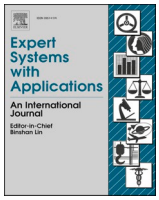
Expert Systems With Applications 271 (2025) 126596

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



An optimized multi-head attention based fused depthwise convolutional model for lung cancer detection

Sadam Kavitha [a,*], Eswar Patnala [a], Hrushikesava Raju Sangaraju [a,b], Rajesh Bingu [a], Salina Adinarayana [b,c], Jagjit Singh Dhatterwal [a]

[a] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Greenfields, Vaddeswaram, Guntur 522302, India
[b] Singapore Institute of Technology, Singapore
[c] Department of Computer Science and Systems Engineering, College of Engineering, Andhra University, Visakhapatnam, Guntur 522302, India

ARTICLE INFO

Keywords:
Depthwise convolutional model
Vision transformer
Equilibrium
Black widow optimization
Multi-head attention
Wiener filter

ABSTRACT

Lung cancer remains one of the leading causes of death globally, with over five million fatalities recorded annually. Early detection through computed tomography (CT) scans is crucial in improving patient outcomes. However, existing deep learning models for lung cancer classification often face challenges such as overfitting and computational inefficiency, limiting their real-world applicability. To overcome these limitations, this study proposes an optimized Multi-Head Attention-based Fused Depthwise Convolutional Neural Network (MHA-DCNN). The IQ-OTH/NCCD Lung Cancer Dataset is utilized for evaluation. The data is pre-processed using a Disperse Wiener filter to remove noise from CT scan images. Relevant features are extracted using a Convolutional Vision Transformer (CViT), and the most significant features are selected using the Enhanced Binary Black Widow Optimization (EBWO) technique. The MHA-DCNN model classifies benign and malignant tumors using these optimized features, while Adaptive Equilibrium Optimization (AEO) is employed for hyperparameter tuning to enhance learning efficiency. Experimental results demonstrate the proposed model's superiority, achieving 99.43% accuracy in classifying normal cases, 99.51% for malignant cases, and 99.66% for benign cases. Precision scores of 99.73% for normal, 99.91% for malignant, and 99.46% for benign tumors further validate the model's reliability. Additionally, F-scores of 99.33% (normal), 99.61% (malignant) and 99.56% (benign) highlight its robust performance. These results underscore the effectiveness of the proposed MHA-DCNN framework, which offers improved classification accuracy while addressing the limitations of existing methods. The framework sets a strong foundation for advancing early lung cancer detection through deep learning, with potential for broader application across diverse medical imaging datasets.

1. Introduction

Lung cancer is the leading cause of death, accounting for five million fatal cases annually (Raza et al., 2023). The inability to diagnose lung cancer until it has progressed to an advanced stage is the fundamental reason for its classification as one of the most hazardous diseases (Nooreldeen & Bach, 2021). The dangerous lung cancer originates in the spongy organs in the chest that endure the breathing process (Lu et al., 2021). Lung cancer is more common among smokers, but it can nonetheless develop in nonsmokers. Some signs of lung cancer include shortness of breath, chest pain, hoarseness, and headaches (Qadhi et al., 2023). This cancer is caused by harmful cells that grow in the lungs. Therapy for lung cancer is surgery, chemotherapy, immunotherapy, radiation and targeted drugs (Ahern et al., 2021).

Treatment of lung cancer has been more effective with early detection of cancer, and several procedures have been devised to facilitate this process. There are two types of lung cancer: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). The most frequent type of lung cancer is non-small cell lung carcinoma. NSCLC, which includes squamous cell carcinoma and adenocarcinoma, accounts for more than 80 % of all lung cancer cases. Sarcomatoid carcinoma and adenosquamous carcinoma are two less common types of non-small cell lung cancer. Small-cell lung cancer spreads faster than NSCLC and is extremely difficult to cure (Wang et al., 2021). It is typically diagnosed as a small lung tumor that has migrated to other parts of the body. SCLC is classified into two types: small cell carcinoma (oat cell carcinoma) and

* Corresponding author.
E-mail address: sadamkavitha78@gmail.com (S. Kavitha).

https://doi.org/10.1016/j.eswa.2025.126596
Received 10 June 2024; Received in revised form 20 December 2024; Accepted 16 January 2025
Available online 22 January 2025

mixed small cell carcinoma (Rudin et al., 2021). In both cases, early detection may help in minimizing the risk of death using radiologic screening.

Many studies are underway, such as non-invasive early lung cancer diagnosis using genomic features fed into machine learning-based models. These models are also used for other detection, like brain tumor segmentation in an effective manner (Liu et al., 2021; Zhu et al., 2023, 2024). Even though the features are inefficient in detection due to interpretation (Chabon et al., 2020). Biomarkers for lung cancer detection by electrochemical biosensors are also created at a high cost (Khanmohammadi et al., 2020). The markers are not so efficient in detection due to laboratory errors. The artificial neural network XGBoost (Liu et al., 2022; Liu and Bao, 2023) is used in lung cancer detection, which processes symptoms as features (Nasser & Abu-Naser, 2019). Clustering-based methods are also created with trained neural networks which have an energy loss issue (Shakeel et al., 2019). Several machine learning and deep learning models have been created to produce adequate results for the identification of lung cancer. Because of the complex design and concerns such as overfitting and interpretation, each model struggles to produce correct results.

Cancer is identified using machine learning and deep learning models, such as support vector machines and AlexNets. This model produces better results, but it is erroneous across many datasets (Naseer et al., 2023). Among various networks, convolutional neural network (CNN) gives better results with less cost in the detection of lung cancer. Moreover, Bio-inspired algorithms like whale (Rana et al., 2020) and adaptive particle swarm algorithms (Xue et al., 2019) are used with the CNN model to enhance the classification results. This model also possesses disadvantages, like the incapability of providing results in different modalities of medical imaging and the fact that it requires large memory (Vijh et al., 2023). As a result, a novel model is required to address the challenges that may develop in existing research studies and achieve efficient outcomes in early cancer detection.

1.1. Motivation

Lung cancer is a hazardous medical condition that has to be detected earlier using different works in recent times. Cancer data is collected using Internet of Things (IoT) medical devices and sent to a server. Because of these limitations, current approaches to detecting lung cancer are ineffective in analyzing large databases, resulting in lower performance. The fundamental issue of these models is that they are ineffective in the current epidemic environment because physical treatment is not practical. Even though the networks have some problems, existing deep learning-based models are used to make the detection accurate. Long-term dependencies, large computing needs that take time during the model's training phase, overfitting, and a lack of feasibility are all potential problems in detection. The work's primary contribution is listed below,

- To introduce an optimized Multi-head attention based fused depthwise convolutional model for effective lung cancer classification
- To present a Disperse Wiener filter for pre-processing the images from the dataset that removes noises faster.
- To deploy a Convolutional Vision Transformer to extract features with high efficiency and resolve overfitting challenges.
- To present an Enhanced binary black widow optimization for selecting relevant features to reduce the complexity.
- To implement a Multi-head attention-based fused depthwise CNN model for classification and hyperparameter tuning by adaptive equilibrium optimization approach.

The paper is organized with the flow from section 1, which gives an introduction to lung cancer and its early detection mechanism. Section 2 follows the introduction and provides an overview of existing detection research. Section 3 includes a full discussion of the proposed methodology. The evaluation results and discussions are presented in section 4. The final outcome and expansion of the research work are presented in Section 5.

2. Literature survey

This section covers some surveys of existing works on lung cancer detection; the works are explained clearly, along with their disadvantages.

A more accurate fuzzy-based lung cancer detection system was developed by Akter et al. (2021). To increase the segmentation accuracy of these images, the suggested technique used median values measured along each row and column in addition to the maxima and minima values. The next round of analysis was classifying the segmented lung nodules as benign or malignant using a neuro-fuzzy classifier. The performance assessment used three metrics: sensitivity, specificity, and accuracy. The suggested methodology yielded findings with 100 % sensitivity, 81 % specificity, 86 % accuracy, and 90 % precision. Among the advancements, the model is less efficient in classification for different detecting specific stages due to ineffective feature selection.

Jena et al. (2021) created a region-based CNN model for the diagnosis of lung cancer using Gaussian distributions. Noises are removed from the dataset's photos using Gaussian and Wiener filters. The region of interest (ROI) was then properly defined using the region-growing segmentation approach. Closed pixels are collapsed with the bigger region, and seed points are chosen in the region-growing segmentation process. Using this method, features were retrieved and then passed to the deep Gaussian mixture model in the region-based convolutional neural network (DGMM-RBCNN). To reduce overfitting, the Gaussian distribution was used for each parameter in the CNN layers. In the evaluation, the aforementioned model achieved an accuracy of 87.79 %. The model has a rather high computational cost and is hampered by geometric distortions.

A deep learning model for the efficient and less complex diagnosis of lung cancer was suggested by Sori et al. (2021). For the detection process, the model proposed a two-path CNN (DFD-Net) denoising first method. An end-to-end approach was taken to provide the denoising and detection phases. A residual learning denoising model (DR-Net) was included during the pre-processing phase in order to eliminate the noise. Next, a two-path convolutional neural network is trained with the denoised image produced by DR-Net to detect lung cancer. The combined integration of local and global aspects was the main goal of the two routes. Each route used various receptive field sizes to help the model in both global and local dependency. In the evaluation, the model attained less accuracy of 0.878 in detection. This inefficient result was made by losing some details in the denoising phase.

A segmentation technique for lung cancer diagnosis using chest radiographs was projected by Shimazaki et al. (2022). The CNN-based architecture was used for the segmentation by measuring the diameter of a particular tumour. The CNN architecture also included encoding and decoding layers to lower feature map resolution. Metrics such as mean false positive indication pre-image (mFPI) and sensitivity were used to assess the model throughout the evaluation phase. Using 0.13 mFPI, the DL-based model yielded a sensitivity of 0.73. Although the model's high sensitivity with low FPs limits its usefulness in lung cancer detection, a screening cohort may have a larger FP count. As they may lower the identification rate of benign nodules/masses, only pathologically established lung tumors and pixel-level annotations by two radiologists are used in this investigation.

Chui et al. (2023) used a modified generative adversarial network (MTL-MGAN) to create a machine learning-based model for lung cancer identification. Information within the prioritized source and target domain is used to select the prioritized dataset to increase the transferability of the multi-round transfer learning process. An intermediate domain is used by the MGAN model to bridge the gaps between the source and target destinations, and more data is used for training. When

the maximum number of rounds is used, the model becomes more complex and operates more slowly. Table 1 presents an overview of previous works and their drawbacks.

From the analysis of existing work, the model needs to be developed to examine every feature for minute detection of lung cancer. Existing models require higher costs to design such detection models. The major flaws of existing works are inaccurate results due to the maximum number of iteration rounds, and lung cancer detection is limited due to high sensitivity with low FPs. Many important details were lost during the denoising phase, making the classification of distinct phases less efficient due to inefficient feature selection. Therefore, the suggested model is required to address the shortcomings of the current research through improved methodologies.

3. Proposed methodology

The proposed model leverages the IQ-OTH/NCCD lung cancer dataset to classify lung cancer as normal, benign, or malignant. The novelty of this study lies in the development of the MHA-DCNN a cutting-edge framework designed to address the critical limitations of existing lung cancer classification models. Traditional deep learning models often face challenges such as overfitting, high computational costs, and an inability to generalize effectively across diverse datasets. The MHA-DCNN overcomes these issues by integrating a multi-head attention mechanism with depthwise convolutions, enabling the model to focus on the most critical regions of CT scan images while maintaining computational efficiency. The fusion of these components ensures that the model processes information with enhanced spatial awareness, significantly mitigating the overfitting problem commonly observed in conventional approaches.

Table 1
Overview of existing works and its drawbacks.

Author & Reference	Approach	Dataset	Performance	Disadvantages
Akter et al. (2021)	Fuzzy based approach	LIDC	Achieve accuracy of 86 %	Models possess less efficiency in classification for different detecting specific stages due to ineffective feature selection.
Jena et al. (2021)	DGMM-RBCNN	LIDC	Achieve 87.79 % of accuracy	The model has a comparatively high computational cost and is also compromised by geometric distortions.
Sori et al. (2021)	DFD-Net	LUNA16 dataset	Obtain an accuracy of 0.878	In the denoising phase, the details were lost.
Shimazaki et al. (2022)	CNN	Chest radiographs	Obtained a sensitivity of 0.73 with 0.13 mFPI	The model's high sensitivity and low FPs limit its usefulness in detecting lung cancer; however, a screening cohort may have a higher FP count.
Chui et al. (2023)	MTL-MGAN	Lung CT Segmentation Challenge 2017	Obtain 8.70 accuracy	Using the maximum number of rounds results in complexity and slower the model.

Furthermore, the study introduces an advanced pre-processing pipeline to improve data quality and feature representation. The Disperse Wiener filter is employed for noise removal, ensuring that the CT scan images are clean and ready for further analysis. This step effectively eliminates distortions that could negatively impact the feature extraction process. The integration of the CViT for feature extraction marks another significant contribution. CViT provides a robust mechanism for capturing global and local features from the pre-processed CT scan images, making it particularly effective in identifying subtle patterns associated with lung tumors.

To enhance the model's efficiency, the study incorporates the EBWO technique for feature selection. EBWO identifies the most relevant features from the extracted dataset, eliminating redundant or irrelevant information. This reduces the computational load and improves the model's classification accuracy by ensuring that only meaningful features contribute to the decision-making process. Additionally, the study employs AEO for hyperparameter tuning, which dynamically adjusts the learning parameters during training. This adaptive approach enhances the model's ability to handle varying data distributions, optimizing its performance across diverse datasets.

The combination of these innovative techniques allows the proposed MHA-DCNN framework to achieve exceptional results. The integration of multi-head attention, depthwise convolutions, and advanced optimization strategies makes this framework a robust and efficient solution for early lung cancer detection. By addressing real-world challenges such as noise, feature selection, and hyperparameter optimization, the proposed model sets a new benchmark for lung cancer classification using CT scan images. This study contributes to the advancement of deep learning methodologies and holds significant potential for practical applications in medical imaging. Fig. 1 is a graphical demonstration of the proposed concept.

At first, medical data was discovered by the dataset that contained the CT scan images, and then the data underwent pre-processing, which eliminated noise from the images. Then, the relevant features are extracted from a pre-processed image. More accurate features are chosen from the retrieved features. The proposed MHA-CDNN model is used to classify benign and malignant tumours based on the selected features. Adjusting the hyperparameter, a physics-based metaheuristic algorithm is utilized to enhance the learning capacity. A detailed description of the process is provided in the following sections.

3.1. Removing noises by using a Disperse wiener filter

In the first phase, lung images are taken from the datasets, which make some noises, leading to improper feature selection. These noises are reduced by applying the well-known wiener filter with the dispersion function as an adaption. The dispersion index is calculated as the ratio of the variance to the mean of pixel intensities in the pre-processed CT scan images. Unlike traditional variance, the dispersion index provides a more reliable estimation of noise levels across images with varying intensity distributions. This method is particularly effective in medical imaging, where the intensity of pixel values can vary significantly. The dispersion index is preferred over traditional variance for noise estimation because it is scale-invariant, ensuring accurate noise reduction even when the images have different brightness levels. It is also more sensitive to localized noise, which aids in improving the quality of the pre-processed images used for feature extraction.

Moreover, the dispersed wiener filter is considered $(m1, m2)$ as the location of a particular pixel for initializing the denoising process. This mean and variance are estimated by the upcoming equation.

$$\mu = \frac{1}{NM} \sum_{m1,m2 \in \aleph} ip(m1, m2) \tag{1}$$

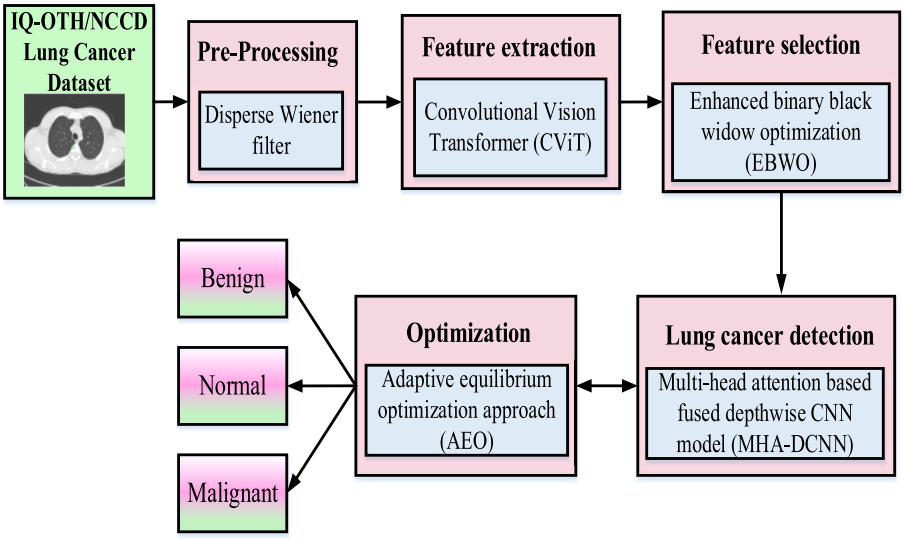


Fig. 1. Graphical workflow of the proposed model.

$$\sigma^2 = \frac{1}{NM}\sum_{m1,m2\in\aleph} ip(m1,m2) - \mu^2 \tag{2}$$

The variance of the noise in the input image is denoted by σ_m, the filter and is adapted by the dispersion function by using the dispersion index in equation (1) instead of variation. The dispersion ratio determines how the set of observed occurrences is scattered. Furthermore, the index is estimated using the variance-to-mean ratio given by the equation below.

$$DisIn = \frac{\sigma^2}{\mu} \tag{3}$$

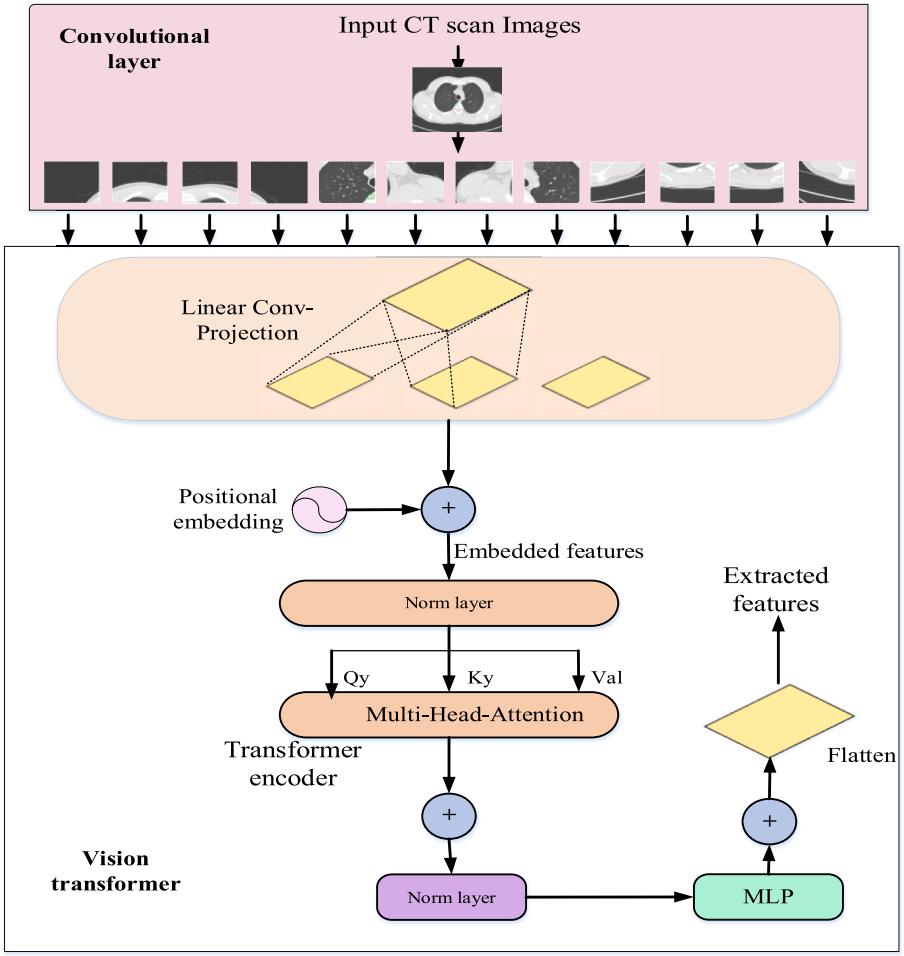


Fig. 2. Overall architectural representation of CViT.

$$DWF[ip(m1,m2)] = \mu + \frac{DisIn - \sigma_m^2}{DisIn}(ip(m1,m2) - \mu) \quad (4)$$

Here, the noise is almost the same as $\frac{\sigma_m^2}{\sigma^2}(ip(m1,m2) - \mu)$ with different signs, while one of the noises is called speckle, which is a multiplicative noise. The noise provides a minimum effect with less intensity of light. The multiplication of $\frac{\sigma_m^2}{\sigma^2}(ip(m1,m2) - \mu)$ with the mean may result in effectively attaining a denoising image. The main advantage of a dispersed wiener filter is that it results in the earlier information about the noise in the input images and then removes the noise to result in the same as the first image (Park et al., 2020).

3.2. Feature extraction by using convolutional vision transformer (CViT)

The pre-processed images are transmitted to the feature extraction step, which is carried out by the highly efficient vision transformer, which combines convolutional layers and attention modules. The resilience of the model is increased by transformer blocks with varied spatial sizes, which produce different impacts while extracting hybrid classification features. The pre-processed image is initially passed to the convolutional layer, which extracts several kernel characteristics. Convolutional layers are used in ViT to enhance CViT's performance and efficiency. In addition, the kernel-based texture characteristics retrieved from the convolutional layers are sent to ViT. The overall architectural representation of CViT is given in Fig. 2.

where *ht* and *wd* represents the height and width, respectively and the patch size is given as *p*. Moreover, the channels are represented by the variable *ch* where the size of the image is (78×78), and the patch size will be (6×6). The number of patches can be computed by using the image size and patch size as $NP = \frac{ht \times wd}{p^2} = \frac{78 \times 78}{(6)^2} = 169$. The 2D matrix is formed from the raw image *Ig* after patch partition *Pt*. This *Pt* is linearly deployed into the 1D embedding vector Pt_{L_P} by the dimension of 64.

Positional embedding reduces the transformer's computing expenses to some extent. In the embedding strategy, the patches are deployed via positional embedding by dividing the picture patches into smaller groups and then applying them to larger image sizes. The sine and cosine functions with different frequencies are used for the performance of position embedding P_{emb}.

$$P_{emb} = \begin{cases} \sin\left(\frac{pos}{1000^{\frac{2j}{maxl}}}\right), j \text{ is even} \\ \sin\left(\frac{pos}{1000^{\frac{2j}{maxl}}}\right), j \text{ is odd} \end{cases} \quad (5)$$

$$Emb_{pat} = Con(Pt_{L\_P}, P_{emb}) \quad (6)$$

Initially, the encoder's input block like Emb_{pat} is concatenated with the MSA output. The output is subsequently sent to a normalization layer, where the dense dropout layer of the MLP is present. Finally, this attention output is combined with a skip connection from the input, increasing the influence of location as the original embedded patch is transferred to the next layer. There are three embedding matrices used in the computation of attention, namely key *ky*, query *qy* and value *val* given in the following equation, which were calculated by using weight matrices Wt_{ky},Wt_{qy} and Wt_{val} respectively.

$$qy = Emb_{pat}.Wt_{qy} \quad (7)$$

$$ky = Emb_{pat}.Wt_{ky} \quad (8)$$

$$val = Emb_{pat}.Wt_{val} \quad (9)$$

here, the weight matrix Wt_{ky}, Wt_{qy}, $Wt_{val} \in \Re^{maxl\ mod\ el \times d_l}$, the single attention function, is executed multiple times in the multi-head attention (MHA) layer where d_l is the sot scaled product. This scalar product is used to prevent the attention value from exploding. The correlation between two image patches is represented by the scoring function made by the attention value. The proposed extraction model has four heads, the MHA of which will be given by the equation below.

$$Atten(qy, ky, val) = soft\max\left(\frac{qyky^T}{\sqrt{d_l}}\right)val \quad (10)$$

$$MHA = Atten,(qy, ky, val)_{\times 4} \quad (11)$$

After paying attention to providing non-linearity in this process, MLP contains a dense layer with Gaussian error linear unit (GELU) activation. The cumulative distribution of Gaussian distribution is denoted by φ then the output features being excited as $Trans_{fea\_shap}$ and flattened by the below equation.

$$GELU(y) = yp(Y \leqslant y) = y\varphi(y) \quad (12)$$

$$Trans_{fea\_shap} = GELU(MHA) \quad (13)$$

$$X_{Trans} = flatten(Trans_{fea\_shap}) \quad (14)$$

The X_{Trans} features are selected for the further phase of classification, this automated process of extraction by the CViT helps to enhance the performance of the classification model (Wu et al., 2021).

3.3. Feature selection using enhanced binary black widow optimization (EBWO)

The approach uses a classifier as a wrapper tool to assess potential solutions. K-nearest neighbours (KNN) were employed in this study to achieve this goal. Binary algorithms are particularly effective in feature selection tasks within a discrete domain, where the goal is to select a subset of features from a larger set. In such tasks, each feature can be included or excluded from the model, which can be naturally represented by binary values (1 for inclusion and 0 for exclusion). Binary algorithms improve the feature selection process by efficiently navigating through the large and discrete search space of possible feature subsets. They reduce the computational complexity by focusing only on relevant features, avoiding overfitting, and ensuring that the model remains interpretable with smaller features.

To adapt binary optimization methods for feature selection, several modifications were made. Specifically, the BBWO algorithm was tailored to work within a binary search space. In BBWO, the feature subset selection is encoded as a binary vector, where each element represents the inclusion (1) or exclusion (0) of a corresponding feature. The algorithm's fitness evaluation is based on classification performance, where subsets of features that improve model accuracy are favored. Additionally, the binary search space allows the algorithm to explore and exploit different combinations of features without the need for continuous value representation. The adjustments guarantee that the BBWO algorithm can effectively search for an optimal feature subset while preserving a balance between exploration and exploitation. These adaptations enable the algorithm to effectively handle the discrete nature of the feature selection problem and improve its overall performance.

In this study, several selection techniques are employed within the Binary Black Widow Optimization (BBWO) algorithm to select relevant features: BBWO Randomization, BBWO Tournament Selection, BBWO-Rank Selection, and BBWO-Roulette Wheel Selection. Each of these methods impacts the feature selection process differently. BBWO Randomization introduces greater diversity by selecting features randomly without considering their fitness, which promotes exploration but may result in slower convergence or suboptimal feature selection. BBWO Tournament Selection increases selection pressure by selecting individuals based on fitness from a randomly chosen subset, leading to quicker convergence. However, this can reduce diversity and cause premature convergence if the selection pressure is too high. BBWO-Rank

Selection improves upon this by selecting individuals based on rank, allowing for a balance between exploiting better solutions and maintaining diversity within the population. Finally, BBWO-Roulette Wheel Selection (fitness proportionate selection) assigns selection probabilities based on fitness, ensuring that higher fitness solutions are more likely to be selected. Still, all individuals have a chance to contribute, promoting exploration and exploitation. These selection methods help the algorithm adapt to different optimization scenarios, ensuring that the most relevant features are chosen while maintaining a balance between exploration and convergence.

First, the starting population is generated randomly for all wrapper-based approaches that produce candidate solutions validated by the KNN. Each member has a fitness value estimated using the equation below.

$$F_j = \alpha \times errte_j + \beta \times \frac{SF}{TF} \tag{15}$$

here, the fitness value of j^{th} feature is denoted by F_j and the error rate of j^{th} feature is referred as $errte_j$.β and α are said to be predefined where $\beta = 1 - \alpha$ and the total number of features and the number of selected features are represented by TF and SF respectively. Then, for the procreation step, two parents were picked using the specified selection mechanism based on the approach used, namely randomization for BBWO and rank selection for BBWO-R. The equation used for creating offspring is given below.

$$\left\{ \begin{array}{l} X_1^j = \sigma \times y_1^j + (1-\sigma) \times y_2^j \\ X_2^j = \sigma \times y_2^j + (1-\sigma) \times y_1^j \end{array} \right\} \tag{16}$$

The first and second offspring bit value is denoted by X_1^j and X_2^j respectively. The parents of the first and second are noted as y_1^j and y_2^j.

$$\left\{ \begin{array}{l} X_{1,2}^j = 1, X_{1,2}^j > 0.5 \\ X_{1,2}^j = 0, \quad else \end{array} \right\} \tag{17}$$

At this point, explaining how the approaches differ from the BWO is vital. Following Equation (16), the sigmoid transfer function was applied. This transformation aims to convert continuous space into discrete space. Rather than employing a transfer function, equation (16) generates random values in the interval [0, 1] for the σ. As a result, the output was scaled from 0 to 1. The threshold value, which was set at 0.5, was then compared to the output to determine whether it was 0 or 1. Another distinction is the method utilized for parent selection. In typical BWO (Al-Saedi & Mawlood-Yunis, 2022), random selection was used; in this study, four selection mechanisms, including randomization, were examined. Cannibalism is followed by father-child cannibalism, which is used for both the parent and the newborn children. The member with the lowest fitness value is considered to be male in cannibalism. The pseudocode of EBWO-based feature selection is depicted in Table 2.

Table 2
Algorithm for EBWO-based feature selection.

Input: Procreating rate, cannibalism rate, mutation rate and maximum number of iteration
Output: Subset of best features
Start
//Initialization
Initiating bit vectors randomly as population and compute fitness value
while terminating criteria are not reach **do**
Compute the number of reproduction *NR* on the basis of procreating rate
Choose the best *NR* solution in pop and save in *pop*
// Procreating and cannibalism
for $j = 1$ to *NR* **do**
Choose two parents from *pop* to procreating
Generate children by the equation (16).
Terminate father
Terminate a few child according to the cannibalism rate (Newly updated solutions)
Save the remaining solution in *pop*2
//Mutation
Compute the count of mutation children *CM* according to mutation rate
for $j = 1$ to *CM* **do**
From pop1 choose a solution as random
Create a new solution by exchanging two bits of the selected solution
Save the new one in *pop*2
//Updating
Unify the population as $pop = pop2 + pop3$
Return best M_{pop}
Return the best solution from pop
Stop

Two bits were traded for each member of the mutation in order to carry out the mutation step. All populations were combined after mutation, and the final candidates were preserved as a new population. Where M_{pop} was the population number, the best member of M_{pop} will be returned. The suggested methods relied on the KNN classifier, with k equal to 5. The widespread usage was used to determine the classifier and k-value. The two main goals of feature selection are to lower the number of selected features and improve classification accuracy. As advised by numerous researchers (Taradeh et al., 2019), the fitness function β and α values were set at 0.99 and 0.01, respectively. In the feature selection phase, the EBWO technique is used to identify the most relevant features from the pre-processed images. The relevance of a feature is determined based on its contribution to the overall classification performance. Features that significantly improve the accuracy of distinguishing between benign, malignant, and normal cases are prioritized. Additionally, EBWO eliminates redundant features, reducing dimensionality and improving computational efficiency. This technique ensures that only the most informative and distinct features are retained, which enhances the model's ability to make accurate predictions without overfitting.

3.4. Lung cancer classification using multi-head attention based fused depthwise CNN model (MHA-DCNN)

The features selected using the EBWO approach are then forwarded to the proposed MHA-DCNN model for classification. This model will be used to classify lung cancers. Fig. 3 shows the graphical structure of the proposed MHA-DCNN.

The core contribution of this study lies in the development of the MHA-DCNN, an innovative deep learning model designed to address the challenges of lung cancer classification. By integrating multi-head attention mechanisms, the model effectively captures complex spatial relationships within CT scan images, enabling it to focus on the most critical features for distinguishing between benign, malignant, and normal cases. The depthwise convolutions further enhance this capability by reducing the computational overhead while maintaining the quality of feature extraction, making the model efficient and scalable for practical use.

To further improve the model's performance, AEO is employed for hyperparameter tuning. AEO dynamically adjusts parameters such as learning rate, batch size, and network depth, ensuring optimal learning during training. This optimization process enhances the model's ability to generalize across diverse datasets and adapt to varying data distributions. These innovations enable the MHA-DCNN to deliver robust and reliable results, making it a valuable tool for early and accurate lung cancer detection. The integration of attention mechanisms, efficient convolutional layers, and advanced optimization strategies sets this model apart from existing approaches, addressing critical limitations such as overfitting and computational inefficiency.

3.4.1. Convolutional layers of CNN

The MHA-DCNN model is a fusion model of CNN (Faruqui et al., 2021), Depthwise convolution (Siddhartha and Santra, 2020) and Multi-head attention. The Convolutional blocks receive the features first, and the module comprises attention and convolution extraction blocks. The

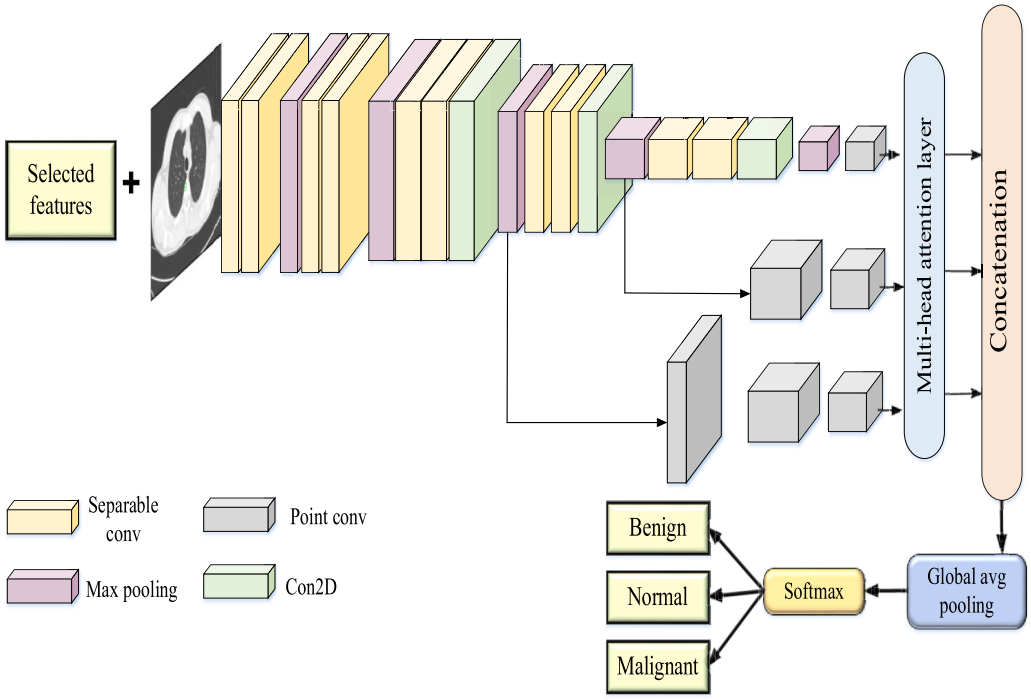


Fig. 3. Graphical structure of the proposed MHA-DCNN.

first two convolutional blocks contain a maximum pooling layer, two depthwise discrete convolution layers, and other components. The maximum pooling layer and two depthwise discrete convolution layers, together with the addition of typical convolutional layers, are combined with the subsequent layers. Batch normalization is used in the depthwise separable convolutional layer before providing the classification results. The Depthwise model is concatenated with the MHA layer to improve the consciousness of classification classes. This model aims to achieve improved classification accuracy using a compact model design. Instead of utilizing a fully linked layer, the model employs global average pooling to guarantee the accuracy of the input data. By minimizing the loss of shallow feature information, avoiding overfitting, and lowering parameter values, the proposed method improves global average pooling.

At the time of convolution sampling, image channels and convolutional kernel should added for receiving features of images and also for avoiding loss in information. After convolution, the image channels are enlarged to 3 to 64 as a limit of 512. Normally, the maximum pooling layer's step size is set to 2, resulting in an image size of one-half after each maximum pooling. The image scale is gradually reduced; image information is continuously compressed, and three channels of feature maps are added. The categorical_crossentropy loss function calculates the difference between the probability distributions acquired during training and the real ones. It shows the difference between the actual and expected result probability. The smallest value of cross entropy is stated to be closer between two probability distributions. The equation for estimating the categorical_crossentropy loss function is as follows.

$$CE = -\frac{1}{m}\sum_{y}[x\ln op + (1-x)\ln(1-op)] \tag{18}$$

here, op is said to be the output value and x is an actual value where the loss function is CE for the sample y and the total number of samples are represent as m. The gradient of bias bs and weight wt is given by the equation below.

$$\frac{\partial CE}{\partial wt_k} = \frac{1}{m}\sum_{y} y_k(\sigma(z) - x) \tag{19}$$

$$\frac{\partial CE}{\partial bs_k} = \frac{1}{m}\sum_{y} (\sigma(z) - x) \tag{20}$$

The weight wt and bias bs are modified faster as bigger the error and gradient.

3.4.2. Depthwise separable convolution

A unique convolution technique that works with depth and space is called depthwise separable convolution. The fundamental idea is to divide an entire convolution process into two stages: point-wise convolution and depthwise convolution. Each input channel is given a convolution kernel by the depthwise convolution for filtering. The outputs of the depthwise convolution are combined using a 1×1 convolution, which is applied via the point-wise convolution. This decomposition leads to a large decrease in computation and model size.

Depthwise convolution: Depthwise convolution generates matching feature maps by applying a single convolution kernel to each input channel. Assuming that the typical convolution kernel L is $(Dp_l Dp_l, N, M)$ the input feature map G is (Dp_G, Dp_G, N), the output feature map O is (Dp_O, Dp_O, M). Here N and M represent the number of input and output channels independently. The feature map's size is denoted by Dp and L is divided into two convolutions such as point-wise convolution $(1, 1, N, M)$ and a depthwise convolution $(Dp_l Dp_l, 1, N)$.

Point-wise convolution: Point-wise convolution projects the channel found by depthwise convolution onto the recently formed channel space using the conventional 1×1 convolution kernel. After depthwise convolution, the $NDp_o \times Dp_o$ feature maps are convolved using N convolution kernels of size $1 \times 1 \times N$ in the point-wise convolution. Then, carry out a weighted combination in the depth direction to produce $O(Dp_O, Dp_O, N))$ feature maps with $MDp_o \times Dp_o \times 1$. The number of convolution kernels determines the number of feature maps.

Analysis of efficiency: The convolution creates a new representation by filtering and combining the information using a convolution

kernel. By dividing filtering and combining into two processes, the depthwise separable convolution eliminates the number of channels and the interaction between the output channels and the kernels. The computational cost is significantly decreased by the kernel size 24. The conventional convolution kernel $L(Dp_l Dp_l, N, M)$ for the input image $G(Dp_G, Dp_G, N)$. The standard convolution equation is given as follows.

$$O_{l,J,m} = \sum_{j,k,n} L_{j,k,n,m} \times G_{l+j-1,l2+k-1,n} \tag{21}$$

The following formula provides an approximation of the standard convolution computation cost.

$$Dp_l \times Dp_l \times N \times M \times Dp_G \times Dp_G \tag{22}$$

The equation for depthwise separable convolution is given as follows,

$$\widehat{O}_{l,l2,n} = \sum_{j,k,n} \widehat{L}_{j,k,n} \times G_{l+j-1,l2+k-1,n} \tag{23}$$

The computation cost of depthwise separable convolution is estimated using the following equation.

$$Dp_l \times Dp_l \times N \times Dp_G \times Dp_G + N \times M \times Dp_G \times Dp_G \tag{24}$$

In the following Equation, the ratio of calculation consumption (RCC) is defined to demonstrate the great computing efficiency of the proposed approach.

$$RCC = \frac{Dp_l \times Dp_l \times N \times Dp_G \times Dp_G + N \times M \times Dp_G \times Dp_G}{Dp_l \times Dp_l \times N \times M \times Dp_G \times Dp_G} = \frac{1}{M} + \frac{1}{Dp_L^2} \tag{25}$$

From equation (27), RCC is said to be smaller than one and nearness to one of M. This shows that maximizing map size will decrease the cost and enhance efficiency. During depthwise separable convolution, the model can learn information about features separated by spaces and channels, improving efficiency. The depthwise separable convolution makes faster classification, and the model is reduced by minimizing the number of computing parameters. This parameter are tuned by the optimization approach described in the further sections.

3.4.3. MHA module

MHA follows the scaled dot-product attention calculation approach, which learns different mappers in the models. Initially, in every dp_{mdl} of query, key and value, the linear transformation will take place and also compute scaled dot product attention in parallel for creating dimensional output do. The concatenation function $concat$ is used to integrate multiple outputs from that output linear transformation to choose a final result. Rather than adding more complexity, MHA enables the model to examine the correlation information in the various representation subspaces. It also enhances the ability to perceive, as shown by the equation below.

$$MH(qy, ky, val) = concat(head_1, \ldots, head_i) wt^p \tag{26}$$

$$head_j = Atten\left(qywt_j^{qy}, kywt_j^{ky:}, valwt_j^{val}\right), wt_j^{qy} \in S^{dp_{mdl} \times dp_l}, wt_j^{ky} \in S^{dp_{mdl} \times dp_l}, wt_j^{val} \in S^{dp_{mdl} \times dp_l}, wt^p \in S^{dp_{mdl} \times gdp_{wl}} \tag{27}$$

The total computing cost lowers as the dimensionality of each attention layer decreases, which also applies to the full-dimensional single-head attention layer. The MHA mechanism is simpler than the scaled dot-product attention mechanism. It also enables the model to learn different representation information, preventing the loss of small target information owing to average value.

3.4.4. Parameter tuning by adaptive equilibrium optimization approach (AEO)

In the proposed approach, AEO optimizes the model's learning capacity through effective hyperparameter tuning. By dynamically adjusting parameters such as learning rate, batch size, and network architecture, AEO enhances the model's efficiency and ensures optimal performance. This optimization significantly contributes to the model's achieving high classification accuracy in the evaluation phase. Specifically, the model correctly classifies normal, malignant tumor cases and benign tumor cases, demonstrating its robustness and reliability across all classes. Furthermore, the AEO is an enhanced approach to equilibrium optimization (EO), which operates based on random dispersion in nonperformer particles. In the search space, dispersal takes place according to the fitness value; such a mechanism has more attraction in optimization. The control volume mass balance model's equilibrium and dynamic states served as the model for EO. Its foundation is the idea of mass conservation in the arriving, leaving, and producing processes within a finite volume. This AEO approach can feasibly use in the hyperparameter tuning to enhance the classification model performance.

3.4.4.1. Mathematical model for AEO. In the search space, consider search agents related to the concentration and initiate the first iteration $it = 1$. This process is done by using the below equation,

$$Cn_j(it = 1) = lb + rd_j(1, \dim) * (up - lp), j = 1, 2, \ldots m \tag{28}$$

The search space is bounded by the lower and upper bounds lb and ub, the number of search agents is m, the problem's dimension is dim.A vector with one dimension rd_j, here j is made up of random values within the interval $[0, 1]$. For a control volume Vol in EO, the location of the j^{th} search agent is updated as

$$\overrightarrow{Cn}_j(new) = \overrightarrow{Cn}_{eql}(it) + \left(\overrightarrow{Cn}_j(it) - \overrightarrow{Cn}_{eql}(it)\right) * \overrightarrow{Et}_j(it) + \frac{\overrightarrow{Gn}_j(it)}{\overrightarrow{\lambda}_j(it) * Vol} \times \left(1 - \overrightarrow{Et}_j(it)\right) \tag{29}$$

From the equilibrium pool $\overrightarrow{Cn}_{eql,pool}$, equilibrium candidates are randomly selected and denoted by $\overrightarrow{Cn}_{eql}$ that contain the best search agents $\overrightarrow{Cn}_{eql(1)}$ $\overrightarrow{Cn}_{eql(2)}$ $\overrightarrow{Cn}_{eql(3)}$ $\overrightarrow{Cn}_{eql(4)}$. The $\overrightarrow{Cn}_{eql(avg)}$ is said to be the average of best search agents. Here, the fitness value of all the best agents is represented as $ft\left(\overrightarrow{Cn}_{eql(1)}\right) ft\left(\overrightarrow{Cn}_{eql(2)}\right) ft\left(\overrightarrow{Cn}_{eql(3)}\right) ft\left(\overrightarrow{Cn}_{eql(4)}\right)$. A sorted list is used to determine the equilibrium candidates and their fitness values for a minimization problem. The following equation denotes the fitness values of every m search agent:

$$ft = (ft_1, ft_2, \ldots, ft_m) \tag{30}$$

The fitness values are then sorted in ascending order using the following equation.

$$[sorted\_ft, sort\_idx] = sort(ft) \tag{31}$$

The following definition of the equilibrium candidates and their fitness is based on the equation above.

$$
\begin{aligned}
ft\left(\overrightarrow{Cn}_{eql(1)}\right) &= sorted\_ft(1)\ and\ \overrightarrow{Cn}_{eql(1)} = \overrightarrow{Cn}(sort\_idx(1))\\
ft\left(\overrightarrow{Cn}_{eql(2)}\right) &= sorted\_ft(2)\ and\ \overrightarrow{Cn}_{eql(2)} = \overrightarrow{Cn}(sort\_idx(2))\\
ft\left(\overrightarrow{Cn}_{eql(3)}\right) &= sorted\_ft(3)\ and\ \overrightarrow{Cn}_{eql(3)} = \overrightarrow{Cn}(sort\_idx(3))\\
ft\left(\overrightarrow{Cn}_{eql(4)}\right) &= sorted\_ft(4)\ and\ \overrightarrow{Cn}_{eql(4)} = \overrightarrow{Cn}(sort\_idx(4))\\
\overrightarrow{Cn}_{eql(avg)} &= \frac{1}{4}\left(\overrightarrow{Cn}_{eql(1)} + \overrightarrow{Cn}_{eql(2)} + \overrightarrow{Cn}_{eql(3)}, \overrightarrow{Cn}_{eql(4)}\right)
\end{aligned}
\tag{32}
$$

The final equilibrium pool is defined from all the fitness values given in the equation below.

$$
\overrightarrow{Cn}_{eql\_pool} = \left\{\overrightarrow{Cn}_{eq(1)}, \overrightarrow{Cn}_{eq(2)}, \overrightarrow{Cn}_{eq(3)}, \overrightarrow{Cn}_{eq(4)}, \overrightarrow{Cn}_{eq(avg)}\right\} \tag{33}
$$

here $\overrightarrow{Et}_j$ is an exponential term that assists the EO in exploitation, and then exploration is computed for j^{th} search agent using the equation below.

$$
\overrightarrow{Et}_j(it) = c_1 sign(s_1 - 0.5)\left[e^{-\overrightarrow{\lambda}_j\left(1-\frac{it}{\max\_it}\right)^{\left(c_2\frac{it}{msx\_it}\right)}}\right] \tag{34}
$$

here, c_1 and c_2 is a parameter used for exploration and exploitation control, respectively. Depending on a random number s_1 between $[0,1]$, *sign* controls the search's direction, $\overrightarrow{\lambda}_j(it)$ is a random vector of dim. The EO will proceed through a maximum of iterations max_*it* in the position update while the current iteration is denoted by *cit* in the interval $[0,1]$ for the j^{th} search agent in *cit*. In the exploration, the generation rate $\overrightarrow{Gn}_j$ helps the phase with participation probability of $\overrightarrow{Cn}_{eql}$. The $\overrightarrow{Gn}_j$ is defined as follows

$$
\overrightarrow{Gn}_j(it) = \overrightarrow{Gn}_{j,0}(it) * \overrightarrow{Et}_j(it) \tag{35}
$$

Then, $\overrightarrow{Gn}_{j,0}(it)$ and $\overrightarrow{GCF}_j(it)$ are estimated using further equations.

$$
\overrightarrow{Gn}_{j,0}(it) = \overrightarrow{GCF}_j(it)\left(\overrightarrow{Cn}_{eql}(it) - \overrightarrow{\lambda}_j(it)\right) \tag{36}
$$

$$
\overrightarrow{GCF}_j(it) = \begin{cases} 0.5s_1 & s_2 \geqslant GP \\ 0 & s_2 \geqslant GP \end{cases} \tag{37}
$$

here, the generation rate control factor is denoted by *GCF* and *GP* refers to the generation probability where s_1, s_2 are the random numbers from $[0, 1]$. The search agents' places in the following iteration are always determined by the equilibrium pool, which comprises the four best positions and an average of the best post positions. During an exploration, this search progressive iteration narrows the boundaries of the search space, forcing the search agent to fall into a local minimum. Search agents select a random distribution for a nonperformer in AEO in order to resolve this problem. The average fitness of all m search agents and the search agents's current fitness are used to help make an adaptive judgment. This is defined as the minimization problem's mathematical model by the following equation.

$$
\overrightarrow{Cn}_j(it+1) = \begin{cases} \overrightarrow{Cn}_j(new) & ft_j(it) < ft_{avg}(it) \\ \overrightarrow{Cn}_j(new) \otimes (0.5 + rd(1,\dim)) & ft_j(it) \geqslant ft_{avg}(it) \end{cases} \tag{38}
$$

here, $\otimes$ denotes the element-wise multiplication function while $ft_j(it)$ is a j^{th} search agent's fitness value at *it*. The average fitness of all search agents at an iteration *it* is denoted by $ft_{avg}(it)$ that estimated by the following equation.

$$
ft_{avg}(it) = \frac{1}{m}\sum_{j=1}^{m} ft_j(it) \tag{39}
$$

The EO's memory-saving procedures compare and update the existing fitness levels. Prior iterations in which one gets a greater fitness value are also passed down to the AEO. This is denoted by the given equation,

$$
\overrightarrow{Cn}_j(it) = \begin{cases} \overrightarrow{Cn}_j(it) & it > 1\ and\ ft_j(it) < ft_j(it-1) \\ \overrightarrow{Cn}_j(it-1) & it > 1\ and\ ft_j(it) \geqslant ft_j(it-1) \\ \overrightarrow{Cn}_j(it) & it = 1 \end{cases} \tag{40}
$$

And

$$
\overrightarrow{ft}_j(it) = \begin{cases} \overrightarrow{ft}_j(it) & it > 1\ and\ ft_j(it) < ft_j(it-1) \\ \overrightarrow{ft}_j(it-1) & it > 1\ and\ ft_j(it) \geqslant ft_j(it-1) \\ \overrightarrow{ft}_j(it) & it = 1 \end{cases} \tag{41}
$$

The proposed model's hyperparameters will be modified in each process using fitness and conditional functions. The pseudocode of the AEO approach is given in Table 3. In the beginning, the search agent count m need to be assigned the number of maximum iterations max_*it*, searching dimension dim and parameters are c_1, c_2, GP, Vol.

By selecting the optimum solution and fitness value, the hyperparameter can be properly modified to improve classification model performance. Furthermore, the optimization becomes more reliable by resolving concerns with standard EO. The AEO algorithm is employed to optimize several key hyperparameters of the model, enhancing its learning capacity. These hyperparameters include the learning rate, batch size, network depth, dropout rate, and filter size in convolutional layers. AEO dynamically adjusts these parameters during training, ensuring that the model converges efficiently and avoids overfitting. By optimizing these parameters, AEO improves the model's ability to generalize to unseen data and enhances its overall performance in classifying lung cancer cases. The performance of the proposed approach is examined, and the findings are presented in the following sections.

4. Result and discussions

The proposed model is examined using various metrics to show off its performance, and the IQ-OTH/NCCD dataset is used for the evaluation process. A detailed description of the dataset is given in the further section.

4.1. Dataset description

The MHA-DCNN model is evaluated using the Iraq-Oncology Teaching Facility/National Center for Cancer Diseases (IQ-OTH/NCCD) dataset, which was acquired from the stated facility. The collection includes CT scans of patients diagnosed with various stages of lung cancer, as well as normal lung scans. In total, 1190 photos provided CT scan slices from around 110 instances. These CT scan images are marked by oncologists and radiologists at these two facilities. The proposed MHA-DCNN method is applied to classify benign tumors, malignant tumors and normal cases, where 'normal' refers to lung tissues without any tumors or abnormalities. This distinction ensures that the model can effectively differentiate between healthy and affected lung tissues while categorizing tumors into benign and malignant categories. Table 4 shows the number of instances in each class of IQ-OTH/NCCD.

Furthermore, these situations are grouped into normal, benign, and malignant. CT scans are available in DICOM (Digital Imaging and Communications in Medicine) format. The 80 to 200 slices range refers

Table 3
Algorithm for AEO based parameter tuning.

Start
//Initialization
Create a random position vector Cn_j for j^{th} search agent by equation (28) for m agents at $it = 1$
for $it = 1 : \max\_it$
Compute fitness value ft for present iteration.
Estimate equilibrium candidates $\overrightarrow{Cn}_{eql(1)}$ $\overrightarrow{Cn}_{eql(2)}$ $\overrightarrow{Cn}_{eql(3)}$ $\overrightarrow{Cn}_{eql(4)}$ and $\overrightarrow{Cn}_{eql(avg)}$ by equation (32)
Build the equilibrium pool $\overrightarrow{Cn}_{eql,pool}$ by equation (33)
Attaining minimum memory utilization by Equation (40) and Equation (46)
for $it = 1 : m$
Choose $\overrightarrow{Cn}_{eql}$ randomly from the equilibrium pool $\overrightarrow{Cn}_{eql,pool}$
Create an exponential term $\overrightarrow{Et}_j$ by equation (34)
Create a generation rate $\overrightarrow{Gn}_j$ by equation (35)
Estimate the average fitness ft_{avg} by equation (39)
Compute the search agent position $\overrightarrow{Cn}_j(new)$ by equation (29)
Update search agent position $\overrightarrow{Cn}_j$ in further iteration by equation (38)
end for
end for
Return the best solution as $\overrightarrow{Cn}_{eql(1)}$ and the best fitness as $ft\left(\overrightarrow{Cn}_{eql(1)}\right)$
Stop

to human chest pictures taken from various angles and sides. Using this dataset, the proposed model was evaluated and validated, and good scores were attained in the testing phase.

4.2. Performance analysis

A variety of criteria, including accuracy, precision, f score, sensitivity, specificity, error rate, receiver operating characteristic (ROC) curve, and computing time, are used to assess the proposed model. The following equations are used to obtain these metrics, which demonstrate the uniqueness of the proposed model with respect to particular terms.

4.2.1. Accuracy

The accuracy refers to the exactness and excellence of the model in percentage, represented by the ratio of correctly classified classes to the entire classes. Accuracy is computed using true and false rates using the following equation: $T.ng-$ true negatives, $F.Ng-$ false negatives, $T.ps-$ true positives and $F.\ ps-$ false positives.

$$Accuracy = \frac{T.\ ng + T.ps}{T.\ ng + F.\ ps + F.\ ps + T.ps} \tag{42}$$

4.2.2. Specificity

The capability of the proposed model to predict true negative results is estimated based on the specificity metric. The equation of specificity is given below,

$$Specificity = \frac{T.ps}{T.ng + F.ps} \tag{43}$$

4.2.3. Sensitivity

The capability of the model to predict true positive results is estimated based on sensitivity. This metric also provides the proposed models recall score, and the specificity equation is provided as follows:

Table 4
Overview of classes in the dataset.

Total images	1190
Total Slices	110
Classes	3
Normal cases	55
Benign cases	15
Malignant	40

$$Specificity = \frac{T.ps}{T.ps + F.\ ng} \tag{44}$$

4.2.4. Precision

The proposed model's classification strength is referred to as precision. The following formula is used to compute the precision.

$$precision = \frac{T.ps}{T.ps\ + F.ps} \tag{45}$$

4.2.5. F-score

The F1 score characterizes the analysis by determining the validity of the classification. The F1 score will be computed by the given equation.

$$F-score\frac{T.ps}{T.ps + \frac{1}{2}(F.ps + F.ng)} \tag{46}$$

4.2.6. Error rate

The error rate defines the degree of classification error by comparing a proposed model to the true model. The equation for error rate analysis is given as follows,

$$ER = 100 - Accuracy \tag{47}$$

4.3. Evaluation of performance

The performance is compared to other existing models using the evaluated measures. First, the proposed model's accuracy is compared to several existing models used to classify lung cancer, as shown in Fig. 4.

From the analysis, the accuracy attained by the proposed model is 99.43 % in classifying the normal class, 99.51 % in the malignant class and 99.66 % in the benign class. Even the existing models Deepa, V. (2023) such as improved deep neural network (IDNN), denoising first two-path convolutional neural network (DFDN), adaptive hierarchical heuristic mathematical model (AHHMM), Naïve Bayes-decision tree-stochastic diffusion search (NB-DT-SDS), Deep-ShrimpNet (D-ShrimpNet) and CNN-Fuzzy Particle Swarm Optimization (CFPSOA) possess less accuracy than proposed model. Moreover, the D-ShrimpNet model possesses 99.16 %, 99.27 %, and 99.35 % accuracy in classifying normal, malignant, and benign cases. Other models attain less than 90 % due to their inefficient feature extractions. Here, the CViT model used in extracting features uniquely presents every related feature with attention. From these features, the most relevant feature is only detected by

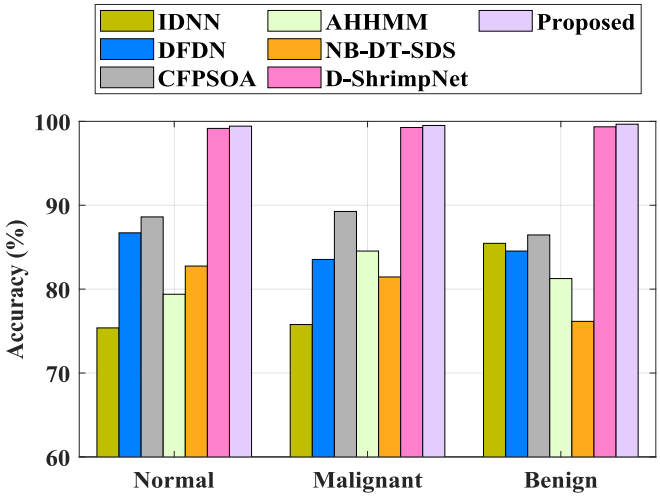


Fig. 4. Accuracy analysis of the proposed model among existing models.

the EBWO, which helps in accurate classification by the proposed model with high feasibility. As a continuation of the accuracy evaluation, the F-score was also examined and compared with the existing models, as shown in Fig. 5.

From the graph, the F-score attained by the proposed model is 99.33 % in classifying the normal class, 99.61 % in the malignant class and 99.56 % in the benign class. Even the existing models (Deepa & Fathimal) possess less F-score than the proposed model. Moreover, the D-ShrimpNet model possesses 99.03 %, 99.36 %, and 99.24 % of F-scores in normal, malignant, and benign class classification, respectively. Other models attain less than 90 % due to their ineffectiveness of classification. Here, from extracted features, only the most relevant feature will excited by the EBWO, which helps in effective classification by the proposed model with higher capability. Further evaluation is made on estimating precision score and compared with the existing models in Fig. 6.

In precision score analysis, the proposed model attained 99.73 % in classifying the normal class, 99.91 % in the malignant class and 99.46 % in the benign class. Even the existing models possess less precision than the proposed model. Moreover, the D-ShrimpNet model possesses 99.44 %, 99.57 % and 99.16 % precision in normal, malignant and benign class classification, respectively. Other models attain less than 85 % due to their overfitting challenges. To address the issue of overfitting, the proposed model employs global average pooling in the MHA layer of the DCNN module. Thus, the proposed model achieves a higher precision score. Here, the sensitivity score of the proposed model is compared with the existing models and depicted in Fig. 7.

The comparative graph shows that the proposed model attained a 99.53 % sensitivity score in classifying the normal class, 99.81 % in the malignant class and 99.76 % in the benign class. Even the existing models possess less sensitivity than the proposed model. Moreover, the D-ShrimpNet model possesses 99.23 %, 99.47 % and 99.34 % sensitivity in normal, malignant and benign classes classification, whereas other models attain less than 85 %. The main strength of the proposed model is its capacity to achieve accurate results through better model consolidation with various attentions. As a result, the proposed model outperformed the other models regarding sensitivity. The specificity score of the proposed model is also compared with the existing models and depicted in Fig. 8.

The proposed model attains 99.97 % specificity in classifying the normal class, 99.89 % in the malignant class and 99.76 % in the benign class. Even the existing models possess less specificity than the proposed model. Moreover, the D-ShrimpNet model possesses 99.87 %, 99.67 % and 99.45 % precision in normal, malignant and benign class classification, respectively. Other models attain less than 80 % due to their ineffectiveness in individual class classification. In the proposed model, the MHA layer is included for parallel attention, and the EBWO approach is used to attain specific features in every class. So, the proposed model attains a higher score in specificity. Some errors may arise during the process of classification; such error rate of the proposed model is also evaluated and compared in Fig. 9.

From the error rate analysis, the proposed model attained a lower error rate of 0.50 in classifying the normal class, 0.41 % in the malignant class, and 0.30 % in the benign class. Even the existing models possess a higher error rate than the proposed model. Moreover, the D-ShrimpNet model has 0.84, 0.73, and 0.65 error rates in normal, malignant, and benign class classifications. Other models attain more than 24.50 error rates due to their complex design and ineffective concentration in classification. In the proposed model, the MHA layer is used in CViT and DCNN modules to focus on every feature. So that the proposed model attains a lower error rate than other existing models, each model has a different computation time due to its complexity. Hence, the computation time of the proposed model is compared with the existing models, as demonstrated in Fig. 10.

Normal deep learning models possess more time to complete the computation process, which is considered computation time. The computation time of the proposed model is less than that of other

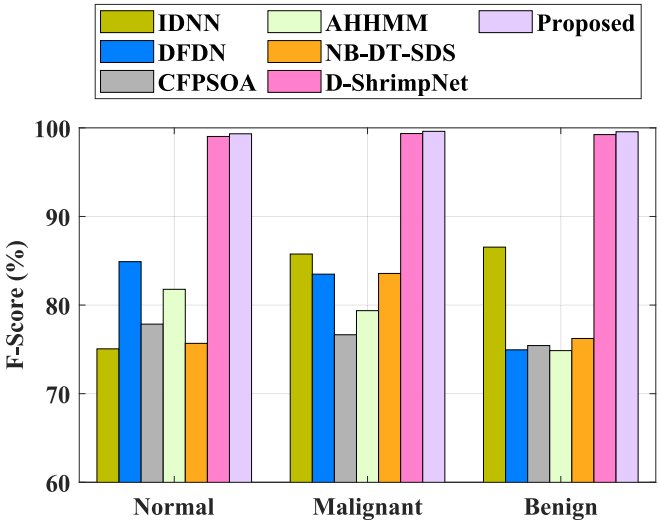


Fig. 5. F-score analysis of the proposed model among existing models.

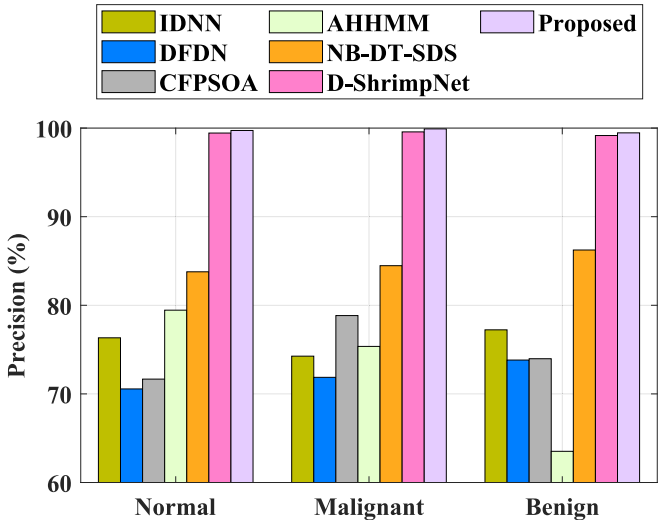


Fig. 6. Comparative analysis of precision score in proposed model among existing models.

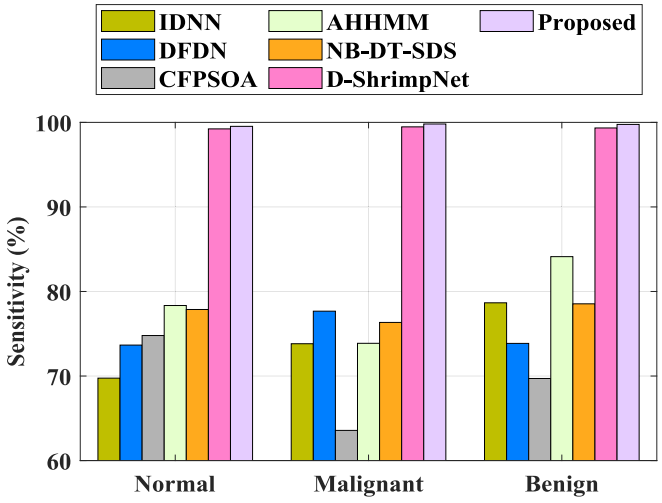


Fig. 7. Comparative analysis of sensitivity score in proposed model vs. existing models.

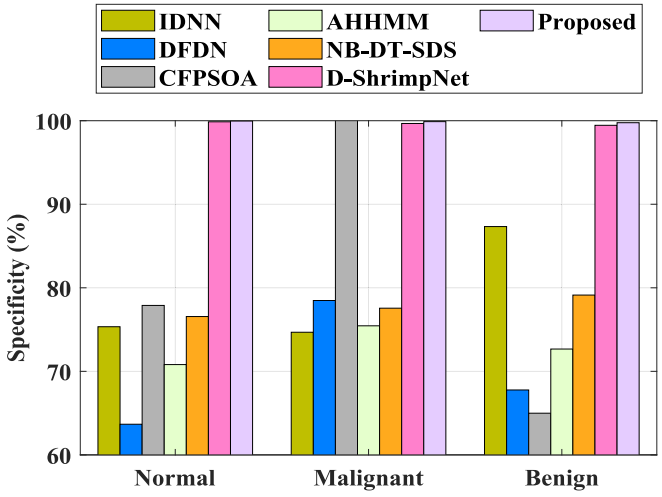


Fig. 8. Specificity score evaluation in proposed model vs. existing models.

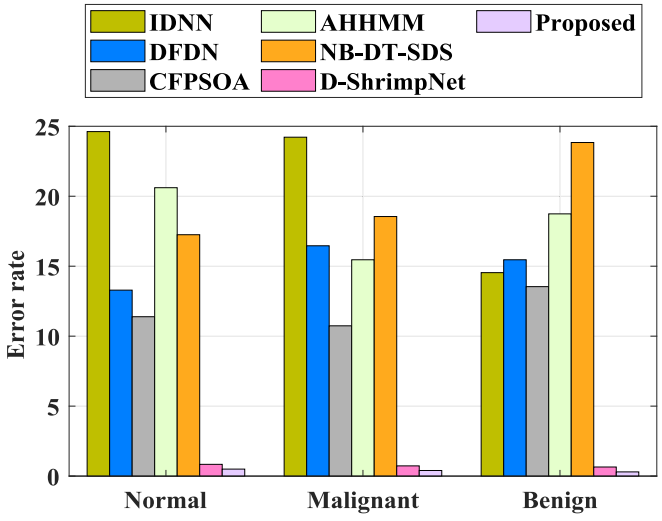


Fig. 9. Comparative analysis of Error rate at different classes by proposed model vs. Existing model.

existing models. The proposed model possess 50 s for entire computation completion while other models possess 209 s by IDNN, 177 s by DFDN, 261 s by CFPSOA, 200 s by AHHMM, 189 s by NB-DT-SDS, 82S by D-ShrimpNet model. Due to parallel processing and effective optimization, the proposed model achieves less complexity and computation time than existing works. The evaluation is extended by ROC analysis for the proposed model and other existing models, plotted in Fig. 11.

The line plot of ROC analysis shows that the proposed model has a larger area under the curve than other current models. The area attained by the proposed model is 0.992 area due to its highly feasible network model. Even other models achieve lower scores, such as 0.931 area in IDNN, 0.97 area in DFDN, 0.9652 area in CFPSOA, 0.958 area in AHHMM, 0.9524 area in NB-DT-SDS, 0.987 area in D-ShrimpNet model.

4.4. Discussion

The proposed MHA-CDNN model provides great results in evaluation by its enhanced fused network models according to the terms of different metrics. The major advancement in the proposed deep learning model is its feasibility in design and the provision of low cost for implementation. Using the IQ-OTH/NCCD dataset, this model divided lung cancer into three categories: normal, malignant, and benign. This dataset is pre-processed using a scattered wiener filter to get the most accurate characteristics for identifying cancer. In the feature extraction phase, each feature is uniquely extracted by the CViT in a parallel process to get faster results. From the extracted features, the most relevant features for defining classes of lung cancer are selected using the EBWO algorithm with less computational complexity. Finally, the proposed MHA-CDNN model with the AEO algorithm is used for highly accurate classification by tuned hyperparameters. With this advancement, the MHA-CDNN model provides good outcomes in the evaluation phase.

The model achieves an accuracy of 99.43 % in classifying the normal class, 99.51 % in the malignant class and 99.66 % in the benign class. Here, the CViT model used in extracting features uniquely presents every related feature with attention. From these features, the most relevant feature is only detected by the EBWO, which helps in accurate classification by the proposed model with high feasibility. The model attained a 99.33 % F-score in classifying the normal class, 99.61 % in the malignant class and 99.56 % in the benign class. Similarly, it attained a 99.73 % precision score in classifying the normal class, 99.91 % in the malignant class and 99.46 % in the benign class due to its feasibility. The extension analysis is made by evaluating metrics like sensitivity, specificity. In both evaluations, the model achieves more than 99 % of the score using the MHA layer included for parallel attention, and the EBWO

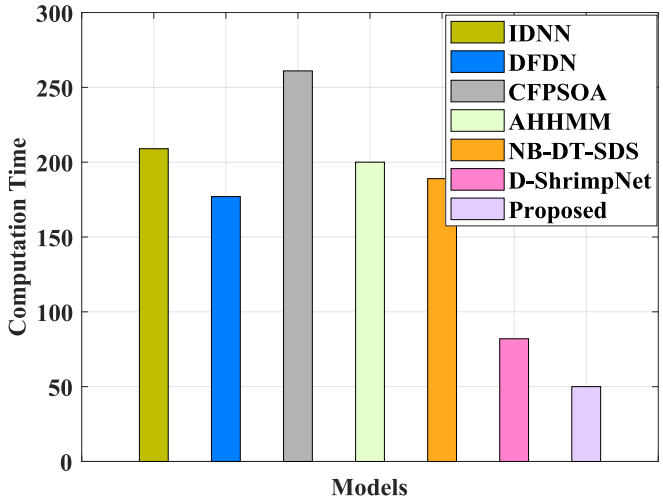


Fig. 10. Analysis of computation time analysis of proposed model vs. Existing model.

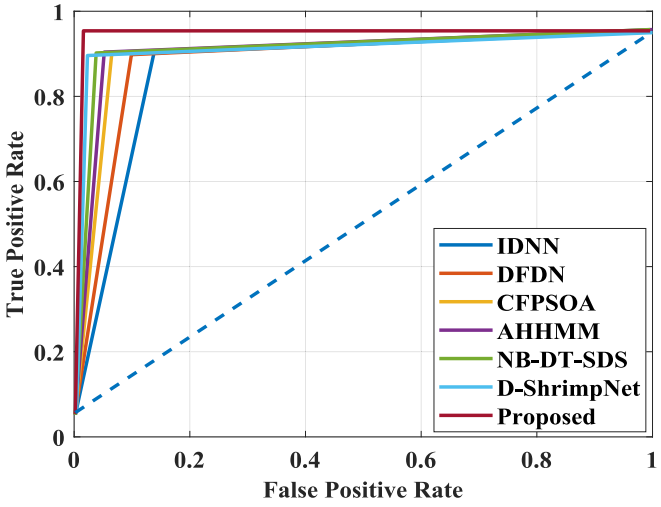


Fig. 11. ROC analysis of proposed model vs. existing models.

approach is used to attain specific features in every class. The model also demonstrated that, compared to other current models, it has a lower error rate and computation time of 0.5 and 50 s, respectively. ROC analysis was also performed to demonstrate that the proposed MHA-CDNN model outperforms state-of-the-art models for lung cancer classification. Table 5 shows a comparison of the proposed model's accuracy to that of existing models.

The proposed model performs better in terms of accuracy than existing models. The proposed model's improved performance is attributed to the Multi-head attention-based DCNN model, which processes the most relevant features selected from EBWO for classification.

5. Conclusion

This research concludes a novel deep-learning model, MHA-CDNN, for effective lung cancer classification. The model utilized the IQ-OTH and NCCD dataset containing CT scan images, pre-processed using a Disperse Wiener filter to remove noise, ensuring data quality. Relevant features were extracted from the pre-processed images using CViT, and the EBWO technique was employed to select the most significant features, enhancing the classification accuracy. The proposed MHA-CDNN model, combined with AEO for hyperparameter optimization, achieved exceptional classification performance, accurately identifying 99.43 % of normal cases, 99.51 % of malignant cases, and 99.66 % of benign cases. The model demonstrated high F-scores of 99.33 % (normal), 99.61 % (malignant), and 99.56 % (benign), along with precision scores of 99.73 %, 99.91 %, and 99.46 %, respectively, for these classes. Furthermore, it achieved impressive specificity and sensitivity scores, highlighting its reliability in lung cancer detection. The model proved superior to existing approaches, with lower error rates (0.5) and reduced computation times (50 s). These strengths indicate the feasibility of the MHA-CDNN in real-world medical applications, offering precise, efficient, and reliable lung cancer classification. CViT and EBWO techniques stand out as a significant contribution, providing advanced feature extraction and selection capabilities.

However, the model has limitations, notably its inability to segment tumors within the lung, which restricts its utility for more detailed diagnostic purposes. Future work will focus on integrating a segmentation module to identify and analyze tumor regions, enhancing the interpretability and diagnostic value of the model. Additionally, the model will be evaluated on diverse and larger datasets to validate its generalizability across different populations and imaging conditions. These extensions will further strengthen the model's applicability and effectiveness in medical imaging. This study offers a robust framework for lung cancer classification, which can be extended and adapted by researchers in the field to develop more comprehensive diagnostic systems. By addressing its current limitations and incorporating future advancements, the proposed model has the potential to make a significant impact in improving cancer diagnosis and treatment planning.

Table 5
Comparative analysis of the proposed model with existing surveys.

Author & Reference	Accuracy (%)
Akter et al. (2021)	86
Jena et al. (2021)	87.79
Sori et al. (2021)	87.8
Chui et al. (2023)	87.0
Proposed	**99.43**

CRediT authorship contribution statement

Sadam Kavitha: Writing – review & editing, Formal analysis, Investigation. **Eswar Patnala:** Project administration, Writing – original draft. **Hrushikesava Raju Sangaraju:** Resources. **Rajesh Bingu:** Conceptualization, Validation. **Salina Adinarayana:** Supervision, Software, Visualization. **Jagjit Singh Dhatterwal:** Methodology, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

Ahern, E., Solomon, B. J., Hui, R., Pavlakis, N., O'Byrne, K., & Hughes, B. G. (2021). Neoadjuvant immunotherapy for non-small cell lung cancer: Right drugs, right patient, right time? *Journal for Immunotherapy of Cancer, 9*(6).

Akter, O., Moni, M. A., Islam, M. M., Quinn, J. M., & Kamal, A. H. M. (2021). Lung cancer detection using enhanced segmentation accuracy. *Applied Intelligence, 51*, 3391–3404.

Al-Saedi, A., Mawlood-Yunis, A. R. (2022). Binary black widow optimization algorithm for feature selection problems. In *International Conference on Learning and Intelligent Optimization*, 93-107. Cham: Springer International Publishing.

Chabon, J. J., Hamilton, E. G., Kurtz, D. M., Esfahani, M. S., Moding, E. J., Stehr, H., Schroers-Martin, J., Nabet, B. Y., Chen, B., Chaudhuri, A. A., & Liu, C. L. (2020). Integrating genomic features for non-invasive early lung cancer detection. *Nature, 580*(7802), 245–251.

Chui, K. T., Gupta, B. B., Jhaveri, R. H., Chi, H. R., Arya, V., Almomani, A., & Nauman, A. (2023). Multiround transfer learning and modified generative adversarial network for lung cancer detection. *International Journal of Intelligent Systems, 2023*, 1–14.

Deepa, V. (2023). Mohamed Fathimal. "Deep-Shrimp Net fostered lung cancer classification from CT images. *Int. J. Image Graph. Signal Process., 15*(4), 59–68.

Faruqui, N., Yousuf, M. A., Whaiduzzaman, M., Azad, A. K. M., Barros, A., & Moni, M. A. (2021). LungNet: A hybrid deep-CNN model for lung cancer diagnosis using CT and wearable sensor-based medical IoT data. *Computers in Biology and Medicine, 139*, Article 104961.

Jena, S. R., George, S. T., & Ponraj, D. N. (2021). Lung cancer detection and classification with DGMM-RBCNN technique. *Neural Computing and Applications, 33*(22), 15601–15617.

Khanmohammadi, A., Aghaie, A., Vahedi, E., Qazvini, A., Ghanei, M., Afkhami, A., Hajian, A., & Bagheri, H. (2020). Electrochemical biosensors for the detection of lung cancer biomarkers: A review. *Talanta, 206*, Article 120251.

Liu, Y., & Bao, Y. (2023). Intelligent monitoring of spatially-distributed cracks using distributed fiber optic sensors assisted by deep learning. *Measurement, 220*, Article 113418.

Liu, Y., Liu, L., Yang, L., Hao, L., & Bao, Y. (2021). Measuring distance using ultra-wideband radio technology enhanced by extreme gradient boosting decision tree (XGBoost). *Automation in Construction, 126*, Article 103678.

Liu, Y., Shi, Y., Mu, F., Cheng, J., & Chen, X. (2022). Glioma segmentation-oriented multi-modal MR image fusion with adversarial learning. *IEEE/CAA Journal of Automatica Sinica, 9*(8), 1528–1531.

Lu, X., Nanehkaran, Y. A., & Fard, M. K. (2021). A method for optimal detection of lung cancer based on deep learning optimized by marine predators algorithm. *Computational Intelligence and Neuroscience, 2021*.

Naseer, I., Masood, T., Akram, S., Jaffar, A., Rashid, M., & Iqbal, M. A. (2023). Lung cancer detection using modified AlexNet architecture and support vector machine. *Computers, Materials & Continua, 74*(1).

Nasser, I. M., & Abu-Naser, S. S. (2019). Lung cancer detection using artificial neural network. *International Journal of Engineering and Information Systems (IJEAIS), 3*(3), 17–23.

Nooreldeen, R., & Bach, H. (2021). Current and future development in lung cancer diagnosis. *International Journal of Molecular Sciences, 22*(16), 8661.

Park, C. R., Kang, S. H., & Lee, Y. (2020). Median modified wiener filter for improving the image quality of gamma camera images. *Nuclear Engineering and Technology, 52*(10), 2328–2333.

Qadhi, O. A., Alghamdi, A., Alshael, D., Alanazi, M. F., Syed, W., Alsulaihim, I. N., & Al-Rawi, M. B. A. (2023). Knowledge and awareness of warning signs about Lung cancer among Pharmacy and Nursing undergraduates in Riyadh, Saudi Arabia-an observational study. *Journal of Cancer, 14*(18), 3378.

Rana, N., Latiff, M. S. A., Abdulhamid, S. I. M., & Chiroma, H. (2020). Whale optimization algorithm: A systematic review of contemporary applications, modifications and developments. *Neural Computing and Applications, 32*, 16245–16277.

Raza, R., Bajwa, U. I., Mehmood, Y., Anwar, M. W., & Jamal, M. H. (2023). dResU-Net: 3D deep residual U-Net based brain tumor segmentation from multimodal MRI. *Biomedical Signal Processing and Control, 79*, Article 103861.

Rudin, C. M., Brambilla, E., Faivre-Finn, C., & Sage, J. (2021). Small-cell lung cancer. *Nature Reviews Disease Primers, 7*(1), 3.

Shakeel, P. M., Burhanuddin, M. A., & Desa, M. I. (2019). Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks. *Measurement, 145*, 702–712.

Shimazaki, A., Ueda, D., Choppin, A., Yamamoto, A., Honjo, T., Shimahara, Y., & Miki, Y. (2022). Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method. *Scientific Reports, 12*(1), 727.

Siddhartha, M., & Santra, A. (2020). COVIDLite: A depth-wise separable deep neural network with white balance and CLAHE for detection of COVID-19. *arXiv preprint arXiv:2006.13873*.

Sori, W. J., Feng, J., Godana, A. W., Liu, S., & Gelmecha, D. J. (2021). DFD-Net: Lung cancer detection from denoised CT scan image using deep learning. *Frontiers of Computer Science, 15*, 1–13.

Taradeh, M., Mafarja, M., Heidari, A. A., Faris, H., Aljarah, I., Mirjalili, S., & Fujita, H. (2019). An evolutionary gravitational search-based feature selection. *Information Sciences, 497*, 219–239.

Vijh, S., Gaurav, P., & Pandey, H. M. (2023). Hybrid bio-inspired algorithm and convolutional neural network for automatic lung tumor detection. *Neural Computing and Applications, 35*(33), 23711–23724.

Wang, M., Herbst, R. S., & Boshoff, C. (2021). Toward personalized treatment approaches for non-small-cell lung cancer. *Nature medicine, 27*(8), 1345–1356.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22–31).

Xue, Y., Xue, B., & Zhang, M. (2019). Self-adaptive particle swarm optimization for large-scale feature selection in classification. *ACM Transactions on Knowledge Discovery from Data (TKDD), 13*(5), 1–27.

Zhu, Z., He, X., Qi, G., Li, Y., Cong, B., & Liu, Y. (2023). Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. *Information Fusion, 91*, 376–387.

Zhu, Z., Wang, Z., Qi, G., Mazur, N., Yang, P., & Liu, Y. (2024). Brain tumor segmentation in MRI with multi-modality spatial information enhancement and boundary shape correction. *Pattern Recognition, 153*, Article 110553.