ISSN Online: 2821-1936

Transactions on Data Analysis in Social Science

Journal Homepage: https://transoscience.ir

Determining the Factors Affecting the Incidence of Hypertension in Pregnant Women Using Data Mining Techniques

Sh. Borhani[1,*], M. Mohammadi Zanjireh[2], F. Haj Ali Asgari[3]

[1] M.Sc. in Software Engineering, IT Supervisor, Virtual School, Tehran University of Medical Sciences, Tehran, Iran
[2] Assistant Professor, Department of Computer Engineering, Faculty of Engineering, Imam Khomeini International University, Qazvin, Iran
[3] Deputy of Administration and Finance, Virtual School, Tehran University of Medical Sciences, Tehran, Iran

ARTICLE INFO

Article History:
Received 16 January 2019
Received in revised form 4 March 2019
Accepted 21 May 2019
Available online 2 June 2019

Keywords:
C5.4 Decision Tree, CHAID, ID3, Neural Network, Bagging, Boosting, Hypertension in Pregnancy, Data Mining

ABSTRACT

Hypertensive disorders in pregnancy are recognized as one of the major complications during gestation, posing serious risks to both the mother and the fetus. These disorders can result in stillbirths and preterm deliveries among otherwise normal pregnancies and are considered the third leading cause of maternal mortality worldwide. However, their exact etiology remains largely unknown. The main objective of this study was to identify the demographic factors influencing the incidence of hypertension in pregnant women using data mining algorithms. The study database included 4,818 records and 80 features, extracted from electronic health records registered in the Tehran University of Medical Sciences health centers through the SIB system of the Ministry of Health and Medical Education. The study followed the CRISP-DM methodology for implementation. Due to class imbalance in the dataset, modeling was performed in two ways: (1) using basic algorithms such as C5.4 decision tree, ID3, CHAID, and artificial neural networks; and (2) using ensemble methods that combined bagging and boosting with the aforementioned algorithms. According to the developed models, the most significant predictors of hypertension in pregnant women included negative Rh factor, maternal age, nutritional habits (consumption of fruits, salt, and type of oil), history of preeclampsia, smoking, marital status, and presence of other hypertensive risk factors. The results showed that the hybrid model combining C5.4 and CHAID decision trees achieved the highest accuracy (75%) in classifying hypertensive cases. The bagging ensemble with C5.4 and ID3 improved accuracy by 4.17%, while the bagging–neural network combination increased it by 30%. Other models employing bagging and boosting techniques did not show notable improvements.

1. INTRODUCTION

Hypertension in pregnancy is among the most common medical complications during gestation and can lead to severe consequences for both the mother and the fetus [1]. Epidemiological studies indicate that hypertensive

* Corresponding Author: borhani@tums.ac.ir
M.Sc. in Software Engineering, IT Supervisor, Virtual School, Tehran University of Medical Sciences, Tehran, Iran

http://dx.doi.org/10.47176/TDASS.2019.59

disorders affect approximately 5–10% of all pregnancies, making them one of the leading causes of maternal mortality in both developing and developed countries [2].

Several risk factors have been identified for pregnancy-induced hypertension, including obesity, family history of hypertension, advanced maternal age, diabetes, pre-existing renal disease, and multiple pregnancies [1, 3]. These factors can increase the risk of preeclampsia and gestational hypertension, highlighting the need for early detection and effective management strategies.

In recent years, the use of data mining and machine learning techniques for predicting hypertensive disorders during pregnancy has gained significant attention. Studies have shown that data-driven models, such as clustering analysis and metabolomic predictive models, can accurately identify the risk of preeclampsia in early pregnancy stages [4, 5]. For instance, one study utilizing metabolic biomarkers and data mining techniques reported an odds ratio of 36 for predicting preeclampsia [5].

However, there remains a need for more precise, locally trained models to enhance prediction accuracy and clinical applicability. Therefore, the present study aims to identify the factors influencing hypertension in pregnant women using data mining methods, which can uncover hidden patterns in clinical data and contribute to the development of early diagnostic tools to reduce associated complications.

2. LITERATURE REVIEW

Hypertension in pregnancy is a significant health concern that can lead to serious complications for both the mother and the fetus. Understanding the multifaceted factors influencing the incidence of hypertension during pregnancy is crucial for developing effective preventive strategies. This literature review synthesizes existing research findings on various determinants of hypertensive disorders in pregnant women, particularly focusing on the utility of data mining techniques in identifying these factors.

Nutritional factors have been shown to play a critical role in the development of hypertension during pregnancy. One notable study indicates that dietary phytate intake inhibits the bioavailability of essential minerals such as iron and calcium in the diets of pregnant women. Deficiencies in these minerals are linked to complications such as pre-eclampsia, a hypertensive disorder [6](Hasan et al., 2016). This suggests that addressing dietary intake could mitigate the risk of hypertension in pregnant women. Furthermore, a study on selenium status found that low levels of selenium are associated with an increased risk of pregnancy-induced hypertension (Rayman et al., 2015) [7]. This underscores the importance of nutrition and micronutrient management in pregnancy as potential avenues for hypertension prevention.

Environmental influences, including exposure to organophosphate flame retardants (PFRs), have emerged as potential risk factors for hypertension in pregnant women. Research indicates that the presence of these chemicals may have implications for maternal health, including blood pressure regulation (Romano et al., 2017) [8]. Understanding the interaction between environmental toxins and traditional risk factors for hypertension is critical, as it may provide insights into comprehensive prevention strategies that consider both environmental and lifestyle factors.

Subclinical thyroid disease has been identified as a predictor of hypertension in pregnant women, highlighting the need to consider thyroid function when assessing risk (Wilson et al., 2012) [9]. Additionally, genetic factors may also contribute to the incidence of hypertension, as evidenced by studies utilizing genetic neural networks for predicting heart disease based on risk factors (Amin et al., 2013) [10]. This approach could be extended to develop predictive models for hypertension in pregnant populations.

Research on anthropometric measures has revealed that waist circumference, waist-to-hip ratio, and waist-to-height ratio are better predictors of hypertension than body mass index (BMI) in general populations (Rayman et al., 2015) [7]. Applying these findings to pregnant women could enhance screening tools for identifying women at higher risk for developing hypertension. The integration of anthropometric data into predictive models for hypertension could facilitate early intervention and management.

The interplay between various cardiovascular disease risk factors, including those specific to pregnancy such as preeclampsia, has been studied extensively. The association between early-onset preeclampsia and later cardiovascular disease risk emphasizes the long-term implications of hypertensive disorders during pregnancy (Veerbeek et al., 2015) [11]. This indicates a need for longitudinal studies that track women post-pregnancy to better understand the long-term effects of hypertensive disorders and the potential for preventive strategies.

Despite the insights gained from existing studies, several knowledge gaps remain. Firstly, while various factors contributing to hypertension in pregnancy have been identified, the interaction between these factors is not well understood. Future research should employ advanced data mining techniques to analyze large datasets, allowing for the identification of complex relationships between genetic, environmental, nutritional, and physiological factors.

Additionally, there is a limited understanding of how access to healthcare and interventions may influence hypertension management in pregnant women, particularly in low-resource settings (Hill et al., 2013) [12]. Investigating the barriers to healthcare access and their impact on hypertension outcomes could provide valuable insights.

Furthermore, while some studies have focused on specific demographics, there is a need for research that encompasses diverse populations to ensure findings are generalizable. Longitudinal studies that follow pregnant women over time could provide a more comprehensive view of the risk factors associated with hypertension and the effectiveness of interventions.

In conclusion, the incidence of hypertension in pregnant women is influenced by a myriad of factors, including nutritional deficiencies, environmental exposures, physiological conditions, and genetic predispositions. The application of data mining techniques offers promising avenues for synthesizing these factors to develop predictive models that can inform interventions. Addressing the identified knowledge gaps through targeted research will be essential for improving outcomes for pregnant women at risk of hypertension.

3. RESEARCH METHODOLOGY

There are various methodologies available for conducting data mining projects, among which the CRISP-DM model (Cross Industry Standard Process for Data Mining) is recognized as one of the most well-known and widely used approaches [13]. The present study has also been carried out based on the CRISP methodology, the stages of which are illustrated in Figure 1.

3.1. System Understanding

The system understanding phase focuses on comprehending the objectives and requirements of the data mining project. In this study, the aim was to investigate the factors influencing hypertension in pregnant women and to access their medical records. Initially, relevant factors were extracted through a comprehensive literature review and examination of published studies. Subsequently, these factors were confirmed and validated through interviews with faculty members of the Department of Obstetrics and Gynecology at Tehran University of Medical Sciences.

3.2. Data Understanding

The data understanding phase begins with the initial collection of data and continues with its description, comprehension, and quality assessment. The dataset used in this study consisted of 4,818 records and 80 features obtained from the electronic health records of pregnant women who visited health centers in southern Tehran. The dataset contained missing records, which were removed, resulting in 2,926 records included in the study. The remaining data were encoded to reduce the number of features. The class feature in this study was the presence of hypertension, which was categorized into two classes: pregnant women without hypertension and pregnant women with hypertension.

1. System Understanding Phase:

- Conducting a comprehensive literature review on data mining and hypertension among the general population and pregnant women.
- Interviewing faculty members of the Department of Obstetrics and Gynecology at Tehran University of Medical Sciences to identify the factors influencing hypertension in pregnant women.
- Examining the accessibility and structure of electronic health records of pregnant women through the national health information system (SIB System: http://sib.tums.ac.ir/).

2. Data Understanding Phase:

Examining and analyzing the data contained in the electronic health records of pregnant women who visited the health centers located in southern Tehran.

3. Data Preparation Phase:

- Collecting and extracting key data fields.
- Examining the relationships among the data fields.
- Removing incomplete or missing records.
- Performing data preprocessing and preparation.
- Selecting the most influential features for disease diagnosis using the Chi-square test.

4. Modeling Phase:

- Selecting the appropriate algorithm for performing data mining.
- Developing a model by combining the C4.5 decision tree algorithm with the CHAID algorithm.
- Developing a model by combining the C4.5 decision tree algorithm with the ID3 algorithm.
- Developing a model based on artificial neural networks.
- Addressing the data imbalance problem using bagging and boosting techniques.

5. Evaluation Phase:

- Comparing and assessing the performance of the algorithms using evaluation metrics such as sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV).

6. Implementation and Deployment Phase:

- Applying the most effective and efficient algorithm identified during the evaluation phase.

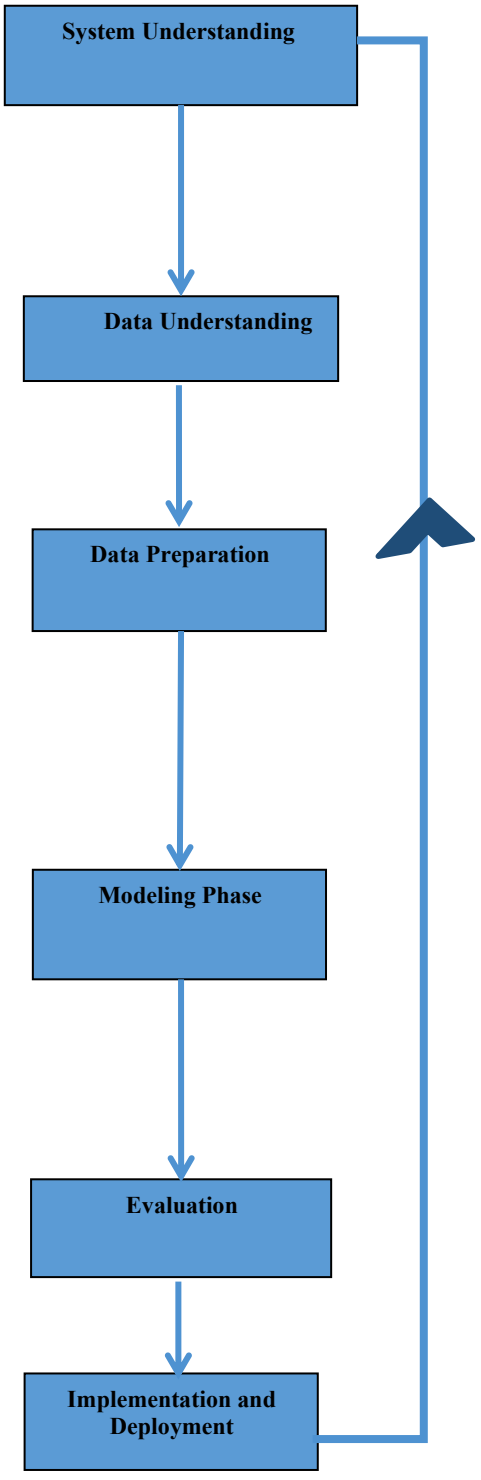

Fig. 1. Activities of the CRISP-DM Methodology Phases

3.3. Data Preprocessing

After extracting the data from the Ministry of Health's SIB system and understanding it with the assistance of obstetrics and gynecology specialists, the data were preprocessed. In general, data preprocessing included two main steps: data integration and data cleaning. Data cleaning involved removing or replacing missing data and normalizing the dataset to bring all variables to a common scale, facilitating easier comparison. The removal and replacement of missing data consisted of three steps: deleting records, imputing missing values, and completely removing irrelevant features.

Certain fields that were interrelated and could be combined were merged with the help of domain experts, as follows:

- Height, weight, and body mass index (BMI) were interdependent. Since BMI can be calculated from height and weight, the height and weight variables were removed.

- The BMI feature and the diagnostic assessment based on BMI represented similar information; therefore, the BMI feature was removed, and the diagnostic assessment feature was retained.

- Features from Stage 3 (presence of diseases such as hepatitis, preeclampsia, asthma, coagulation disorders, AIDS, known thrombophilia, diabetes, tuberculosis, epilepsy, multiple sclerosis, reproductive system anomalies, iron-deficiency anemia, sickle-cell anemia, ischemic heart disease, valvular heart disease, congenital heart anomalies, gastrointestinal disorders, history or presence of breast cancer, kidney disease, thyroid disorders, minor maternal or paternal thalassemia, connective tissue disorders, chronic hypertension, psychiatric disorders, thromboembolism, chronic hypertension) and Stage 4 (lung auscultation, abdominal masses, pale conjunctiva, limb edema, limb bruising, skin rashes, breast masses, arrhythmias, scleral jaundice, abnormal thyroid texture, abnormal thyroid size, thyroid nodules, systolic murmur, diastolic murmur, abnormal breast texture, abnormal or asymmetrical breast appearance, any non-milk breast discharge, abdominal scars, skeletal anomalies, limb erythema, gastrointestinal disorders) and lung auscultation findings (hepatomegaly, splenomegaly, limb pallor) all indicated the presence of disease and were thus combined into a single feature labeled "presence of disease."

- The features "presence of hypertension" and "type of hypertension" both indicated hypertension status; they were combined under the feature "type of hypertension" for this study.

3.4. Feature Selection

After data preprocessing, the feature selection process was performed. To select the relevant features, all fields were first examined, and those with higher informational value were retained, while features with limited contribution to model performance were removed to prevent a decline in modeling quality. Feature selection in this study was conducted using decision trees and the Chi-square test, a non-parametric statistical method, in combination with expert opinions from faculty members. Ultimately, out of the 80 features available in the database, 28 features were selected. The selected features for modeling are presented in Table 1.

Table 1. Key Features Selected Using Decision Trees and the Chi-Square Test

Role	Name	Feature Type	Range/Values
Label	Type of Hypertension	Categorical	Absent (2881), Present (34)
Regular	Gestational Age (weeks)	Nominal	Range1 [-∞ - 10.8] (278), Range2 [10.8 - 18.6] (735), Range3 [18.6 - 26.4] (783), Range4 [26.4 - 34.2] (737), Range5 [34.2 - ∞] (382)
Regular	Age	Nominal	Range1 [-∞ - 16.4] (67), Range2 [16.4 - 23.8] (624), Range3 [23.8 - 31.2] (1278), Range4 [31.2 - 38.6] (718), Range5 [38.6 - ∞] (228)
Regular	Number of Pregnancies	Nominal	Range1 [-∞ - 2] (2496), Range2 [2 - 4] (394), Range3 [4 - 6] (23), Range4 [6 - 8] (1), Range5 [8 - ∞] (1)

Regular	Type of Insurance	Categorical	None (919), Present (1996)
Regular	Population Type	Categorical	Rural (107), Urban (2604), Suburban (204)
Regular	Nationality	Categorical	Non-Iranian (452), Iranian (2463)
Regular	Education Level	Categorical	Below Diploma (596), Diploma (1784), Associate Degree (129), Bachelor's (266), Master's (33), Illiterate (101), Doctorate (5), Not Specified (1)
Regular	Marital Status	Categorical	Never Married (11), Married (2847), Unknown (56), Never Married (1)
Regular	Occupation	Categorical	Housewife (2770), Employed (89), Unknown (56)
Regular	BMI-based Diagnosis	Categorical	Normal (666), Obese (1379), Overweight (832), Underweight (38)
Regular	Fruit Intake	Categorical	2–4 servings or more (2634), Less than 2 servings (247), Rarely/Never (34)
Regular	Vegetable Intake	Categorical	3–5 servings or more (2375), Less than 3 servings (483), Rarely/Never (57)
Regular	Daily Milk and Dairy Consumption	Categorical	Less than 2 servings (551), 3–4 servings or more (2289), Rarely/Never (75)
Regular	Table Salt Use	Categorical	Rarely/Never (2316), Sometimes (364), Always (235)
Regular	Fast Food and/or Carbonated Drinks	Categorical	1–2 times per month (249), Rarely/Never (2561), ≥2 times per week (105)
Regular	Oil Consumption	Categorical	Only semi-solid, solid, or animal oil (232), Only liquid vegetable oil (normal or frying) (2464), Mixed oils (219)
Regular	Physical Activity	Categorical	No activity (2233), Active (682)
Regular	Smoking	Categorical	No tobacco use (2822), Tobacco user (93)
Regular	Allergy	Categorical	No (2884), Yes (31)
Regular	Factor 1: Parents with Hypertension	Categorical	No (2757), Yes (158)
Regular	Factor 2: At least one parent with early-onset coronary artery disease	Categorical	No (2813), Yes (102)
Regular	Factor 3: Dyslipidemia	Categorical	No (2824), Yes (91)
Regular	Factor 4: At least one parent with kidney or endocrine disease	Categorical	No (2819), Yes (96)
Regular	Factor 5: Sleep Apnea	Categorical	No (2911), Yes (4)
Regular	Negative Family History	Categorical	No (2773), Yes (142)
Regular	History of Preeclampsia	Categorical	No (2873), Yes (42)
Regular	History of Gestational Diabetes	Categorical	No (2868), Yes (47)

3.5. Modeling

In this study, modeling was conducted in two stages. Initially, modeling was performed using the basic algorithms: ID3 decision tree, CHAID, and artificial neural networks. Subsequently, due to the imbalanced nature of the dataset, bagging was applied in combination with CHAID, ID3, and neural network algorithms, and AdaBoost was also employed alongside CHAID, ID3, and neural networks.

3.5.1. Modeling with Basic C4.5 and ID3 Decision Tree Algorithms

Evaluation of the models developed using the basic C4.5 and ID3 decision tree algorithms indicated that the model accuracy for women without hypertension was 99.07%, and for women with hypertension, it was 62.50%. Table 2 presents the classification accuracy of the C4.5 and ID3 decision tree algorithms. In the table, the value 2,870 represents women who did not have hypertension and were correctly predicted as "Absent" by the model. The value 27 represents women who actually had hypertension but were incorrectly predicted as "Absent." The value 3 represents women who did not have hypertension but were incorrectly predicted as "Present." The value 5 represents women who had hypertension and were correctly predicted as "Present." The overall accuracy of the model was 97.98% ± 0.34%. In this model, education level was identified as the most important feature.

Table 2. Evaluation of C4.5 and ID3 Decision Tree Algorithms

Predicted \ Actual	Absent	Present	Class Accuracy
Absent (Predicted)	2870	27	99.07%
Present (Predicted)	3	5	62.50%
Recall per Class	99.90%	62.15%	

3.5.2. Modeling with Basic C4.5 and CHAID Decision Tree Algorithms

According to the evaluation results, the model developed using the basic C4.5 and CHAID decision tree algorithms achieved an accuracy of 98.97% for women without hypertension and 75% for women with hypertension. Table 3 presents the evaluation of the C4.5 and CHAID decision tree combination. In the table, the value 2,871 represents women who did not have hypertension and were correctly predicted as "Absent" by the model. The value 30 represents women who actually had hypertension but were incorrectly predicted as "Absent." The value 1 represents women who did not have hypertension but were incorrectly predicted as "Present." The value 3 represents women who had hypertension and were correctly predicted as "Present." The overall accuracy of the model was 98.93% ± 0.24%. In this model, a history of preeclampsia was identified as the most important feature.

Table 3. Evaluation of the C4.5 and CHAID Decision Tree Combination

Predicted \ Actual	Absent	Present	Class Accuracy
Absent (Predicted)	2871	30	98.97%
Present (Predicted)	3	1	75%
Recall per Class	97.99%	9.09%	

3.5.3. Modeling with the Artificial Neural Network Algorithm

According to the evaluation results, the model developed using the artificial neural network (ANN) algorithm achieved an accuracy of 98.93% for women without hypertension and 20% for women with hypertension. Table 4 presents the evaluation of the ANN model. In the table, the value 2,874 represents women who did not have hypertension and were correctly predicted as "Absent" by the model. The value 31 represents women who actually had hypertension but were incorrectly predicted as "Absent." The value 8 represents women who did not have hypertension but were incorrectly predicted as "Present." The value 2 represents women who had hypertension and were correctly predicted as "Present." The overall model accuracy was 98.66% ± 0.32%.

This algorithm required the longest execution time among all the methods because it evaluates all classes of each feature separately. In ANN modeling, a very low learning rate can lead to underfitting, whereas a very high learning rate can cause overfitting; therefore, the optimal learning rate must be determined through trial and error. Based on the analyses conducted in this study, the optimal parameters for the ANN were determined as a learning rate of 0.3, a momentum of 0.2, and 500 training epochs.

Table 4. Evaluation of the Artificial Neural Network

Predicted \ Actual	Absent	Present	Class Accuracy
Absent (Predicted)	**2874**	**31**	**98.93%**
Present (Predicted)	**8**	**2**	**20%**
Recall per Class	**99.72%**	**6.06%**	

3.5.4. Handling Imbalanced Data and Modeling

In this study, the dataset was highly imbalanced, with the normal class (majority) containing a large number of records and the minority class (non-normal) containing significantly fewer records. Consequently, the main class distribution was highly skewed, with the minority class much smaller than the majority class. To address this imbalance, bagging was employed in combination with the C4.5 decision tree, ID3, CHAID, and artificial neural network algorithms. Additionally, the AdaBoost algorithm was applied together with C4.5, ID3, and CHAID decision tree algorithms.

3.5.5. Bagging-Based Modeling Using C4.5 and CHAID Decision Tree Algorithms

Table 5 presents the evaluation of the bagging algorithm in combination with C4.5 and CHAID decision tree algorithms. According to the evaluation, the model achieved an accuracy of 98.03% for women without hypertension and 71.43% for women with hypertension. In the table, the value 2,870 represents women who did not have hypertension and were correctly predicted as "Absent" by the model. The value 28 represents women who actually had hypertension but were incorrectly predicted as "Absent." The value 2 represents women who did not have hypertension but were incorrectly predicted as "Present." The value 3 represents women who had hypertension and were correctly predicted as "Present." The overall model accuracy was 97.98% ± 0.22%.

Table 5. Evaluation of Bagging with C4.5 and CHAID Decision Tree Algorithms

Predicted \ Actual	Absent	Present	Class Accuracy
Absent (Predicted)	**2870**	**28**	**98.03%**
Present (Predicted)	**2**	**5**	**71.43%**
Recall per Class	**99.93%**	**15.15%**	

In this model, a history of preeclampsia was identified as the most important feature. Based on the evaluation metrics, the accuracy of the bagging-enhanced model showed no improvement over the basic algorithm for women without hypertension (increase of 0.04%) and for women with hypertension.

3.5.6. Bagging-Based Modeling Using C4.5 and ID3 Decision Tree Algorithms

Table 6 presents the evaluation of the bagging algorithm in combination with C4.5 and ID3 decision tree algorithms. According to the evaluation, the model achieved an accuracy of 99% for women without hypertension and 67.66% for women with hypertension. In the table, the value 2,870 represents women who did not have hypertension and were correctly predicted as "Absent" by the model. The value 29 represents women who actually had hypertension but were incorrectly predicted as "Absent." The value 2 represents women who did not have hypertension but were incorrectly predicted as "Present." The value 4 represents women who had hypertension and were correctly predicted as "Present." The overall model accuracy was 98.93% ± 0.24%.

Table 6. Evaluation of Bagging with C4.5 and ID3 Decision Tree Algorithms

Predicted \ Actual	Absent	Present	Class Accuracy
Absent (Predicted)	2870	29	99%
Present (Predicted)	4	2	67.66%
Recall per Class	99.93%	12.12%	

Based on the evaluation metrics, the accuracy of the bagging-enhanced model showed an improvement of 0.03% for women without hypertension and 4.17% for women with hypertension compared to the basic algorithm.

3.5.7. Bagging-Based Modeling Using the Artificial Neural Network Algorithm

Table 7 presents the evaluation of the bagging algorithm combined with the artificial neural network (ANN). According to the results, the model achieved an accuracy of 98.94% for women without hypertension and 50% for women with hypertension. In the table, the value 2,880 represents women who did not have hypertension and were correctly predicted as "Absent" by the model. The value 31 represents women who actually had hypertension but were incorrectly predicted as "Absent." The value 2 represents women who did not have hypertension but were incorrectly predicted as "Present." The value 2 represents women who had hypertension and were correctly predicted as "Present." The overall model accuracy was 98.87% ± 0.31%.

Table 7. Evaluation of Bagging with Artificial Neural Network

Predicted \ Actual	Absent	Present	Class Accuracy
Absent (Predicted)	2880	31	98.94%
Present (Predicted)	2	2	50%
Recall per Class	99.93%	6.06%	

The resulting network in this method was a two-layer ANN comprising an input layer with 68 nodes (corresponding to the number of features plus 1 bias node), a hidden layer with 36 nodes, and an output layer with 2 nodes (corresponding to the number of classes). The output layer included the weights assigned to each node in both the hidden and output layers.

Based on the evaluation metrics, the accuracy of the bagging-enhanced ANN showed an improvement of 0.30% for women without hypertension and 0.21% for women with hypertension compared to the basic ANN algorithm.

3.5.8. Boosting-Based Modeling Using CHAID and C4.5 Decision Tree Algorithms

According to the evaluation, the model achieved an accuracy of 98.97% for women without hypertension and 75% for women with hypertension. In the table, the value 2,871 represents women who did not have hypertension and were correctly predicted as "Absent" by the model. The value 30 represents women who actually had hypertension but were incorrectly predicted as "Absent." The value 1 represents women who did not have hypertension but were incorrectly predicted as "Present." The value 3 represents women who had hypertension and were correctly predicted as "Present." The overall model accuracy was 98.93% ± 0.24%. The evaluation results for this model are presented in Table 8.

Table 8. Evaluation of AdaBoost with C4.5 and CHAID Decision Tree Algorithms

Predicted \ Actual	Absent	Present	Class Accuracy
Absent (Predicted)	**2871**	**30**	**97.98%**
Present (Predicted)	**3**	**1**	**75%**
Recall per Class	**97.99%**	**9.09%**	

In this approach, the most important feature identified by the generated trees was a history of preeclampsia. Other trees highlighted features such as age, smoking status, and additional variables. The application of the AdaBoost algorithm combined with C4.5 and CHAID decision trees did not improve the predictive performance for either women without hypertension or women with hypertension; the results were similar to the basic CHAID algorithm and did not resolve the data imbalance issue. Similarly, applying AdaBoost with C4.5 and ID3 decision trees also showed no improvement over the basic CHAID algorithm in evaluation results.

4. EVALUATION

In this study, model evaluation was conducted using confusion matrix metrics, including sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV). Table 9 presents a comparison of the models using the basic algorithms: the combination of C4.5 and ID3 decision trees, the combination of C4.5 and CHAID decision trees, and the artificial neural network (ANN).

Table 9. Comparison of Models Using Basic Algorithms

Metrics	C4.5 + ID3 Decision Tree	C4.5 + CHAID Decision Tree	Artificial Neural Network
Sensitivity	62.50%	75%	20%
Specificity	87%	68.89%	87.09%
Accuracy	97.98%	93.98%	98.66%
Positive Predictive Value	15.62%	9.09%	6.06%
Negative Predictive Value	99.90%	97.99%	99.72%

Based on the evaluation metrics including accuracy, sensitivity, specificity, and positive and negative predictive values the combined C4.5 and ID3 decision tree model outperformed the other models. Table 10 presents a comparison of the models after addressing data imbalance using the bagging technique with the combined algorithms: C4.5 + CHAID, C4.5 + ID3, and ANN.

Table 10. Comparison of Models Using Bagging and Boosting Algorithms

Metrics	Bagging: C4.5 + ID3 Decision Tree	Bagging: C4.5 + CHAID Decision Tree	Bagging: Artificial Neural Network
Sensitivity	66.67%	71.43%	50%
Specificity	86.96%	87%	87.27%
Accuracy	98.93%	93.98%	98.87%
Positive Predictive Value	12.12%	9.09%	6.06%
Negative Predictive Value	99.93%	97.99%	99.93%

As observed, the bagging model combining C4.5 and ID3 decision trees achieved a 4.17% improvement compared to the corresponding basic algorithm, while the bagging model with the neural network showed a 30% improvement relative to its basic version. Considering the information provided in Tables 9 and 10, it can be concluded that the combination of C4.5 and CHAID as a base algorithm outperformed other algorithms, primarily due to its higher overall accuracy.

5. DISCUSSION AND FINAL CONCLUSION

Based on the results obtained from the modeling conducted in this study, the following observations can be made regarding the factors affecting hypertension in pregnant women:

Using the basic C4.5 + ID3 decision tree algorithm, factors such as maternal diet (salt intake, type of oil consumed, fruit consumption), smoking, background factors including a history of preeclampsia and presence of risk factors 1, 2, and 3 for hypertension, education level, marital status, and negative family history (Ehrash) were found to influence the incidence of hypertension with an accuracy of 62.50%. Among these features, education level emerged as the most influential factor.

Using the basic C4.5 + CHAID decision tree algorithm, similar factors influenced hypertension, including maternal diet, smoking, background conditions, education level, marital status, and negative family history, with a prediction accuracy of 75%. In this model, the most influential feature identified was a history of preeclampsia, which formed the primary split in the decision tree.

In the basic artificial neural network model, to determine the most influential features, a feature matrix was constructed. Results showed that BMI, marital status, diet (fruit, vegetables, salt, daily milk and dairy products, type of oil consumed), smoking, history of pre-existing conditions such as preeclampsia, and risk factors 1 and 2 contributed to hypertension with an overall accuracy of 20%. This algorithm required careful tuning of the learning rate to prevent underfitting or overfitting.

Using bagging with the C4.5 + CHAID decision tree, factors such as age, BMI, education level, marital status, diet (fruit, vegetables, salt, fast food/sodas, type of oil), physical activity, smoking, and negative family history influenced hypertension, with an accuracy of 71.43%. In the resulting trees, history of preeclampsia was the most prominent feature.

Using bagging with C4.5 + ID3, factors including BMI, marital status, diet, smoking, negative family history, history of pre-existing conditions, and risk factors 1 and 2 influenced hypertension, with an accuracy of 66.67%. In this model, marital status was the dominant feature in the trees.

Bagging combined with neural networks identified BMI, marital status, diet, smoking, pre-existing conditions, risk factors 1 and 2, physical activity, and age as influential factors, with an accuracy of 50%.

Using boosting with C4.5 + CHAID, factors such as education level, marital status, diet, smoking, negative family history, history of pre-existing conditions, and risk factors 1, 2, and 3 influenced hypertension, with an accuracy of 75%. The most influential feature was history of preeclampsia.

Using boosting with C4.5 + ID3, the influential factors were similar, with an accuracy of 56.55%, and education level was identified as the most critical feature.

Overall, the C4.5 + CHAID decision tree demonstrated the best performance for predicting hypertension in pregnant women, particularly for those who developed the condition, with an accuracy of 75%. The use of bagging with C4.5 + ID3 increased accuracy by 4.17%, and bagging with neural networks improved accuracy by 30%. However, bagging with C4.5 + CHAID decreased overall accuracy by 3.57%. These findings suggest that with a more balanced dataset, bagging with neural networks could potentially yield even better results, as evidenced by the 30% improvement over the basic neural network model.

In conclusion, the most important features associated with the risk of hypertension in pregnant women include negative family history, age, diet (fruit, salt, type of oil), history of preeclampsia, smoking, marital status, and risk factors 1, 2, and 3 for hypertension. These features can be considered major risk factors for developing hypertension during pregnancy.

6. RECOMMENDATIONS

It is recommended that future studies utilize datasets with more uniform and balanced feature distributions to achieve improved predictive performance. Increasing the number of hypertensive cases in the dataset would likely enhance the performance of classification algorithms. Additionally, to gain a better understanding of influential features, future research could combine decision tree algorithms (ID3 and CHAID) and neural networks with association rule mining to generate rules based on feature interactions that contribute to hypertension risk in pregnant women.

Transparency Statement

The data supporting this study are available upon reasonable request to the corresponding author, subject to ethical and confidentiality considerations.

Acknowledgments

We would like to express our gratitude to all individuals who contributed to this project.

Declaration of Interest

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

REFERENCES

[1] Umesawa, M., & Kobashi, G. (2017). Epidemiology of hypertensive disorders in pregnancy: Prevalence, risk factors, predictors and prognosis. Hypertension Research, 40(3), 213–220. https://doi.org/10.1038/hr.2016.126

[2] Hutcheon, J. A., Lisonkova, S., & Joseph, K. S. (2011). Epidemiology of pre-eclampsia and the other hypertensive disorders of pregnancy. Best Practice & Research Clinical Obstetrics & Gynaecology, 25(4), 391–403. https://doi.org/10.1016/j.bpobgyn.2011.01.006

[3] Mustapha, W. M., Sadique, S., Nabiee, R., & Mustapha, W. M. (2012). A comprehensive review of hypertension in pregnancy. Journal of Pregnancy, 2012, 105918. https://doi.org/10.1155/2012/105918

[4] Marić, I., Tsur, A., Aghaeepour, N., Montanari, A., Stevenson, D. K., Shaw, G. M., & Winn, V. D. (2017).

Cluster analysis to estimate the risk of preeclampsia in the high-risk Prediction and Prevention of Preeclampsia and Intrauterine Growth Restriction (PREDO) study. PLoS ONE, 12(3), e0174399. https://doi.org/10.1371/journal.pone.0174399

[5] Kenny, L. C., Broadhurst, D. I., Dunn, W., Brown, M., North, R. A., McCowan, L., Roberts, C., Cooper, G. J. S., Kell, D. B., & Baker, P. N. (2010). Robust early pregnancy prediction of later preeclampsia using metabolomic biomarkers. Hypertension, 56(4), 741–749. https://doi.org/10.1161/HYPERTENSIONAHA.110.157297

[6] Hasan, S. M. A., Hassan, M., Saha, S., Islam, M., Billah, M., & Islam, S. (2016). Dietary phytate intake inhibits the bioavailability of iron and calcium in the diets of pregnant women in rural Bangladesh: A cross-sectional study. BMC Nutrition, 2, 44. https://doi.org/10.1186/s40795-016-0064-8

[7] Rayman, M. P., Bath, S. C., Westaway, J. A. F., Williams, P., Mao, J., Vanderlelie, J. J., Perkins, A. V., & Redman, C. W. G. (2015). Selenium status in UK pregnant women and its relationship with hypertensive conditions of pregnancy. British Journal of Nutrition, 113, 249–258. https://doi.org/10.1017/S000711451400364X

[8] Romano, M. E., Hawley, N. L., Eliot, M. N., Calafat, A. M., Jayatilaka, N. K., Kelsey, K. T., McGarvey, S. T., Phipps, M. G., Savitz, D. A., Werner, E. F., & Braun, J. M. (2017). Variability and predictors of urinary concentrations of organophosphate flame retardant metabolites among pregnant women in Rhode Island. Environmental Health, 16, 40. https://doi.org/10.1186/s12940-017-0247-z

[9] Wilson, K. L., Casey, B. M., McIntire, D. D., Halvorson, L. M., & Cunningham, F. G. (2012). Subclinical thyroid disease and the incidence of hypertension in pregnancy. Obstetrics & Gynecology, 119, 315–320. https://doi.org/10.1097/AOG.0b013e318240de6a

[10] Amin, S., Agarwal, K., & Beg, R. (2013). Genetic neural network-based data mining in prediction of heart disease using risk factors. In Proceedings of the 2013 IEEE Conference on Information and Communication Technologies (pp. 1227–1231). IEEE. https://doi.org/10.1109/CICT.2013.6558288

[11] Veerbeek, J. H. W., Hermes, W., Breimer, A. Y., van Rijn, B. B., Koenen, S. V., Mol, B. W., Franx, A., de Groot, C. J. M., & Koster, M. P. H. (2015). Cardiovascular disease risk factors after early-onset preeclampsia, late-onset preeclampsia, and pregnancy-induced hypertension. Hypertension, 65, 600–606. https://doi.org/10.1161/HYPERTENSIONAHA.114.04850

[12] Hill, J., Hoyt, J., van Eijk, A. M., D'Mello-Guyett, L., ter Kuile, F. O., Steketee, R., Smith, H., & Webster, J. (2013). Factors affecting the delivery, access, and use of interventions to prevent malaria in pregnancy in sub-Saharan Africa: A systematic review and meta-analysis. PLoS Medicine, 10(7), e1001488. https://doi.org/10.1371/journal.pmed.1001488

[13] Wiemer, H., Drowatzky, L., & Ihlenfeldt, S. (2019). Data mining methodology for engineering applications (DMME): A holistic extension to the CRISP-DM model. Applied Sciences, 9(12), 2407. https://doi.org/10.3390/app9122407