Al-Nahrain Journal for Engineering Sciences **NJES** 28(1)97-120, 2025
http://doi.org/10.29194/NJES.28010097

Navigating the Challenges and Opportunities of Tiny Deep Learning and Tiny Machine Learning in Lung Cancer Identification

Yasir Salam Abdulghafoor[1], Auns Qusai Al-Neami[2], Ahmed Faeq Hussein[3]

Authors affiliations:
1*) Dept Biomedical Eng., College of Engineering, Al-Nahrain University
yasir.salam.phd2023@ced.nahrainuniv.edu.iq
2) Dept Biomedical Eng., College of Engineering, Al-Nahrain University
auns.q.hashim@nahrainuniv.edu.iq
3) Dept Biomedical Eng., College of Engineering, Al-Nahrain University
ahmed.f.hussein@nahrainuniv.edu.iq

Paper History:

Received: 13th Aug. 2024

Revised: 8th Sep. 2024

Accepted: 20th Oct. 2024

Abstract

Lung cancer is the most common dangerous disease that, if treated late, can lead to death. It is more likely to be treated if successfully discovered at an early stage before it worsens. Distinguishing the size, shape, and location of lymphatic nodes can identify the spread of the disease around these nodes. Thus, identifying lung cancer at the early stage is remarkably helpful for doctors. Lung cancer can be diagnosed successfully by expert doctors; however, their limited experience may lead to misdiagnosis and cause medical issues in patients. In the line of computer-assisted systems, many methods and strategies can be used to predict the cancer malignancy level that plays a significant role to provide precise abnormality detection. In this paper, the use of modern learning machine-based approaches was explored. More than 70 state-of-the-art articles (from 2019 to 2024) were extensively explored to highlight the different machine learning and deep learning (DL) techniques of different models used for the detection, classification, and prediction of cancerous lung tumors. The efficient model of Tiny DL must be built to assist physicians who are working in rural medical centers for swift and rapid diagnosis of lung cancer. The combination of lightweight Convolutional Neural Networks and limited resources could produce a portable model with low computational cost that has the ability to substitute the skill and experience of doctors needed in urgent cases.

Keywords: Lung Cancer, Tiny Machine Learning, Tiny Deep Learning, Automated Diagnosis.

"التنقل بين التحديات والفرص التي يوفرها التعلم العميق الدقيق والتعلم الآلي الدقيق في التعرف على سرطان الرئة

ياسر سلام عبد الغفور، انس قصي هاشم، احمد فائق حسين

الخلاصة

ان سرطان الرئة هو أكثر مرض شائع وخطير والذي اذا عولج بصورة متاخرة ممكن ان يقود الى الموت. من الارجح معالجته بنجاح اذا تم اكتشاف هذا المرض في مرحلة مبكرة قبل ان تصبح الحالة مرضية اكثر سوءا. ان اختلاف شكل و حجم وموقع العقد اللمفاوية يمكن ان يكشف عن انتشار المرض حول هذه العقد، ولهذا فان التعرف على سرطان الرئة في المراحل المبكرة يعتبر مساعدا وعلى نحو ملحوظ للاطباء. ان سرطان الرئة يمكن تشخيصه بنجاح عن طريق الاطباء ذوي الخبرة، على كل حال، ان الخبرة المحدودة للاطباء ممكن ان تؤدي الى عدم تشخيص المرض وتسبب مشاكل طبية للمريض، في مجال انظمة الكمبيوتر او الحاسوب المساعده، هناك العديد من الطرق والستراتيجيات للتنبؤ عن مستوى السرطان الخبيث والتي تلعب دور مهم لتوفير تشخيص دقيق للحالات الغير طبيعية، في بحث المراجعة هذا تم الاطلاع على استخدام الطرق الحديثة لتعلم الالة، اكثر من 71 بحث حديث للسنوات (2019-2024) تم الاطلاع عليها على نطاق واسع لتسليط الضوء على تقنيات مختلفة لتعلم الالة والتعلم العميق ولنماذج مختلفة استخدمت لتشخيص وتصنيف وللتنبؤ باورام الرئة الخبيثة، يجب بناء نموذج كفوء للتعلم العميق الدقيق لمساعدة الاطباء الذين يعملون في المراكز الطبية النائية من اجل تشخيص سريع لمرض سرطان الرئة، ان الدمج مابين الشبكات العصبية الملتوية الخفيفة الوزن مع المصادر المحددة ممكن ان ينتج نموذج محمول منخفض الكلفة الحسابية وله القابلية على تعويض خبرة ومهارات الاطباء اللازمه والمطلوبة في مثل هذه الحلات الطارئة..

1. Introduction

Numerous techniques have been introduced to evaluate clinical records to obtain useful information and create a meaningful projection of the prognosis of patients with cancer in response to the public's on-going scares about the identification of cancer. On the basis of clinical datasets, precisely predicting the proficiency of a given medication or identify a combination of effective therapies on clinical practice is immanent. Even if traditional methods of machine learning (ML), such Artificial Neural Networks and Support Vector Machines, have demonstrated promise, more space for advancement undoubtedly still exists.

Lung cancer is one of the deadliest diseases known to science because it causes many fatalities worldwide. A study estimated that 2.21 million cases of lung cancer were found in 2020, and that lung cancer killed 1.8 million people [1]. According to the World Health Organization's 2020 report, which predicted the mortality rate to be 1.80 million, lung cancer is the deadliest type of cancer overall [2].

Furthermore, 80%–90% of lung cancer cases resulted from smoking, making it the main and primary cause of lung cancer. In addition to smoking, the following factors increase the risk of lung cancer to 10%, 9–15%, and 1%–2%: radon exposure, exposure to carcinogens (such as uranium and asbestos), outdoor air contamination, respectively. The total risk of developing lung cancer may reach 100% because of the interactions between exposures [3, 4].

Small-cell lung cancer (SCLC) and non-SCLC (NSCLC) are the two main categories of lung malignancies. NSCLC accounts for 80%–85% of all lung cancer cases, and it is the most recurrently type, progressing and sophisticating more slowly than SCLC [5]. The stages I–IV of lung cancer progression can be discriminated, with stages I and IV representing the least and hardest stages, respectively. **Fig. 1** [6, 7] shows the chest X-ray sample images of SCLC and NSCLC. The majority of NSCLC stages I and II can be administered and ablated surgically. Chemotherapy, targeted therapy, immunotherapy, and other treatments are used to treat NSCLC in stages III and IV when surgery is not a choice among other medications [8-10].

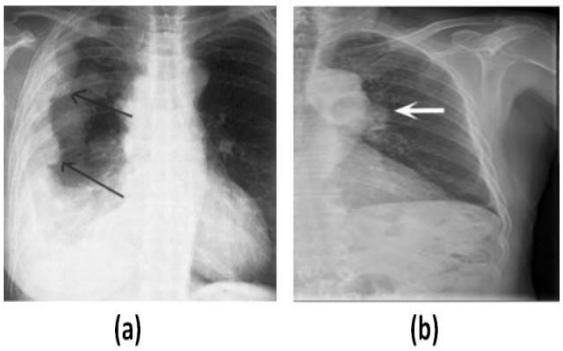


Figure (1): SCLC & NSCLC chest X-ray sample images, a) is SCLC, and b) is NSCLC

Lung malignancies usually do not show any noticeable signs. However, when the disease advances, other symptoms that may arise include shortness of breath, chest pain, coughing, and abrupt weight loss. Meanwhile, 75% of patients with lung cancer receive their initial diagnosis at the progressed stage (III or IV) because lung cancer typically remains nonvisible during the early stages (only 16% of cases are discovered at this time) [11].

ML is a branch of artificial intelligence (AI) that employs arithmetical algorithms to find patterns in data and make forecasting [12]. It has been widely used in sophisticated approaches for early detection, cancer type classification, signature extraction, tumor microenvironment deconvolution, prognosis prediction, and drug response assessment [13, 14]. AI's ML branch of research uses statistical and mathematical procedures to teach computers how to learn from data and solve problems on the basis of whether the data are labelled or not, and learning can be either supervised or unsupervised [15, 16].

The United Nations has established the 2030 agenda for sustainable development, which is a comprehensive framework focused on fostering peace and prosperity, anchored by 17 sustainable development goals [17]. These goals serve as a universal call to action for all countries to strive for a future that balances environmental, economic, and social sustainability. In response, Edge ML (EML) and Tiny ML (Tiny ML) have risen as sustainable alternatives, enabling the execution of ML models on smaller, lower-powered devices like mobile phones, wearable devices, and Internet of Things (IoT) devices [18].

This study provides a comprehensive review of state-of-the-art ML and deep learning (DL) techniques for lung cancer diagnosis, focusing specifically on Tiny ML and Tiny DL models. These models were highlighted as key solutions for resource-constrained medical environments, offering low-power, efficient alternatives to traditional DL models. In this paper, optimization techniques, such as quantization, pruning, and clustering, which reduce computational costs, were discussed. Future research directions aimed at improving model accuracy and practical deployment in real-world healthcare settings were outlined.

Designing accurate and efficient models for these devices is challenging due to their limited computing and memory resources. Model compression techniques, including pruning, quantization, and knowledge distillation, have been widely used to address these challenges by reducing the size and computational complexity of the models. Moreover, most of these model compression techniques target uni-modal models to be compressed for sustainable edge hardware deployment [19].

A typical ML method is unable to provide outcomes that can be trusted because medical photos of variant people differ fundamentally. Recently, DL techniques have been successfully used in several fields, most notably in medical image analysis. These methods are feasible and streamlined for analyzing medical imaging to determine diseases, particularly cancer.

The paper is organized as follows: Section 2 provides an in-depth review of the applications of AI in biomedical image processing, with a focus on DL techniques and their relevance to lung cancer diagnosis. Section 3 discusses Tiny ML, outlining its potential for developing lightweight models suitable for resource-constrained environments. Section 4 details various methods for optimizing computational resources in Tiny ML, such as quantization, pruning, and clustering techniques. Section 5 explores Tiny DL (Tiny DL), with a focus on compact DL models designed for embedded systems. Finally, Section 6 presents the application of Tiny ML in healthcare, followed by a discussion of the limitations of existing models and a conclusion summarizing the paper's findings and potential future directions.

2. Scope of AI in Biomedical Image Processing

DL techniques are adjacent to ML techniques that permit the training of a model on the basis of the result and guess the outcome by using a given data set. Neural Networks (NNs) with many layers, such as an input layer, multiple hidden layers, and an output layer, are used in DL techniques. DL models are taught with higher precision because they have multiple layers. Four categories of DL models can be recognized by their learning approaches: supervised learning models, unsupervised learning models, semi-supervised learning models, and reinforced learning models [20].

In many different scopes, the algorithms of DL can be applied to enhance picture identification performance, with prominent and fructified outcomes. One type of DL application outside of the clinical domain for biometric palm vein attribute distinction is the Convolutional Neural Network (CNN) [21]. It was applied to text categorization and classification of texts into shared datasets [22, 23]. By implementing trait extraction with Karhunen–Loève Transform and Haar wavelet, NNs were utilized as a novel response to the expanding fixture for dial emotion discrimination [24]. Strong stock market prediction models were constituted by the implementation of efficient methods of ML [25].

Researchers in medical sciences were inspired by the success of DL networks to implement them to medical data images for errands such determination of lung disease. The outcomes showed that deep networks can potently extract beneficial features that characterize between different image classes [26]. **Fig.2** [6, 27-29] shows different medical image modalities of lung cancer and COVID-19. The most widely employed DL framework is CNN. Its ability to extract variant type attributes from images has led to its application in the classification of different medical photos [30].

Deep neural networks can decipher and handle complex problems like remedying of natural language, image manipulation and image processing (Fig. 3) [31], and machine vision. DL can also be used for identifying dynamic image traits [32]. DL is progressing for results improvement [33]. These inventions were made possible for data rather than handmade traits that depend on domain-related information [34].

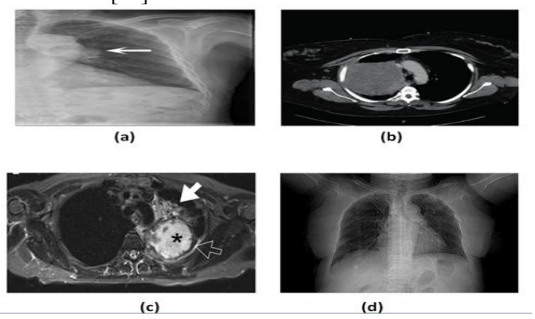


Figure (2): Different samples of medical images modalities, a) chest X-ray lung cancer image, b)CT lung cancer image, c) MRI lung cancer image, d) chest X-ray image of Covid19

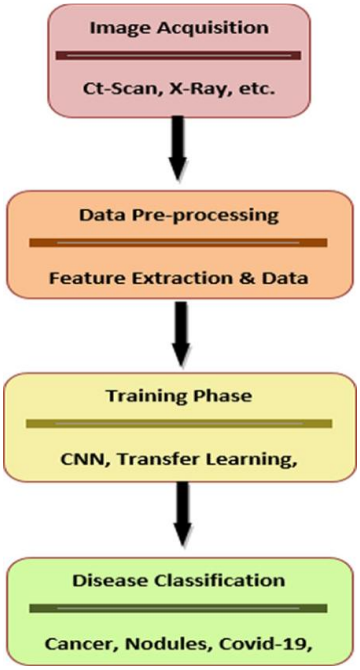


Figure (3): Methodology of deep learning for image processing

The CNN group of ML models contain multiple layers of convolution that may be taught to discriminate image traits. Most of these layers are salutary for image processing in computer vision applications. During the training process, deep CNN can close the contrast between its thrash model and seeded components. During this procedure, the characteristics of various components are produced in numerous lapped layers, as shown in Fig. 4 [35]. This method has the ability of precise structural component identification and classification while testing a fresh input image [36-38]. Various techniques of ML and DL used for lung tumor detection are shown in Tables 1 and 2.

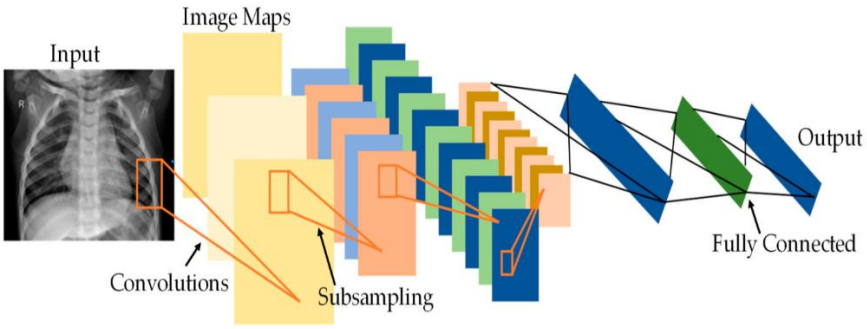


Figure (4): Design of classical CNN for medical diseases prognosis.

Table (1): Different machine learning techniques used to detect lung tumours.

Authors	Year	Technique	Results	Limitation
Punithavathy et al. [39]	2019	Using Support Vector Machine SVM techniques for lung cancer classification for Computed Tomography/Positron Emission Tomography CT/PET images	98.1% achieved accuracy	Using traditional hardware
Saleh et al. [40]	2021	Using Convolution Neural Networks with Support Vector Machine CNN-SVM classifier using CT images	97.91% accuracy	Limited dataset& using traditional hardware
Morgado et al. [41]	2021	Using image phenotypes and the mutation status investigation	range values of area under curve AUC is 0.725 -0.737	Limited features selection using traditional hardware
Zhang et al. [42]	2021	Using Machine Learning ML for radiomics approach in CT images	specificity 96%, AUC = 78%, and sensitivity = 92%	Individual limitations& using traditional hardware
Hussain et al. [43]	2022	Using Naïve Bayes NB, Decision Tree DT, SVM classifiers with Gaussian Radial Basis Function RBF polynomial in MRI images	Accuracy = 100%	Limited dataset, low severity level, using traditional hardware
Perumal et al. [44]	2022	Using Partial Least Square Linear Discriminant PLS-LD binary classifier with Raman spectroscopy	specificity = 83%, Accuracy = 85%, sensitivity = 87%, and	Small patient cohort, using traditional hardware.
Ishii et al. [45]	2022	Using Artificial Neural Networks ANN to classify photomicroscopic images into 4 groups	Batches accuracy is 0.94%to 0.99% prediction accuracy is 0.75%to 0.95&	Criteria of cases is limited, limited number of cases, using traditional hardware
Carrillo Perez et al. [46]	2022	Using ML techniques to study Ribonucleic Acid RNA modalities in whole slide imaging	Area Under the Precision-Recall Curve AUPRC of 0.980 +-0.016, AUC of 0.993+-0.004 & F1 score of 96.81+-1.07.	Limited diagnosis capabilities, using traditional hardware
Nancy et al. [47]	2022	Using different ML techniques with Contrast Limited Adaptive Histogram Equalization CLAHE algorithm in CT images	Good performance of PSO SVM	Limited training data, using traditional hardware
Kwon et al. [48]	2023	Use three ML algorithm: Adaptive Boosting AB, Multi-Layer Perceptron MLP, Linear Regression LR. detecting cancer marker cell free Deoxyribonucleic Acid DNA & Copy Number Variation CNV	Higher value of AUC in combined analysis	The process is not efficient, limited data, using traditional hardware
Huang et al. [49]	2023	Using the Multi-Instance Learning (MIL) model was to classify lung neoplasms using scleral images screen	Specificity of 0.828 +- 0.095, AUC of 0.897 +-0.041, sensitivity of 0.836 +-0.048,	Limited data, limited diagnosis. limited algorithm

Kumar & Rao [50]	2023	Using a based Weighted Classifier of Labeled Priority for lung tumor detection in MRI images	96.6% is achieved specificity,	Limited diagnosis, limited data, using traditional hardware
Earnest et al. [51]	2023	Using different ML techniques were employed to predict the quality and timeliness of Victorian Lung Cancer Registry VLCR dataset	AUC IN SVM 0.89, in K Nearest Neighbour KNN is 0.85, SVM is better than KNN,	Limited tools and data, not vulnerable population, using traditional hardware
Mohan & Thayyile [52]	2023	Using different ML techniques in x-ray, CT images	Random Forest RF technique is better performance	Limited training data, using traditional hardware
Dirik [53]	2023	Using a different ML algorithms for lung cancer detecting	91%. of accuracy	Limited training data, using traditional hardware
Bhuiyan et al. [54]	2024	Using different models of ML models and compare between them such Light Gradient Boosting Machine light GBM, SVM, LR, Adaptive Boosting AdaBoost & Extreme Gradient Boosting XGBoost	XGBoost is the best with accuracy of 96.92%	Limited number of cases, using traditional hardware, need to innovative technologies such block chain

Table (2): Various techniques of traditional deep learning for lung cancer detection

Authors	Year	Technique	Results	Limitation
Xu et al. [55]	2019	Using CNN and Recurrent Neural Networks RNN for solitary sclerosing papillary SSP tumor detection in CT images	Probability value $p<0.05$, AUC =0.74%,	Less interpretable feature presentationusing traditional hardware
Jakimovisky & Davcev [56]	2019	Using a double and regular Convolution Deep Neural Networks CDNN for lung tumor detection in CT images.	0.876 % accuracy of regular CDNN.	Lower certainty of classification, DNN need to be modified, using traditional hardware
Park & Monahan [57]	2019	Using CNN with genetic algorithm for lung cancer detection in X-ray images	97.15% is achieved accuracy	Need to user interface, using traditional hardware
Subramanian et al. [58]	2020	Using Visual Geometry GroupVGG-16 Net, LeNet& AlexNet, with SoftMax classifier to predict lung cancer in CT Images	99.5% of Accuracy	Internet of Thing IoT application and cloud computing need to be developed, using traditional hardware
Al-Yasriy et al. [59]	2020	using AlexNet CNN to classify lung cancer in CT images	95% for Specificity, accuracy ups to 93.548%. 95.714% for sensitivity	Limited training dataset, limited classification tools, using traditional hardware
Elnakib et al. [60]	2020	Using different compact Deep Learning DL model including Alex, VGG16, and VGG19 networks. for lung tumor detection in CT images	Specificity of 95%, accuracy of 96.25%, sensitivity of 97.5%,	Needed to fusion of different models, high false positive and false negative cases, using traditional hardware
Lin et al. [61]	2020	Proposed Taguchi parametric optimization with (two-dimension 2D CNN) in CT images	The proposed method is finer than the original 2D CNN by 6.86% and 5.29%	Depth layer &optimal size algorithm should be developed, using traditional hardware
Kalra et al. [62]	2020	Using a convolutional neural network and designed a 3D CNN model in CT images	Specificity with 97.68 %.0.97% accuracy precision with 87.31 %, recall with 74.46 %	Need to initial lung segmentation and to deeper layer and extensive parameters tuning, using traditional hardware
Amma et al. [63]	2020	Using architecture of Visual Geometry Group with 16 layers for lung cancer identification in CT images	Probability rate of lung cancer was approx. 59% and of not occurring lung cancer was approx. 40%.	Limited convolution layers, using traditional hardware
Zhan [64]	2021	Developed (CNN) sensors to train the model by CT images used dynamic sampling techniques &	0.84% accuracy rate	limited data source, low specificity, model require further research and validation,

		transfer learning.		using traditional hardware
Mohammed & Cinar[65]	2021	Using different CNN models include: AlexNet, ResNet18, GoogleNet, and Residual Net ResNet50 model for lung cancer detection in CT images	Accuracy range of all models are 88.4 &to 100%	Data is limited, the model need to compare with other models, using traditional hardware
Abd Al-Ameer et al. [66]	2022	proposed a different CNN model including Inception V3, Random Forest, using histopathological	Specificity measure 96.88%., accuracy 97.09%, F-score measure 97.09%, precision 96.89%, recall 97.31%,	Limited features &convolution layers, need to user interface, using traditional hardware
Ren et al. [67]	2022	Proposed a hybrid framework called Latent Constraint Generative Adversarial Network LCGANT framework with 2 parts to generate and to classify lung cancer images in histopathological images	99.84% +-0.156% (F1-score).,99.84% -+0.156% (accuracy), 99.84% +-0.156% (sensitivity 99.84%+- 0.153% (precision).),	Limited dataset, low resolution synthetic images, using traditional hardware
Humayun et al. [68]	2022	Employ a different CNN models, compared for 20 epochs structure in ImageNet dataset	The accuracy of VGG 16 is 98.83 %, VGG 19 is 98.05 %, and Xception is97.4 %.	Limited clinical data, using traditional hardware
Said et al. [69]	2023	They proposed UNet with Transformer UNETR network of 2 parts for segmentation and classification lung cancer using CT images	98.77% as classification accuracy and segmentation accuracy of 97.83%.	Requires a high-rendition GPU to run easinessly, using traditional hardware
Üzülmez & Çifçi [70]	2024	They suggested 4-layer CNN with uncertainty quantification and compare it with other CNNS such ResNet50, AlexNet, VGG16 and inceptionv3 using CT images using CT images	The proposed 4-layer CNN is the best by achieving accuracy of 0.971	Using traditional hardware, limited training data,

DL and other AI-based methods have been developed in recent years to detect lung cancer early [71]. By using medical imagery, such as X-rays, CT scans, and MRI scans, DL techniques have greatly improved medical diagnosis. A doctor's physical diagnosis based only on the photographs may differ from another's. DL-based techniques yielded more accurate outcomes [72]. In contrast to the conventional approach, ML has demonstrated remarkable success in the medical domain, primarily in the areas of disease detection and diagnosis. Most recently, the DL technique reduced the challenge of manual feature extraction and improved classification accuracy [73]. In accordance with previous studies (Tables 1 and 2), some questions were asked here: What are the features (memory, Central Process Unit CPU and Graphical Process Unit GPU) of computers or hardware that have been used in these studies? How much power and energy have been consumed in the hardware system? How much computing time has been taken to provide results? Are these research studies beneficial for physicians who are working in rural medical centers?

The limitations of the existing works

Previous studies have shown several limitations for ML and DL techniques (Tables 1 and 2). The most common limitations are the limited or small medical training data that had been used, which could reflect on the model's accuracy. The traditional techniques also used traditional hardware that requires high computational cost (large memory and long processing time). Even the deployment of traditional DL models in devices of edge computing, such as Raspberry-Pi or Jetson Nano, is considered challenging due to high computational costs in terms of space and time [74]. Tens of millions of parameters are typically present in well-known deep NNs, which require a large amount of memory to run the most advanced model. Furthermore, deep NNs require high-performance hardware resources, which makes it difficult to implement the most advanced model on portable devices [75].

Even while ML and CNN models have shown encouraging outcomes in earlier research apart from the technical limitations, a number of issues still need to be resolved. One significant drawback is the computational cost of these models, which are unsuitable for devices with limited resources due to their millions of parameters that require fine tuning and longer time. For this reason, researchers started using Tiny ML and Tiny DL techniques. They developed lightweight CNN models that can be applied on limited constrained resources and used in many medical applications and the diagnosis of different diseases, especially lung tumor/cancer prediction.

3. Tiny ML

Tiny ML is a nascent field that combines ML and embedded systems. Tiny ML tools are effectively

provided to develop models of ML, which can be implemented on limited resource devices. The process of Tiny ML prevalence begins with data collection from the hardware, where an inference engine is needed. The data could either be closely imported into tools of friendly user. Edge Impulse Studio can be logged on onboard storage. The collected data set is trained by the ML model, and then the model is forwarded into a shape that can be implemented straightaway on the Microcontroller Unit (MCU). The trained model is employed by the MCU for inference in subsequent iterations. Two key challenges need to be identified to open the full possibility of ML for IoT systems.

A. Interoperability: No unified standard exists for employing Tiny ML, and the MCU market is relatively segmented. Manual application and hardware-specific optimizations are required because the implementation of specific platforms is not scalable [76].

B. Characterization of performance: Tiny ML is limited by nonexistence of a standardized framework. The hardware's performance is difficult to appraise in a vendor-agnostic and neutral manner. When rendition boosts are documented, it is so hard for disbanding either they are involved to implementation of software or to hardware and whether these winnings popularize through different implementations. TensorFlow Lite (TFL) localized these two contestations [77]. TFL libraries facilitate the most feasible implementations of ML models, so it has become synonymous to Tiny ML

Tiny ML is a prototype that easily implements the ML actuating at the brim devices with base rate memory requirements and processor [78]. Hence, a few milliwatts power or less is expected to consume within such systems [79]. Tiny ML practiced challenges are tremendous. In recent NNs, the order of billions of number of required parameters have increased among the best presently plenteous technologies [80]. With chunkier networks having wider applicability and choicer results, the size of these networks is proportional to the energy required to run them, making this direction of upward NNs uncomfortable at a wide range [81]. Research direction is another reason why Tiny ML can be considered as essential rather than overgrown.

The solution development of Tiny ML needs two core traditional workflows: ML-guided and Hard Ware (HW)-guided, and co-design a third that can be considered more fangled method. The design of ML framework and its hardware embodiment can be arbitrated by classical workflows [82, 83]. First, a suitable model can be created, trained, and tested by the connoisseurs of ML for the scope of problem.

They ameliorate the parameters of the model, and a satisfactory device is deployed by this solution. Second, sophistication targets the produced ameliorated hardware because the bandstand of the hardware is not pre-tidied, and then utilization particularly tiny techniques and models. The co-design represents the holistic methodology, whereas ML-guided and HW-guided represent Tiny ML solutions (Fig. 5) [84]

In the workflow of ML-oriented (Fig. 5a), the majority of experience are in the layout, rehearsal, adaptation, and assessment of ML models. The hardware bandstand selection is sustained or bounded because of the necessity or specific industrial requirements [84, 85]. Embedded devices ported by modern NN models are a typical example of this workflow [86]. The efficient implementation of power amortization, usage memory, and latency is required for extensive experimental investigations and cloud solutions compared with such devices of short supply resources.

The ML-guided workflow can be identified by the next stages:

Model layout: ML experts mold, rehearse, and validate an exhaustive model comfortable for the scope of problem. The hardware bandstand is ignored in this stage to actualize maximum rendition and generalization, but it is highly hinged on the complexion of this scope.

• Model optimization: Various strategies are included in this stage to bargain efficiency rendition.

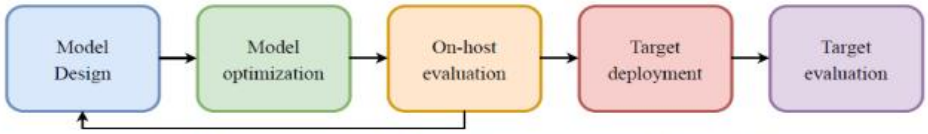
• **On-host assessment:** In the specifications, the ameliorated model is evaluated against the parameters; rendition needed, and it is redesigned if there is any destitution.

• **Objective deployment:** Specific amelioration is implemented to the model to increase inference efficiency, and the specific features of the hardware device are leveraged.

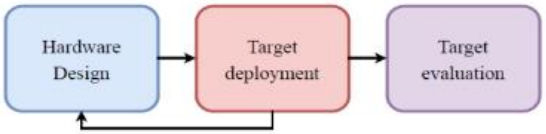
• **Objective assessment:** Assessment of the system in output is performed.

In the HW-oriented approach (Fig. 5b), the design of improved hardware bandstand that are ameliorated for embedded applications is mainly focused by the developers. Dealing with the bottlenecks in subsistent architecture with respect to ML framework computations is required to deploy present and greeter algorithms of ML. Such hardware expedition modules and NNs needed to be designed to enhance throughput depreciation, e.g., decreasing the degree of complication in convolution layers [87, 88]; efficient, feature-rich, and low-robustness perceptron [89]; and enhanced caches of data [90]. In other cases, new hardware platforms optimized for embedded applications are designed by developers with outreached competencies of digital signal processing [91]. This sophistication of ameliorated libraries is required [92, 93] to extract the most performances.

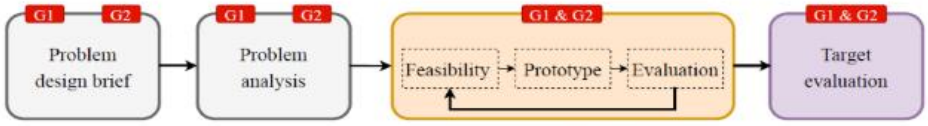
As shown in Fig. 5c of co-design workflow, both sides of the development from the start is integrated by the approach to obtain more benefits from further rendition enhancement and resource depreciation. In particular, while the previous workflows steps are separated by model optimization and hardware design in Figs. 5a and Fig. 5b, they are co-optimized and intertwined. In some situations, detailed architectures are developed for rimmed algorithms of ML on Fine-grained Balanced Graph Attention [94]. In other situations, neural computations of networks, upon request of an expeditor, are allowed through HW-enlightened rehearsed methodologies using compute in-memory hardware [95].



(a) The **ML-oriented** workflow for implementing a TinyML-based system.



(b) The **HW-oriented** workflow for implementing a TinyML-based system.



(c) The **co-design** workflow for implementing a TinyML-based system.

Figure (5): Tiny ML based system work flow.

4. Methods for Resource Optimization

Arithmetical resources of microcontrollers, such as usage processor and memory, can be saved by different approaches and methods [96] when used in Tiny ML devices. Simply sending dense arithmetical assignments headway to yate ledge [97] is one of the overall routes for decreasing arithmetical gestation on basal edge devices. However, this method may lead to increased depreciation of power in Tiny ML devices because powering an on-device NN is less energy amortizing than the process of uploading and extraditing data [98]. This may be climactically leastwise for battery-operated devices. So, a resource problem is preferably resolved within the edge device by arithmetical means. The processes of pruning, quantization, and clustering methods can achieve this reduction in processor usage and size of the ML model.

4.1 Quantization

a microcontroller, floating-point operation is performed. This method is usually needed for running NNs. In the output and heuristics mode, NNs typically use high-exactness 32-bit floating-point data [99]. However, a great deal of memory is required for these floating-point NN operations, productivity power of system, and celerity of timepiece from a microcontroller [100]. In some trims, hardware floating-point operations cannot be performed by microcontrollers, such as M4F processors beginning with the hardware floating-point unit in the Advanced RISC Machine Arm Cortex-M processor series [101]. Still, employing the C library of the arm software floating-point by computational means can solve this problem through emulation software of floating-point (EFP), or fixed data point format can be converted by the floating-point data [73, 102]. The model's memory footprint by 75% is lowered by quantization of 32-bit floating-point data to 8-bit fixed point data, and the microcontroller could be run much faster by integer operations [103]. In [104], the quantization of fine-tuned CNN with different activations and weight bit-width and 30 epochs of subdivision was tested. The error rate of categorization increased from 6.98% to 8.30% only, as shown using four-bit fixed-point activation and weight values, compared with the values of floating point. Good results were documented in [105] when inspecting variant datasets with four bit precision quantization, with 50% memory and 75% provident computation reported with plopping in accuracy by 5%. However, the results indicated that the accuracy began to drop more briskly when 3-bit or 2-bit ultralow precision was used.

A model's bit-width activation and weighing can be ameliorated partly by the CPU restriction memory of a microcontroller [106] by mixed-precision quantization. In this method, each layer can be quantized partly to variant bit widths to avert data loss and enable accuracy [107]. However, a major computational challenge is typical bit widths investigating for all layers.

4.2 Binarization

Binarization is another form of quantization, whereby weightings, operands, and activations are reduced to a single bit, and the compact level of bit width is enabled [108, 109]. In binarized neural networks, all calculations use the binarized weightings of the activation and weights and bit-wise. XNOR operations are substituted by arithmetic operations. As a result, the power efficiency increased, and the memory required (32) is reduced by 1-bit operations and the kits of memory entrance (32). In [110], the authors proposed dual enclosure NNs particularly styled for limited enclosure devices.

4.3 Pruning

The pruning of unused features of an NN can lower computational complexity. Pruning techniques can be divided into two main classes: modulated and non-modulated pruning [111]. In modulated pruning, the conduits or filters are eliminated. In non-modulated pruning, the weight capitation relevance is eliminated by concocting it to zero [96]. In addition, a different pruning approach is possibly combined. For example, in [112], the authors introduced a procedure whereby neural architecture search was combined with unstructured and structured pruning approaches, which spontaneously meticulous, strewn, and lightweight CNN architecture was found. The pragmatism of NN model's weightings zeroing out is named quantum pruning when a sixfold improvement in model compression can be brought and it performs to a sprinkled model [113]. The method's downside is that it also leads to the use of sparse convolution libraries and complications of sprinkled matrix that require additional arithmetic power [114, 115].

The procedure of weight pruning comforts the use of microcontrollers because of the importance of the benefits of model size compression, and it can be more significant than the additional arithmetical problem from sprinkled complications. The forms of layers and matrices of weight is changed by eliminating the combinations of weight links in modulated pruning method [112]. When conduits or filters are eliminated, the network's heuristics rapidity increases, and the size of the model decreases. A lightweight network is produced by channel-level pruning, but when the bade of the whole network is slashed, it can downgrade the model's rendition and nicety. Hence, methods of unstructured pruning are recommended to be used whenever possible. In [116], a 3.54-fold mean size model was reduced by 88%, and performance speedup was reported when the variant weight and burl pruning groups were tested with a two-manner single instruction multiple data (SIMD) unit for 16-bit stationary-point mathematics by Arm Cortex-M4 microcontroller, 512kB flash storage, and 128kB Static Random-Access Memory SRAM. The authors suggested a pruning procedure, named Scalpel, which is a combination of burl pruning and SIMD-aware pruning of weighing. The benefits were smaller memory and better efficiency for the model than the basal procedures of pruning.

4.4 Clustering

The process of reducing the number of individual weight values is known as clustering, whereby a smaller number of centroid weight values are replaced with the model's weight values that are computed from the grouped weights of the original model [113]. Weight clustering does model compression by reducing memory usage. So, the original CNN is five times bigger than the compressed model. If the weight clustering process is compared with quantization process, higher accuracy and compression ratio is brought by weight clustering, but the two can still be used effectively together [117]. The k-means clustering algorithm is used with the weight clustering process [117, 118]

4.5 Software and Libraries of TinyML

A number of bandstands, libraries, and frameworks are chosen, and they can be updated with the technology of Tiny ML as follows:

4.5.1 TFL: TLF

Is an open source of a DL framework that supports edge-aware learning inference. Edge-aware ML at the device may be approached by this framework by employing five significant constrains (size, time, correlation, power amortization, and privacy) [119]. Other software libraries of tiny ML are available, such as NanoEdge, Pytorch, Edge impulse, micro Tensor Virtual Machine mTVM, u Tensor, STMicroelectronics STM32cubeAI, and Embedded Learning Library.

5. Tiny DL

Lightweight DL is a branch of ML and AI possessing compactness and efficiency while developing DL models. It is a typical model for IoT devices, low-powered mobile phones, and embedded systems because of low latency and speed. Pruning and quantization processes are employed to remove nonessential parameters and reduce the accuracy of model weight to make lightweight DL models. Fine-tuning a pre-trained DL model on a smaller dataset is implemented using Transfer Learning [120].

At a historical rate, many devices of IoT depended on microcontrollers is briskly augmented, reaching 250 B [96], granting applicability to many applications, including precision agriculture, automated retail, smart manufacturing, and personalized healthcare. A marque modern prospect of Tiny ML is given by these decreased-cost, decreased-stamina microcontrollers when these tiny devices are run by DL models. Data analytics can be performed directly close to the sensor. Thus, the AI applications' domain can be dramatically expanded.

Microcontrollers have a budget of extremely constrained resource, particularly storage (Flash) and (SRAM). The mobile devices are three times quantum larger than on-chip memory, and cloud GPUs is 5- or 6-times quantum larger than on this memory chip, making the deployment of DL extremely hard. M7 MCU-ARM Cortex solely has 1 MB Flash storage 320 kB SRAM, which make the-shelf DL models impossible to be run off. In ResNet-50 [121], the bound of storage is exceeded by 100. In MobileNetV2 [122], the crest of memory bound is exceeded by 22. A huge hiatus exists between the required and affordable hardware ability till the int8 muzzled model of MobileNetV2 still encroaches the bound of memory by 5.3 times.

In contrast to cloud and mobile devices, microcontrollers have no operating system, which is considered a bare-metal device. Therefore, laying out the DL model and the heuristic library are needed to combine to administrate the tiny resources efficiently and suit the virulence budget of storage and memory. The presence of streamlined design of network [106, 123, 124] and procedures of neural architecture search [112, 124-126] overemphasized GPU or smartphones, whereby any storage and memory are plentiful. Therefore, the produced models do not suit microcontrollers

When they just ameliorate to decrease latency or Floating-Point Operations per Second, and ML on microcontrollers have been studied in limited literature [127-130]. Microcontrollers' inference by DL is a fast-growing area. The presented frameworks have many limitations, such as TensorFlow Lite Micro [131], Cortex Microcontroller Software Interface Standard Neural Network CMSIS-NN [132], CMix-NN [133], and Micro Tensor Virtual Machine MicroTVM [134]. One method is to compress off-the-shelf networks by pruning [135-140] and quantization [97, 141-146]. Redundancy is removed, and complexity is reduced. An effective compression method is served by tensor decomposition [147-149]. Close design of streamlined and mobile-pally network is considered another method [106, 121, 122, 150]. Recently, efficient network design is dominated by Neural Architecture Search (NAS) [124, 125, 151-154]. The finesse of the search space is highly related to the performance of NAS [155]. Traditionally, inference prospectus design for NAS design search space is followed by people. For example, MobileNetV2 [122] is originated by the broadly used mobile-preparing search space [124, 125]. The input of 224 resolution with a same norm number of conduit form is used by both while investigating for block deepness, kernel sizes, and elaborating ratios.

Tiny ML can be applied by NN model compression because computational resources are relatively lacking [156]. The overstocking resources and bandwidth of memory entrance in subsumed systems are limited.

Nowadays, the compression methods of neural network, such as trimming, muzzling, forfeiture of model designing procedures, and procedure of transfer learning, have been widely used. The concept of overfitting in ML is originally emerged by pruning; and removing the convolutional kernels that have less impact or less weight conduits in the model NN is considered its main concept. The authors in [157] proposed a procedure that link filter trimming and weight trimming when the NN parameter redundancy is reduced. However, recovering the training operation is difficult, and its accuracy has no guarantee. The quantization process of the model not only augments the rapidity of actuating the enclosure model but also reduces its storage [158].

When the quantized model is applied, storage is reduced, and the embedded model running speed is increased. The trio weight muzzling procedure [159] is less precise than the asymmetric trio weight muzzling method [160], but the influence of compact is average. The forfeit model design includes SqueezeNet [161] and ShuffleNet [133]. The main notion is to collect the conduits and then calculate them for arithmetical cost reduction. However, fine model design methods highly require skill and experience.

Removing a huge number of excrescence or almost unnecessary neurons from model of the neural network could decrease the computational power for the subsumed device and decrease the amortizing of the power.

5.1 NN Compression Concept

5.1.1 Group Convolution

The spatial complexity of the model [155] is reduced using a number of deeply separable convolutional compression parameters [162].

The concept of process can be seen in **Fig. 6**. The size of systematic convolution input trait map is C × H × W. There are M output conduits, if C × K × K represents the size of convolution kernel. Afterwards, the map of feature is subdivided into G groups in accordance with the concept of group convolution, Then, the overall number of the original convolution parameters C × K × K × M becomes C/G × K × K × M, so the overall number of parameters (1-1/G) is decreased.

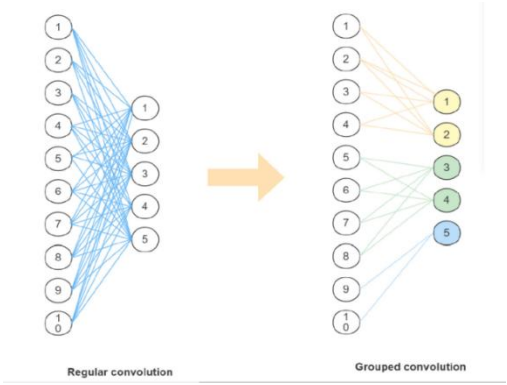


Figure (6): Set up of grouped or aggregated convolutional neural networks.

In conclusion, the configuration of aggregated convolution has a very crucial role in parameter compression. Aggregated convolution could be employed by more researchers because it is easy, and therefore, no experience and skill are needed compared with the procedures of fine-granulated model design. Furthermore, model overfitting can be prevented, and the precision of the model can be improved because of the regularization effect that exists in using grouped CNN.

5.1.2 Pruning

In this procedure, the pruning operation is performed when the group convolution is completed. Puring is a customary and broadly used approach in NNs. It emanates from the problem of the overfitting classical ML. The regular trimming of CNN is better than some weights trimmed in **Fig. 7**. So, this employ the matrix booting, opposite that, the implementation of GPU actuating is difficult and incremental.

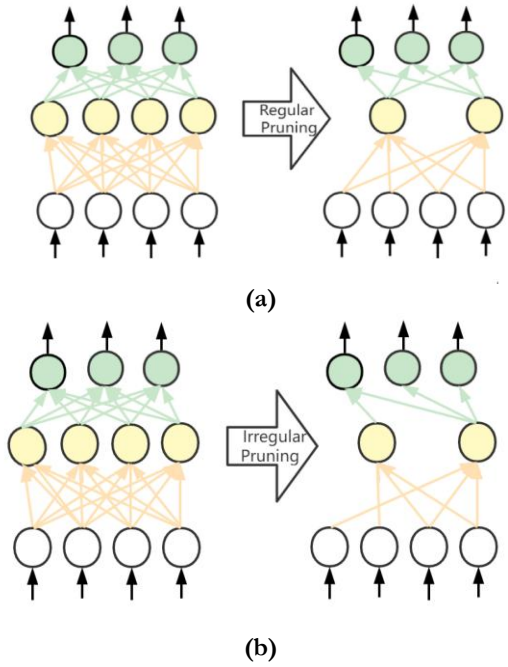


Figure (7): The pragmatism diagrams of (a)regular trimming & (b)irregular trimming.

Depending on previous precautions, the determination of the importance of each channel in each layer [163] adopts the ongoing schematizing pertinence of Batch Normalization layers. Pruning [164] removes the conduits and parameters that have lower case weights in CNN.

The weights and elements are updated to be more suitable to the original data set by performing recovery training to eliminate the amount of redundancy after pruning. afterwards, the data are recorded and remapped to another interval by quantization.

Inconsistent trio weight muzzling is a type of regular manipulating, in which the hazard of overfitting can be reduced. Then, the model is converted to TFL format by using the TFL converter [160]. Table 3 shows various studies of lightweight CNN that have been used in the detection of lung cancer, lung disease, and other diseases.

6. Healthcare and Tiny ML

The sector of healthcare is inexorably touched by the development of Tiny ML, which is essential in each worldwide community. Different organs in the human body can emit signals, so the human body is considered as a signal issuer. These data can be collected and used to abate impasses involved in healthcare by deploying sensors. The deployment of Tiny ML could solve these issues.

Previous surveys [165-167] indicated that the healthcare sector would be revolutionized by the integration of these edge devices, and enhance the well-being of people. Quality and reliable healthcare monitoring can be provided by the latter; furthermore, many health products will be improved. Moreover, modern hues of medical censoring,

prognostics, and treatment is enabled by the potential of technology, the finesse of care is improved, and patient findings is ultimately improved. Especially with processing of data will be in real-time and in low profile devices by the ability of TinyML. Patients are treated and monitored by healthcare professionals in a more potent and streamlined manner [168] when this new possibilities is opened. The previous works' analysis permit the merging of commonality pragmatism adopted to Tiny ML application in healthcare, as seen in Fig. 8[169]. Biological signals are provided by the human body, and these signals are identified by specific sensors and then transported to the enclosure system that includes DL or ML algorithms combined with a cloud habitat. This combination could give permission to the device to reveal deficiency type, censor, forecast, or help the patient.

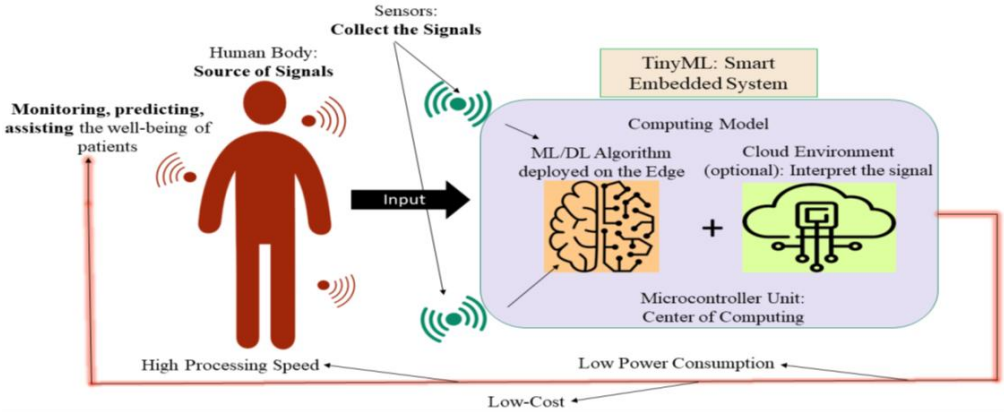


Figure (8): Evincing of the general pragmatism adopted to implement TinyML technology in the healthcare sector.

Table (3): The literature studies of using a various light weight neural networks for detection of lung cancer, lung diseases and other diseases

Authors	Year	Method	Results	Limitation
Nasrullah et al. [170]	2019	Proposed novel deep learning for lung nodule detection by Region-based Convolutional Neural Network RCNN based on learned features extracted by Compound MixNet CMixNet then classified by gradient booster machine using CT images	Specificity (91%), sensitivity (94%).	Long Training time, using traditional hardware
Sanagala et al. [171]	2019	The study proposed a fast and lightweight CNN for lung cancer detection and comparing the results with traditional CNN using chest CT images,	Accuracy of 99.5%,	*Using traditional hardware
Pasa et al. [172]	2019	The study proposed a compressed CNN model with 5 convolution layers and& average pooling layer using x ray chest images	reducing the computational, memory and power requirements significantly, the output with saliency maps and grad-CAMs found that a good visual explanation	Using traditional hardware. limited dataset, dataset need to pre-trained.
Rajaraman et al. [173]	2020	Study demonstrated use pruned CNN and fine tuning for detecting pulmonary manifestations of COVID-19 with chest X-rays.	99.01% of accuracy and AUC of 0.9972 in detecting COVID-19 findings on Chest X-rays CXRs images	Using traditional hardware.
Shuvo et al. [174]	2020	Proposed a light weight CNN for lung disease detection using scalogram based features utilizes Empirical Mode Decomposition EMD & Continuous Wavelet Transform CWT	98.70% for six-class pathological categorization are done & weighted accuracy values of 98.92% for three-class chronic categorization	Using traditional hardware
Sumari et al. [175]	2021	Proposed CNN with Gray-Level Co-occurrence Matrix GLCM for Covid19 detection using quicker & detailed diagnosis features in chest x-ray images	97.06% of accuracy	Using traditional hardware.
Srinivasu et al. [14]	2021	This study proposed light weight DL based mobilenetv2 and Long Short-	Method has 85% accuracy more than other methods and faster	The model's precision is low for

		Term Memory LSTM for skin disease detection and compared the performance with Feedforward Time-Delay Neural Network FTNN, CNN &VGG. using dermatoscopic images	recognizing & lesser computation than the conventional MobileNet model results in minimal computational efforts.	poor illumination photograph conditions, the proposed approach is less efficient than laboratory test, using traditional hardware
Gouher et al. [176]	2021	The study proposed a lightweight CNN called UNet that used to detect the lung cancer from chest X-ray images dataset	Accuracy of 90%	Using traditional hardware, and limited training dataset
Shukla et al. [177]	2021	This paper focuses on detecting lung nodule in CT scan images using lightweight CNN MobileNet &SqueezeNet	Reduces the computational memory cost of hardware	*Using traditional hardware
Kumar et al. [178]	2021	The study proposed a light weight "MobiHisNet," deployed on Raspberry pi2 for histopathological image classification	Experiments on breast cancer MobiHisNet on edge devices shown that a higher accuracy, lesser complexity, and lesser memory requirements	*Using histopathological images
Saddam Bekhet et al. [179]	2021	The proposed light weight CNN that could run smoothly on a normal Central Process Unit CPU (0.54% of AlexNet parameters) for Covid19 detection in CT images	96% accuracy.	*Using traditional hardware
Guo et al. [180]	2021	Proposed a light weight CNN (SequeezeNet & shufflenet) models compared with ResNet 18 for thyroid tissue recognition in Single Photon Emission Computed Tomography SPECT images	The accuracy and sensitivity rates are 96.69% and 94.75%, the specificity and precision rates are 99.6% and 99.96%, respectively, and there is no significant difference compared with other models. ($P > 0.05$).	Limited dataset. the residual thyroid tissue was less finely classified, using traditional hardware
Jha et al. [181]	2021	The study proposed a light weight CNN named NanoNet, used architecture for the partition of video capsule colonoscopy and endoscopy images using a dataset containing endoscopy biopsies and surgical instruments	Experiments demonstrate the performance of model is increased rendition of architecture in terms of a trade-off between model complication, rapidity, model parameters, and metric renditions resulting a tiny size with 36000 parameters compared with traditional model that having millions parameters	Using traditional hardware
Gunraj et al. [182]	2022	Introduce a novel lightweight neural network architecture called COVID-Net CT S, which is smaller & faster than COVID-Net CT.	Accuracy of 99.0%, COVID-19 sensitivity99.1%, positive predictive Valu98.0%, specificity99.4%, and negative predictive value of 99.7%.	The model is not suitable for downstream tasks ,using traditional hardware
Mehrrotraa et al. [183]	2022	Proposed a light weight CNN and ML algorithm for tuberculosis & Covid19 detection in x-ray chest images.	AUC of 0.94 and 1 with accuracy of 87.90% and 99.10% respectively in categorizing Tuberculosis TB infected images from Normal and COVID images.	Limited dataset, manual feature extraction & cross population test techniques are needed, not suitable for real time application, using traditional hardware
Tsivgoulis et al. [184]	2022	Proposed modified light weight SequeezeNet v1, v2 and compared with SequeezeNet for lung cancer detection in 2D&3D CT images	For 2D images dataset, accuracy are 93% & 94% for SequeezeNetv1 & SequeezeNetv2 respectively and for 3D dataset the accuracy are 94%	*Using traditional hardware

			&96%, the proposed models is 1.3-1.5 faster time and 0.8-2.5 better performance than SqueezeNet	
Ukwandu et al. [185]	2022	Proposed lightweight deep learning procedure for the prediction of COVID-19 using the MobileNetV2 model in CT images	MobileNet model delivers a high efficiency and competitive accuracy with lesser computational cost	*Using traditional hardware
Al-Naqeeb & Al-Shamma [73]	2022	This work developed a lightweight CNN model, identified as DuaNet	99.87% of accuracy	*Using traditional hardware
Awasthi et al. [186]	2022	Proposed Left Ventricle Network LVnet for left ventricle segmentation and it was compared with other models, like MiniNetV2, UNet and Fully Convolutional Dense Dilated Network (FCdDN).in Ultra Sound US cardiac images	Show improvement in segmentation performance high as 5% &18.5% with & without papillary musckes, method requires only 5% of the memory by a UNet model.	Post processing quality is limited, the method not suitable for real time application, using traditional hardware
Heidari, et al. [187]	2022	Proposed light weight DL based block chain CNN with TR to reduce the layers of CNN in different dataset using python programming	Technique outperform F1 (2.9%), precision (2.7%), recall (3.1%), and accuracy (2.8%)	Using traditional hardware
Arvind et al. [188]	2023	Proposed modified light weight Unet with multiple dropouts in deconvolution layers for lung cancer detection using x-ray chest images	Accuracy of 92.71%	*Using traditional hardware, insufficient resolution of small size organ images, limited optimization & equalization, limited training parameters &cost functions
Hao et al. [189]	2023	The study proposes a lightweight model, named Global Spatial Context Enhanced U-Net GSCEU-Net, using modified convolution module Shifted Convolution SConv with MLP and GSC module for encoding and Efficient Channel Attention ECA for decoding, the proposed model used to detect skin lesion using microscopic images	Compared to U-Net, the proposed model, reducing the computational complexity by 170 times & the parameter count by 190 times	Model's training convergence low rapidity, the floating-point operations have not touched the fastest level, using traditional hardware,
Wang et al. [190]	2023	Proposed light weight SequeezeNet model with SVM classifier for breast cancer detection using mammography images.	Sensitivity of 94.30%.and accuracy of 94.10% a	Data pattern was not captured effectively. fine-tuning of the model's hyperparameters was not pursued, the model is noisy, using traditional hardware
Hou & Navarro-Cía[191]	2023	Proposed light weight Efficient Net of limited parameters with advanced normalization tools for Covid19 detection in CT images	Classification accuracy increases from 91.15% to 95.50% and (AUC) from 96.40% to 98.54%.	Limited dataset low quality, dataset was not cleaned, limited similarity metrics, the model need to fine tuning, using traditional hardware
Liu & Li [192]	2023	The study tested several popular CNN architectures by using biopsy samples images for lung cancer detection	Accuracy range (0. 8808–0. 9121).	*Using traditional hardware *limited training dataset *biopsy samples

				dataset collection more cost
Raza et al. [193]	2023	Proposed light weight EfficientNetB1 model with modified classifier layer for lung cancer detection in CT images.	Accuracy of 99.1%	*Using traditional hardware *limited training dataset *comparative study was achieved by the same type of net
Mothkur et al. [194]	2023	Proposed light weight SequeezeNET & MobileNet for lung cancer detection in CT images	Accuracy of 85.21%	*Using traditional hardware *need to fine tuning the lightweight NN *need to various segmentation techniques
Roy & Satija [195]	2023	The study proposed lightweight inception network, namely, Respiratory Disease Lightweight Inception network (RDLINet). To classify a wide spectrum of respiratory diseases using lung sound signals	Accuracy of 96.6%,	*Using traditional hardware
Islam et al. [196]	2023	Study proposed the different CNN architecture VGG19,ResNet50,& light weight MobilNetV2 for lung cancer detection using chest CT images	MobileNetV2 provided the highest level of accuracy with less overfitting compared to other models	*Using traditional hardware
Al-Ofary & IIhan [197]	2023	Proposed to use AlexNet with light weight SequeezeNet & shuffle Net with 2 different classifier softmax and SVM for lung & colon cancer in CT images	The ShuffleNet occurred the better accuracy of 99.93%	*Using traditional hardware
Xiao et al. [75]	2023	Proposed light weight fast NN called FastNet using weight accumulation for tumor identification in mobile assisted device using histopathological images	The cost was reduced, efficiency was improved the highest accuracy of 97.34%,	*Using histopathological images
Biswas & Barma[198]	2023	Proposed light weight MicroMobileNet employed on mobile device for cancer detection in histopathological images	Accuracy upto 98.43%, the new network has been implemented on an edge device with high speed (140ms) and very low memory (7.4 MB).	*Using histopathological images
Awan et al. [199]	2023	Proposed light weight NN called EfficientB5 with fine tuning for lung cancer detection using chest x ray images	Recall (99.5%) for biclassification. remarkable accuracy (99.5%) and	The algorithm presented was innovated to detect another lung diseases, disease severity levels was not identified, using traditional hardware
Sait & Rahaman [200]	2023	proposed light weight MobileNetv3 for lung cancer detection with A Optimization AO algorithm of fine tuning for reducing training time, the study used DenseNet 121 for feature extraction	Kappa value of 95.8 with less elements & parameters, accuracy of 98.6%	Need to liquid neural networks and ensemble learning techniques, using traditional hardware
Asif et al. [201]	2023	The study proposed Lightweight Stacked Ensemble model, known as LWSE by combining MobileNet & light weight CNN with MLP classifier for different infection chest diseases using CT & x ray images	An outstanding accuracy of 98.83% on the CT dataset and an accuracy of 96.40% and 97.89% on the CXR dataset, low computational cost, faster performance than other models	Using traditional hardware

Hadi et al. [202]	2023	Proposed CORONA Net light weight NN included CNN for features extraction, and Digital Wavelet Transform DWT for features reduction and LSTM for Covid19 classification in chest x ray images	The proposed method achieves a high performance in comparison with the existing deep learning methods	Using traditional hardware
Raiaan et al. [203]	2023	Proposed a light weight shallow CNN called ResNet 10 with 3 blocks of convolution layers with 64 batch size and 0.0001 learning rate to classify fundus images in 5 classes with 3 augmentation techniques	MobileNetV2, VGG16, Xception, VGG19, InceptionV3 and ResNet50 achieved testing accuracies of 91.42%, 90.16%, 89.57%, 88.21%, 87.68% and 87.23%, respectively. Proposed RetNet-10 model performed the best, with a testing accuracy of 98.65%.	Using traditional hardware
Lang et al. [204]	2023	This paper, proposed a Lightweight Contextual and Channel Fusion network (LCCF-Net) for medical segmentation, this net included different blocks for feature decoding to reduce parameters features to reduce parameters	Results show that the proposed method better than other state-of-the-art methods for kidney tumor recognition, retinal vessel detection and COVID 19 segmentation	Using traditional hardware
Hareesh & Bellamkonda[205]	2023	Proposed using different DL techniques such CNN VGG16 and light weight MobileNet model for lung cancer detection in CT images, using 8 batch size,0,01 learning rate &50 epochs	Achieved a peak accuracy of CNN is 95.15%. accuracy of VGG-16 is95.88%. Furthermore, MobileNet demonstrated exceptional performance with an accuracy of 98.39%	Using traditional hardware
Wanasinghe et al. [206]	2024	The study proposed light weight CNN to extract features from Melodic Mel spectrogram, Mel frequency cepestral coefficients &chromogram to classify lung disease through lung sound detection using lung sound recording	The highest accuracy achieved in the developed classification is 91.04% for 10 classes.	Need to augmentation &preprocessing techniques, using traditional hardware Need to quantization technique, need to validation model.
Nahiduzzaman, et al. [207]	2024	Proposed a Lightweight Parallel Depth-wise separable Convolutional Neural Network (LPDCNN) for features extraction and computational cost reduction also the method used a Ridge Regression Extreme Learning Machine (Ridge-ELM) for precise classification of three lung cancer types using CT images.,	The framework offers exclusionary efficacy, with a time of testing just 0.003s in binary categorization, outstanding results are obtained with average recall and accuracy values of 98.25 ± 1.031 % and 98.40 ± 0.822 %, respectively, recall and accuracy values of 99.70 ± 0.671 % and 99.70 ± 0.447 %%, respectively, for four-class categorization.	Using traditional hardware

7. Results and discussion

The previous studies enumerated in Table 3 showed that the most common limitations are limited training dataset and the use of traditional hardware (super computers). Most studies have applied lightweight CNN as part of Tiny DL techniques, but they still use the traditional hardware or super computer. The computational time may be reduced, but the traditional supercomputers need a large space memory to process data as fast as possible, and large computer memory requires high cost. The traditional hardware can be considered a big challenge for the physicians who work in rural areas. The traditional hardware used in ML and Tiny ML techniques requires a high-speed processor, which means high cost is needed. This type of computer also consumes a large amount of power and energy. Moreover, most of the datasets that have been used in studies were collected from CT devices or PET/CT and a few literature used X-ray data set. Medical CT images show physicians more details about the location and size of lung tumors. These CT medical data are more beneficial for diagnosis. However, patients are exposed to more radiation, and this type of examination requires high cost. A Kumar et al. [178] and S Biswas and S Barma [198] used Tiny DL and applied lightweight models on a mobile device and a single-board computer, which are considered limited constrained devices. They used histopathological and microscopic data that can be considered complicated

data, requiring a long time and high cost for collection. This type of data is not always available compared with X-ray chest image data. Then, using X-ray or chest X-ray image data is preferable, Using X-ray data images as the first medical imaging choice is possible for doctors with limited experience. X-ray images may provide doctors rapid primary diagnosis if they use an efficient Tiny DL model. Even although less details are shown in X-ray images, using these data in efficient Tiny DL models could avoid the high cost of CT exam and receiving a high dose of radiation in further examinations.

8. Conclusion

This review concludes the potential challenges and opportunities in employing TDL and Tiny ML for the identification and classification of lung cancer. Despite significant advances, most existing studies still rely heavily on traditional DL models that require extensive computational resources, such as large memory and high-powered hardware, which are costly and energy consuming. This poses a substantial difficulty for physicians working in resource-constrained environments such as rural medical centers.

This review underscores the importance of designing lightweight, efficient models such as compressed CNNs that can operate effectively on limited-resource devices like mobile phones and single-board computers. By relating lightweight CNNs with limited devices, offering physicians in remote areas a portable, low-cost, and accurate diagnostic tool for lung cancer detection is possible. This methodology could significantly reduce reliance on expert physicians and expensive, high-powered computers. Moreover, such solutions can support rapid decision-making, which is crucial for the early diagnosis and treatment of lung cancer.

Future work in this area may help achieve accuracy levels comparable to those of traditional DL models while minimizing computational costs and resource demands, thus democratizing access to advanced diagnostic technologies.

References

[1] R. Sharma, "Mapping of global, regional and national incidence, mortality and mortality-to-incidence ratio of lung cancer in 2020 and 2050," *International Journal of Clinical Oncology,* vol. 27, no. 4, pp. 665-675, 2022.

[2] W. H. O. C. I. W. H. O. https://www.who.int/news-room/fact-sheets/detail/cancer. (Accessed 22 Jan 2024).

[3] A. J. Alberg and J. M. Samet, "Epidemiology of lung cancer," *Chest,* vol. 123, no. 1, pp. 21S-49S, 2003.

[4] U. S. E. P. A. I. E. Division, *A Citizen's guide to radon: the guide to protecting yourself and your family from radon*. US Environmental Protection Agency, Indoor Environments Division, 2002.

[5] M. Šutić *et al.*, "Diagnostic, predictive, and prognostic biomarkers in non-small cell lung cancer (NSCLC) management," *Journal of personalized medicine,* vol. 11, no. 11, p. 1102, 2021.

[6] S. H. Song, C. W. Ha, C. Kim, and G. M. Seong, "Complete spontaneous remission of small cell lung cancer in the absence of specific treatment: A case report," *Thoracic Cancer,* vol. 12, no. 19, pp. 2611-2613, 2021.

[7] R. Advani *et al.*, "Phase I and pharmacokinetic study of BMS-188797, a new taxane analog, administered on a weekly schedule in patients with advanced malignancies," *Clinical cancer research,* vol. 9, no. 14, pp. 5187-5194, 2003.

[8] E. S. Kim *et al.*, "The BATTLE trial: personalizing therapy for lung cancer," *Cancer discovery,* vol. 1, no. 1, pp. 44-53, 2011.

[9] M. Ladanyi and W. Pao, "Lung adenocarcinoma: guiding EGFR-targeted therapy and beyond," *Modern pathology,* vol. 21, no. 2, pp. S16-S22, 2008.

[10] A. Pallis *et al.*, "Targeted therapies in the treatment of advanced/metastatic NSCLC," *European journal of cancer,* vol. 45, no. 14, pp. 2473-2487, 2009.

[11] S. Walters *et al.*, "Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004–2007," *Thorax,* vol. 68, no. 6, pp. 551-564, 2013.

[12] K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, and N. Waddell, "Deep learning in cancer diagnosis, prognosis and treatment selection," *Genome Medicine,* vol. 13, pp. 1-17, 2021.

[13] S. Benzekry, "Artificial intelligence and mechanistic modeling for clinical decision making in oncology," *Clinical Pharmacology & Therapeutics,* vol. 108, no. 3, pp. 471-486, 2020.

[14] P. N. Srinivasu, J. G. SivaSai, M. F. Ijaz, A. K. Bhoi, W. Kim, and J. J. Kang, "Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM," *Sensors,* vol. 21, no. 8, p. 2852, 2021.

[15] I. El Naqa and M. J. Murphy, *What is machine learning?* Springer, 2015.

[16] M. M. e. al, "Adaptive Computation and Machine Learning," *MIT PRESS.Cambridge* 2018.

[17] U. Nations. "The Sustainable Development Goals Report." https://sdgs.un.org/goals (accessed 22/3, 2024).

[18] V. Janapa Reddi *et al.*, "Edge impulse: An mlops platform for tiny machine learning," *Proceedings of Machine Learning and Systems,* vol. 5, 2023.

[19] H. Du, *Deep Learning Techniques for Analyzing Clinical Lung Cancer Data*. Wake Forest University, 2019.

[20] M. A. Thanoon, M. A. Zulkifley, M. A. A. Mohd Zainuri, and S. R. Abdani, "A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images," *Diagnostics,* vol. 13, no. 16, p. 2617, 2023.

[21] X. Chen, L. Yao, T. Zhou, J. Dong, and Y. Zhang, "Momentum contrastive learning for few-shot COVID-19 diagnosis from chest CT

images," *Pattern recognition,* vol. 113, p. 107826, 2021.

[22] A. T. Sadiq and S. M. Abdullah, "Hybrid intelligent technique for text categorization," in *2012 international conference on advanced computer science applications and technologies (ACSAT)*, 2012: IEEE, pp. 238-245.

[23] D. Ezzat and H. A. Ella, "GSA-DenseNet121-COVID-19: a hybrid deep learning architecture for the diagnosis of COVID-19 disease based on gravitational search optimization algorithm," *arXiv preprint arXiv:2004.05084,* 2020.

[24] G. Savitha and P. Jidesh, "A holistic deep learning approach for identification and classification of sub-solid lung nodules in computed tomographic scans," *Computers & Electrical Engineering,* vol. 84, p. 106626, 2020.

[25] H. N. Abdullah and H. K. Abduljaleel, "Deep CNN based skin lesion image denoising and segmentation using active contour method," *Engineering and Technology Journal,* vol. 37, no. 11A, pp. 464-469, 2019.

[26] K. He *et al.*, "Synergistic learning of lung lobe segmentation and hierarchical multi-instance classification for automated severity assessment of COVID-19 in CT images," *Pattern recognition,* vol. 113, p. 107828, 2021.

[27] M. R. Regmi *et al.*, "An aggressive progression of a lung mass: a rare case of sarcomatoid carcinoma," *European Journal of Medical Case Reports,* vol. 4, no. 5, pp. 177-179, 2020.

[28] J. Biederer *et al.*, "MRI of the lung (3/3)—current applications and future perspectives," *Insights into imaging,* vol. 3, pp. 373-386, 2012.

[29] A. T. Abdulahi, R. O. Ogundokun, A. R. Adenike, M. A. Shah, and Y. K. Ahmed, "PulmoNet: a novel deep learning based pulmonary diseases detection model," *BMC Medical Imaging,* vol. 24, no. 1, p. 51, 2024.

[30] H. H. Abid and M. E. Abdulmunim, "Segmentation brain tumor and diagnosing using watershed algorithm," *Am. J. Eng. Res,* vol. 5, no. 11, pp. 31-35, 2016.

[31] S. T. Ahmed and S. M. Kadhem, "Using Machine Learning via Deep Learning Algorithms to Diagnose the Lung Disease Based on Chest Imaging: A Survey," *International Journal of Interactive Mobile Technologies,* vol. 15, no. 16, 2021.

[32] A. K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, A. Khanna, and J. J. Rodrigues, "Identifying pneumonia in chest X-rays: A deep learning approach," *Measurement,* vol. 145, pp. 511-518, 2019.

[33] A. S. Abdalrada, O. H. Yahya, A. H. M. Alaidi, N. A. Hussein, H. T. Alrikabi, and T. A.-Q. Al-Quraishi, "A predictive model for liver disease progression based on logistic regression algorithm," *Periodicals of Engineering and Natural Sciences,* vol. 7, no. 3, pp. 1255-1264, 2019.

[34] E. Hussain, M. Hasan, M. A. Rahman, I. Lee, T. Tamanna, and M. Z. Parvez, "CoroDet: A deep learning based classification for COVID-19 detection using chest X-ray images," *Chaos, Solitons & Fractals,* vol. 142, p. 110495, 2021.

[35] T. Rahman *et al.*, "Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray," *Applied Sciences,* vol. 10, no. 9, p. 3233, 2020.

[36] A. K. Das, S. Kalam, C. Kumar, and D. Sinha, "TLCoV-An automated Covid-19 screening model using Transfer Learning from chest X-ray images," *Chaos, Solitons & Fractals,* vol. 144, p. 110713, 2021.

[37] P. Rajesh, A. Murugan, B. Murugamantham, and S. Ganesh, "Lung cancer diagnosis and treatment using AI and Mobile applications," 2020.

[38] S. T. A. S. M. K, "Using Machine Learning via Deep Learning Algorithms to Diagnose the Lung Disease Based on Chest Imaging: A Survey'," *IJIM,* vol. 15, no. 16, 2021.

[39] K. Punithavathy, S. Poobal, and M. Ramya, "Performance evaluation of machine learning techniques in lung cancer classification from PET/CT images," *FME Transactions,* vol. 47, no. 3, pp. 418-423, 2019.

[40] A. Y. Saleh, C. K. Chin, V. Penshie, and H. R. H. Al-Absi, "Lung cancer medical images classification using hybrid CNN-SVM," *International Journal of Advances in Intelligent Informatics,* vol. 7, no. 2, pp. 151-162, 2021.

[41] J. Morgado *et al.*, "Machine learning and feature selection methods for egfr mutation status prediction in lung cancer," *Applied Sciences,* vol. 11, no. 7, p. 3273, 2021.

[42] T Zhang et al, ",' Simultaneous Identification of EGFR, KRAS, ERBB2, and TP53 Mutations in Patients with Non-Small Cell Lung Cancer by Machine Learning-Derived Three-Dimensional Radiomics' " *MDPI,* 2021, doi: . https://doi.org/10.3390/cancers13081814,2021.

[43] L. Hussain *et al.*, "Lung cancer prediction using robust machine learning and image enhancement methods on extracted gray-level co-occurrence matrix features," *Applied Sciences,* vol. 12, no. 13, p. 6517, 2022.

[44] J. Perumal, P. Lee, K. Dev, H. Q. Lim, U. Dinish, and M. Olivo, "Machine Learning Assisted Real-Time Label-Free SERS Diagnoses of Malignant Pleural Effusion due to Lung Cancer," *Biosensors,* vol. 12, no. 11, p. 940, 2022.

[45] S Ishii et al, "' Machine learning-based gene alteration prediction model for primary lung cancer using cytologic images," *', in Wiley Online Library (wileyonlinelibrary.com),* 2022.

[46] F. Carrillo-Perez, J. C. Morales, D. Castillo-Secilla, O. Gevaert, I. Rojas, and L. J. Herrera, "Machine-learning-based late fusion on multi-omics and multi-scale data for non-small-cell lung cancer diagnosis," *Journal of Personalized Medicine,* vol. 12, no. 4, p. 601, 2022.

[47] P Nancy et al, "' Optimized Feature Selection and Image Processing Based Machine Learning Technique for Lung Cancer Detection'," *International Journal of Electrical and Electronics Research (IJEER), 0822SI-IJEER-2022-05;,* 2022.

[48] H J Kwon et al, " Enhancing Lung Cancer Classification through Integration of Liquid Biopsy Multi-Omics Data with Machine Learning Techniques'," *MDPI Journal,,* ,2023, doi: ,. https:// doi.org/10.3390/cancers15184556.

[49] Q. Huang *et al.*, "Machine Learning System for Lung Neoplasms Distinguished Based on Scleral Data," *Diagnostics,* vol. 13, no. 4, p. 648, 2023.

[50] MS Kumar&KV Rao, " A Labelled Priority based Weighted Classifier for Feature Extraction for Accurate Lung Tumour Detection using Machine Learning Technique," *', International Journal of Intelligent Systems and Applications in Engineering,,* 2023.

[51] A Earnest et al, " Machine Learning Techniques to Predict Timeliness of Care among Lung Cancer Patients',", *Healthcare* ,2023, doi: ' https:// doi.org/10.3390/healthcare11202756.

[52] Kumar Mohan &Bhraguram Thayyile, "Machine learning techniques for lung cancer risk prediction for text dataset'," *',international journal of data informatics and intelligent computing',,* 2023.

[53] M Dirik, " Machine learning-based lung cancer diagnosis'." *Turkish Journal of Engineering,* 2023, doi: https://dergipark.org.tr/en/pub/tuje ,.

[54] M. S. Bhuiyan *et al.*, "Advancements in early detection of lung cancer in public health: a comprehensive study utilizing machine learning algorithms and predictive models," *Journal of Computer Science and Technology Studies,* vol. 6, no. 1, pp. 113-121, 2024.

[55] Y Xu et al, "' Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging," *Clinical cancer research,* June 1, 2019, doi: , doi: 10.1158/1078-0432.CCR-18-2495, Clin Cancer Res; 25(11).

[56] G Jakimovisky & D Davcev, "Using Double Convolution Neural Network for Lung Cancer Stage Detection'," *Appl. Sci.,* 2019, doi: doi:10.3390/app9030427,.

[57] H Park & C Monahan, "' Genetic Deep Learning for Lung Cancer Screening," *arXiv: ,* vol. v1, 27 Jul 2019.

[58] RR Subramanian et al', "' Lung Cancer Prediction Using Deep Learning Framework," *International Journal of Control and Automation,* vol., Vol. 13, no., No. 3,, pp. pp. 154-160, (2020).

[59] HF Al-Yasriy et al, " Diagnosis of Lung Cancer Based on CT Scans Using CNN',," in *2nd International Scientific Conference of Al-Ayen University (ISCAU-2020)*, 2020.

[60] A Elnakib et al, "Early Lung Cancer Detection Using Deep Learning Optimization'," *IJOE,* vol. Vol. 16,, no. No. 6, 2020.

[61] C. L. e. al';, "Using 2D CNN with Taguchi Parametric Optimization for Lung Cancer Recognition from CT Images'," *Appl. Sci.,* 2020, doi: doi:10.3390/app10072591,.

[62] A. K. e. al', "An Approach for Lung Cancer Detection using Deep Learning," *', International Research Journal of Engineering and Technology (IRJET),,* vol., Volume: 07, no. Issue: 09 | Sep 2020.

[63] - TA Amma et al, "' Lung Cancer Identification and Prediction Based on VGG Architecture'," *International Journal of Research in Engineering, Science and Management,* vol. Volume-3, no., Issue-7, , July-2020.

[64] X Zhan, . , "' A Convolutional Neural Network-Based Intelligent Medical System with Sensors for Assistive Diagnosis and Decision-Making in Non-Small Cell Lung Cancer'," *Sensors,* 2021,, doi: https://doi.org/ 10.3390/s21237996,.

[65] SHM Mohammed& ACinar', "' Lung cancer classification with Convolutional Neural Network Architectures," *Qubahan academic Journal,,* 2021, doi: https://doi.org/10.48161/qaj.v1n1a33,.

[66] A. A. Abd Al-Ameer, G. A. Hussien, and H. A. Al Ameri, "Lung cancer detection using image processing and deep learning," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 28, no. 2, pp. 987-993, 2022.

[67] Z Ren et al, 1614., "' A Hybrid Framework for Lung Cancer Classification',," *Electronics, ,* 2022, doi: https://doi.org/10.3390/ electronics11101614,.

[68] M Humayun et al, "' A Transfer Learning Approach with a Convolutional Neural Network for the Classification of Lung Carcinoma',," *Healthcare,* 2022,, doi: https:// doi.org/10.3390/healthcare10061058.

[69] Y. Said, A. A. Alsheikhy, T. Shawly, and H. Lahza, "Medical images segmentation for lung cancer diagnosis based on deep learning architectures," *Diagnostics,* vol. 13, no. 3, p. 546, 2023.

[70] S. Üzülmez and M. A. Çifçi, "Early diagnosis of lung cancer using deep learning and uncertainty measures," *Journal of the Faculty of Engineering and Architecture of Gazi University,* vol. 39, no. 1, pp. 385-400, 2024.

[71] N Vijayan & J Kuruvilla " The impact of transfer learning on lung cancer detection using various deep neural network architectures',," presented at the 19th India Council International Conference (INDICON),, 2022.

[72] M Agarwal et al, "An Efficient and Optimized Convolution Neural Network for Covid and Lung Disease Detection',," presented at the 8th International Conference on Communication and Electronics,, 2023.

[73] D. Al-Naqeeb and O. Al-Shamma, "DuaNet: A Novel Lightweight CNN Model for Classifying Five-class Lung Diseases," in *2022 International Conference on Data Science and Intelligent Computing (ICDSIC)*, 2022: IEEE, pp. 202-207.

[74] S. S. Skandha, M. Agarwal, K. Utkarsh, S. K. Gupta, V. K. Koppula, and J. S. Suri, "A novel genetic algorithm-based approach for compression and acceleration of deep learning convolution neural network: an application in computer tomography lung cancer data," *Neural Computing and Applications,* vol. 34, no. 23, pp. 20915-20937, 2022.

[75] P. Xiao *et al.*, "FastNet: A Lightweight Convolutional Neural Network for Tumors Fast Identification in Mobile Computer-Assisted Devices," *IEEE Internet of Things Journal,* 2023.

[76] S.A. R. ZAIDI et al, " Unlocking Edge Intelligence Through Tiny Machine Learning (TinyML)',," *Digital Object Identifier 10.1109/ACCESS.2022.3207200,,* 2022.

[77] e. a. R. David, " , , , "TensorFlow lite micro: Embedded machine learning for TinyML systems," in *in Proc. Mach. Learn. Syst*, 2021., vol. vol. 3, p. pp. 800811.

[78] L.CAPOGROSSO et al, " A Machine Learning-oriented Survey on Tiny Machine Learning'," *arXiv:IEEE,* vol.,v2, ,26 Sep 2023.

[79] P. Warden and D. Situnayake, "Tinyml: Machine learning with tensorflow lite on arduino and ultra-low-power microcontrollers.," *O'Reilly Media,,* 2019.

[80] H. S. N. - A. Pramod, and A. K. Tyagi, , "'Machine learning and deep learning: Open issues and future research directions for the next 10 years,'" *'' Computational analysis and deep learning for medical care: Principles, methods, and applications,,* pp. pp. 463–490, , 2021.

[81] e. a. N. C. Thompson, "The computational limits of deep learning,"," *arXiv preprint arXiv:2007.05558,,* 2020.

[82] L. Dutta and S. Bharali, "Tinyml meets iot: A comprehensive survey," *Internet of Things,* vol. 16, p. 100461, 2021.

[83] V. Rajapakse, I. Karunanayake, and N. Ahmed, "Intelligence at the extreme edge: A survey on reformable tinyml," *ACM Computing Surveys,* vol. 55, no. 13s, pp. 1-30, 2023.

[84] L. Capogrosso, F. Cunico, D. S. Cheng, F. Fummi, and M. Cristani, "A machine learning-oriented survey on tiny machine learning," *IEEE Access,* 2024.

[85] H. Ren, D. Anicic, and T. A. Runkler, "TinyReptile: TinyML with federated meta-learning," in *2023 International Joint Conference on Neural Networks (IJCNN)*, 2023: IEEE, pp. 1-9.

[86] e. a. - L. Heim, " "Measuring what really matters: Optimizing neural networks for tinyml,"," *arXiv preprint arXiv:2104.10645,* , 2021.

[87] J. Chang, Y. Choi, T. Lee, and J. Cho, "Reducing MAC operation in convolutional neural network with sign prediction," in *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, 2018: IEEE, pp. 177-182.

[88] M. Olyaiy, C. Ng, and M. Lis, "Accelerating DNNs inference with predictive layer fusion," in *Proceedings of the ACM International Conference on Supercomputing*, 2021, pp. 291-303.

[89] e. a. - W.-C. Lin, "'An efficient and low-power mlp accelerator architecture supporting structured pruning, sparse activations and asymmetric quantization for edge computing,," presented at the 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS), in 2021.

[90] e. a. Y.-C. Zhou, "'An enhanced data cache with in-cache processing units for convolutional neural network accelerators," " presented at the 15th International Conference on Solid- State & Integrated Circuit Technology (ICSICT). , in 2020.

[91] e. a. - F. Conti, "'Energy-efficient vision on the pulp platform for ultra-low power parallel computing,"," *Workshop on Signal Processing Systems (SiPS). IEEE,,* pp. pp. 1–6., in 2014

[92] A. Garofaloet al ", "Pulp-nn: accelerating quantized neural networks on parallel ultra-low-power risc-v processors,," *' Philosophical Transactions of the Royal Society A,* vol. vol. 378,, no. no. 2164, pp. ', , p. 20190155, 2020.

[93] P.-n. A Garofalo et al, "A computing library for quantized neural network inference at the edge on risc-v based parallel ultra low power clusters,'" presented at the 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS), in 2019.

[94] S. Prakash *et al.*, "CFU Playground: Full-Stack Open-Source Framework for Tiny Machine Learning (TinyML) Acceleration on FPGAs," in *2023 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 23-25 April 2023 2023, pp. 157-167, doi: 10.1109/ISPASS57527.2023.00024.

[95] C. Zhou *et al.*, "ML-HW Co-Design of Noise-Robust TinyML Models and Always-On Analog Compute-in-Memory Edge Accelerator," *IEEE Micro,* vol. 42, no. 6, pp. 76-87, 2022, doi: 10.1109/MM.2022.3198321.

[96] R. Immonen and T. Hämäläinen, "Tiny Machine Learning for Resource-Constrained Microcontrollers," *Journal of Sensors,* vol. 2022, p. 7437023, 2022/11/10 2022, doi: 10.1155/2022/7437023.

[97] e. a. Zhuang Liu, "Learning efficient convolutional networks through network slimming," in *In ICCV*, , 2017.

[98] H. Ren, D. Anicic, and T. A. Runkler, "TinyOL: TinyML with Online-Learning on Microcontrollers," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 18-22 July 2021 2021, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9533927.

[99] K. Dokic, M. Martinovic, and D. Mandusic, "Inference speed and quantisation of neural networks with TensorFlow Lite for Microcontrollers framework," in *2020 5th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, 25-27 Sept. 2020 2020, pp. 1-6, doi: 10.1109/SEEDA-CECNSM49515.2020.9221846.

[100] e. a. - Y. Zhang, "'Hello edge: keyword spotting on microcontrollers,," http://arxiv.org/abs/1711.07128.," 2017.

[101] "ARM Cortex-M,." Wikipedia. (accessed.

[102] M. Courbariaux, Y. Bengio, and J.-P. David, "Training deep neural networks with low

precision multiplications," *arXiv preprint arXiv:1412.7024,* 2014.
[103] C. Zhang, "How to run deep learning model on microcontroller with CMSIS-NN (part 3)„" *https://www.dlology.com/ blog/ how-to-run-deep-learning-model-on-microcontrollerwith-cmsis-nn-part-3/*. "" 2018.
[104] D. Lin, S. Talathi, and S. Annapureddy, "Fixed Point Quantization of Deep Convolutional Networks," presented at the Proceedings of The 33rd International Conference on Machine Learning, Proceedings of Machine Learning Research, 2016. [Online]. Available: https://proceedings.mlr.press/v48/linb16.html.
[105] e. a. - S. Zhuo, , , ""An empirical study of low precision quantization for TinyML," *http://arxiv.org/abs/2203.05492.,* " 2022,.
[106] A. Capotondi, M. Rusci, M. Fariselli, and L. Benini, "CMix-NN: Mixed Low-Precision CNN Library for Memory-Constrained Edge Devices," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 67, no. 5, pp. 871-875, 2020, doi: 10.1109/TCSII.2020.2983648.
[107] H. A. Rashid, P. R. Ovi, C. Busart, A. Gangopadhyay, and T. Mohsenin, "TinyM(2)Net: A Flexible System Algorithm Co-designed Multimodal Learning Framework for Tiny Devices," (in eng), *ArXiv,* Feb 9 2022.
[108] L. Mocerino and A. Calimera, "Fast and Accurate Inference on Microcontrollers With Boosted Cooperative Convolutional Neural Networks (BC-Net)," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 68, no. 1, pp. 77-88, 2021, doi: 10.1109/TCSI.2020.3039116.
[109] e. a. M. Courbariaux, ""Binarized neural networks: training deep neural networks with weights and activations constrained to +1 or -1„" *http://arxiv.org/abs/1602.02830.,* " 2016,.
[110] e. a. B. McDanel, ""Embedded binarized neural networks,"" *http://arxiv.org/abs/1709 .02260.,* 2017,.
[111] S. Anwar, K. Hwang, and W. Sung, "Structured Pruning of Deep Convolutional Neural Networks," *J. Emerg. Technol. Comput. Syst.,* vol. 13, no. 3, p. Article 32, 2017, doi: 10.1145/3005348.
[112] e. a. - I. Fedorov, ""Sparse: sparse architecture search for CNNs on resourceconstrained microcontrollers„" *" Advances in Neural Information Processing Systems,,* vol. vol. 32, , 2019.
[113] ". "TensorFlow model optimization,"." Google, . (accessed.
[114] J. Yu, A. Lukefahr, D. Palframan, G. Dasika, R. Das, and S. Mahlke, "Scalpel: Customizing dnn pruning to the underlying hardware parallelism," *ACM SIGARCH Computer Architecture News,* vol. 45, no. 2, pp. 548-560, 2017.
[115] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *arXiv preprint arXiv:1608.08710,* 2016.
[116] J. Yu, A. Lukefahr, D. Palframan, G. Dasika, R. Das, and S. Mahlke, "Scalpel: Customizing DNN Pruning to the Underlying Hardware Parallelism," presented at the Proceedings of the 44th Annual International Symposium on Computer Architecture, Toronto, ON, Canada, 2017. [Online]. Available: https://doi.org/10.1145/3079856.3080215.
[117] T. Z. S. Ye, K. Zhang et al. , ""A unified framework of DNN weight pruning and weight clustering/quantization using ADMM,"" *http://arxiv.org/abs/1811.01907.,* , 2018,.
[118] L. Meng and N. Suda ", "Optimizing Power and Performance for Machine Learning at the Edge: Model Deployment Overview,"" *ARM AI - AI Virtual Tech Talks Series,* pp. pp. 1–35„ 2020.
[119] N. Schizas, A. Karras, C. Karras, and S. Sioutas, "TinyML for Ultra-Low Power AI and Large Scale IoT Deployments: A Systematic Review," *Future Internet,* vol. 14, no. 12, p. 363, 2022. [Online]. Available: https://www.mdpi.com/1999-5903/14/12/363.
[120] Hari Kishan Kondaveeti et al, in *Advancement in Business Analytics Tools for Higher Financial Performance book,*, -,2023, ch. Lightweight Deep Learning: Introduction, Advancements, and Applications capter,.
[121] e. a. Martín Abadi "Tensorflow: A system for large-scale machine learning," *In OSDI, ,* 2016.
[122] e. a. Han Cai, " Once for All: Train One Network
and Specialize it for Efficient Deployment," *In ICLR,,* 2020.
[123] e. a. Han Cai, " ProxylessNAS: Direct Neural Architecture Search on Target Task
and Hardware," *In ICLR,,* 2019.
[124] e. a. f. Tianqi Chen, "An automated end-to-end optimizing compiler for deep learning," *In OSDI,,* 2018.
[125] e. a. Jungwook Choi, " Pact: Parameterized clipping activation for quantized neural networks," *arXiv preprint arXiv:1805.06085,,* 2018.
[126] Matthieu Courbariaux and Yoshua Bengio., "Binarynet: Training deep neural networks with weights and activations constrained to+ 1 or-1," *. arXiv preprint arXiv:1602.02830,* 2016.
[127] e. a. Yunchao Gong, "Compressing deep convolutional networks using vector quantization.," *arXiv preprint arXiv:1412.6115,* , 2014.
[128] e. a. Song Han, "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding.," *In ICLR,,* 2016.
[129] e. a. Song Han, ". Learning both Weights and Connections for Efficient Neural Networks.," *In NeurIPS,,* 2015.
[130] e. a. Kaiming He, "Deep Residual Learning for Image Recognition.," *In CVPR,,* 2016.
[131] e. a. Yihui He, "AMC: AutoML for Model Compression and Acceleration on Mobile Devices," *In ECCV, ,* 2018.
[132] Y. H. a. . ". Channel pruning for accelerating very deep neural networks," *In ICCV,,* 2017.

[133] Zhang X, et al. „ "Shufflenet: an Extremely Efficient Convolutional
Neural Network for Mobile Devices[J]," 2017.
[134] Y.-D. K. al, "Compression
of deep convolutional neural networks for fast and low power mobile applications,". *arXiv preprint arXiv:1511.06530,,* 2015.
[135] e. a. Liangzhen Lai, " Cmsis-nn: Efficient neural network kernels for arm
cortex-m cpus," *arXiv preprint arXiv:1801.06601,,* 2018.
[136] T. Lawrence and L. Zhang, "IoTNet: An Efficient and Accurate Convolutional Neural Network for IoT Devices," *Sensors,* vol. 19, no. 24, p. 5541, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/24/5541.
[137] e. a. Vadim Lebedev, " Speeding-up convolutional neural networks using fine-tuned cp-decomposition.," *arXiv preprint arXiv:1412.6553,,* 2014.
[138] E. L. a. N. D. Lane.. " Neural networks on microcontrollers: saving memory at inference via operator reordering," *arXiv preprint arXiv:1910.05110,,* 2019.
[139] e. a. Ji Lin, " Runtime neural pruning.," *In NeurIPS,* , 2017.
[140] e. a. - Haoxiao Liu, "DARTS: Differentiable Architecture Search.," *In ICLR,,* 2019.
[141] e. a. Zechun Liu, "MetaPruning: Meta Learning for Automatic Neural Network Channel Pruning," *In ICCV,,* . 2019.
[142] e. a. Ningning Ma, "ShuffleNet V2: Practical Guidelines for
Efficient CNN Architecture Design.," in *In ECCV*, 2018.
[143] e. a. Ilija Radosavovic, " Designing network design spaces," in *arXiv preprint arXiv:2003.13678,*, 2020.
[144] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks," in *Computer Vision – ECCV 2016*, Cham, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., 2016// 2016: Springer International Publishing, pp. 525-542.
[145] e. a. Manuele Rusci, "Memory-driven mixed low precision quantization for enabling deep network inference on microcontrollers," *In MLSys, ,* 2020.
[146] e. a. Mark Sandler, " MobileNetV2: Inverted Residuals and Linear Bottlenecks.," in *In CVPR*, 2018.
[147] e. a. Mingxing Tan, " MnasNet: Platform-Aware Neural Architecture Search for Mobile," in *In CVPR,*, .. 2019.
[148] e. a. H. KuanWang, "Hardware-Aware Automated Quantization with Mixed Precision," in *In CVPR*, 2019.
[149] e. a. Bichen Wu, " FBNet: Hardware-Aware Efficient ConvNet Design via
Differentiable Neural Architecture Search," in *In CVPR,*, . 2019.
[150] e. a. Xiangyu Zhang, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices.," in *In CVPR,*, 2018.
[151] e. a. - Shuchang Zhou, "Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients," *arXiv preprint arXiv:1606.06160,,* . 2016.
[152] e. a. Chenzhuo Zhu, "Trained ternary quantization.," *arXiv preprint arXiv:1612.01064,,* 2016.
[153] B. Z. a. Q. V. Le., "Neural Architecture Search with Reinforcement Learning.," *In ICLR,,* 2017.
[154] e. a. Barret Zoph, "Learning Transferable Architectures for Scalable Image Recognition.," in *In CVPR,* , 2018.
[155] Y. Li, Z. Li, T. Zhang, P. Zhou, S. Feng, and K. Yin, "Design of a Novel Neural Network Compression Method for Tiny Machine Learning," presented at the Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering, Xiamen, China, 2022. [Online]. Available: https://doi.org/10.1145/3501409.3501526.
[156] W. L. e. al., "A review of deep neural network model
compression techniques for embedded applications[J]," *Journal of Beijing
Jiaotong University,,* 2017.
[157] L. Jin, W. Yang, S. Wang, Z. Cui, X. Chen, and L. Chen, "A hybrid pruning method for convolutional neural network compression," *Minicomput. Syst,* vol. 39, pp. 2596-2601, 2018.
[158] Su Loach.. , " Research and application of lightweight target detection algorithm
based on deep learning[D]," *South China University of Technology,* 2020.
[159] M. Dehghanian and M. S. M. Mosadegh, "Ternary Weighted Function and Beurling Ternary Banach Algebra l 1 w (S)," in *Abstract & Applied Analysis*, 2011.
[160] D. J., "Research on model compression and forward acceleration techniques for embedded deep neural networks [D]." *University of Science and Technology of China,,* 2018.
[161] e. a. Iandola FN, "Squeezenet: Alexnet-level Accuracy
with 50x Fewer Parameters and <0.5mb Model Size[J],", 2016.
[162] Chollet F., " Xception: Deep Learning with Depthwise Separable Convolutions[C]// " in *{ieee} Conference on Computer Vision and Pattern Recognition ({cvpr})*, 2017: : Institute of Electrical and Electronics Engineers Inc. .
[163] S. L. . "Research and application of lightweight target detection algorithm based on deep learning[D]," *South China University of Technology,,* 2020.
[164] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," presented at the Proceedings of the 32nd International Conference on Machine Learning, Proceedings

of Machine Learning Research, 2015. [Online]. Available: https://proceedings.mlr.press/v37/ioffe15.html.

[165] M. S. Diab and E. Rodriguez-Villegas, "Embedded machine learning using microcontrollers in wearable and ambulatory systems for health and care applications: A review," *IEEE Access,* vol. 10, pp. 98450-98474, 2022.

[166] P. P. Ray, "A review on TinyML: State-of-the-art and prospects," *Journal of King Saud University-Computer and Information Sciences,* vol. 34, no. 4, pp. 1595-1623, 2022.

[167] V. Tsoukas, E. Boumpa, G. Giannakas, and A. Kakarountas, "A review of machine learning and tinyml in healthcare," in *Proceedings of the 25th Pan-Hellenic Conference on Informatics*, 2021, pp. 69-73.

[168] C. Nicolas, B. Naila, and R.-C. Amar, "Tinyml smart sensor for energy saving in internet of things precision agriculture platform," in *2022 Thirteenth International Conference on Ubiquitous and Future Networks (ICUFN)*, 2022: IEEE, pp. 256-259.

[169] Y. Abadade, A. Temouden, H. Bamoumen, N. Benamar, Y. Chtouki, and A. S. Hafid, "A comprehensive survey on tinyml," *IEEE Access,* 2023.

[170] N. Nasrullah, J. Sang, M. S. Alam, M. Mateen, B. Cai, and H. Hu, "Automated lung nodule detection and classification using deep learning combined with multiple strategies," *Sensors,* vol. 19, no. 17, p. 3722, 2019.

[171] S. S. Sanagala, S. K. Gupta, V. K. Koppula, and M. Agarwal, "A fast and light weight deep convolution neural network model for cancer disease identification in human lung (s)," in *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, 2019: IEEE, pp. 1382-1387.

[172] F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, and D. Pfeiffer, "Efficient deep network architectures for fast chest X-ray tuberculosis screening and visualization," *Scientific reports,* vol. 9, no. 1, p. 6268, 2019.

[173] S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio, and S. K. Antani, "Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays," *Ieee Access,* vol. 8, pp. 115041-115050, 2020.

[174] S. B. Shuvo, S. N. Ali, S. I. Swapnil, T. Hasan, and M. I. H. Bhuiyan, "A lightweight cnn model for detecting respiratory diseases from lung auscultation sounds using emd-cwt-based hybrid scalogram," *IEEE Journal of Biomedical and Health Informatics,* vol. 25, no. 7, pp. 2595-2603, 2020.

[175] P. Sumari, S. J. Syed, and L. Abualigah, "A novel deep learning pipeline architecture based on CNN to detect Covid-19 in chest X-ray images," *Turkish Journal of Computer and Mathematics Education (TURCOMAT),* vol. 12, no. 6, pp. 2001-2011, 2021.

[176] M. F. M. Gouher, K. Ramanjaneyulu, C. M. Bhuma, M. B. Mohammad, and U. Lekhana, "A LIGHTWEIGHT CNN FOR LUNG NODULE DETECTION AND CLASSIFICATION FROM CHEST RADIOGRAPHS."

[177] V. V. K. Shukla, M. Tanmisha, R. Aluru, B. Nagisetti, and P. Tumuluru, "Lung nodule detection through ct scan images and dnn models," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 2021: IEEE, pp. 962-967.

[178] A. Kumar, A. Sharma, V. Bharti, A. K. Singh, S. K. Singh, and S. Saxena, "MobiHisNet: a lightweight CNN in mobile edge computing for histopathological image classification," *IEEE Internet of Things Journal,* vol. 8, no. 24, pp. 17778-17789, 2021.

[179] S. Bekhet, M. H. Alkinani, R. Tabares-Soto, and M. Hassaballah, "An Efficient Method for Covid-19 Detection Using Light Weight Convolutional Neural Network," *Computers, Materials & Continua,* vol. 69, no. 2, 2021.

[180] Y. Guo *et al.*, "Classification and diagnosis of residual thyroid tissue in SPECT images based on fine-tuning deep convolutional neural network," *Frontiers in Oncology,* vol. 11, p. 762643, 2021.

[181] D. Jha *et al.*, "Nanonet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy," in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 2021: IEEE, pp. 37-43.

[182] H. Gunraj, A. Sabri, D. Koff, and A. Wong, "Covid-net ct-2: Enhanced deep neural networks for detection of covid-19 from chest ct images through bigger, more diverse learning," *Frontiers in Medicine,* vol. 8, p. 729287, 2022.

[183] R. Mehrrotraa *et al.*, "Ensembling of efficient deep convolutional networks and machine learning algorithms for resource effective detection of tuberculosis using thoracic (chest) radiography," *IEEE Access,* vol. 10, pp. 85442-85458, 2022.

[184] M. Tsivgoulis, T. Papastergiou, and V. Megalooikonomou, "An improved SqueezeNet model for the diagnosis of lung cancer in CT scans," *Machine Learning with Applications,* vol. 10, p. 100399, 2022.

[185] O. Ukwandu, H. Hindy, and E. Ukwandu, "An evaluation of lightweight deep learning techniques in medical imaging for high precision COVID-19 diagnostics," *Healthcare Analytics,* vol. 2, p. 100096, 2022.

[186] N. Awasthi, L. Vermeer, L. S. Fixsen, R. G. Lopata, and J. P. Pluim, "LVNet: Lightweight model for left ventricle segmentation for short axis views in echocardiographic imaging," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control,* vol. 69, no. 6, pp. 2115-2128, 2022.

[187] A. Heidari, S. Toumaj, N. J. Navimipour, and M. Unal, "A privacy-aware method for COVID-

19 detection in chest CT images using lightweight deep conventional neural network and blockchain," *Computers in Biology and Medicine,* vol. 145, p. 105461, 2022.

[188] S. Arvind, J. V. Tembhurne, T. Diwan, and P. Sahare, "Improvised light weight deep CNN based U-Net for the semantic segmentation of lungs from chest X-rays," *Results in Engineering,* vol. 17, p. 100929, 2023.

[189] S. Hao *et al.*, "GSCEU-Net: An End-to-End Lightweight Skin Lesion Segmentation Model with Feature Fusion Based on U-Net Enhancements," *Information,* vol. 14, no. 9, p. 486, 2023.

[190] J. Wang, M. A. Khan, S. Wang, and Y. Zhang, "SNSVM: SqueezeNet-guided SVM for breast cancer diagnosis," *Computers, Materials & Continua,* vol. 76, no. 2, 2023.

[191] Y. Hou and M. Navarro-Cía, "A computationally-inexpensive strategy in CT image data augmentation for robust deep learning classification in the early stages of an outbreak," *Biomedical Physics & Engineering Express,* vol. 9, no. 5, p. 055003, 2023.

[192] L. Liu and C. Li, "Comparative study of deep learning models on the images of biopsy specimens for diagnosis of lung cancer treatment," *Journal of Radiation Research and Applied Sciences,* vol. 16, no. 2, p. 100555, 2023.

[193] R. Raza *et al.*, "Lung-EffNet: Lung cancer classification using EfficientNet from CT-scan images," *Engineering Applications of Artificial Intelligence,* vol. 126, p. 106902, 2023.

[194] R. Mothkur and B. Veerappa, "Classification of lung cancer using lightweight deep neural networks," *Procedia Computer Science,* vol. 218, pp. 1869-1877, 2023.

[195] A. Roy and U. Satija, "RDLINet: A Novel Lightweight Inception Network for Respiratory Disease Classification Using Lung Sounds," *IEEE Transactions on Instrumentation and Measurement,* 2023.

[196] MA Islam et al, "' Deep Convolutional Neural Networks and Transfer Learning Based Approach for Lung Cancer Detection from CT Scan Images'„" https://ieeexplore.ieee.org/xpl/conhome/10227140/proceeding,IEEE,, 2023.

[197] S. Al-Ofary and H. O. Ilhan, "Classification of PCA based Reduced Deep Features by SVM for Diagnosing Lung and Colon Cancer," in *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2023: IEEE, pp. 1-5.

[198] S. Biswas and S. Barma, "MicrosMobiNet: A Deep Lightweight Network With Hierarchical Feature Fusion Scheme for Microscopy Image Analysis in Mobile-Edge Computing," *IEEE Internet of Things Journal,* 2023.

[199] T. Awan, K. B. Khan, and A. Mannan, "A compact CNN model for automated detection of COVID-19 using thorax x-ray images," *Journal of Intelligent & Fuzzy Systems,* vol. 44, no. 5, pp. 7887-7907, 2023.

[200] A. R. W Sait, ', „„ , "' Lung Cancer Detection Model Using Deep Learning Technique," *Applied Sciences (2076-3417),* vol. Vol 13,, no. Issue 22,, p. p12510, 2023, doi: DOI 10.3390/app13221251.

[201] S. Asif, M. Zhao, F. Tang, and Y. Zhu, "LWSE: a lightweight stacked ensemble model for accurate detection of multiple chest infectious diseases including COVID-19," *Multimedia Tools and Applications,* pp. 1-37, 2023.

[202] M. U. Hadi, R. Qureshi, A. Ahmed, and N. Iftikhar, "A lightweight CORONA-NET for COVID-19 detection in X-ray images," *Expert Systems with Applications,* vol. 225, p. 120023, 2023.

[203] M. A. K. Raiaan *et al.*, "A lightweight robust deep learning model gained high accuracy in classifying a wide range of diabetic retinopathy images," *IEEE Access,* 2023.

[204] J. Lang and Y. Liu, "LCCF-Net: Lightweight contextual and channel fusion network for medical image segmentation," *Biomedical Signal Processing and Control,* vol. 86, p. 105134, 2023.

[205] P. Hareesh and S. Bellamkonda, "Deep Learning-Based Classification of Lung Cancer Lesions in CT Scans: Comparative Analysis of CNN, VGG-16, and MobileNet Models," in *International Conference on Image Processing and Capsule Networks*, 2023: Springer, pp. 373-387.

[206] T. Wanasinghe, S. Bandara, S. Madusanka, D. Meedeniya, M. Bandara, and I. de la Torre Díez, "Lung Sound Classification with Multi-Feature Integration Utilizing Lightweight CNN Model," *IEEE Access,* 2024.

[207] M. Nahiduzzaman, L. F. Abdulrazak, M. A. Ayari, A. Khandakar, and S. R. Islam, "A novel framework for lung cancer classification using lightweight convolutional neural networks and ridge extreme learning machine model with SHapley Additive exPlanations (SHAP)," *Expert Systems with Applications,* vol. 248, p. 123392, 2024.