Journal of Quality Measurement and Analysis *JQMA* 21(1) 2025, 267-285
e-ISSN: 2600-8602 *http://www.ukm.my/jqma*
https://doi.org/10.17576/jqma.2101.2025.17

MULTI-PHASE DUAL-ENCODER MODEL FOR ANOMALY DETECTION IN MEDICAL IMAGING

(Model Berbilang Fasa dengan Dua Pengekod untuk Pengesanan Anomali dalam Pengimejan Perubatan)

NUR RUSYIDAH AZRI, SARATHA SATHASIVAM* & MAJID KHAN MAJAHAR ALI

ABSTRACT

An error in medical diagnosis is enough to change a life, and many are suffering from incorrect diagnoses. Deep learning models in healthcare face several challenges due to limited and imbalanced datasets, where anomaly samples often dominated in publicly available data and it is problematic to collect independent data due to strict privacy regulations. Furthermore, current models often perform poorly on small sample datasets and lack robustness across various types of medical imaging. To address these issues, we developed a novel model that capable of detecting anomalies in medical imaging across both small and large datasets. This model features a multiple phase of training and validation, with dual encoders and a shared decoder architecture. Our results demonstrate that this model outperforms established classification methods in medical imaging datasets, including BraTS2021, RESC, and BreastMNIST, by achieving superior accuracy in distinguishing normal and anomalous images based on reconstruction error metrics. Furthermore, the research explores the interpretability of latent space features using explanation methods along with visualization techniques. By automating the diagnostic process, our model aligns with Malaysia's Healthcare Government Plan 2023 to reform the health system over the next 15 years and reduces workload for healthcare professionals. Our model can serve as a foundation for developing reliable diagnostic tools with interpretable latent space to understand the model's decision.

Keywords: diagnostic automation; dual-encoder model; medical data analysis; one-class classification; unsupervised deep learning

ABSTRAK

Kesilapan dalam diagnosis perubatan boleh memberi kesan besar terhadap kehidupan pesakit, dan ramai yang telah menderita akibat salah diagnosis. Model pembelajaran mendalam berdasarkan rangkaian neural dalam bidang kesihatan menghadapi beberapa cabaran disebabkan oleh set data yang terhad dan tidak seimbang, di mana sampel anomali sering mendominasi dalam data yang tersedia secara umum, sementara pengumpulan data baharu menghadapi kekangan kerana peraturan privasi yang ketat. Tambahan pula, model yang sedia ada sering menunjukkan prestasi yang kurang baik pada set data bersaiz kecil dan tidak cukup mantap untuk diaplikasikan pada pelbagai jenis pengimejan perubabatan. Bagi mengatasi isu-isu ini, kami telah membangunkan model baharu yang mampu mengesan anomali dalam pengimejan perubatan merentasi set data bersaiz kecil dan besar. Model ini menampilkan banyak fasa latihan dan penilaian, dengan dua pengekod yang berkongsi penyahkod tunggal. Hasil kajian menunjukkan bahawa model ini mengatasi hasil klasifikasi sedia ada dalam set data pengimejan perubatan, termasuk BraTS2021, RESC, dan BreastMNIST, dengan menunjukkan ketepatan yang lebih tinggi dalam membezakan imej normal dan anomali berdasarkan analisis ralat pembinaan semula. Selain itu, kajian ini meneroka kebolehan mentafsir ciri ruang laten menggunakan kaedah penjelasan bersama teknik visualisasi. Automasi diagnosis yang diperkenalkan oleh model kami selaras dengan Rancangan Kesihatan Kerajaan Malaysia 2023 untuk mereformasi sistem kesihatan dalam tempoh 15 tahun akan datang, selain mengurangkan beban kerja professional kesihatan. Model ini juga berpotensi menjadi asas untuk membangunkan alat diagnostik yang boleh dipercayai dengan ruang laten yang boleh ditafsirkan bagi memahami hasil klasifikasi model.

Kata kunci: analisis data perubatan; automasi diagnosis; klasifikasi satu kelas; model dua-pengekod; pembelajaran mendalam tanpa pengawasan

1. Introduction

Diagnostic errors in healthcare can have severe consequences, yet they occur frequently across various medical settings. In 2016, the story of Rufino Borrego was widely reported as a sobering reminder of human cost of diagnostic errors (The Nation 2016). He was misdiagnosed with muscular dystrophy at the age of 13 and spent 43 years on wheelchair. His condition was later identified by a neurologist in 2010 as myasthenia which is a treatable disease and should not spend his life on a wheelchair for that long. These errors are not isolated incidents. It is estimated that more than 7 million diagnostic error occur annually in United States emergency departments, contributing to nearly 800,000 deaths or disabilities each year (Hussain *et al.* 2019). Public concerns about diagnostic errors in Malaysia were further validated after a high-profile misdiagnosis led to a patient's death in 2023 (BERNAMA 2023). This tragic case highlighted the urgent need for advanced, reliable, and automated diagnostic tools. Under the Healthcare Government Plan 2023, Malaysia's 15-year initiative to digitalize healthcare by utilizing artificial intelligence (AI) to improve diagnostic accuracy.

Artificial intelligence has been considered as a potential approach to eliminate human errors from diagnosis (Baurasien et al. 2023) and have capabilities to determine patient's health status (Alowais *et al.* 2023). Recent advancements in deep learning for medical anomaly detection have shown much promise due to their ability to learn and differentiate hidden patterns within complex datasets (Chatterjee *et al.* 2024; Chen *et al.* 2022b). These models are also applicable for other applications such as intrusion detection (Su *et al.* 2020; Tang *et al.* 2019) and bank risk assessment (Addo *et al.* 2018), because of their architectures consist of multiple layers that progressively extract increasingly abstract features from the data. However, it is concerning that the accuracy of these models heavily depends on the amount and quality of training data (Bhatt *et al.* 2024; Brownlee 2020), usually requiring large datasets with over 10,000 samples. The challenge of training these models with small datasets has been less explored, leaving a gap in healthcare-based anomaly detection capabilities. This difficulty arises from the nature of medical data collection, where normal data are limited and significantly outnumbered by anomalies from rare conditions in publicly available datasets (Churová *et al.* 2021). Researchers are usually limited to use of medical data given by medical institutions, as medical records are highly sensitive and governed by privacy regulations such as the Personal Data Protection Act 2010.

Current techniques such as transfer learning and data augmentation are widely used to enhance model performance when dealing with limited and imperfect data. Transfer learning utilizes a pre-trained model as a starting point for a related task, while data augmentation increases the training data size by creating variations of the existing samples through rotation, cropping, or adjusting the colour balance of the original samples. However, these techniques have drawbacks: transfer learning can struggle with mismatches when the new task differs significantly from the original training context (Zhuang *et al.* 2021), and data augmentation may generate non-representative samples that mislead the training process (Medvedieva *et al.* 2024). As a result, a significant research gap remains in developing models that can effectively handle limited and imbalanced medical datasets without relying on conventional techniques. Most existing anomaly detection methods are designed for large, balanced datasets, making them unsuitable for real-world medical applications where acquiring extensive normal data is challenging due to privacy regulations and the rarity of certain conditions. Additionally, these methods often lack robustness when applied across different medical imaging modalities, resulting in inconsistent performance. Furthermore, interpretability is essential for ensuring that clinicians can trust and understand the model's decisions in diagnostic contexts. Hence, a novel approach is needed that achieves high accuracy and generalization with small datasets, reliably detects anomalies, and provides explanations that clinicians can understand and use confidently.

The paper is structured as follows: the next section reviews related work, focusing on recent theories and applications of autoencoders in anomaly detection and healthcare. We then elaborate our developed methodology, including the dual-encoder and decoder architecture, mathematical formulations, and the process for determining anomaly scores and thresholds, to highlight the novelty of our model. Subsequently, we describe the datasets used for evaluation. Later sections present our analysis, comparing the model’s effectiveness with other well-known medical anomaly detection methods. Finally, we conclude by discussing our findings, limitations, and implications for future research in medical diagnostics.

2. Related Work

The exploration of autoencoders in AI has led to major advancements in machine learning, particularly in anomaly detection. Autoencoders are designed to reproduce input data and have evolved into effective tools for feature extraction (Bengio *et al.* 2013). These models excel at capturing the underlying patterns of data, making them suitable for both unsupervised and supervised learning task. Autoencoders range from simple reconstruction tools to complex systems capable of selective feature extraction (Berahmand *et al.* 2024), surpassing traditional methods by handling high-dimensional data in a non-linear manner. Unlike support vector machines or decision trees, they require less manual feature engineering which can operate directly on raw data without extensive preprocessing, contributing to higher operational efficiency. Moreover, compared to convolutional or recurrent neural networks, autoencoders are well-suited for anomaly detection as they specifically engineered for feature compression and reconstruction. The development of dual autoencoder architecture has enhanced the capabilities of collaborative filtering in identifying anomalies (Dong *et al.* 2020). Recent innovations have expanded the potential of autoencoders in several groundbreaking ways, which will be discussed in this subsection.

2.1. *Autoencoders in anomaly detection*

Several variations and enhancements of autoencoders have been developed to tackle distinct challenges within anomaly detection. Dual autoencoder architectures have shown promising results in addressing complex data with imbalanced datasets. Wu, Cui & Welsch (2020) proposed a model that integrates generative adversarial networks (GANs) with dual autoencoders for fraud detection, training each autoencoder separately on normal and fraudulent datasets. This approach allowed the model to learn feature representations specific to each class, which are then combined in a dual encoding framework. Similarly, Fan *et al.* (2020) demonstrated the effectiveness of dual autoencoders in attributed networks, capturing interactions between network structure and node attributes for improved anomaly detection. Additionally, Chen *et al.* (2022a) show the potential of autoencoders in cybersecurity by explored an encoder-decoder-encoder architecture designed to learn from the distribution of normal samples only.

Further advancements in autoencoder-based approaches address the challenge of limited data in anomaly detection. For instance, Zhao *et al.* (2019) introduced a variational autoencoder

(VAE) combined with a convolutional neural network (CNN) to generate artificial vibration signals for fault diagnosis, achieving higher accuracy with small and unbalanced datasets. Gong *et al.* (2019) developed a memory-augmented autoencoder that retrieves relevant patterns during reconstruction, enhancing anomaly detection through a dual-function encoder that creates latent representations for both query and informative latent features. Similarly, Zhao *et al.* (2022) utilized a modified Wasserstein autoencoder to generate synthetic fault data, adding a gradient penalty to minimize differences between the encoder's output and a Gaussian baseline. Zhang *et al.* (2023) further advanced data augmentation by simulating rare anomalies using VAEs to enhance anomaly detection accuracy.

In image and video anomaly detection, autoencoders are widely applied. Chang *et al.* (2020) utilized a deep autoencoder to separate spatial and temporal information, improving anomaly detection in video surveillance. Sinha *et al.* (2020) used a convolutional autoencoder to detect geographic anomalies by identifying spatial deviations in remote sensing data. A hybrid approach integrating convolutional autoencoders with GANs has been shown to refine anomaly detection by leveraging pixel-level reconstruction errors alongside GAN discriminator feedback (Carrara *et al.* 2021). This approach achieved significant results in industrial inspection, particularly for detecting manufacturing defects (Zhou *et al.* 2021). For industrial image analysis, Pinon *et al.* (2023) showed the effectiveness of patch-based autoencoders in detecting anomalies in materials like wood and carpet. Furthermore, Cui *et al*. (2023) applied a multi-scale approach to anomaly detection across various materials, such as metal, wood, and plastic. These studies highlight the versatility of autoencoders in various applications.

2.2. *Application in healthcare*

Anomaly detection using autoencoders has gained traction in healthcare for diagnostic and health monitoring tools. For instance, Kraljevski *et al.* (2020) utilized convolutional autoencoders to derive health indicators from piezoelectric sensors in microfluidic valves, facilitating real-time detection of system malfunctions. During the COVID-19 pandemic, Oh *et al.* (2020) introduced a patch-based CNN with gradient-weighted class activation maps for chest X-ray diagnostics, enhancing accuracy on limited datasets by generating interpretable saliency maps. Similarly, Li *et al.* (2021) employed a stacked autoencoder for COVID-19 CT image classification, to address challenges related to small data availability and patient privacy concerns. Vaiyapuri *et al.* (2022) improved cervical cancer classification using metaheuristic optimization and stacked sparse autoencoders with preprocessing techniques like bilateral filtering and Kapur entropy-based segmentation.

Recent developments in autoencoder models emphasize denoising and extracting features from latent spaces. El-Shafai *et al*. (2022) combined a denoising autoencoder with transfer learning, improving pneumonia detection. Yu *et al.* (2023) introduced an approach to enhance performance by projecting data into mutually orthogonal subspaces for feature extraction. Georgescu (2023) used masked autoencoders to reconstruct missing fragments, to handle limited labelled data in medical imaging. Tian *et al*. (2024) explored memory-augmented cross-attention in transformers, while Shames and Kamil (2024) applied transfer learning for early COVID-19 detection using chest CT images. These advancements highlight autoencoders' versatility and efficacy in medical anomaly detection.

2.3. *Limitations of current approaches*

Current approaches to anomaly detection in healthcare applications face several limitations. A major issue is the complexity of models such as dual autoencoders, which involve multiple layers of non-linear transformations that complicate the understanding of relationships between input data and model outputs. This opacity is compounded by the lack of a clear semantic mapping between the latent space and the original input features (Higgins *et al*. 2017), making it difficult to trace the specific characteristics or patterns that influence model predictions (Guo *et al*. 2022). This lack of interpretability limits the adoption of autoencoder-based model in high-stakes applications where transparency is essential. Moreover, the training of deep autoencoder models can be computationally intensive due to their complex structures, which may not be suited for real-time anomaly detection without substantial hardware resources (Chen & Guo 2023).

Recent studies integrating autoencoders with metaheuristic approaches have shown poor performance on small datasets due to overfitting issues, as seen in the combination of ResNet50 with genetic algorithms (Abid *et al*. 2023). Several GAN-based models applied to biomedical imaging have also been evaluated across multiple datasets, yet they have not consistently delivered reliable performance on smaller datasets, where overfitting and model collapse are prevalent (Esmaeili *et al*. 2023). Furthermore, (Bao *et al*. 2024) observed that no single anomaly detection method performs equally well across all datasets, highlighting the need for more adaptable models. Models like shared autoencoders are often optimized for low-dimensional medical imaging, limiting their broader applicability in clinical settings (Jia & Liu 2023). These limitations highlight the ongoing challenges in developing robust, scalable, and interpretable anomaly detection models for healthcare.

Our proposed multi-phase dual-encoder model offers a unique solution to these challenges by enhancing both accuracy and interpretability. The model features a dual-encoder architecture that captures detailed and high-level features at the image and pixel levels, improving feature extraction over traditional single-encoder designs. The multi-phase training and validation process iteratively select the best subsets of training data and fine-tunes parameters, effectively mitigating overfitting and improving performance on small datasets. Additionally, dynamic threshold optimization based on sensitivity and specificity ensures balanced and reliable anomaly detection. By integrating interpretability techniques, the model provides transparent and understandable results, making it suitable for clinical application. These innovations collectively enhance the model's robustness and applicability across diverse medical imaging modalities.

3. Methodology

This study presents a dual-encoder model with a shared decoder to improve anomaly detection in medical imaging. The model architecture integrates two encoders: one designed to extract high-level contextual features and the other to capture fine-grained details at the pixel level. The multi-phase process involves training the model on normal samples only while using a validation set with both normal and anomalous data to iteratively refine training subsets, optimize parameters, and determine the anomaly detection threshold. This structured approach ensures the model generalizes well, even when working with limited and imbalanced datasets. For interpretability, the model applies heatmap visualizations to localize anomalies and Local Interpretable Model-Agnostic Explanations (LIME) to clarify which features influence its predictions, facilitating understanding and trust of the model.

3.1. *Model architecture*

Before applying data to the model, several preprocessing steps were implemented to ensure consistency across all datasets. All images were resized to 68*68 pixels to ensure compatibility with the model's fixed architecture, resulting in an initial input vector with 4,624 neurons (68×68). This resizing ensures that the input vector size does not exceed double the first encoder output, to maintain essential visual details and minimized distortion. Additionally, the number of neurons for each encoder and decoder was fixed, ensuring a consistent architecture throughout the model. Grayscale conversion was applied to reduce computational load since the model uses two parallel encoders. Pixel values were also normalized to a range between 0 and 1 to ensure uniformity and prevent numerical instability. Following preprocessing, images were compiled into a single NumPy array for efficient loading and memory management. These standardized steps ensure the model receives clean, uniformly processed data, and promoting more stable and reliable datasets.

The two parallel encoders work concurrently to extract complementary features. The image-level encoder captures structural patterns across the entire image, while the pixel-level encoder divides the image into smaller patches to detect localized details and subtle variations. Figure 1 illustrates this architecture, starting with the input image, which is initially flattened into a one-dimensional vector. This vector is represented as $\boldsymbol{x} \in \mathbb{R}^{A\times B\times 1}$, where A and B are the height and width, respectively, and 1 signifies the single grayscale intensity channel.

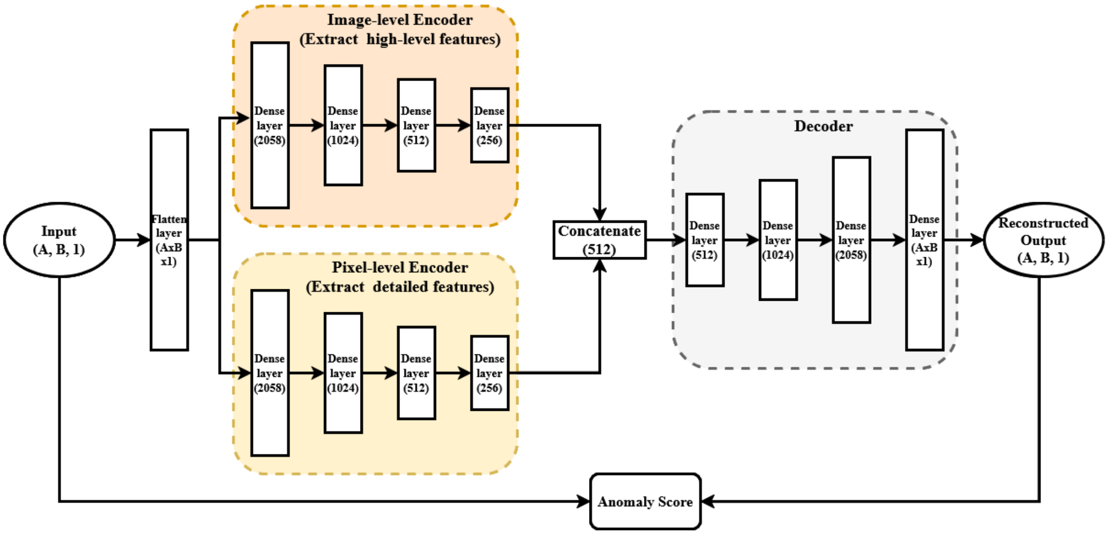


Figure 1: Architecture of the proposed dual-encoder model

Each encoder comprises four dense layers, with neurons progressively reduced from 4,624 to 256. The image-level encoder output is denoted as $\boldsymbol{h}_{img}$ and the pixel-level encoder output as $\boldsymbol{h}_{pix}$, defined by the following encoding functions:

$$\boldsymbol{h}_{img} = f_{img}(\boldsymbol{x}; \boldsymbol{W}_{img}, \boldsymbol{b}_{img}) \tag{1}$$

$$\boldsymbol{h}_{pix} = f_{pix}(\boldsymbol{x}; \boldsymbol{W}_{pix}, \boldsymbol{b}_{pix}) \tag{2}$$

where f_{img} and f_{pix} are the encoding functions, and $\boldsymbol{W}$ and $\boldsymbol{b}$ are the learnable weights and biases, respectively. The general form of an encoder was introduced by (Hinton & Salakhutdinov 2006), is expressed as as $h = f(Wx + b)$.

The features extracted from the dual encoders are fused into a unified representation of the latent space, encapsulating information from the input data while reducing dimensionality. The fused representation is obtained by combining the outputs of the image level encoder $\boldsymbol{h}_{img}$ and pixel-level encoder $\boldsymbol{h}_{pix}$ using a weighted approach:

$$\boldsymbol{h}_{concat} = \beta \boldsymbol{h}_{img} + (1 - \beta)\boldsymbol{h}_{pix} \tag{3}$$

where β is a parameter that balances the contributions of each encoder. The decoder then uses this latent space to reconstruct the original input through its own transformation function, g:

$$\widehat{\boldsymbol{x}} = g(\boldsymbol{h}_{concat}; \boldsymbol{W}_{dec}, \boldsymbol{b}_{dec}) \tag{4}$$

where $\widehat{\boldsymbol{x}}$ is the reconstructed image that should closely match the input $\boldsymbol{x}$. The decoder mirrors the encoder's structure, gradually expanding from 512 neurons to match the original image dimensions.

The model employs the Leaky ReLU activation function in the encoders to allow a small gradient flow when neuron input is negative, ensuring sustained learning and preventing inactive neurons. This function is defined as (Maas *et al.* 2013):

$$f_{Leaky\ ReLU}(x) = \begin{cases} x & if\ x > 0 \\ \alpha x & if\ x \leq 0 \end{cases}$$

where $\alpha = 0.01$. The decoder also uses Leaky ReLU, except for the final layer, which uses a sigmoid activation function to match the grayscale intensity range of the input image, constraining outputs between 0 and 1 (Rumelhart *et al.* 1988):

$$f_{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

Dropout regularization is applied in both encoders and the decoder to prevent overfitting, expressed as (Srivastava *et al*. 2014):

$$\boldsymbol{h}_{dropout}^{(l)} = \boldsymbol{r}^{(l)} \odot \boldsymbol{h}^{(l)} \tag{5}$$

where $\boldsymbol{r}^{(l)}$ is a random vector that selectively retains neuron outputs during training. The dropout rates and other parameters are dynamically optimized by the Keras Tuner, ensuring the model's adaptability and generalization across various datasets.

The anomaly score is calculated based on the reconstruction error, serving as an indicator for potential anomalies in the images. The model's objective is to minimize reconstruction error, which is quantified through Mean Squared Logarithmic Error (MSLE):

$$\mathcal{L}(\boldsymbol{x}, \widehat{\boldsymbol{x}}) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} log((\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i)^2) \tag{6}$$

where n represents the number of pixels. This error metric allows the model to concentrate on small reconstruction errors that are likely to indicate anomalies. To optimize the feature extraction process and ensure that the model captures relevant details, the training framework focuses on minimizing reconstruction loss using only normal samples. This ensures the model learns typical data patterns, making deviations more detectable as potential anomalies.

Throughout the training process, the Adam optimizer is employed to update model weights and biases, taking advantage of its efficient handling of sparse gradients and noise. Each training epoch processes data in batches, where the model iteratively refines its parameters to minimize reconstruction loss. The batch loss is calculated as an average of squared logarithmic difference across all samples as Eq. (6). This repeated training loop over multiple epochs progressively tunes the model to recognize and compress normal patterns effectively.

Algorithm 1 Dual-Encoder Model for Anomaly Detection

```
Input: Anomalous images $\boldsymbol{A}$, Normal images $\boldsymbol{N}$, Training attempt $TA$,
Output: Reconstructed images $\hat{\boldsymbol{x}}$, Evaluation metrics
1: Load $\boldsymbol{A} = \{a_1, a_2, \dots, a_{N_a}\}$ and $\boldsymbol{N} = \{n_1, n_2, \dots, n_{N_n}\}$,    $a_i, \boldsymbol{n_i} \in R^{A\times B\times \mathbb{1}}$
2. Define best model $Best = none$
3. for $TA \neq 0$ do
4.     Randomly shuffle and split $\boldsymbol{N}$ into training $\mathcal{TR}$, validation $\mathcal{N}_{val}$, and test $\mathcal{N}_{tst}$ sets;
       split $\boldsymbol{A}$ into validation $\mathcal{A}_{val}$ and test $\mathcal{A}_{tst}$ sets
5.     Define validation set $\mathcal{V} = \mathcal{N}_{val} \cup \mathcal{A}_{val}$ and $\mathcal{T} = \mathcal{N}_{tst} \cup \mathcal{A}_{tst}$; define dropout rates
       and $\beta$
6.     for $\boldsymbol{x} \in \mathcal{TR}$ do
7.         Flatten input $\boldsymbol{x} \in R^{A\times B\times \mathbb{1}}$ to $\boldsymbol{x_{flat}} \in R^{AB}$
8.         Capture global structure patterns: four dense layers with decreasing neurons
           $\boldsymbol{h}_{img}{}^{(1)} = LReLU(\boldsymbol{W}_{img}{}^{(1)}\boldsymbol{x_{flat}} + \boldsymbol{b}_{img}{}^{(1)})$ ; $\boldsymbol{h}_{DO\_img}{}^{(1)} = \boldsymbol{r}^{(1)} \odot \boldsymbol{h}_{img}{}^{(1)}$
           $\boldsymbol{h}_{img}{}^{(2)} = LReLU(\boldsymbol{W}_{img}{}^{(2)}\boldsymbol{h}_{DO\_img}{}^{(1)} + \boldsymbol{b}_{img}{}^{(2)})$ ;
               $\boldsymbol{h}_{DO\_img}{}^{(2)} = \boldsymbol{r}^{(2)} \odot \boldsymbol{h}_{img}{}^{(2)}$
           $\boldsymbol{h}_{img}{}^{(3)} = LReLU(\boldsymbol{W}_{img}{}^{(3)}\boldsymbol{h}_{DO\_img}{}^{(2)} + \boldsymbol{b}_{img}{}^{(3)})$ ;
               $\boldsymbol{h}_{DO\_img}{}^{(3)} = \boldsymbol{r}^{(3)} \odot \boldsymbol{h}_{img}{}^{(3)}$
           $\boldsymbol{h}_{img}{}^{(4)} = LReLU(\boldsymbol{W}_{img}{}^{(4)}\boldsymbol{h}_{DO\_img}{}^{(3)} + \boldsymbol{b}_{img}{}^{(4)})$
9.         Capture local structure pattern: four dense layers with decreasing neurons
           $\boldsymbol{h}_{pix}{}^{(1)} = LReLU(\boldsymbol{W}_{pix}{}^{(1)}\boldsymbol{x_{flat}} + \boldsymbol{b}_{pix}{}^{(1)})$; $h_{DO\_pix}{}^{(1)} = \boldsymbol{r}^{(4)} \odot \boldsymbol{h}_{pix}{}^{(1)}$
           $\boldsymbol{h}_{pix}{}^{(2)} = LReLU(\boldsymbol{W}_{pix}{}^{(2)}h_{DO\_pix}{}^{(1)} + \boldsymbol{b}_{pix}{}^{(2)})$; $h_{DO\_pix}{}^{(2)} = \boldsymbol{r}^{(5)} \odot \boldsymbol{h}_{pix}{}^{(2)}$
           $\boldsymbol{h}_{pix}{}^{(3)} = LReLU(\boldsymbol{W}_{pix}{}^{(3)}h_{DO\_pix}{}^{(2)} + \boldsymbol{b}_{pix}{}^{(3)})$; $h_{DO\_pix}{}^{(3)} = \boldsymbol{r}^{(6)} \odot \boldsymbol{h}_{pix}{}^{(3)}$
           $\boldsymbol{h}_{pix}{}^{(4)} = LReLU(\boldsymbol{W}_{pix}{}^{(4)}h_{DO\_pix}{}^{(3)} + \boldsymbol{b}_{pix}{}^{(4)})$
10.        Concatenate features $\boldsymbol{h}_{concat} = \beta\boldsymbol{h}_{img} + (1-\beta)\boldsymbol{h}_{pix}$
11.        Reconstruct the pattern: four dense layers with increasing neurons
           $\boldsymbol{z}^{(1)} = LReLU(\boldsymbol{W}_{dec}{}^{(1)}\boldsymbol{h}_{concat} + \boldsymbol{b}_{dec}{}^{(1)})$; $\boldsymbol{z}_{DO}{}^{(1)} = \boldsymbol{r}^{(7)} \odot \boldsymbol{z}^{(1)}$
           $\boldsymbol{z}^{(2)} = LReLU(\boldsymbol{W}_{dec}{}^{(2)}\boldsymbol{z}_{DO}{}^{(1)} + \boldsymbol{b}_{dec}{}^{(2)})$; $\boldsymbol{z}_{DO}{}^{(2)} = \boldsymbol{r}^{(8)} \odot \boldsymbol{z}^{(2)}$
           $\boldsymbol{z}^{(3)} = LReLU(\boldsymbol{W}_{dec}{}^{(3)}\boldsymbol{z}_{DO}{}^{(2)} + \boldsymbol{b}_{dec}{}^{(3)})$; $\boldsymbol{z}_{DO}{}^{(3)} = \boldsymbol{r}^{(9)} \odot \boldsymbol{z}^{(3)}$
           $\hat{\boldsymbol{x}}_{\boldsymbol{flat}}{}^{(4)} = \sigma(\boldsymbol{W}_{dec}{}^{(4)}\boldsymbol{z}_{DO}{}^{(3)} + \boldsymbol{b}_{dec}{}^{(4)})$
12.        Calculate the training reconstruction error using Eq. (6)
13.    for $\boldsymbol{x} \in \mathcal{V}$ do
14.        Repeat step 6 – 11
15.        Evaluate and determine the optimal threshold $\theta$ by maximizing sensitivity +
           specificity
16.    if evaluation on this model is higher than previous model, then $Best = this\ model$
17.    $TA = TA - 1$
18. for $\boldsymbol{x} \in \mathcal{T}$ do
19.    Repeat step 6 – 11
20.    Apply the optimal threshold $\theta *$ to classify $\boldsymbol{x} \in \mathcal{T}$ based on the reconstruction error
       in Eq. (6)
21. Calculate evaluation metrics: accuracy, sensitivity, F1-score, geometric mean, AUROC
```

```
22. Display the difference of the images $|\boldsymbol{x} - \hat{\boldsymbol{x}}|$, and generate anomaly maps by
    thresholding the difference to highlight regions with significant reconstruction error
23. Train a Random Forest Classifier on the latent feature $\boldsymbol{h}_{concat}$, for the selected samples,
    use LIME to explain the classifier's prediction
24. Compute gradient of the anomaly score by $G = \frac{\partial \mathcal{L}(x,\hat{x})}{\partial x}$
25. Create a heatmap by $H = Average(|G|)$
```

3.2. *Anomaly detection strategy*

The anomaly detection process in this study is centred on reconstruction error. Once the model completes training on normal samples, it calculates the reconstruction error for each input image during validation and testing phases. The underlying assumption is that anomalous data will have a higher reconstruction error due to the model's unfamiliarity with these patterns, while normal data will exhibit lower reconstruction errors as they align with learned patterns from the training phase. Anomaly detection relies on establishing a reconstruction error threshold θ which is determined during the validation phase. This threshold is chosen by evaluating reconstruction errors of normal and anomalous samples, defined within a range:

$$\theta \in [min(\mathcal{L}_{normal}), max(\mathcal{L}_{anomalous})] \tag{7}$$

By systematically testing different thresholds, the optimal $\theta *$ is determined by maximizing the combined sensitivity and specificity, ensuring both high recall of anomalies and accurate classification of normal samples. The optimal threshold $\theta *$ is given by:

$$\theta^* = arg\,\max_{\theta}(sensitivity + specificity) \tag{8}$$

This threshold is then applied in the testing phase, where an image is flagged as anomalous if its reconstruction error exceeds $\theta *$. To determine how well this approach distinguish between normal and anomalous images, we evaluate the model's performance using the area under the receiver operating characteristic curve (AUROC) and F1-score as primary metrics, because accuracy can be misleading on imbalanced datasets. AUROC is particularly valuable for comparing models on such datasets, as it reflects the balance between true positive rate (sensitivity) and the false positive rate. A higher AUROC suggests a better-performing model, demonstrating its ability to accurately identify true positives while minimizing false positives. On the other hand, the F1-score provides a single metric that balances precision and sensitivity, capturing the model's capability to detect anomalies accurately while minimizing both false positives and false negatives.

In addition to AUROC and F1-score, additional metrics like precision, sensitivity, and geometric mean are used to gain a comprehensive understanding of the model's performance. Precision ensures that normal samples are not flagged unnecessarily, thus minimizing false alarms. This is particularly important in healthcare, where false alarms could lead to unnecessary follow-up procedures, increased patient anxiety, and added strain on medical resources. Specificity reflecting the model's ability to correctly classify normal data, while the geometric mean accounts for both sensitivity and specificity to provide a balanced assessment of accuracy across classes. Together, these metrics provide a holistic view of the model's ability in anomaly detection, meeting the sensitive requirements of healthcare applications.

3.3. *Interpretability and transparency*

Interpretability and transparency enable clinicians to understand the reasoning behind model predictions, to build trust and effectively integrate the model into diagnostic workflows. We use heatmap generation to pinpoint areas in each image that significantly contribute to reconstruction error, overlaying color-coded regions on the original image to indicate areas most associated with anomalies. For each input image $\boldsymbol{x}$, the gradient of the reconstruction error $\mathcal{L}$ with respect to each pixel is computed as:

$$G = \frac{\partial \mathcal{L}(\boldsymbol{x},\widehat{\boldsymbol{x}})}{\partial \boldsymbol{x}} \tag{9}$$

The gradient G highlights the spatial areas the model considers anomalous, with larger values indicating higher contributions to the anomaly score. This approach is inspired by the gradient-based saliency methods proposed by (Simonyan *et al.* 2014). Averaging the absolute gradient across channels yields the heatmap H:

$$H = Average(|G|) \tag{10}$$

which visually represents the areas with the most significant reconstruction contributions.

Additionally, LIME provides further insight by identifying input features that most influence anomaly classification. As introduced by (Ribeiro *et al.* 2016), it approximates the model's complex decisions by fitting locally interpretable linear models around specific instances. For a given image $\boldsymbol{x}$, LIME perturbs the input to create similar instances $\boldsymbol{x}'$ and calculates the model's predictions on these perturbed samples. The explanation is derived from a weighted linear model fit to the perturbed instances, where the weight ω for each instance $\boldsymbol{x}'$ is based on its similarity to $\boldsymbol{x}$:

$$\omega(\boldsymbol{x},\boldsymbol{x}') = exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}\prime\|^2}{\rho^2}\right) \tag{11}$$

where ρ controls the width of the neighbourhood. This technique helps clinicians understand the specific visual characteristics that influence the model's anomaly detection decisions.

3.4. *Experimental setup*

The dual-encoder model for anomaly detection was implemented using Python 3.10 with Keras 2.10 as the core deep learning framework, and TensorFlow 2.10 as the backend. Keras was chosen for its user-friendly interface, while TensorFlow provided the necessary computational power for efficient neural network training. The development environment was Spyder 5.4.3 within Anaconda Navigator 2.6.1. The experiments were conducted on a high-performance workstation configured for deep learning tasks. The system included an Intel® Core™ i5-14400F CPU for handling neural network operations and a GeForce GTX 1660 GPU to accelerate training and inference through parallel processing. To manage large datasets efficiently, the workstation was equipped with 32GB of DDR5 RAM at 5200MHz and a 512GB SSD for rapid data retrieval and smooth computation.

This study utilizes three medical imaging benchmarks for anomaly detection across varied modalities, including brain MRI (BraTS2021), retinal OCT (RESC), and breast ultrasound (BreastMNIST). Each dataset was sourced from reputable medical institutions and research initiatives, providing a robust foundation for studying machine learning in medical imaging.

Strict data handling practices were applied to prevent data leakage and to maintain the predefined splits for training, validation, and test sets as specified by the data providers. The number of normal and anomalous samples within each set was preserved, with images reshuffled within splits as the recommended sample counts. No sample was shared across sets, ensuring that the model's performance reflects its generalization capabilities and preserving the integrity of the evaluation process.

These datasets were selected to evaluate the model adaptability to varying dataset sizes, imaging modalities, and specific challenges. BraTS2021 was chosen to assess the model's ability to handle large datasets, testing its scalability in high-volume medical imaging tasks. RESC examines the model's capability to generalize across medical domains and intermediate dataset sizes. Finally, BreastMNIST represents smaller datasets with limited normal samples, offering a realistic scenario wherenormal samples are often fewer than anomalous ones. There are only 18 normal and 138 anomalous samples in the BreastMNIST test set. On the other hand, BraTS2021 test set includes 640 normal and 3,075 anomalous samples.

Table 1: Summary of three medical imaging benchmarks

Dataset	Total Samples	Training (Normal)	Validation	Test	Image Size (Pixels)
BraTS2021	11,298	7,500	83	3,715	240*240
RESC	6,217	4,297	115	1,805	512*1,024
BreastMNIST	381	147	78	156	224*224

Both BraTS2021 and RESC were designed for unsupervised learning models trained exclusively on normal samples, while BreastMNIST includes both normal and anomalous samples in the training set. For consistency, anomalous samples in the BreastMNIST training set were removed, reducing the training size but align with the model's design to train on normal samples only.

While model parameters are adjusted during training, key hyperparameters such as batch size, epochs, training attempts, Keras Tuner iterations, and seed values are set at the beginning. These hyperparameters shape the model's training strategy, with each dataset using configurations tailored to its size, image complexity, and modality. Preliminary experiments identified optimal hyperparameter values, supporting stability and performance across datasets. A consistent TensorFlow seed of 25 was applied across datasets, while 10 trials with Keras Tuner were sufficient to dynamically tune dropout rates. Other hyperparameters are dataset-specific to maximize model robustness and adaptability across different imaging modalities, as summarized in Table 2.

Table 2: Dataset-specific hyperparameter configurations for model training

Dataset	Batch Size	Number of Epoch	Numpy Seed	Random Seed	Attempt for Training and Validation
BraTS2021	32	150	20	10	4
RESC	128	150	10	10	8
BreastMNIST	64	200	15	25	10

3.5.1. *BraTS2021 dataset*

The BraTS2021 dataset (Louis *et al*. 2021) is a benchmark dataset originating from the Brain Tumor Segementation Challenge, which is part of the MICCAI (Medical Image Computing and Computer Asisted Intervention) conferences. The dataset consists of brain MRI images, focusing primarily on glioma segmentation. The dataset includes expertly annotated regions by radiologists, making it a valuable resource for supervised and unsupervised learning tasks (Shelatkar *et al.* 2022).

3.5.2. *RESC dataset*

The RESC dataset (Hu *et al.* 2019) is a retinal optical coherence tomography dataset, known as the Retinal Edema Segmentation Challenge. RESC aims to facilitate research on retinal health, including the detection of diabetic macular edema and other retinal abnormalities. The dataset's high resolution and clinical relevance have made it a staple in ophthalmology-related machine learning research (Wang *et al*. 2021; Zhou *et al*. 2020).

3.5.3. *BreastMNIST dataset*

The BreastMNIST dataset (Yang *et al.* 2023) is part of the large MedMNIST collection, designed to provide a standardized benchmark for machine learning research in medical imaging. It contains ultrasound images of the brest, categorized into benign and malignant samples. The dataset is curated from the Breast Cancer Digital Repository, ensuring high-quality labeling for anomaly detection purposes.

4. Analysis and Results

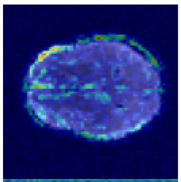
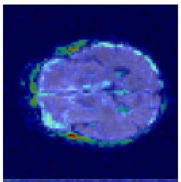
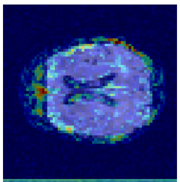
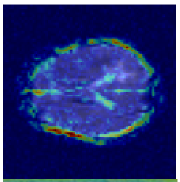
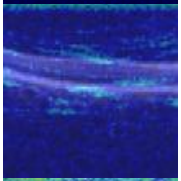
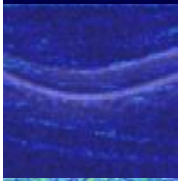
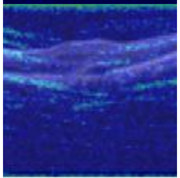
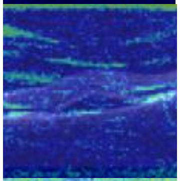
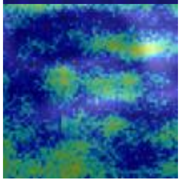
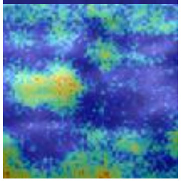
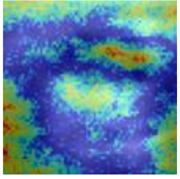
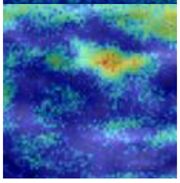
The analysis of the experimental results highlights the strong performance of our dual-encoder model across multiple metrics, as summarized in Table 3. The model demonstrates high discriminative ability on BraTS2021 and RESC datasets, with AUROC values exceeding 94% and precision nearing 100%, suggesting effective differentiation between normal and anomalous samples within these imaging modalities. The high F1-score and geometric mean values across all datasets indicate balanced classification performance, although these metrics are slightly lower on BreastMNIST. The model's ability to detect true anomalies (sensitivity) is robust for MRI and OCT images, but reduced on ultrasound images because of the limited training data available for BreastMNIST. However, the results for BreastMNIST remain acceptable, demonstrating its potential for anomaly detection even with smaller datasets. Moreover, the high precision value further implies that the model has a low false-positive rate, making it reliable for clinical use where minimizing unnecessary follow-up tests is desirable.

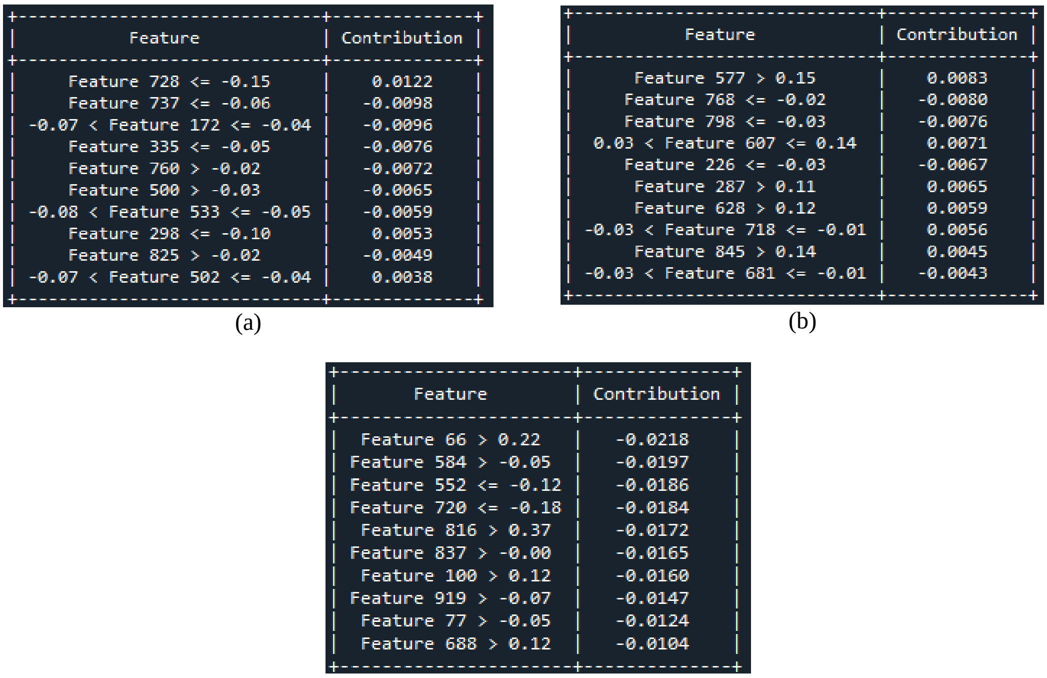
Table 4 presents selected corresponding Gradient-weighted Class Activation Maps (Grad-CAM) for normal and anomalous images in each dataset. These visualizations highlight the spatial regions contributing to the anomaly score. For BraTS2021 MRI images, the Grad-CAM maps show red/orange areas near the edges that do not correspond to actual anomalies, while the central light blue areas overlap with true anomalies. This indicates that while the model identifies the correct region, it also highlights irrelevant areas, suggesting room for improvement in spatial focus. For RESC OCT images, the Grad-CAM maps lack prominent red/orange regions in the anomalous images, with light blue areas dispersed across the image. This diffuse focus indicates that the model captures generalized features but may not target specific anomalies effectively. However, the strong performance on RESC suggests that these generalized features still contribute to accurate anomaly detection. For BreastMNIST ultrasound images, the Grad-CAM maps display a mix of light blue and occasional red/orange patches in the anomalous images, reflecting the variability in ultrasound imaging and the inherent difficulty of precise anomaly localization. This variability likely explains the slightly lower sensitivity observed for BreastMNIST. Overall, the Grad-CAM visualizations provide insight into the model's strengths and limitations in anomaly localization across different imaging modalities.

Table 3: Performance metrics of our model across datasets

Dataset	AUROC	F1-score	Precision	Sensitivity	Geometric Mean
BraTS2021	94.58	94.55	99.85	89.85	94.46
RESC	94.08	93.65	99.12	88.74	93.93
BreastMNIST	87.47	90.83	95.19	86.84	87.47

Table 4: Corresponding gradient-weighted class activation maps for normal and anomalous images

Dataset	Normal Image		Anomalous Image	
BraTS2021				
RESC				
BreastMNIST				



```
+------------------------------+--------------+
|           Feature            | Contribution |
+------------------------------+--------------+
|     Feature 728 <= -0.15     |    0.0122    |
|     Feature 737 <= -0.06     |   -0.0098    |
| -0.07 < Feature 172 <= -0.04 |   -0.0096    |
|     Feature 335 <= -0.05     |   -0.0076    |
|     Feature 760 > -0.02      |   -0.0072    |
|     Feature 500 > -0.03      |   -0.0065    |
| -0.08 < Feature 533 <= -0.05 |   -0.0059    |
|     Feature 298 <= -0.10     |    0.0053    |
|     Feature 825 > -0.02      |   -0.0049    |
| -0.07 < Feature 502 <= -0.04 |    0.0038    |
+------------------------------+--------------+
```

(a)

```
+------------------------------+--------------+
|           Feature            | Contribution |
+------------------------------+--------------+
|      Feature 577 > 0.15      |    0.0083    |
|     Feature 768 <= -0.02     |   -0.0080    |
|     Feature 798 <= -0.03     |   -0.0076    |
|  0.03 < Feature 607 <= 0.14  |    0.0071    |
|     Feature 226 <= -0.03     |   -0.0067    |
|      Feature 287 > 0.11      |    0.0065    |
|      Feature 628 > 0.12      |    0.0059    |
| -0.03 < Feature 718 <= -0.01 |    0.0056    |
|      Feature 845 > 0.14      |    0.0045    |
| -0.03 < Feature 681 <= -0.01 |   -0.0043    |
+------------------------------+--------------+
```

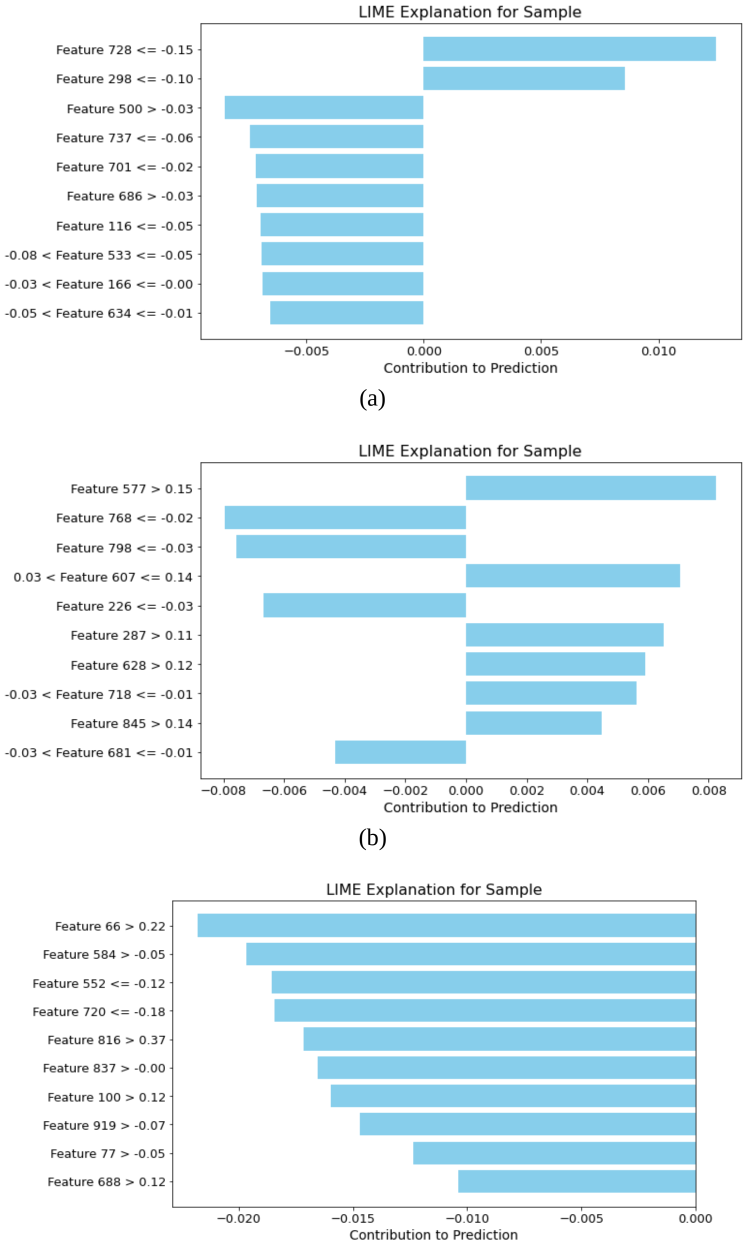
(b)

```
+-----------------------+--------------+
|        Feature        | Contribution |
+-----------------------+--------------+
|  Feature 66 > 0.22    |   -0.0218    |
| Feature 584 > -0.05   |   -0.0197    |
| Feature 552 <= -0.12  |   -0.0186    |
| Feature 720 <= -0.18  |   -0.0184    |
|  Feature 816 > 0.37   |   -0.0172    |
| Feature 837 > -0.00   |   -0.0165    |
|  Feature 100 > 0.12   |   -0.0160    |
| Feature 919 > -0.07   |   -0.0147    |
|  Feature 77 > -0.05   |   -0.0124    |
|  Feature 688 > 0.12   |   -0.0104    |
+-----------------------+--------------+
```

(c)

Figure 2: Top contributing features identified by LIME for (a) BraTS2021, (b) RESC, and (c) BreastMNIST

The LIME analysis across these datasets identifies and visualizes influential latent features, providing insight into the factors driving the model's anomaly predictions. Figure 2 shows the top latent features influencing anomaly detection, while Figure 3 provides the corresponding feature contribution plots, showing both positive and negative influences on the model's predictions. For example, in Figure 2(a), the condition Feature 728 ≤ -0.15 with a contribution value of 0.012 indicates that when Feature 728 is less than or equal to -0.15, it pushes the prediction towards the anomalous class. Conversely, Feature 737 ≤ -0.06 with a contribution of -0.010 suggests that when Feature 737 is less than or equal to -0.06, it pushes the prediction towards the normal class. This analysis shows that the model relies on multiple features to make predictions in these datasets.



(a)

(b)

(c)

Figure 3: LIME feature contribution plot for (a) BraTS2021, (b) RESC, and (c) BreastMNIST

In the BraTS2021 dataset, high-contribution features such as Feature 728 and Feature 298 serve as strong indicators for anomaly detection. These features likely represent complex patterns associated with tumour-like structures in brain regions, showcasing the model's ability to learn relevant latent features. Features with negative contributions, such as Feature 737 and Feature 172, help the model balance its predictions by identifying normal areas. For the RESC dataset, the model relies on a broader array of subtle features, as indicated by the dispersed nature of influential latent features. This reflects the complexity of retinal OCT images, where anomalies may manifest as subtle variations in texture or shape. For the BreastMNIST dataset, the LIME results primarily reveal negative contributions, suggesting that the model focuses on distinguishing normal patterns rather than detecting specific indicators of anomalies. This behaviour aligns with the inherent variability in ultrasound imaging, where anomalies are more challenging to localize precisely, likely contributing to the slightly lower sensitivity observed.

Table 5: Comparison of AUROC results across different methods

Publication	Methods	BraTS2021	RESC	BreastMNIST
ACCV'18	GANomaly	74.79 ± 1.93	52.56 ± 3.95	66.43 ± 1.33
ICML'18	DeepSVDD	86.98 ± 0.66	74.17 ± 1.29	**73.66 ± 2.11**
MedIA'19	f-AnoGAN	77.26 ± 0.18	77.42 ± 0.85	65.19 ± 2.01
CVPR'19	CutPaste	78.81 ± 0.67	**90.23 ± 0.61**	72.82 ± 2.27
ICCV'21	PaDiM	79.02 ± 0.38	75.87 ± 0.56	67.33 ± 0.49
ICLR'22	UTRAD	82.92 ± 2.32	89.39 ± 1.92	70.58 ± 5.99
CVPR'22	PatchCore	**91.65 ± 0.36**	**91.55 ± 0.10**	62.42 ± 0.68
BMVC'22	CFLOW	74.82 ± 5.32	74.95 ± 5.81	54.34 ± 2.47
WACV'22	CS-Flow	**90.91 ± 0.83**	87.34 ± 0.58	**81.87 ± 0.55**
CVPR'23	SimpleNet	82.52 ± 3.34	76.15 ± 7.46	72.50 ± 2.62
JARASET'24	FT-CNN	88.66 ± 1.07	87.52 ± 2.18	68.18 ± 0.67
	Our Model	**94.58**	**94.08**	**87.47**

[a]values in bold indicate the top three AUROC scores for each dataset

Table 6: Comparison of F1-score results across different methods

Publication	Methods	BraTS2021	RESC	BreastMNIST
ACCV'18	GANomaly	72.91 ± 1.06	66.41 ± 2.56	57.34 ± 2.02
ICML'18	DeepSVDD	**88.95 ± 4.10**	70.11 ± 0.10	78.16 ± 2.40
MedIA'19	f-AnoGAN	73.51 ± 1.52	78.86 ± 2.09	59.66 ±2.23
CVPR'19	CutPaste	76.23 ± 1.36	**88.11 ± 0.74**	73.33 ± 1.46
ICCV'21	PaDiM	79.94 ± 0.05	73.47 ± 0.54	68.77 ± 0.98
ICLR'22	UTRAD	79.10 ± 1.47	71.99 ± 1.38	78.65 ± 6.95
CVPR'22	PatchCore	82.24 ± 0.04	**91.60 ± 0.09**	66.67 ± 1.08
BMVC'22	CFLOW	76.78 ± 6.49	60.13 ± 0.86	**79.67 ± 0.32**
WACV'22	CS-Flow	**88.53 ± 0.94**	80.94 ± 0.04	**86.21 ± 0.64**
CVPR'23	SimpleNet	85.07 ± 1.26	75.16 ± 5.79	72.86 ± 1.33
JARASET'24	FT-CNN	84.95 ± 0.64	77.38 ± 4.02	61.49 ± 1.83
	Our Model	**94.55**	**93.65**	**90.83**

[a]values in bold indicate the top three F1-scores for each dataset

The performance of our model was evaluated against several existing anomaly detection methods across the three datasets, as summarized in Tables 5 and 6. Table 5 presents the AUROC scores, while Table 6 provides the F1-scores results, highlighting the comparative effectiveness of each method. Our model consistently achieved high AUROC values for the BraTS2021, RESC, and BreastMNIST datasets, with scores of 94.58%, 94.08%, and 87.47%, respectively. These results outperform other methods, including GANomaly, DeepSVDD, and CutPaste, as shown in Table 5, showcasing the robustness of our dual-encoder approach. For instance, on the RESC dataset, GANomaly achieved only 52.56% AUROC, while our model reached 94.08%, demonstrating a significant improvement in detection accuracy. The model's superior performance is further validated by F1-scores, which reflect a balanced trade-off between precision and sensitivity. Our model achieved 94.55% for BraTS2021, 93.65% for RESC, and 90.83% for BreastMNIST, consistently outperforming all the comparison methods in this metric. This performance can be attributed to the dual-encoder's ability to effectively reduce dimensionality while preserving essential features necessary for accurate anomaly detection. The high F1-scores also indicate that the model minimizes both false positives and false negatives, making it particularly reliable for clinical applications where diagnostic errors can have significant consequences.

5. Conclusion

The combination of image-level and pixel-level encoders in our dual-encoder architecture enables the model to capture features at different levels of abstraction, ensuring adaptability across diverse datasets. The model demonstrated state-of-the-art performance in anomaly detection, consistently outperforming existing methods in terms of AUROC and F1-score across BraTS2021, RESC, and BreastMNIST datasets. The high precision and sensitivity achieved validate the model's reliability in detecting anomalies while minimizing false positives and negatives, making it suitable for clinical applications. Benchmarking against existing methods, including GANomaly, DeepSVDD, and PatchCore, highlighted the unique strengths of our methodology, particularly the dual-encoder design and multi-phase training process, which iteratively optimizes performance and adapts to dataset variability. Unlike conventional approaches, the model handles both small and large datasets without relying on transfer learning or data augmentation, showcasing its robustness and scalability. However, challenges remain in the interpretability of the latent space and the model's ability to accurately localize anomalies. Future work should integrate mechanisms, such as anomaly-focused loss functions or spatial attention modules, to improve the model's focus on anomalous regions, enhancing both localization accuracy and interpretability. Additionally, adaptive hyperparameter tuning methods and expansion to other medical imaging modalities could further validate the model's utility and versatility in anomaly detection.

Acknowledgments

This research was supported by the Ministry of Higher Education Malaysia (MOHE) through the MyBrainSc Scholarship and Universiti Sains Malaysia.

References

Abid M.H., Ashraf R., Mahmood T. & Faisal C.M.N. 2023. Multi-modal medical image classification using deep residual network and genetic algorithm. *PLoS ONE* **18**(6): e0287786.

Addo P.M., Guegan D. & Hassani B. 2018. Credit risk analysis using machine and deep learning models. *Risks* **6**(2): 38.

Alowais S.A., Alghamdi S.S., Alsuhebany N., Alqahtani T., Alshaya A.I., Almohareb S.N., Aldairem A., Alrashed M., Bin Saleh K., Badreldin H.A., Al Yami M.S., Al Harbi S. & Albekairy A.M. 2023. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education* **23**(1): 689.

Bao J., Sun H., Deng H., He Y., Zhang Z. & Li X. 2024. BMAD: Benchmarks for medical anomaly detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4042–4053.

Baurasien B.K., Alareefi H.S., Almutairi D.B., Alanazi M.M., Alhasson A.H., Alshahrani A.D., Almansour S.A., Alshagag Z.A., Alqattan K.M. & Alotaibi H.M. 2023. Medical errors and patient safety: Strategies for reducing errors using artificial intelligence. *International Journal of Health Sciences* **7**(S1): 3471–3487.

Bengio Y., Courville A. & Vincent P. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(8): 1798–1828.

Berahmand K., Daneshfar F., Salehi E.S., Li Y. & Xu Y. 2024. Autoencoders and their applications in machine learning: a survey. *Artificial Intelligence Review* **57**(2): 28.

BERNAMA. 2023. Moh probing alleged misdiagnosis leading to patient's death at hsah. https://www.bernama.com/en/news.php?id=2252540 (25 Oktober 2024).

Bhatt N., Bhatt N., Prajapati P., Sorathiya V., Alshathri S. & El-Shafai W. 2024. A data-centric approach to improve performance of deep learning models. *Scientific Report* **14**(1): 22329

Brownlee J. 2020. Impact of dataset size on deep learning model skill and performance estimates https://machinelearningmastery.com/impact-of-dataset-size-on-deep-learning-model-skill-and-performance-estimates/ (26 Oktober 2024).

Carrara F., Amato G., Brombin L., Falchi F. & Gennaro C. 2021. Combining GANs and AutoEncoders for efficient anomaly detection. *25th International Conference on Pattern Recognition*, pp. 3939-3946.

Chang Y., Tu Z., Xie W. & Yuan J. 2020. Clustering driven deep autoencoder for video anomaly detection. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, Proceedings, Part XV*, pp. 329-345.

Chatterjee S., Saad F., Sarasaen C., Ghosh S., Krug V., Khatun R., Mishra R., Desai N., Radeva P., Rose G., Stober S., Speck O. & Nürnberger A. 2024. Exploration of interpretability techniques for deep COVID-19 classification using chest x-ray images. *Journal of Imaging* **10**(2): 45.

Chen L., Li Y., Deng X., Liu Z., Lv M. & Zhang H. 2022a. Dual auto-encoder GAN-based anomaly detection for industrial control system. *Applied Sciences* **12**(10): 4986.

Chen X., Wang X., Zhang K., Fung K.M., Thai T.C., Moore K., Mannel R.S., Liu H., Zheng B. & Qiu Y. 2022b. Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis* **79**: 102444.

Churová V., Vyškovský R., Maršálová K., Kudlácek D. & Schwarz D. 2021. Anomaly detection algorithm for real-world data and evidence in clinical research: Implementation, evaluation, and validation study. *JMIR Medical Informatics* **9**(5): e27172.

Cui Y., Liu Z. & Lian S. 2023. A survey on unsupervised anomaly detection algorithms for industrial images. *IEEE Access* **11**: 55297–55315.

Dong B., Zhu Y., Li L. & Wu X. 2020. Hybrid collaborative recommendation via dual-autoencoder. *IEEE Access* **8**: 46030–46040.

El-Shafai W., El-Nabi S.A., El-Rabaie E.S.M., Ali A.M., Soliman N.F., Algarni A.D. & Abd El-Samie F.E. 2022. Efficient deep-learning-based autoencoder denoising approach for medical image diagnosis. *Computers, Materials and Continua* **70**(3): 6107-6125.

Esmaeili M., Toosi A., Roshanpoor A., Changizi V., Ghazisaeedi M., Rahmim A. & Sabokrou M. 2023. Generative adversarial networks for anomaly detection in biomedical imaging: A study on seven medical image datasets. *IEEE Access* **11**: 17906–17921.

Fan H., Zhang F. & Li Z. 2020. Anomalydae: Dual autoencoder for anomaly detection on attributed networks. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain*, pp. 5685–5689.

Georgescu M.I. 2023. Masked autoencoders for unsupervised anomaly detection in medical images. *Procedia Computer Science* **225**: 969–978.

Gong D., Liu L., Le V., Saha B., Mansour M.R., Venkatesh S. & Van Den Hengel A. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1705-1714.

Guo X., Gichoya J.W., Purkayastha S. & Banerjee I. 2022. CVAD: An anomaly detector for medical images based on cascade VAE. *Workshop on Medical Image Learning with Limited and Noisy Data*, pp. 187-196.

Higgins I., Matthey L., Pal A., Burgess C., Glorot X., Botvinick M., Mohamed S. & Lerchner A. 2017. Beta-VAE: Learning basic visual concepts with a constrained variational framework. *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017*.

Hinton G.E. & Salakhutdinov R.R. 2006. Reducing the dimensionality of data with neural networks. *Science* **313**(5786): 504–507.

Hu J., Chen Y. & Yi Z. 2019. Automated segmentation of macular edema in OCT using deep neural networks. *Medical Image Analysis* **55**: 216–227.

Hussain F., Cooper A., Carson-Stevens A., Donaldson L., Hibbert P., Hughes T. & Edwards A. 2019. Diagnostic error in the emergency department: Learning from national patient safety incident report analysis. *BMC Emergency Medicine* **19**: 77.

Jia H. & Liu W. 2023. Anomaly detection in images with shared autoencoders. *Frontiers in Neurorobotics* **16**: 1046867.

Kraljevski I., Duckhorn F., Tschoepe, C. & Wolff, M. 2020. Convolutional autoencoders for health indicators extraction in piezoelectric sensors. *2020 IEEE Sensors*, pp. 1–4.

Li D., Fu Z. & Xu J. 2021. Stacked-autoencoder-based model for COVID-19 diagnosis on CT images. *Applied Intelligence* **51**(5): 2805–2817.

Louis D.N., Perry A., Wesseling P., Brat D.J., Cree I.A., Figarella-Branger D., Hawkins C., Ng H.K., Pfister S.M., Reifenberger G., Soffietti R., Von Deimling A. & Ellison D.W. 2021. The 2021 WHO classification of tumors of the central nervous system: A summary. *Neuro-Oncology* **23**(8): 1231–1251.

Maas A.L., Hannun A.Y. & Ng A.Y. 2013. Rectifier nonlinearities improve neural network acoustic models. *Proceedings of the 30th International Conference on Machine Learning* **30**(1).

Medvedieva K., Tosi T., Barbierato E. & Gatti A. 2024. Balancing the scale: Data augmentation techniques for improved supervised learning in cyberattack detection. *Eng* **5**(3): 2170–2205.

Oh Y., Park S. & Ye J.C. 2020. Deep learning COVID-19 features on CXR using limited training data sets. *IEEE Transactions on Medical Imaging* **39**(8): 2688–2700.

Pinon N., Trombetta R. & Lartizien C. 2023. Anomaly detection in image or latent space of patch-based auto-encoders for industrial image analysis. *arXiv preprint arXiv:2307.02495.*

Ribeiro M.T., Singh S. & Guestrin C. 2016. "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144.

Chen S. & Guo W. 2023. Auto-encoders in deep learning–A review with new perspectives. *Mathematics* **11**(8): 1777.

Rumelhart D.E., Hinton G.E. & Williams R.J. 1988. Learning internal representations by error propagation. In Collins A. & Smith E.E. (eds.). *Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence*: 399–421. California: Morgan Kaufmann Publishers.

Shames M.A. & Kamil M. 2024. Lung infection detection via CT images and transfer learning techniques in deep learning. *Journal of Advanced Research in Applied Sciences and Engineering Technology* **47**(1): 206–218.

Shelatkar T., Urvashi D., Shorfuzzaman M., Alsufyani A. & Lakshmanna K. 2022. Diagnosis of brain tumor using light weight deep learning model with fine-tuning approach. *Computational and Mathematical Methods in Medicine* **2022**: 2858845.

Simonyan K., Vedaldi A. & Zisserman A. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. *Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014.*

Sinha S., Giffard-Roisin S., Karbou F., Deschatres M., Karas A., Eckert N., Coléou C. & Monteleoni C. 2020. Variational Autoencoder Anomaly-Detection of Avalanche Deposits in Satellite SAR Imagery. *Proceedings of the 10th International Conference on Climate Informatics*, pp. 113-119.

Srivastava N., Hinton G., Krizhevsky A., Sutskever I. & Salakhutdinov R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15**(1): 1929–1958.

Su T., Sun H., Zhu J., Wang S. & Li Y. 2020. BAT: Deep learning methods on network intrusion detection using NSL-KDD dataset. *IEEE Access* **8**: 29575–29585.

Tang C., Luktarhan N. & Zhao Y. 2020. SAAE-DNN: Deep learning method on intrusion detection. *Symmetry* **12**(10): 1695.

The Nation. 2016. Man spends 43 years in wheelchair on wrong diagnosis. https://www.nationthailand.com/in-focus/30296185 (25 Oktober 2024).

Tian Y., Pang G., Liu Y., Wang C., Chen Y., Liu F., Singh R., Verjans J.W., Wang M. & Carneiro G. 2024. Unsupervised anomaly detection in medical images with a memory-augmented multi-level cross-attentional masked autoencoder. In Cao X., Xu X., Rekik I., Cui Z. & Ouyang X. (eds.). *Machine Learning in Medical Imaging*. *Lecture Notes in Computer Science*:**14349**: 11-21. Cham: Springer.

Vaiyapuri T., Alaskar H., Syed L., Aljohani E., Alkhayyat A., Shankar K. & Kumar S. 2022. Modified metaheuristics with stacked sparse denoising autoencoder model for cervical cancer classification. *Computers and Electrical Engineering* **103**: 108292.

Wang Z., Zhong Y., Yao M., Ma Y., Zhang W., Li C., Tao Z., Jiang Q. & Yan B. 2021. Automated segmentation of macular edema for the diagnosis of ocular disease using deep learning method. *Scientific Reports* **11**(1): 13392.

Wu E., Cui H. & Welsch R.E. 2020. Dual autoencoders generative adversarial network for imbalanced classification problem. *IEEE Access* **8**: 91265–91275.

Yang J., Shi R., Wei D., Liu Z., Zhao L., Ke B., Pfister H. & Ni B. 2023. MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data* **10**(1): 41.

Yu Q., Li C., Zhu Y. & Kurita T. 2023. Convolutional autoencoder based on latent subspace projection for anomaly detection. *Methods* **214**: 48–59.

Zhang C., Wang X., Zhang J., Li S., Zhang H., Liu C. & Han P. 2023. VESC: a new variational autoencoder based model for anomaly detection. *International Journal of Machine Learning and Cybernetics* **14**(3): 683–696.

Zhao D., Liu S., Gu D., Sun X., Wang L., Wei Y. & Zhang H. 2019. Enhanced data-driven fault diagnosis for machines with small and unbalanced data based on variational auto-encoder. *Measurement Science and Technology* **31**(3): 035004.

Zhao K., Jiang H., Liu C., Wang Y. & Zhu K. 2022. A new data generation approach with modified Wasserstein auto-encoder for rotating machinery fault diagnosis with limited fault data. *Knowledge-Based Systems* **238**: 107892.

Zhou K., Xiao Y., Yang J., Cheng J., Liu W., Luo W., Gu Z., Liu J. & Gao S. 2020. Encoding Structure-Texture Relation with P-Net for Anomaly Detection in Retinal Images. In Vedaldi A., Bischof H., Brox T. & Frahm J.M. (eds.). *Computer vision—ECCV 2020*: 360-377. Cham: Springer International Publishing.

Zhou Q., Mei J., Zhang Q., Wang S. & Chen G. 2021. Semi-supervised fabric defect detection based on image reconstruction and density estimation. *Textile Research Journal* **91**(9–10): 962-972.

Zhuang F., Qi Z., Duan K., Xi D., Zhu Y., Zhu H., Xiong H. & He Q. 2021. A comprehensive survey on transfer learning. *Proceedings of the IEEE* **109**(1): 43–76.

School of Mathematical Sciences
Universiti Sains Malaysia
11800 USM Penang
Pulau Pinang PM, MALAYSIA
E-mail: rusyidahazri@student.usm.my, saratha@usm.my[*]*, majidkhanmajaharali@usm.my*

Received: 25 November 2024
Accepted: 27 December 2024

[*] Corresponding author