Explainable Hybrid Deep Learning for Automated Diagnosis of Canine Mammary Tumors

Elham Shawky Salama*
Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt

Heba Askr*
Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt

Ashraf Darwish*
Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt

Aboul Ella Hassanien*
Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt

*Scientific Research School of Egypt (SRSEG), https://egyptscience-srge.com/

ABSTRACT
Automatic classification of canine mammary tumors (CMT) plays a vital role in ensuring accurate diagnosis, reliable predictive evaluation, and reducing manual intervention in in the diagnostic process. This paper introduces an Explainable Artificial Intelligence (XAI) system for automated CMT classification, which combines the DenseNet201 deep learning (DL) architecture for feature extraction with a Random Forest (RF) classifier to distinguish between benign and malignant tumors in CMT images. The proposed system follows a structured four-phase pipeline: (1) data acquisition and Laplacian filter preprocessing to enhance tumor edges; (2) feature extraction using DenseNet201 and classification with a Random Forest (RF) model; (3) testing and evaluation of the proposed system; and (4) interpretability analysis using the Shapley Additive exPlanations (SHAP) XAI technique. Experimental results demonstrate that the DenseNet201-RF hybrid model achieves an accuracy of 93.9%, surpassing benchmark classifiers while offering interpretability for clinical validation. The proposed system enables accurate automatic classification of canine mammary tumors while integrating SHAP-based XAI for interpretability.

General Terms
Machine Learning, Animals, Tumors.

Keywords
Canine mammary tumors, Deep learning, DenseNet201, Explainable artificial intelligence, Random Forest, SHAP.

1. INTRODUCTION
Cancer is one of the most dangerous diseases, resulting from abnormal cells growing in an uncontrolled manner, altering the functioning of normal cells and generating abnormal cell proliferation and uncontrolled growth, causing significant morbidity and mortality [1]. When the tumor invades other parts of the body or spreads to nearby tissues, it is referred to as cancer or malignant [2]. In female dogs, mammary tumors are the most common neoplasm and are regarded as a significant public health issue. They affect numerous mammalian species. The death rate for CMT is over three times higher than that of humans. However, if breast cancer is diagnosed early, it has a better chance of being curable. Understanding how breast cancer is diagnosed in various animals might aid in our comprehension of the pathophysiology of this complex disease [3], and [4].

One of the most common forms of cancer in female dogs is CMT and it presents a serious clinical challenge in veterinary medicine. Studies in animal science have shown that the annual incidence of CMT in dogs varies across regions [5]. In California, USA, there were 145 incidences reported per 100,000 female canines. In the UK, it was 1340.7 per 100,000 dogs. In mainland China, it was 46.71%. CMT make up 54% of all tumors in animals. Because of reproductive health initiatives, the occurrence of CMT has declined in developed nations [6]. Early disease detection can help doctors begin treatment earlier, resulting in disease control to the greatest extent possible, ultimately saving the lives of suffering animals. The predominant technique for identifying malignant cells is the tissue sample analysis, which employs a specialized needle to extract a core sample from the suspicious area, yielding histopathological images. Malignant cells can be detected with the aid of appropriate image analysis. Nevertheless, the human process of interpreting these images is considered a gap which takes a lot of time, and the precision may differ based on the medical professionals' backgrounds. To address the gap, researchers are developing computer-aided diagnostic tools to reduce prognosis time and get superior outcomes more rapidly. These tools improve classification accuracy by minimizing interpretative inter-variability [7].

CMT are prevalent in female dogs, with 50% being malignant. World Health Organization (WHO) classified 18.25% CMT by histological type. The predominant varieties of CMT are adenocarcinoma, papillary carcinoma, solid carcinoma, complicated carcinoma, and carcinosarcoma. The non-malignant breast tumors mostly consist of fibroadenomas, ductal papillomas, benign mixed tumors, and simple adenomas. Often, many tumor forms are seen in different mammary glands of the same canine [8]. The primary factors affecting the incidence and advancement of CMT are breed, age, genetic predisposition, hormonal levels, diet, and cyclooxygenase-2 expression, as illustrated in Figure 1. CMT can occur in any dog breed; however, purebred dogs exhibit a higher prevalence. Specific breeds, including Cocker Spaniels, Poodles, and German Shepherds, are prone to the development of CMT. Canines aged nine years and older are at an increased risk of developing malignant tumors. Genetic mutations in genes such as BRCA1/2 and p53 significantly increase the probability of CMT development. Hormonal variables, including elevated levels of estrogen and progesterone, as well as more frequent estrous cycles, substantially elevate the risk. Dietary elements, such as increased concentrations of insulin-like growth factors and other hormones, facilitate tumor advancement and malignancy.

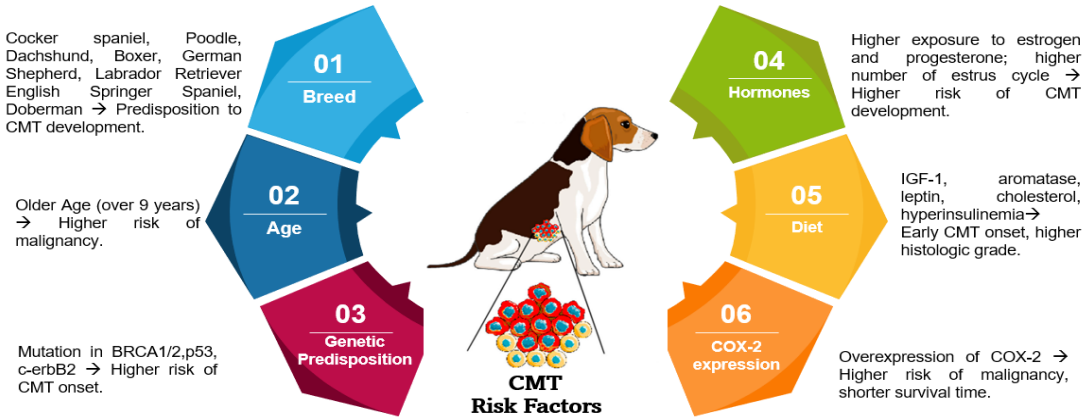


Fig 1: Main risk factors of canine mammary tumors

Finally, the overexpression of COX-2 correlates with increased tumor malignancy and decreased survival duration, rendering these aspects essential for comprehending and mitigating CMT risk [9]. DL has emerged as a revolutionary computer-aided diagnostic tool in the medical field, transforming the detection, analysis, and diagnosis of diseases [10] and [11]. DL uses neural networks to analyze large datasets, enabling the automated assessment of medical imaging with remarkable precision and rapidity. DL exhibits its effectiveness in pattern identification, anomaly detection, and disease prediction, often surpassing human expertise. Furthermore, DL facilitates real-time diagnosis, tailored treatment plans, and predictive analytics by combining data from multiple sources, such as genetic data and electronic medical records. By analyzing massive datasets from diagnostic imaging modalities such as ultrasonography, X-rays, and histopathology slides, DL algorithms may accurately identify tumors, classify malignancies, and differentiate between benign and malignant growths [12].

The existing literature indicates that the lack of extensive CMT image datasets has led to a limited number of publications on the application of DL in CMT identification, highlighting the urgent need for larger CMT datasets. In addition, although XAI techniques have started to become popular in more general DL applications, there is still little evidence of their incorporation into CMT detection models. The interpretability and reliability of these models, which are essential for their use in clinical settings, are restricted by the lack of XAI in the literature [13]. To mitigate this gap, this paper presents an explainable DL model for CMT classification having the following contributions:

1. The paper introduces a hybrid DenseNet201-RF that utilizes the power of the pretrained models to address the feature extraction and Random Forest (RF) for classification in the CMT domain.
2. This paper employs SHAP as an XAI technique to enhance the interpretability and explainability of the proposed model's results, effectively addressing the black-box nature of DL predictions, particularly in the context of CMT classification, and addressing a gap in the current literature.
3. This paper utilizes data augmentation strategy to present a variety of CMT cases, addressing the limitations in dataset size for CMT research.
4. This paper introduces the Laplacian filter as a preprocessing stage for input CMT images to enhance the classification performance.
5. The proposed model exhibits superior performance relative to other competitive pretrained and state-of-the-art models, positioning it as a highly promising solution for CMT classification challenges.

The following is the structure of the remaining sections of this work. Section 2 summarizes the literature review on CMT using DL. The proposed model methodology is introduced in Section 3. Section 4 introduces the results. Section 5 introduces the comparison with literature. Finally, Section 6 presents the conclusion and future work.

2. RELATED WORK

This section provides a review of the current literature on CMT prediction, along with an evaluation of the limitations inherent in each research. Kumar et al. [14] presented the first research in the literature for the binary classification of CMT using a dataset of CMT histological images. For the binary classification of CMT, the model surpassed most cutting-edge techniques and maintained 93% for binary classification of CMT, but their work was based only on 352 images which did not reflect the different classes of the CMT.

Aubreville et al. [15] presented a new dataset of 21 canine mammary carcinoma (CMC) whole-slide images (WSIs) that were thoroughly annotated for mitotic figures (MF). A multi-expert review procedure carefully annotated the dataset, which contains 13,907 MF, and 36,379 hard negatives. Machine learning (ML) was used to refine the annotations and find previously missed MF. Furthermore, two-dimensional projection and representation learning were used to improve annotation consistency. ResNet attained a mean F1 score of 0.791 on the test set. Burrai et al. [16] evaluated the application of pretrained DL models in conjunction with several ML classifiers for the classification of CMT cases. The researchers utilized a dataset of 1,056 images from 20 benign and 24 malignant CMT. The EfficientNet model achieved an accuracy between 0.82 and 0.85.

The current literature indicates that existing research on DL models for the identification and classification of CMT is insufficient, revealing considerable research gaps. In addition, the limited amount of the datasets utilized in the CMT research inadequately represent the diversity and enough CMT classes. Kumar et al. [14] employed merely 352 images, while Isik et al. [17] utilized only 146 images, both of which are insufficient for depicting the complexities of CMT in actual applications. Likewise, Burrai et al. [16] utilized a marginally bigger sample of 1,056 images, nevertheless, faced limitations in precision.

Furthermore, although Aubreville et al. [15] presented a more extensive dataset for the detection of mitotic figures in canine mammary cancer, their emphasis was on annotation methodologies rather than a wider classification of CMT. These limitations underscore the deficiency of comprehensive DL research on CMT and emphasize the importance of models developed using larger datasets. Moreover, XAI models are mostly unexamined in this domain, emphasizing the necessity for future investigations to create interpretable, high-performing models for CMT classification.

The existing literature indicates a scarcity of research on the classification of various CMT utilizing DL models. Moreover, the existing literature relies on relatively small datasets that fail to represent the different classes of CMT, hence highlighting a substantial research gap in veterinary diagnosis.

3. PROPOSED SYSETM

This section presents the methodology of the proposed system. As shown in Figure 2, the proposed system mainly consists of four phases: (1) data acquisition and preprocessing; (2) feature extraction, learning, and classification using the proposed hybrid DenseNet201-RF model; (3) performance evaluation; and (4) the explanation of the results. These phases will be illustrated in detail in the following subsections.

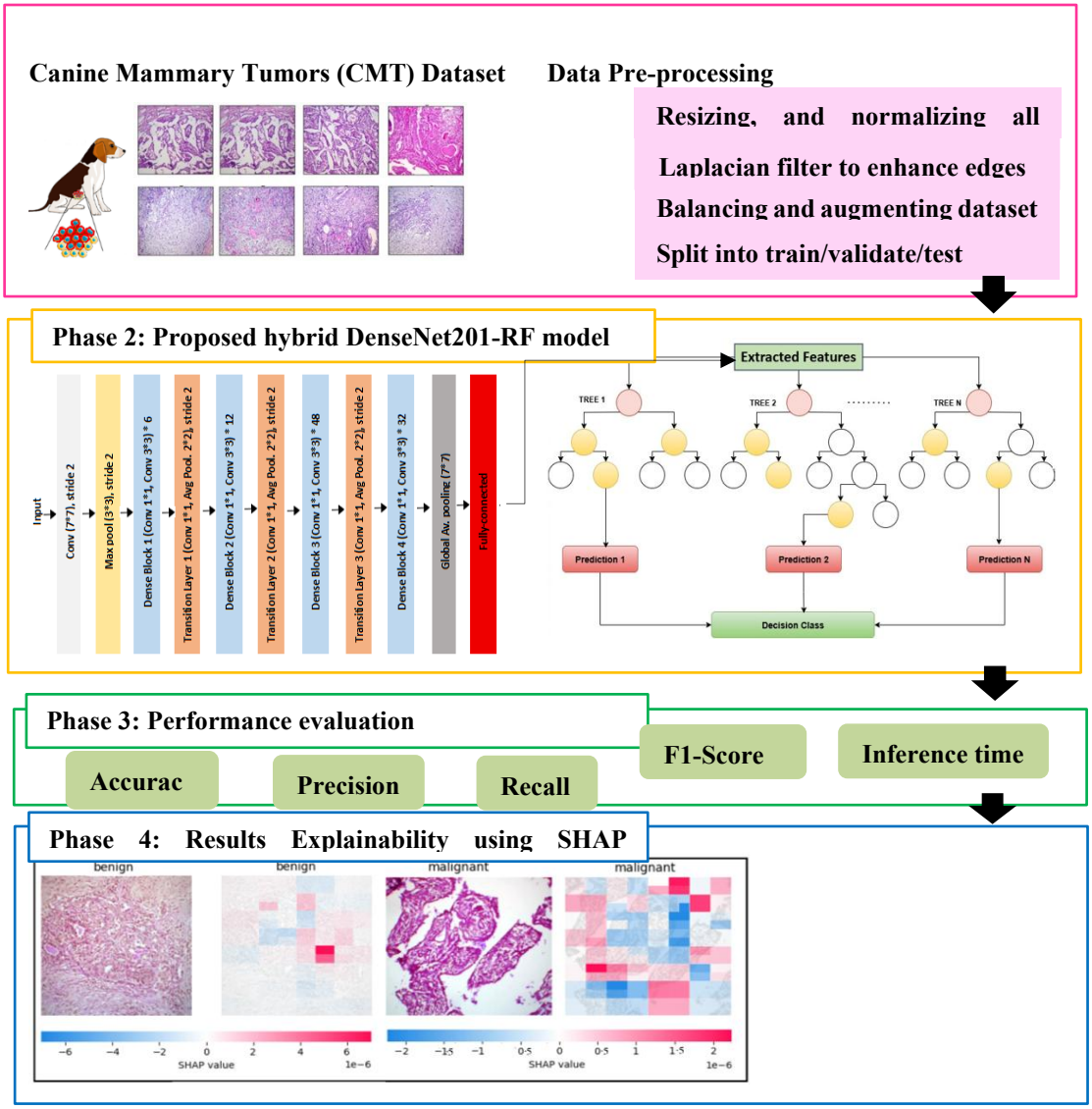


Fig 2: The four phases of the proposed system

3.1 Phase 1: Data acquisition and preprocessing

The dataset used in this paper is the CMT image dataset which consists of 1,056 JPEG images (480 benign, and 576 malignant) from CMT submitted to the University of Sassari's Department of Veterinary Medicine. Tissue samples were fixed, paraffin-embedded, and histopathologically analyzed. Pathologists classified the CMT tissues using histopathological classification. A video pool was selected from each video, resulting in 1024 × 768 high-resolution RGB images from 20 benign, and 24 malignant cases [16]. Figure 3 presents samples from this dataset.

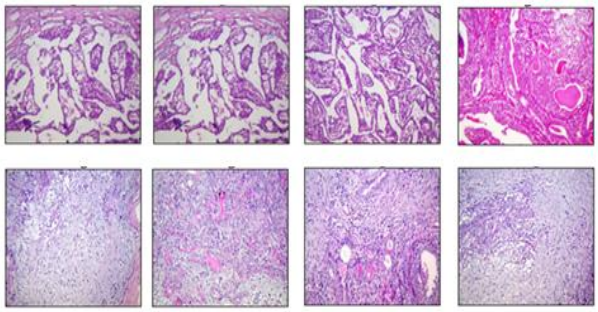


Fig 3: Samples images selected from the CMT dataset: malignant class (top row), and benign class (bottom row)

3.1.1 Scaling, Resizing and sharpening the data

To enhance the quality of the input images and prepare them for the format of the classification phase, a set of preprocessing steps is employed in the proposed work. The first step is to scale the pixel values to a range from 1 to 255 as a normalization for pixel values keeping input pixel values small and consistent. The second step is to resize the images to a standard dimension for consistent analysis to provide fixed input dimensions to the DL model.

Since the proposed work utilizes histopathological CMT images which have very small details, a Laplacian filter is applied to give the images a sharper appearance and enhance the edges of input images using OpenCV library [18]. The Laplacian of an image is a second-derivative method for enhancement which can be calculated using Equation (1).

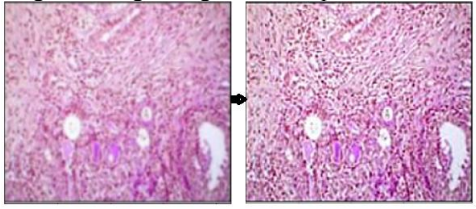
$$\nabla^2 f(x,y) = \frac{\nabla^2 f(x,y)}{\nabla x^2} + \frac{\nabla^2 f(x,y)}{\nabla y^2} \quad , \qquad (1)$$

Where f denotes the image, and f(x, y) denotes the current pixel value. The Laplacian filter has many discrete convolution kernel types such as basic filters which highlight edges and high boost filters which combine the original image with an edge-enhanced version which gives more aggressive detail enhancement [19]. A high-boost filter shown below is preferred in the proposed work since the CMT images have fine details that should be sharpened to enhance the classification process.

0	-1	0
-1	5	-1
0	-1	0

Figure 4 shows the sample resulting images after applying the high-boost Laplacian filter for malignant and benign samples selected from the CMT dataset. It shows how the edges are refined to be sharper and clearer compared to the original image samples, enhancing the clarity and distinction of edges in digital images.

Original Benign Image Laplacian Filtered Image



Original Malignant Image Laplacian Filtered Image

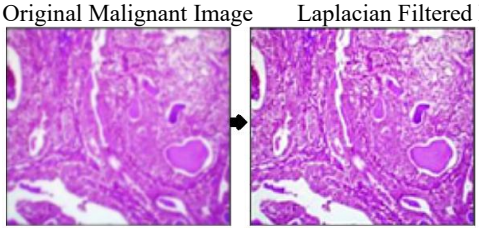


Fig 4: Laplacian filter input, and output for malignant (left), benign (right) samples from the CMT dataset

3.1.2 Balancing Dataset

In the CMT dataset utilized in the proposed work, there are 480 benign samples, and 576 malignant samples with total 1,057 images. As can be concluded, there is an imbalanced number of samples in each class which will result in biased classification results to the class with greater number of samples. This will produce a model that cannot generalize to new samples and hence will affect the model performance [20]. To avoid this effect, the difference between numbers of samples in different classes is calculated. Then, duplicated versions of samples from the class with a smaller number of samples are then created such that the count of these new samples equals that difference as shown in Table 1. The original dataset contains 480 samples in benign class, and 576 samples in malignant class. After balancing, the benign samples equal the number of samples in the malignant class; 576 samples.

Table 1. Number of samples in of each class in the CMT dataset before and after balancing

Class	Before Balancing	After Balancing
Benign	480	576
Malignant	576	576
Total	**1,056**	**1,152**

3.1.3 Data Augmentation

Data augmentation is a crucial process for enhancing the accuracy of DL models by generating new data from existing ones, thereby enhancing their generalization capabilities, and improving model robustness. Data augmentation is deployed in this paper by generating rotated, scaled, and translated versions of the original images [21]. Table 2 introduces the data augmentation methods applied in the proposed system.

Table 2. Configurations of data augmentation methods applied to the CMT dataset

Augmentation method	Value
Rotation	5 Degrees
Shifting in x-axis	0.01
Shifting in y-axis	0.01

3.1.4 Data Splitting

The CMT dataset was divided into three subsets—training, validation, and testing —using a ratio of 80:10:10, respectively.

Table 3. Distribution of CMT dataset with split ratio of 80:10: 10 for training, validation, and testing respectively

Splitting Ratio	Class	Original Dataset	After Balancing	After Augmentation
Training (80%)	Benign	384	460	460*50 epochs = 23,000
	Malignant	460	460	460*50 epochs = 23,000
Validation (10%)	Benign	48	58	58*50 epochs = 2,900
	Malignant	58	58	58*50 epochs = 2,900
Testing (10%)	Benign	48	58	58
	Malignant	58	58	58
Total		**1,056**	**1,152**	**51,916**

To preserve the class distribution across these subsets, stratified sampling was employed during the splitting process. Table 3 gives the distribution of the samples in three cases: original dataset, after balancing, and after augmentation. After splitting, there are 460, 58, and 58 samples from the benign class in the training, validation, and testing splits respectively with a total 576 samples. For the malignant class, the same distribution of train, validate and test splits are for the malignant class since malignant, and benign classes have the same number of samples after balancing. The training process is employed using 50 epochs which creates a new augmented version of the data in every epoch, creating fifty augmented versions of the balanced data.

In augmentation, only the training and validation samples are augmented, and the testing samples remain the same to avoid data leakage and have a realistic evaluation of the model. In the training split, 23,000 samples for benign and 23,000 samples for malignant are generated. In the validation split, 2,900 samples for benign, and 2,900 samples for malignant are generated. While the testing split contains 58 samples for benign, and 58 samples for malignant. This results in a total 51,916 samples for building the proposed system.

3.2 Phase 2: the proposed hybrid DenseNet201-RF model

Figure 5 depicts the proposed hybrid DenseNet201-RF, that utilizes DenseNet-201 DL model for feature extraction and Random Forest (RF) classifier for classification of the CMT input images. The proposed model contains two stages: feature extraction, and classification.

3.2.1 Feature Extraction using DenseNet201

Feature extraction in image processing minimizes data dimensionality while retaining critical information from the original dataset, enhancing the accuracy of classification models and optimizing recognition rates. This process involves identifying and isolating relevant data and organizing the information into feature vectors, which serve as inputs for the ML algorithms. In DL, transfer learning models are pretrained on extensive image datasets and employed as a method for feature extraction, enabling the application of previously acquired knowledge [16]. The proposed work uses the learned weights of the DenseNet201 model to collect the features from the CMT dataset.

DenseNet201 is a DL model that is pretrained the ImageNet dataset. It has four dense blocks and three transition blocks, as shown in Figure 6. DenseNet201 begins with a convolutional layer to extract initial features from the input histopathological images, followed by multiple DenseNet201 blocks, each containing batch normalization, ReLU activation, convolutional layers, and average pooling for feature extraction and dimensionality reduction. Transition blocks are placed between DenseNet201 blocks to transition and further reduce feature map dimensions, consisting of batch normalization, ReLU activation, convolutional layers, and average pooling. The output from the DenseNet201 blocks is passed to a fully connected layer, which aggregates the extracted features. DenseNet201 creates direct connections between all layers, promoting seamless information flow without redundancy. This extensive connectivity facilitates feature propagation and mitigates the vanishing gradient problem, thereby enhancing performance and learning efficiency. The interconnected architecture of DenseNet201 enables enhanced learning and optimal feature utilization [22].

3.2.2 Classification using Random Forest

Random Forest classifier is composed of multiple decision trees, as shown in Figure 7. It can be regarded either as a single classification technique, or several classification methods within a classifier, or as an independent approach with various parameters. Each tree produces a unique classification result, and the forest selects the class that receives the majority vote across all trees. To construct these trees, the Random Forest classifier generates random subsets of the extracted features and identifies key combinations of attributes. After collecting and preprocessing the input CMT images, the DenseNet201 model is used to extract the features. These extracted features are fed as input to the RF classifier to decide whether the input samples are benign or malignant class. By merging the strengths of DL (for feature extraction) with ensemble learning (for classification) leads to a powerful, accurate, and robust hybrid classifier.

3.3 Phase 3: performance evaluation

It is crucial to assess the efficacy and efficiency of the model after training. Using the testing dataset (10%), this phase involves analyzing the prediction capabilities of the model and examining performance indicators that reflect its usefulness in real-world scenarios. Several key performance measures are used in this phase to evaluate the DenseNet201 model [23], such as accuracy, precision, recall, and inference time. Equation (2) defines accuracy as the proportion of correctly predicted samples to the total predictions, serving as an indicator of the model's overall predictive efficacy.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (2)$$

On the other hand, precision is the ratio of accurately predicted positive cases for a class to all positive predictions made for that class, as stated in Equation (3):

$$Precision = TP / (TP + FP) \quad (3)$$

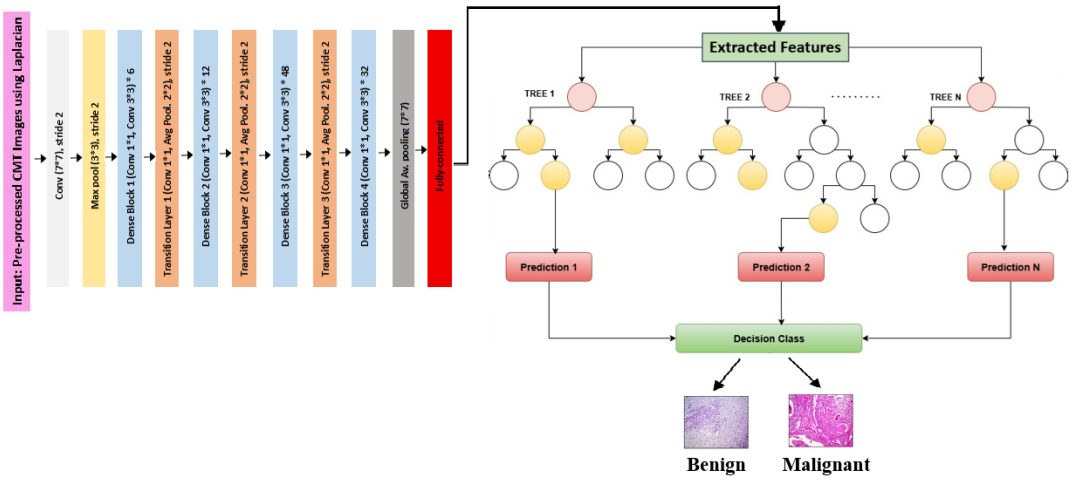


Fig 5: The proposed DenseNet201-RF model

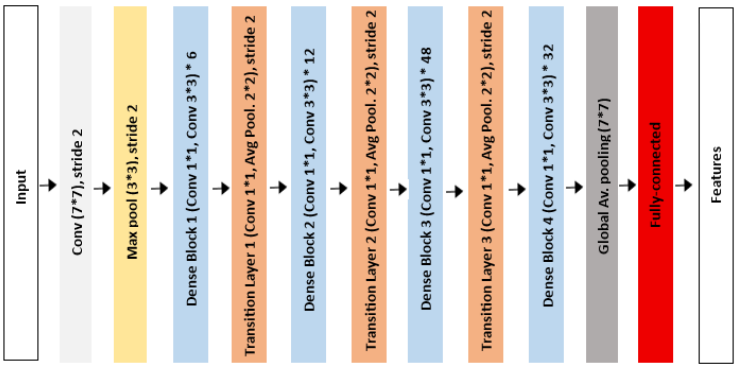


Fig 6: The DenseNet201 model architecture

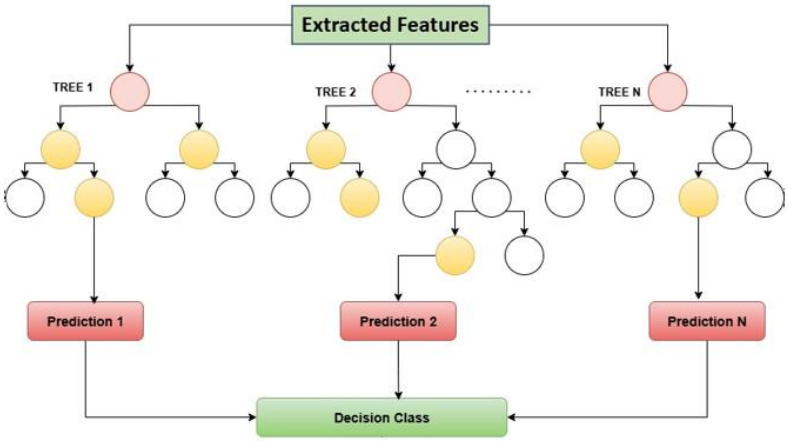


Fig 7: The Random Forest classifier architecture

Equation (4) states that sensitivity, sometimes referred to as recall, quantifies the proportion of real positive samples that are correctly classified as positive.

$$Sensitivity, or\ Recall\ = TP\ /\ (TP\ +\ FN) \quad (4)$$

The F1 score is computed by weighted averaging of precision and recall, as demonstrated in Equation (5):

$$F1\ score\ =\ 2 \times ((precision \times recall)\ /\ (precision + recall)) \quad (5)$$

The number of false positive samples is represented by FP, the number of false negative samples by FN, the number of true positive samples by TP, and the number of true negative samples by TN. When evaluating the effectiveness of DL models, inference time is crucial, particularly in situations requiring real-time processing, or operating with limited resources. It describes how long it takes a model to process an input and produce an output. To ensure optimal hardware utilization, and energy economy, evaluating inference time aids in selecting the best model depending on latency and resource constraints [24]. As specified in Equation (6), start timing at the exact moment just before the input is sent to the model.

$$T_{start} = current\_time() \quad (6)$$

The function $current\_time()$ can be determined using either $time.time()$, or $time.perf\_counter()$. The trained model is provided with the input data X_{test} to generate predictions, as outlined in Equation (7):

$$Y_{pred} = Model\ (X_{test}) \quad (7)$$

Here, Y_{pred} represents the model's output, or prediction. After

obtaining the output, use Equation (8) to record the end time.

$$T_{end} = current\_time() \quad (8)$$

Next, the inference time $T_{inference}$ as shown in Equation (9), is calculated by subtracting the start time from the end time.

$$T_{inference} = T_{end} - T_{start} \quad (9)$$

3.4 Phase 4: result explainability using SHAP

XAI is a new AI discipline recently used by many applications to change the black box nature of ML or DL models and produce human-understandable explanations. The need for trustworthy interpretability techniques has increased as ML and DL techniques are being used more often in critical applications. Because it guarantees consistency in feature importance attributions and is applicable to various model types, SHAP stands out among other XAI models.

Because SHAP provides both local and global interpretability, it is especially suited for analyzing complex DL models, in contrast to some other explanation techniques that might yield unstable or biased results [25]. In order to determine the most important aspects of the decision-making process, The SHAP XAI technique is used in this paper to interpret the class-type predictions in CMT problems. The model's behavior can be better understood by utilizing SHAP's dependable feature attribution, which helps with the creation of better DL models for early warning and prediction of CMT classification problems.

4. EXPERIMENTAL RESULTS AND DISCUSSION

This section explains the experimental setup, the parameter settings, the experiments deployed for CMT dataset, and the analysis of the results.

4.1 Experimental setup

The deployment of the proposed model is developed using personal computer with Intel(R) Core (TM) i7 with 8 GB RAM, 1 Tera Hard Disk, and 64-bit Windows 10 operating system. The code is developed using the python programming language version 3 with TensorFlow framework version 2.10.1 with 960M graphics processing units (GPUs). The proposed system is deployed by setting specific parameters during the training, such as learning rate, optimizer, batch size, loss function, number of epochs, early stopping, rescaling the image size of the input image, and splitting ratio. The exact values of these parameters are listed in Table 4.

Table 4. Parameter settings of the training parameters of the DenseNet201 model.

Parameter	Settings
Learning rate	0.001
Optimizer	Adam optimizer
Batch size	64
Loss function	Categorical cross entropy
Number of Epochs	50
Patience in early stopping	10
Image size	224 width * 224 heights
Splitting ratio	80% train, 10% validate, 10% test

4.2 Performance evaluation of the DenseNet201-RF model for CMT classification

To evaluate the proposed DenseNet201-RF model, the input CMT images are split into three segments, training, validation, and testing. Ten percent is used for validation, ten percent is used for testing, and eighty percent is used for training. The testing data collected after training is used to evaluate the performance of the model and guarantee its ability to generalize to novel data. To create increasingly abstract CMT representations, the DenseNet201 blocks process the images from each epoch. Because each layer has direct access to the gradients and feature maps of earlier levels, the hybrid DenseNet201 architecture promotes effective learning. To monitor the performance of the model using previously unknown validation data, prevent overfitting and adjust the hyperparameters of the model, a validation step is commonly incorporated into the training process. Before predicting the matching class, the DenseNet201 model evaluates the best features of each image in the testing set.

Figure 8 presents the training and validation performance of the proposed DenseNet201 model across 50 epochs, as indicated by the loss and accuracy curves. The training loss consistently declines, signifying excellent model learning, and the validation loss stabilizes with slight variations, showing strong convergence without considerable overfitting. The training accuracy steadily increases, approaching 95%, while the validation accuracy closely trails at around 91–92%. The alignment of training and validation metrics illustrates the resilience and dependability of the DenseNet201 model in effectively categorizing CMT, attaining high accuracy while ensuring stability throughout the training and validation stages.

The alignment of training and validation metrics illustrates the resilience and dependability of the DenseNet201 model in effectively categorizing CMT, attaining high accuracy while ensuring stability throughout the training and validation stages. After testing the DenseNet201-RF model, the accuracy, precision, recall, and inference time value are 93.9, 93.02, 92.94, and 93.91 respectively.

4.3 Influence of SHAP for explainability

The DenseNet201-RF model was combined with SHAP XAI technique to enable post-hoc interpretability of model predictions. SHAP heatmaps illustrate the contribution of different image regions to the model's predictions. Regions highlighted in red indicate positive contributions toward the predicted class, whereas blue regions denote negative contributions. Critical image regions that influence classification results are found and examined for both correctly and incorrectly classified samples using SHAP visualizations as shown in Figure 9. Correct benign predictions emphasize well-organized glandular structures, while correct malignant predictions focus on irregular cellular architecture and densely packed nuclei (Figures 9.a and 9.b). For cases that are misclassified (Figures 9.c and 9.d), benign samples predicted as malignant exhibit localized atypical patterns that resemble malignant features, whereas malignant samples predicted as benign show emphasis on relatively organized regions that reduce malignancy confidence. The model's focus is demonstrated to be on features that cause errors, like those present in benign adenomas, providing insight into the limitations of the model. The results show that the model identifies clinically relevant histopathological patterns and transparently explains both correct and incorrect decisions.

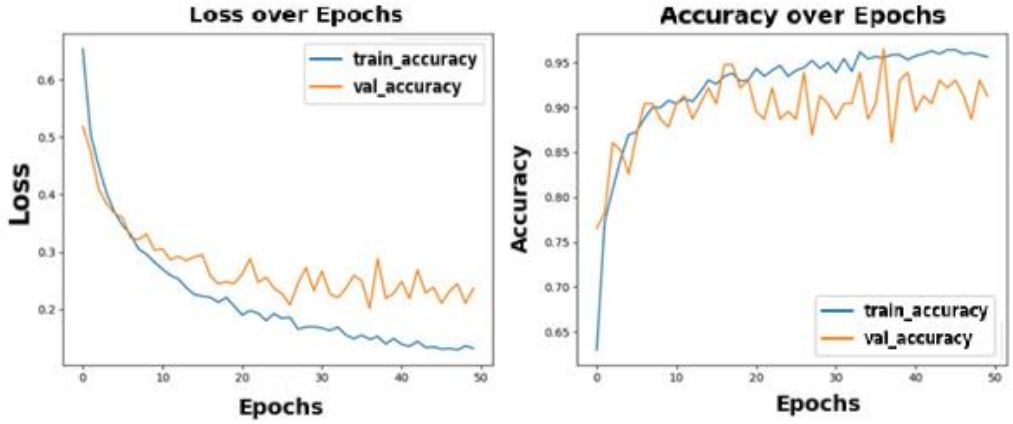


Fig 8: The accuracy and loss curves of the DenseNet201 model

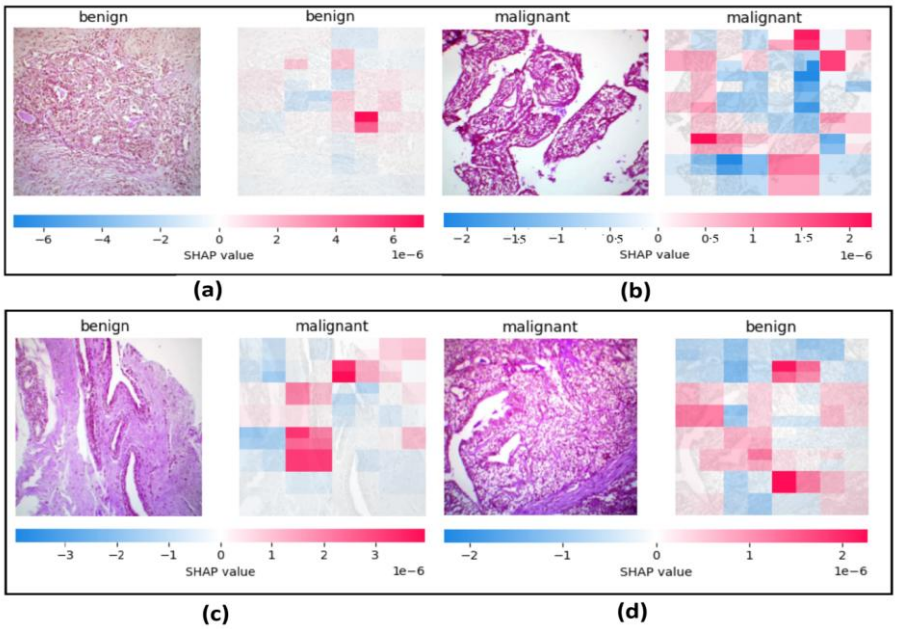


Fig 9: Explanation of the proposed hybrid DenseNet201-RF model on samples from the CMT dataset using SHAP XAI model

The incorporation of SHAP improves the transparency of the system and its dependability by clarifying the decision-making procedure of the model and differentiating the feature contributions to each prediction. Because it fosters trust and facilitates expert oversight, this interpretability is especially beneficial for clinical applications, which helps the model be adopted in delicate domains like medical diagnostics.

4.4 Ablation study

4.4.1 Comparing DenseNet201 with five DL models for CMT classification

An ablation study comparing the DenseNet201 model with five pretrained DL models for CMT classification was carried out utilizing the same dataset across all experiments. These DL models are VGG16, InceptionV3, InceptionResNetV2, Xception, and MobileNetV2. Feature extraction is performed using one of the DL models, and classification is performed by appending a SoftMax layer at the end of each model. VGG16 is a 16-layer convolutional neural network using 3×3 filters with a deep and uniform architecture [26]. InceptionV3 employs inception modules to capture multi-scale features efficiently while reducing computation [27]. InceptionResNetV2 combines inception modules with residual connections to improve training speed and accuracy on large datasets [28]. Xception uses depthwise separable convolutions to capture spatial and channel-wise features efficiently, enhancing accuracy and reducing computational cost [29]. MobileNetV2 is a lightweight network for mobile devices, using inverted residuals and linear bottlenecks to achieve high accuracy with low computational cost [30]. Table 5 compares these five DL models with the DenseNet201 model in terms of the number of layers and the number of computational parameters in millions (M).

Table 5. Comparison between VGG16, InceptionV3, MobileNetV2, Xception, InceptionResNetV2, and DenseNet201 models

Model	# of layers	# of parameters
VGG16	16	138 M
InceptionV3	48	23.8
MobileNetV2	53	3.4 M
Xception	71	22.9 M
InceptionResNetV2	164	55.8 M
DenseNet201	201	20 M

DenseNet201 has the highest number of layers, making it highly suitable for complex image classification tasks, especially where fine details matter (e.g., histopathological

images of CMT). Moreover, it has fewer parameters, achieving high accuracy without a huge computational cost.

Table 6 gives the confusion metrics of the five DL models with the DenseNet201 model emerging as the most effective model. Because DenseNet201 has the fewest false positives and a good balance between true positives and true negatives, it is highly effective at reducing misclassifications.

Table 6. Confusion matrices of the six pretrained DL models: VGG16, InceptionV3, InceptionResNetV2, Xception, MobileNetV2, and DenseNet201

Benign	51 (TP)	8 (FN)
Malignant	18 (FP)	39 (TN)
	Benign	**Malignant**

(a) **VGG16**

Benign	47 (TP)	12 (FN)
Malignant	7 (FP)	50 (TN)
	Benign	**Malignant**

(b) InceptionV3

Benign	44 (TP)	15 (FN)
Malignant	3 (FP)	54 (TN)
	Benign	**Malignant**

(c) InceptionResNetV2

Benign	49 (TP)	10 (FN)
Malignant	8 (FP)	49 (TN)
	Benign	**Malignant**

(d) Xception

Benign	46 (TP)	13 (FN)
Malignant	4 (FP)	53 (TN)
	Benign	**Malignant**

(e) MobileNetV2

Benign	52 (TP)	7 (FN)
Malignant	8 (FP)	49 (TN)
	Benign	**Malignant**

(f) DenseNet201

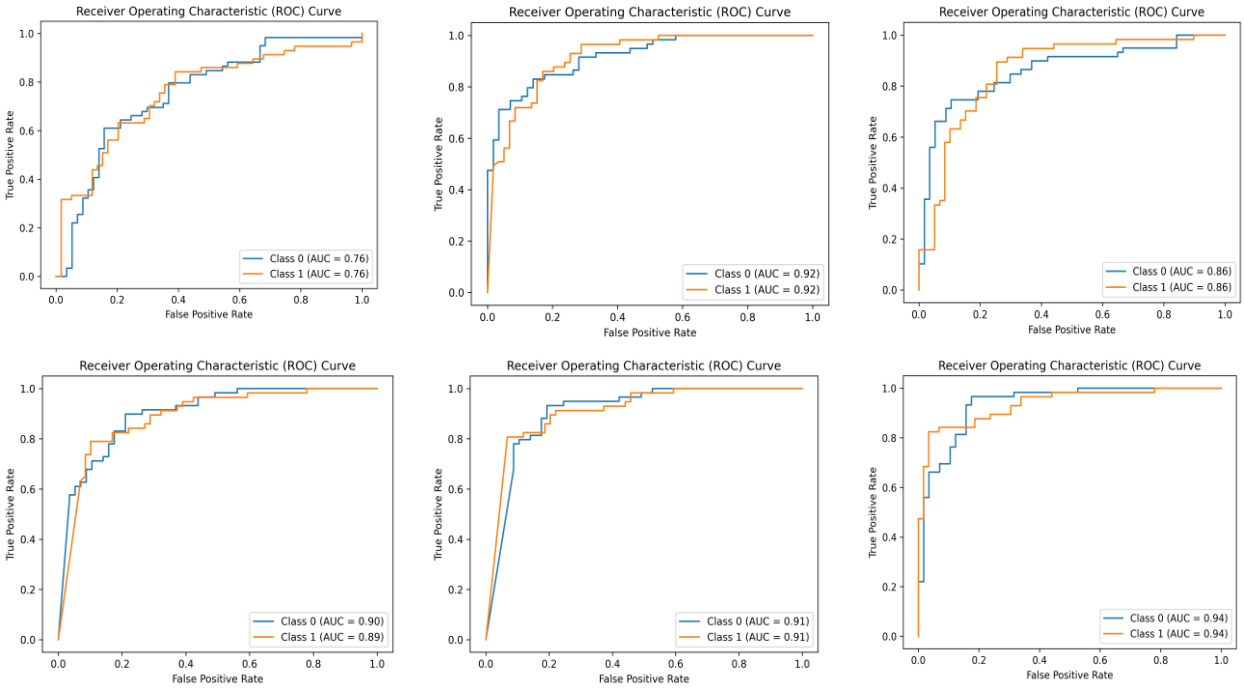


Fig 10: The ROC curves of the DenseNet201 model versus five pretrained DL models

Due to its precision and efficiency, DenseNet201 is the optimal model for CMT classification scenarios requiring both expeditious implementation and effective classification accuracy. Moreover, the ROC (Receiver Operating Characteristic) curves and AUC (Area Under the Curve) scores presented in Figure 10 demonstrate the capacity of the six pretrained DL models to differentiate between classes. The best classification performance and robustness are demonstrated by the proposed DenseNet201 model, which surpasses the rest with the highest AUC scores of 0.94 for both Class 0 (benign) and Class 1 (malignant). AUC ratings of roughly 0.92 and 0.91 are achieved by models such as InceptionV3, and MobileNetV2 respectively; however, DenseNet201 consistently outperforms them in class distinction with little overlap. Compared to both classes, VGG16 has a much lower AUC of 0.76. Table 7 shows how well the proposed DenseNet201 model performs, surpassing all other pretrained models in terms of accuracy (87.07%), precision (87.5%), recall (87%), and F1 score (87%). These results demonstrate its capacity to produce the most consistent classification results, confirming its supremacy over rivals such as MobileNetV2 (85.34% accuracy) and Xception (84.48% accuracy).

Table 7. Comparison between the proposed DenseNet201 model versus five DL models

Model	Accuracy	Precision	Recall	F1 score	Inference time (in Minutes)
VGG16	77.58	78.5	77	77.5	12.05
InceptionV3	83.62	84	84	83.5	6.25
InceptionResNetV2	84.48	86	85	84.5	10.37
Xception	84.48	84.5	84.5	84	**4.08**
MobileNetV2	85.34	86	85.5	85	4.48
DenseNet201 (proposed)	**87.07**	**87.5**	**87**	**87**	7.94

Despite having a somewhat longer inference time (7.94 minutes) than some models, such as Xception (4.08 minutes) and InceptionResNetV2 (4.48 minutes), the proposed DenseNet201 model offers a well-balanced combination of classification performance and computational time. Additionally, DenseNet201 is well known for its efficient inference time, which makes it more suitable for real-time, or resource-constrained applications than deeper and more computationally expensive models like InceptionResNetV2. The proposed DenseNet201 model is the top-performing model among the assessed architectures, as evidenced by its AUC scores, balanced confusion matrix performance, and quick inference time.

4.4.2 Comparing proposed hybrid DenseNet201-RF with three ML classifiers

This section examines the impact of employing various ML classifiers on the performance of classifying the CMT using DenseNet201 as a feature extractor. These ML classifiers include support vector machines (SVM), k-Nearest Neighbors (KNN), and a Decision Trees (DT), in addition to the SoftMax classifier. Support Vector Machine (SVM) is a supervised machine learning algorithm that finds the optimal hyperplane to separate data points of different classes with the maximum margin [16]. K-Nearest Neighbors (KNN) is a simple, supervised machine learning algorithm that classifies a data point based on the majority class among its *k* closest neighbors in the feature space [31]. Decision Tree is a supervised machine learning algorithm that splits data recursively based on feature values to create a tree-like model for classification or regression [32].

Table 8 provides a comparison of three ML classifiers associated with the DenseNet201 as feature extractor, beside the SoftMax classifier. After extracting meaningful features from the input CMT images using the DenseNet201 model, these classifiers are employed to classify the images into benign or malignant categories, alongside the SoftMax classifier. The comparison is in terms of accuracy, precision, recall, and inference time.

Table 8. Comparison between different classifiers associated with DenseNet201 as feature extractor

Classifier	Accuracy	Precision	Recall	F1 score
KNN	88.16	89.16	89.16	88.48
Decision Tree	81.74	82.74	82.74	81.72
SVM	89.56	89.95	89.95	89.56
SoftMax	87.07	87.5	87	87
Random Forest	93.9	93.02	92.94	93.91

The RF classifier achieves the highest performance. This superior performance of RF stems from its ensemble learning approach, where multiple decision trees independently make predictions, and the final classification is determined through majority voting [33].

4.4.3 Testing the proposed DenseNet201-RF model using the BreakHis human breast cancer dataset

Histopathological images of canine mammary tumors closely resemble human breast cancer, showing similar glandular architecture and malignant features such as nuclear pleomorphism, high mitotic rate, and stromal invasion [34], [35], and [36]. Because of these shared morphological and molecular characteristics, human breast cancer serves as a valuable comparative model for canine mammary tumors research.

The human breast cancer dataset deployed in this experiment is called BreakHis dataset, which comprises 7,909 microscopic tumor images from 82 patients, captured at 40×, 100×, 200×, and 400× magnifications. It includes 2,480 benign and 5,429 malignant samples. Figure 11 gives a sample image of each class. The same configurations and preprocessing steps that were deployed for the CMT dataset are employed on BreakHis dataset, as shown in Figure 12.

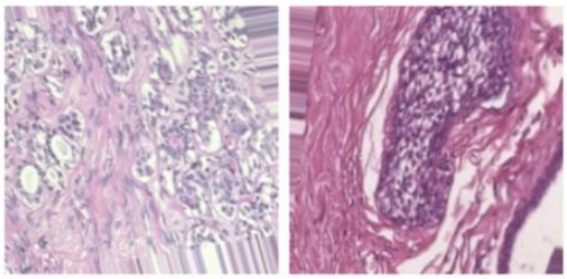


Fig 11: Sample images from the BreakHis dataset. Left: benign class. Right: malignant class

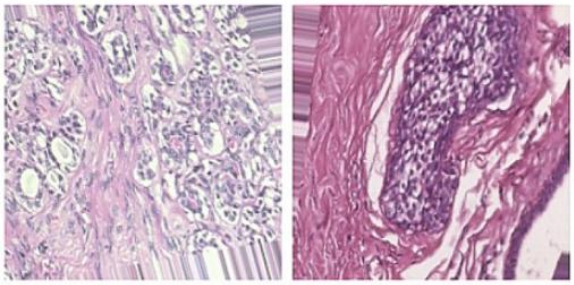


Fig 12: Applying Laplacian Filter for sample images from BreakHis dataset in figure 11

Table 9 gives the results of deploying the DenseNet201, and DenseNet201-RF on the BreakHis dataset. DenseNet201 model is used as feature extractor and classifier. While using DenseNet201-RF model, the DenseNet201 model is utilized as feature extractor, and the RF is utilized as classifier. The results demonstrate promising performance, highlighting the potential of integrating deep learning models are feature extractors and machine learning techniques as classifiers in human breast cancer research.

Table 9: The performance of the proposed DenseNet201-RF model on the BreakHis dataset compared to DenseNet201 model

Classifier	Accuracy	Precision	Recall	F1 score
DenseNet 201	87.29	88.35	87.30	87.21
DenseNet 201-RF	91.91	90.82	90.91	90.82

5. COMPARISON WITH LITERATURE

The proposed DenseNet201-RF model is compared to related works in literature that utilized the same dataset [16], as illustrated in Table 10. With an accuracy of 93.9%, the DenseNet201-RF model shows a significant improvement. This performs better than reference models that use SVM as the classifier on the same dataset, such as Inception-SVM (81.0%), VGG16-SVM (82.0%), and EfficientNet-SVM (85.0%). The ensemble learning capability of the RF classifier, which uses the various decision boundaries of several decision trees to improve classification robustness is responsible for the DenseNet201-RF model's superiority.

The proposed improvements, especially the combination of balanced, augmented, and sharpened images with RF integration, are essential to this performance. The efficiency of the selected ensemble approach for CMT classification is confirmed by the DenseNet201-RF, which not only outperforms other architectures but also attains higher accuracy than DenseNet201 models combined with other classifiers.

Table 10. Comparison of the proposed DenseNet201-RF model versus the state-of-the art models based on the same CMT dataset.

Method	Accuracy
InceptionV3 (feature extractor) + SVM (classifier) [16]	81%
VGG16 (feature extractor) + SVM (classifier) [16]	82%
EfficientNet (feature extractor) +SVM (classifier) [16]	85%
DenseNet201(feature extractor) + SoftMax (classifier) + Balanced [**Proposed**]	87.07%
DenseNet201(feature extractor) + SoftMax (classifier) + Balanced +Augmented [**Proposed**]	89.63%
DenseNet201 (feature extractor) + SoftMax (classifier) + Balanced+ Augmented+ Laplacian [**Proposed**]	91.5%
DenseNet201 (feature extractor) + Random Forest (classifier) + Balanced + Augmented + Laplacian [**Proposed**]	93.9 %

6. CONCLUSION AND FUTURE WORK

This research presents an explainable hybrid DenseNet201-RF model for the classification of CMT images. The proposed model depends on four critical phases: image preprocessing, feature extraction, model classification, and performance evaluation. Moreover, a technique in XAI called SHAP elucidates the results. This technique mitigates the black-box nature of DL predictions, thereby improving the reliability and interpretability of the results.

To compare the efficacy of the proposed DenseNet201-RF model in categorizing CMT, many machine learning classifiers such as SVM, DT, and KNN beside the SoftMax classifier were utilized subsequent to the feature extraction procedure. Among these, the DenseNet201-RF model achieved the highest accuracy of 93.9%, outperforming the other classifiers and demonstrating the effectiveness of using the RF classifier to maximize the classification performance of the model.

On the other hand, the proposed DenseNet201-RF model outperformed other state-of-the-art models, which only achieve 85% accuracy using the same CMT dataset, with a reasonable inference time and a 93.9% accuracy rate. This illustrates the dependability and effectiveness of CMT classification problems using the proposed model.

The integration of XAI improves its use in clinical and veterinary diagnostics by providing insights into the decision-making process and closes a significant gap in the existing literature. Future research is suggested to use more diverse samples of the CMT classes. Additionally, efforts could be focused on creating lightweight models deployed for mobile applications for real-time use in veterinary and clinical settings. Investigating further explainable AI methods would enhance interpretability and reliability, which are essential for acceptance in medical settings.

Data Availability

The findings of the proposed work are supported by a dataset publicly available at: https://zenodo.org/records/7716066. The BreakHis dataset is publicly available at: https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

CRediT author statement

Aboul Ella Hassanein: Idea, Senior administration, Supervision, Validation, review and editing. **Ashraf Darwish:** Investigation, Supervision, Writing- Reviewing and Validation. **Elham Shawky Salama:** Data curation, Writing-results, Results visualization and Software. **Heba Askr:** Methodology, Validation, Visualization, Writing –original draft, and Writing – review and editing.

Funding

No funding received.

Ethics statement

This paper was conducted using publicly available data and previously published research articles; no new experiments involving human participants or animals were performed by the authors. All sources of information have been properly cited to ensure academic integrity and transparency. The paper adheres to ethical guidelines for scholarly publishing, including the avoidance of plagiarism, data fabrication, and inappropriate citation practices. No conflicts of interest or ethical concerns are associated with this work.

7. REFERENCES

[1] M. Kumar and U. Mehta, "Enhancing the performance of CNN models for pneumonia and skin cancer detection using novel fractional activation function," Applied Soft Computing, vol. 168, p. 112500, 2025. DOI: https://doi.org/10.1016/j.asoc.2024.112500

[2] T. Umamaheswari and Y. M. Mohanbabu, "CNN-FS-IFuzzy: A new enhanced learning model enabled by adaptive tumor segmentation for breast cancer diagnosis

using 3D mammogram images," Knowledge-Based Systems, vol. 288, p. 111443, 2024. DOI: https://doi.org/10.1016/j.knosys.2024.111443

[3] E. Vazquez et al., "Canine mammary cancer: State of the art and future perspectives," Animals, vol. 13, no. 19, p. 3147, 2023. DOI: https://doi.org/10.3390/ani13193147

[4] A. F. Oliveira-Lopes, M. M. Götze, B. E. Lopes-Neto, D. D. Guerreiro, I. C. Bustamante-Filho, and A. A. Moura, "Molecular and Pathobiology of Canine Mammary Tumour: Defining a Translational Model for Human Breast Cancer," Veterinary and Comparative Oncology, vol. 22, no. 3, pp. 340-358, 2024. DOI: https://doi.org/10.1111/vco.12996

[5] I. Dolka, M. Czopowicz, D. Stopka, A. Wojtkowska, I. Kaszak, and R. Sapierzyński, "Risk factor analysis and clinicopathological characteristics of female dogs with mammary tumours from a single-center retrospective study in Poland," Scientific Reports, vol. 14, no. 1, p. 5569, 2024. DOI: https://doi.org/10.1038/s41598-024-56194-z

[6] N. Nosalova et al., "Canine mammary tumors: classification, biomarkers, traditional and personalized therapies," International Journal of Molecular Sciences, vol. 25, no. 5, p. 2891, 2024. DOI: https://doi.org/10.3390/ijms25052891

[7] J. Zhang et al., "Prognosis prediction based on liver histopathological image via graph deep learning and transformer," Applied Soft Computing, vol. 161, p. 111653, 2024. DOI: https://doi.org/10.1016/j.asoc.2024.111653

[8] I. Kaszak, O. Witkowska-Piłaszewicz, K. Domrazek, and P. Jurka, "The novel diagnostic techniques and biomarkers of canine mammary tumors," Veterinary Sciences, vol. 9, no. 10, p. 526, 2022. DOI: https://doi.org/10.3390/vetsci9100526

[9] M. Goldschmidt, L. Peña, R. Rasotto, and V. Zappulli, "Classification and grading of canine mammary tumors," Veterinary pathology, vol. 48, no. 1, pp. 117-131, 2011. DOI: https://doi.org/10.1177/0300985810393258

[10] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017), "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," preprint:1711.05225. DOI: https://doi.org/10.48550/arXiv.1711.05225

[11] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017), "Dermatologist-level classification of skin cancer with Deep Neural Networks," Nature, 542(7639), 115–118. DOI: https://doi.org/10.1038/nature21056

[12] H. Askr et al., "Exploring the anticancer activities of Sulfur and magnesium oxide through integration of deep learning and fuzzy rough set analyses based on the features of Vidarabine alkaloid," Scientific Reports, vol. 15, no. 1, p. 2224, 2025. DOI: https://doi.org/10.1016/j.antiviral.2023.105740

[13] P. S. Basran and R. B. Appleby, "What's in the box? A toolbox for safe deployment of artificial intelligence in veterinary medicine," Journal of the American Veterinary Medical Association, vol. 262, no. 8, pp. 1090-1098, 2024. DOI: https://doi.org/10.2460/javma.24.01.0027

[14] A. Kumar et al., "Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer," Information Sciences, vol. 508, pp. 405-421, 2020. DOI: https://doi.org/10.1016/j.ins.2019.08.072

[15] M. Aubreville, C. A. Bertram, T. A. Donovan, C. Marzahl, A. Maier, and R. Klopfleisch, "A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research," Scientific data, vol. 7, no. 1, p. 417, 2020. DOI: https://doi.org/10.6084/m9.figshare.13182857

[16] G. P. Burrai et al., "Canine mammary tumor histopathological image classification via computer-aided pathology: an available dataset for imaging analysis," Animals, vol. 13, no. 9, p. 1563, 2023. DOI: https://doi.org/10.3390/ani13091563

[17] A. H. Işık, Ö. Özmen, Ö. C. Eskicioğlu, N. Işık, and S. Melenli, "Classification and Diagnosis of Mammary Tumors in Dogs Using Deep Learning Techniques," Traitement du Signal, vol. 40, no. 4, 2023. DOI: https://doi.org/10.18280/ts.400444

[18] G. Bradski, "The opencv library," Dr. Dobb's Journal: Software Tools for the Professional Programmer, vol. 25, no. 11, pp. 120-123, 2000.

[19] Depeursinge, A., Andrearczyk, V., Whybra, P., van Griethuysen, J., Müller, H., Schaer, R., Zwanenburg, A. (2020), "Standardised convolutional filtering for radiomics," preprint:2006.05470. DOI: https://doi.org/10.48550/arXiv.2006.05470

[20] B. Kitchenham, "A procedure for analyzing unbalanced datasets," IEEE transactions on Software Engineering, vol. 24, no. 4, pp. 278-301, 2002. DOI: https://doi.org/10.1109/32.677185

[21] Q. Zhu, L. Fan, and N. Weng, "Advancements in point cloud data augmentation for deep learning: A survey," Pattern recognition, vol. 153, p. 110532, 2024. DOI: https://doi.org/10.1016/j.patcog.2024.110532

[22] A. M. Roy and J. Bhaduri, "DenseSPH-YOLOv5: An automated damage detection model based on DenseNet and Swin-Transformer prediction head-enabled YOLOv5 with attention mechanism," Advanced Engineering Informatics, vol. 56, p. 102007, 2023. DOI: https://doi.org/10.1016/j.aei.2023.102007

[23] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," Computers & Operations Research, vol. 152, p. 106131, 2023. DOI: https://doi.org/10.1016/j.cor.2022.106131

[24] M. M. H. Shuvo, S. K. Islam, J. Cheng, and B. I. Morshed, "Efficient acceleration of deep learning inference on resource-constrained edge devices: A review," Proceedings of the IEEE, vol. 111, no. 1, pp. 42-91, 2022. DOI: https://doi.org/ 10.1109/JPROC.2022.3226481

[25] O. Buyuktepe, C. Catal, G. Kar, Y. Bouzembrak, H. Marvin, and A. Gavai, "Food fraud detection using explainable artificial intelligence," Expert Systems, vol. 42, no. 1, p. e13387, 2025. DOI: https://doi.org/10.1111/exsy.13387

[26] Hossain, A. A., Nisha, J. K., & Johora, F. (2023), "Breast cancer classification from ultrasound images using

VGG16 model-based transfer learning,". International Journal of Image, Graphics and Signal Processing, 15(1), 12. DOI: DOI: https://doi.org/10.5815/ijigsp.2023.01.02

[27] Sharma, A., & Parvathi, R. (2025), "Enhancing Cervical Cancer Classification: Through a Hybrid Deep Learning Approach Integrating DenseNet201 and InceptionV3," IEEE Access. DOI: https://doi.org/10.1109/ACCESS.2025.3527677

[28] Ferreira, C. A., Melo, T., Sousa, P., Meyer, M. I., Shakibapour, E., Costa, P., & Campilho, A. (2018, June), "Classification of breast cancer histology images through transfer learning using a pretrained inception resnet v2," In International conference image analysis and recognition (pp. 763-770). Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-93000-8_86

[29] Sharma, S., & Kumar, S. (2022), "The Xception model: A potential feature extractor in breast cancer histology images classification," ICT Express, 8(1), 101-108. DOI: https://doi.org/10.1016/j.icte.2021.11.010

[30] Wibowo, A., Hartanto, C. A., & Wirawan, P. W. (2020), "Android skin cancer detection and classification based on MobileNet v2 model," International Journal of Advances in Intelligent Informatics, 6(2), 135-148. DOI: https://doi.org/10.26555/ijain.v6i2.492

[31] Sharma, M., Singh, S. K., Agrawal, P., & Madaan, V. (2016), "Classification of clinical dataset of cervical cancer using KNN," Indian J. Sci. Technol, 9(28), 1-5. DOI: https://doi.org/10.17485/ijst/2016/v9i28/98380

[32] Tarawneh, O., Otair, M., Husni, M., Abuaddous, H. Y., Tarawneh, M., & Almomani, M. A. (2022), "Breast cancer classification using decision tree algorithms," International Journal of Advanced Computer Science and Applications, 13(4).

[33] A. Murmu and P. Kumar, "DLRFNet: deep learning with Random Forest network for classification and detection of malaria parasite in blood smear," Multimedia Tools and Applications, vol. 83, no. 23, pp. 63593-63615, 2024. DOI: https://doi.org/10.1007/s11042-023-17866-6

[34] Razavirad, A., Rismanchi, S., Mortazavi, P., & Muhammadnejad, A. (2024). Canine Mammary Tumors as a Potential Model for Human Breast Cancer in Comparative Oncology. Veterinary Medicine International, 2024(1), 9319651. DOI: DOI: https://doi.org/10.1155/2024/9319651

[35] Fuertes-Recuero, M., García San José, P., Valdivia, G., Suarez-Redondo, M., Penelo, S., Arenillas, M., ... & Ortiz-Díez, G. (2026). Preoperative Clinical Predictors of Histologic Malignancy and Carcinoma Grade in 286 Canine Mammary Nodules from 92 Bitches: A Retrospective Study Tumour. Animals, 16(3), 421. DOI: https://doi.org/10.3390/ani16030421

[36] Salas, Y.; Márquez, A.; Diaz, D.; Romero, L. Epidemiological Study of Mammary Tumors in Female Dogs Diagnosed during the Period 2002–2012: A Growing Animal Health Problem. PLoS ONE 2015, 10, e0127381. DOI: https://doi.org/10.1371/journal.pone.0127381