IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Breast Cancer–Detection System Using PCA, Multilayer Perceptron, Transfer Learning, and Support Vector Machine

HUAN-JUNG CHIU[1], TZUU-HSENG S. Li [1*], (Member, IEEE), PING-HUAN KUO[2*]

[1]aiRobots Laboratory, Department of Electrical Engineering, National Cheng Kung University, Tainan 70101, Taiwan.
[2]Department of Intelligent Robotics, National Pingtung University, Pingtung 90004, Taiwan.

Corresponding author: Ping-Huan Kuo (phkuo@mail.nptu.edu.tw) and Tzuu-Hseng S. Li (thsli@mail.ncku.edu.tw).

This work was supported by the Ministry of Science and Technology, Taiwan, under Grants MOST 106-2218-E-153-001-MY3 and MOST 106-2221-E-006-009-MY3.

ABSTRACT This study proposed a new processing method to predict breast cancer on the basis of nine individual attributes, including age, body mass index, glucose, insulin, and a homeostasis model assessment. First, principal component analysis (PCA) was used to identify valuable parts of the data and further reduce the dimensions of the data. The cumulative proportion of the top five major components was 99.89%. The multilayer perceptron network (MLP) method was then used to extract characteristics included in the data, and the structure of the network was designed for the exploration of how data developed as the dimensions increased or decreased. As such, the model was established to first explore (high dimensional) and then develop (low dimensional) data. After training and learning, the models could segregate the representative attributes and numbers, and the characteristic data were then used as classifiers through transfer learning techniques using support vector machines. To verify the proposed method, the experiment performed *k*-fold cross-validation 50 times on average. Experimental results verified the proposed method with 10-fold cross-validation using the dataset of Manuel Gomes from the University Hospital Centre of Coimbra, and an accuracy of 86.97% was achieved. The results indicate that the proposed series of processes and methods can effectively and powerfully examine the incidence of breast cancer. Furthermore, the data processed using only the PCA method as well as the characteristics extracted through the PCA method then combined with MLP after learning were analyzed. The differences displayed for the visual technique characteristics of the t-distributed stochastic neighbor embedding were compared.

INDEX TERMS multilayer perceptron network, principal component analysis, support vector machine, transfer learning.

I. INTRODUCTION

Globally, breast cancer is the most common malignant tumor among women [1] and is the second most common cause of death [2]. Breast cancer is cancer in the breast tissue, and its signs and symptoms include breast lumps, epidermal tissue dimples, shape changes, and red plaques appearing on the epidermis. If cancer spreads, it can cause osteocope, swollen lymph nodes, and dyspnea [3]. Early prediction and diagnosis of breast cancer can help medical personnel to provide appropriate treatments or relapse monitoring after surgery. Furthermore, they can control the treatment pain for patients, decrease mortality risks, and increase survival rates. Thus, the medical field seeks risk factors that may cause breast cancer as well as potentially influential relationships among risk factors, including direct or indirect relationships in numerous diagnosed clinical cases. The direct relationships of key risk factors with breast cancer incidence were sought to help with early prediction and prevention of breast cancer. Consequently, studies identifying risk factors of breast cancer include discussions of obesity, sex hormones, the endocrine system, and adipose tissue [1, 4, 5].

In studies from recent years, prediction systems based on machine learning have used data including risk factors such as X-ray images and heredity profiles, as well as various clinical data and learning algorithms for breast cancer prediction. Various types of studies have been conducted for the risk

prediction of breast cancer, such as mammographic studies [6-8], the discussion of hormones [9], genetic research [10-12], and studies based on images that have used the most popular deep learning method, as addressed in [13-15]. In the review of studies using incidence prediction for related risk factors, Artificial Intelligence (AI), or algorithms related to machine learning were used. The method of cancer gene mapping analysis was described in [16], where multimodal autoencoder (MAE) classifiers constructed using various risk factor data in breast cancer diagnosis were used to predict the survival rates of breast cancer prognosis. Abdikenov et al. [17] proposed a Pareto optimality-based prognostic model to understanding changes in hyper-parameters in various performance metrics, and [18] presented a wrapper method that embeds Bayesian classifiers for hybrid feature selection of breast cancer datasets. Whitney et al. [19] used the most popular integrated methods of deep convolutional neural networks (CNNs) and transfer learning to compare breast magnetic resonance Magnetic Resonance Imaging (MRI) tumor classification with traditional methods. The hybrid CNN-SVM model that was proposed in [20] used the MNIST numerical database that contains 70,000 examples for the experiment. Each example has $28 \times 28 = 784$ high-dimensional attributes. To reduce the number of dimensions and extract features, the experiment used a multilayer neural network with deep supervised learning. This concept entails the use of a large quantity of extremely high-dimensional data, which is the basic element of a neural network. The database and methods used in this study differ from the aforementioned databases and methods; the quantity of data in the database selected in this study was lower than that in the aforementioned databases. The aforementioned CNN method is not applicable when the quantity of data is limited. For the CNN method to be effectively used, a considerable quantity of data is required for training. This is why the database and method used in this study differ from those in other studies; here, an enhanced data exploration method that uses the multilayer perceptron network (MLP) model is proposed. The internal MLP model first increases the size and then reduces dimension to the original size.

In this article, a series of processes and methods are presented, and the novel and traditional algorithms were integrated. First, principle component analysis (PCA) was employed to find valuable data and reduce the data dimensions. The multilayer perceptron network (MLP) method was then used to extract the data characteristics through network learning included. After training and learning, the models could segregate representative attributes and numbers. MLP can process nonlinear separable problems. This is mainly attributed to the perceptron and its ability to achieve nonlinear classification through multilayer combination and the activation function. Models with complete training data can express their features' nonlinear characteristic and completely preserve their nonlinear expression through transfer learning techniques. Here, nonlinear expression refers to the distribution of an appropriate weight to each layer by using a weight tuning method that is specifically adjusted to training data, which are transferred to another model for use. Through the transfer learning method, the characteristic data determined through the MLP method can be transferred to become corresponding information for the SVM. The integrated techniques that combine the MLP with SVM can be applied effectively to examine breast cancer prediction approaches. In addition, the breast cancer dataset of the Faculty of Medicine of the University of Coimbra was used for *k*-fold cross-validation.

This study extended relevant medical studies that have examined lung nodules. An evaluation model based on the combination of a fuzzy system and neural networks was used for screening and testing through image input and self-defined fuzzy membership functions. Then, neural networks were employed to perform a final evaluation [21]. The regional reconstruction of computed tomography scan images was improved using a self-adaptive variation of the partial differential equation model [22]. Based on electrocardiogram signal diagnosis, neural network training provides effective solutions that can strengthen the generalizability of neural networks and maintain sensitivity and accuracy. This method can be applied in electrocardiogram examinations to identify certain abnormal heart activities that are difficult to detect [23].

The major contributions of this study are as follows: (1) The PCA-based scheme integrated MLP and SVM to construct the proposed learning algorithm to predict breast cancer; (2) low-dimensional feature extraction was applied to realize the network structure being designed for exploration first (high dimension) and development second (low dimension) in the MLP; (3) a complete blueprint of the breast cancer diagnosis medical system was constructed for the machine learning algorithm for breast cancer prediction, and the modules thereof were elaborated to provide the key technologies for the latest implementations; (4) the concept of transfer learning was used to construct a link between models and applied to small databases; and (5) the *k*-fold cross-validation was designed for the experiments combined with multiple randomized mechanisms with even distribution to test four breast cancer databases and verify the feasibility of the algorithm.

The remainder of this paper is organized as follows. Section II fully introduces the breast cancer diagnosis medical system. Section III proffers the proposed model. Section IV mainly explains the results of the *k*-fold cross-validation experiments. Section V presents the comparison researches and also discusses transfer learning and nonlinear methods. Section VI provides the conclusion of the study.

II. SYSTEM ARCHITECTURE

This system architecture provides a complete breast cancer diagnosis medical system. In Fig. 1, six major blocks are included, and the presented method is combined with the AI model block. Novel and traditional algorithms are integrated, and a human professional training experience method is adopted. The entire system encompasses the six major blocks: inquiry, diagnosis, data center, AI model, decision-making center, and confirmed diagnosis. In the inquiry block, inquiries from suspected patients are accepted, and basic information for the preliminary evaluation is the input of the entire system. In the diagnosis block, data are subjected to direct measurements, such as mammographic images and heart rhythms, and indirect measurements, such as blood and urine. The procedure obtains an appropriate amount of blood from the patient and uses an automatic blood cell analyzer.

A smart microscope [24] can be used to analyze various data that are commonly seen in routine blood analysis, including a report on 31 features in the entire blood and seven reportable parameters of blood [25]. A ZigBee radio frequency based [26] on the e-health Device Healthcare support platform and the Internet of things (IoT) are used in combination to [27] transmit data to the data center to process sensitive information. Regarding the accuracy requirement for various data that were obtained from the whole blood analysis, performance assessment documents from the blood analysis equipment vendor [28] indicated a high correlation between the consistency of the automatic method and that of the manual method. The Pearson correlation coefficient was greater than 0.9 for the 178 types of bodily fluid when strict biological sampling specifications were used. The inspection procedures intended for use by the medical laboratory complied with the Section 5.5 requirements of the international ISO 15189:2012 [29] standard. First, standards, guides, methods, or journal publication methods that have been verified or are recognized internationally must be used in the inspection procedure. The measurement data for the selected inspection procedure include true value, accuracy, precision (including gauge repeatability and intermediate measurement precision), measurement uncertainty, and analytical specificity (including disruptors, analysis sensitivity, detection limit, limit of quantitation, measurement interval, diagnostic specificity, and diagnostic sensitivity). "Measurement uncertainty" uses the coefficient of variation (CV) or standard deviation (SD) to indicate deviations in measurement. This is based on ISO [30] Guide to the Expression of Uncertainty in Measurement, which states that when completing a measurement uncertainty evaluation for a test item, the set acceptable standard (target uncertainty and the laboratory's quality goal or quality specification) and the scope of the allowable total error (TEa) must be reviewed to determine if they reach "target measurement uncertainty." According to [31], targetm easurement uncertainty can be interpreted as "the allowable error upper limit for test results

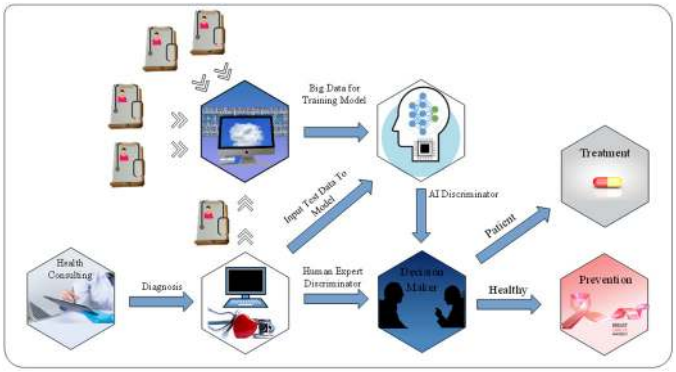


FIGURE 1. The architecture of the system.

used clinical diagnostics." The review acceptance standard is divided into three performance specifications that are based on individual biological variation (CVi). The desirable level of specification uses ≤1/2 CVi as the acceptance standard, along with a minimum level of ≤3/4 CVi or the optimal level of ≤1/4 CVi [32]. If the CVi cannot be obtained for the test item, ≤1/3 TEa is used as the acceptance standard [33]. The data obtained require subsequent analysis and interpretation. Next, the data are passed to the data center block, AI model block, and decision-making center. The data center stores copious diagnosis data to form a big data database. In the AI model block, big data provides training to complete the prediction models for breast cancer detection. The diagnosed data are then inputted with single cases to determine the morbidity probabilities. In the decision-making center, the predictions obtained with AI models and their diagnosis data are received by doctors with clinical experience for the final decision. In this stage, the AI model organized and analyzed patient data to provide valuable information recommendations to assist patient treatment or accurately predict possible locations of lesions at a preliminary stage.

The AI model plays an assistive role for physicians, who remain the actual decision makers. Based on judgments corresponding to personal medical literacy, the AI model provides a rapid and accurate approach to help physicians interpret information. The data-based learning features fast speeds and low costs and reduces misjudgments. In addition, physicians can apply inquiry approaches and their experience to treat patients efficiently. The confirmed diagnosis block contains treatments after morbidity and periodical prevention with no morbidity.

The method in the AI, model block is the processing method that was developed in this study to examine the prediction of breast cancer incidence. The following subsections describe the selected machine learning algorithms one by one.

A. PRINCIPAL COMPONENT ANALYSIS

PCA was originally proposed by Person in 1902 [34] and was defined and named after Hotelling [35], who developed it independently. PCA has different names in different fields. It is termed eigenvalue decomposition in linear algebra, singular value decomposition in matrices, and proper orthogonal decomposition in mechanical engineering.

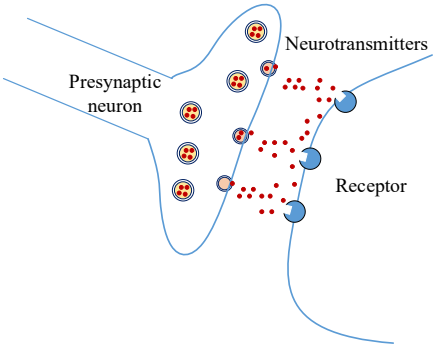


FIGURE 2. Synaptic Plasticity Effect of Neurons.

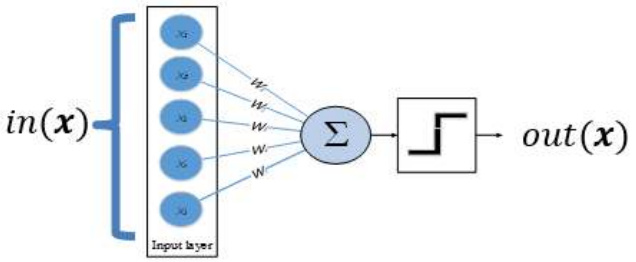


FIGURE 3. Artificial Neuron of Single Layer Perceptron.

The basic concept of PCA is the identification of a set of axes (vectors) that conveys the maximum quantity of information in the characteristic space of a multidimensional dataset. The multiple attributes (or multiple dimensions) of one piece of data are expressed as a linear combination of these basis vectors. Assuming that the vector space constituted by the selected axes (or vectors) has *N* basic components, for all data in that space, the multiple attributes of the data are expressed as linear combinations of these *N* components. The maximum amount of information signifies that the largest variances can be obtained after the dataset is projected to the examined axis. The actual meaning of the projection is the dot product calculation of the two vectors. One vector is the multiple attributes of one piece of data, and the other vector is the respective axis. Thus, the product is spanned by unit vectors that constitute basis vectors. The projected points are expressed in the linear subspace of the selected vectors. This means that the projection reduces the original dimension. All data are projected to that axis or are the most scattered in this direction; namely, the largest difference between the maximum and the minimum means that the axis in question represents the first major component.

B. MULTILAYER PERCEPTRON NETWORK

The well-known American neurologist Frank Rosenblatt [36, 37] proposed the perceptron, which is inspired by brain nerve cell adaptation theory [38], which is the cell assembly theory of synaptic plasticity [39], in the learning process. Synaptic plasticity [40, 41] describes the connection or synapse between cells in neuroscience. Its characteristic of adjustable connection strength is depicted in Fig. 2 [42] and Fig. 3.

In the original McCulloch & Pitts model (MCP) [38] of the artificial neurons model, a supervised learning method was introduced to allow the adjustment of artificial neurons to the correct weights from the training data. A binary perceptron also called a threshold function maps its input $\mathbf{x} \in R^d$ to an output $f(\mathbf{x})$. Where $\mathbf{x}=[x_1 \quad x_2 \quad \cdots \quad x_d]^{\mathrm{T}}$ are the different components of the input vector, and $\mathbf{w}=[w_1 \quad w_2 \quad \cdots \quad w_d]^{\mathrm{T}}$ are the weights of each input component connected to the perceptron. The perceptron uses additional and constant inputs, and weight b is also called deviation. The magnitudes of individual weights w_i influences perceptron output, and the equation is expressed in (1).

$$f(\mathbf{x}) = \begin{cases} 1, & if \quad \mathbf{w}^{\mathrm{T}}\mathbf{x} + b \geq 0 \\ 0, & else, \end{cases} \tag{1}$$

The effect of the neuron depends on the activation function, and the activation function is a sign function expressed in (2).

$$sign\,(z) = \begin{cases} +1, & if \quad z \geq 0, \\ -1, & else. \end{cases} \tag{2}$$

The decision-making boundary of the perceptron is (3). The linearity of this decision-making function depends on the input $\mathbf{x}$, so it is called the linear classifier.

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0 \tag{3}$$

The multilayer perceptron network [43] is the combination of multiple layers and multiple perceptrons [44]. The hidden layer is the characteristic capture of data using perceptrons. Data dimensions can be increased or reduced according to the increments and decrements in the perceptrons, and they come from the learning of the provided data. Different types of studies and applications were then developed based on the algorithm of perceptrons [45–49].

C. SUPPORT VECTOR MACHINE

The basic SVM concept serves mainly to identify a decision-making boundary for distinguishing two types and maximizing this decision boundary. The SVM is a supervised learning method used in classification and regression analysis. The statistical method is used to calculate the minimization to estimate the hyperplane of a classification problem. Consequently, SVM itself is a classifier starting from linear separability and finally expanding to nonlinear functions. Its advantages are as follows: The classification model can be established with a small number of samples, and it is applicable in high-dimensional spaces. It has the kernel functions with separability, and can be used for various applications after the kernel functions are changed.

The most basic linear separable function and the symbols that are described in [50-53] are re-organized as follows. The input vector in the input space is expressed as $\mathbf{x}_i$ and $1 \leq i \leq n$. Each input vector $\mathbf{x}_i$ in a high-dimensional space R^d, i.e., $\mathbf{x}_i \in R^d$. In this study, the parameters of training data were $\mathbf{x}_i \in R^{80}$, $1 \leq i \leq 104$. In addition, its corresponding binary label $y_i \in \{-1, +1\}$ means the vectors of the input space are

divided into two categories corresponding to $y_i = 1$ or $y_i = -1$. The data can be expressed as $\{\mathbf{x}_i, y_i\}$, $i = 1,....,n$, $y_i \in \{-1,+1\}$. Suppose a hyperplane separates the input space of two classes and the $\mathbf{x}_i$ points that lay on the hyperplane satisfy $\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b = 0$, where $\mathbf{w} \in R^d$ and b are normal vector to the hyperplane and hyperplane offset, respectively. The optimized hyperplane with the largest margin and the shortest distance to the closest positive (negative) input point that satisfies $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1, \forall i$ is sought. Furthermore, it introduced some nonnegative term ξ_i, called slack variables, that if the optimized hyperplane is unable to linearly separable in classes of input data. The variable is regarded as the measure of misclassifications and is in cooperation with the positive constant term C to determine the tradeoff between the maximization of margin and training error. The soft-margin SVM of the optimization problem is expressed in (4).

$$\begin{aligned} &\underset{\mathbf{w},b,\xi}{minimize}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i^n \xi_i\right\}, \\ &subject\ to\ y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i \end{aligned} \tag{4}$$

The aim is to seek the optimal hyperplane in (4), which is a quadratic programming problem that can be solved using the Lagrangian form of (5) and transformed by Karush-Kuhn-Tucker conditions into the dual problem of (6).

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n}\alpha_i y_i(\mathbf{w}^T\mathbf{x}_i + b) + \sum_{i=1}^{n}\alpha_i \tag{5}$$

$$\begin{aligned} &\underset{\alpha \in R^N}{maximmize} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,\,j=1}^{n}\alpha_i\alpha_j y_i y_j k_{i,j}(\mathbf{x}_i, \mathbf{x}_j) \\ &subject\ to \sum_{i=1}^{n} y_i\alpha_i = 0,\ 0 \leq \alpha_i \leq C, \forall i \end{aligned} \tag{6}$$

where α_i is the Lagrange multiplier of $\mathbf{x}_i$ instance and the kernel function is expressed as $k_{i,j}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$. Finally, most α_i values approach 0 during optimization process, and $\mathbf{x}_i$ that correspond to α_i^* that are larger than 0 are called support vectors. Excluding the nonsupport vectors, N_x represents the number of support vectors, and for all i, $\alpha_i > 0$. The hyperplane expression segregated in this method is shown in (7).

$$\begin{aligned} &\mathbf{w} = \sum_{i=1}^{N_x}\alpha_i^* y_i\mathbf{x}_i, \\ &b = \frac{1}{N_x}\sum_{j=1}^{N_x}(y_i - \sum_{i=0}^{N_x}\alpha_i^* y_i k_{i,j}(\mathbf{x}_i, \mathbf{x}_j)) \end{aligned} \tag{7}$$

Instead, classification $f(\boldsymbol{q})$ calculates the $\boldsymbol{q}$ in the kernel function of every support vector through the new query vector $\boldsymbol{q}$. In the function, b represents the offset of the hyperplane along the normal vector and is obtained through SVM training. The function is expressed in (8).

$$f(\boldsymbol{q}) = sign\left\{\sum_{i=1}^{N_x}\alpha_i y_i k\ (\boldsymbol{q}, \mathbf{x}_i) + b\right\} \tag{8}$$

The detailed verification and derivation process are described in [50–56], and the core functions are developed into different forms [57, 58].

D. TRANSFER LEARNING

Transfer learning is simply the transfer of the parameters obtained from trained models to a new model. The original goal of transfer learning is to solve the deficient number of markers in the data and uneven distribution. The marking of sample data necessitates copious labor time, and problems related to the marking of data result in the predicament of low learning performance models in the training of supervised learning. Because of this, the existing marker data are transferred to the unmarked data. Thus, transfer learning literally means the transfer of models that have already completed learning to new models for the enhancement or acceleration of the new models' training. Alternatively, it can be viewed as the preprocessing of the establishment of new models. The required characteristic parameters can be extracted in advance through such preprocessing or transferring modes so that the training does not have to start from scratch. Relevant proofs appear in the study [59]. This method allows for shorter training and collection of mass data. The characteristic parameters of the marker data of the same source field do not require the same training process to be transferred to the marker data of a different target field [60]. Moreover, for favorable effects, the limitation is that the source field and the target field are required to be correlated. If they exhibit too little correlation, the learning performance cannot be enhanced, and the results are potentially worse. With the rapid development of artificial neural networks, transfer learning constitutes a critical technique in the AI learning field, and it combines mass data with deep neural networks [61–63].

E. BATCH NORMALIZATION

Batch normalization for optimization techniques in deep neural network learning models was developed in [64]. The proposition is to overcome the problem of training difficulties in saturated nonlinear models. In addition, the extremely low learning rates slow down the training speed of the models, resulting in problems with saturated nonlinear models. The study proposed a solution for the aforementioned problem through normalized input layers: batch normalization. The advantages of batch normalization are the increase in learning efficiency, the lack of excessive dependence in the given initial values, and the effective decrease in excessive learning.

The following complies with the explanation in [64]. A small-batch amount B with size m is adopted, and m activation values exist in the minibatch, expressed as $B = \{x^b_{1...m}\} = \{x^b_1 \quad x^b_2 \quad \dots \quad x^b_m\}$. In our experiment, the parameter m represents 512 , and $\hat{x}^b_{i...m}$ and $y_{i...m}$ are the normalized values and their corresponding linear transformations, which correspond to (9) and (10). The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3036912, IEEE Access

referenced transformation is used as Batch Normalized Transform and is expressed as $BN_{\gamma,\beta}: x^{b}{}_{1\ldots m} \rightarrow y_{1\ldots m}$. The ε is a constant value that is added to the minibatch variance to maintain the numerical value stability. The mean and variance of minibatch are expressed in (11) and (12), respectively.

In learning, normalization is performed according to each minibatch, which is the unit for batches. The output of the former activation layer is normalized by subtracting the mean value of the batch and then dividing it with the standard deviation. Furthermore, two learning parameters are added to the input data. They are the representative standard deviation parameter γ and the representative mean value parameter β, which measures the degree of dispersion. For the stochastic gradient descent, the denormalization is executed with these two weight parameters instead of the parameter activation. Thus, the weights of the original network training are maintained, and network stability was preserved. The values trained in every layer are allowed to be transmitted in an effective and stable range. This method receives extensive application and recognition in the deep neural network learning field thereafter [60, 63–68].

$$\hat{x}_i^b = \frac{x^b{}_i - \mu B}{\sqrt{\sigma_B^2 + \varepsilon}} \tag{9}$$

$$y_i \leftarrow \gamma x_i^b + \beta \equiv BN_{\gamma,\beta}(x^b{}_i) \tag{10}$$

$$\mu B = \frac{1}{m}\sum_{i=1}^{m} x^b{}_i \tag{11}$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m} (x^b{}_i - \mu B)^2 \tag{12}$$

III. THE PROPOSED MODEL

The steps of the proposed process and method are explained here and displayed in Fig. 4. The data obtained from the dataset are first analyzed through PCA method, and the dimension is reduced. The characteristics in the data are then extracted in the MLP learning model, and, after learning, transferred through transfer learning to the SVM, which serves as the classifier. The concrete process is shown in the flow chart in Appendix Fig. A1. Finally, patients and healthy individuals can be distinguished effectively.

A. DIMENSION REDUCTION USING PRINCIPAL COMPONENT ANALYSIS

This section explains the initial step of the proposed process and method. PCA can effectively be used to extract features from the data sets of small instances and to preprocess the MLP to normalize the data. PCA, a crucial preprocessing method, can efficiently reduce the dimensions of data. The original dimension of the data is reduced through the PCA conversion, retaining as much useful information as possible. After the experiment, the cumulative proportion of the top five major components obtained is 99.89%.To enable the

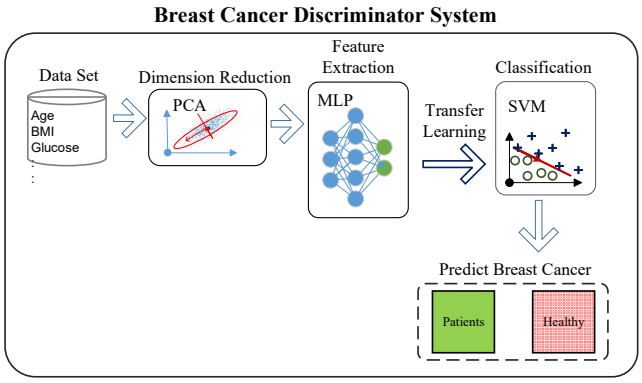


FIGURE 4. Flowchart of multilayer perceptron (MLP) transfer to the support vector machine (SVM) for extracting the feature with a supervised learning method when the proposed method is applied to breast cancer discrimination. The process includes PCA preprocessing, MLP feature extraction, and SVM learning through transfer.

generalizability of the model, we employed *k*-fold cross-validation, dividing data randomly and evenly into *k* sets. Then, a set was used as the testing data and the remaining *k-1* sets were used as the training data. During the training stage, testing data were not introduced. This process was repeated until each set had been used as the testing data. Neither testing nor training data exhibited information-leak problems. Furthermore, the experiment performed *k*-fold cross-validation 50 times on average to enhance the generalizability of the model. As such, after the dimension is reduced to 5, 99.89% of the original data continues to be expressed, but the individual dimension does not retain its original meaning after the transformation. Accordingly, we conducted experiments without PCA and compared their processes with that of the experiment with PCA. The experiments without PCA were multilayer perceptron network (MLP), support vector machine (SVM), k-nearest neighbors (k-NN), decision trees (DTs), random forest (RF), MLP transferred to RF (MLP2RF), and the proposed MLP transfer to SVM (MLP2SVC). Without PCA, the demonstration in Appendix Table A1 indicates that the performance of various algorithms with 10-fold cross-validation was dissatisfactory. At the last line, the 10-fold cross-validation testing value of MLP2SVC without PCA was 61.29%.

B. MULTILAYER PERCEPTRON MODEL

This section explains the function of the MLP. Although the data become more centralized after the dimension reduction in subsection A, fewer characteristics exist. To increase the separability of the characteristics, the MLP network learning is used to extract the more definite characteristics. A further explanation is that the data are used to train a model that can extract the data characteristics. The network structure is designed for exploration first (high Fig. 5 dimension) and development second (low dimension), and the structure is shown in Fig. 5.

First, the high dimensions are explored to 32 nodes through two fully connected hidden layers, and some of the nodes (10%) are deactivated using the random dropout layer. In the fully connected layer, each node is connected to all

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3036912, IEEE Access

nodes of the previous layer to extract all local features. All local features are combined through the weight matrix. However, connecting all nodes in the fully connected layer results in two problems: (1) increasing computation time, which leads to low efficiency, and (2) overfitting. To address the problems, a dropout layer must be inserted. In the dropout layer, connections of nodes to the previous nodes are randomly neglected. These randomly abandoned nodes temporarily lack their weight update in the forward propagation process. In backward propagation, the nodes also lack weight update. The nodes are reduced to 16, and the parameters from the previous layer are normalized through the batch normalization layer. As mentioned in Subsection E of Section II, the parameters mainly increase the training accuracy without causing overfitting. The normalized parameters are sent to the next layer, which is the hidden layer with five nodes. Finally, the output layer has two nodes. In addition, a more detailed illustration of the parameter numbers in every hidden layer and their dropout functions is tabulated in Table I.

C. COMBINATION OF PCA AND MLP MODEL

This section focuses on the experimental conditions displayed by the combination of the two methods of PCA and MLP. The experiment used the data expression that was obtained from the original data (Fig. 6(a)), the original data after MLP (Fig. 6(b)), the original data after PCA (Fig. 6(c)), and the original data after PCA and MLP (Fig. 6(d)). Other than the original data (Fig. 6(a)), the extracted features of all other data had five dimensions. The raw data and those separately obtained through the MLP and PCA methods are visible respectively in Fig. 6(a), Fig. 6(b), and Fig. 6(c). In either Fig. 6(b) or Fig. 6(c), the point distribution was more concentrated than that in Fig. 6(a). The overall range decreased considerably. The visualization technique of the t-distributed stochastic neighbor embedding (t-SNE) [69] is shown in Fig. 6. It cannot provide class-specific cluster comparison between the MLP method (Fig. 6(b)) and the PCA method (Fig. 6(c)) because PCA is an unsupervised dimensionality reduction technique. However, through careful comparison of Fig. 6(b) and Fig. 6(c), we noted that the points of Fig. 6(c) were more concentrated, and the number of the getaway center was smaller than that of Fig. 6(b). Comparing Fig. 6(b) and Fig. 6(d) and their red and blue dots indicates the two categories in the data. The groups of the two categories are clearly seen to be more distinctly separated through the combination of PCA and MLP (Fig. 6(d)). Although blue Dots continue to appear in the red dot group, in comparison, Fig. 6(b) is more group centralized. The differences in the coordinate axes in Fig. 6 (b) and (d) are notable. In Fig. 6(b), the x and y axes are (-150–100) and (-75–75). However, after processing through the combination of PCA and MLP, the x and y axes changed to (-100–50) and (-75–100). Although little difference is exhibited in the y-axis interval, 40% of the original interval was decreased in the x-axis. Through statistical SD and t test [70] analysis, the advantages and disadvantages of the blue

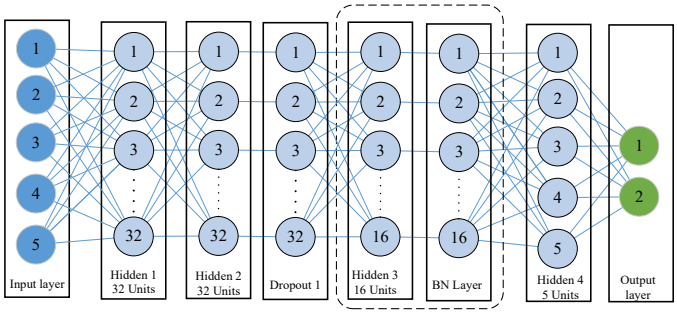


FIGURE 5 Structure of the multilayer perceptron for extraction of the feature with the supervised learning method when applied to 5 attributes.Hidden layers have six layers, including four fully connected layers, one dropout layer, and one batch normalization layer.

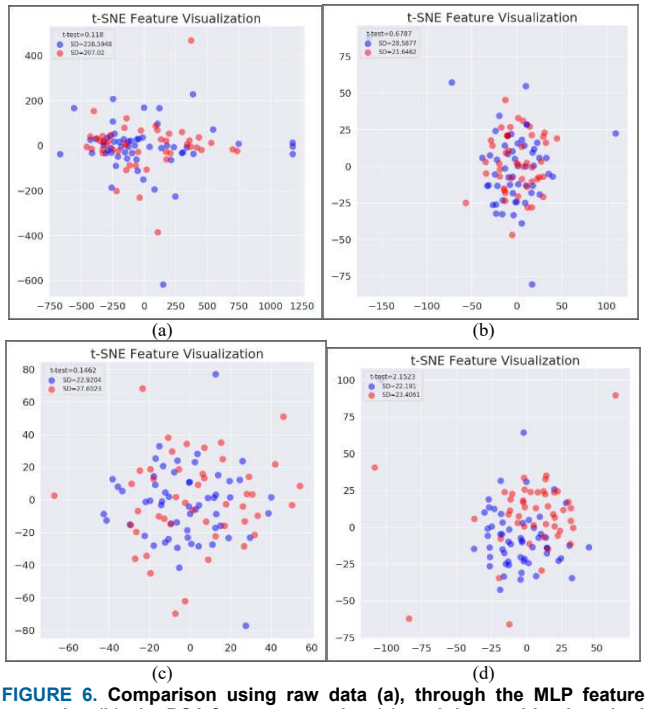


FIGURE 6. Comparison using raw data (a), through the MLP feature extraction (b), the PCA feature extraction (c), and the combined method of PCA and MLP (d) are features set visualization through t-SNE.

TABLE I. Settings of MLP.

Type	Settings	Parameters
Input layer	5	N/A
Dense 1	Number of nodes:32	192
	Act.: ReLU	
Dense 2	Number of nodes:32	1056
	Act.: ReLU	
Dropout	Set:0.1	
Dense 3	Number of nodes:16	528
	Act.: ReLU	
Batch normalization	Number of nodes:16	64
Dense 4	Number of nodes:5	85
	Act.: ReLU	
Dense 5	Number of nodes:2	12
	Act.: Softmax	

and red dots were revealed from SD determination. However, the comparison indicated that Fig. 6(d) had superior performance through the *t* test method. By using the aforementioned indicators to judge the two groups, a clearer separation was achieved. The combination of these two methods proves to be superior.

D. TRANSFERRING MULTILAYER PERCEPTRON NETWORK MODEL

The transfer technique is required to transfer the MLP model that has finished training to other methods as a classifier. Transfer learning is used to transfer the output layer of the MLP model that has finished learning to an SVM classification method. The differences in Table I and Appendix Table A2 were compared. The data pass through characteristic extraction one more time after the exclusion of the final output layer. This makes it possible to select a more flexible and powerful classifier.

E. CLASSIFIER USING SUPPORT VECTOR MACHINE

The SVM classifier is chosen for the final identification. After the comparison of direct classification training with MLP and transfer learning through MLP, the SVM was chosen as the classifier.

IV. THE EXPERIMENTAL RESULTS

This section explains and analyzes the experimental results. First, the data sources are introduced, and the data collation is then explained. Compared with other studies regarding breast cancer detection and prediction, this study focused on the exploration of an algorithm. To enable a fair comparison of the detection accuracy, identical databases were adopted. Ten-fold cross validation was used as the validation method. Training and testing data were separated to avoid learning overfitting and data leakage. Feature extraction of the algorithm could be used to determine the potential links in correlations. Finally, the conclusion of the 10-fold cross-validation is analyzed.

A. DATA DESCRIPTION

The data are from the Breast Cancer Coimbra Dataset (BCCD) of the Machine Learning Repository at the University of California, Irvine [71]. The data were obtained from Manuel Gomes of the University Hospital Centre of Coimbra [72]. From the Department of Obstetrics and Gynecology of the University Hospital Centre of Coimbra, data of patients with breast cancer between 2009 and 2013 were collected. Women newly diagnosed with breast cancer were recruited. The breast cancer diagnosis of each patient was based on a positive mammography result and histologically confirmed. All of these patients' breast cancer was native. The data were collected before any surgery or treatment. Patients who had received treatment before the inquiry were excluded. Furthermore, healthy female volunteers were selected and included as the control group. None of the recruited patients had received prior cancer therapy. At the time of inclusion, the patients exhibited no infection, other acute diseases, or complications. The goal was to evaluate the hyperresistinemia and metabolic anomaly of breast cancer. Every piece of data includes 10 attributes: age (years), BMI (kg/m^2), glucose (mg/dL), insulin (μU/mL), homeostasis model assessment, leptin (ng/mL), adiponectin (μg/mL), resistin (ng/mL), chemokine monocyte chemoattractant protein 1(MCP-1) (pg/dL), and the last label determines whether the person is a confirmed breast cancer patient. A total of 116 pieces of data was compiled with 64 diagnosed patients and 52 healthy individuals. Another three data sets were from the database of the Machine Learning Repository at the University of California, Irvine with the clinical breast cancer cases at University of Wisconsin Hospitals as the sources. Regarding the Breast Cancer Wisconsin (Original) data set collected from 1989 to 1991, a total of 699 pieces of data were compiled with 241 diagnosed patients and 458 healthy individuals. For samples collected through breast fine-needle aspiration, 11 cytological features or attribute messages, each categorized from level 1 to 10, were assigned to determine whether the samples were benign or malignant. The 11 attribute messages were as follows: sample code number, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, and class (2 for benign, 4 for malignant). They were applied in the multi-surface method of pattern separation for medical diagnosis of breast cytology [73].

The Breast Cancer Wisconsin (Diagnosis) data set collected samples from 1989 to 1995; however, the number of attributes assigned to each individual was 32. A total of 569 pieces of data were compiled with 212 diagnosed patients and 357 healthy individuals. Interactive image processing techniques and linear program based classifiers were used to digitalize breast cancer cells [74] and accurately and automatically analyze the cell nuclei size, shape, and textures. For each nuclei, the following 10 features were calculated: radius, perimeter, area, compactness, smoothness, concavity, concave point, symmetry, fractal dimension, and texture. In addition, the mean, maximum (or minimum), and standard errors of each feature were added along with the sample code number and diagnosis (M = malignant and B = benign).

The Breast Cancer Wisconsin (Prognosis) data set has collected data of consecutive patients since 1989. A total of 198 pieces of data were compiled with 47 recurring and 151 nonrecurring individuals. Differing from previous data sets, this one also contains outcome (R = recurring, N = nonrecurring), time (recurrence time if field 2 = R, disease-free time if field 2 = N), tumor size – diameter of the excised tumor in cm, and lymph node status – number of positive axillary lymph nodes observed at time of surgery. Each individual has 34 attributes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3036912, IEEE Access

B. DATA PREPROCESSING

The data are sorted into *x*-input and *y*-output datasets. The *x*-input includes the nine attributes for every piece of data, and the *y*-output includes the label corresponding to the *x*-input that determines whether the person is a confirmed breast cancer patient. After the data are sorted into two datasets, the individual data groups are divided into corresponding training data and test data.

C. K-FOLD CROSS-VALIDATION

K-fold is a common cross-validation method. All data are randomly and evenly distributed into *k* sets. One set is selected as the test data, and other sets are used as training data. This is repeated until every set has been used as test data, which means the test has been performed *k* times. In the experiments, *k* was set as 10, meaning that the 10-fold cross-validation method was used to validate the power of the proposed method. The 116 pieces of data were tested 10 times through 10-fold cross-validation. Each time, one-tenth of the total data was extracted as the test data randomly without repetition, and the rest constituted training data. The first six times, the sample numbers were 11, and the sample numbers were 12 at other times.

For the results, Appendix Fig. A2 illustrates the comparison with other methods including MLP, SVC, and RF. The red bars in the histogram show that the proposed method of MLP transfer to SVM (MLP2SVC) is mostly superior to or as effective as other methods in multiple comparisons. In the first, second, third, eighth, and ninth tests, it outperforms other methods but is inferior to the MLP method in the tenth test. In addition, in Appendix Table A3, three methods are added in numerical representation: *k*-NN, DT, and MLP transferred to RF (MLP2RF). Notably, the MLP2RF method uses the RF method, instead of SVM, as the classifier under the proposed structure. Appendix Fig. A2 and Appendix Table A3 demonstrate that, for the 10-fold cross-validation value was 86.99%. MLP performs well the fifth and tenth times but exhibits the worst performances the first and sixth times. By contrast, MLP transfer to SVM (MLP2SVC) was more stable with smaller oscillations. This shows that, for the performance accuracy, using the MLP model after learning and transferring it to SVM as the classifier provides superior performance. In addition, MLP transferred to RF (MLP2RF) was added for comparison— its accuracy was not superior to that of MLP2SVC. The method is seen to be superior to only the RF method but inferior to the SVC method. For the comparison of six methods, the average accuracy of the 10-fold cross-validation of the proposed method is 86.97%, meaning that it is clearly superior to other methods.

In addition, Appendix Fig. A3 shows the nonnormalized and normalized confusion matrixes of the proposed method in the 10-fold cross-validation. The confusion matrixes can be used to analyze the quantities of incorrect and correct classifications observed. The results are presented in the Appendix Fig. A3. In each test, two confusion matrixes were generated. Nonnormalized confusion matrixes are directly presented using the number of individuals. For example, in Appendix Fig. A3 (a), the total number of individuals tested was 12, six of whom were healthy and six had been diagnosed with the disease. The test results indicated that in the healthy group, five were healthy and one had been diagnosed, whereas in the group diagnosed with the disease, four had been diagnosed and two were healthy. In addition, normalized confusion matrixes demonstrate the percentages of incorrect and correct classifications in each class. To verify the performance, we made the experiments numerous times and determined the mean, which is explained in a following section. The 10th test, shown in Appendix Fig. A3(s), exhibits the worst performance: The number of test samples is 11, and the number of misjudgments is 4, indicating that a maximum 30% rate for misjudgments was not achieved in the worst conditions. The test with the best performance is the fourth test shown in Appendix Fig. A3(g) and (h). The number of test samples is 12, and the number of misjudgments is 0 with an accuracy of 100%.

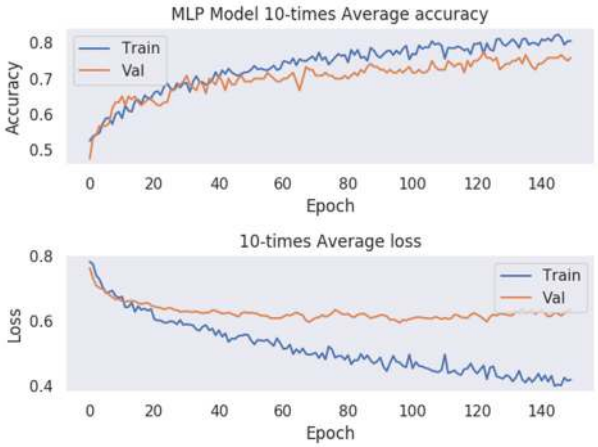


FIGURE 7 **Analysis using the 10-times average of each epoch for overfitting of the multilayer perceptron model architecture.**

The conditions during MLP training require clarification. To avoid overfitting during MLP network iteration training, we conducted an experiment to observe changes in accuracy and losses in training data and verification data. The experiment used all data to the MLP for 150-iteration training. The proportion of training data and validation data was random for 9/10 of the all data and the remaining 1/10 data. This method produced 10 sets of nonrepeating training data and validation data. After testing these 10 sets of data, the mean was obtained from each iteration to draw the curve, as displayed in Fig. 7. This figure shows that in the accuracy index, the training data and the verification data increased as the number of iterations increased. In the loss index, the training data decreased as the number of iterations increased. However, the verification data maintained their level after the 20th iteration. We inferred that as the number of iterations increased, the loss was maintained at a fixed value and did not increase. This signified that the selected network

architecture gradually reached convergence and did not overfit.

D. EXPERIMENTAL RESULTS

Methods that used the data from the Breast Cancer Coimbra Dataset and data from Manuel Gomes from the University Hospital Centre of Coimbra are compared in Table II. The six most accurate methods were compared. The test set that was used in this article and the comparison methods were not consistent. By comparing the various methods presented in Table II, we found that the experiment's breast cancer database had a different proportion distribution for its training and test data. There was 75% training data and 25% test data, and 80% training data and 20% training data.

Furthermore, the methods for calculating accuracy were different. Some used the means of multiple tests for accuracy, and some used just one test for accuracy. In the experiment that was conducted for this study, we conducted 10 unrepeated data groupings to precisely balance the model performance. The grouping proportion was 90% training data and 10% training data. The 10 groups of data were tested after training, and the means of the 10 tests were used for accuracy. Notably, the method proposed by Saritas et al. [75] exhibits a similar, albeit slightly lower, level of accuracy to that of the proposed method in this study. The differences in performance can be clearly seen in the data for the other methods. To make the experiment more accurate, databases that included other breast cancer tests and their corresponding information were analyzed and tabulated in Table III. To ensure accuracy during the experiment when using the databases in Table III, the method proposed by this study was used 50 times, and the means were used for accuracy. The 10-fold cross-validation method was used for each process. All training and testing data were random. Four types of data test results corresponding to Fig. 8 and average and SD statistical analysis were added to the heading. Notably, among the four databases, two (green and magenta lines) had an average accuracy of 0.97 and two (red and blue lines) had average accuracies of 0.81 and 0.82, respectively. In Table III, the distribution of No. of instances corresponds to the projected accuracy. The results suggested that the proposed method had an accuracy exceeding 80% in the relatively few instances (116 and 198). The proposed method had favorable performance regardless of whether the No. of attributes was 10 or 34. The key was the PCA preprocessing

TABLE II. Comparisons of BCCD Data Accuracy for Several Methods.

Method	Accuracy
Average K-mean + Local Outlier Rectifier V.2.0 [76]	56.90%
Random forest model [77]	74.3%
Artificial neural network and naive Bayes Classification algorithm [75]	86.95%
Extreme learning machine [78]	80.00%
PySpark and its machine learning frameworks [79]	83.0%
Classification via regression [80]	80.0%
Proposed model	86.97%

TABLE III. Comparison of results of using the proposed method with 50 times 10-fold cross-validation for various breast cancer data sets

Dataset	No. of Attributes	No. of Instances	No. of Classes	Avg. Acc. over 50-times
Breast Cancer Coimbra Data Set	10	116	2	0.82
Breast Cancer Wisconsin (Original) [81]	11	699	2	0.97
Breast Cancer Wisconsin (Diagnosis) [82]	32	569	2	0.97
Breast Cancer Wisconsin (Prognosis) [83]	34	198	2	0.81

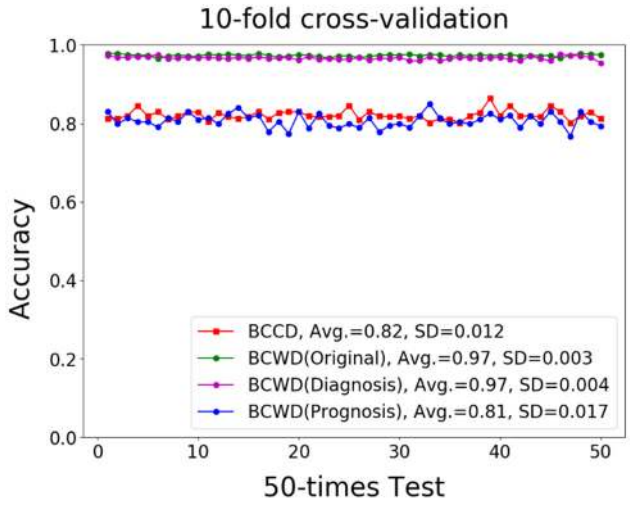


FIGURE 8 Average accuracy comparsion of BCCD, BCWD(Original), BCWD(Diagnosis), and BCWD(Prognosis) with 10-fold cross-validation performed 50 times.

step in the proposed method. When the No. of instances increased by three- to six-fold (699 and 569 instances), the method attained 97% accuracy, indicating that it exhibited excellent performance as the number of instances increased. The differences between numbers of attributes did not affect the method's performance. Although the Breast Cancer Wisconsin (Original) and (Diagnosis) data sets share University of Wisconsin Hospitals as their source of clinical breast cancer cases, their No. of attributes differ. The main reason was that the Breast Cancer Wisconsin (Original) data set categorizes each of the 11 attributes into 1–10 levels during sample collection, whereas the Breast Cancer Wisconsin (Diagnosis) data set uses 32 attributes to present accurate cell digitization. Therefore, the same data have different feature expressions. Despite the difference in the number of features (attributes), the proposed method extracted effective features.

The highest accuracy achieved during 10-fold cross-validation (presented in Appendix Table A3) was 86.97%, as applied to the BCCD data set. Furthermore, a comparison of the results of six methods (Table II) using the same data set suggested that the proposed method is clearly superior. For further comparison of its performance (see Fig. 8), three additional breast cancer databases were added to test the

proposed method and 10-fold cross-validation was performed 50 times, and the results indicated that the proposed method performs well.

V. DISCUSSIONS

This study normalized the reduced dimensions from PCA and used MLP to obtain the features before transfer learning was used for conversion to a new model. Finally, SVM was used as the classifier. The aforementioned calculation method was integrated with a constructed medical care IoT system architecture. Four data sets of actual medical treatment were collected for use in predicting breast cancer occurrence. The means of 50-times 10-fold cross-validation were compared with those of other methods to determine the method's accuracy in predicting breast cancer occurrence. The results revealed that the proposed method is effective and superior to other methods.

For transfer learning model implementation, the network learned a set of rich and discriminative features for recognizing hundreds to thousands of object classes. Thus, these filters could be reused for tasks other than what the transfer learning model was originally trained for. This study used the transfer learning model to differentiate two categories. The result indicated that this process could be applied to classification problems with few categories, extending the applicability of transfer learning.

Using transfer learning to improve the model's effectiveness, particularly for situations with little target data, is a very convenient and effective method. First, a neural network is trained from a large amount of source data. The features that are learned through the NN are universal and can be generalized. Generally, only two methods of using the pretraining model to realize transfer learning are available: feature extraction (converting features of the source data and the target data to the same space) and fine-tuning (parameter-sharing mechanism for source data and the target data). The more suitable method is based on the quantity of target data and the data characteristics. The following compares the similarities and differences of these two strategies. In the feature extraction strategy, only the weight or method of the last (newest) layer changes during the training process when training a new model. The feature extraction and classifiers (fully connected layer) are viewed as different individuals. The posttransfer model with classifiers removed is the feature extractor. For fine-tuning, all weights change except for the weight of the last layer in the original model. The features of the low layers (the first few layers) in a frozen model have less generalizability and universality. These layers are trained together with higher layers (the last few layers) to fine-tune the weight, which is helpful for identifying new categories.

Regarding the nonlinear methods addressed in Table II, the design concept of artificial neural networks (ANNs) is based on a simulation of the biological nerve conduction mechanism that was mentioned at the beginning of this paper and has the same concept as MLP. In the ANNs in [84] and [85], the neurons in each layer had an input and output. The activation function was used to balance the importance of neuron output. The key function of nonlinear activation was to convert the linear into nonlinear relationship. Sigmoid function, tangent, Leaky, Rectified Linear Unit, Exponential Linear Unit, and Maxout are commonly used for the activation function. Adding some nonlinear elements to a neural network can make the neural network better at solving complex problems.

Regression analysis is generally in the scope of statistics and is often built on the basis of a large quantity of observed data. Regression analysis attempts to build a regression function relationship between the dependent variable (target) and the independent variable (prediction). Regression analysis is often seen in predictive model building, time sequence model building, and finding the correlation between variables. Regression analysis estimates the correlation between two or more variables and can present the significant relationship between the dependent variable and independent variable. Regression analysis can express the differing impact of multiple independent variables on the dependent variable. Depending on whether the function expression between the independent and dependent variables is linear or nonlinear, regression analysis can be divided into linear regression or nonlinear regression.

Next, we discuss DTs. A DT is a type of simple nonlinear model that is suitable for predicting classification and regression data types. The model mainly uses the tree branch concept in a tree-shaped decision model. Data sets are continuously broken into smaller subsets to grow the DT. Two types of tree-shape nodes can appear: the internal node (which indicates one feature) and the leaf node (which indicates a category). To make a determination, the sample's features are judged by the internal nodes until it reaches the leaf node where the sample is classified.

The most recent transfer learning work, [86], proposed a spectrum-sensing framework based on deep learning. Transfer learning was incorporated into the framework and applied to different scenarios with different wireless signals and transmissions to improve sensing performance and robustness. Literature [87] proposed a method to improve fuzzy transfer learning based on the fuzzy system (particularly based on fuzzy rule models). An innovative fuzzy rule method was proposed that combines the infinite Gaussian mixture model with active learning, which improves the performance and versatility of the constructed model. In [88], transfer learning was applied to aggregate data from multiple users in the hand gesture recognition field (based on the electromyogram). Deep learning calculation was used to learn to determine features from large data sets, thereby reducing the record load while improving gesture recognition ability. Reference [89] proposed a transfer learning method that differs from past methods. This method reduces the deviation that is produced by the learning delivery function caused by data shortages in certain categories. Past transfer learning methods based on domain adaptation calculation in subspace reconstruction ignored categories. This proposed category reconstruction adaptation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3036912, IEEE Access

method is called class-specific reconstruction transfer learning. This method uses intraclass dependency and mutual relationships to optimize the model's excellent transfer loss function. Literature [90] indicated that deep learning is effective in predicting machine equipment malfunction and explored the transfer deep network. The deep network was trained through the machine equipment's historic malfunction data, and the transfer of the deep network was used to predict new subjects. This reference proposed a deep transfer learning (DTL) network based on the sparse autoencoder (SAE). The DTL method uses three types of transfer strategies: weight transfer, hidden feature transfer learning, and weight update. SAEs that are trained with historic malfunction data are transferred to new subjects.

To verify the proposed method, we performed cross-validation method 50 times in the experiment. In the cross-validation, the experimental results were stable, and thus, overfitting was avoided. The *k*-fold cross-validation method involved the random and even division of data into k sets. Then, a set was used as the testing data, whereas the remaining *k*-1 sets were used as the training data. During the training stage, testing data were not introduced. This process was repeated until each set had been used as the testing data.

We reviewed the data distribution of each designed stage, and Fig. 6 facilitates the explanation. The blue and red dots represent the two categories of data. The original data have eight attributes for each individual, representing eight dimensions. However, to visualize the presentation, we used visualization techniques of the t-SNE method to degrade the multidimensional space into a two-dimensional plane. In addition to observing the blue and red dot positions to identify the individual distribution, we used statistical SD and *t*-test methods to analyze the values of the two groups. The *t*-test verified whether the two groups of independent data exhibited significant differences. A *t*-test result approaching zero indicated no difference between the two groups. The results are placed in the upper-left corner of each panel of Fig. 6. Fig. 6(a) depicts the original data analysis with a *t* value of 0.118. The SD values of the two categories were 238.5 and 207.0, respectively, which were used as the control group. Data preprocessed through PCA are presented in Fig. 6(c). The *t* value became 0.146 and the SD values were 22.9 and 27.6 for the two categories, respectively. Although little difference was observed between the *t* values, the two categories' SD values exhibited substantial improvements. The curves exhibit observable changes in the horizontal and vertical axes. The data after MLP processing are presented in Fig. 6(d) with a *t* value of 2.152 and SD values of 22.1 and 23.4 for the two categories, respectively. The *t* value improved substantially, and the SD of the blue group improved slightly. The individual distribution groups largely separated. Then, the MLP learning was transferred to the SVM, which created the final decision boundary and divided the data into two categories.

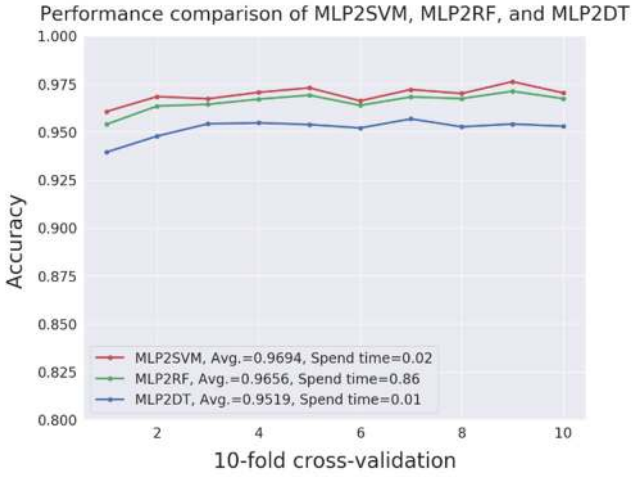


FIGURE 9 Performance comparison of MLP2SVM, MLP2RF, and MLP2DT with 10-fold cross-validation performed 50 times for each fold average. Statistics present the average overall accuracy as well as the average time spent on each fold cross-validation (unit: sec).

On the basis of transfer learning, after MLP training (approximately 13 seconds), the data were transferred to classifiers. We selected the most commonly used classifiers, namely SVM, RF, and DT, and compared them using the following methods. We ran 50 iterations using the 10-fold cross-validation method and recorded the accuracy and time spent in each fold cross-validation. After 50 runs, the results of each fold were averaged. The accuracy is plotted in Fig. 9, in which "Avg." represents the mean of 10-fold accuracy and "spend time" represents the time spent for each fold cross-validation calculation on average (unit: sec). As shown in Fig. 9, the average accuracy for each fold cross-validation reached 92% or higher for each of the classifiers. According to the time spent by the classifiers, MLP2RF had the most salient "spend time" compared with the other two classifiers; however, the accuracy performance of MLP2RF was slightly lower than that of MLP2SVM. Although MLP2DT required the shortest time, its accuracy performance was the worst. In comparison, the proposed MLP2SVM exhibited outstanding performance in both time spent and accuracy. Although the accuracy performance had a difference of merely 0.38%, the temporal difference reached 43 folds. The results verified that the selected SVM classifier was the most suitable option under the framework.

The decision maker in the system remains the doctor. A limitation of this system is that it requires comprehensive evaluation by a physician for diagnosis. The AI model provides objective information analysis of data. Specific data originate from a large number of clinical patients, and the professional judgments that doctors make regarding the diagnosis category is the basis that is provided to the AI model. Recommendations and prompts are provided to doctors by the AI model as measurement indicators. Thus, the AI model can assist doctors in reducing errors and increasing diagnosis efficiency. For example, judgment errors can be caused easily by inaccurate measurement

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3036912, IEEE Access

values, improper medical record management, or poor communication. The AI model offers a safe and efficient method for assisting doctors in breast cancer diagnosis.

VI. CONCLUSION

A novel processing method has been proposed in this study to effectively predict breast cancer incidence. Experiments verified the proposed method, and the accuracy reached 86.94% after comparison with the other five methods in 10-fold cross-validation, which outperforms other methods. In addition, every confusion matrix in the 10-fold cross-validation was examined to verify the accuracy of this method. The learning effects of the characteristic extraction of MLP have been then verified after comparison with the use of SVM for classification before and after MLP learning. The results of the verification indicated that the proposed method can effectively and feasibly predict breast cancer incidence. The proposed method exhibited limitations in that a trained model fails to modify itself when new data are imported and that, under the MLP network training framework, training time increases as the data size increases. Future studies should employ solutions aimed at continuous learning and accelerating network training so as to eliminate existing limitations in the proposed method.

APPENDIX

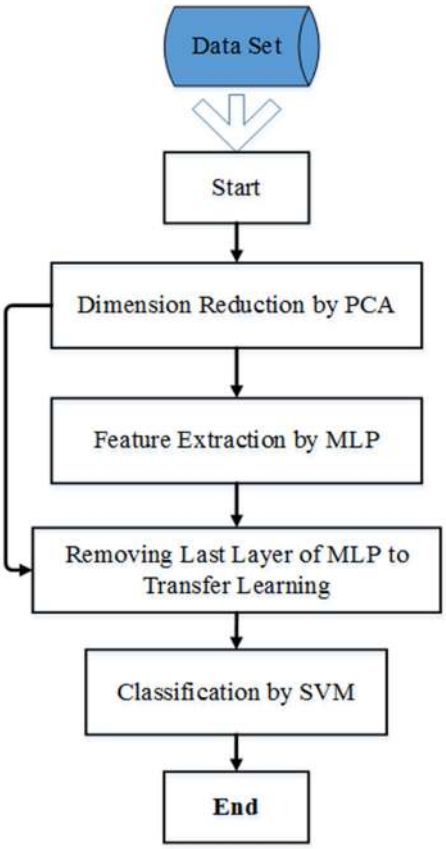


FIGURE A1. System flowchart of the training process.

TABLE A1. Comparison for various models on-10-fold cross-validation without PCA.

Test	MLP	SVC	kNN	DT	RF	MLP2RF	Without PCA
1	41.67	50.00	58.33	66.67	58.33	66.67	50.00
2	58.33	50.00	58.33	66.67	66.67	58.33	50.00
3	66.67	58.33	58.33	50.00	66.67	66.67	58.33
4	58.33	58.33	58.33	91.67	91.67	58.33	58.33
5	75.00	58.33	50.00	83.33	75.00	66.67	83.33
6	58.33	58.33	58.33	58.33	75.00	41.67	90.91
7	54.55	54.55	63.64	81.82	81.82	72.73	54.55
8	45.45	54.55	81.82	63.64	63.64	72.73	54.55
9	54.55	54.55	54.55	54.55	72.73	45.45	54.55
10	54.55	54.55	54.55	63.64	72.73	45.45	54.55
Avg	56.74	55.15	59.62	68.03	72.42	59.47	61.29

TABLE A2. Settings transfer of MLP.

Type	Settings	Parameters
Input layer	5	N/A
Dense 1	Number of nodes: 32 Act.: ReLU	192
Dense 2	Number of nodes: 32 Act.: ReLU	1056
Dropout	Set: 0.1	
Dense 3	Number of nodes: 16 Act.: ReLU	528
Batch normalization	Number of nodes: 16	64
Dense 4	Number of nodes: 5 Act.: ReLU	85

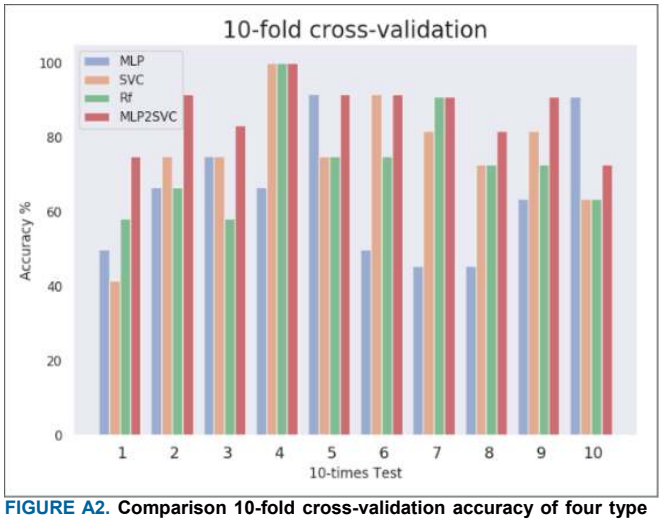
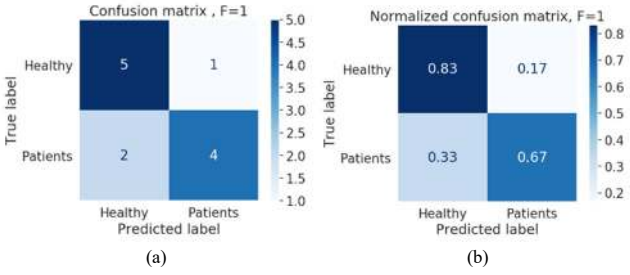
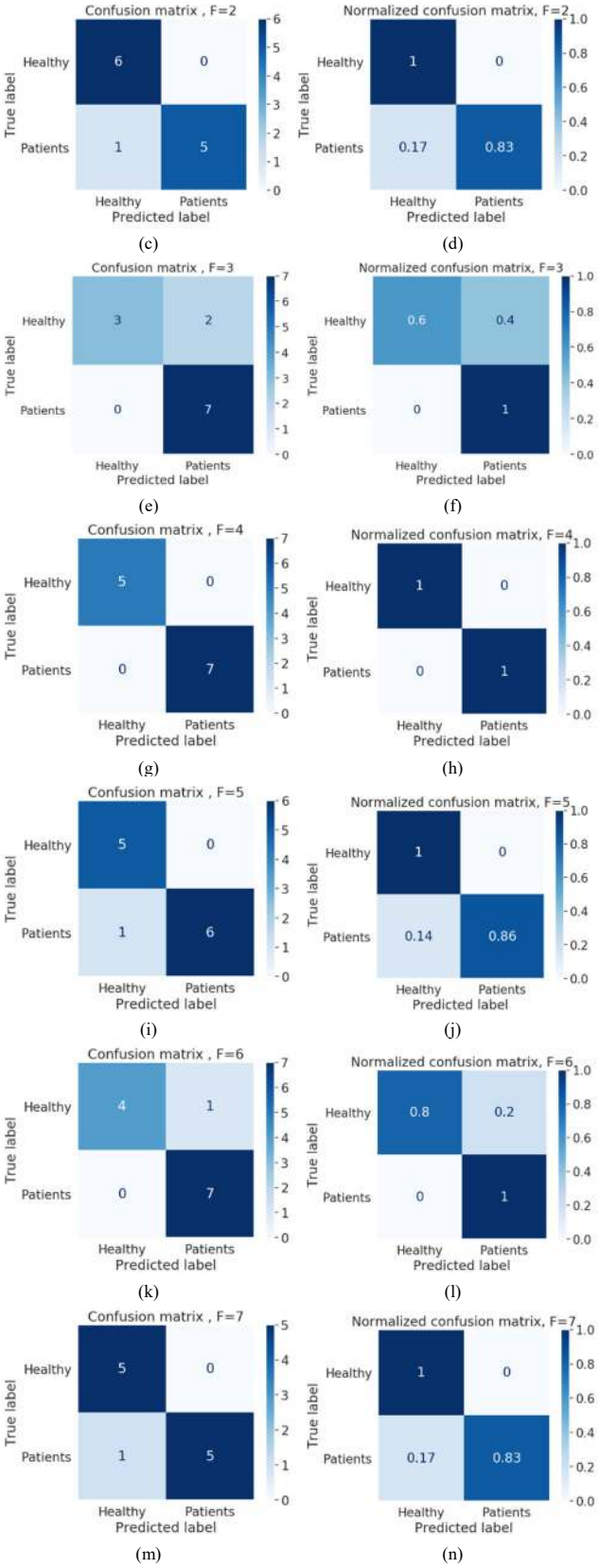


FIGURE A2. Comparison 10-fold cross-validation accuracy of four type modes.



(a) (b)



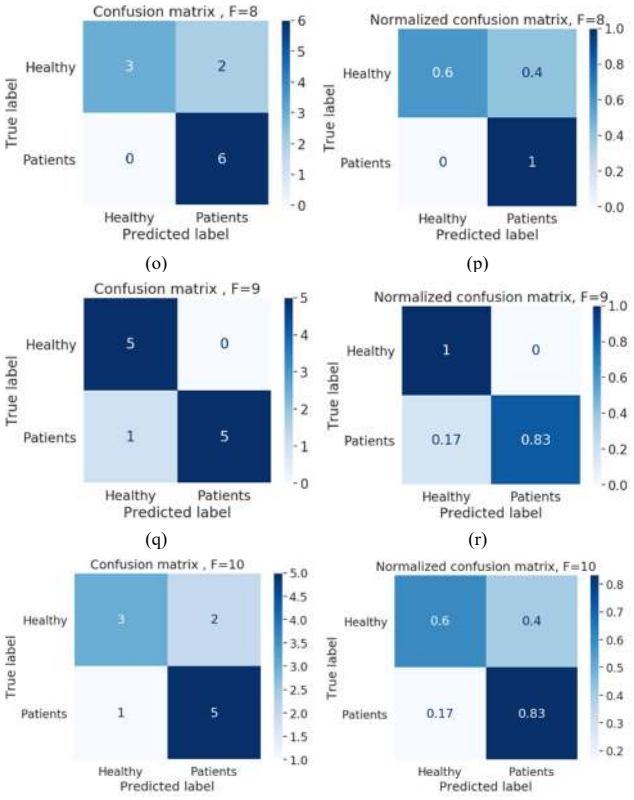


FIGURE A3. Nonnnormalized and normalized confusion matrixes of the proposed method in the 10-fold cross-validation.

TABLE A3. Comparison for various models on-10-fold cross-validation.

Test	MLP	SVC	kNN	DT	RF	MLP2RF	Proposed
1	50.00	41.67	41.67	50.00	58.33	75.00	75.00
2	66.67	75.00	58.33	58.33	66.67	83.33	91.67
3	75.00	75.00	66.67	58.33	58.33	58.33	83.33
4	66.67	100	91.67	66.67	100	100	100
5	91.67	75.00	75.00	66.67	75.00	91.67	91.67
6	50.00	91.67	83.33	66.67	75.00	83.33	91.67
7	45.45	81.82	54.55	72.73	90.91	81.82	90.91
8	45.45	72.73	72.73	72.73	72.73	72.73	81.82
9	63.64	81.82	54.55	63.64	72.73	45.45	90.91
10	90.91	63.64	72.73	63.64	63.64	63.64	72.73
Avg	64.55	75.83	67.12	63.94	73.33	75.53	86.97

REFERENCES

[1] V. G. Vogel, "Epidemiology, genetics, and risk evaluation of postmenopausal women at risk of breast cancer," *Menopause*, vol. 15, no. 4, 2008.

[2] J. Ferlay *et al.*, "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012," *Int. J. Cancer*, vol. 136, no. 5, pp. E359–E386, Mar.2015.

[3] C. Saunders and S. Jassal, *Breast cancer*, Oxford. Oxford University Press, 2009.

[4] A. G. Renehan, M. Tyson, M. Egger, R. F.Heller, and M. Zwahlen, "Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies," *Lancet*, vol. 371, no. 9612, pp. 569–578, 2008.

[5] J. Crisóstomo *et al.*, "Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer," *Endocrine*, vol. 53, no. 2, pp. 433–442, 2016.

[6] J. A. Tice, S. R. Cummings, E. Ziv, and K. Kerlikowske, "Mammographic breast density and the gail model for breast cancer

risk prediction in a screening population," *Breast Cancer Res. Treat.*, vol. 94, no. 2, pp. 115–122, 2005.
[7] A. R. Brentnall *et al.*, "Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort," *Breast Cancer Res.*, vol. 17, no. 1, p. 147, 2015.
[8] Z. Wang *et al.*, "Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion With CNN Deep Features," *IEEE Access*, vol. 7, pp. 105146–105158, 2019.
[9] R. K. Ross, A. Paganini-Hill, P. C. Wan, and M. C. Pike, "Effect of hormone replacement therapy on breast cancer risk: estrogen versus estrogen plus progestin," *JNCI J. Natl. Cancer Inst.*, vol. 92, no. 4, pp. 328–332, Feb.2000.
[10] J. Tyrer, S. W. Duffy, and J. Cuzick, "A breast cancer prediction model incorporating familial and personal risk factors," *Stat. Med.*, vol. 23, no. 7, pp. 1111–1130, Apr.2004.
[11] A. Collins and I. Politopoulos, "The genetics of breast cancer: risk factors for disease," *Appl. Clin. Genet.*, vol. 4, pp. 11–19, Jan.2011.
[12] Y. Zhang *et al.*, "ELMO: An Efficient Logistic Regression-Based Multi-Omic Integrated Analysis Method for Breast Cancer Intrinsic Subtypes," *IEEE Access*, vol. 8, pp. 5121–5130, 2020.
[13] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, "A deep learning mammography-based model for improved breast cancer risk prediction," *Radiology*, vol. 292, no. 1, pp. 60–66, May2019.
[14] G. G. N.Geweid andM. A.Abdallah, "A Novel Approach for Breast Cancer Investigation and Recognition Using M-Level Set-Based Optimization Functions," *IEEE Access*, vol. 7, pp. 136343–136357, 2019.
[15] X.Li, M.Radulovic, K.Kanjer, andK. N.Plataniotis, "Discriminative Pattern Mining for Breast Cancer Histopathology Image Classification via Fully Convolutional Autoencoder," *IEEE Access*, vol. 7, pp. 36433–36445, 2019.
[16] M. R. Karim, G. Wicaksono, I. G. Costa, S. Decker, and O. Beyan, "Prognostically relevant subtypes and survival prediction for breast cancer based on multimodal genomics data," *IEEE Access*, vol. 7, pp. 133850–133864, 2019.
[17] B. Abdikenov, Z. Iklassov, A. Sharipov, S. Hussain, and P. K. Jamwal, "Analytics of Heterogeneous Breast Cancer Data Using Neuroevolution," *IEEE Access*, vol. 7, pp. 18050–18060, 2019.
[18] Q. Wuniri, W. Huangfu, Y. Liu, X. Lin, L. Liu, and Z. Yu, "A Generic-Driven Wrapper Embedded With Feature-Type-Aware Hybrid Bayesian Classifier for Breast Cancer Classification," *IEEE Access*, vol. 7, pp. 119931–119942, 2019.
[19] H. M. Whitney, H. Li, Y. Ji, P. Liu, and M. L. Giger, "Comparison of breast mri tumor classification using human-engineered radiomics, transfer learning from deep convolutional neural networks, and fusion methods," *Proc. IEEE*, vol. 108, no. 1, pp. 163–177, 2020.
[20] X.-X. Niu and C. Y. Suen, "A novel hybrid CNN–SVM classifier for recognizing handwritten digits," *Pattern Recognition*, vol. 45, no. 4, pp. 1318–1325, 2012.
[21] G. Capizzi, G. L. Sciuto, C. Napoli, D. Połap, and M. Woźniak, "Small Lung Nodules Detection Based on Fuzzy-Logic and Probabilistic Neural Network With Bioinspired Reinforcement Learning," *IEEE Trans. Fuzzy Syst., vol. 28, no. 6, pp. 1178–1189, 2020*.
[22] W. Wei, B. Zhou, D. Połap, and M. Woźniak, "A regional adaptive variational PDE model for computed tomography image reconstruction," *Pattern Recognit., vol. 92, pp. 64–81, 2019.*
[23] F. Beritelli, G. Capizzi, G. Lo Sciuto, C. Napoli, and M. Woźniak, "A novel training method to preserve generalization of RBPNN classifiers applied to ECG signals diagnosis," *Neural Networks,* vol. 108, pp. 331–338, 2018.
[24] "Study on the Application of Internet of Things-Based Intelligent Microscope in Blood Cell Analysis," *J. Comput. Theor. Nanosci.*, vol. 14, no. 2, 2017.
[25] Sysmex America, Inc., "Sysmex XE-5000 Specifications," in *XE-5000™ Automated Hematology System Leaflet*, [online document], 2010. Available: http://photos.labwrench.com/equipmentManuals/10820-4303.pdf
[26] P. D. P.Adi and A.Kitagawa, "ZigBee Radio Frequency (RF) Performance on Raspberry Pi 3 for Internet of Things (IoT) based Blood Pressure Sensors Monitoring," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, 2019.
[27] A.Agirre, A.Armentia, E.Estévez, and M.Marcos, "A Component-Based Approach for Securing Indoor Home Care Applications," *Sensors (Basel).*, vol. 18, no. 1, p. 46, Dec.2017.
[28] A.PARIS, T.NHAN, E.CORNET, J.-P.PEROL, M.MALET, and X.TROUSSARD, "ORIGINAL ARTICLE: Performance evaluation of the body fluid mode on the platform Sysmex XE-5000 series automated hematology analyzer," *Int. J. Lab. Hematol.*, vol. 32, no. 5, pp. 539–547, Oct.2010.
[29] Online Browsing Platform (OBP), "Examination processes," *ISO 15189:2012 Medical laboratories — Requirements for quality and competence*, para. 5.5, 2012. [Online]. Available:https://www.iso.org/obp/ui/#iso:std:iso:15189:ed-3:v2:en
[30] International Bureau of Weights and Measures, "Evaluation of Measurement Data – Guide to the Expression of Uncertainty in Measurement," *ISO Guide 98-3*, 2008. [Online]. Available as JCGM 100:2008:http://www.bipm.org/en/publications/guides/gum.html
[31] R. da Silva and A. Williams, "Eurachem/CITAC Guide: Setting and Using Target Uncertainty in Chemical Measurement. Leoben, Austria: Eurachem; 2015." 2016. [Online]. Available: http://www.citac.cc/STMU_2015_EN.pdf
[32] C. G. Fraser, *Biological variation: from principles to practice*. Amer. Assoc. for Clinical Chemistry, pp. 55, 2001.
[33] J. O. Westgard, "Method validation," *Basic Method Validation. 2nd ed. Madison, WI Westgard QC*, pp. 156–157, 2003.
[34] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, Nov.1901.
[35] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, 1933.
[36] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychol. Rev.*, vol. 65, no. 6, p. 386, 1958.
[37] F. Rosenblatt, "Principles of neurodynamics. perceptrons and the theory of brain mechanisms," https://apps.dtic.mil/dtic/tr/fulltext/u2/256582.pdf, 1961.
[38] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, 1943.
[39] D. O. Hebb, *The organization of behavior*. na, 1949.
[40] T. V. P. Bliss and G. L. Collingridge, "A synaptic model of memory: long-term potentiation in the hippocampus," *Nature*, vol. 361, pp. 31–39, 1993.
[41] M. F. Bear and R. C. Malenka, "Synaptic plasticity: LTP and LTD," *Curr. Opin. Neurobiol.*, vol. 4, no. 3, pp. 389–399, 1994.
[42] J. A. Kauer and R. C. Malenka, "Synaptic plasticity and addiction," *Nat. Rev. Neurosci.*, vol. 8, no. 11, pp. 844–858, 2007.
[43] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.
[44] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm," *Mach. Learn.*, vol. 37, no. 3, pp. 277–296, 1999.
[45] D. O. Hebb, *The organization of behavior: A neuropsychological theory*. Psychology Press, John Wiley and Sons, Inc., New York, 1949.
[46] E. Romero and J. M. Sopena, "Performing feature selection with multilayer perceptrons," *IEEE Trans. Neural Networks*, vol. 19, no. 3, pp. 431–441, 2008.
[47] B. Gas, "Self-organizing multilayer perceptron," *IEEE Trans. Neural Networks*, vol. 21, no. 11, pp. 1766–1779, 2010.
[48] J. Tang, C. Deng, and G. BinHuang, "Extreme Learning Machine for Multilayer Perceptron," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 27, no. 4, pp. 809–821, 2016.
[49] W. Liu, P. Gao, Y. Wang, W. Yu, and M. Zhang, "A unitary weights based one-iteration quantum perceptron algorithm for non-ideal training sets," *IEEE Access*, vol. 7, pp. 36854–36865, 2019.
[50] B. Scholkopf *et al.*, "Input space versus feature space in kernel-based methods," *IEEE Trans. Neural Networks*, vol. 10, no. 5, pp. 1000–1017, 1999.
[51] A. Bordes, S. Ertekin, J. Weston, and L. Bottou, "Fast Kernel Classifiers with Online and Active Learning," *J. Mach. Learn. Res.*, vol. 6, pp. 1579–1619, 2005.

[52] C.-F. Lin and S.-D.Wang, "Fuzzy support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 464–471, 2002.
[53] C. J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.
[54] C. J. C. Burges and B. Schölkopf, "Improving the accuracy and speed of support vector machine," in *Advances in Neural Information Processing Systems*, 1997, pp. 375–381.
[55] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. their Appl.*, vol. 13, no. 4, pp. 18–28, 1998.
[56] B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds., *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA, USA: MIT Press, 1999.
[57] B. Schölkopf, A. J. Smola, F. Bach, *et al*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
[58] B. Haasdonk, "Feature space interpretation of SVMs with indefinite kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 4, pp. 482–492, 2005.
[59] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pp. 3320–3328, 2014.
[60] S. J. Pan and Q. A. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.
[61] M. Long, Y. Cao, J. Wang, and M. I.J ordan, "Learning transferable features with deep adaptation networks," *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, pp. 97–105, 2015.
[62] D. M. Don, N. A. Goldstein, D. M. Crockett, and S. D. Ward, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 133, no. 4, pp. 562–568, 2016.
[63] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, 2016.
[64] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning, ICML 2015*, 2015, vol. 1, pp. 448–456.
[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, vol. 2016-Decem, pp. 770–778.
[66] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[67] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
[68] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.
[69] L.van der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, pp. 3221–3245, 2014.
[70] Student, "The probable error of a mean," In: S. Kotz and N. L. Johnson (eds) *Breakthroughs in Statistics*. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY. https://doi.org/10.1007/978-1-4612-4380-9_4.
[71] "Breast Cancer Coimbra Dataset of the Machine Learning Repository in University of California, Irvine." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra.
[72] M. Patrício *et al.*, "Using Resistin, glucose, age and BMI to predict the presence of breast cancer," *BMC Cancer*, vol. 18, no. 1, pp. 1–8, 2018.
[73] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 87, no. 23, pp. 9193–9196, Dec. 1990.
[74] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Proc.SPIE*, 1993, vol. 1905.
[75] M. Saritas and A.Yasar, "Performance analysis of ANN and naive Bayes classification algorithm for data classification," *Int. J. Intell. Syst. Appl. Eng.*, vol. 7, no. 2 SE-Research Article, Jun.2019.
[76] R. O. Badiang, B. D. Gerardo, and R. P. Medina, "Relocating local outliers produced by K-means and K-medoids using local outlier rectifier V.2.0," in *2019 International Conference on Advanced Computer Science and information Systems (ICACSIS)*, 2019, pp. 89–94.
[77] Y. Li, "Performance evaluation of machine learning methods for breast cancer prediction," *Appl. Comput. Math.*, vol. 7, no. 4, p. 212, 2018.
[78] M. F. Aslan, Y. Celik, K. Sabanci, and A. Durdu, "Breast cancer diagnosis by different machine learning methods using blood analysis data," *Int. J. Intell. Syst. Appl. Eng.*, vol. 6, no. 4 SE-Research Article, Dec.2018.
[79] P. D. Hung, T. D. Hanh, and V. T. Diep, "Breast cancer prediction using Spark MLlib and ML packages," in *Proceedings of the 2018 5th International Conference on Bioinformatics Research and Applications*, 2018, pp. 52–59.
[80] F. Sardouk, A. D. Duru, O. Bayat, *et al*, "Classification of breast cancer using data mining," *Am. Sci. Res. J. Eng. Technol. Sci.*, vol. 51, no. 1, pp. 38–46, 2019.
[81] "Breast Cancer Wisconsin (Original) Dataset of the Machine Learning Repository in University of California, Irvine." [Online]. Available:https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original).
[82] "Breast Cancer Wisconsin (Diagnostic) Dataset of the Machine Learning Repository in University of California, Irvine." [Online]. Available:https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
[83] "Breast Cancer Wisconsin (Prognostic) Dataset of the Machine Learning Repository in University of California, Irvine." [Online]. Available:https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic)
[84] D. Anderson and G. McNeill, "Artificial neural networks technology," *Kaman Sci. Corp.*, vol. 258, no. 6, pp. 1–83, 1992.
[85] J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress," *Neurocomputing,* vol. 74, no. 1, pp. 239–255, 2010.
[86] Q. Peng, A. Gilman, N. Vasconcelos, P. C. Cosman, and L. B. Milstein, "Robust Deep Sensing Through Transfer Learning in Cognitive Radio," *IEEE Wirel. Commun. Lett.*, vol. 9, no. 1, pp. 38–41, 2020.
[87] H. Zuo, J. Lu, G. Zhang, and F. Liu, "Fuzzy Transfer Learning Using an Infinite Gaussian Mixture Model and Active Learning," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 2, pp. 291–303, 2019.
[88] U. Côté-Allard et al., "Deep Learning for Electromyographic Hand Gesture Signal Classification Using Transfer Learning," I*EEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 4, pp. 760–771, 2019.
[89] S. Wang, L. Zhang, W. Zuo and B. Zhang, "Class-Specific Reconstruction Transfer Learning for Visual Recognition Across Domains," in *IEEE Transactions on Image Processing*, vol. 29, pp. 2424-2438, 2020, doi: 10.1109/TIP.2019.2948480.
[90] C. Sun, M. Ma, Z. Zhao, S. Tian, R. Yan and X. Chen, "Deep Transfer Learning Based on Sparse Autoencoder for Remaining Useful Life Prediction of Tool in Manufacturing," in *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2416-2425, April 2019, doi: 10.1109/TII.2018.2881543.

Huan-Jung Chiu received the B.S. degree in the Department of Mechanical and Marine Engineering, National Kaohsiung Marine Institute of Technology, Kaohsiung, Taiwan, in 2001, and the M.S. degree from the Department of Electrical Engineering, National Cheng Kung University (NCKU), Tainan, Taiwan, in 2012. He is currently pursuing a Ph.D. degree in NCKU. His current research interests include fuzzy control, intelligent systems, humanoid robot, image processing, machine learning, and robotic application.

Tzuu-Hseng S. Li (S'85–M'90) received the B.S. degree from the Tatung Institute of Technology, Taipei, Taiwan, in 1981, and the M.S. and Ph.D. degrees from National Cheng Kung University (NCKU), Tainan, Taiwan, in 1985 and 1989, respectively, all in electrical engineering. Since 1985, he has been with the Department of Electrical Engineering, NCKU, where he is currently a Distinguished Professor. From 1996 to 2009, he was also a Researcher with the Engineering and Technology Promotion Center, National Science Council, Tainan. From 1999 to 2002, he was the Director of the Electrical Laboratories with NCKU. From 2009 to 2012, he was the Dean of the College of Electrical Engineering and Computer Science with National United University, Miaoli City, Taiwan. From 2009 to 2016, he was the Vice President of the Federation of International Robot-Soccer Association. Since 2014, he has been the Director of the Center for Intelligent Robotics and Automation, NCKU. His current research interests include artificial and biological intelligence and applications, fuzzy system and control, home service robots, humanoid robots, mechatronics, 4WIS4WID vehicles, and singular perturbation methodology. Dr. Li was a recipient of the Outstanding Automatic Control Award in 2006 from the Chinese Automatic Control Society (CACS) in Taiwan and received the Outstanding Research Award in 2017 from the Ministry of Science and Technology, Taiwan. He was a Technical Editor of the IEEE/ASME *Transactions on Mechatronics* and an Associate Editor of the *Asia Journal of Control*. He is currently an Editor-in-Chief of *iRobotics*, and Associate Editors of the *International Journal of Electrical Engineering*, the *International Journal of Fuzzy Systems*, and the *IEEE Transactions on Cybernetics*. He was elected as the President of the CACS from 2008 to 2011 and the Robotics Society of Taiwan from 2012 to 2015. He was elevated to CACS Fellow and RST Fellow in 2008 and 2018, respectively.

Ping-Huan Kuo received the B.S., M.S., and Ph.D. degrees from the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, in 2008, 2010, and 2015, respectively. Since 2017, he has been with the Department of Intelligent Robotics, National Pingtung University, where he is currently an Assistant Professor. His major research interests include fuzzy control, intelligent algorithms, humanoid robot, image processing, robotic application, big data analysis, machine learning, deep learning applications.