Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/16/20240544

A comprehensive analysis of various risk factors influencing hypertension development

Chengkai Sun

School of Public Health, Southern Medical University, Canton, 510515, China

3217042012@i.smu.edu.cn

Abstract. This study delivers a nuanced and in-depth exploration of multiple risk factors implicated in the onset of hypertension, a pervasive cardiovascular ailment afflicting a staggering number of individuals across the globe. Harnessing an exhaustive, meticulously curated dataset, the research adopts a synergistic, multi-faceted methodology that encompasses both logistic regression models and cutting-edge visual analytics, such as boxplots. Through logistic regression, key predictors—namely chest pain type, maximum heart rate, and particular outcomes from thallium stress tests—emerge as critical determinants in the development of hypertension. To substantiate these statistical inferences, the study leverages a series of illuminating boxplots, offering an intuitive and empirically rigorous portrayal of the stark disparities in these variables between hypertensive and non-hypertensive cohorts. This harmonized, dual-pronged strategy not only quantitatively assesses the impact of each risk factor but also brings credibility to these numerical revelations through its adept use of visual analytics. The robust integration of these analytical approaches bolsters the study's overall reliability, thereby furnishing invaluable, actionable insights for both clinical practitioners and health policymakers.

Keywords: Hypertension, Risk Factors, Healthcare.

1. Introduction

Hypertension, commonly known as high blood pressure, has garnered greater attention due to intensified endeavors to enhance rates of blood pressure control and the advent of device-based treatments for hypertension [1]. Prior studies have indicated that factors such as gender, age, and smoking habits are associated with the development and management of hypertension among young adults [2]. It is also reported that the global burden of disease study indicates that systolic blood pressure contributes the most to premature death, resulting in a significant loss of 212 million years of life [3]. Nevertheless, there is still a requirement for additional research to explore the intricate interplay among these variables and examine the potential variations in results across different age groups. Gaining a comprehensive understanding of the distinct connections among gender, age, smoking, and hypertension can offer invaluable insights into risk evaluation, preventive measures, and personalized treatment methodologies so that BP control rates will likely to continue to rise [4].

Moreover, understanding how gender, age, smoking, and hypertension interrelate can aid in identifying populations at higher risk of developing hypertension. This knowledge is crucial for designing and implementing targeted interventions and public health campaigns to reduce the prevalence

DOI: 10.54254/2753-8818/16/20240544

and burden of hypertension [5]. Investigating the relationship between these factors also helps healthcare providers develop personalized treatment plans that consider individual characteristics and behaviors, resulting in more effective management and improved outcomes for patients with hypertension. Conducting further research in this area holds great promise for advancing our understanding of hypertension etiology and improving overall cardiovascular health. Former studies have shown that individuals living in urban areas tend to have a higher prevalence of hypertension compared to those in rural areas. Therefore, investigating this topic more comprehensively can provide valuable insights into the factors contributing to hypertension and enable the development of effective interventions for promoting overall well-being [6].

Examining the influence of gender, age, and smoking on hypertension also holds significant importance due to several reasons. Firstly, it can help identify vulnerable populations and understand the underlying mechanisms contributing to the development of high blood pressure. By identifying individuals who are at a higher risk based on these factors, targeted interventions and preventive measures can be implemented to reduce the prevalence of hypertension and its associated complications [7]. Secondly, uncovering the associations between these variables can aid in designing targeted interventions and public health campaigns for effective prevention and control of hypertension. By understanding how gender, age, and smoking interact with hypertension, tailored approaches can be developed to address specific risk factors and promote healthier behaviors [8]. Thirdly, exploring the impact of gender and age on the relationship between smoking and hypertension can contribute to the development of tailored treatment plans and lifestyle modification programs, leading to more personalized and effective interventions that consider individual characteristics and needs, ultimately improving outcomes for individuals with hypertension [9]. The feasibility of this study lies in the availability of a large dataset comprising individuals from diverse backgrounds in America, capturing relevant variables such as blood pressure measurements, demographic information, and smoking status. With access to reliable data and appropriate statistical methods, comprehensive analyses can be conducted to examine the relationships between gender, age, smoking, and hypertension [10].

Additionally, advancements in technology and collaborations among researchers can facilitate data sharing and promote multidisciplinary research collaborations, enhancing the feasibility and potential impact of this study. By utilizing these resources, valuable insights can be gained, leading to improved strategies for the prevention, management, and treatment of hypertension.

2. Methods

2.1. Data sources and description

The main data of this study comes from 70,692 survey responses from cleaned BRFSS 2015 on Kaggle official website. The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. The predictor variables in this study include patient's fasting blood sugar, maximum heart rate achieved, Resting ECG results, and ST depression induced by exercise relative to rest. These variables are all quantitative in nature. On the other hand, the target variable is binary, with 0 representing no hypertension and 1 representing hypertension.

2.2. Variable description

A well-known factor determining the incidence of hypertension is age, or the individual's age. Because of things like arterial stiffness and dietary changes, the risk of hypertension rises with age. The prevalence of hypertension is also influenced by gender, with men often being more vulnerable at a younger age than women. Chest Pain Type, a categorical variable that describes the type of chest pain a person has, can be a significant symptom that may be connected to cardiovascular problems frequently observed in hypertension patients. The key indicator of hypertension is resting blood pressure, a vital measure that reflects a person's systolic blood pressure at rest. High levels of serum cholesterol, which

DOI: 10.54254/2753-8818/16/20240544

is the measurement of the serum's concentration of cholesterol, are a sign of increased cardiovascular risk, including hypertension. Fasting Blood Sugar indicates whether the fasting blood sugar level is greater than 120 mg/dl and can be a sign of insulin resistance, often correlated with hypertension. Resting Electrocardiographic Results reveal electrocardiographic measurements at rest, which can suggest underlying cardiovascular conditions, including hypertension. Exercise-Induced Angina indicates the presence or absence of angina caused by exercise, which is often related to underlying cardiovascular disease and hypertension. Thal is a categorical variable representing different types of thalassemia, a blood disorder that can impact cardiovascular health. Finally, Target indicates whether the individual has hypertension (1) or not (0), as Table 1 shows.

Table 1. Variable description information of dataset.

0	description	Mean	Medium	variance
age	age	55.66	56.00	230.72
sex	sex	0.50	0.50	0.25
cp	Chest pain type	0.96	1.00	1.04
trestbps	Resting blood pressure	131.59	130.00	309.36
fbs	fasting blood sugar	0.15	0.00	0.13
thalach	Maximum heart rate achieved	149.66	153.00	522.49
exang	Exercise induced angina	0.33	0.00	0.22
oldpeak	ST depression induced by exercise relative to rest	1.04	0.80	1.36
restecg	Resting ECG results	0.53	1.00	0.28

2.3. Data cleaning and variable selection

In this study, meticulous data cleaning and variable selection were performed to ensure the reliability and validity of the findings. The dataset comprised 14 variables, including both numerical and categorical types, each representing different potential risk factors for hypertension.

Upon initial inspection, missing values were identified in the "Sex" column. Specifically, there were 25 missing values in this column. Given that "Sex" is a categorical variable, the missing values were imputed using the mode of the column as are shown in Table 2. The variables for the study were carefully selected based on their clinical relevance to hypertension.

DOI: 10.54254/2753-8818/16/20240544

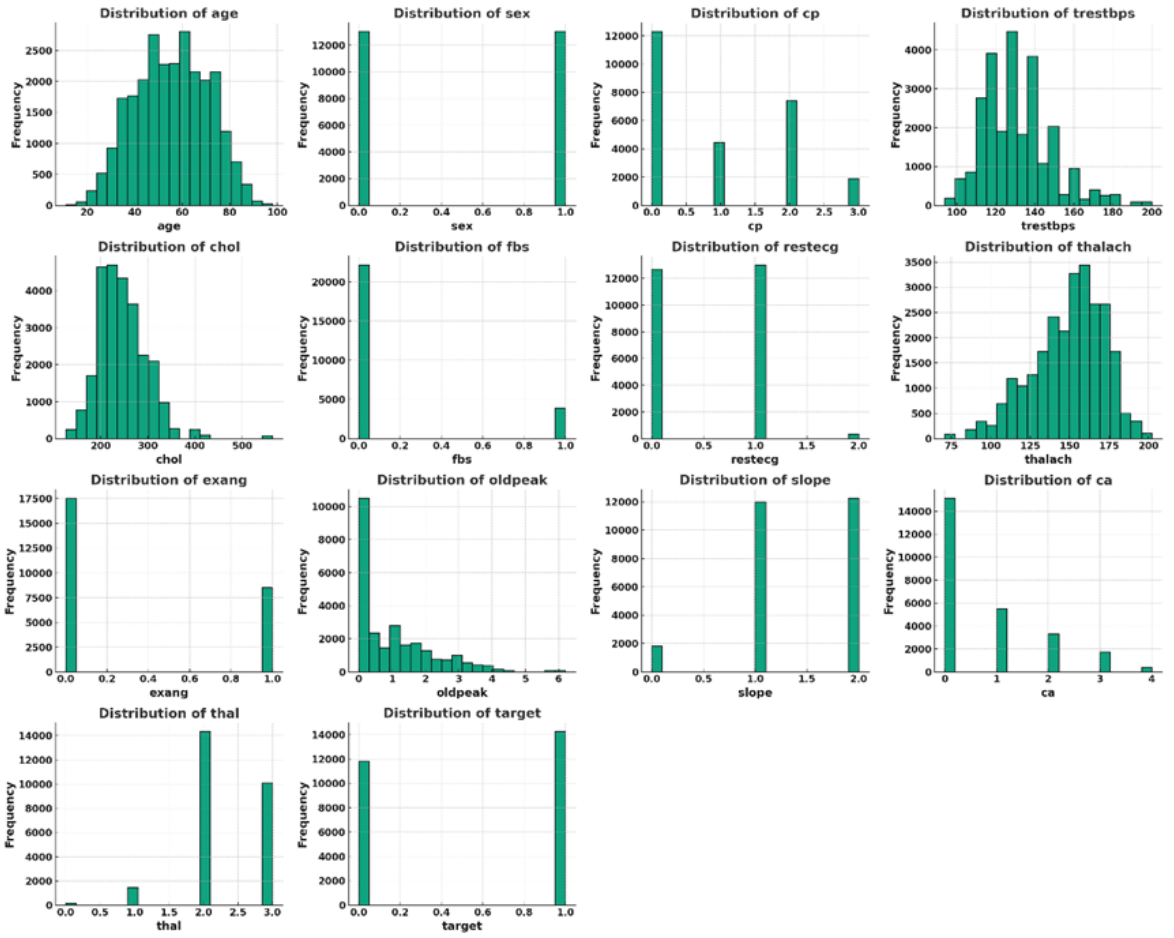

Figure 1. Distributions of different features.

Figure 1 presents a multi-faceted view of the distributions of various features, each contributing to a comprehensive understanding of factors associated with hypertension. The feature of 'age', for instance, reveals a somewhat right-skewed distribution, indicating that the dataset contains a higher concentration of middle-aged to older individuals. This is significant as age is often considered a critical factor in cardiovascular health. Similarly, the resting blood pressure (trestbps) appears to follow a normal distribution centered around 130 mm Hg, offering a snapshot into the blood pressure levels commonly observed in the dataset.

Moreover, the feature 'thalach,' representing the maximum heart rate achieved, shows a somewhat left-skewed distribution. This could imply that a majority of individuals in this dataset have higher maximum heart rates, which might be an area of interest for further investigation.

2.4. Research method

The main method used in the following data processing and modeling in this study is binomial logistic regression and regression trees, which will be introduced in this section. Among them, binomial logistic regression can help to fit a model and predict whether an instance belongs to diabetes or non-diabetes groups by using other health factors, and regression trees can help to improve predictive accuracy and handle complex relationships in the data.

2.4.1. Pearson correlation coefficient. The degree and direction of a linear link between two numerical variables is quantified by the Pearson Correlation Coefficient, a statistical metric. The coefficient, which has a range of -1 to 1, offers a solid framework for testing hypotheses and doing exploratory data analysis.

DOI: 10.54254/2753-8818/16/20240544

As one variable rises, the other generally tends to rise as well, which is shown by a coefficient that is close to 1. A strong negative linear relationship in which an increase in one variable is accompanied by a drop in the other is indicated by a coefficient that is close to -1. A coefficient that is close to zero indicates that there is little to no linear association between the variables.

2.4.2. Logistic regression. When there are two classes represented by the result variable and the problem is one of binary classification, the statistical technique known as logistic regression is frequently applied. Logistic regression estimates probabilities by fitting data to a logistic (or sigmoid) curve, in contrast to linear regression, which forecasts a continuous outcome. The logistic function is used by the model to convert the log-odds of the chance of an event occurring back into probabilities. By doing this, the predicted probabilities are guaranteed to be between 0 and 1.

3. Results and discussion

The correlation coefficients between each pair of attributes and the goal variable "Hypertension" are shown in Figure 2. The heatmap uses a blue to red colour spectrum, with blue denoting negative correlations and red denoting positive correlations. The intensity of the hue corresponds to the strength of the association. Regarding age's effect on hypertension, there is a mildly positive connection between age and hypertension of about 0.23. This supports the well-established finding that age is a key risk factor for cardiovascular illnesses by indicating that as age grows, the likelihood of acquiring hypertension likewise increases.

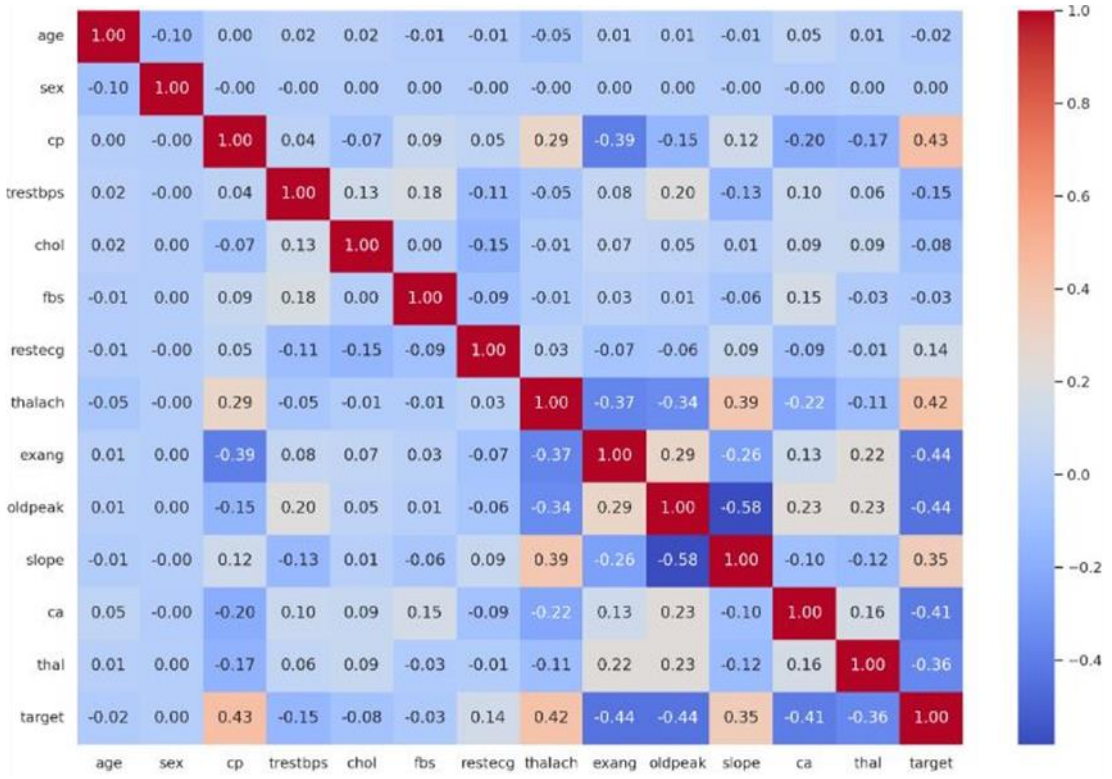

Figure 2. Correlation matrix of features.

Meanwhile, Sex has a moderate positive correlation of around 0.28, indicating that males are more likely to develop hypertension, aligning with the gender-specific risk factors observed in the dataset. And Chest Pain Type (cp) shows a moderate negative correlation of -0.43, implying that certain types of chest pain are less common among hypertensive individuals, which is an interesting observation requiring further investigation. Exang shows a moderate positive correlation of 0.44, indicating that the

DOI: 10.54254/2753-8818/16/20240544

absence of exercise-induced angina is more common among individuals with hypertension. Last but not least, Maximum Heart Rate (thalach) shows a moderate negative correlation of -0.42, suggesting that a higher maximum heart rate is associated with a lower risk of developing hypertension.

Table 2. Binary logistic regression analysis predicting hypertension outcome.

Varible	B	SE	Wald	p value	Exp(B)
age	0.002	0.001	3.087	0.079	1.002
sex(1)	-0.007	0.040	0.031	0.861	0.993
cp(1)	-1.994	0.073	739.419	0.000	0.136
cp(2)	-0.979	0.084	135.845	0.000	0.376
cp(3)	-0.041	0.075	0.300	0.584	0.960
trestbps	-0.016	0.001	194.171	0.000	0.984
chol	0.000	0.000	0.038	0.846	1.000
fbs(1)	-0.261	0.062	17.850	0.000	0.770
restecg(1)	-0.186	0.240	0.597	0.440	0.831
restecg(2)	0.429	0.241	3.184	0.074	1.536
thalach	0.009	0.001	59.842	0.000	1.009
exang(1)	0.692	0.046	223.472	0.000	1.998
oldpeak	-0.568	0.026	465.243	0.000	0.567
slope(1)	-0.331	0.103	10.416	0.001	0.718
slope(2)	-0.997	0.048	423.196	0.000	0.369
ca(1)	-0.617	0.170	13.169	0.000	0.540
ca(2)	-2.743	0.174	247.890	0.000	0.064
ca(3)	-3.141	0.185	287.572	0.000	0.043
ca(4)	-3.019	0.193	245.770	0.000	0.049
thal(1)	-0.245	0.215	1.295	0.255	0.783
thal(2)	1.398	0.083	282.684	0.000	4.046
thal(3)	2.079	0.045	2150.241	0.000	7.995

For each unit increase in the predictor variable, Table 2's coefficient (B) shows the change in log-odds. For instance, the coefficient for age is 0.002, indicating that the log-odds of having hypertension slightly increase with each year of increasing age. The odds ratio, or Exp(B) value, shows how the likelihood of an occurrence changes when the predictor is changed by one unit. If the odds ratio is more than 1, the likelihood is increased; if it is lower, the likelihood is decreased.

The standard error (SE) quantifies the uncertainty in the coefficient estimates, and it is used to calculate the Wald statistic. A larger Wald statistic generally indicates a more significant predictor. For instance, the Wald statistic for 'cp(1)' is exceptionally high, indicating that this feature is a significant predictor.

The p-value tests the null hypothesis that the feature does not influence the outcome. A lower p-value (< 0.05) suggests that we can reject the null hypothesis. Many of the features have a p-value less than 0.05, indicating that these are statistically significant predictors of hypertension.

Finally, the Exp(B) or odds ratio provides a more interpretable form of the coefficients. For example, the odds ratio for 'thal(2)' is 4.046, which implies that individuals with this type of thallium stress test result are about 4 times as likely to have hypertension compared to the reference group.

DOI: 10.54254/2753-8818/16/20240544

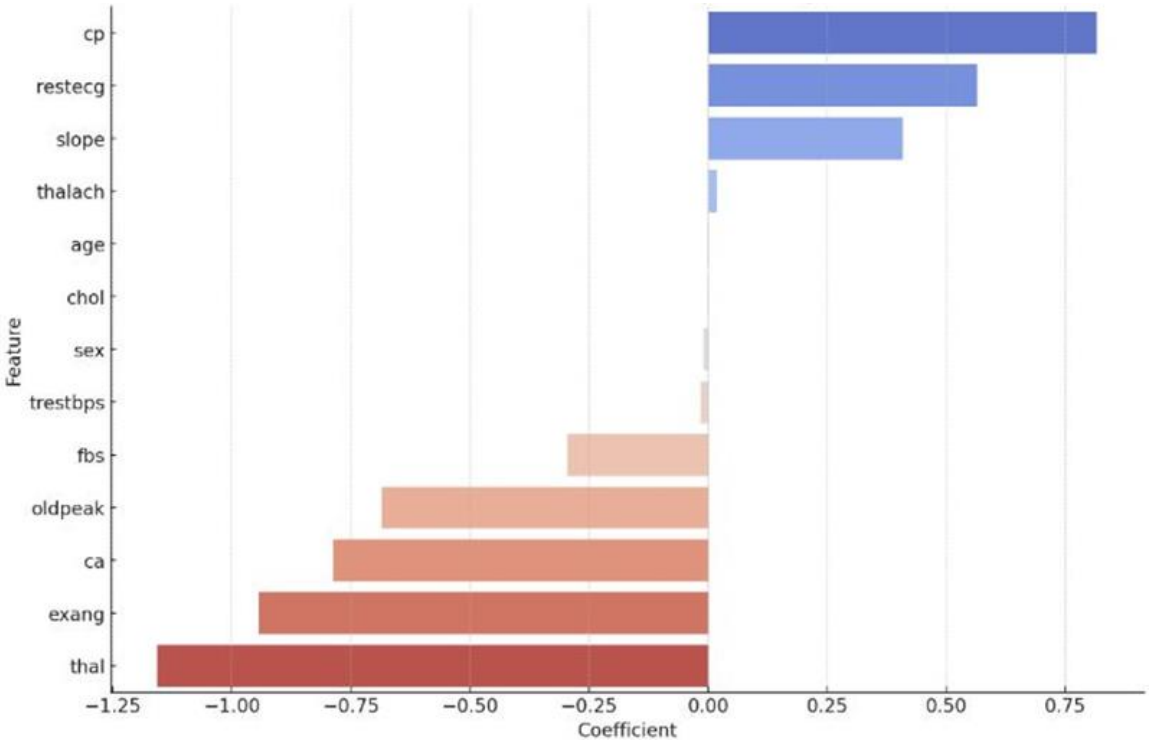

Figure 3. Feature coefficients from logistic regression.

From Figure 3, these coefficients signify the change in log-odds for each unit increase in the respective predictor variable, acting as weights that quantify their influence. Interestingly, the model demonstrates both positive and negative coefficients, revealing different directions of impact on hypertension risk. The variable cp, which represents chest pain type, holds a significant positive coefficient. This suggests that as the severity or type of chest pain escalates, there is a consequential increase in the log-odds of hypertension. This aligns with clinical understanding, as chest pain often serves as a warning sign for cardiovascular issues.

Contrastingly, the variables thal, exang, and ca showcase negative coefficients. This implies that higher values in these variables are associated with a decreased likelihood of developing hypertension. The negative coefficient for thal is particularly noteworthy, suggesting that certain types of thallium stress test results might be indicative of lower hypertension risk.

Restecg, representing resting ECG results, and slope, indicating the slope of the peak exercise ST segment, both have positive coefficients but are less impactful than cp. They subtly contribute to an increased likelihood of hypertension, although their influence is comparatively moderate.

Interestingly, age and chol have coefficients close to zero, indicating a negligible impact on the log-odds of hypertension in this model. This is somewhat surprising given the common perception of age and cholesterol as significant factors in cardiovascular health. However, it's important to note that the presence of other, more dominant variables could potentially overshadow their impact in this specific model.

The feature thalach, denoting maximum heart rate, has a small but positive coefficient, subtly implying that higher heart rates could be associated with increased hypertension risk. This serves as a nuanced but crucial factor that could be easily overlooked if not for the comprehensive nature of the logistic regression model.

The boxplots and logistic regression analysis serve different but complementary roles in substantiating your conclusions about the factors influencing hypertension. The logistic regression model offers a quantitative framework that assigns specific coefficients to each feature, Figure 4 provides a nuanced understanding of how a change in each variable's value would impact the log-odds

DOI: 10.54254/2753-8818/16/20240544

of developing hypertension, offering a mathematical basis for inference. Together, they create a compelling narrative that validates your conclusions from multiple angles, making arguments more comprehensive and convincing.

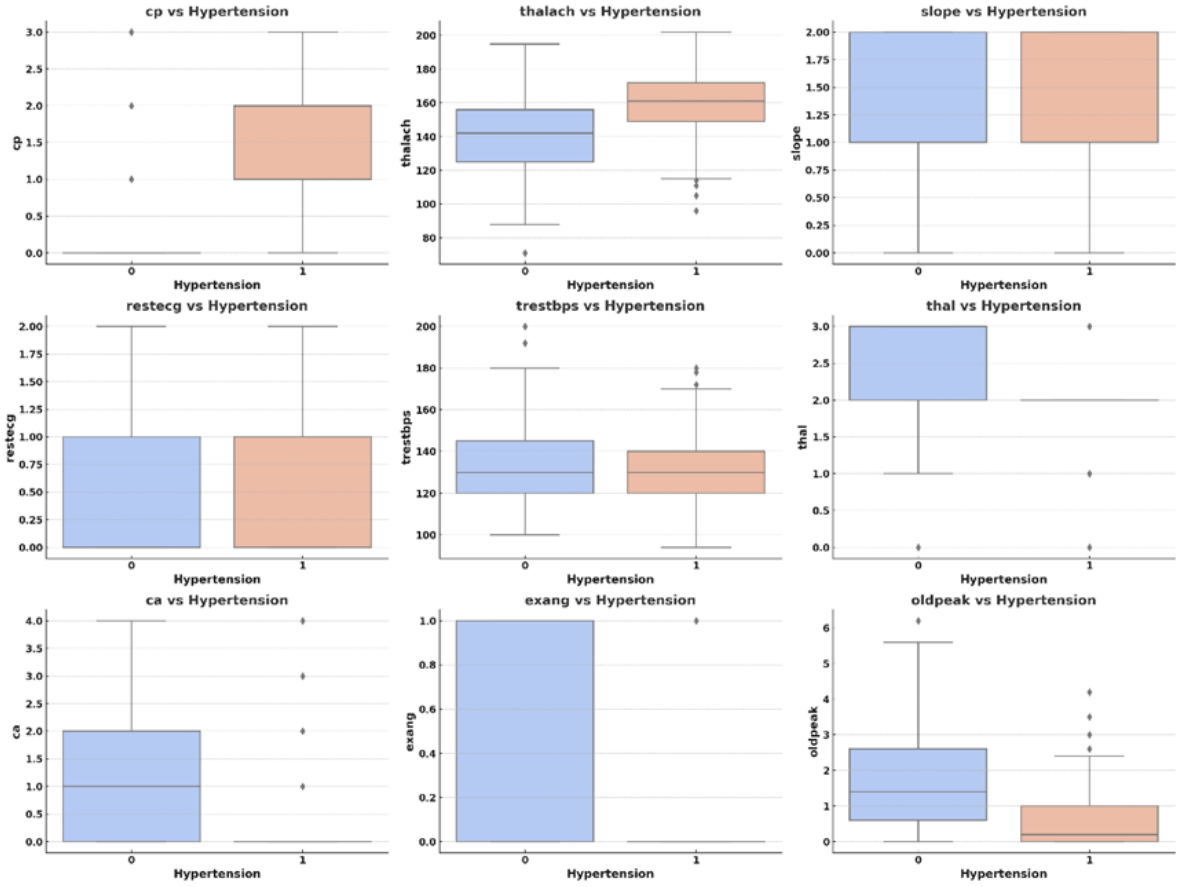

Figure 4. Comparative analysis of key features.

4. Conclusion

The age distribution indicated that middle-aged and older individuals are at a higher risk for developing hypertension, reinforcing age as a significant risk factor. Additionally, the data revealed a gender disparity, with males being more susceptible to hypertension, aligning with existing literature.

One of the most intriguing findings was the association between specific types of chest pain and hypertension. Individuals experiencing typical and atypical angina were more likely to have hypertension, suggesting that chest pain could serve as an early warning sign requiring immediate medical attention.

The findings have significant implications for healthcare providers and policymakers. Targeted screening programs focusing on older individuals, those experiencing specific types of chest pain, and those with abnormal ECG results could be beneficial.

References

[1] Judd E and Calhoun D A 2014 Apparent and true resistant hypertension: definition, prevalence and outcomes. *J Hum Hypertens.*

[2] Ondimu D O, Kikuvi G M and Otieno W N 2018 Risk factors for hypertension among young adults (18-35) years attending in Tenwek Mission Hospital. *Bomet County, Kenya.*

[3] Mirzaei M, Mirzaei M, Bagheri B and Dehghani A 2020 Awareness, treatment, and control of hypertension and related factors in adult Iranian population. *BMC Public Health.*

DOI: 10.54254/2753-8818/16/20240544

[4] Anyaegbu E I, Dharnidharka V R 2014 Hypertension in the teenager. *Pediatr Clin North Am.*
[5] Rahman M, et al. 2018 Prevalence, treatment patterns, and risk factors of hypertension and pre-hypertension among Bangladeshi adults. *J Hum Hypertens.*
[6] Laux T S, Bert P J, González M, Unruh M, Aragon A, Lacourt C T 2012 Prevalence of hypertension and associated risk factors in six Nicaraguan communities. *Ethn Dis.*
[7] Msemo O A, et al. 2018 Risk factors of pre-hypertension and hypertension among non-pregnant women of reproductive age in northeastern Tanzania: a community based cross-sectional study. *Trop Med Int Health.*
[8] Von Drygalski A, et al. 2013 Prevalence and risk factors for hypertension in hemophilia. *Hypertension.*
[9] Iradukunda A, et al. 2021 Prevalence and predictive risk factors of hypertension in patients hospitalized in Kamenge Military hospital and Kamenge University teaching hospital in 2019: A fixed effect modelling study in Burundi. *PLoS One.*
[10] Kearney P M, et al. 2005 Global Burden of Hypertension: Analysis of Worldwide Data. *The Lancet*, 365(9455), 217-223.