

Predicting cardiovascular disease using different blood pressure guidelines

Christopher M. Bopp¹, William Briggs², Catherine Orlando¹, Raed Seetan¹

¹Department of Computer Science, Slippery Rock University, Slippery Rock, United States

²Department of Mathematics and Statistics, Slippery Rock University, Slippery Rock, United States

Article Info

Article history:

Received Jun 21, 2022

Revised Feb 20, 2023

Accepted Mar 9, 2023

Keywords:

Cardiovascular disease

Heart disease

Supervised learning

ABSTRACT

The criteria used to categorize patients as either hypertensive or normotensive were changed in 2017 by the American Heart Association and the American College of Cardiology (AHA/ACC). The updated guidelines lowered the criteria by which individuals are classified as hypertensive; systolic blood pressure (SBP) cut-off from ≥ 140 mmHg to ≥ 130 mmHg and diastolic blood pressure from ≥ 90 mmHg to ≥ 80 mmHg. The purpose of this study was to investigate what effect these changes in diagnostic criteria had on the ability of supervised learning to predict cardiovascular disease. Three models were developed and tested. Two models using a binned hypertension measure based on either the AHA/ACC new released guidelines or the Joint National Committee on the Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC7) original guidelines. The third model used SBP as a continuous variable. Data from 68,657 patients was processed through decision tree algorithm to determine which model offered the best accuracy. For both female and male subjects, the model with SBP returned the best area under the receiver operator characteristic curve and overall better sensitivity and specificity values. Our results showed that changing the criteria by which individuals are classified as hypertensive or normotensive negatively impacted the ability of decision tree to predict cardiovascular disease in both females and males.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Christopher M. Bopp

Department of Computer Science, Slippery Rock University

Slippery Rock, PA 16057, United States

Email: cmb1055@sru.edu

1. INTRODUCTION

Cardiovascular disease (CVD) is the number one cause of death in the developed world [1]. The clinical end points of cardiovascular disease are myocardial infarction or stroke and often individuals do not know that they are at risk until such an event occurs [2]. Early efforts to predict cardiovascular disease used logistic regression analysis and often sought to predict cardiovascular death or myocardial infarction based upon the presence of cardiovascular disease risk factors. Common risk factors include a history of smoking, high serum cholesterol, high blood pressure, age and male sex [3]–[5]. In 1998, the Framingham Heart Study found that a model using total cholesterol, high-density lipoprotein cholesterol, age, sex, systolic blood pressure (SBP), diabetes mellitus, and smoking status was able to separate those who experience cardiovascular disease events from those who do not [6]. Despite these efforts, half of myocardial infarction and strokes occur in people who are not predicted to be at increased risk for cardiovascular disease [7]–[9]. For this reason, researchers have begun to apply machine learning algorithms to the task of cardiovascular disease prediction.

In 2017, Weng *et al.* [10] compared an established algorithm from the American College of Cardiology against four machine learning algorithms. They found that all four algorithms outperformed the established algorithm; the use of these techniques increased predictive accuracy as assessed by area under the receiver operating characteristic curve (AUC) by 1.7%-7.6%. These four algorithms also produced higher sensitivity and specificity values compared to traditional modeling techniques. Alaa *et al.* [11] combined 5 common machine learning algorithms into a model that outperformed traditional prediction equations, including the Framingham Risk Score and Cox proportional hazard models, by up to 5%. Martins *et al.* [12] applied data mining techniques for cardiovascular disease prediction and found that the best technique was decision tree.

Of the laboratory-based cardiovascular disease risk factors, blood pressure is the easiest to measure. Blood pressure also has the advantage of not requiring a blood draw or finger stick for assessment. Of the modifiable risk factors, hypertension also contributes to more cardiovascular disease deaths annually [13]. Between 2003 and 2017, blood pressure was categorized using recommendations from the 7th Report of the Joint National Committee on the Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC7) [14]. Under these guidelines, individuals with SBP ≥ 140 mmHg or diastolic blood pressure (DBP) ≥ 90 mmHg were categorized as hypertensive. Patients were categorized as prehypertensive if their SBP was 120-139 mmHg, or their DBP was 80-89. Individuals with SBP < 120 and DBP < 80 mmHg were categorized as normal.

In 2017, the American Heart Association and the American College of Cardiology (AHA/ACC) released updated guidance [15]. Individuals are now categorized as hypertensive with SBP ≥ 130 mmHg or DBP ≥ 80 mmHg. Individuals with SBP between 120 and 129 mmHg or DBP between 80 and 89 mmHg are now categorized with elevated blood pressure. Normal blood pressure remained the same. These new guidelines have not been universally accepted; the European Society of Hypertension's 2021 practice guidelines recommend using the older values to diagnose hypertension [16]. The purpose of this study was to determine if categorizing patients as hypertensive using the JNC7 guidelines or the AHA/ACC guidelines provides an advantage over the use of SBP alone in predicting cardiovascular disease. This study has a secondary aim to establish clinical cut points, the criteria used to diagnose hypertension, using receiver operator characteristic (ROC) curves.

2. RESEARCH METHOD

2.1. Data source

The main goal of this study was to determine how the accuracy of the decision tree algorithm to predict cardiovascular disease is affected by the inclusion hypertension status as a categorical variable (high or normal) compared to the inclusion of SBP as a continuous variable. To investigate that, the CardioTrain data set was retrieved from Kaggle on September 23, 2021 [17]. The original data set includes data from 70,000 patients from the European Union and includes the following continuous attributes: patient ID, age (days), height, weight, gender, SBP, DBP. Age in years and body mass index (BMI) were calculated and included in the continuous attributes. The data set also includes categorical attributes. Cholesterol and glucose were categorized as normal (1), above normal (2) and well above normal (3). Smoking, alcohol use, physical activity and cardiovascular disease were also included as binary variables; 0 indicated that this variable was false; 1 indicated that this factor was present. Individuals were characterized as obese if their BMI was ≥ 30 kg/m². Hypertension categories were created based on the AHA/ACC or JNC7 guidelines [14], [15].

2.2. Data processing

The original CardioTrain data set contained 70,000 records. After cleaning the data, 68,657 records remained. BMI was used as an initial measure to remove data. Using Tukey's hinge method, outliers (values below 14 kg/m²) were identified and removed. The highest BMI ever recorded was 105 kg/m². Values above this were removed (n=29). Blood pressure readings and mean arterial pressure served as another method to remove noisy data. Blood pressure is recorded as SBP over DBP; SBP must be greater than DBP. Any instance where SBP was less than or equal to DBP was removed (n=1,248) as were any instances where either blood pressure value was 0 or negative (n=3). Records where SBP were not physiologically possible (n=50) were also removed.

2.3. Statistical analysis

Binary logistic regression: data were entered into SPSS 28 (IBM, Armonk, New York, USA) for analysis. Binary logistic regression is a type of regression that finds relationships between the dependent variables and a dichotomous independent variable. Therefore, the independent variable is either 0 or 1 and the logarithm of the dependent variables are used as predictors. We use binary logistic regression to see which of the risk factors are significant in the model and therefore should not be omitted from future models.

Separate analyses were performed for each gender. Three binomial logistic regressions were performed with the presence of cardiovascular disease as the dependent variable and age, gender, cholesterol, glucose, obesity, and smoking entered in the model. A variable related to SBP was also included in each analysis; hypertension status per AHA/ACC guidelines was included in the first analysis, hypertension status per the JNC7 guidelines in the second and continuous SBP values in the third. Each analysis calculated odds ratios and 95% confidence intervals for the respective CVD risk factor.

ROC analysis: each instance of binary logistic regression generated predicted probability scores that were used in ROC analysis to determine AUC for each model. ROC analysis with cardiovascular disease as the state variable and systolic and DBP as the test variable was used to calculate Youden's index as in (1). Youden's Index will yield the associated blood pressure value cutoff that generated the best sensitivity and specificity values for each dataset. A fourth model (Youden) using these clinical cut points to define hypertension was developed and compared to existing models.

$$\text{Youden's Index} = \text{Sensitivity} + \text{Specificity} - 1 \quad (1)$$

Classification: decision tree algorithm was used to generate a contingency table for each model. In decision tree recursive partitioning is used to split the training set into segments by minimizing impurity. A node is considered pure if 100% of cases in the node fall into a specific category [18]. Each model underwent a test/train validation (50%/50%) and a cross validation with 10 sample folds. The values in the contingency table were used to calculate sensitivity, specificity, positive likelihood ratio, negative likelihood ratio, positive predictive value (PPV), negative predictive value (NPV), and accuracy were calculated per Sheffler [19].

Sensitivity is the proportion of true positives out of all individuals with CVD.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (2)$$

Specificity is the percentage of true negatives out of all patients that do not have CVD

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negatives} + \text{False Positives}} \quad (3)$$

Positive Predictive Value determines what percentage of positive tests are truly positive

$$\text{PPV} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (4)$$

Negative Predictive Value determines what percentage of negatives tests are truly negative

$$\text{NPV} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Negatives}} \quad (5)$$

Positive likelihood ratio equals the probability that a positive test would be expected in a patient with the disease (true positive) divided by the probability that a positive test would be expected in a patient without the disease (false positive)

$$\text{Positive Likelihood Ratio} = \frac{\text{Sensitivity}}{1 - \text{Specificity}} \quad (6)$$

Negative Likelihood Ratio is the probability of a patient with CVD testing negative (false negative) divided by the probability of a patient without CVD testing negative (true negative).

$$\text{Negative Likelihood Ratio} = \frac{1 - \text{Sensitivity}}{\text{Specificity}} \quad (7)$$

2.4. Supervised learning

The dataset was imported into SPSS Modeler Version 18.3 (IBM, Armonk, NY, USA). Each data set was exposed to 13 classification algorithms. The software returned the AUC and model accuracy for the top 4 models.

3. RESULTS

Table 1 shows subject characteristics for the CardioTrain dataset split by sex. In female patients, smoking was not a significant contributor to any model. Smoking remained in the analysis for consistency and because removal of smoking from the model did not alter any results. Fasting plasma glucose was entered as a categorical variable with three levels (normal, high and very high). Although high glucose was only significant in the AHA/ACC based model for both female and male subjects, very high glucose was a significant contributor, so this variable was kept in all models. The use of the AHA/ACC guidelines increased rates of hypertension in both females and males. Tables 2 and 3 include odds ratios and 95% confidence intervals for each risk factor in the three models of interest in females and males, respectively. Each blood pressure variable was a significant contributor to their respective models. In female subjects, individuals with hypertension were 3.1 AHA/ACC or 5.5 JNC7 times more likely to have cardiovascular disease compared to those without hypertension. For each mmHg increase in SBP, the odds of cardiovascular disease increased by 6.3%. In males, individuals with hypertension were 2.6 AHA/ACC or 5.6 JNC7

times more likely to have cardiovascular disease compared to those without hypertension. For each mmHg increase in SBP, the odds of cardiovascular disease increased by 6.5%. Although smoking is traditionally a risk factor, males who smoke had lower odds of being diagnosed with cardiovascular disease. The same was seen in females and males with glucose in the “very high” category more discussion in the discussion section.

Table 1. Subject characteristics of the CardioTrain dataset

Variable	Female (n=44,721)	Male (n=23,936)
Categorical variables – n (%)		
Cardiovascular disease prevalence	22,008 (49.2%)	11,959 (50.0%)
Obese	13,266 (29.7%)	4,692 (19.6%)
Normal	33,029 (73.9%)	18,495 (77.3%)
Cholesterol	6,256 (14%)	3,045 (12.7%)
High	5,436 (12.1%)	2,432 (10.2%)
Very High	3,7802 (84.5%)	20,574 (86.0%)
Fasting Plasma Glucose	3,347 (7.5%)	1,719 (7.2%)
Normal	3,572 (8.0%)	1,643 (6.9%)
High	793 (2.0%)	5,244 (21.9%)
Very High	35,586 (79.6%)	20,397 (85.2%)
Current smoker	14,777 (33%)	8,762 (36.6%)
Hypertension	Continuous variables - Mean (SD)	
Age (Years)	53.4 (6.7)	53.1(6.9)
Body mass index (kg/m ²)	27.9(5.6)	26.7(4.5)
SBP (mmHg)	125.9 (16.7)	128.1(16.5)
DBP (mmHg)	80.8 (9.6)	82.2(9.3)

Table 2. Odds ratios and 95% confidence intervals for all risk factors included in each model for the CardioTrain dataset in females. * Indicates significant at p<0.05

Variable	AHA/ACC			JNC7			SBP			
	OR	95% CI		OR	95% CI		OR	95% CI		
		Lower	Upper		Lower	Upper		Upper	Lower	
Age	1.066*	1.063	1.07	1.068*	1.064	1.072	1.057*	1.053	1.060	
Obese	1.577*	1.508	1.649	1.307*	1.246	1.371	1.228*	1.170	1.289	
Cholesterol	High	1.666*	1.567	1.770	1.366*	1.281	1.456	1.431*	1.340	1.528
	Very high	3.570*	3.292	3.873	3.095*	2.841	3.372	3.019*	2.771	3.290
Glucose	High	1.088*	1.003	1.179	1.037	0.952	1.130	1.017	0.931	1.110
	Very high	.728*	.665	.797	0.768*	0.699	0.845	0.740*	0.673	0.814
Current smoker	0.911	.781	1.063	.854	0.726	1.005	0.881	0.747	1.038	
Hypertension	AHA/ACC	3.104*	2.938	3.279						
	JNC7				5.554*	5.294	5.827			
	SBP							1.063*	1.061	1.065

ROC analysis plots sensitivity against 1-specificity, the model that approaches the closest to the upper left-hand corner (closest to sensitivity and specificity values of 1) offers the best predictive performance. As is seen in Figure 1, the model based on the AHA/ACC guidelines performed worst. Based on visual inspection, the model based on the JNC7 guidelines performed similarly to the model based on SBP, although AUC values reported in Table 4 does show that SBP outperformed the JNC7 guidelines. The model with SBP as a continuous variable performed better than the models using hypertension status binned by either the AHA/ACC or JNC7 criteria for both females and males. In females, the AUC for SBP

was 0.793 (Std. Error=0.002, 95% CI=0.788-0.797) compared to an AUC of 0.722 (Std Error=0.002, 95% CI=0.717-0.726) for the AHA/ACC model and an AUC of 0.778 (Std. Error=0.002, 95% CI=0.773-0.782) for the JNC7 model. In males, the AUC for SBP was 0.693 (Std. Error=0.002, 95% CI=0.690-0.704) compared to an AUC of 0.772 (Std Error=0.002, 95% CI=0.766-0.778) for the AHA/ACC model and an AUC of 0.785 (Std. Error=0.002, 95% CI=0.779-0.791) for the JNC7 model.

Table 3. Odds ratios and 95% confidence intervals for all risk factors included in each model for the CardioTrain dataset in males. * indicates significant at $p < 0.05$

Variable	AHA/ACC			JNC7			SBP			
	OR	95% CI		OR	95% CI		OR	95% CI		
		Lower	Upper		Lower	Upper		Upper	Lower	
Age	1.055*	1.050	1.059	1.050*	1.046	1.055	1.045*	1.040	1.049	
Obese	1.823*	1.699	1.955	1.474*	1.367	1.589	1.402*	1.299	1.512	
Cholesterol	High	1.938*	1.779	2.110	1.526*	1.392	1.672	1.529*	1.394	1.678
	Very high	3.484*	3.111	3.903	3.012*	2.671	3.396	2.955*	2.619	3.335
Glucose	High	1.155*	1.034	1.291	1.080	0.960	1.216	1.079	0.957	1.217
	Very high	0.650*	0.573	0.737	0.680*	0.596	0.777	0.651*	0.596	0.744
Current smoker	0.857*	.803	.915	0.787*	0.734	0.845	0.881	0.747	1.038	
Hypertension	ACC/AHA	2.666*	2.455	2.895						
	JNC7				5.614*	5.273	5.977			
	SBP							1.065*	1.062	1.067

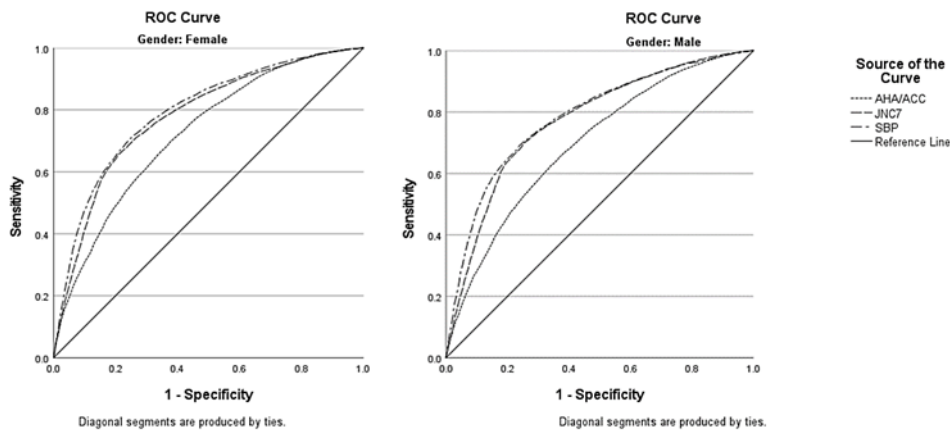


Figure 1. ROC curves for each model using the CardioTrain dataset, separated by sex

Table 4. AUC for each model, separated by sex

	Female		Male	
	AUC	95% CI	AUC	95% CI
AHA/ACC	0.722	0.717-0.726	0.697	0.690-0.704
JNC7	0.778	0.773-0.782	0.772	0.766-0.778
SBP	0.793	0.788-0.797	0.785	0.779-0.791
Youden (not shown)	0.775	0.771-0.780	0.765	0.759-0.771

Individual ROC analyses performed with systolic and DBP as the test variable and cardiovascular disease as the state variable generated coordinate points for the ROC curve which allowed for the determination of Youden's Index and the associated systolic and DBP levels in females and males. For SBP, Youden's index in females and males was 0.425 and 0.423, respectively. For both, 129.5 mmHg was the SBP value associated with Youden's Index. For DBP, Youden's index was 0.307 and 0.327 for females and males, respectively. For both genders, the associated DBP value associated with Youden's index was 82.5 mmHg and 84.5 mmHg for females and males respectively. Using these newly defined cut points, a 4th model (Youden's) was created to define hypertension (≥ 130 mmHg SBP; ≥ 83 mmHg DBP in females and ≥ 85 mmHg DBP in males) did not improve AUC over that associated with either the JNC7 model or the SBP model see Table 4.

Classification: The model based on SBP performed better in model accuracy for both genders according to AUC and classification analyses compared to either of the models where hypertension status was categorized by AHA/ACC or JNC7. Between the two binned hypertension variables, the JNC7 guidelines performed better than the AHA/ACC guidelines in this dataset for both genders. Table 5 displays contingency table data from the cross-validation phase of the classification analysis including true positives, true negatives, false positives and false negatives.

Table 5. Contingency tables for each model by sex

		True positive	True negative	False positive	False negative
AHA /ACC	Female	15,088	14,118	8,595	6,920
	Male	7,088	7,995	3,992	4,871
JNC7	Female	14,505	17,932	4,781	7,503
	Male	8,113	9,133	2,844	3,846
SBP	Female	14,978	17,723	4,990	7,030
	Male	8,326	9,049	2,928	3,633
Youden's	Female	15,765	16,709	6,004	6,243
	Male	8,769	8333	3,644	3,190

Table 6 shows classification measures derived from contingency tables generated from these values, including sensitivity, specificity, PPV, NPV, positive likelihood ratio, negative likelihood ratio and accuracy. Sensitivity is the proportion of true positive tests out of all the individuals with a positive test. Positive predictive value (also known as precision) is the probability that a subject with a positive screening test truly has CVD. Negative predictive value is the probability that subjects with a negative test truly do not have CVD.

In females, the use of the clinical cut points established by Youden's Index offered increased model accuracy (72.6%) compared with AHA/ACC (65.3%) and JNC7 (72.5) models but did not beat SBP (73.1%). In males, the accuracy of this new model (71.4%) was greater than that established by AHA/ACC (63.3%) but did not exceed that of the JNC7 (72.1%) or SBP (72.6%) models. In females, this model increased sensitivity at the expense of specificity; in males the opposite occurred, sensitivity was lower than that of the SBP model, but specificity was higher.

Table 6. Classification analysis showing contingency table derived measures

	AHA/ACC		JNC7		SBP		Youden	
	Female	Male	Female	Male	Female	Male	Female	Male
Accuracy	65.3%	63.3%	73.3%	72.2%	73.3%	73.3%	73.3%	71.1%
Sensitivity	0.69%	0.59%	0.66%	0.68%	0.68%	0.70%	0.72%	0.73%
Specificity	0.62%	0.667%	0.790%	0.76%	0.78%	0.76%	0.74%	0.70%
Positive likelihood ratio	1.8	1.8	3.1	2.85	3.1	2.85	2.7	2.4
Negative likelihood ratio	0.5	0.61	0.43	0.42	0.4	0.4	0.38	0.38
Positive predictive value	0.64%	0.64%	0.75%	0.74%	0.75%	0.74%	0.72%	0.70%
Negative predictive value	0.67%	0.62%	0.71%	0.70%	0.71%	0.71%	0.72%	0.72%

Table 7 shows the results from the two best supervised learning algorithms in female and male subjects for each model. In female subjects, and using the AHA/ACC model, the neural net algorithm matched the AUC found from ROC analyses (0.722) and performed slightly better in measures of accuracy (66.0% to 65.3%). For the JNC7 model, neural net was again superior to the other models tested; the resulting AUC value was 0.777 and the algorithm demonstrated 72.4% accuracy. Both values were slightly lower than those produced by decision tree. For the SBP model, neural net again returned the best AUC (0.797) and accuracy (73%); supervised learning algorithms met the accuracy and exceeded the AUC of the decision tree classification analysis. In male subjects, logistic regression and linear support vector machine (LSVM) were the best performing algorithms in both the AHA/ACC model and the JNC7 model. The AUC values for each model matched the AUC of the decision tree algorithm (AHA/ACC AUC=0.697; JNC7 AUC=0.772). The accuracy of these algorithms was slightly higher in the AHA/ACC model (64.1%) versus 63.0% in the classification analysis, however, the accuracy of these algorithms in the JNC7 model (71.9%) was slightly lower than the 72.0% accuracy returned in the classification analysis. In the SBP model, the algorithms returning the best AUC and accuracy values were logistic regression and neural net. As with the JNC7 findings, supervised learning provided a slightly higher AUC (0.787 vs. 0.785) and a slightly lower accuracy (72.3% vs. 73%).

Table 7. AUC and accuracy determined for each model using supervised learning algorithms

Model	AUC	Accuracy	Algorithm
Female			
AHA/ACC	0.722	66.0%	Neural net
	0.722	65.8%	LSVM
JNC7	0.777	72.4%	Neural net
	0.775	72.2%	Tree-AS
SBP	0.797	73.0%	Neural net
	0.695	64.2%	LSVM
Male			
AHA/ACC	0.697	64.1%	Logistic regression
	0.697	64.1%	LSVM
JNC7	0.772	71.9%	Logistic regression
	0.772	71.9%	LSVM
SBP	0.787	72.3%	Logistic regression
	0.784	72.0%	Neural net

4. DISCUSSION

Our results show that changing the criteria by which individuals are classified as hypertensive or normotensive negatively impacted the ability of decision tree and other algorithms to predict cardiovascular disease in both females and males. Both models using binning by hypertension status performed worse than the model in which SBP was entered as a continuous variable. Although it is convenient to create arbitrary cut points, and to classify individuals as hypertensive based on these values, this analysis shows that this technique may decrease the prognostic ability of the model. Altman and Royston recommend that dichotomizing continuous variables be avoided [20]. Consider the case of two individuals with SBP values of 192 mmHg and 131 mmHg, both would be considered hypertensive by the AHA/ACC model. Using the SBP model, where each mmHg above the cut point increased risk by 6.3% would clearly show that the individual at 192 mmHg demonstrates 4 times greater odds of developing cardiovascular disease ($62 \times 0.065 = 4.03$) compared to the individual at 131 mmHg who only has a slight increase (6.3%) in the odds of developing cardiovascular disease.

To further examine the effects of binning versus using continuous variables in a model a second dataset was evaluated with supervised learning. The Framingham dataset was downloaded from Kaggle on February 20, 2022 [21], and it has been used in several recent studies [22]–[24]. The Framingham heart study is a longitudinal cohort study of cardiovascular disease risk, which includes continuous variables for age, cholesterol, SBP, BMI and fasting plasma glucose [25]. These variables were dichotomized using standard clinical cut points. Gender and smoking status are also included as categorical variables. Two models were produced, one with continuous variables and one where the continuous variables were dichotomized. Each model was run through 5 supervised algorithms, discriminant analysis, chi-square automatic interaction detection (CHAID), Tree-AS, decision list and quick, unbiased, efficient statistical tree (QUEST). In all cases, the model with the continuous variables performed better than the binned model in terms of AUC and accuracy measures.

Most traditional risk factors did contribute to these models as expected, including age, cholesterol, obesity, and hypertension status [26]. Traditionally those that smoke or those with elevated blood glucose values would be at increased risk of cardiovascular disease. In females, smoking was not a significant contributor to any model, possibly because the rates of smoking in females were so low in this sample (2%). In males, smoking was a significant contributor to all models, but smokers showed decreased odds of cardiovascular disease; this finding remained significant after correction for age. Those categorized in the very high blood glucose category also saw decreased odds of cardiovascular disease. It is possible that those individuals with these risk factors were younger or see their healthcare practitioners more often and thus get better care, but more research is needed to clarify these findings.

There is still much room for improvement in the ability of these algorithms to predict cardiovascular disease. The best performing model classified only 73% of cases correctly. Expanding these techniques to a larger dataset with more potential risk factors may increase the effectiveness of these techniques. Qi *et al.* found that machine learning techniques applied to electronic medical records (EMR) produced the best prediction models (AUC 0.902) when the EMR data contained both longitudinal and cross-sectional patient data [27]. Distinct types of CVD may require different modeling approaches. Krittanawong found that boosting algorithms best predicted coronary artery disease while support vector machine algorithms best predicted stroke [28]. Kwon *et al.* reported that feedforward neural network and gradient boosting machine algorithms predicted adverse cardiac events following invasive coronary treatments [29].

ROC analysis not only provided support for the increased accuracy of the continuous SBP model, but also provided guidance on clinical cut points that could be utilized in this dataset [30]. Unfortunately, dichotomizing hypertension status based on these new cut points did not improve accuracy over the model

which used SBP as a continuous variable. These new cut points would have limited applicability outside of this dataset as they were based on this particular group of subjects.

Supervised learning algorithms in SPSS Modeler appeared to meet or exceed the accuracy of the original decision tree models. Neural net appears to function the best in females and LSVM and Tree-AS also performed well. In males, logistic regression was the best algorithm while LSVM and neural net were close seconds. Once again, SBP as a continuous variable performed the best.

5. CONCLUSION

In conclusion, our study shows that changing the criteria by which individuals are classified as hypertensive or normotensive negatively impacted the ability of decision tree to predict cardiovascular disease in both females and males by 7.2% and 9.1%, respectively. While artificially binning continuous variables may offer simplicity and convenience, this practice does indeed negatively impact the predictive capabilities of supervised learning algorithms. Artificially binning hypertension status also decreased the ability of decision tree to predict cardiovascular disease in females and males by 7.8% and 9.6%, respectively. Supervised learning appears to offer advantages over traditional cardiovascular disease risk prediction techniques when applied to available datasets. The ultimate goal of such procedures is to identify those most at risk to direct additional attention and resources to mitigate the risk of cardiovascular disease. It remains to be seen if incorporating supervised and machine learning algorithms into medical practice will increase prognostic ability when the disease status of the patient is unknown.




REFERENCES

- [1] C. H. Han, H. Kim, S. Lee, and J. H. Chung, "Knowledge and Poor Understanding Factors of Stroke and Heart Attack Symptoms," *International Journal of Environmental Research and Public Health*, vol. 16, no. 19, p. 3665, Sep. 2019, doi: 10.3390/ijerph16193665.
- [2] CDC, "Heart Disease Facts," *Cdc.Gov*. Oct. 2022. [Online]. Available: <https://www.cdc.gov/heartdisease/facts.htm> (accessed: Oct, 3, 2022).
- [3] S. H. Walker and D. B. Duncan, "Estimation of the Probability of an Event as a Function of Several Independent Variables," *Biometrika*, vol. 54, no. 1/2, p. 167, Jun. 1967, doi: 10.2307/2333860.
- [4] J. Truett, J. Cornfield, and W. Kannel, "A multivariate analysis of the risk of coronary heart disease in Framingham," *Journal of Chronic Diseases*, vol. 20, no. 7, pp. 511–524, Jul. 1967, doi: 10.1016/0021-9681(67)90082-3.
- [5] H. E. Bays *et al.*, "Ten things to know about ten cardiovascular disease risk factors," *American Journal of Preventive Cardiology*, vol. 5, p. 100149, Mar. 2021, doi: 10.1016/j.ajpc.2021.100149.
- [6] R. B. D'Agostino, S. Grundy, L. M. Sullivan, and P. Wilson, "Validation of the Framingham Coronary Heart Disease Prediction Scores," *JAMA*, vol. 286, no. 2, pp. 180–187, Jul. 2001, doi: 10.1001/jama.286.2.180.
- [7] P. M. Ridker *et al.*, "Rosuvastatin to Prevent Vascular Events in Men and Women with Elevated C-Reactive Protein," *New England Journal of Medicine*, vol. 359, no. 21, pp. 2195–2207, Nov. 2008, doi: 10.1056/NEJMoa0807646.
- [8] A. Chaulin, "Elevation Mechanisms and Diagnostic Consideration of Cardiac Troponins under Conditions Not Associated with Myocardial Infarction. Part 1," *Life*, vol. 11, no. 9, p. 914, Sep. 2021, doi: 10.3390/life11090914.
- [9] M. Bahls *et al.*, "Physical activity, sedentary behavior and risk of coronary artery disease, myocardial infarction and ischemic stroke: a two-sample Mendelian randomization study," *Clinical Research in Cardiology*, vol. 110, no. 10, pp. 1564–1573, Oct. 2021, doi: 10.1007/s00392-021-01846-7.
- [10] S. F. Weng, J. Repps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PLOS ONE*, vol. 12, no. 4, p. e0174944, Apr. 2017, doi: 10.1371/journal.pone.0174944.
- [11] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. F. Rudd, and M. van der Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants," *PLOS ONE*, vol. 14, no. 5, p. e0213653, May 2019, doi: 10.1371/journal.pone.0213653.
- [12] B. Martins, D. Ferreira, C. Neto, A. Abelha, and J. Machado, "Data Mining for Cardiovascular Disease Prediction," *Journal of Medical Systems*, vol. 45, no. 1, p. 6, Jan. 2021, doi: 10.1007/s10916-020-01682-8.
- [13] G. Danaei *et al.*, "The Preventable Causes of Death in the United States: Comparative Risk Assessment of Dietary, Lifestyle, and Metabolic Risk Factors," *PLoS Medicine*, vol. 6, no. 4, p. e1000058, Apr. 2009, doi: 10.1371/journal.pmed.1000058.
- [14] A. V. Chobanian, "The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure The JNC 7 Report," *JAMA*, vol. 289, no. 19, pp. 2560–2572, May 2003, doi: 10.1001/jama.289.19.2560.
- [15] P. K. Whelton *et al.*, "2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Pr," *Hypertension*, vol. 71, no. 6, Jun. 2018, doi: 10.1161/HYP.0000000000000065.
- [16] G. S. Stergiou *et al.*, "2021 European Society of Hypertension practice guidelines for office and out-of-office blood pressure measurement," *Journal of Hypertension*, vol. 39, no. 7, pp. 1293–1302, Jul. 2021, doi: 10.1097/HJH.0000000000002843.
- [17] U. Svetlana, "Cardiovascular Disease dataset," *Kaggle.com*, Oct. 2022. [Online]. Available: <https://www.kaggle.com/sulianova/cardiovascular-disease-dataset/metadata>
- [18] IBM Corp., "IBM SPSS decision trees 21.0," p. 118, 2012. Accessed: Oct, 3, 2022. [Online]. Available: [http://library.uvm.edu/services/statistics/SPSS21Manuals/IBM SPSS Decision Trees.pdf](http://library.uvm.edu/services/statistics/SPSS21Manuals/IBM%20SPSS%20Decision%20Trees.pdf).
- [19] J. Shreffler and M. R. Huecker, "Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios," *StatPearls*, 2020. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/32491423>. Accessed: October 3, 2022.
- [20] D. G. Altman and P. Royston, "The cost of dichotomising continuous variables," *BMJ*, vol. 332, no. 7549, p. 1080.1, May 2006, doi: 10.1136/bmj.332.7549.1080.
- [21] A. Bhardwaj, "Framingham heart study dataset," *Kaggle.com*, Oct. 2022. [Online]. Available: <https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset> (accessed Oct. 3, 2022).
- [22] N. Masih and S. Ahuja, "Application of data mining techniques for early detection of heart diseases using Framingham heart study dataset," *International Journal of Biomedical Engineering and Technology*, vol. 38, no. 4, pp. 334–344, 2022, doi:




- 10.1504/IJBET.2022.123149.
- [23] A. M. Kuruvilla and N. Balaji, "Heart disease prediction system using Correlation Based Feature Selection with Multilayer Perceptron approach.," *IOP Conference Series: Materials Science and Engineering*, vol. 1085, no. 1, p. 012028, Feb. 2021, doi: 10.1088/1757-899X/1085/1/012028.
- [24] H. Agrawal, J. Chandiwala, S. Agrawal, and Y. Goyal, "Heart Failure Prediction using Machine Learning with Exploratory Data Analysis," in *2021 International Conference on Intelligent Technologies (CONIT)*, Jun. 2021, pp. 1–6. doi: 10.1109/CONIT51480.2021.9498561.
- [25] S. S. Mahmood, D. Levy, R. S. Vasan, and T. J. Wang, "The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective," *The Lancet*, vol. 383, no. 9921, pp. 999–1008, Mar. 2014, doi: 10.1016/S0140-6736(13)61752-3.
- [26] G. Liguori, Y. Feito, C. Fountaine, and B. Roy, Eds. *ACSM's guidelines for exercise testing and prescription*, 11th ed. Philadelphia: PA, USA: Wolters Kluwer, 2021.
- [27] Q. Li, A. Campan, A. Ren, and W. E. Eid, "Automating and improving cardiovascular disease prediction using Machine learning and EMR data features from a regional healthcare system," *International Journal of Medical Informatics*, vol. 163, p. 104786, Jul. 2022, doi: 10.1016/j.ijmedinf.2022.104786.
- [28] C. Krittawong *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Scientific Reports*, vol. 10, no. 1, p. 16057, Sep. 2020, doi: 10.1038/s41598-020-72685-1.
- [29] O. Kwon *et al.*, "Electronic Medical Record–Based Machine Learning Approach to Predict the Risk of 30-Day Adverse Cardiac Events After Invasive Coronary Treatment: Machine Learning Model Development and Validation," *JMIR Medical Informatics*, vol. 10, no. 5, p. e26801, May 2022, doi: 10.2196/26801.
- [30] I. Unal, "Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach," *Computational and Mathematical Methods in Medicine*, vol. 2017, pp. 1–14, 2017, doi: 10.1155/2017/3762651.

BIOGRAPHIES OF AUTHORS






Christopher M. Bopp    holds a Ph.D. in human physiology from Kansas State University and is a recent graduate of the Master of Science in Health Informatics program at Slippery Rock University. His research interests include cardiovascular disease prediction and prevention. He can be contacted at email: cmbopp@gmail.com.






William Briggs    holds a Bachelors in Petroleum and Natural Gas Engineering and a minor in math and is a recent graduate of the Master of Data Analytics program at Slippery Rock University. His research interests include Data Mining. He can be contacted at email: billybriggs577@gmail.com.



Catherine Orlando    is a recent graduate of the Master of Science in Health Informatics program at Slippery Rock University. Her research interests include Data Mining. She can be contacted at email: corlando12@uri.edu.



Raed Seetan    is an associate professor at the Computer Science Department, Slippery Rock University. His research interests include Bioinformatics, Data Mining and Machine Learning. He can be contacted at email: raed.seetan@sru.edu.