การประชุมวิชาการทางวิศวกรรมไฟฟ้า ครั้งที่ 48
The 48th Electrical Engineering Conference (EECON-48)
วันที่ 19-21 พฤศจิกายน 2568 ณ โรงแรมฟูราม่า จังหวัดเชียงใหม่

# Enhanced Breast Cancer Classification Using an Ensemble Learning Model

**Sitthichat Meekhwan, Thanakarn Suangun*, Buntueng Yana**

Department of Electrical Engineering, School of Engineering, University of Phayao, Thailand,
mr.sitthichat.n@gmail.com, *thanakarn.su@up.ac.th, mr.buntueng@gmail.com

## Abstract

Breast cancer remains a significant global health crisis. And it is a leading cause of cancer-related mortality among women. The early and accurate diagnosis is a critical factor for improving patient prognosis and survival rates. While various machine learning classifiers have been applied to this problem, their performance has often been constrained by issues such as data imbalance and the limitations of individual algorithms. This study looks at these problems by creating and testing an ensemble learning method.

The WBCD dataset has 212 cancer samples and 357 benign samples to start with. To fix the imbalance in the cancer class, we first used the SMOTE technique to add to the dataset[1-4]. Subsequently, we apply t-SNE [5, 6] for dimensionality reduction, transforming the high-dimensional feature space into a lower-dimensional representation. This preprocessing step is intended not only to aid visualization but also to reduce model training and testing time. We propose an ensemble model that integrates K-NN, RF, and XGBoost as the foundation of the model, with LR serving as the meta-classifier to aggregate the model predictions.

We benchmarked the proposed model's performance against the individual base models. Experimental results demonstrate that the proposed ensemble model achieves superior classification accuracy. We achieve the accuracy at 98.50 percents. Furthermore, our findings confirm that the application of t-SNE significantly reduces the computational time required for training and testing. Ensemble models present a powerful and efficient model for breast cancer classification.

**Keywords:** Breast cancer classification, t-SNE, SMOTE, ensemble model

## 1. Introduction

Breast cancer is still a major health issue that kills a lot of women around the world. The World Health Organization (WHO) reports that every year, over 2.3 million new cases of breast cancer are found [7] and about 685,000 people die from it. The 2022 global cancer statistics report backs this up by showing that breast cancer is the most frequent type of cancer in women in 185 countries. It makes up 1.6% of all cancer cases and has a death rate of up to 6.9%[8]. This is a big problem for public health. Crucially, early diagnosis of breast cancer enables patients to receive prompt treatment. It is significantly reducing mortality rates.

Traditional breast cancer diagnosis often relies on the histopathological analysis of tissue samples by pathology experts. Even though this is still the standard way to do things, it can be hard and take a long time due of things like inter-observer variability. Different pathologists may look at the same sample and come to different results. Doing the same thing again can also make you fatigued, which could impair how accurate the diagnosis is. Machine Learning (ML) methods are an interesting way to improve this procedure. ML can be a very useful tool for doctors because it can quickly and objectively analyse complex data. The ML algorithms could be able to speed up the time it takes to make a diagnosis and make the results more reliable overall. So, the goal of this study is to create a better ML model that will make breast cancer categorisation more accurate and faster.

Many studies have investigated how different ML approaches may classify breast cancer. It is general knowledge that the accuracy rates that come out of this process depend a lot on the algorithms used and the specific features of the datasets used. Single-classifier models rarely work as well as they should. We use the Wisconsin Breast Cancer Diagnosis (WBCD) dataset to test how well our proposed method performs. This dataset is a common benchmark in the area, so we can compare it directly to other work.

## 2. literature review

To develop a robust classification model, two key stages are critical: effective data preprocessing and implementing advanced classification algorithms. This review examines seminal works in both areas, focusing first on dimensionality reduction and feature selection techniques. And second, we use the power of models that learn together.

### 2.1 Choosing Features and Reducing Dimensions

Medical diagnostics often use high-dimensional data, but it's hard to deal with because it makes calculations harder and raises the risk of overfitting. Before processing, one of the most important things to perform is to get rid of some data dimensions while maintaining the most important ones. A lot of individuals use Variable Importance Measures (VIM), Principal Component Analysis (PCA), and t-Distributed Stochastic Neighbour Embedding (t-SNE) to do this.

The PCA is a linear dimensionality reduction method that takes a high-dimensional feature space and turns it into a lower-dimensional features space by making a new collection of uncorrelated variables, or principal components, that keep as much of the original data's variance as workable. Haq et al. [9] looked at how well PCA, Relief, and auto-encoders worked for feature

*Corresponding Author: Thanakarn Suangun

selection. Their results revealed that Relief works well, but they also said that PCA is especially good for some models, such as linear Support Vector Machines (SVMs). Sahu et al. [10] used PCA to make high-dimensional data smaller before using additional methods. They concluded PCA is a crucial part of making data smaller, which makes the model run more efficiently.

VIM Another way to choose features is VIM, which finds the most important predictors in a dataset. It often makes use of the built-in features of a Random Forest (RF) algorithm. Huang et al. [11] used VIM based on the Gini index to quantify the importance of data characteristics. This technique made it possible to choose the most distinguishing aspects of breast cancer tumours. This made the dataset easier to work with and made their final model more accurate.

t-SNE allows visualization of a dataset in two or three dimensions by reducing its number of dimensions. The main goal of t-SNE is to show a complicated data structure in a space with fewer dimensions while keeping the integrity of the local neighbourhood. This method works well at showing hidden clusters. Neto et al. [5] successfully used t-SNE to reduce data dimensionality. This highlights its usefulness as an unsupervised method. Similarly, Mera et al. [6] used t-SNE with the objectives of reducing data complexity, preserving overall data structure, and visualizing data clusters for human interpretation. In their research, they employed t-SNE as a complementary analysis tool alongside Probabilistic Latent Semantic Analysis (PLSA), demonstrating its power in uncovering subtle relationships within a dataset.

### 2.2 Ensemble Learning for Enhanced Classification

When applying the ML technique classification task, there are two important steps: data pre-processing and learning model selection. While single classifiers are often effective, ensemble models have gained prominence for base models' ability to deliver superior performance. As Naseem et al. [12] state, the goal of using an ensemble model is to achieve performance that surpasses that of any individual constituent model. By combining classifiers, an ensemble can learn more complex patterns and produce results that are more robust and accurate. For instance, Naseem et al. [12] developed an ensemble combining SVM, Logistic Regression (LR), Naive Bayes (NB), Decision Tree (DT), and an Artificial Neural Network (ANN), achieving an impressive accuracy of 98.83%.

Ensemble voting, a common and powerful ensemble strategy, combines predictions from multiple individual models to produce an output prediction. Sahu et al. [10] employed an ensemble that combined the predictive capabilities of RF, SVM, LR, and Gradient Boosting (GB). In their final stage, a soft voting mechanism was used to aggregate predictions. This method operates by averaging the class probabilities from each base model to determine the final classification. A key strength of this approach is that soft voting inherently ensures that predictions from models with higher confidence. The model has a greater influence on the outcome. When compared to individual classifiers, this technique leads to improved generalization and a reduced risk of overfitting. Ultimately, combining diverse classifier types enhances the model's overall robustness. And this strategy makes it more resilient to the individual biases and errors of any single model

## 3. Methodology

The method of this study is systematically designed to build. As outlined in the workflow in Fig. 1, the process encompasses five key stages, from initial data handling to final comparative analysis.

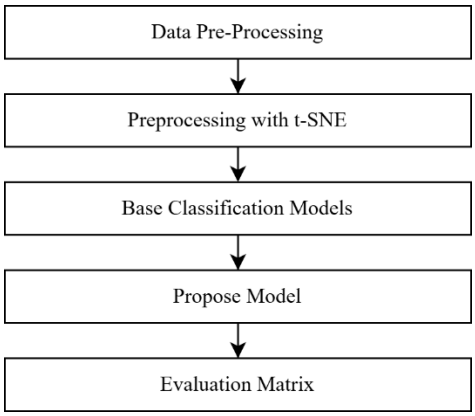


Fig. 1 Overview of the proposed methodology

We begin with a data preprocessing phase where the inherent class imbalance in the WBCD dataset is addressed using the Synthetic Minority Over-sampling Technique (SMOTE). The preprocessing is followed by the application of t-SNE for dimensionality reduction because we want to optimize the computation time and model accuracy. Next, a suite of diverse base models is trained and tested to establish a performance benchmark. These baseline results then provide the foundation for developing our proposed model: a novel ensemble that strategically combines the predictive power of top-performing base learners. To ensure the reliability of our findings, a stratified 10-fold cross-validation scheme is applied. Finally, the proposed model is compared with the base models using a range of standard evaluation metrics to quantify its effectiveness.

### 3.1 Data Pre-processing

This study utilizes the publicly available, WDBC dataset. The dataset comprises 569 samples, with 212 classified as malignant (M) and 357 as benign (B). While analyzing the dataset, the distribution reveals a class imbalance. The data distribution can introduce bias into the ML model and lead to unreliable predictions. To address this imbalance, we employ SMOTE, a widely adopted method, for this purpose. SMOTE works by generating new synthetic samples for the minority class based on the feature space similarities of existing minority samples. Following the best practices recommended by

Ahmad et al. [1] and Rahman et al. [2], the SMOTE procedure was applied only to the training data within each fold of our cross-validation process. This critical step prevents data leakage from the synthetic samples into the test set. While ensuring that our performance metrics provide an unbiased estimate of the model's generalization ability. This process balanced the classes and expanded the training set, creating a more robust foundation for model training.

### 3.2 Preprocessing with t-SNE

The original WDBC dataset contains 30 features, creating a high-dimensional space that is difficult for humans to visualize and can increase model computation time. To address this, we apply t-SNE for dimensionality reduction. The primary goal of using t-SNE is to convert the high-dimensional data into a low-dimensional space while preserving the local neighbourhood structure. This allows for effective visualization of the data. And it is possible to qualitatively assess the separability of the malignant and benign clusters, as shown in Fig. 2. A secondary aim is to investigate the impact of this dimensionality reduction on the model's training and testing time.

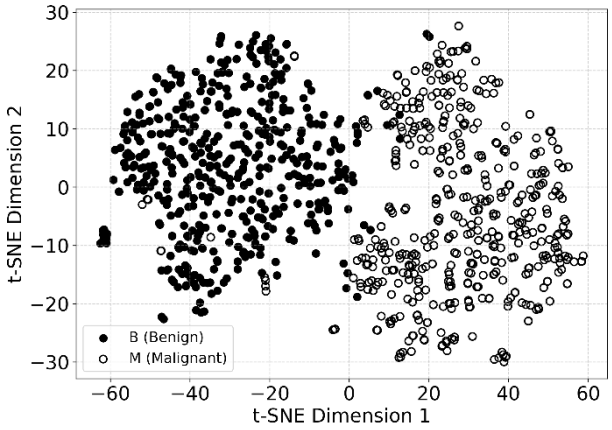


Fig. 2 Visualisation of the dataset after applying t-SNE

This graph shows the transformation from a high-dimensional space to a two-dimensional space.

### 3.3 Base Classification Models

To establish a performance baseline, we evaluated six well-established ML algorithms. We selected these models to represent a diverse range of learning strategies.

1. Support Vector Machine: It is a classification algorithm that finds an optimal hyperplane that maximizes the margin between two classes in the feature space.
2. Random Forest: An ensemble learning model comprising a multitude of decision trees. It operates by constructing trees on random subsets of the data and features, and outputs the class that is the mode of the classes' output by individual trees, improving accuracy and controlling for overfitting.
3. Extreme Gradient Boosting: An advanced and efficient implementation of the gradient boosting framework. It builds decision trees sequentially, where each new tree corrects the errors of the previous one, resulting in a highly accurate predictive model.
4. Logistic Regression: A linear model used for binary classification. It models the probability of a discrete outcome by passing a linear combination of the input features through a sigmoid function.
5. K-Nearest Neighbors: An instance-based, non-parametric algorithm. It classifies a new data point based on the majority class of its 'k' nearest neighbors in the feature space.
6. Artificial Neural Network: A computational model inspired by biological neural networks. It consists of interconnected layers of "neurons" that learn complex, non-linear relationships between input and output data

### 3.4 Propose Model

Building upon the base models, we propose an ensemble classifier to enhance predictive performance. The architecture of our proposed model is detailed in Fig. 3. Stacking is an ensemble technique that combines multiple classification models by training a meta-model to make the final prediction based on the predictions generated by the base learners.

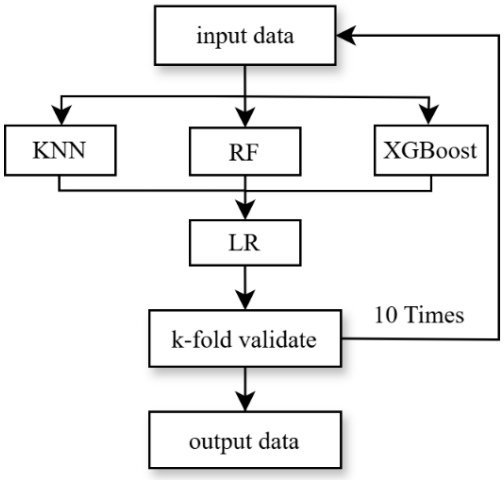


Fig. 3 Proposed model

Our model uses K-NN, RF, and XGBoost as the base learners. This selection intentionally combines a K-NN instance-based model with two powerful, rule-based tree ensembles: RF and XGBoost. The rationale is that their diverse learning approaches capture different aspects of the data, and their weaknesses can be mutually offset. For instance, while K-NN can be sensitive to noise, tree-based models like RF are more robust to such data.

The predictions from these three base models are then used as input features to train a meta-model. For this role, we selected LR because it is an excellent choice for a meta-model. It is computationally efficient, its simple

linear nature reduces the risk of overfitting on the base model predictions, and it is highly effective at learning the optimal weights to combine the predictions from the base learners.

### 3.5 Evaluation Matrix

To ensure the robustness and generalizability of our results, we employ a 10-fold cross-validation strategy, a standard and reliable technique in ML [13]. In this method, the dataset is partitioned into ten equal-sized subsamples. One subsample is retained as the test set, and the remaining nine are used for training. This process is repeated ten times, with each subsample used exactly once as the test data. The average of the results from all ten folds is the final performance. This method makes sure that the model's performance isn't affected by how the data is split at the start, and it gives a more accurate prediction of how well it will work on new data.

We use several standard evaluation metrics based on the model's parts to measure how well it classifies things. These are True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The matric that we measure in this work are accuracy, precision, sensitivity, and specification. These matric are computed using these equations.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{Specification} = \frac{TN}{TN+FP} \quad (4)$$

We recorded computational time as a key factor in evaluating the model during training and testing.

## 4. Results and Discussions

This section presents the performance evaluation of the proposed ensemble model against the six base models. We analyzed the results in two key areas: first, the overall classification performance on the original dataset, and second, the specific impact of using t-SNE for dimensionality reduction on both model accuracy and computational efficiency.

### 4.1 Comparative Classification Performance

The primary aim of this research was to develop a model with superior diagnostic accuracy. We trained and evaluated all models using a 10-fold cross-validation scheme. The comprehensive performance results, including accuracy, precision, sensitivity, specificity, and testing time, are summarized in Table 1.

Table 1 Comparison models on t-SNE dataset.

| MODEL | Acc (%) | Pre (%) | Sen (%) | Spe (%) | Testing time(ms) |
|---|---|---|---|---|---|
| SVM | 97.30 | 97.20 | 97.37 | 97.26 | 53 |
| RF | 98.00 | 98.02 | 97.96 | 97.98 | 81 |
| XG Boost | 97.60 | 97.84 | 97.36 | 97.59 | 62 |
| LR | 97.00 | 96.85 | 97.13 | 96.97 | 70 |
| K-NN | 97.10 | 97.17 | 96.88 | 97.01 | 88 |
| ANN | 96.30 | 95.48 | 97.54 | 96.48 | 43 |
| Propose | **98.50** | **98.23** | **98.76** | **98.49** | 81.3 |

Table 1 shows that the proposed model does the best on almost all the evaluation metrics. It had a top-notch accuracy of 98.50% and a sensitivity of 98.76%. This result is better than that of the best single classifiers, RF (98.00% accuracy) and XGBoost (97.60% accuracy). The model's very high sensitivity is very impressive because it means that it can reliably detect real malignant cases very well, which is very important in a clinical diagnostic scenario.

The superior performance of the proposed model can be attributed to its sophisticated stacking architecture. It effectively combines the strengths of diverse algorithms: the robustness of rule-based ensembles like RF and XGBoost and the local similarity detection of the instance-based K-NN model. The Logistic Regression meta-learner then learns the optimal way to weigh the predictions from these base models, creating a final decision that is more robust and accurate than any single model could achieve on its own.

### 4.2 Impact of t-SNE on Accuracy and Testing Time

A secondary objective was to investigate the trade-off between dimensionality reduction and performance. As illustrated in Figure 4, applying t-SNE to reduce the feature set from 30 to 2 resulted in a slight reduction in accuracy across all models.

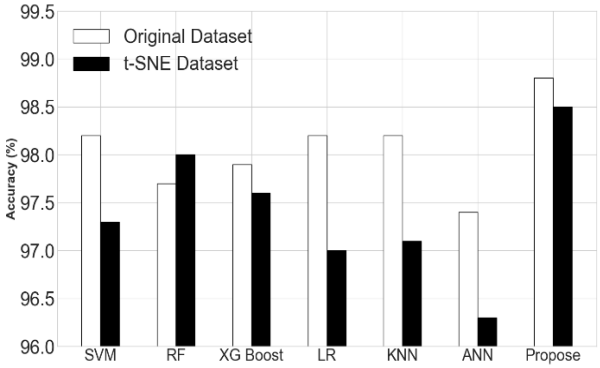


Fig. 4 Model cccuracy: original vs. t-SNE data

As expected, the dimensionality reduction process resulted in a slight loss of discriminative information and a corresponding decrease in accuracy for all models. However, the proposed model consistently outperformed the base models on both the original and the t-SNE-transformed data. Fig. 4 presents a bar graph comparing the accuracy of the proposed model against the base

models for both datasets, illustrating this superior performance even with the reduced feature set.

We also plot the training and testing time of all models before and after utilizing t-SNE as illustrated in Fig. 5 and Fig. 6, respectively.

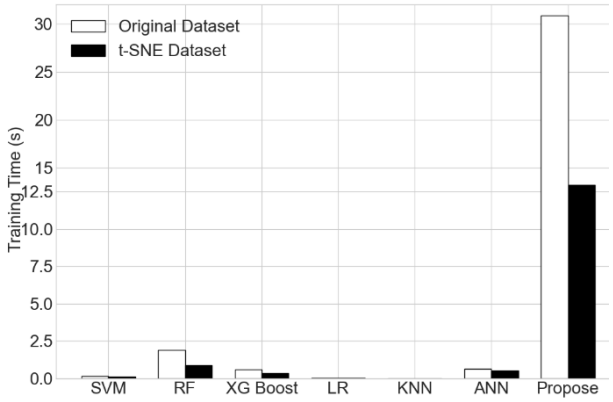


Fig. 5 Training time of before and after utilizing t-SNE

As expected, applying t-SNE to reduce the feature dimensions resulted in a significant decrease in computational time for all models. The results in Fig. 5 and Fig. 6 confirm that the t-SNE preprocessing step effectively reduces the computational overhead.

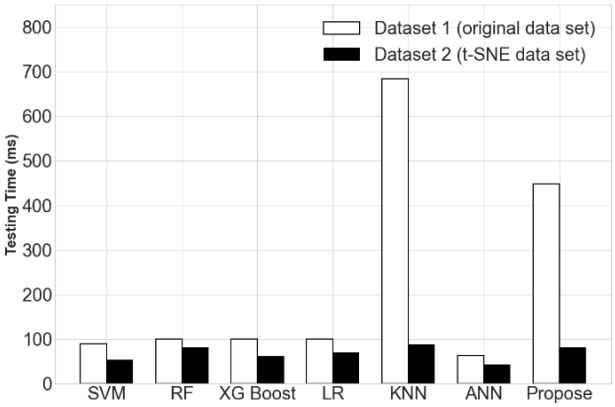


Fig. 6 Testing time of before and after utilising t-SNE

Conversely, the primary benefit of using t-SNE is a dramatic reduction in model training and testing time, as shown in Fig. 5 and Fig. 6. The K-NN method gets the biggest boost in computing efficiency. This is because K-NN's computational complexity depends a lot on how many dimensions the data has. By cutting the number of features from 30 to 2, the distance computations for each test point become much faster. This impact is less noticeable in tree-based models like RF and XGBoost since they mostly do value comparisons at nodes instead of distance computations in high-dimensional space. These results show that t-SNE is a very useful preprocessing step for applications where speed is very important.

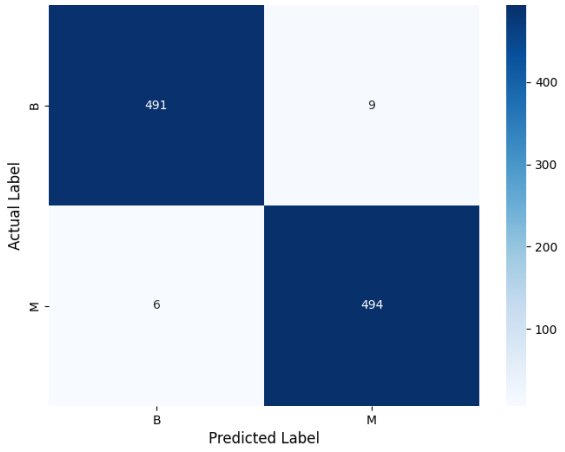


Fig 6 Confusion matrix of ensemble model

In Fig. 6 Confusion matrix of ensemble model is a confusion matrix of ensemble model for binary classification. It refers to performing the model to classify ALL.

## Conclusions

This research successfully developed and validated an ensemble model for breast cancer diagnosis that shows superior performance over several individual classifiers. The experimental results confirm two primary conclusions.

First, the proposed model, which integrates RF, XGBoost, and K-NN as base learners with a Logistic Regression meta-classifier, achieved an outstanding accuracy of 98.5%. This performance surpasses that of all evaluated base models, establishing our ensemble approach as a highly effective and accurate diagnostic tool. The model's high sensitivity underscores its potential clinical value in correctly identifying malignant cases.

Second, this study quantified the trade-off between accuracy and computational efficiency introduced by dimensionality reduction. Using t-SNE, we shortened the model's training and testing time by transforming the high-dimensional data into a low-dimensional space. This effect was strongest for the suggested ensemble model, which went from taking 448ms to test on the original dataset to 81.3ms on the t-SNE dataset. The ensemble model's performance was still better than the base models, even though it took less time and was slightly less accurate.

In short, our work shows that a strong ensemble model can accurately classify breast cancer. It also shows that t-SNE pretreatment can help find a great balance between high accuracy and much better computing efficiency. In the future, researchers could investigate using this model on other, larger medical imaging datasets and look into using additional non-linear dimensionality reduction methods.

## 5. References

1. Ahmad, A., et al. *Vehicle Recognition using Multi-Layer Perceptron and SMOTE Technique*. in *2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*. 2022.
2. Rahman, M.A., et al., *Enhancing Early Breast Cancer Detection Through Advanced Data Analysis.* IEEE Access, 2024. **12**: p. 161941-161953.
3. Anklesaria, S., et al. *Breast Cancer Prediction using Optimized Machine Learning Classifiers and Data Balancing Techniques*. in *2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA*. 2022.
4. Chen, H., et al., *Auxiliary Diagnosis of Breast Cancer Based on Machine Learning and Hybrid Strategy.* IEEE Access, 2023. **11**: p. 96374-96386.
5. Neto, A.C., A.L.M. Levada, and M.F.C. Haddad, *Supervised t-SNE for Metric Learning With Stochastic and Geodesic Distances.* IEEE Canadian Journal of Electrical and Computer Engineering, 2024. **47**(4): p. 199-205.
6. Mera, M., et al., *A Study of the Relationship Between Driving and Health Based on Large-Scale Data Analysis Using PLSA and t-SNE.* IEEE Access, 2024. **12**: p. 99614-99659.
7. Shandilya, S. and C. Chandankhede. *Survey on recent cancer classification systems for cancer diagnosis*. in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. 2017.
8. Bray, F., et al., *Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA: A Cancer Journal for Clinicians, 2024. **74**(3): p. 229-263.
9. Haq, A.U., et al., *Detection of Breast Cancer Through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques.* IEEE Access, 2021. **9**: p. 22090-22105.
10. Sahu, P.K. and T. Fatma, *Optimized Breast Cancer Classification Using PCA-LASSO Feature Selection and Ensemble Learning Strategies With Optuna Optimization.* IEEE Access, 2025. **13**: p. 35645-35661.
11. Huang, Z. and D. Chen, *A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm.* IEEE Access, 2022. **10**: p. 3284-3293.
12. Naseem, U., et al., *An Automatic Detection of Breast Cancer Diagnosis and Prognosis Based on Machine Learning Using Ensemble of Classifiers.* IEEE Access, 2022. **10**: p. 78242-78252.
13. Ni, X., et al., *Classification of Aviation Incident Causes using LGBM with Improved Cross-Validation.* Journal of Systems Engineering and Electronics, 2024. **35**(2): p. 396-405.

**Sitthichat meekhwan** received the bachelor's degree in electrical engineering from the University of Phayao (UP), Phayao, Thailand, in 2019. He is currently studying for a master's degree at the Faculty of Engineering. His current research interests in Machine learning, deep learning, and robotic

**Thanakarn Suangun** received B.Eng., M.Sc. in Communication Engineering, and Ph.D. degrees in Electrical Engineering from King Mongkut's University of Technology North Bangkok (KMUTNB), Thailand, in 2006, 2009, and 2020, respectively. Since 2012, he has been a lecturer at the Department of Electrical Engineering, School of Engineering, University of Phayao, Thailand. His research interests include RF/microwave, multiband small antennas for communication applications, and automation.

**Buntueng Yana** (Member, IEEE) received B.Eng. and M.Eng. degrees in Electrical Engineering from Chiang Mai University, Thailand, in 2006 and 2009, respectively, and a Ph.D. in Information Science and Technology from Osaka University, Japan, in 2019. He is currently a lecturer with the Department of Electrical Engineering at the University of Phayao, Thailand, where he began his academic career. His research interests include robotics, automatic control, power generation, and the application of machine learning to medical applications.