VIETNAMESE BRA SIZE CLASSIFICATION WITH MACHINE LEARNING

Hong Nhung Luu Thi
Hung Yen University of Technology and Education
luuhongnhung228@gmail.com

Truong-Phuc Nguyen
Hung Yen University of Technology and Education
n.t.phuc149@gmail.com

Tien-Dat Nguyen
Hung Yen University of Technology and Education
nguyendatdtqn@gmail.com

Thi-Nguyen Le
Hanoi University of Industry
le.nguyenthi@haui.edu.vn

Nhat-Trinh Nguyen
Ha Noi University of Sscience and Ttechnology
trinh.nguyennhat@hust

Hoang Luu
Hung Yen University of Technology and Education
luugiaphucloc@gmail.com

February, 2024

ABSTRACT

Artificial intelligence technology is rapidly developing and is widely applied in the garment industry. This study aims to use machine learning models to classify three groups of female students' breasts based on dimensions related to breast sizes: large, medium, and small. To do that, we first collect 460 samples from female students in the North of Vietnam by using 21 measurements. These measurements are used to train classification models by using Logistic Regression, Support Vector Machines, Random Forest, XGBoost, and artificial neural networks. Experimental results by using k-fold cross-validation show that neural networks achieve promising results compared to other classification models. The investigation of feature contribution is also conducted to show the role of features for each classifier. The promising experimental results facilitate to build a demo system where audiences can experiences with bra size classification.

1 Introduction

Recently, the integration of Artificial Intelligence (AI) into various business sectors is becoming increasingly widespread, especially in the context of strong technological development, such as Industry 4.0 in Vietnam and Industry 5.0 in Western countries. One of the notable applications of AI in the current era is in the field of smart fashion, where the ability to create and apply machine learning models based on collected data is becoming more common.

The term "smart fashion" is becoming a keyword to describe the integration of technology into the clothing classification process. Particularly, with the diversity in forms and styles of fashion items, data is becoming richer, simultaneously creating challenges in product classification for both buyers and sellers. This has driven the demand for using machine learning models, which help in making quick and accurate predictions and classifications.

This study focuses on the application of machine learning models to classify the bra sizes of women living in Northern Vietnam. With the creativity of AI, we hope to make accurate and quick predictions, helping to optimize the bra size selection process, bringing a smart and convenient fashion shopping experience to consumers. At the same time, the study also opens up prospects for the widespread application of AI methods in the modern fashion industry.

Yi [4] developed an AlexNet neural network model to address the task of classifying and recognizing clothing styles. Zhang and colleagues [5] used transfer learning techniques based on the Inception-V3 structure to complete the

recognition of clothing styles. To address the issue of recognizing detailed clothing styles, Li and colleagues [6] also proposed an improved linear CNN model for clothing style recognition.

2 Related Work

Women's breast shapes are highly diverse, categorized based on breast form or fundamental sizes. The characteristic sizes of female breasts serve as crucial foundations for grouping female breasts to facilitate the process of bra design and sizing. Several studies on breast shapes and classification of women's breasts have been conducted based on age characteristics and regional peculiarities in different countries. However, each study in each country evaluates and selects different shapes and characteristic sizes. For example, Martin divides female breasts into four types based on breast form: flat breasts, hemispherical breasts, conical breasts, and pendulous breasts. Similarly, based on breast form, the Wakoru Institute of Human Science in Japan categorizes female breasts into six groups: flat breasts, cone-shaped breasts, hemispherical breasts, projecting breasts, droplet type 1, and droplet type 2.

Lim and colleagues conducted a study on bra cup sizes and breast classification based on circumference and volume. The satisfaction of wearing bras among 182 women aged twenty was analyzed. The results indicated that using breast circumference to establish cup size was appropriate. Female breasts were grouped based on breast circumference and breast volume. According to breast volume, female breast groups were classified as follows: flat breasts with breast volume under $200cc$, cone-shaped breasts ranging from 200 to $300cc$ and 300 to $400cc$, small hemispherical breasts less than $200cc$ and 200 to $300cc$, 300 to $400cc$, projecting breasts at 200 to 300cc, 300 to 400cc, 400cc and above, and droplet type breasts at 400cc. Breast volume and circumference were correlated with each other. Rong Zheng et al [7] proposed a bra sizing system for Chinese women that uses bust circumference and bust width/depth ratio as the main parameters of sizing. Breast shape is classified based on 8 parameters: BMI, breast volume, inner chest arc, outer chest arc, height, nipple direction, upper chest slope, lower chest shape . Among the parameters related to breast shape, bust circumference and chest depth/width ratio are the most important [7]. Kweon and Sohn classified the breasts of women in their twenties into flat breasts, cone breasts, hemispherical breasts, protruding breasts, and teardrop breasts (figure 1.5) [6]. Cho and Sohn classified women's breasts in their 20s into large breasts, medium breasts, and small breasts [8]. Lim grouped breasts into 3 types: 75A, 75AA, and 75B [5]. In Vietnam, author Tran Thi Minh Kieu and her colleagues surveyed the breast shape of North Vietnamese female students aged 18-25 and the compatibility of the visual image of some bra types with the breast types of women. North Vietnamese female students [10]. The study is based on the wearer's perception and expert assessment of the fit of bra cups for different breast types. Author Pham Thi Tham and her colleagues researched the influence of chest angle on women's bra design. The shape index has been determined based on the chest angle and several other dimensions [11]. Currently, there are many artificial neural networks used for classification in the textile and garment field such as: Using Networks (R-FCN) to classify clothes and applying the improved model HSR-FCN for high efficiency in enhanced recognition of deformed clothing and shorter training time, the accuracy rate is 3

Classification of female breasts has been done for many ages, women in many countries, regions, etc. based on shape or some size parameters. In Vietnam, a number of studies on anthropometric characteristics of the human body and building a number system for different subjects have been carried out. However, the chest has many sizes, these sizes greatly affect the comfort of wearing a bra. And classifying breasts is quite difficult. Studies on the breast subgroups of Vietnamese women in general and North Vietnamese female students in particular have not been conducted fully and in detail. To analyze breast anthropometric characteristics and breast grouping, determining breast sizes is essential. The method for determining the typical size of the chest has not been fully presented in studies. Several methods have been used to classify breasts such as: Classification using Kmean-Clustering (NC group), classification using Random Forest (), ... This study aims to use AI technology to classify breasts. To achieve this goal, the research team measured the breast sizes of 460 female students in Northern Vietnam, selected 8 important sizes using the Random Forest method, evaluated the importance of the sizes and Compare with some other classification methods such as:....to find a highly effective and suitable classification method.

3 Method

This section shows the proposal for the prediction problem. It first states the problem and then describes data collection. It next summarizes feature selection and finally overviews classification methods.

3.1 Problem Statement

The prediction is formulated as a multi-class classification problem. Let $\mathcal{D}$ be a set of n instances, $\mathcal{Y} = \{y_1, y_2, ..., y_k\}$ be the label set with k class labels, $\boldsymbol{\mathcal{X}} = \mathbb{R}^{n \times m}$ denotes the m dimensional feature space corresponding to n instances.

The task is to train a classifier (a mapping function) $f : \boldsymbol{\mathcal{X}} \rightarrow \mathcal{Y}$ from the training set $\mathcal{D} = \{\boldsymbol{x}_i, y_j \| 1 \leq i \leq n \text{ and } 1 \leq j \leq k\}$, where $\boldsymbol{x}_i \in \boldsymbol{\mathcal{X}}$ is a feature vector of the instance i^{th} and $y_j \subset \mathcal{Y}$ is the label of $\boldsymbol{x}_i$. In this study, we consider three classes ($k = 3$): small, medium, and large which can represent a wide range of Vietnamese woman.

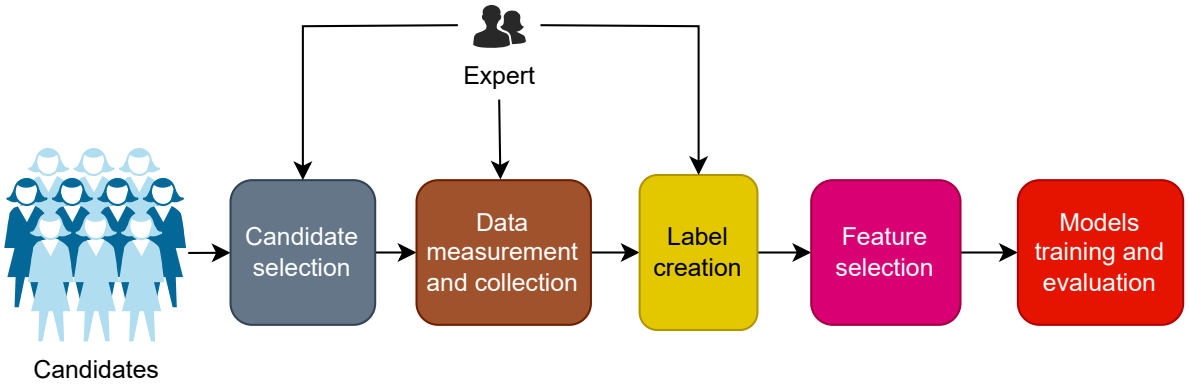


Figure 1: The pipeline of the prediction.

3.2 Data Preparation

3.2.1 Candidate selection

The first step of building AI models for bra classification is to collect data. To do that, 460 female students from 3 universities in northern Vietnam are selected for collecting training samples. The students are in the ages of 18-24 with normal health, body mass index from 14.5-24.3 kg/m^2. Female students volunteer to participate in the study and are guaranteed that they have not breast surgery, not pregnant, and not during their menstrual period. We ensure to only collect measurement metrics on these students without using any personal information. The minimum number of samples is determined according to the following formula.

$$n_{0.01} = \frac{t^2 \times \delta^2}{m^2} = \frac{2.3264^2 \times 4.48^2}{0.5^2} = 434.047 \quad (1)$$

where n is the sample size, t is the reliability (determined with p = 0.99; t = 2.3264), m is the error (m = 0.5 cm), σ is the standard deviation (σ = 4.48 cm). m and σ are determined through a preliminary survey of the chest sizes of 30 female students [CITE]. The actual number of instances is 460 which is larger than the calculated instance size in Eq. (1) to ensure the quality of the prediction and eliminate the randomness when training classifiers.

3.2.2 Data measurement and collection

Once the candidates have been selected, we process to collect necessary information for training AI models. Based on the careful discussion with domain experts, 21 measurements are identified. They include three body parameters: height, weight, and BMI and 18 breast sizes. The three body measurements can be measured directly and 18 breast sizes can be estimated by using the Geomagic Design X software (footnote) on scanning data obtained from the MB Scan 3D device with marked measurement landmarks. The Scan3D MB2019 device (footnote or citation) has been set up to measure breast shapes and sizes to meet the goal of low cost, fast measurement speed, high accuracy, and no inconvenience when measuring. The 3D measurement dimensions of the female chest are compared with the contact measurement results. The results show that the Scan3D MB2019 system can be used to scan 3D chest data.

The collection is done as follows. The students stand on a turntable place in a measuring chamber in a position consistent with ISO 20685 standards, in which the distance of the device from the center of the projector lens to the turntable axis is 850 mm. Need a figure to show the measurement. Breast dimensions determined by non-contact 3D measurement are compared with contact measurement results on the same subject. Standing posture and anthropometric landmarks on the female student's chest are used for both contact and non-contact 3D measurements. The dimensions selected for measurement and comparison are the chest circumference group, the chest curve group, and the distance between points. 3D measuring equipment is calibrated before measuring. The calibration method is tested on mannequins and 12 female students with 12 sizes, each size measured 3 times. This number of dimensions is enough to ensure the ANOVA analysis [CITE]. The results of 20 contact and 20 non-contact 3D measurements for mannequin chest dimensions are compared and analyzed by ANOVA to evaluate the differences (paper 1).

To determine the chest measurements, it is necessary to establish the measurement landmarks. The body measurement landmarks are marked on human skin during the measurement process. The accuracy of the measurement results depends largely on the accuracy of the landmarks. In this study, landmarks are marked in contact due to the difficulty in determining some points on the breast. Use concentric circles made of black decals with a thickness of 0.1 mm and a diameter of 5 mm to stick at locations on the chest. The measurements of each student is carefully stored for later processing steps. Table 1 shows 21 measurements of each student.

Table 1: Determine bust size using non-contact 3D measurement and contact measurement.

Size (Symbol)	**Method for determining size**
Height (h)	The subject's standing height was measured with a straight ruler in a standard standing position ISO 20685, measure the distance from the standing floor to the highest point at the top of the head, inside midsagittal plane when the head is held in the Frankfurt plane (which is the plane passing through the rim upper edge of the external auditory canal and lower edge of the orbital rim, perpendicular to the longitudinal plane the vertical dimension contains the axis of the body).
Weight (w)	The weight of the measured object is determined by an electronic scale with a measuring range of 0÷100 kg, error ± 0.01 kg.
Mass index (bmi)	BMI is calculated from height and weight measurement results according to the formula [121]: $bmi = w/h^{(2.15)}$. In there , *w* is weight (kg); *h* is height (m).
Upper bust circumference (vnt)	On the subject's 3D body scan data, the size of the chest circumference (L1) is determined by creating a plane parallel to the body axis cutting through the body at a horizontal position armpit. The contour circumference surrounding the outside of the stem and lying on the cutting plane is circumference on the chest of the body.
Chest circumference (vn)	Chest size (L2) is determined by creating a plane parallel to the axis the body cuts through the torso at the fullest point of the chest. The border perimeter encloses the points outside of the torso and lying on the cutting plane is the body's chest girth.
Chest anklet (vcn)	The size of the chest circumference (L3) is determined by the circumference of the cross section by plane perpendicular to the body axis through the lowest point of the chest.
Upper chest prolapse (snt)	Upper chest prolapse (D1) is calculated by the distance from the nipple point (P4) to the plane cut through the point on the chest (P1).
Right lower chest prolapse (sndp)	Right lower chest prolapse (D2) is the distance from the right nipple point (P4) to the cutting plane across the lowest point of the chest.
Left lower chest prolapse (sndt)	Left lower chest ptosis (D2') is the distance from the left chest tip point (P4') to the cutting plane across the lowest point of the chest.
Distance of 2 points nipple (cn)	The distance between the two nipples is determined by measuring the distance from the nipple point In the case of two nipple points no lying on the same horizontal plane, the distance from the chest is calculated as the total distance from the point of the chest tip (P4), (P4') to the vertical line between the front body.
Thoracic arch in right (ccnp)	Right breast arch determined by the curve from the innermost point of the right breast (P6) through point P7 to the outermost point of the left breast (P5).
Left pectoral arch(ccnt)	The left pectoral arch is determined similarly to the right pectoral arch but on the left.
External thoracic arch right (cnnp)	Right external chest arch is calculated from the midpoint of the breast (P5) to the point of the right nipple (P4), measure along the chest line.
External thoracic arch left (cnnt)	The left external chest arc is calculated from the midpoint of the breast (P5') to the point of the left nipple (P4'), measured along the chest line.
Internal thoracic arch right (cntp)	The right internal chest arc is calculated from point (P6) to the right nipple point (P4) according to the chest arc line.
Internal thoracic arch left (cntt)	The left internal chest arc is calculated from the point (P6') to the left nipple point (P4') along the thoracic arc line.
Distance from sternum to tip of right chest (xup)	Distance from sternum to right nipple measured from the midpoint of the front neck (P1) to the point right nipple (P4)
Distance from the sternum to the tip of the left chest (xut)	The distance from the sternum to the left nipple is measured from the middle of the front neck (P1) to the pointleft nipple (P4').
Difference size (cl)	Equal to the difference between bust circumference and bust circumference.
Volume left breast (ttt)	The volume of the left breast of the body (ttt) is performed by marking the breast based on landmarks around the chest. Then the breast is separated from the body, filling and calculating the volume of the left breast.
Volume of right breast (ttp)	The volume of the right breast of the body (ttp) is performed by marking the breast breast based on landmarks around the chest. Then the breast is separated from the body, filling and calculating the volume of the right breast.

3.2.3 Label creation

The label creation of collected instances is done in two steps: data pre-processing and label creation.

Data pre-processing The data pre-processing process is to ensure the quality of collected instances before doing data annotation and training AI models. The pre-processing step consists of two smaller steps: duplication removal and measurement checking. For the first step, we carefully check all collected instances to avoid any duplication. If two instances are similar in terms of measurements, we only keep one. The next step is to check each measurement to avoid missing values. Any missing values are corrected by domain experts with a careful investigation using stored

measurements of each subject in Section 3.2.2. Finally, the processing step yields 460 samples, in which each sample has 21 completed measurements used to later steps.

Table 2: Statistical characteristics of small sizes.

Statistic	Max	Min	Mean	Median	Coefficient of variation	Standard deviation	95% Confidence Interval
ttp	648.50	221.20	388.23	368.60	287.50	77.01	(376.54; 399.92)
cl	18.20	5.30	11.46	11.80	11.00	2.27	(11.12; 11.81)
cnnp	13.50	7.50	10.55	10.50	10.50	1.26	(10.36; 10.74)
vn	84.50	72.50	78.66	78.50	78.00	2.73	(78.24; 79.07)
vtn	86.40	71.20	77.04	77.00	78.00	2.97	(76.59; 77.50)
cnnt	13.00	7.30	10.11	10.30	11.20	1.21	(9.93; 10.30)
cntp	10.70	6.40	8.57	8.50	8.50	0.80	(8.45; 8.69)
vcn	73.6	61.20	67.20	67.00	68.00	2.65	(66.80; 67.60)

Table 3: Statistical characteristics of medium sizes.

Statistic	Max	Min	Mean	Median	Coefficient of variation	Standard deviation	95% Confidence Interval
ttp	648.50	221.20	388.23	368.60	287.50	77.01	(376.54; 399.92)
cl	18.20	5.30	11.46	11.80	11.00	2.27	(11.12; 11.81)
cnnp	13.50	7.50	10.55	10.50	10.50	1.26	(10.36; 10.74)
vn	84.50	72.50	78.66	78.50	78.00	2.73	(78.24; 79.07)
vtn	86.40	71.20	77.04	77.00	78.00	2.97	(76.59; 77.50)
cnnt	13.00	7.30	10.11	10.30	11.20	1.21	(9.93; 10.30)
cntp	10.70	6.40	8.57	8.50	8.50	0.80	(8.45; 8.69)
vcn	73.60	61.20	67.20	67.00	68.00	2.65	(66.80; 67.60)

Table 4: Statistical characteristics of large sizes.

Statistic	Max	Min	Mean	Median	Coefficient of variation	Standard deviation	95% Confidence Interval
ttp	705.80	325.40	502.98	521.30	435.20	73.18	(491.17; 145.79)
cl	18.40	9.00	13.61	13.60	13.00	1.58	(13.36; 13.87)
cnnp	15.40	9.40	12.47	12.60	13.50	1.15	(12.29; 12.66)
vn	92.00	80.00	86.49	86.50	86.00	2.58	(86.07; 86.91)
vtn	90.60	74.00	83.00	83.20	83.50	3.17	(82.49; 83.51)
cnnt	15.30	8.50	11.90	12.10	12.50	1.11	(11.72; 12.08)
cntp	12.60	8.10	9.89	9.90	10.50	1.05	(9.72; 10.06)
vcn	78.50	66.50	72.88	73.10	75.00	2.44	(72.48; 73.27)

Label creation Based on the description of the data dimensions table, we can categorize the data into three types: Small, Medium, and Large. This data partitioning process is performed based on 8 features and their corresponding labels. First, for the Small data type, we filter each feature sequentially by value range from largest to smallest as described in the table. Then, we combine the feature tables based on the "stt" column. Tables that do not match the "stt" column will be discarded (using the merge function). Once we have the table for Small data, we remove the samples belonging to the Small type from the original data and continue the same process for the remaining two data types. Finally, we will have three tables corresponding to each data type. However, this process still does not handle correctly with other descriptions such as mean, median, coefficient of variation, and standard deviation, as well as the number of samples in each label. According to the description, the number of samples in the Small label is 141, Medium is 169, and Large is 150. To address this issue, we will create a dictionary containing the correct mean values for each label description. Then, we will randomly select samples so that the number of samples matches the description and calculate their mean. I use a threshold of 0.1 to select samples with means closest to the description. After experimenting and observing multiple times, and adjusting the mean, median, coefficient of variation, and standard deviation values, we will have a dataset divided closely to the description and with the correct number of samples in each label.

Table 5: An example of data samples in three classes. The meaning of measurements are mentioned in Table 1.

h	w	bmi	vtn	vn	vcn	cn	cnnp	cnnt	cntp	cntt	ccnp	ccnt	snt	sndp	sndt	xup	xut	cl	ttp	ttt	label
158.5	44.0	17.5	75.4	81.1	74.8	14.5	13.8	14.4	8.6	8.4	21.5	21.2	8.9	6.5	6.5	21.8	21.1	6.3	325.1	335.7	Small
163.0	43.0	16.2	76.0	79.0	64.0	16.5	13.1	12.6	9.2	9.1	19.8	18.2	8.4	3.5	3.7	21.0	20.5	15.0	521.6	513.5	Medium
152.0	46.0	19.9	77.5	85.5	70.4	18.9	13.5	12.5	10.3	10.5	20.4	20.1	8.5	5.5	4.2	19.5	20.5	15.1	625.8	585.4	Large

3.2.4 Data splitting

After doing some initial data processing, we perform data separation. The data is split with 80% of the data for the training set, used to train the model, and the remaining 20% of the data for the test set, used to fine-tune parameters

and evaluate the model. In addition, we also ensure that there are 3 classes in the test data set to be able to evaluate in the most intuitive way. In addition, we also standardize the data, bringing this data to the same scale, in order to bring characteristic values from many different scales to a single scale, with the same scale and scale. value range. This helps improve model performance, reducing the influence of magnitude and differences between data columns. Outliers are data points whose values are too large or too small compared to the rest of the data set. They can be the result of measurement error, noise, or the manifestation of significant differences in the data set. Removing outliers improves prediction and minimizes the negative impact of large or small values on the machine learning model building process. We use a method called IQR Method (Interquartile Range) to remove these data points considered outliers. This method uses the interquartile range to identify outliers. Points outside this range are considered outliers and eliminated. Finally we collected the following data table:

3.3 Feature Selection

Feature selection plays an critical role for training classification models. Good features can represent well training samples while bad features can affect the quality of classifiers. The feature selection is done in three settings: using all measurements, using eight important measurements, and using selected measurements automatically selected by machine learning methods.

Using all measurements The first setting uses all 21 measurements in Table 1 as 21 features.

Eight important features As mentioned, 21 features in Table 1 can provide good measurements to distinguish samples in three classes. However, there exist noisy features that may reduce the performance of classifiers. To improve the quality of classifiers, we use eight well selected features from [123].

Extract typical breast size: Feature selection/feature extraction is very important in data exploration. Caret Packages and Random Forests in R are applied to extract attributes on collected data about female student chest sizes. First, the data were explored using principal component analysis PCA. Then, the process of extracting characteristic breast sizes is carried out through 3 stages: eliminating redundant sizes in the data set, ranking breast sizes according to importance and finally selecting the sizes. measure in the data set. Ranking breast sizes in the data set according to the importance of the sizes is also done using the LVQ (Learning Vector Quantization) method. To select dimensions in the chest data set using this method, the RFE (Recursive Feature Elimination) algorithm is applied according to the steps [123]:

- Import and initialize data set F.
- Select classifier C.
- Calculate the weight of each breast size in dataset F based on the accuracy of classification prediction.
- Delete the minimum weighted chest size of fj and update F.
- Repeat steps 3 and 4 until F only has one bust size left.
- Rank the importance of breast sizes

The most important breast sizes for data grouping were selected based on the results of this ranking.

Selected feature using machine learning The final setting uses selected features from machine learning methods. To decide which features to choose for this dataset, we decided to train a classifier from the Random Forest model Liu et al. [2012] with the above dataset. We then list the importance of the model based on the trained weights and proceed to extract features with important feature values greater than 5 percent (0.05). There are 6 features in total of 21 features from table 1 that satisfy the above conditions including: Upper bust circumference, Chest circumference, Chest anklet, Difference size, Volume of right breast and Volume left breast. These are the features selected after the feature selection process ends.

3.4 Data Augmentation

We conduct the data augmentation process to enrich the training data source. All data sets will be applied data scaling to ensure increased convergence speed when training the model, reducing the influence of outliers as well as features with large value ranges. However, the distribution of labels in the data set is not balanced, we perform label balancing with SMOTE cha [2002]. Immediately after the label balancing process is completed, we conduct data augmentation to generate more training data. By splitting the data into classes and using this data to train an unsupervised Gaussian Mixture model McLachlan et al. [2019] with a data generation rate of 1.5 times compared to the original data, we will get the following data generated based on data provided from the classes. These data will be labeled with the label

value of the input data of the data generation model. Performing this process for 3 labels and merging them with the original data and removing outliers , we will get the data set after data augmentation process.

3.5 Classification Models

Once the data has been formed and the labels have been created, we train classifiers to predict the bra sizes. To do that, we utilize the following machine learning methods: Logistic Regression, Support Vector Machine, Random Forest, XGBoost, and a simple neural network.

Logistic Regression Logistic Regression cox [1958] is a popular machine learning model applied in binary classification problems. This model belongs to the category of linear models and is designed to predict the probability of belonging to a particular class. The Logistic Regression model uses the sigmoid function to convert the output to a value between 0 and 1. The sigmoid function is represented by the formula:

$$g(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

where z is a linear combination of input variables and corresponding weights.

Logistic Regression gives the predicted probability of a sample belonging to a particular class. If the probability is greater than a threshold (usually 0.5), the sample belongs to class 1, otherwise it is class 0. Logistic Regression describes the relationship between the independent variable and the probability of belonging to a particular class as follows.

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)}} \tag{3}$$

where Y is the dependent variable (class), X is the independent variable vector, and β denotes the coefficients learned from the training process. Logistic Regression is trained through the process of optimizing a loss function, usually log-likelihood or entropy loss, using methods such as gradient descent. The coefficients β are adjusted to optimize the model's predictive ability.

Support Vector Machine Support Vector Machine (SVM) Cortes and Vapnik [1995] is a powerful machine learning model widely used in many applications, especially in classification and regression problems. SVM is especially suitable for data sets with the nonlinear nature, and it is capable of handling both binary and multi-class classification problems. SVM focuses on finding the best hyperplane to divide data points into different classes. The best hyperplane is the hyperplane with the largest distance from the data points closest to it, also known as support vectors. SVM also works well in the high-dimensional space and can use kernel functions to map data points into an advanced space, helping to create complex hyperplanes. Given a training data set $(X, Y) = \{(x_1, y_1), (x_2, y_2), ...(x_n, y_n)\}$ where x_i is the feature vector and y_i is the label belonging to class -1 or 1, the goal is to find a hyperplane of the form $w \cdot X + b = 0$ that divides the data points to maximize the distance from every point to the hyperplane.

$$L(w, b) = \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i \cdot (w \cdot x_i + b)) + \lambda \|w\|^2 \tag{4}$$

where λ is the regularization parameter to control the complexity of the model. The goal of the loss functino is to optimize w and b by minimizing the loss function. This can be done through optimization algorithms like Gradient Descent or variations like Stochastic Gradient Descent (SGD). Constraints are applied to ensure that each data point lies on the correct side of its correct hyperplane.

$$y_i \cdot (w \cdot X_i + b) \geq 1 \quad \text{for all } i \tag{5}$$

Karush-Kuhn-Tucker (KKT) Gordon and Tibshirani conditions are important in SVM optimization and include conditions on partial derivatives, constraints, and the magnitude of slack variables. Besides the binary classification problem, SVM can be extended to the multi-class classification problem through methods such as One-vs-One (OvO) or One-vs-All (OvA), depending on the requirements. specific needs of the problem. With nonlinear data, SVM processes using kernel functions such as Polynomial Kernel, Radial Basis Function (RBF) Kernel, or Sigmoid Kernel. The kernel maps data from the original feature space to the enhanced space, making finding more flexible hyperplanes.

Random Forest Random Forests Liu et al. [2012] is an important supervised machine learning algorithm. It can be used for both classification and regression problems. The idea of Random Forests is to create decision trees on randomly selected data samples, get predictions from each tree, and select the best solution by voting. The advantage of Random Forest is that it is less likely to have overfitting issues, but the disadvantage is the speed due to the use of many decision trees like in Figure 2. Random Forest is implemented over 6 steps as follows.

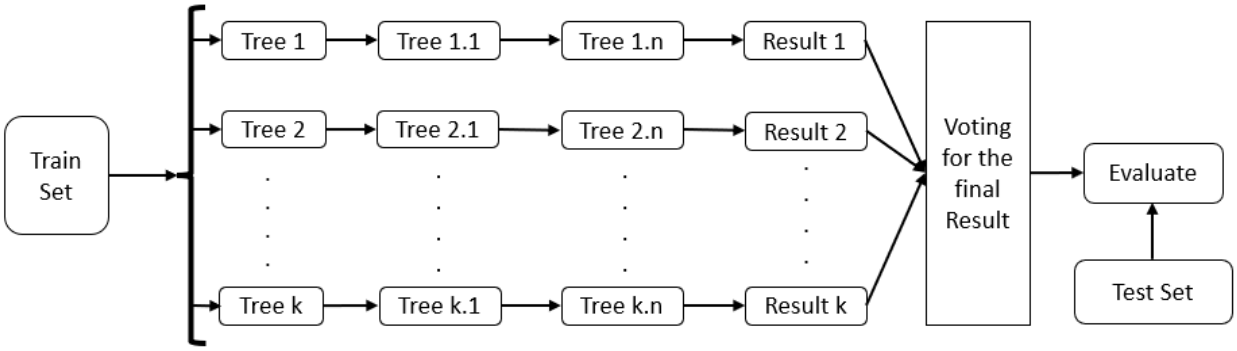


Figure 2: The Random Forest model.

Step 1: Data splitting D into trainning and testing sets

Step 2: The desired size of the ensemble (number of trees in the random forest) is set as B, and a method for measuring disorder, like Gini or Entropy, is chosen.

Step 3: Create B bootstrap samples Abney [2002] for trainning set.

Step 4: For each bootstrap sample b=1...B

a. Build modified decision tree $\mathbf{T}_b$ on bootstrap b recursively performing the following steps until the data can no longer be split (all samples at the current node have the same label).

1. A subset of features M_{sub} is randomly selected from the dataset M. Typically $M_{sub} = \sqrt{M}$
2. A feature f^* and a corresponding threshold x_f for f^* value are choosen from M_{sub} . Choose $\{f^*, x_{f^*}\}$ values to minimize the chosen disorder measure.
3. Divide data at current node $\mathbf{D}_{node}$ based on $\{f^*, x_{f^*}\}$ into left and right:
   A. $\mathbf{D}_{left} = \mathbf{D}_{node} | \mathbf{x}_{f'} \leq x_f$.
   B. $\mathbf{D}_{right} = \mathbf{D}_{node} | \mathbf{x}_{f^* x_f}$. Transfer $\mathbf{D}_{left}$ and $\mathbf{D}_{right}$ to children node. After, saving decision tree is created T_b.

Step 5: Returns the trained combinator of the tree $T_{1...B}$.

Step 6: Predict on the test data set, pass the data for each tree in the set $T_{1...B}$ and create B prediction sets. Combine the B set of predictions either through majority voting (classification) or by calculating the mean/median (regression) to arrive at the final result.

XGBoost XGBoost Chen et al. [2015] (eXtreme Gradient Boosting), is a machine learning algorithm of the Gradient Boosting type, similar to Random Forest. XGBoost is designed to optimize performance and speed, helping to efficiently process large data sets while also allowing the use of GPUs for faster calculations. XGBoost has shown promising results in many classification problems CITATION and XGBoost also has strong customization capabilities with many parameters and can handle missing values, helping to process errors. The disadvantage of XGBoost is that it has high computational complexity and is susceptible to overfitting, which happens when trained on small datasets or with too many trees. The XGBoost algorithm is described in the steps below.

Input: Training data set $\{(x_i, y_i)\}_{i=1}^N$, loss function $L(y, F(x))$, weak learners $\boldsymbol{M}$ and the learning rate α.

1. Initialize the initial model with a constant constant value

$$\hat{f}_{(0)}(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i, \theta). \tag{6}$$

2. Iterates through values from m=1 to $\boldsymbol{M}$

a. Calculate gradients and hessians Van Den Bos [1994]

$$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x) - \hat{f}_{(n-1)}(x)}. \tag{7}$$

$$\hat{h}_m(x_i) = \left[\frac{\partial^2 \mathcal{L}(y_i, f(x_i))}{partial f(x_i)^2} \right]_{f(x) = \hat{f}_{(m-1)}(x)}. \tag{8}$$

b. Take a base learner (or weak learner, for example decision tree, etc.) using the training data set $\left\{x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right\}_{i=1}^{N}$ by solving the optimization problem

$$\hat{\phi}_m = \arg\min \sum_{i=1}^{N} \frac{1}{2}\hat{h}_m(x_i)\left[\phi(x_i) - \frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right]^2. \tag{9}$$

$$\hat{f}_m(x) = \alpha\hat{\phi}_m(x). \tag{10}$$

c. Update model parameters:

$$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_m(x). \tag{11}$$

3. Then calculate the output of the model $\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x)$.

Neural networks Neural Networks are computational models that mimic the complex functions of the human brain. The neural networks consists of interconnected nodes or neurons that process and learn from data, enabling task such as pattern recognition and decision making in machine learning. In supervised learning, the neural network is guided by a teacher who has access to both input-output pairs. The network creates outputs based on inputs without taking into account the surroundings. By comparing these outputs to the teacher-known desired outputs, an error signal is generated. In order to reduce errors, the network's parameters are changed iteratively and stop when performance is at an acceptable level.

4 Setting and Evaluation Metrics

4.1 Setting

All classifiers were trained by using k-fold cross-validation ($k = 5$) in two settings: original data and data augmentation. Before training, we scaled the data samples to ensure the minimal influence of variables with large amplitude values and speed up the convergence of the training process. For data augmentation, we scaled the values of samples and balanced the labels in the data set using the SMOTE method cha [2002]. The data augmentation process was done with a ratio of 1.5 compared to the original data on each classes using Gaussian Mixture Walker and Duncan [1967]. Outliners were also removed to ensure the high-quality of the augmented dataset.

The 5 model hyperparameters used for training are listed in the following table 6:

Table 6: Models hyperparameter settings

Models	Number of features	Hyperparamter settings
Logistic Regression	6 features	C=0.001, class-weight='balanced', penalty='l2', solver='newton-cholesky'
SVM		C=50, γ=0.01, kernel='linear', probability=True, class-weight='balanced'
Random Forest		criterion='entropy', max-depth=6, min-samples-leaf=3, min-samples-split=5
XGBoost		n-estimator=50, min-child-weight=5, max-depth=5, lr=0.02, γ=0.2
ANN		4x Dense(512), 2x DropOut(0.2), activation='relu', optimizer='adam', loss='categorical-crossentropy', epoch=10, batchsize=32
Logistic Regression	8 features	C=0.2, class-weight='balanced', penalty='l2', solver='newton-cholesky'
SVM		C=1, γ=0.09, kernel='rbf', probability=True, class-weight='balanced'
Random Forest		criterion='entropy', max-depth=7, min-samples-leaf=4, min-samples-split=5
XGBoost		n-estimator=70, min-child-weight=6, max-depth=5, lr=0.03, γ=0.2
ANN		4x Dense(512), 2x DropOut(0.2), activation='relu', optimizer='adam', loss='categorical-crossentropy', epoch=10, batchsize=32
Logistic Regression	21 features	C=0.1, class-weight='balanced', penalty='l2', solver='newton-cholesky'
SVM		C=0.09, γ=0.01, kernel='rbf', probability=True, class-weight='balanced'
Random Forest		criterion='entropy', max-depth=6, min-samples-leaf=3, min-samples-split=5
XGBoost		n-estimator=50, min-child-weight=3, max-depth=3, lr=0.02, γ=0.2
ANN		4x Dense(512), 2x DropOut(0.2), activation='relu', optimizer='adam', loss='categorical-crossentropy', epoch=10, batchsize=32

4.2 Evaluation Metrics

Accuracy Accuracy is used to evaluate the performance of a classification model. It is calculated by dividing the number of correct predictions by the total number of predictions.

$$\text{Accuracy} = \frac{\text{Number of corredictions}}{\text{Total number of predictions}} \tag{12}$$

Using accuracy is an initial step that should be used when evaluating a machine learning model, but it cannot assess the entire model. To achieve the best performance, it needs to be combined with other evaluation methods. For binary classification, accuracy can also be calculated using TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

Precision In addition Accuracy, precision is also used to evaluate the performance of a classification model. Precision is defined as the ratio of true positive points among those classified as the total of false positives and true negatives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

High precision means that the accuracy of the found points is high. High recall means that the True Positive Rate is high, which implies that the rate of missing actual positive points is low. This is very important in machine learning problems. For example, in a problem of detecting spam emails, precision is important because if an email is predicted as spam, users may lose trust if the email is actually important. Therefore, it is necessary to optimize precision to minimize the number of true negatives.

Recall Recall, also known as sensitivity or true positive rate, is an important measure in classification problems, focusing on the model's ability to detect all samples of a specific class. Recall is defined as the ratio of the number of true positive predictions (True Positives) to the total number of actual positive points (True Positives + False Negatives). Recall is determined by the formula:

$$\text{Recall} = \frac{\text{T P}}{\text{T P + F N}} \tag{15}$$

Recall is important in problems where missing positive samples (False Negatives) is undesirable and can have significant consequences. For example, in the classification of cancer patients, missing a patient can lead to severe consequences. In some situations, recall can be considered the main objective, and the model can be adjusted to ensure it has the ability to detect all positive samples. For a classification model, evaluating recall is crucial to ensure sensitivity in detecting positive class samples.

F1-Score The F1-score is a performance evaluation metric for classification models, combining precision and recall. It is often used in situations where both precision and recall are important and need to be balanced, calculated using the formula:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{16}$$

The F1-score can help evaluate the balance between precision and recall. If a model has high precision but low recall (or vice versa), the F1-score will reflect this imbalance.

ROC curve A ROC Curve (Receiver Operating Characteristic Curve) is a graph showing the performance of a classification model at all classification thresholds. ROC shows two parameters: TPR (True Positive Rate) and FPR (False Positive Rate). This is computed though TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative). True Positive Rate(TPR) is a synonym for recall and defined as follows.

$$\text{TPR} = \frac{\text{TP}}{\text{TP + FN}} \tag{17}$$

Negative is defined as follows.

$$\text{FPR} = \frac{\text{FP}}{\text{FP + TN}} \tag{18}$$

Besides, AUC (Area Under the Curve) is used to represent the level of classification. AUC has a threshold from (0,0) to (1,1). A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. AUC is defined as follows

$$\text{AUC} = \text{P(score(x+))} > \text{P(score(x-))} \tag{19}$$

5 Results and Discussion

This section reports the comparison of classification with five classifiers in two settings: using original data and using data augmentation.

Performance with original data Table 7 shows the performance of the five classifiers when using the original data in three feature groups. The general trend indicates that Random Forest, XGBoost, and ANN are strong methods for bra classification. A possible reason is that Random Forest and XGBoost are quite strong methods that use ensemble learning for classification. In the setting of small data samples, ensemble learning methods have shown its efficiency. In addition, ANN uses feed-forward networks that also work well for various classification problems. Logistic regression and SVM are behind the three strong methods.

Table 7: Results with original data. **Bold** values are the best, underline values are second best.

Model	Number of features	Accuracy	Precision	Recall	F1-score
Logistic Regression	6 features	73.53	74.11	73.53	73.08
SVM		73.31	74.51	73.31	73.13
Random Forest		72.66	74.41	72.66	71.67
XGBoost		73.52	**76.54**	73.52	72.15
ANN		**74.65**	76.07	**74.65**	**74.19**
Logistic Regression	8 features	74.84	75.52	74.84	74.79
SVM		73.98	75.09	73.98	73.48
Random Forest		74.18	76.47	74.18	73.42
XGBoost		**81.62**	78.06	**81.62**	76.40
ANN		77.50	**78.38**	77.50	**77.49**
Logistic Regression	21 features	74.40	75.32	74.40	74.40
SVM		75.72	78.26	75.72	75.34
Random Forest		76.60	79.01	75.40	77.02
XGBoost		76.59	79.09	76.59	75.90
ANN		**84.75**	**85.42**	**84.75**	**84.39**

The scores in Table 7 also show that the performance of classifiers increases when using more features, in which Random Forest and ANN with 21 features achieve the best results. This is because the number of training samples is quite small (460), so using a small set of features seems to limit the performance of the classifiers. On the other hand, increasing the number of features (21) can help to find good mapping functions that reflect the distribution of data. An interesting point is that the feature selection methods (automatic and human selection) are inefficient compared to using all 21 features. As discussed, the small number of samples may limit the efficiency of the feature selection methods.

Performance with data augmentation Table 8 reports the performance of the five classifiers with data augmentation. We can observe the significant reduction of performance compared to the scores in Table 7. This shows that the data augmentation methods seems to be inappropriate for this dataset by adding noise samples that affect the original data points. However, even with the significant reduction of performance, ANN still shows competitive scores followed by Random Forest. It again confirms the efficiency of ANN for this classification problem. An interesting point is that there are tiny differences among the three feature groups. A possible reason is that adding more augmented samples saturates the contribution of features. It us quite easy to understand that the SMOTE method cha [2002] creates new samples by using values of the original samples. Therefore, there are no differences among augmented samples that challenge the classifiers.

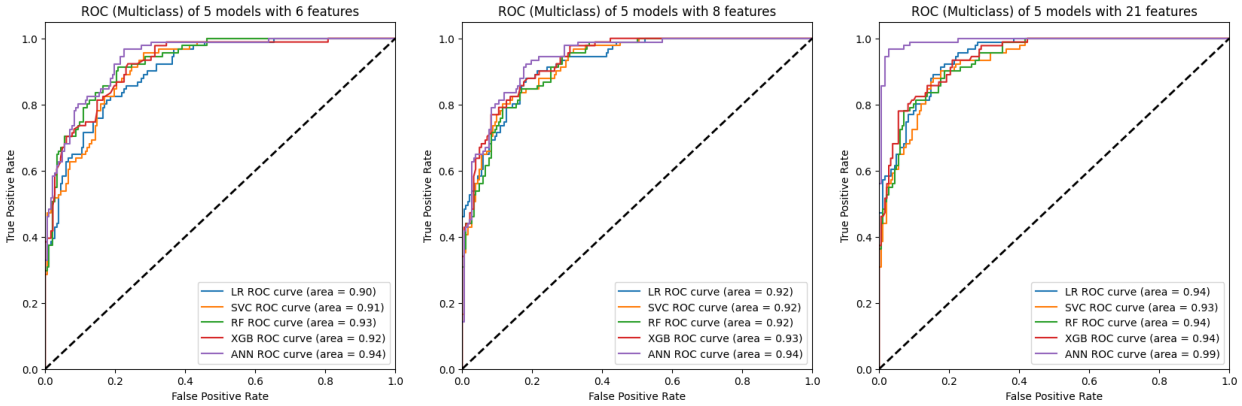
Performance with ROC curves The performance of classifiers were also investigated by observing the ROC curves. To do that, we observed the ROC scores of the five classifiers and then visualized the observation in Figure 3. Due to data augmentation is not so efficient, we only show the ROC curves by observing the original data.

The ROC curves are consistent with the performance reported in Table 7. The gap of performance between ANN and others classifiers is small when using six and eight features. However, the ROC curve in Figure 3(c) in which ANN with 21 features outputs better performance than others.

Feature contribution The contribution of features was observed for the classifiers. To do that, the weight of each feature was observed in each fold and the final weight was computed by the average over five folds. Figure 4 shows the contribution of features when training the five classifiers. We can observe that except the logistic regression, features positively contribute to classifiers. For example, only two features have negative weights for SVM. It is similar to XGBoost, in which there are two features that have no contribution. On the other hand, the weight of features are positive for Random Forest and ANN. It supports the results in Tables 7 and 8 in which Random Forest, XGBoost, and ANN output good performance.

Table 8: Results with data augmentation. **Bold** values are the best, underline values are second best.

Model	Number of features	Accuracy	Precision	Recall	F1-score
Logistic Regression	6 features	71.14	71.41	71.14	70.98
SVM		72.84	73.35	72.84	72.18
Random Forest		72.50	73.98	72.50	71.04
XGBoost		72.60	74.38	72.60	70.95
ANN		**75.38**	**76.94**	**75.38**	**74.03**
Logistic Regression	8 features	70.84	71.19	70.84	70.69
SVM		72.95	73.55	72.95	72.18
Random Forest		75.10	75.46	75.10	74.13
XGBoost		74.92	75.94	74.92	73.39
ANN		**76.98**	**77.13**	**76.98**	**76.19**
Logistic Regression	21 features	73.55	73.68	73.55	73.48
SVM		76.35	76.68	76.35	76.08
Random Forest		73.21	73.85	73.21	72.40
XGBoost		73.88	75.83	73.88	72.39
ANN		**76.53**	**77.31**	**76.53**	**76.27**



(a) ROC curves with 6 features. (b) ROC curves with 8 features. (c) ROC curves with 21 features.

Figure 3: Performance with ROC curves of classifiers.

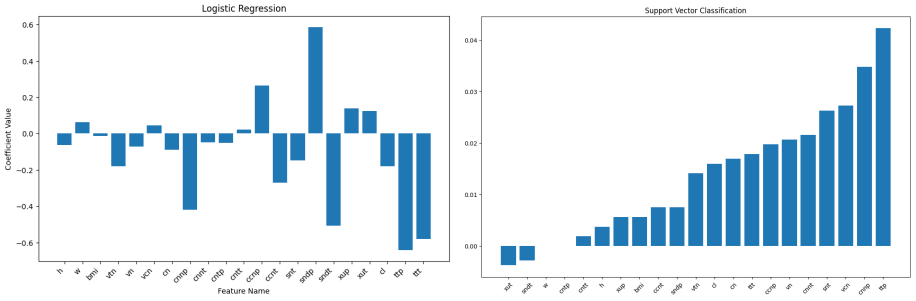
6 Demonstration

We have developed a demo system for breast classification for Vietnamese women. You also can access our demo system at: here. You can watch our user guideline video on YouTube here. The system takes women's breast measurements as inputs and returns the predicted bra sizes. We provide three testing samples that belong to 6, 8, and 21 measurements stated in Section 3.3. Figure 5 shows the interface of the demo system.

After accessing to the link, users can experience the system by selecting one of three testing samples provided in advance. Once the sample has been selected, the system shows features (measurements) of this sample on the web interface. Each text box represents the value of each feature. The users can also change the value of each feature. After that, the system returns the prediction of the selected sample when users click on the "Predict" button. Note that we selected the best model corresponding to each feature group in Table 7 for the deployment.

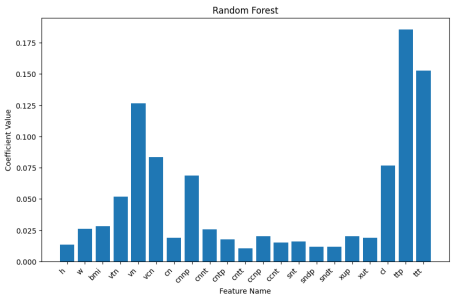
7 Conclusion

This paper introduces an investigation of automatically classifying the size of bras. To do that, we collect 460 samples measured by 21 measurements from female students in Northern Vietnam. The samples are used to trained five classifiers for the classification. Experimental results on the collected dataset show two important points. First, artificial neural networks achieve promising results compared to other strong classification methods such as SVM, Random Forest, or XGBoost. Second, data augmentation is inefficient to improve the performance of classifiers. The paper also

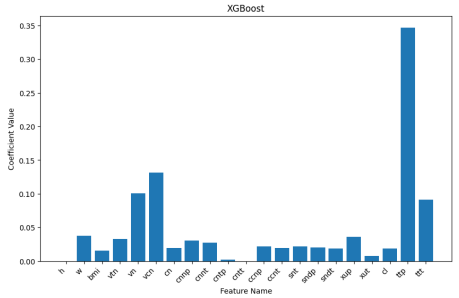


(a) Logistic Regression

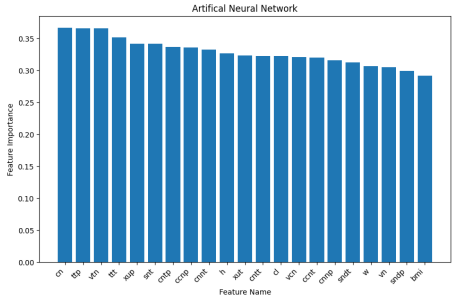
(b) SVM



(c) Random Forest



(d) XGBoost



(e) ANN

Figure 4: Feature contribution of classifiers.

provides a demo system where audiences can experience with the best model for the bra size classification. The system plays is an important step in the bra production process.

Future work will increase the number of training samples to improve the performance of classifiers. The system will be also deployed as a component in the complete pipeline of bra size production process by receiving measurements from machines for automation classification.

Acknowledgement

This work was supported by Hung Yen University of Technology and Education under the grant number UTEHY.2023.L....

References

Yanli Liu, Yourong Wang, and Jian Zhang. New machine learning algorithm: Random forest. In *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3*, pages 246–252. Springer, 2012.

Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

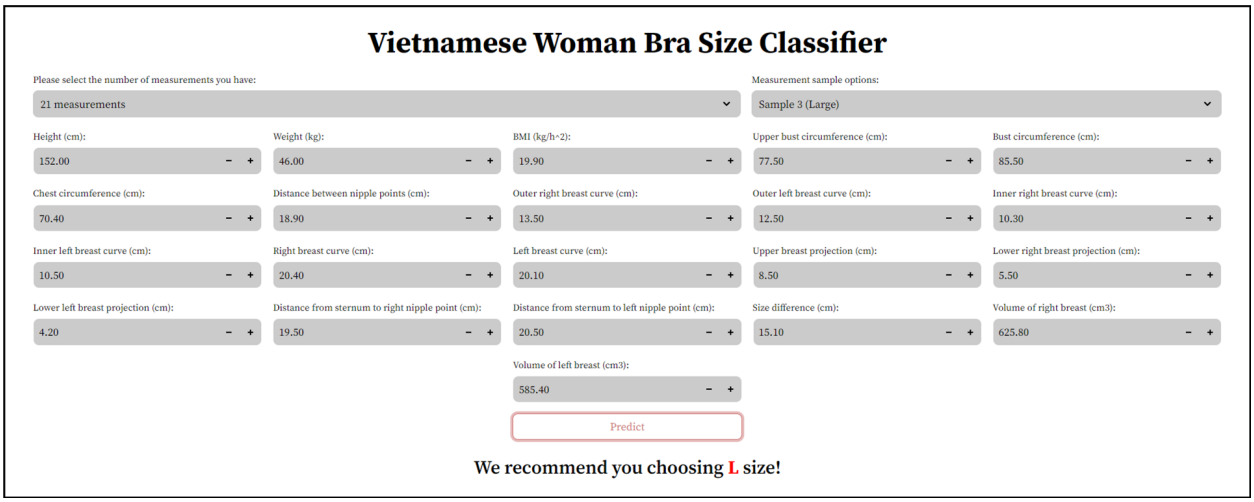


Figure 5: The demo system for Vietnamese woman bra size classification.

Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232, 1958.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

Geoff Gordon and Ryan Tibshirani. Karush-kuhn-tucker conditions. *Optimization*, 10(725/36):725.

Steven Abney. Bootstrapping. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 360–367, 2002.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

A Van Den Bos. Complex gradient and hessian. *IEE Proceedings-Vision, Image and Signal Processing*, 141(6):380–382, 1994.

Strother H Walker and David B Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179, 1967.