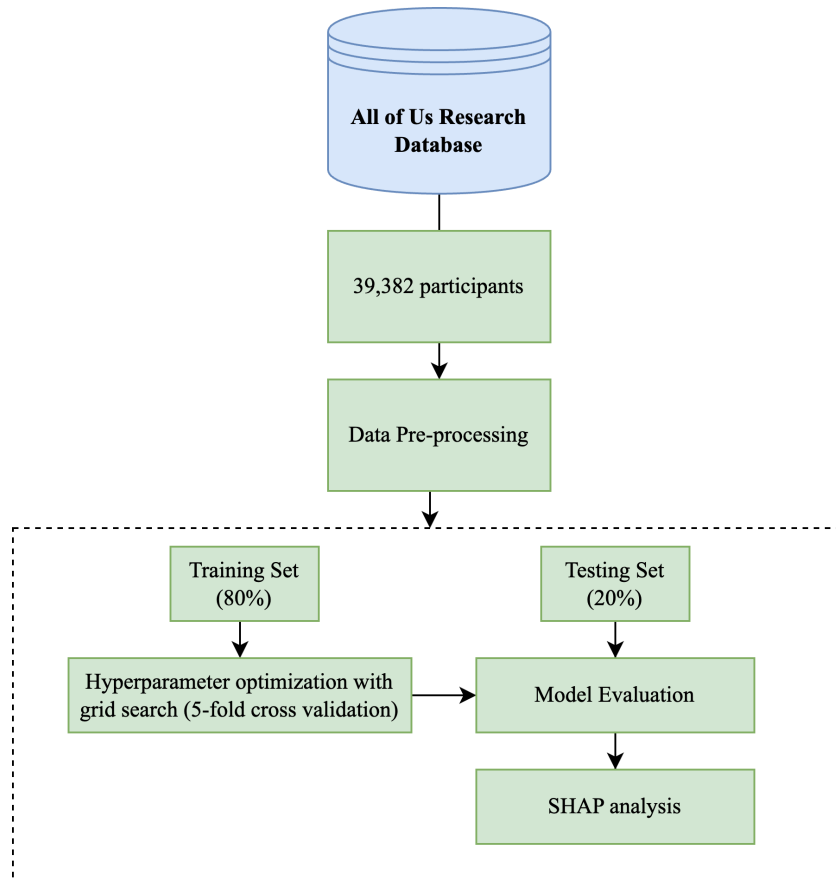
This study aims to leverage the electronic health records of Black women collected to develop and train machine learning models for predicting hypertension cases. Furthermore, we will evaluate the efficacy of these models using standard ML performance metrics to identify the most effective predictive model. The ultimate goal is to enhance understanding and prediction of hypertension within this population, contributing to tailored health interventions.

### 1.3.2 Second Research Question

*What EHR-derived factors are the most significant predictors of hypertension in Black women?*

Once we determine the best model, we aim to use SHAP to evaluate the model and decide which features are the best predictors for our model. This process could potentially uncover new predictors for hypertension in Black women, furthering our understanding of the condition amongst this demographic.

## CHAPTER 2. METHODOLOGY



**Figure 1.** Overall Workflow

## 2.1 Data Collection

This study used data from the All of Us Research Program's Registered Tier Dataset v7, available to authorized users on the Researcher Workbench[10]. The All of Us Research Program is an initiative by the National Institutes of Health that aids in the collection and study of data from more than one million people living in the United States. This initiative is designed to aid research in becoming more inclusive of populations that are historically underrepresented in medical research, including but not limited to our focal population of black women. The data utilized in this study is derived from participants belonging to underrepresented populations who voluntarily provide a comprehensive array of health-related information, including electronic health records and biospecimens. This dataset encompasses a broad spectrum of medical information, such as laboratory reports, laboratory measurements, appointment details, and results from various diagnostic tests. Additionally, it includes information regarding the medical conditions of the participants. Moreover, all the data in the program has been de-identified to maintain the confidentiality and privacy of the participants involved in the research. The participants who enroll in the program are further asked to fill out a set of core surveys related to their background, overall health, and lifestyle choices. The participants of the program may also choose to fill out additional surveys on other topics.

Through the data workbench provided to us by the program, we created a cohort of black/african-american women aged 18-91. This cohort comprised the data of 39,382 distinct participants, including clinical and physical measurements such as age, systolic blood pressure, diastolic blood pressure, heart rate, body weight, body height, and body

mass index (BMI). Survey results for each participant from the “The Basics,” “Lifestyle,” and “Overall Health” survey sets were included in the dataset to capture the socio-economic and demographic data of the participants. “The Basics” portion of the survey collects information regarding the demographic background of the participants while also including questions about the participant’s work and home life. The questions inquired about participants’ place of birth, racial and ethnic backgrounds, gender identity, sexual orientation, biological sex, level of education, marital situation, home environment, disabilities, income, and employment status. The “Lifestyle” section of the survey pertained to asking the participants about their use of tobacco, alcohol, and drugs. The “Overall Health” section of the survey asks questions regarding the participant's health and daily activity, inquiring about their general health, mental health, physical health, and quality of life, along with questions regarding the participant's social activities. This section of the survey further asks questions pertaining to women’s health by asking questions related to the participants menstrual cycles, medical procedures, and pregnancy status.

A copy of these surveys with the specific questions asked can be found on the All of Us Research Hub website in the Survey Explorer tool. The Georgia Tech Institutional Review Board waived IRB approval for the study as the All of Us dataset consists of de-identified data, and interacting All of Us data does not count as human subjects research.

## **2.2 Data Pre-processing**

The blood pressure (BP) of the participants in the dataset was measured in three successive measurements. We created the mean of the second and third measurements of the

systolic and diastolic measurements to help us identify the most accurate blood pressure reading. We took this approach as it adheres to the recommendations made by the American Heart Association BP measurement[11], as the first blood pressure measurement is usually the highest, and because other students took a similar approach to ensure more accurate blood pressure readings [6]. We amended the dataset to add a Hypertension column to indicate whether the participant had hypertension. We defined an individual as hypertensive if they had systolic BP  $\geq 140$ mmHg or diastolic BP  $\geq 90$ mmHg as defined by the Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC7) guideline [12].

Subsequent data processing involved eliminating features with fewer than 37000 non-null observations. We did this for two reasons: First, to decrease the number of features to enhance model performance and reduce the risk of overfitting while avoiding principal component analysis (PCA), which could compromise model interpretability and make it difficult to identify effective predictors of hypertension. Second, imputing data with a high proportion of null values may result in biased estimates, negatively affecting model performance[13]. Additionally, we performed a Chi-squared test to evaluate the significance of each categorical feature with the hypertension variable. Categorical features exhibiting a P-value greater than 0.05 were deemed non-significant and subsequently removed from the dataset, as they were not considered predictive of hypertension. The remaining continuous features were evaluated with the Mann-Whitney test for nonnormal values such as age. In contrast, the remaining continuous features were assessed using the Student's t-test, as they were normally distributed. Continuous features with missing values were imputed with the mean of those values and then standardized using the StandardScaler

function from sklearn. The categorical values were imputed with the most frequent ones and encoded using one hot encoding.

## 2.3 Model Selection and Evaluation Metrics

We explored multiple ML models, such as XGBoost, random forest, and GBM, as our literature review indicated that they produced the best results in this use case. The dataset was divided into training and testing subsets, with 80% of the data allocated for the training of the models and the remaining 20% reserved for assessing model performance through the application of predetermined evaluation metrics. We conducted hyperparameter optimization using GridSearchCV for each model, using a 5-fold cross-validation.

### 2.3.1 Extreme Gradient Boosting(XGBoost)

XGBoost is a tree-based supervised machine learning algorithm that uses gradient boosting to solve real-world problems. It utilizes a tree-based learning algorithm to build models sequentially and incorporates regularization techniques to prevent overfitting, making it a robust model for datasets with imbalanced classes.

The model is trained using the following objective function:

$$L(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{j=1}^m \Omega(f_j) \quad (1)$$

Where  $(L(\theta))$  is the overall loss function,  $(l(y_i, \hat{y}_i))$  is the loss function for the prediction of the  $(i)$ -th instance,  $(\hat{y}_i)$  is the predicted value, and  $(\Omega(f_j))$  is the regularization term for the  $(j)$ -th tree to prevent overfitting.

### 2.3.2 Random Forest(RF)

Random Forest is an ensemble learning method that constructs multiple decision trees during the training process and outputs the mode of their predictions for classification tasks. This algorithm is also robust against overfitting and improves prediction accuracy and variance by averaging the results of numerous trees. The prediction for a new instance is given by:

$$\hat{y} = \frac{1}{N} \sum_{j=1}^N f_j(x) \quad (2)$$

Where  $(N)$  is the number of trees,  $(f_j(x))$  is the prediction from the  $(j)$ -th tree, and  $(x)$  is the input feature vector.

### 2.3.3 Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) is another powerful ensemble technique that builds models sequentially, similar to XGBoost. However, GBM focuses on optimizing a loss function by adding new models that predict the residuals of the existing models, allowing it to capture complex patterns in data effectively. The objective function for GBM is similar to that of XGBoost and can be expressed as:

$$L(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{j=1}^m \Omega(f_j) \quad (3)$$

Where the key difference is that for each iteration  $(m)$ , the model is updated as follows:

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \eta f_m(x_i) \quad (4)$$

Where  $(\eta)$  is the learning rate, and  $(f_m(x_i))$  is the new model added at iteration  $(m)$ .

#### 2.3.4 Evaluation Metrics

We evaluated Model performance using accuracy, AUROC, and specificity. The evaluation metrics are calculated with the following equations:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$\text{AUROC} = \int_0^1 TPR(FPR^{-1}(t)) dt \quad (6)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (7)$$

TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, FN is the number of false negatives, TPR is the true positive rate, and FPR is the false positive rate.

## 2.4 Identifying Predictors of Hypertension

To interpret the significance of electronic health record (EHR) features in relation to hypertension among Black women, we employed Shapley values through the use of Shapley Additive Explanations (SHAP)[9, 14]. Each SHAP value,  $(\Phi_i)$ , represents the estimated contribution of feature  $(i)$  to the prediction of hypertension. In SHAP analysis, the log odds of the predicted probability for a particular feature $(i)$  are calculated as follows:

$$P_i = S(\Phi_0 + \Phi_i) \quad (8)$$

$$S(x) = \frac{1}{1 + e^{-x}} \quad (9)$$

Where the base SHAP value  $(\Phi_0)$  represents the logit of the baseline prevalence in the population,  $(P_0)$ , which is given as:

$$\text{logit}(P_0) = \ln \left( \frac{P_0}{1 - P_0} \right) = \Phi_0 \quad (10)$$

To obtain  $(P_0)$  itself, we apply the logistic function to  $(\Phi_0)$ :

$$P_0 = S(\Phi_0) \tag{11}$$

In our case,  $(P_0)$  was set to 0.318, indicating the prevalence of hypertension in the pre-processed EHR records.

## CHAPTER 3. RESULTS

### 3.1 A Majority of EHRs Consisted of an Understudied Population of Low-Income Black Women Without a College Degree

The pre-processed dataset comprised 39,382 unique entries of black women, with 216 EHR features. Among these, 12,528 (31.8%) were classified as hypertensive, while 26,854 (68.2%) were categorized as non-hypertensive. The majority of the electronic health records (EHRs) represented low-income black women, with 30.3% earning less than \$10,000 and 70% lacking a college degree (see Table 1). Our analysis revealed that hypertensive participants tended to be older and had a higher body weight compared to their non-hypertensive counterparts. Additionally, hypertension was more prevalent among individuals from lower-income backgrounds and those with lower educational attainment. We preprocessed the data and then used it to train ML models to predict hypertension through k-fold cross-validation. We used this approach to allow us to identify critical features that contribute to or predict hypertension amongst black women.

**Table 1.** Sociodemographic and clinical characteristics of the participants

Category P-Value	Subcategory	Overall	Hypertensive	No-Hypertension	
<b>n</b>		39382	12528	26854	
<b>Age, median [Q1,Q3]</b> <0.001		51.0 [37.0,60.0]	55.0 [46.0,62.0]	48.0 [33.0,58.0]	
<b>Weight, mean (SD)</b> <0.001		89.0 (25.0)	91.1 (25.6)	88.0 (24.7)	
<b>Height, mean (SD)</b> 0.020		163.7 (7.2)	163.6 (7.1)	163.8 (7.2)	
<b>BMI, mean (SD)</b> 0.058		34.5 (91.1)	36.2 (146.7)	33.7 (46.0)	
<b>Heart Rate, mean (SD)</b> <0.001		75.3 (12.4)	74.9 (13.0)	75.5 (12.1)	
<b>Education, n (%)</b> <0.001	Prefer Not To Answer	404 (1.0)	131 (1.0)	273 (1.0)	
	Skip	966 (2.5)	398 (3.2)	568 (2.1)	
	College graduate or advanced degree	8613 (21.9)	2329 (18.6)	6284 (23.4)	
	Highest Grade: College One to Three	11985 (30.4)	3833 (30.6)	8152 (30.4)	
	Highest Grade: Twelve Or GED	12167 (30.9)	3940 (31.4)	8227 (30.6)	
	Less than a high school degree or equivalent	5247 (13.3)	1897 (15.1)	3350 (12.5)	
	<b>Annual Income, n (%)</b> <0.001	More than 200k	283 (0.7)	44 (0.4)	239 (0.9)
	150k - 200k	370 (0.9)	78 (0.6)	292 (1.1)	
	100k - 150k	1029 (2.6)	263 (2.1)	766 (2.9)	
75k - 100k	1263 (3.2)	312 (2.5)	951 (3.5)		
50k - 75k	2550 (6.5)	691 (5.5)	1859 (6.9)		
35k - 50k	2759 (7.0)	832 (6.6)	1927 (7.2)		
25k - 35k	3148 (8.0)	929 (7.4)	2219 (8.3)		
10k - 25k	6458 (16.4)	2123 (16.9)	4335 (16.1)		
Less than 10k	11924 (30.3)	3927 (31.3)	7997 (29.8)		
Prefer Not To Answer	6501 (16.5)	2198 (17.5)	4303 (16.0)		
Skip	3097 (7.9)	1131 (9.0)	1966 (7.3)		

### 3.2 GBM Was the Best ML Model to Predict Hypertension Amongst Black Women

According to Table 2, all the algorithms performed with decent accuracy (>68%), AUROC (>64%), and specificity (>98%) scores. GBM achieved the highest score in terms of accuracy (68%) and AUROC (64%), making it the best-performing model overall. Random Forest had the best score for specificity (100%), making it the best for correctly identifying individuals who did not have hypertension. We selected GBM as the model to conduct SHAP analysis with as it scored the best on accuracy and AUROC while scoring the second-highest specificity (99.4%).

**Table 2.** Performance of selected machine learning models

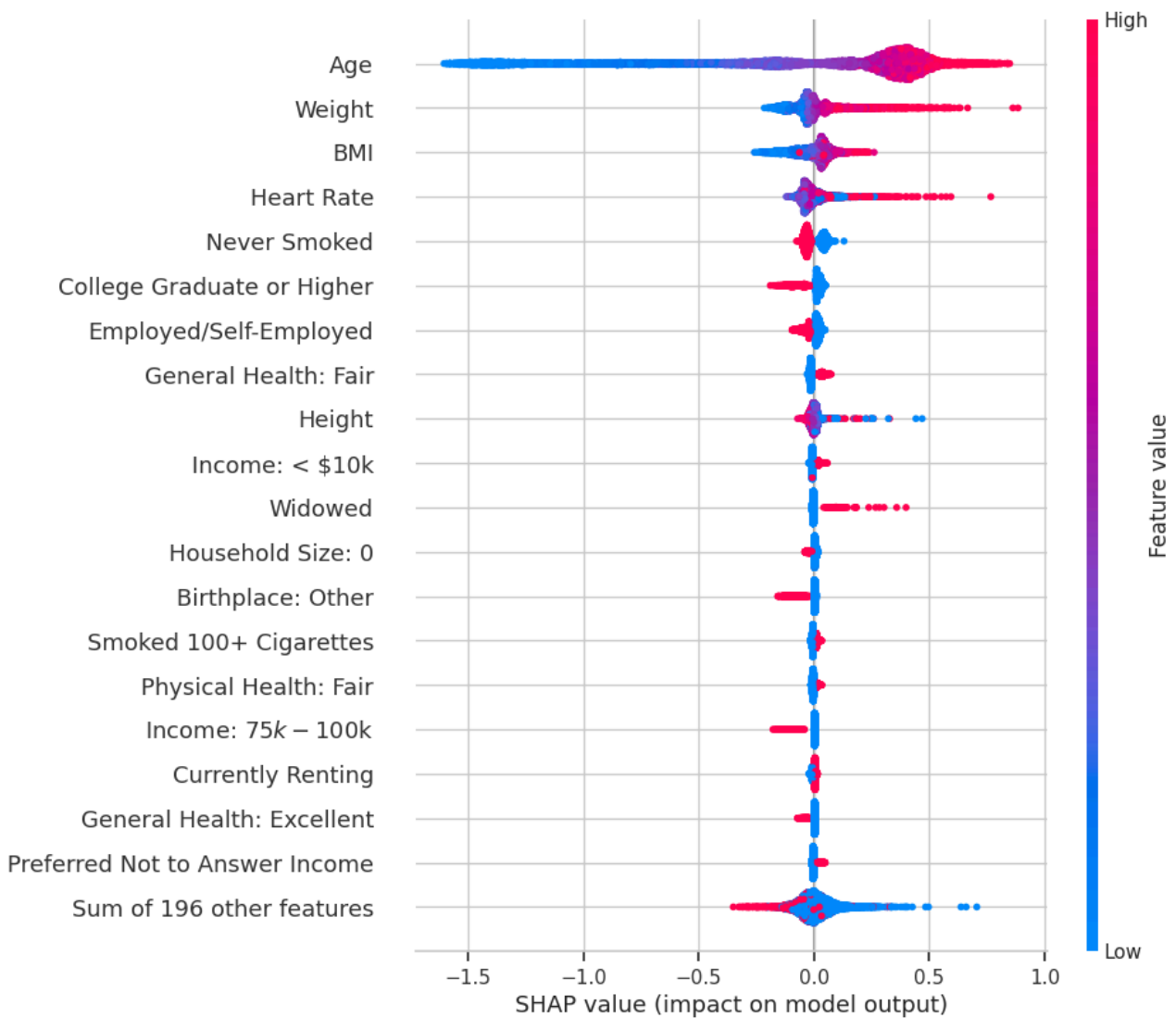
<b>Model</b>	<b>Accuracy</b>	<b>AUROC</b>	<b>Specificity</b>
GBM	0.680	0.647	0.994
Random Forest	0.679	0.640	1.000
XGBoost	0.678	0.646	0.988

### **3.3 Identification of Established and Novel Features Associated with Hypertension Amongst Black Women**

This study aimed to identify the sociodemographic and clinical characteristics that serve as significant predictors of hypertension among black women while also examining the directionality of these associations through the use of Shapley values. Shapley values estimate each feature's contribution to the model's overall predictive capability. In Figure 2, the color of the bar represents the value of the feature for numerical values such as age, weight, BMI, and heart rate, with red being a high value and blue being a low value. For categorical features, red indicates more individuals with the feature, while blue indicates the opposite. The SHAP values' direction indicates each feature's impact on the model's prediction. Positive SHAP values (to the right of 0) indicate that the feature increases the model's prediction for hypertension. In contrast, negative SHAP values (to the left of 0) indicate that the feature decreases the model's prediction for hypertension.

We discovered that the 20 features with the highest mean absolute Shapley values were broadly consistent with established sociodemographic determinants of hypertension, including advanced age [15], low income [16], lower educational attainment [17], and unemployment, alongside lifestyle factors such as smoking[18]. Moreover, the model uncovered associations with features that have received limited attention in previous research. Notably, we found that being a widowed black woman was positively correlated with an increased risk of hypertension. In contrast, we observed that the risk of hypertension had a negative association for immigrant black women, where women reported their birthplace as

“Other” were born outside of the United States. Additionally, our model identified that black women who self-reported their general health condition as “Fair” had a small yet positive association with the risk of developing hypertension. In contrast, women who self-reported their general health as “Excellent” had a negative association with the risk of developing hypertension.



**Figure 2.** Top 20 most predictive features based on their SHAP value to predict Hypertension.

## CHAPTER 4. DISCUSSION

In this study, we explored the feasibility of employing interpretable machine learning models to predict the risk of hypertension in black women. To our knowledge, this is the first study utilizing machine learning techniques to forecast hypertension and its associated factors using data that exclusively focuses on black women. We demonstrated that machine learning models trained on electronic health records (EHRs) could moderately predict hypertension rates among this demographic. Furthermore, our analysis revealed several well-known risk factors linked to hypertension in black women, as well as underexplored and novel characteristics. Notably, we identified that widowed black women are at a higher risk for hypertension, while immigrant black women appear to be at a lower risk.

The results of our evaluation determined that GBM outperformed the other two classifiers in terms of accuracy and AUROC, making it the superior algorithm for this dataset and problem. The evaluations proved that using scarce EHRs to predict hypertension amongst a specific population is a feasible idea; however, using such a model in a clinical setting would require a more significant amount of data and tuning alongside the oversight of a medical professional. The model's performance was consistent with other studies that used hypertension, such as Shrivastava et al. [5], as our high specificity scores suggest that the model is better at predicting non-hypertensive cases than it is at predicting hypertensive cases.

Our study found that being widowed increases the risk of hypertension for black women. This result is consistent with other studies, such as in the case of Schwandt et al. [19], where the study demonstrated that widowed or single black women were at higher risk of developing hypertension

compared to women who were still married. Furthermore, our study revealed that black immigrant women are at a lower risk of developing hypertension, indicating that the “Healthy Immigrant Effect” [20] — the concept suggesting that immigrants arrive in their new country in better health than those born domestically — applies to black women as well. This observation is supported by a study performed by Brown et al. [21], where the authors found that foreign-born, non-Hispanic black people had a lower prevalence of hypertension compared to black people who were born in the United States.

Additionally, our model found that black women who reported their general health as “Excellent” were less likely to be identified as hypertensive. In contrast, those who identified their health as “Fair” were more likely to be recognized as hypertensive. The general health portion of the All of Us research survey allows participants to self-report the state of their general health in terms of “Excellent, Very Good, Good, Fair, Poor.” This result would suggest that self-reported general health scores of black women can be seen as a somewhat reliable predictor of hypertension for them but would require further research into the matter.

Due to limitations with the dataset, our model’s predictive accuracy was left lacking. To address the issue with accuracy while also attempting to validate our findings, we recommend future work to train hypertension prediction models using other classical machine learning algorithms with more extensive data on black women’s EHRs. We believe it would be beneficial to include data from different populations, such as white women, alongside the dataset of black women. This would allow us to gain better insights into bias regarding model accuracy, where ML models might favor one class, such as white women, over the other, such as black women. In addition, this could allow us to understand better the difference between the quality of EHRs of white and black women and

the impact the dataset's quality has on the ML models. We further recommend conducting a comprehensive study to assess the reliability of self-reported health scores as predictors of hypertension among Black women. This investigation would involve cross-referencing self-reported data with clinical measurements and biomarker analyses. Such an approach would provide valuable insights into the validity of self-reported health assessments as a relevant factor in hypertension prediction within this demographic.

## CHAPTER 5. CONCLUSION

Our study demonstrates that interpretable machine learning models trained on electronic health records (EHRs) can predict hypertension in Black women with moderate accuracy. We identified age, weight, and body mass index (BMI) as the most significant risk factors associated with hypertension. Additionally, our findings indicate that widowed Black women are at a greater risk of developing hypertension, while immigrant Black women appear to be at a reduced risk. We also discovered that self-reported general health scores may serve as potential predictors of hypertension among Black women, although further research is necessary to validate this. Moreover, enhancing the prediction performance of EHR-based machine learning models for hypertension prediction in Black women would require additional data and increased medical oversight.

## **Acknowledgments**

We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data [and/or samples and/or cohort] examined in this study.

## Reference list

- [1] Rahul Aggarwal, Nicholas Chiu, Rishi K. Wadhwa, Andrew E. Moran, Inbar Raber, Changyu Shen, Robert W. Yeh, and Dhruv S. Kazi. 2021. Racial/Ethnic Disparities in Hypertension Prevalence, Awareness, Treatment, and Control in the United States, 2013 to 2018. *Hypertension* 78, 6, <https://doi.org/10.1161/HYPERTENSIONAHA.121.17570>
- [2] A. L. Hines, H. Zare, and R. J. Thorpe, Jr. 2022. *Racial Disparities in Hypertension Among Young, Black and White Women* J Gen Intern Med, Vol. 37. United States. <https://doi.org/10.1007/s11606-021-07073-0>
- [3] K. R. Finger, I. Mabry-Hernandez, Q. Ngo-Metzger, T. Wolff, C. A. Steiner, and A. Elixhauser. 2006. *Delivery Hospitalizations Involving Preeclampsia and Eclampsia, 2005–2014* Healthcare Cost and Utilization Project (HCUP) Statistical Briefs, Agency for Healthcare Research and Quality (US), Rockville (MD).
- [4] Natalie Hernandez-Green, V Morgan Davis, Oluyemi Farinu, Kaitlyn Hernandez-Spalding, Kennedy Lewis, S Merna Beshara, Sherilyn Francis, Joy Lethenia Baker, Sherrell Byrd, Andrea Parker, and Rasheeta Chandler. 2024. Using mHealth to reduce disparities in Black maternal health: Perspectives from Black rural postpartum mothers. *Women's Health* 20 <https://doi.org/10.1177/17455057241239769>
- [5] Anurag Shrivastava, Midhun Chakkaravarthy, and Mohd Asif Shah. 2023. A new machine learning method for predicting systolic and diastolic blood pressure using clinical characteristics. *Healthcare Analytics* 4 <https://doi.org/10.1016/j.health.2023.100219>

- [6] S. M. S. Islam, A. Talukder, M. A. Awal, M. M. U. Siddiqui, M. M. Ahamad, B. Ahammed, L. B. Rawal, R. Alizadehsani, J. Abawajy, L. Laranjo, C. K. Chow, and R. Maddison. 2022. Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries. *Front Cardiovasc Med* 9(20220331), <https://doi.org/10.3389/fcvm.2022.839379>
- [7] D. S. Char, N. H. Shah, and D. Magnus. 2018. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *N Engl J Med* 378, 11, <https://doi.org/10.1056/NEJMp1714229>
- [8] Yongchao Huang, Suzanne Alvernaz, Sage J. Kim, Pauline Maki, Yang Dai, and Beatriz Peñalver Bernabé. 2024. Predicting Prenatal Depression and Assessing Model Bias Using Machine Learning Models. *Biological Psychiatry Global Open Science* 4, 6, <https://doi.org/10.1016/j.bpsgos.2024.100376>
- [9] A. A. Huang, and S. Y. Huang. 2023. Shapely additive values can effectively visualize pertinent covariates in machine learning when predicting hypertension. *J Clin Hypertens (Greenwich)* 25, 12, (20231116), <https://doi.org/10.1111/jch.14745>
- [10] All of Us Research Workbench. Registered Tier from [www.workbench.researchallofus.org](http://www.workbench.researchallofus.org)
- [11] Paul Muntner, Daichi Shimbo, Robert M. Carey, Jeanne B. Charleston, Trudy Gaillard, Sanjay Misra, Martin G. Myers, Gbenga Ogedegbe, Joseph E. Schwartz, Raymond R. Townsend, Elaine M. Urbina, Anthony J. Viera, William B. White, Jackson T. Wright, Hypertension on behalf of the American Heart Association Council on, Young Council on Cardiovascular Disease in the, Nursing Council on Cardiovascular and Stroke, Intervention Council on Cardiovascular Radiology and, Cardiology Council on Clinical, and Research and Council on Quality of Care and Outcomes. 2019.

Measurement of Blood Pressure in Humans: A Scientific Statement From the American Heart Association. *Hypertension* 73, 5, [https://doi.org/ 10.1161/HYP.0000000000000087](https://doi.org/10.1161/HYP.0000000000000087)

[12] A. V. Chobanian, G. L. Bakris, H. R. Black, W. C.ushman, L. A. Green, J. L. Izzo, Jr., D. W. Jones, B. J. Materson, S. Oparil, J. T. Wright, Jr., and E. J. Roccella. 2003. The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: the JNC 7 report. *Jama* 289, 19, (20030514), [https://doi.org/ 10.1001/jama.289.19.2560](https://doi.org/10.1001/jama.289.19.2560)

[13] J. H. Lee, and J. C. Huber, Jr. 2021. Evaluation of Multiple Imputation with Large Proportions of Missing Data: How Much Is Too Much? *Iran J Public Health* 50, 7, [https://doi.org/ 10.18502/ijph.v50i7.6626](https://doi.org/10.18502/ijph.v50i7.6626)

[14] Scott M. Lundberg, and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR* abs/1705.07874

[15] T. W. Buford. 2016. Hypertension and aging. *Ageing Res Rev* 26(20160201), [https://doi.org/ 10.1016/j.arr.2016.01.007](https://doi.org/10.1016/j.arr.2016.01.007)

[16] D. Edmund Anstey, Jessica Christian, and Daichi Shimbo. 2019. Income Inequality and Hypertension Control. *Journal of the American Heart Association* 8, 15, [https://doi.org/ 10.1161/JAHA.119.013636](https://doi.org/10.1161/JAHA.119.013636)

[17] M. Zacher. 2023. Educational Disparities in Hypertension Prevalence and Blood Pressure Percentiles in the Health and Retirement Study. *J Gerontol B Psychol Sci Soc Sci* 78, 9, [https://doi.org/ 10.1093/geronb/gbad084](https://doi.org/10.1093/geronb/gbad084)

- [18] A. Viridis, C. Giannarelli, M. F. Neves, S. Taddei, and L. Ghiadoni. 2010. Cigarette smoking and hypertension. *Curr Pharm Des* 16, 23, [https://doi.org/ 10.2174/138161210792062920](https://doi.org/10.2174/138161210792062920)
- [19] Hilary M. Schwandt, Josef Coresh, and Michelle J. Hindin. 2010. Marital Status, Hypertension, Coronary Heart Disease, Diabetes, and Death Among African American Women and Men: Incidence and Prevalence in the Atherosclerosis Risk in Communities (ARIC) Study Participants. *Journal of Family Issues* 31, 9, [https://doi.org/ 10.1177/0192513X10365487](https://doi.org/10.1177/0192513X10365487)
- [20] Andreea C. Brabete. 2017. *Chapter 8 - Examining Migrants' Health From a Gender Perspective* The Psychology of Gender and Health, Academic Press, San Diego. <https://doi.org/https://doi.org/10.1016/B978-0-12-803864-2.00008-0>
- [21] A. G. M. Brown, R. F. Houser, J. Mattei, D. Mozaffarian, A. H. Lichtenstein, and S. C. Folta. 2017. Hypertension among US-born and foreign-born non-Hispanic Blacks: National Health and Nutrition Examination Survey 2003-2014 data. *J Hypertens* 35, 12, [https://doi.org/ 10.1097/hjh.0000000000001489](https://doi.org/10.1097/hjh.0000000000001489)