SCIPEDIA Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería

Revista Internacional Métodos numéricos para cálculo y diseño en ingeniería RIMNI CIMNE

Binary and Multi-Classification Models for Breast Cancer Diagnosis Using Automated Deep Learning and Mammography Images with Different Augmentation Cases

Nur Syafiqah Charim[1], Nadiah Ruza[2], Mohd Hafiz Arzmi[3,4] and Saiful Izzuan Hussain[1,*]

[1]Department of Mathematical Sciences, Faculty of Science & Technology, Universiti Kebangsaan Malaysia, Bangi, 43600, Malaysia

[2]School of Business and Economics, Universiti Putra Malaysia, Serdang, 43400, Malaysia

[3]Department of Fundamental Dental and Medical Sciences, Kulliyyah of Dentistry, International Islamic University Malaysia, Kuantan, 25200, Malaysia

[4]Cluster of Cancer Research Initiative IIUM (COCRII), Kulliyyah of Dentistry, International Islamic University Malaysia, Kuantan, 25200, Malaysia

ABSTRACT

Mammography is a very efficient medical imaging procedure that is used to detect and diagnose breast cancer. However, the use of mammography for the early detection and identification of cancer is very complicated and represents a considerable workload for radiologists. Machine learning (ML) can help address these challenges by providing accurate, automated diagnosis, but traditional ML methods are complex and resource-intensive. Google AutoML Vision offers a simplified approach, enabling healthcare professionals with minimal programming skills to develop effective diagnostic models. The aim of this study was to evaluate the ability of automated deep learning using mammography images using Google AutoML with different augmentation cases. In this work, two models were created: one for binary classification and another for multi-classification. The binary classification model includes two scenarios: non-cancerous and malignant, while the multi-classification approach includes three scenarios: normal, benign and malignant. The average accuracy of the two classifications was evaluated and compared. The average accuracy of the binary and multi-classification models was 77.98% and 79.29%, respectively. These results suggest that Google AutoML can simplify the use of ML models in the clinical setting and provide a reliable diagnostic tool that can reduce the workload of radiologists. This study shows that AutoML has the potential to streamline diagnostic workflows in healthcare and make machine learning more accessible and effective in medical practise.

OPEN ACCESS

Received: 01/08/2024

Accepted: 29/11/2024

DOI
10.23967/j.rimni.2025.10.56851

Keywords:
Breast cancer
deep learning
AutoML
malignant

*Correspondence: Saiful Izzuan Hussain (sih@ukm.edu.my).

1 Introduction

Breast cancer is a prevalent global health issue, with numerous new cases identified annually. According to the American Cancer Society, it is estimated that more than 310,720 women will be diagnosed with invasive breast cancer in the United States in 2024. In addition, there will be an estimated 56,500 cases of ductal carcinoma *in situ* (DCIS). Breast cancer is the most common cancer in women in the United States and accounts for approximately 30% of all newly diagnosed cancers in women each year. The lifetime prevalence of breast cancer in women in the United States is about 13%, which means that the chance of developing breast cancer is 1 in 8. The incidence of breast cancer is highest in middle-aged and older women, with the average age at diagnosis being 62. Occasionally, a small proportion of cases occur in women under the age of 45 [1].

In recent years, the prevalence of breast cancer has steadily increased with an annual growth rate of 0.6%. In women younger than 50, however, the increase is slightly higher at 1.0% per year. Although breast cancer incidence has risen, its mortality rate has steadily declined since 1989, decreasing by 42% as of 2021. The decrease in numbers is mainly due to early detection of the disease through screening, increased public awareness and advances in medical interventions [1].

Early detection of breast cancer is crucial for improving treatment outcomes and increasing survival rates. It improves the overall effectiveness of treatment and reduces the mortality rate associated with this common disease. Early-stage breast cancer is often manageable with less invasive treatments, such as lumpectomy instead of mastectomy, and many patients may avoid chemotherapy entirely. This not only improves patients' quality of life, but also reduces the physical and emotional burden associated with more intensive therapies. Detecting disease in its early stages is strongly associated with improved survival rates. When breast cancer is diagnosed before metastasis, the relative five-year survival rate is around 90%. This statistic emphasises the critical importance of early detection, as the chances of survival decrease significantly if the disease progresses and spreads to other parts of the body. Furthermore, early detection of breast cancer has been instrumental in reducing the mortality rate associated with this disease [2]. The mortality rate for breast cancer has fallen by 43% since 1989, largely due to the widespread introduction of screening programmes and increased public awareness. According to a study published in [3], the introduction of these programmes has helped to prevent more than 460,000 deaths from breast cancer. This emphasises the importance of early detection strategies in saving lives.

In general, breast tumours can be divided into two categories: benign and malignant. Benign tumours are non-threatening, while malignant tumours are dangerous because they have the potential to develop into cancer [4]. Due to the different prognoses and treatment options associated with the different tumour types, each tumour must be identified separately. An accurate diagnosis requires the precise identification of each breast cancer variant. Imaging techniques such as mammography, ultrasound, magnetic resonance imaging (MRI) and computed tomography (CT) are generally used to detect breast cancer and are performed by specialists such as radiologists, sinologists or pathologists. Deciphering the different patterns in mammography is a challenging task that requires a high degree of concentration, regularity and expertise. This results in misdiagnosis, misinterpretation of tumours, and failure to detect malignancies. However, there are ways to eliminate inconsistencies in the classification of breast cancer through the use of CAD diagnostic systems and machine learning [5].

Machine learning methods are increasingly utilized in intelligent healthcare systems, particularly for breast cancer detection. Machine learning can help doctors detect dangerous tumours such as breast tumours accurately, quickly and easily. An accurate diagnosis allows the doctor to treat the patient quickly to prevent death from breast cancer. Although there are many studies on the

effectiveness of machine learning techniques in the early detection of breast cancer, there is still more room to improve machine learning techniques for real-world applications as not many studies have been conducted on deep learning applications. This study was conducted to show that machine learning techniques are capable of achieving better results in the early detection of breast cancer.

A major obstacle to the use of machine learning in medical diagnostics is the complexity associated with the creation and application of these models. Traditional machine learning methods require extensive knowledge of data science, programming and expertise, which is a challenge for many healthcare organisations. The development process can require a significant investment of time and computer capacity to successfully train and evaluate models. To overcome these difficulties, we use Google AutoML, a tool specifically designed to streamline the machine learning process. Google AutoML provides a machine learning method for developing models that enables users with basic machine learning skills to create high-quality models. Google AutoML simplifies the process of model training by automating complicated operations such as feature selection, hyperparameter tuning and model evaluation. This accessibility makes it easier for clinicians and researchers to utilise advanced machine learning techniques.

The use of Google AutoML in this particular case is significant on several levels. First and foremost, it enables a greater number of healthcare providers to take advantage of new diagnostic technologies by providing them with accessible machine learning tools so that no specialised knowledge is required. It also improves the effectiveness and accuracy of breast cancer detection, leading to better outcomes for patients. Google AutoML allows medical professionals to focus more on patient care than technological intricacies by streamlining the development of robust diagnostic models. Using Google AutoML can lead to more consistent and reliable diagnostic results while reducing the potential unpredictability that can arise from human interpretation. Ensuring this consistency is crucial for the timely detection and treatment of breast cancer, ultimately leading to higher survival rates and improved quality of care for patients. This study aims to demonstrate the tangible benefits of using Google AutoML to automate machine learning in the timely detection of breast cancer. It illustrates the ability of this technology to completely transform healthcare diagnostic procedures.

To summarise, while mammography is an essential tool for breast cancer screening, it is often a huge burden for radiologists due to its complexity and the sheer volume of images to interpret. Traditional ML methods typically require extensive programming skills and significant computing resources, which can be a barrier for many healthcare organisations. This study explores the use of Google AutoML, an automated machine learning platform, to simplify the development of breast cancer diagnostic models from mammogram images. Google AutoML enables clinicians and healthcare professionals with minimal programming skills to develop robust deep learning models. This ease of use makes the benefits of machine learning accessible to a wider healthcare audience, enabling faster deployment and potentially better outcomes for patients. The study makes an important contribution by comparing binary and multi-classification models in the context of breast cancer detection, demonstrating the applicability of AutoML in a medical setting. In this way, this research advances the application of automated ML in healthcare diagnostics and shows how such technologies can transform medical workflows without the need for extensive technical knowledge.

2 Literature Review

Many researchers have conducted studies on the application of deep learning in mammography. Reference [6] investigated the recent developments in deep learning techniques applied to mammography for breast cancer detection. The review highlighted the challenges in integrating these technologies

into the clinical setting, such as the need to ensure privacy, improve model interpretability, and demonstrate generalizability. The potential of these models for medical image analysis is immense.

Several research papers have used deep learning techniques to examine mammograms, although there is still room for improvement in this area. Reference [7] used a Deep Convolutional Neural Network (CNN) and discovered that this approach effectively categorizes breast cancer when analysing mammogram images. Machine learning-based models have been reported to show high performance in detecting breast microcalcifications by mammography compared to other machine learning methods using semi-automatic segmentation methods [6]. Reference [8] compared the performance accuracy between CNN and CAD for breast cancer detection. CNN outperformed CAD at low sensitivity and is comparable to CAD at high sensitivity. Reference [9] used deep artificial neural network (ANN) methods to successfully detect cancer in mammography images with an accuracy equivalent to radiologists. Another study on deep learning algorithms using mammography images was conducted by [10]. This indicates a growing interest in deep learning to assist radiologists and improve the resolution of images acquired with mammograms [11,12]. Reference [13] showed that CNN deep learning methods can automatically detect nuances in mammogram images to distinguish between images with benign tumours that were previously detected as malignant tumours. This method can help ensure that computerized clinical devices do not perform incorrect examinations.

Reference [14] used deep learning techniques to improve the precision of breast cancer detection based on mammography images. Deep learning approaches were found to be inherently adaptive to achieve high accuracy on mammography platforms and showed significant potential in mitigating false positives and negatives from mammography. Reference [15] discovered that deep learning models applied to mammography images outperformed alternative methods in accurately assessing breast cancer risk. These results suggest that neural network models outperform other models in accurately categorizing benign and malignant cancers. Reference [16] suggested using EfficientNet within the CNN framework, which requires a minimal number of parameters. The main advantage of EfficientNet lies in its efficiency and precision in categorizing mammography images, which is due to its limited number of parameters. Experimental results show that EfficientNet achieves an overall accuracy of 86.5%. Reference [17] conducted a study in which they used deep learning methods to evaluate mammograms. Their study highlights the importance of data preparation and augmentation in achieving exceptional classification accuracy and demonstrates the sophisticated techniques used in this area.

AutoML has gained great popularity in smart healthcare systems due to its ease of use, lack of need for programming skills and ability to create ML models of exceptional quality. GoogleAutoML is capable of training deep learning models using large datasets. Reference [18] used Apple Create ML and Google Cloud AutoML in a series of clinical situations involving the lung and colon. The study evaluated and compared the two platforms based on their accuracy. The analysis showed that both platforms had exceptional accuracy and no statistically significant difference was found between them. Reference [19] compare the performance of Google AutoML with Apple CreateML in classifying animal images. The authors found that Google AutoML outperforms CreateML in terms of accuracy and average precision in various datasets, demonstrating its effectiveness in image classification tasks. In a study conducted by [20], the diagnosis of invasive ductal carcinoma, a type of breast cancer, was investigated using Google Cloud AutoML Vision. A total of 278,124 histopathological images were used, and another 378,215 photos were created by rotating the original image at a certain angle for modelling. The study achieved an average accuracy of 91.6%, surpassing the results of previous studies. Reference [21] used Google AutoML to recognize COVID-19 X-ray images and achieved an accuracy of up to 98.41% for both binary and multi-class classifications. The classifications included healthy,

pneumonia and COVID-19 categories. Accuracy was determined based on an evaluation of 1125 chest X-ray images.

In summary, studies have shown that Google AutoML can train robust deep learning models and offers a user-friendly interface that does not require extensive programming knowledge. Research has shown that Google AutoML performs competitively even when compared to traditional machine learning methods in several areas, including breast cancer, lung cancer and COVID-19 detection. Overall, the results suggest that AutoML platforms can significantly improve diagnostic accuracy and streamline medical workflows, ultimately improving patient outcomes and driving the adoption of AI in healthcare diagnostics.

3 Materials and Methods

3.1 Dataset

This study uses the INBreast dataset, as shown in Fig. 1. The INbreast database contains mammography images originally collected by the Breast Centre of the Centro Hospitalar de S. João (CHSJ) in Porto. The data collection ran from August 2008 to July 2010 and comprised 115 cases with a total of 410 images [22].

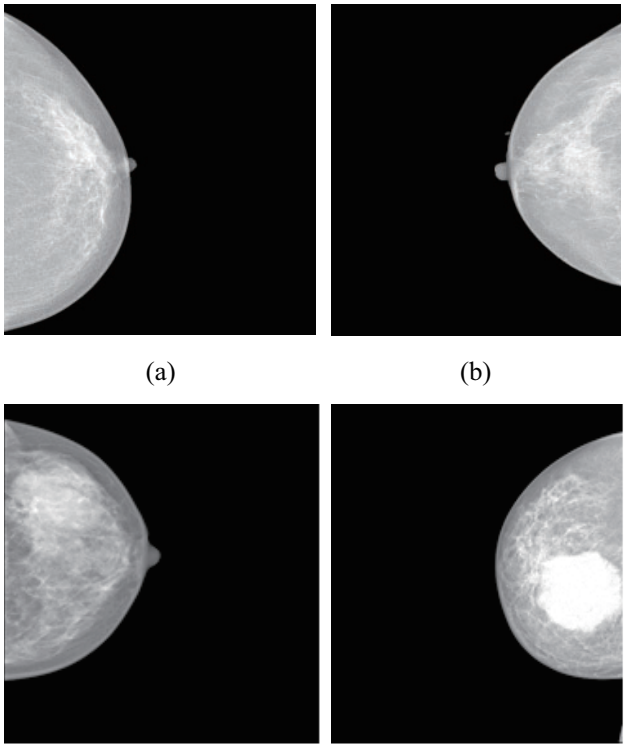

Figure 1: Examples of images used: (a) normal; (b) benign; (c) and (d) malignant

The database contained images of normal, benign and malignant mammograms. Two models were trained using data from these images. Model 1 consisted of 2050 images, where each 410 images were rotated by four angles: 45°, 90°, 180° and 270°. Model 2 consisted of 610 images where all 410 images were flipped horizontally. Other augmentation methods, such as contrast adjustment or the addition

of noise, were not used in this study because this is medical imaging and these transformations could alter important features that are critical for diagnosis. Mammography images contain subtle details that are critical for distinguishing between benign and malignant disease, and excessive augmentation could obscure these features. By limiting augmentation to geometric transformations, we ensured that the integrity of the original images was preserved while mitigating the effects of class imbalance.

Each model was evaluated with two scenarios, non-cancerous *vs*. malignant and non-cancerous *vs*. benign *vs*. normal. For the non-cancerous and malignant scenarios, Model 1 was trained on 1345 images of non-cancerous and 705 images of malignant, while Model 2 contained 402 images of non-cancerous and 208 images of malignant. For the second scenario, Model 1 consisted of 1010 images of benign, 705 images of malignant and 335 images of normal breast tissue, while Model 2 contained 202 images of benign, 208 images of malignant and 200 images of normal breast tissue.

3.2 Automated Deep Learning

Automated deep learning has three critical processes that need to be performed: hyperparameter optimization, combinatorial optimization, and transfer learning.

1. Hyperparameters of the machine learning model were adjusted by using hyperparameter optimization.
2. Performance was optimized by combining different features and machine learning models by using combinatorial optimization.
3. Model configuration was sped up by transferring learning data and parameters from one application to another by using transfer learning.

The performance of the model dataset in classifying cancer was tested using the Receiver Operating Characteristics (ROC) measure. The predictions of the dataset were evaluated using the free-response ROC (FROC) to assess the ability of the model to detect and accurately localize cancerous lesions. The FROC curve is a graphical representation that illustrates the relationship between sensitivity, which measures the proportion of correctly identified lesions, and the number of false positives.

3.3 Google AutoML

AutoML is an artificial intelligence method that utilizes the concept of deep learning. Deep learning is a special type of algorithm used in machine learning. In the study, Google Cloud AutoML Vision was selected as the primary platform for the development of breast cancer detection models. AutoML Vision is an image processing program that uses machine learning algorithms to detect and detect objects in images. It is available on the Google Cloud Platform [23]. AutoML Vision enables users with minimal expertise to create personalized deep learning models for specific tasks, leading to high-quality results. AutoML Vision uses standardized, pre-trained and ready-to-use AI applications, similar to Google's Vision API in its range of AI/ML solutions. Google AutoML Vision allows individuals without machine learning expertise to use their machine learning knowledge and skills to run a machine learning model. The cloud-based models are supported by state-of-the-art GPUs that allow the user to complete a customized task within a few hours, depending on the size of the dataset. This infrastructure ensures that the model training and deployment processes are both efficient and capable of handling large datasets, making it ideal for medical applications where high computational demands are common.

AutoML divides the dataset into three different default subsets to prepare the model for creation: 80% for preparation, 10% for validation and 10% for off-box checking. Changes to the default settings can be made at any time.

Integrating a machine learning model into the radiology workflow is a challenging endeavour. To develop an image processing model with machine learning, one must have advanced knowledge of programming languages such as Python and MATLAB. In addition, expertise in image recognition, data processing and predictive analytics approaches is essential. Learning and mastering these skills can be challenging for a radiologist, and it may be necessary to hire a graduate expert or technical assistant to handle routine clinical work.

The speed at which tasks are completed and the quality of computer equipment are additional barriers that should be emphasized and addressed. Machine learning models can be computationally intensive when run on local computer channels. Without the presence of graphics processors on a personal computer, the batch processing process can take several days to complete the task of classifying the photos. The creation and construction of a machine learning model presents an additional challenge when it comes to feature engineering and pre-processing the data. Feature engineering is used to eliminate certain components, such as data cleaning, noise filtering, blocking of elements and segmentation. The quality of the model depends on the performance of the functional technique and the accompanying features used in the software implementation. The presence of non-uniform boundaries in the objects to be measured and the moderate intensity of staining in the tissues are factors that contribute to the confusion of pathologists in the assessment of heterogeneity.

3.4 Model Built

The following measures were taken to develop the experimental breast cancer detection model. The training of the model was preceded by the creation of an image dataset. This was done by uploading the extended dataset to Google Cloud Storage:

1. The collection of images to be uploaded to the cloud storage was prepared. At least 100 images per label were required.

2. The images were imported into the Google Cloud AutoML Vision UI to create a dataset where all data was correctly labelled.

3. Training the data and selecting the process.

AutoML Vision automates the training process. The usual training approach for the dataset is to use 80% of the data for preparation, 10% for validation and 10% for evaluation. This process aims to identify the most effective algorithms for detecting breast anomalies in the training dataset. AutoML Vision applies validation and testing methods to datasets to evaluate the performance of the model.

3.5 Model Performance

The performance of the model was evaluated using three metrics: sensitivity, specificity and accuracy. These metrics were adopted to characterize a diagnostic test in machine learning and were used to quantify and understand the reliability and suitability of a model. Sensitivity measures the ability to accurately detect a positive disease by indicating the accuracy of disease prediction. High sensitivity refers to a low risk of patients being misclassified as normal due to misdiagnosis. Specificity is a statistical measure of the likelihood that people who do not have a disease will be correctly identified. In the context of medical diagnosis, while the reported value for specificity can be very high, the figures for sensitivity are of greater importance and significance. Accuracy refers to the exact

number of correctly predicted labels for the specified disease. It shows the ability of the model to accurately predict both diseases and non-diseases. To assess sensitivity, specificity and accuracy, it is necessary to understand the following terms: true positive (TP), true negative (TN), false negative (FN) and false positive (FP). When a patient is diagnosed with a disease and the diagnostic test confirms the presence of the disease, the expected results are called true positive (TP). TN stands for the absence of a disease in a patient and the agreement of the diagnostic tests with this observation. TP and TN show a direct correspondence between the expected results and the actual condition, which is called the standard of truthfulness. However, it is easy to make a false diagnosis during medical examinations. A false positive (FP) occurs when a diagnostic test indicates that a patient has a disease that is not present. Similarly, a false negative (FN) occurs when the diagnostic test gives a negative result even though the disease is present. False-positive (FP) and false-negative (FN) results indicate that the experimental results are in direct contrast to the actual situation.

The use of TP (true positive), TN (true negative), FN (false negative) and FP (false positive) in the quantification of sensitivity, specificity and accuracy is explained below.

Sensitivity = TP/(TP + FN) = (Total true predicted positive assessments)/(Total of all positive assessments)

Specificity = TN/(TN + FP) = (Total negative predicted assessments)/(Total of all negative assessments)

Accuracy = (TN + TP)/(TN + TP + FN + FP) = (Total correct predicted assessments)/Total of all assessments)

4 Results and Discussion

4.1 Model Evaluation

Several experiments were performed to identify and categorize the different forms of breast cancer using mammography images in two different models: binary classification and multi-classification models. The binary classification strategy was used to create automated deep learning models that distinguish between two scenarios: non-cancerous and malignant. The multi-classification technique was used to distinguish between three scenarios: normal, benign and malignant. The execution of all tasks was performed using Google AutoML Vision. An evaluation was performed to determine the effectiveness of the proposed approach for both binary classification and multi-classification. Of all submitted mammography images, 80% were used for preparation, 10% for analysis and the remaining 10% for testing. A confidence threshold of 0.5 was used in this study.

The confidence threshold represents the certainty of the model to the expected probability of detecting the selected category. The scoring technique assesses the predictive accuracy of the test pattern at many levels, using a compatibility threshold and assigning a score between 0.0 and 1.0. Figs. 2 and 3 show a graphical representation of the accuracy and recognition scores derived from previous computational models. Each point on the curve represents a specific accuracy and recognition value at a confidence threshold. Fig. 2 shows the integral of the precision-recall curve for the binary classification model. The scenarios for (a) and (b) include 2050 images and compare non-cancerous cases with malignant cases. Similarly, scenarios (c) and (d) use 610 images and also compare non-cancerous cases with malignant cases.

Fig. 3 shows the area under the precision-recall curve for the multi-classification model. The scenarios for (a) and (b) are normal *vs*. benign *vs*. malignant for 2050 images, while (c) and (d) are normal *vs*. benign *vs*. malignant for 610 images.

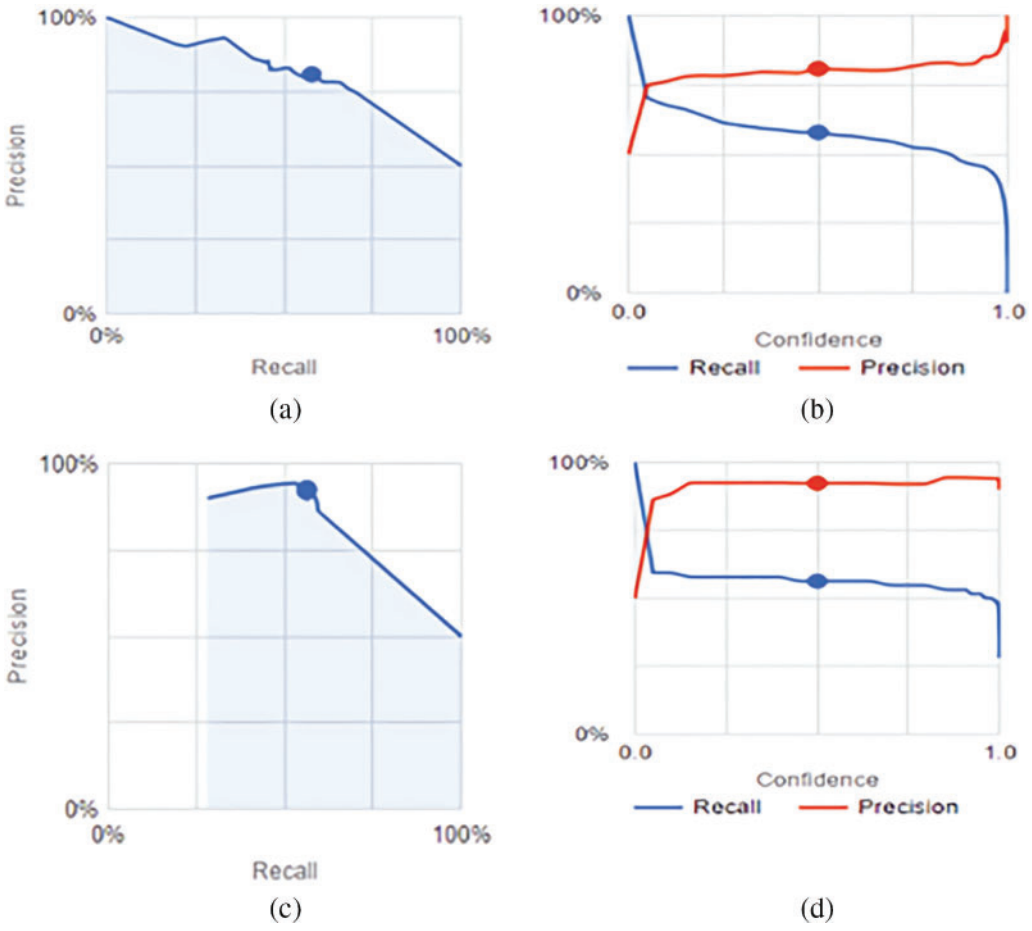

Figure 2: Area under precision-recall curve and precision-recall curve for binary models

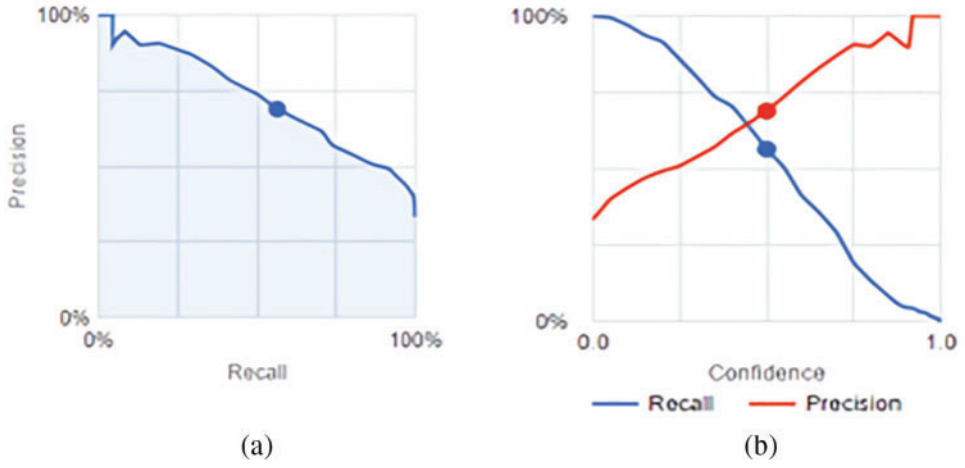

Figure 3: (Continued)

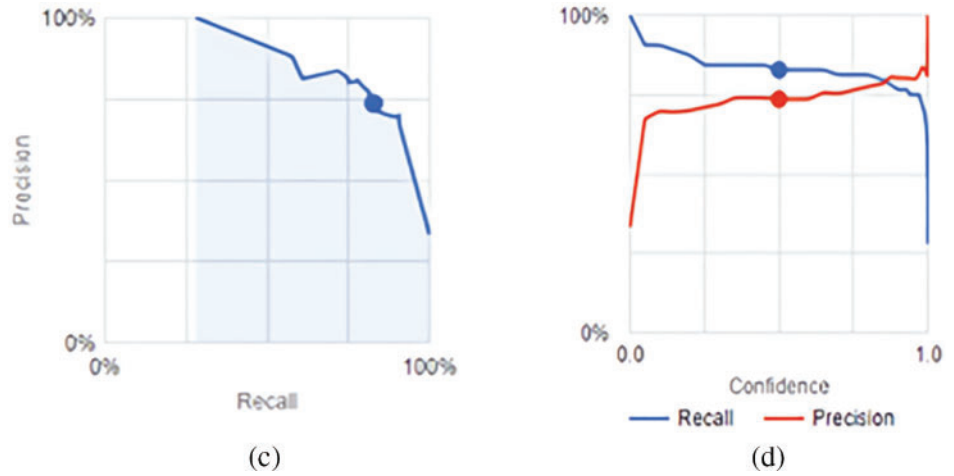

Figure 3: Area under precision-recall curve and precision-recall curve for multi-classification model

The average precision of the models was assessed using the area under the precision-recall curve (AuPRC). To obtain a more detailed assessment of performance, confusion matrix analyses were performed for both models. Fig. 4 shows the confusion matrices for both the binary and multi-class classification models.

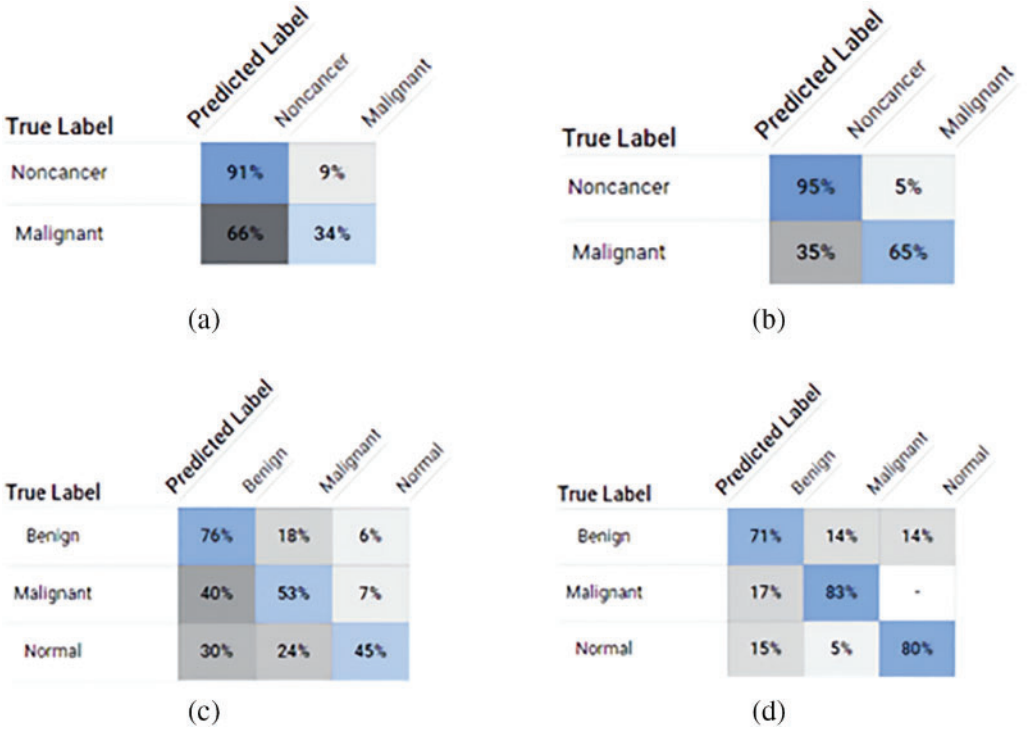

Figure 4: Confusion matrix for (a) and (b) non-cancer *vs*. malignant; (c) and (d) benign *vs*. malignant *vs*. normal

In all models, all classes were classified into the correct category with a probability of at least more than 34%. In model 1, the classification between non-cancerous and malignant was the most difficult. 66% of malignant classifications were likely to be classified as non-cancerous, while 9% of non-cancerous cases were incorrectly classified as malignant. For the two-scenario (and) model, the prediction of Model 2 (Fig. 4b) was higher than that of Model 1 (Fig. 4a), where only 35% of malignant classifications were classified as non-malignant.

The confusion matrix in Fig. 4c and d contains three scenarios, namely benign, normal and malignant. In Model 1 with three scenarios, 18% of the benign category was misclassified as malignant and 8% as normal. 40% of the malignant category was misclassified as benign and 7% as normal. In the normal category, 30% was classified as benign and 24% as malignant. In Model 2 with three scenarios, 14% of the benign category was misclassified as malignant and 14% as normal. 17% of the malignant category was classified as benign, while 15% and 5% of the normal category were classified as benign and malignant, respectively. This shows that Model 2 had a higher precision than Model 1. The precision and recall results of the binary and multi-classification models are shown in Table 1. Precision is defined as the actual correct and correctly predicted items. A model with high precision would be associated with fewer false positives. A model with high recall, on the other hand, would be associated with fewer false negatives.

Table 1: Precision and recall for each AutoML-based model

Model	Scenarios	Precision	Recall
1	Non-cancer *vs*. malignant	0.8082	0.5784
	Benign *vs*. malignant *vs*. normal	0.6886	0.5637
2	Non-cancer *vs*. malignant	0.9231	0.5625
	Benign *vs*. malignant *vs*. normal	0.7361	0.8281

Sensitivity, specificity, precision and accuracy for the binary and multi-classification models were calculated from the AutoML results for precision and recall and presented in Table 2.

Table 2: Sensitivity, specificity, and accuracy classification for each scenario

Model	Scenario	Sensitivity	Specificity	Accuracy
1	Non-cancer *vs*. malignant	0.6667	0.7262	0.7157
	Benign *vs*. malignant *vs*. normal	0.5873	0.7660	0.7108
2	Non-cancer *vs*. malignant	0.8824	0.8298	0.8438
	Benign *vs*. malignant *vs*. normal	0.8261	0.9024	0.8750

The accuracy of Models 1 and 2 for the two scenarios, i.e., non-cancerous and malignant, respectively, was 71.57% and 84.38%. This shows that Model 2 was more accurate for two scenarios compared to Model 1 for two scenarios in discriminating between breasts without cancer and breasts with malignant tumours. The sensitivity value of Model 2 was 88.24%, which was higher than 66.67% for Model 1 with two scenarios. The accuracy of Models 1 and 2 with three scenarios, i.e., benign *vs*. malignant *vs*. normal, was 71.08% and 87.50%, respectively. Model 2 with three scenarios had the highest accuracy among the tested models in discriminating between normal breasts and breasts with benign and malignant lumps.

The sensitivity of this model was the highest among the three-scenario models, reaching 82.61%. According to the results obtained, all models had an accuracy of over 71%, with the highest accuracy reaching 87.50%. The results suggest that the models have the potential to be used in hospitals or health centres to detect breast cancer, reducing the burden on physicians and speeding up the processing of medical images. Although the accuracy of these models did not exceed 90%, their potential for clinical application in the detection of breast tumours is significant.

Based on the data collected, all models had an accuracy of over 71%, with the highest accuracy being 87.50%. These results suggest that the models can be effectively used in medical settings such as hospitals or health centres to identify breast cancer. This would reduce the workload of clinicians and reduce the time spent on image processing. Although the accuracy of the models does not exceed 90%, they are very promising for practical use in breast cancer screening.

The lack of better accuracy may be related to the quality and quantity of raw data. Accurate and properly annotated data of the highest calibre is crucial for the efficient training of robust machine learning models. How the data is enriched also plays an important role in improving the performance of the model. For example, improving the dataset by applying rotation to the photos, as shown in previous studies, can significantly increase the performance of the model by providing a wider range of training examples. To improve the precision and reliability of these models, future research should prioritize addressing the limitations associated with the data and optimizing data augmentation techniques. This will improve the seamless integration of these models into clinical practice.

5 Conclusion

Breast cancer diagnoses continue to rise globally, underscoring the need for timely and accurate detection methods. Early diagnosis is critical to reducing treatment duration, preventing disease progression, and lowering mortality rates. A variety of models have been tested to differentiate between people with and without cancer using deep learning techniques. For the diagnosis of breast cancer, mammography images of patients with breast cancer are used in our study, as these images are easily accessible and affordable. CT scans, while effective, are costly and less accessible, especially in rural or underserved areas. CT scans are usually only available in large hospitals and healthcare facilities in urban areas. In addition, CT scans expose patients to a higher radiation dose compared to mammography. Therefore, using a deep learning method with mammography images is more favourable as it is relatively easy to implement compared to CT scans.

Deep learning models using mammography images have been used on the Google Cloud Platform to successfully identify and distinguish different types of breast tumours. Our study showed that the multi-classification models have higher accuracy in discriminating between malignant and benign tumours as well as cancer-free patients compared to the binary classification models. The AutoML model shows significant potential to streamline workflows, ease clinician workloads, enhance diagnostic accuracy, and improve patient care outcomes. Future research should aim to enhance the demographic representation of the AutoML model by incorporating data from diverse patient populations. This can be achieved by expanding the image datasets for each label, thereby increasing the precision of the model. In addition, a comparative analysis between the use of CT scans and mammography images can be performed to evaluate their respective effectiveness.

However, the model's performance is constrained by the limited availability of images, particularly of patients with normal breasts and malignant tumours. It is important to emphasise this limitation because the accuracy of the model improves as more photos are collected for each label. Further

investigation, especially with data from breast cancer patients, is needed for the present study and future research endeavours.

Future studies should investigate advanced augmentation techniques to better reflect the variability observed in clinical practice. Including images with different patient demographics, breast density types and tumour characteristics will improve the generalisability and robustness of the model and ensure that it works effectively in different clinical scenarios. Furthermore, addressing the class imbalance is crucial for improving the model's ability to accurately detect malignant cases. Techniques such as cost-sensitive learning or synthetic data generation methods such as SMOTE could be used for this purpose. These approaches would help to mitigate the imbalance and their combination with over-sampling techniques or the integration of datasets with a higher number of malignant cases could further improve the performance of the model.

Future studies should also explore more sophisticated augmentation techniques. Adjusting variables such as contrast, brightness or noise could better simulate the variability in clinical practice and provide the model with a wider range of training examples. In addition, the use of customised deep learning models or AutoML platforms that provide more control over the hyperparameters would allow for more precise tuning of the model for specific use cases. In addition, external validation with independent datasets from different medical institutions is essential to ensure the applicability of the model in practice. Extending the validation process to other imaging modalities such as ultrasound or MRI would also test the adaptability of the model.

Comparing AutoML-based models with traditional deep learning approaches such as CNNs, DNNs or transfer learning will provide valuable insights into the strengths and limitations of each method. This will highlight how AutoML compares to more complex, manually developed models in terms of both accuracy and computational efficiency. Finally, the inclusion of explanatory methods, such as explainable AI (XAI), in future studies will allow clinicians to understand how the model makes its predictions. This transparency is crucial to increase confidence in AI systems in medical diagnostics and ensure their successful adoption in clinical settings.

Despite the limitations of model customisation and fine-tuning, AutoML effectively simplifies the implementation of deep learning models without the need for extensive programming knowledge. This indicates a significant potential of AutoML for medical applications, especially for healthcare professionals who may not have technical knowledge of programming or machine learning. Overall, the study emphasises the potential of automated machine learning platforms to improve diagnostic processes and facilitate access to advanced AI tools in the clinical setting.

Acknowledgement: We would like to extend our sincere gratitude to the Ministry of Higher Education (MOHE) for funding this research through the Fundamental Research Grants Scheme. We also thank the editor and reviewers of RIMNI for their valuable feedback and insights, which have greatly enhanced the quality of this manuscript.

Funding Statement: This research was funded by the Ministry of Higher Education (MOHE) through the Fundamental Research Grants Scheme, under grant number FRGS/1/2020/STG06/UKM/03/1.

Author Contributions: Saiful Izzuan Hussain: Conceptualization, methodology, writing—review & editing, and supervision. Nur Syafiqah Charim: Data curation, analysis, and writing—review & editing the original draft. Nadiah Ruza: Validation, writing—review & editing. Mohd Hafiz Arzmi: Validation, writing—review & editing. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly accessible and can be referred to via DOI: 10.1016/j.acra.2011.09.014.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. American Cancer Society. How common is breast cancer? May 23, 2024. [cited 2024 Nov 28]. Available from: https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html.
2. Elmore JG, Armstrong K, Lehman CD, Fletcher SW. Screening for breast cancer. JAMA. 2005 Mar;293(10):1245–56. doi:10.1001/jama.293.10.1245.
3. World Health Organization. WHO launches new roadmap on breast cancer. 2023 Feb 3. [cited 2024 Nov 28]. Available from: https://www.who.int/news/item/03-02-2023-who-launches-new-roadmap-on-breast-cancer.
4. Azman BM, Hussain SI, S I, Azmi NA, Abd Ghani MZ, Norlen NID. Prediction of distant recurrence in breast cancer using a deep neural network. Métodos numéricos para cálculo y diseño en ingeniería: Revista internacional. 2022;38(1):1–11.
5. Sadaf A, Crystal P, Scaranelo A, Helbich T. Performance of computer-aided detection applied to full-field digital mammography in detection of breast cancers. Eur J Radiol. 2011 Mar;77(3):457–61. doi:10.1016/j.ejrad.2009.08.024.
6. Wang J, Yang X, Cai H, Tan W, Jin C, Li L. Discrimination of breast cancer with microcalcifications on mammography by deep learning. Sci Rep. 2016 Jan;6(1):1–9. doi:10.1038/srep27327.
7. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. Sci Rep. 2018 Jan;8(1):1–7. doi:10.1038/s41598-018-19811-z.
8. Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, et al. Large scale deep learning for computer-aided detection of mammographic lesions. Med Image Anal. 2017 Jan;35:303–12. doi:10.1016/j.media.2016.07.007.
9. Becker Magda Marcon AS, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol. 2017 Jul;52(7):434–40. doi:10.1097/RLI.0000000000000361.
10. Kim EK, Kim HE, Han K, Kang BJ, Sohn YM, Woo OH, et al. Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study. Sci Rep. 2018 Jan;8(1):1–8. doi:10.1038/s41598-018-19939-y.
11. Hamidinekoo A, Denton E, Rampun A, Honnor K, Zwiggelaar R. Deep learning in mammography and breast histology: an overview and future trends. Med Image Anal. 2018 Jan;47:45–67. doi:10.1016/j.media.2018.03.006.
12. Burt JR, Torosdagli N, Khosravan N, RaviPrakash H, Mortazi A, Tissavirasingham F, et al. Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. Brit J Radiol. 2018;91(1089):20170545. doi:10.1259/bjr.20170545.
13. Aboutalib SS, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S. Deep learning to distinguish recalled but benign mammography images in breast cancer screening. Clin Cancer Res. 2018 Dec 1;24(23):5902–9. doi:10.1158/1078-0432.CCR-18-1115.
14. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep learning to improve breast cancer detection on screening mammography. Sci Rep. 2019 Jan;9(1):1–12. doi:10.1038/s41598-019-41592-0.
15. Yala Constance Lehman A, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology. 2019 Jul;292(1):60–6. doi:10.1148/radiol.2019181517.

16. Ruza N, Hussain SI, Mohamed SKC, Arzmi MH. Early detection of breast cancer in mammograms using the lightweight modification of EfficientNet B3. Métodos numéricos para cálculo y diseño en ingeniería: Revista internacional. 2023;39(3):1–11.
17. Paraschiv E, Stanciu A. Applying deep learning methods for mammography analysis and breast cancer detection. Appl Sci. 2023 Mar;13(7):Art. no. 4272. doi:10.3390/app13074272.
18. Borkowski AA, Wilson CP, Borkowski SA, Thomas LB, Deland LA, Grewe SJ, et al. Google auto ML versus apple create ML for histopathologic cancer diagnosis: which algorithms are better? J Pathol Inform. 2019 Jan;10(1):1–10. doi:10.4103/jpi.jpi_34_19.
19. Wang T, Zhu Y. The create ML-based compared to the AutoML-based approach to animal image classification. In: Proceedings of the International Conference on Cloud Computing, Performance Computing, and Deep Learning, 2022; San Diego, CA; vol. 12287, p. 412–8. doi:10.1117/12.2624988.
20. Zeng Y, Zhang J. A machine learning model for detecting invasive ductal carcinoma with google cloud AutoML vision. Comput Biol Med. 2020 Jan;122:103861. doi:10.1016/j.compbiomed.2019.103861.
21. Hussain SI, Ruza N. Automated deep learning of COVID-19 and pneumonia detection using Google AutoML. Intell Autom Soft Comput. 2022;31(2):1–6. doi:10.32604/iasc.2022.015207.
22. Moreira C, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. Acad Radiol. 2012 Feb;19(2):236–48. doi:10.1016/j.acra.2011.09.014.
23. Google. Cloud vision AI. Google Cloud. [cited 2024 Nov 28]. Available from: https://cloud.google.com/vision.