LUNG DISEASE CLASSIFICATION FROM RESPIRATORY SOUNDS USING CNN-TRANSFORMER KNOWLEDGE DISTILLATION

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
ARTIFICIAL INTELLIGENCE

Submitted by

ALANKAR MAHAJAN

2K21/AFI/04

Under the supervision of

Dr. ARUNA BHAT

(Associate Professor)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

MAY, 2023

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Alankar Mahajan**, Roll No.– **2K21/AFI/04** student of M.Tech A.I (Department of Computer Science and Engineering), hereby declare that the project Dissertation titled **"Lung Disease Classification from Respiratory Sounds using CNN-Transformer Knowledge Distillation"** which is submitted by me to Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirements for the award of the degree of Master of Technology in Artificial Intelligence, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship, or other similar title or recognition.

Place: Delhi — Alankar Mahajan

Date: 30/05/2023 — (2K21/AFI/04)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled **"Lung Disease Classification from Respiratory Sounds using CNN-Transformer Knowledge Distillation"** which is submitted by Alankar Mahajan, Roll No.– 2K21/AFI/04, Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirements for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi **Dr. Aruna Bhat**

Date: 30/05/2023 **(Associate Professor)**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I am extremely grateful to my project guide **Dr. Aruna Bhat** for her continuous guidance and mentorship that she provided me during the project. She showed me the path to achieve my target by explaining all the tasks to be done and explaining to me the importance of this project as well as its industrial relevance. She was always ready to help me and clear my doubts regarding any hurdles in the project. Without her constant support and motivation, this project would not have been successful.

Place: Delhi — Alankar Mahajan

Date: 30/05/2023 — 2K21/AFI/04

ABSTRACT

Because of the availability of enormous volume of data and computing resources, deep neural networks have gained popularity as they find application in both industry and academia. For certain tasks where a large training dataset cannot be obtained, transfer learning has shown that complex models that are already trained on large datasets can be applied to specific tasks by fine-tuning. However, in the healthcare domain, there is always a shortage of publicly available data and resources. Also, large pre-trained models are complex and often have a higher memory requirement, making them difficult to deploy. To overcome this issue, Knowledge Distillation has been used widely in the healthcare domain. Knowledge Distillation has been successful in compressing large and complex models making them easier for deployment.

To explore about the applications of Knowledge Distillation in Healthcare domain, we conducted a Systematic Literature Review. We analyzed recent studies based on some research questions that we formulated. After our analysis, we found some research gaps.

In this study, we explore the potential of utilizing transformers in a limited data setting. We propose a framework based on Knowledge Distillation to train transformers for classifying lung diseases using respiratory sounds. Our proposed framework combines the attributes of both Convolutional Neural Network (CNN) and Transformers i.e., the translation equivariance and inductive biases of CNNs with the ability of transformers to handle long range dependencies. We have used Wavegram-Logmel-CNN as the teacher and Audio Spectrogram Transformer (AST) as the student model. The results show that our proposed framework improves the accuracy of the Transformer model. We also discuss the future scope for further improvements.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
CNN	Convolutional Neural Network
NLP	Natural Language Processing
KD	Knowledge Distillation
ViT	Vision Transformer
GAN	Generative Adversarial Network
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
GRU	Gated Rectified Unit
AST	Audio Spectrogram Transformer
RQ	Research Question
IC	Inclusion Criteria
EC	Exclusion Criteria

CHAPTER 1: INTRODUCTION

1.1 OVERVIEW

In this study, we have proposed a framework to train Transformer model for the task of audio classification of respiratory sounds. The training is done using limited data. Since transformers cannot generalize well on small data, we incorporate a teacher-student approach by using a CNN as a teacher to train our transformer model. This framework also combines the results shows that our approach improves the performance of the transformer model.

1.2 PROBLEM STATEMENT

Deep learning models have become the state-of-the-art in various domains such as computer vision, natural language processing (NLP), speech processing etc. With the availability of large amount of data and compute resources, large and complex deep learning models can be applied to solve different tasks. In certain situations where limited data is available, a model trained from scratch may not be the best approach as the model may fail to capture relationships among features and learn the data points instead. This can be solved using a model training technique called Transfer Learning. Using this, a complex model which is already trained on a large dataset to learn representations (feature extraction) from the data is used by slightly fine-tuning it on the target task. This enables complex models such as [1], [2], [3] and many more which have been trained on [4] to be used on problems where enough data cannot be gathered. This technique works because it has been found that shallower layers of complex models tend to learn general features like corners, edges of objects which may be generalized to a lot of different tasks, whereas the deeper layers focus on the finer details of the image [5].

There are still limitations to transfer learning, the major one being the domain sensitivity of the models. If the data distribution between the data on which pretraining is done and that of target task is significant, pretrained models perform poorly. In the healthcare domain, medical datasets have different distribution than ordinary images in ImageNet dataset. Also, there is not enough data available publicly that may come close to the scale of ImageNet data as medical institutions have to maintain the privacy of patients. It is also desired that models should have lower inference time (latency) and should not have a higher memory requirement as they are to be deployed on edge devices with low processing capabilities for real time applications. Thus, model compression method is required. Acceleration techniques such as pruning which removes redundant parameters and quantization which reduces the bits required to store parameters are incorporated in deep learning frameworks. These methods help in reducing the model to some extent. Another method that has become popular is knowledge distillation where a large model called the teacher generally supervises the training of a simpler model called the student. Knowledge can either be the probability distribution obtained from the final layer or features extracted from intermediate layers of the teacher. This knowledge is used to train the student model. The student model is either a less complex model that does not require high processing capability or is identical to the teacher model. Such model can be easily deployed on edge devices. This makes knowledge distillation a suitable model compression technique for healthcare applications.

1.3 OBJECTIVE

The motivation of our research comes from the identification of gaps that we found while conducting a systematic review of recent work that has been done in healthcare sector involving knowledge distillation (explained in Chapter 2). Since the scale of medical datasets is not the same as datasets like ImageNet, Transformers models fail to generalize well on such data. Also, CNNs can perform well on small datasets due to their equivariance to translation and in-built biases, we can use representations learned by CNNs in conjunction with training a transformer so that it may generalize well. Therefore, the CNN acts as teacher and Transformer becomes the student.

CHAPTER 2: RELATED WORK

In this section, we have explained the systematic review that we conducted on recent work that has incorporated knowledge distillation to train deep learning models. The systematic review is organized in the form of research questions to analyse the selected studies. Using our systematic review, we found some gaps which laid the basis of our research.

2.1 BACKGROUND

This section describes all the related terminologies that are need to understand the subsequent sections of our systematic review and the successive chapters of our thesis. This includes all the relevant information regarding deep learning, its different architectures, knowledge distillation and its types and various performance metrics.

2.1.1 Deep Learning

Deep Learning uses a class of machine learning algorithms called neural networks which consist of processing units called neurons which are interconnected to each other. These are modeled after the brain as the it consists of interconnected neurons through information is passed in the form of electric signals [6]. A Feedforward neural network consists of three layers: input, hidden and output layer. A neural network with a single hidden layer is called a shallow network. The network becomes deeper as the hidden layers keep on increasing. Neurons of one layer are interconnected to its preceding and succeeding layer (except for input and output layer). Each neuron receives an input and performs multiplication with a weight value assigned to it. To add non linearity to the calculations, these values are passed through an activation function. The weights of these neurons are the parameters of the model that are to be updated so that network can make predictions as accurate as possible. The weights are updated by back propagating the residuals.

Convolutional neural networks (CNN) are used to solve problems which involve images. Feedforward networks have all neurons interconnected thus spatial information of an image will not be preserved. A CNN uses convolution operation in which a kernel (also called filter) slides over the image. The output obtain after a convolution operation is called a feature map. It is passed as input to pooling layer that helps in reducing the dimensionality and extracting the important features. These features are flattened and passed to fully connected layers similar to the layers in feed forward neural networks to obtain predictions on the given task.

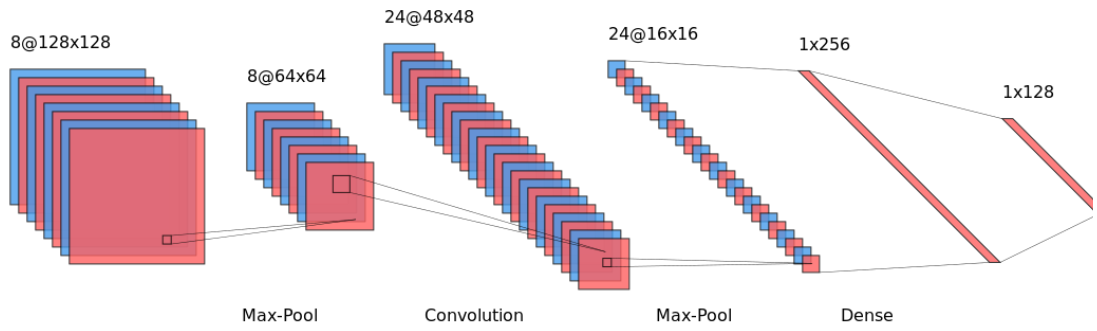


Figure 16 Architecture of a simple CNN.

Recurrent neural networks (RNN) are used for sequential data (time series or text data). rely on a similar kind of translational invariance as CNN. The input to an RNN cell is dependent on the output from its previous states also called the hidden states. These are not suitable for tasks involving long term dependencies because of the problem of vanishing gradients, where change in weights (model parameters) become so small that there is no new information learned by the model. LSTM (Long Short-Term Memory) solves this issue to some extent by dividing the hidden state into hidden and cell state. The hidden states help in choosing which features to use for each time stamp by using three gates namely forget gate, update gate and output gate. This increases the complexity and make LSTM computationally expensive. Gated Rectified Unit (GRU) reduces the complexity by using only two gates instead of three.

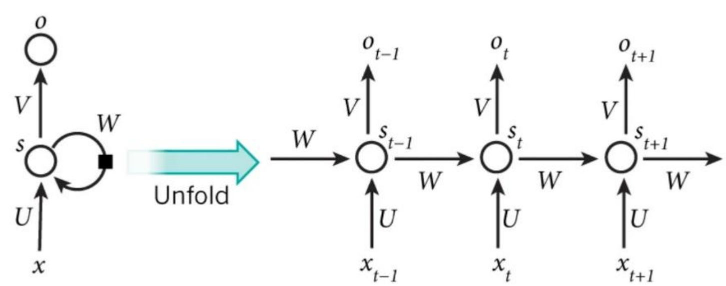


Figure 17 Architecture of RNN [6]. In case of multiple inputs at different timestamps, an unfolded RNN is obtained by connecting RNN cells in series which takes input x at timestamp t and hidden state s_{t-1} from the previous node to produce output o_t. The current hidden state s_t is passed to the next node.

GANs (Generative Adversarial Network) are generative models consisting of two networks called the Generator and Discriminator. Generator is tasked with learning distribution of data from training and generating data from random noise. Discriminator tries to distinguish whether the input given is real or is produced by the generator [7]. Both the models are trained in parallel.

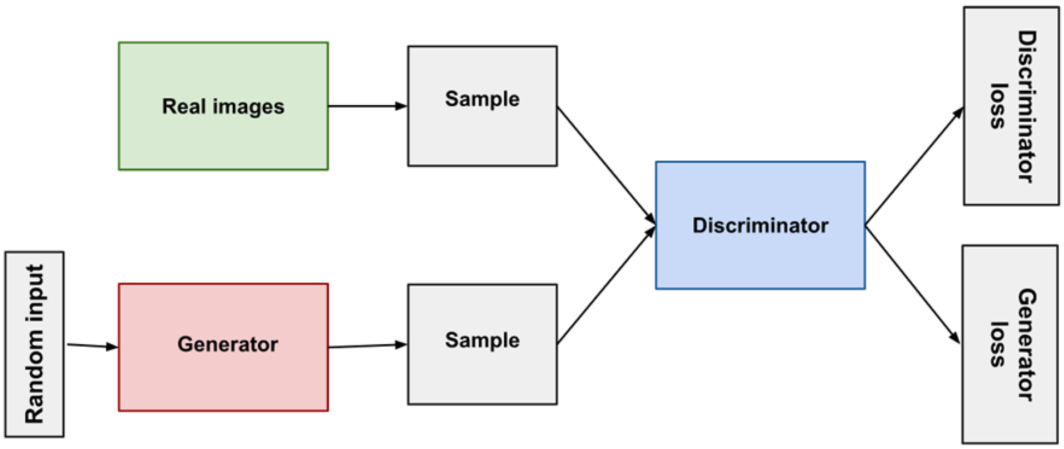


Figure 18 Architecture of GAN [107].

The concept of self-attention and Transformer model [8] is the reason for the major breakthroughs that have occurred in the field of NLP recently. Self-attention involves weighing significance of each word in a sentence by comparing each word in a sentence to every other word based on the context. The Transformer uses this self-attention mechanism by combining multiple self-attention modules where each module focuses on different part of the sentence. Inspired by this, [9] proposed Vision Transformer (ViT). ViT works by slicing the image into 16 x 16 patches and maps these patches into a one-dimensional vector which is passed to a Transformer encoder. Thus, vision transformer converts images to tokens which resembles converting sentences to tokens in case of NLP Transformer.

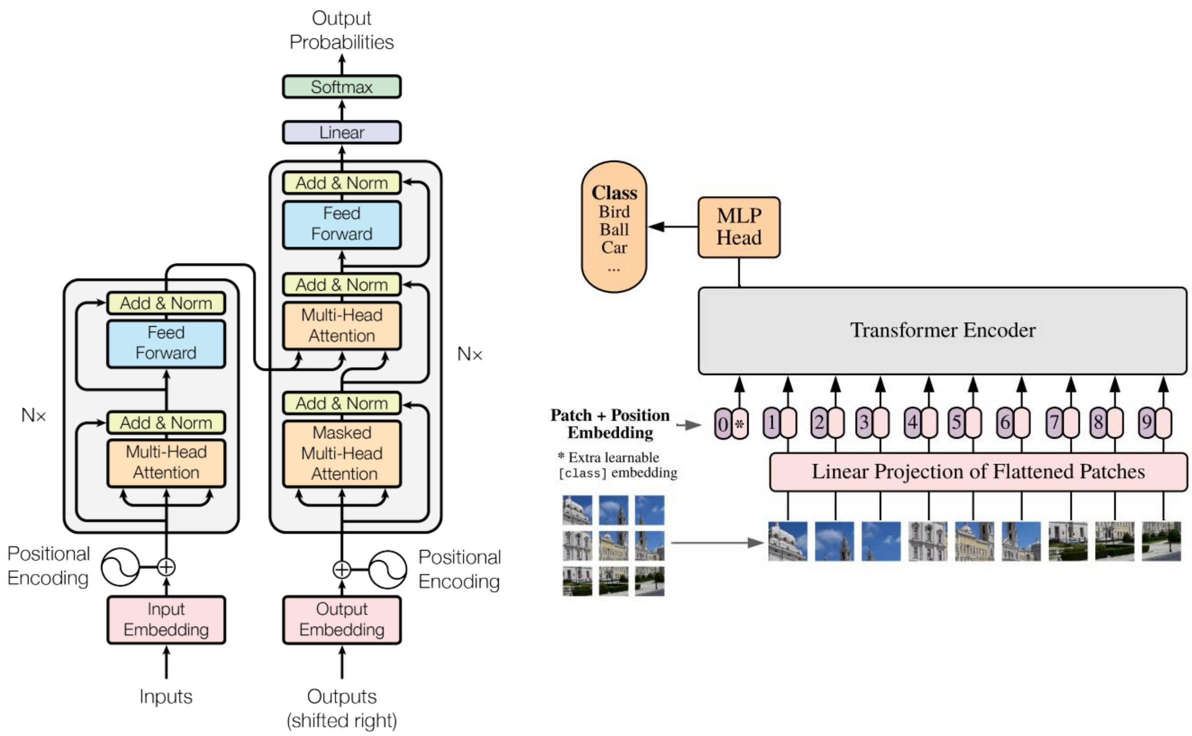


Figure 19 Architecture of NLP Transformer (Left) [8] and Vision Transformer (Right) [9].

2.1.2 Knowledge Distillation

Knowledge distillation is a model compression technique that is used to train a simple model called the student by transferring the representations learned by a complex model called the teacher.

The idea of model compression was given by [10] in which features from a complex ensemble model were compressed into a single model by pseudo labeling an unlabeled dataset and training the simple model on that dataset. Based on this, [11] generalized this idea by introducing Knowledge distillation in which the logits from a complex model were used as “soft labels” to train a simpler model (Figure 5).

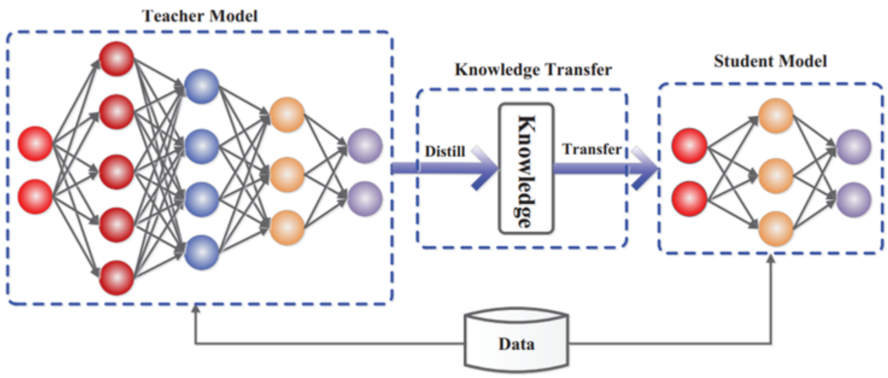


Figure 20 General Framework for Knowledge Distillation [106].

Based on the mechanism used for training (Figure 6), knowledge distillation can be classified as. Offline Distillation*:* Teacher model is either a pretrained model or a model fine-tuned on the target task. Knowledge from teacher is used to supervise the student. Online Distillation: Teacher and student are updated simultaneously for each epoch. Self-Distillation: Teacher and student model are same. Knowledge from deeper layers is used to train shallow layers.

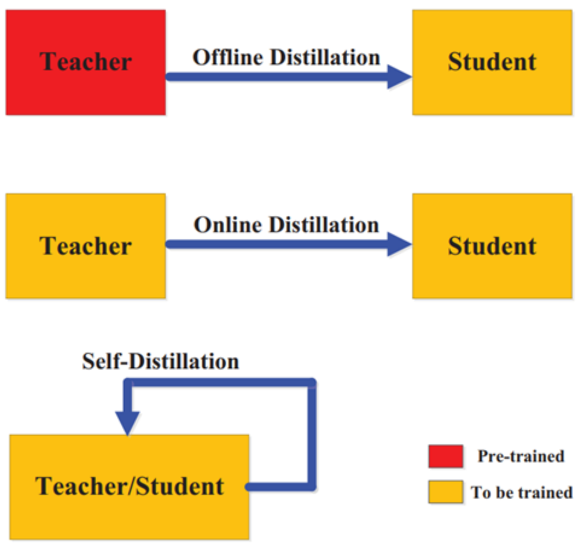


Figure 21 Types of knowledge distillation based on training mechanism [106].

Knowledge Distillation can also be classified based on the type of knowledge extracted from teacher to train student (Figure 7), it can be classified as: Response Based: Logits from final layer of the teacher act as labels to train the student. Feature Based: Instead of using logits, student learns from the feature maps of the teacher. Relation Based: Student learns by modelling relationship between feature maps of teacher. It is done using correlation analysis, similarity matrix, graph analysis of feature maps, etc.

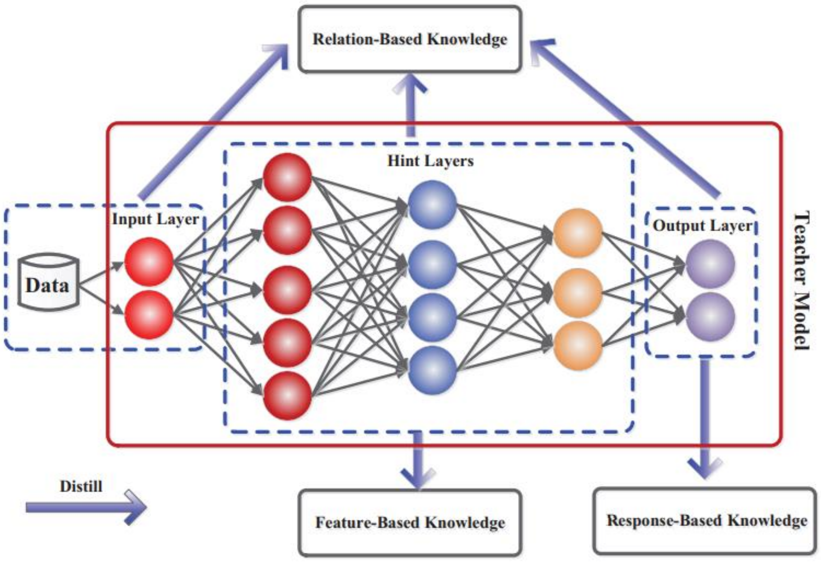


Figure 7 Different types of knowledge in teacher model and their sources [106].

2.1.3 Performance Metrics

- *Accuracy:*

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{1}$$

- *Precision:*

$$Precision = \frac{T_P}{T_P + F_P} \tag{2}$$

- *Sensitivity (Recall):*

$$Recall = \frac{T_P}{T_P + F_N} \tag{3}$$

- *Specificity:*

$$Specificity = \frac{T_N}{T_N + F_P} \tag{4}$$

where T_P = true positives (positive cases predicted as positives), T_N = true negatives (negative cases predicted as negative), F_P = false positives (negative cases predicted as positive), F_N = false negatives (positive cases predicted as negative).

- *F1-Score:*

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

- *AUC-ROC:* It is the area under the curve of ROC (Receiver Operating Characteristic) where TPR (Recall) is plotted against FPR (1-Specificity).

- *Kappa Score:* It is used to measure performance of a classification model against a model that randomly classifies an observation.

$$Kappa\ Score = 1 - \frac{1 - p_o}{1 - p_e} \tag{6}$$

where p_o = observed agreement, p_e = expected agreement.

- *Mean Squared Error:*

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{7}$$

- *Mean Absolute Error:*

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{8}$$

where N = total number of samples, y_i = actual value, $\hat{y}_i$ = predicted value.

- *Dice Similarity Coefficient:* Measure of similarity between two sets.

$$DSC = \frac{2\,|X \cap Y|}{|X| + |Y|} \tag{9}$$

- *IOU (Intersection Over Union):* Measures accuracy of predicted bounding box. It is also known as *Jaccard Correlation Coefficient*.

$$IOU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{10}$$

2.2 SYSTEMATIC LITERATURE REVIEW

This systematic literature review provides an overview of the recent studies which involve knowledge distillation using deep learning in the healthcare domain. We adopt the methodology of our study from [12]. First, we define the research questions for our SLR. Next, we define the search process to obtain relevant articles from academic databases. Then we apply the inclusion and exclusion criteria to filter out studies that are relevant to our SLR. These steps are explained in the following subsections.

2.2.1 Research Questions Formulation

The following research questions are constructed to meet objectives of our SLR.

RQ1. What are the different deep learning architectures being used as teacher and student?

RQ2. On what tasks are these models being applied to?

RQ3. What data preprocessing techniques are used?

RQ4. What datasets are being used to train the models?

RQ5. What metrics have been used to analyze the performance?

2.2.2 Search Method

The scientific databases that are used to find literature for our study are PubMed and ScienceDirect. These databases have been searched using keywords. The keywords used to create the search query are “knowledge distillation”, “deep learning”, “healthcare”. After searching the databases, we obtain 657 studies on which screening is applied. Before screening, duplicate records are removed. While screening, we filter out studies that are not relevant to our study based on our inclusion criteria and obtain eligible studies. We apply the exclusion criteria to obtain final set of 34 studies to analyze for our SLR. Figure 8 shows the process of literature selection.

2.2.3 Defining Inclusion and Exclusion Criteria

The studies meeting the inclusion criteria (IC) were included in our SLR. These are defined as follows:

- IC1: Studies match the search queries.
- IC2: Studies which have open access or can be accessed by an institutional proxy (account or network).

The studies meeting the exclusion criteria (EC) were not included in our SLR. These are defined as follows:

- EC1: Studies which do not train deep learning models using knowledge distillation.
- EC2: Studies which are not related to healthcare domain.
- EC3: Studies which are published before January 2019.

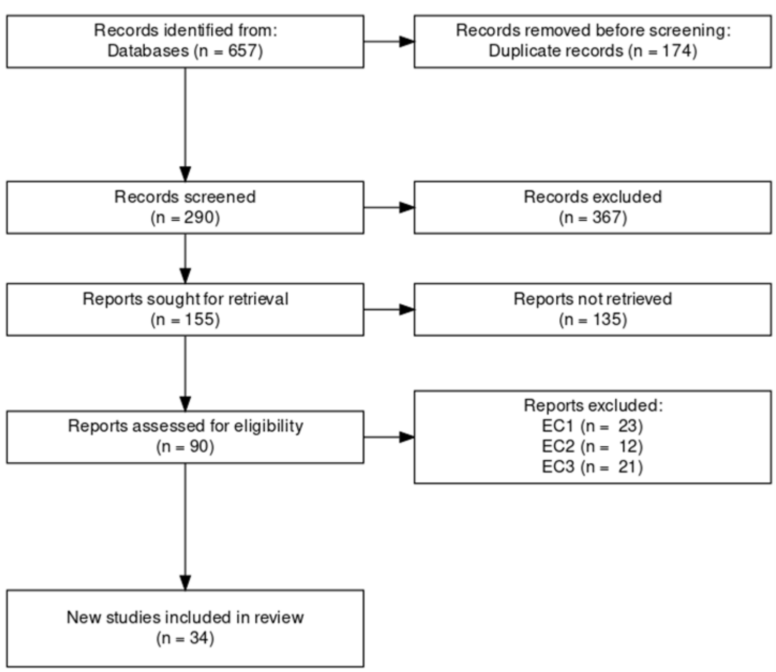


Figure 22 Flowchart depicting the procedure of literature selection.

2.2.4 Discussion

After searching the databases and applying the inclusion and exclusion criteria, a total of 34 research articles are selected. These studies, along with their Year of Publishing (YoP), publisher and type are listed in Table 1. For the rest of this section, we denote each study by their assigned ID as shown in Table 1.

Table 8 Studies selected for Systematic Literature Review

ID	Research Study	YoP	Publisher	Type
RS1	[13]	2021	Wiley	Journal
RS2	[14]	2021	IEEE	Journal
RS3	[15]	2020	ACM	Conference
RS4	[16]	2021	IEEE	Conference
RS5	[17]	2021	IEEE	Journal
RS6	[18]	2021	Elsevier	Journal
RS7	[19]	2021	ACM	Conference
RS8	[20]	2021	AMIA	Journal
RS9	[21]	2021	ACM	Conference
RS10	[22]	2020	IEEE	Conference
RS11	[23]	2020	IEEE	Conference
RS12	[24]	2021	IEEE	Journal
RS13	[25]	2021	Springer	Conference
RS14	[26]	2020	AAAI	Conference
RS15	[27]	2022	IEEE	Journal
RS16	[28]	2021	IEEE	Conference
RS17	[29]	2022	SPIE	Journal
RS18	[30]	2021	IOP Publishing	Journal
RS19	[31]	2021	Elsevier	Journal
RS20	[32]	2022	Springer	Journal
RS21	[33]	2022	Nature Portfolio	Journal
RS22	[34]	2019	IEEE	Journal
RS23	[35]	2021	Springer	Journal
RS24	[36]	2021	IEEE	Conference
RS25	[37]	2021	Elsevier	Journal
RS26	[38]	2022	IEEE	Conference
RS27	[39]	2021	IEEE	Conference
RS28	[40]	2022	Nature Portfolio	Journal
RS29	[41]	2022	Nature Portfolio	Journal
RS30	[42]	2022	Elsevier	Journal
RS31	[43]	2022	Elsevier	Journal
RS32	[44]	2022	Elsevier	Journal
RS33	[45]	2022	Elsevier	Journal
RS34	[46]	2022	Elsevier	Journal

A descriptive analysis of selected studies is presented. Based on figure 9, half of the stud
ies are published in the year 2021. Out of the 34 studies, 22 (65 percent) are published in journals and 12 (35 percent) in conferences. The maximum number of studies i.e., 12 are obtained from IEEE Xplore Digital Library out of which 7 are conference articles and 5 are journal articles.

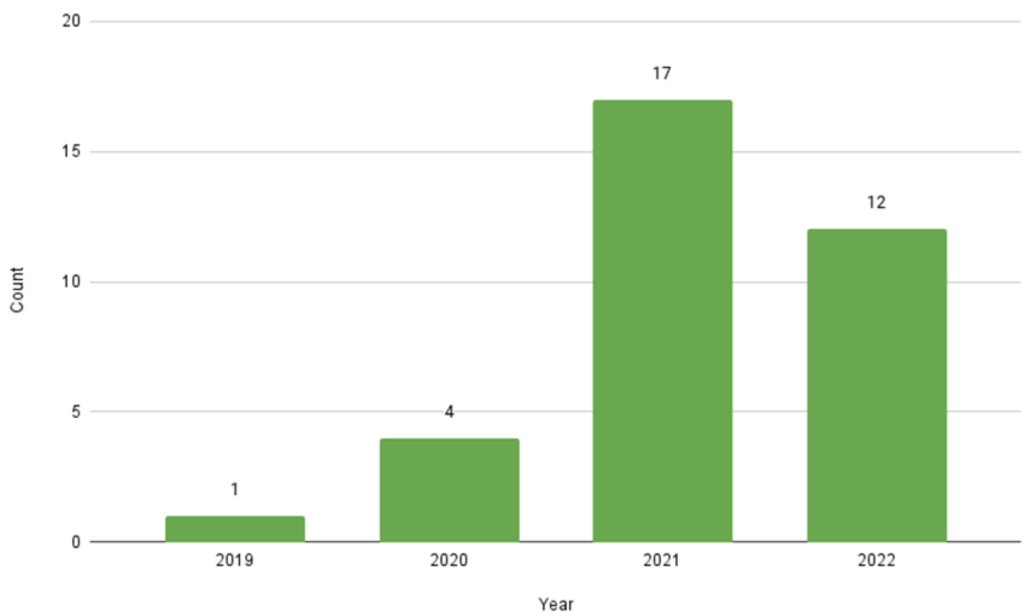


Figure 23 Selected studies grouped by year of publishing.

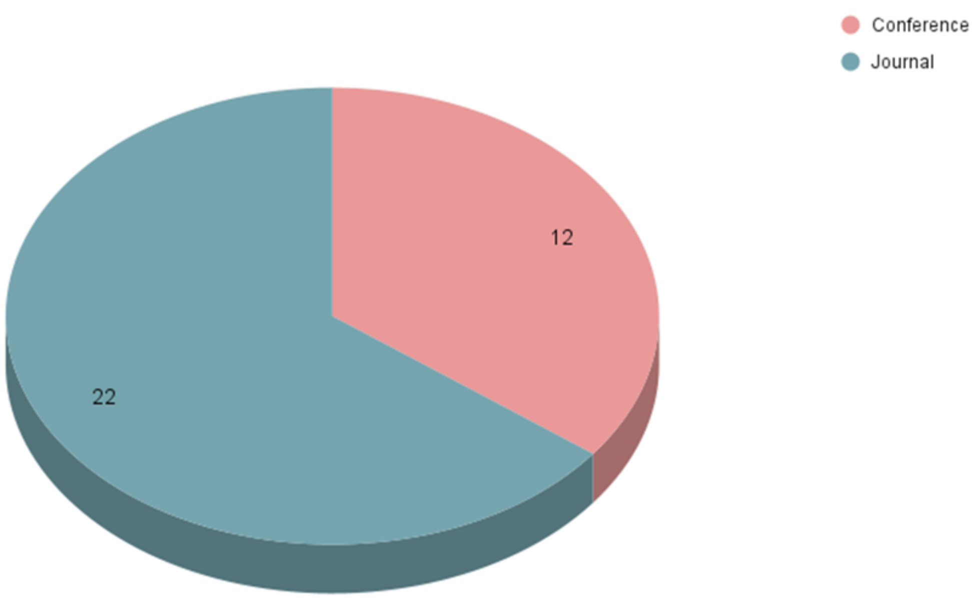


Figure 24 Type of articles included in our study.

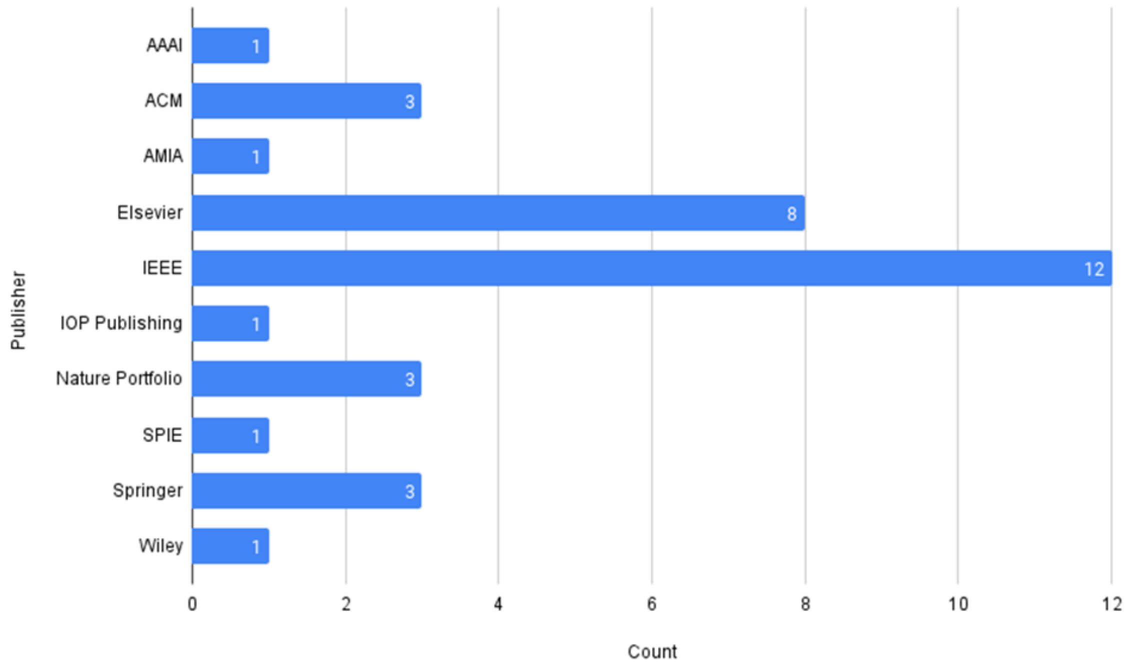


Figure 25 Count of selected studies by publisher.

2.2.5 Research Questions

RQ1: Network Architectures

All the teacher and student models are mentioned in Table 2. Majority of the studies are using identical architectures for both teacher and student. Study RS1, RS5, RS11 do not use any student model because they use self-distillation technique for training the model. Remaining studies use a custom CNN model or a pretrained student model that is lighter than their respective teacher model. This reduces the number of parameters.

Table 9 Teacher and Student Deep Learning models used in the studies

ID	Model (Teacher)	Model (Student)
RS1	[47] + CNN	None (Self Distillation)
RS2	[48]	[49]
RS3	ResNet-146	8-Layer CNN
RS4	[1] + LSTM (Trained on multiple tasks)	Same
RS5	[50]	None (Self Distillation)
RS6	[1] (different variants are trained)	Same

RS7	GRU with self-attention	Same
RS8	[51] pretrained on radiology corpus	Same
RS9	[1] (different models were trained on different subsets)	Same (knowledge of all subsets is distilled into a unified model)
RS10	Encoder-Decoder Model based on [52]	Same with reduced number of filters
RS11	Custom CNN	None (Self Distillation)
RS12	Ensemble of [47], [53], [54]	[55], [56]
RS13	Custom Graph CNN	Same
RS14	[57] + [58] with residual connections	Same
RS15	[59]	Custom model based on [58]
RS16	Ensemble of Custom CNN models	Custom CNN with decision tree regularization
RS17	Ensemble of [58], [1]	Custom CNN
RS18	[60]	[58]
RS19	[1] (Pretrained)	[1] (weights were initialized using He initialization)
RS20	[47] + LSTM	[47], [1] + LSTM
RS21	[9]	Same
RS22	[61], [2], [62] used for neighbour generation	Custom NN
RS23	[2]	Same (neurons reduced in the last layer)
RS24	3D CNN based on VGG7	Same
RS25	[2]	Same (filters reduced by a factor of 4), Custom CNN
RS26	[47]	[63]
RS27	Custom CNN + GRU	Same model with reduced depth of convolution layers
RS28	[47]	Same
RS29	[64]	Same model with less layers
RS30	[9] + [58]	[1] + [58]
RS31	[1]	Custom CNN

RS32	[65], [2], ResNet-121	[66], [52], [47], [3]
RS33	Custom GAN	Same GAN with low rank convolutions
RS34	[9], [1]	[67] (number of filters are halved)

RQ2: Tasks

The selected studies have been categorized based on application areas (tasks) in Table 3. The most popular application areas are Medical Image Classification and Medical Image Segmentation.

Table 10 Classification of selected studies based on the task

Task	**ID**
Medical Image Classification	RS1, RS3, RS4, RS5, RS6, RS11, RS12, RS13, RS19, RS20, RS21, RS22, RS23, RS24, RS25, RS26, RS28, RS31, RS32, RS34
Medical Image Segmentation	RS2, RS9, RS10, RS14, RS17, RS18, RS30
Sound Analysis	RS16
Radiology Protocol Selection	RS8
Medical Named Entity Recognition	RS29
Medical Image Synthesis	RS33
Medical Image Registration	RS15
Electronic Medical Record Analysis	RS7
ECG Analysis	RS27

RQ3: Preprocessing Methods

The different preprocessing methods used in the studies are listed in Table 4. We are not considering resizing image according the deep learning model as it is a necessary step and will be same for any deep learning pipeline.

Table 11 Preprocessing Techniques used in the selected studies

ID	Pre-processing Techniques
RS1	2D rotations by 90, 180 and 270 degrees
RS2	Random Rotation, Random Flip
RS3	Rotation, Zoom, Width Shift, Shear, Horizontal Flip
RS4	Normalization
RS5	Random Scaling, Random Rotation, Random Flip
RS6	Center Crop
RS7	
RS8	
RS9	
RS10	Random Rotation, Shear Translation, Random Flip
RS11	
RS12	Random Brightness, Random Contrast, Affine Transformation, Horizontal Flip, Vertical Flip
RS13	
RS14	Center Crop, Random Flip, Random Rotation
RS15	
RS16	Sample rate conversion, Duration trimming, Mel-spectrogram extraction
RS17	Random Rotation
RS18	
RS19	Horizontal Flip, Vertical Flip, Random Rotation
RS20	Geometry Correction, Intensity Correction
RS21	Histogram Equalization, Gaussian Blurring, Normalization
RS22	
RS23	
RS24	
RS25	Histogram Equalization, Random Flip
RS26	Normalization, Random Erasing
RS27	Band Pass Filtering, Normalization
RS28	
RS29	
RS30	
RS31	

RS32	Rotation, Random Zoom, Width Shift, Height Shift, Shear
RS33	
RS34	Random Flip, Random Crop

RQ4: Datasets

Table 5 contains all the datasets used in the selected studies. The authors of studies RS1, RS4, RS8, RS18, RS28, RS34 have collected their own data from hospitals, clinics and medical institutes and these datasets are not available publicly. Rest of the studies uses datasets that are publicly available.

Table 12 Datasets used in selected studies

ID	Dataset
RS1, RS4, RS8, RS18, RS28, RS34	Self-Acquired
RS2	[68], [69]
RS3	[70]
RS5	[71]
RS6	[72]
RS7	[73] (for pretraining), [74, 75]
RS9	Lesion-10, Disease-48
RS10	CVC-ClinicDB
RS11	[76]
RS12	[77, 78]
RS13	[79], [80]
RS14	[81, 82]
RS15	[83], [68]
RS16	[84]
RS17	[85]
RS19	[76]
RS20	[86]
RS21	[87], [88, 89]
RS22	[90], [91], [92]
RS23	[93, 94, 95]
RS24	OASIS, AIBL

RS25	[71], [96]
RS26	[97]
RS27	[73]
RS29	[98], [99], [100]
RS30	[101], [73]
RS31	[102]
RS32	[103]
RS33	[104], [105]

RQ5: Performance Metrics

Since the studies are diverse, there is no efficient metric for comparison. Table 6 shows all the metrics that have been used in the studies to analyze performance.

Table 13 Performance metrics used in selected studies

ID	Metrics
RS1	Accuracy: 0.986 Precision: 0.990 Sensitivity: 0.987 Specificity: 0.986 F1-Score: 0.988 AUC: 0.991
RS2	LiTS: Dice: 0.959 KiTS: Dice: 0.965
RS3	Accuracy: 0.968 F1-Score: 0.968 Precision: 0.96 Sensitivity: 0.976 Specificity: 0.957
RS4	Accuracy: Task 1: 0.825 Task 2: 0.511 BLEU: Task 3: 0.481
RS5	Accuracy: 0.787
RS6	Kappa: 0.641 F1-Score: 0.913 AUC: 0.938
RS7	MSE: 283.931

	MAE: 10.701
RS8	Precision: 0.67 Recall: 0.62 F1-Score: 0.63
RS9	Mean Average Precision: Shot-Based: 0.630 Region-Based: 0.640 Feature-Based: 0.642
RS10	IOU: polyps: 0.798 merkel: 0.863 ulcer: 0.461 bleeding: 0.573
RS11	AUC: 0.879
RS12	Accuracy: MobileNetV3-Small: 0.935 EfficientNet-B0: 0.928
RS13	ABIDE: Accuracy: 0.66 F1-Score: 0.67 AUC: 0.70 TADPOLE: Accuracy: 0.85 F1-Score: 0.60 AUC: 0.70
RS14	Dice: 0.901
RS15	SLIVER: Dice: 0.912 Jacc: 0.836 LiTS: Dice: 0.867 Jacc: 0.764
RS16	Specificity: 0.80 Sensitivity: 0.59
RS17	Dice: 0.945
RS18	Dice: 0.856 IOU: 0.739 HD95: 5.67
RS19	Accuracy: 0.857 F1-Score: 0.732 Precision: 0.751 Recall: 0.748
RS20	Accuracy: DenseNet-121: 0.821

	ResNet-50: 0.753 ResNet-18: 0.795
RS21	SIIM ACR: AUC: 0.913 COVID: AUC: 0.966
RS22	MNIST: Accuracy: 0.985(LeNet) QuickDraw: Accuracy: 0.897(LeNet) CelebA: Accuracy: 0.937(VGG16)
RS23	Accuracy: VGG-16(64 neurons): 0.963 VGG-19(64 neurons): 0.965
RS24	OASIS: AUC: 0.858 AIBL: AUC: 0.777
RS25	EyePACS: Accuracy: VGG/4: 0.766 Custom CNN: 0.768 Messidor: Accuracy: VGG/4: 0.798 Custom CNN: 0.791
RS26	Accuracy: 0.925
RS27	Sensitivity: 0.827 F1-Score: 0.828
RS28	Accuracy: 0.764
RS29	CADEC: F1-Score: 0.677 ADE: F1-Score: 0.874 SMM4H: F1-Score: 0.345
RS30	Syanpse:

	Dice: 81.61 HD: 19.23 ACDC: Avg. Dice: 91.29
RS31	Accuracy: 0.917
RS32	Accuracy: XceptionNet: 0.722 MobileNet: 0.641 MobileNetV2: 0.612 DenseNet: 0.628
RS33	DRIVE SSIM: 0.975 PSNR: 42.2337 IXI SSIM: 0.965 PSNR: 36.3135
RS34	Accuracy: 0.983 Precision: 0.977 Recall: 0.977 Specificity; 0.995 F1-Score: 0.977

2.2.5 Outcomes of the Review

Of all the studies, 10 (38%) use a complex teacher model which is either pretrained or has its parameters initialized (by random or some methods) and a simple student model. This helps in compressing the complex model making its deployment easier. The exception here is [38] in which the teacher model is [47], which is lighter than the student model [63].

18 (53%) studies use identical model architecture for teacher and student (Figure 11). This does not help in compressing model parameters but address the privacy concern of medical data to some extent. The direct association between the data and student is removed as student is trained using soft labels from the teacher model rather than actual data. The authors of [24], [28] and [29] have implemented this more efficiently by combining federated learning with knowledge distillation. In Federated Learning, local models (similar or different) are trained on decentralized devices on local data and their parameters are pooled into a single global model.

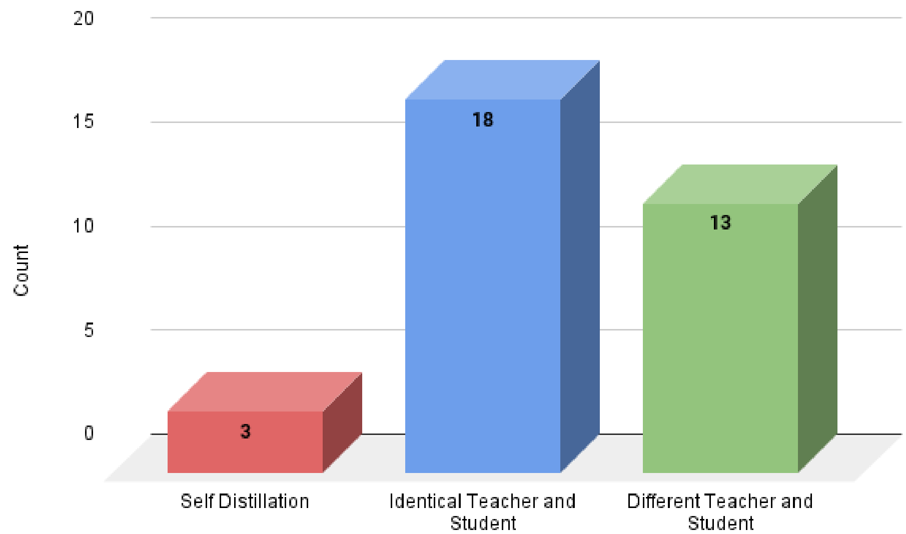


Figure 26 Studies grouped by type of knowledge distillation.

Most of the studies have used knowledge distillation to train and apply deep learning models to medical image classification and segmentation tasks (Figure 12). ResNet-50 [1] is the most frequently used model for classification task. The reason it works so well is because resnets introduced the concept of skip connection where output from previous layer is used as input to next layers after skipping some layers. [58] is most frequently used for segmentation tasks as it requires fewer examples to learn and does not need multiple passes on images as compared traditional CNN models that use sliding window method to detect boundary of an object.

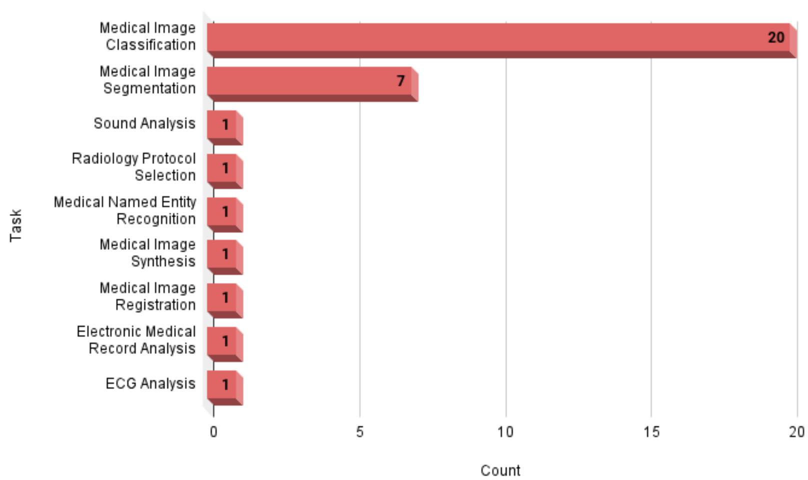


Figure 27 Selected studies grouped by task.

Random Flip, Random Crop and Random Rotation are the most common preprocessing techniques used in the studies. For data involving audio samples, standard preprocessing methods such as duration trimming, Mel Spectrogram extraction have been used. Almost half of the studies have not used any preprocessing or data augmentation technique. Since medical datasets are small, more emphasis should be on preprocessing and augmentation to avoid overfitting. The authors of [13], [16], [20], [30], [40] and [46] have created their own dataset but none of them have are publicly available. Rest of the studies have used publicly available datasets.

There are some interesting studies that extend the general teacher student framework. The authors of [14] have proposed a novel framework for medical image segmentation by using structural information from intermediate feature maps of teacher network. In [16], the authors have adopted a Multi Task approach by training different teacher models on 3 different tasks and their corresponding combinations. This creates an ensemble of teacher models. All the knowledge obtained from this ensemble is used to train the student. To tackle the problem of class imbalance in data, the authors of [21] presented the idea of using relational subsets as labels rather than actual classes.

The subsets are created based on sampling probability, frequently appeared regions and common features on which ensemble of CNNs are trained. In [28], a decision tree has been used as a regularization method to train an ensemble of custom CNNs. The authors of [30] have extended self-distillation framework by creating an ensemble of teacher and its assistant networks. The student model from previous epochs is used as assistant to overcome the knowledge gap between teacher and student. The authors of [24] have proposed a framework for robust and efficient training of student by progressively selecting the best performing teacher out of an ensemble of teachers at each epoch and using its knowledge to guide the student.

2.2.6 Research Gaps and Future Work

Along with our analysis and findings described in section 4, we have identified some points on which work can be done in the future. These are listed as follows:

- As Knowledge Distillation is mainly used to train deep learning models, which are known to perform better as more data is provided to them. In case of knowledge distillation, this will help teacher model in generalizing better which in turn will improve the performance of the student. Very few public datasets used in the selected studies such as [102], [72] and [76] can be considered large scale. Thus, there is a need of creating large medical datasets under proper regulations regarding the privacy of the patients.

- Federated Learning, can be used along with knowledge distillation to train models where privacy of patient records is a necessity. We have seen this being implemented in [36], [40] and [41]. This combination of federated learning and knowledge distillation can be explored further.

- We have observed that despite being applied to different diseases, most of the work use Knowledge distillation to solve classification and segmentation tasks. Since the domain of healthcare is diverse, the use of knowledge distillation can be explored further in tasks such as medical image synthesis, patient health record analysis and many more.

- Most of the studies have not mentioned any data preprocessing methods. This can lead to overfitting as medical datasets are small and deep learning models tend to learn small data rather than learning representation. Thus, in case of data shortage, data augmentation should be given more importance. In addition to standard methods, [7] can be used to generate and use synthetic images for applications involving medical images.

- In general teacher student framework, the student model is usually a small network with basic operations as compared to teacher. This comes with a drop in model performance and generalizability. Vision transformer and its variants have gained popularity because of its generalizability which occurs as a result of it being pretrained on huge amount of data. Since these models cannot be directly applied to small datasets, we can use them as student and a strong classifier as teacher. Thus, the student will be more generalizable. The authors of [38] have implemented this framework but it needs further exploration as it comes with a tradeoff of increased processing requirement.

CHAPTER 3: METHODOLOGY

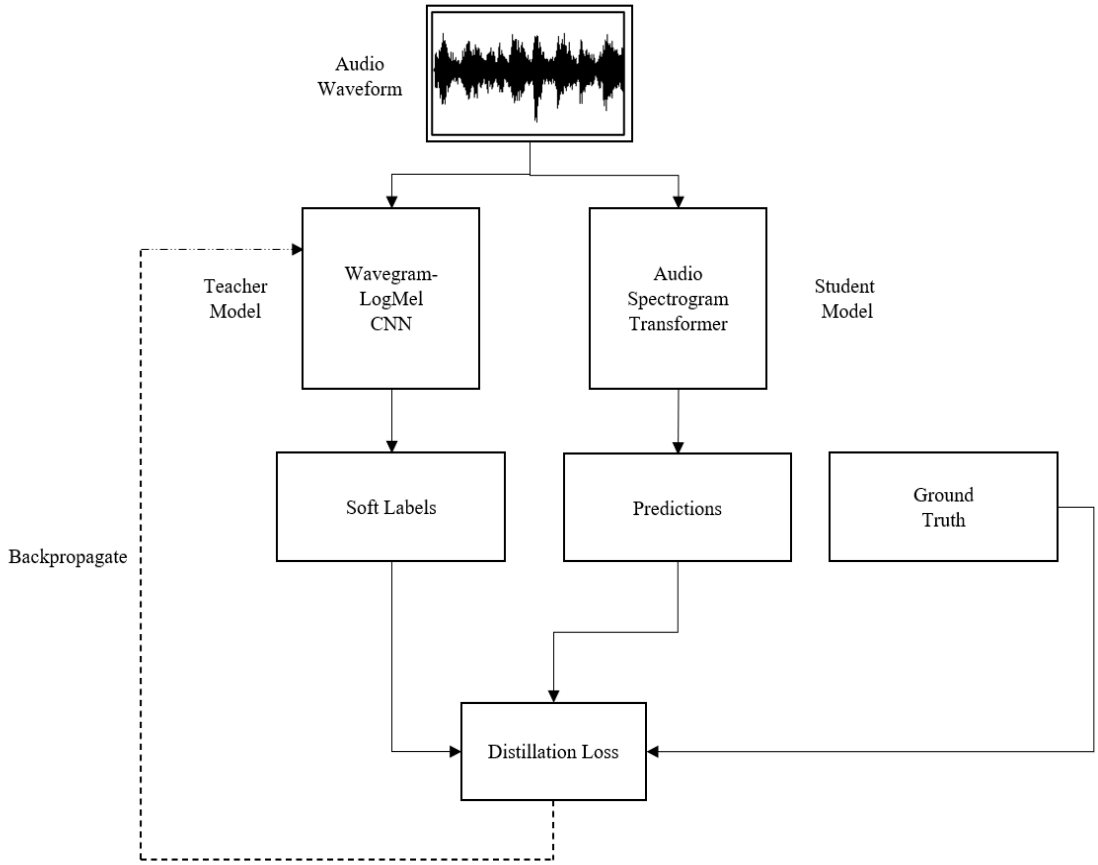


Figure 28 Proposed Framework.

3.1 Dataset

The ICBHI Respiratory Sound Database [110] is a collection of respiratory sound recordings and associated data that can be used for research in the field of respiratory medicine. It comprises a total of five and a half hours of recorded data in 920 samples which are obtained from 126 patients. All the audio files are further annotated by experts to identify the presence of crackles, wheezes, a combination of both or no adventitious sound. The duration of audio samples ranges from ten seconds to ninety seconds. The audio samples are recorded at different sampling rates between 4000 Hz to 44100 Hz.

3.2 Teacher Model

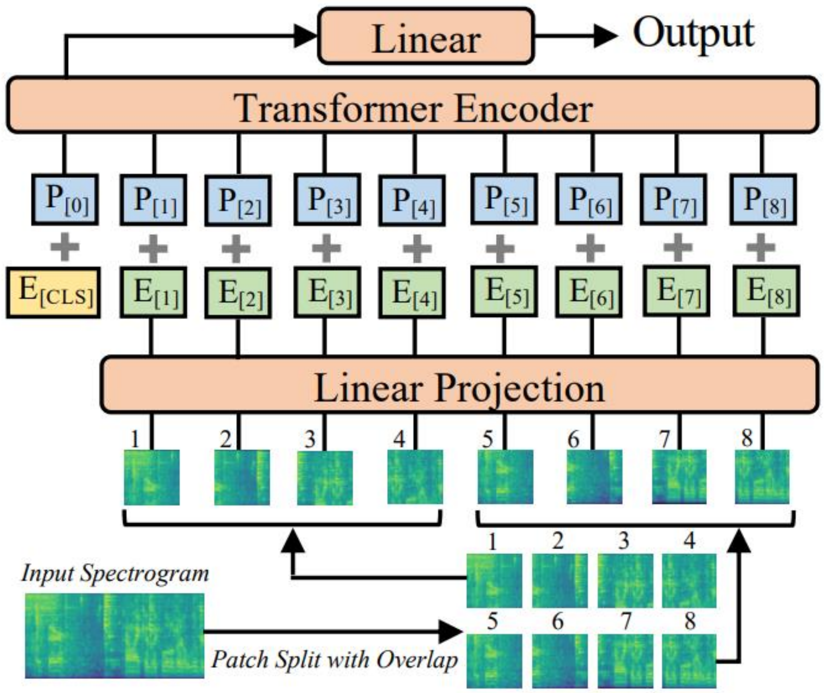


Figure 29 AST Architecture.

For the student model, we have used Wavegram-LogMel-CNN [108]. In this model, audio waveforms are passed through a series of convolution layers. The model includes a 1D convolution layer followed by three convolution blocks. The first layer uses a filter length of eleven and a stride of five. Each of the subsequent convolutional blocks includes two convolutional layers, followed by a Max-pooling layer for down sampling, with a stride of four for each layer. The resulting vector representation is called the wavegram. Meanwhile, the logmel spectrogram for the audio sample is also extracted and passed through a 2D CNN block. The wavegram and logmel spectrogram feature map are then combined through concatenation, enabling the CNN model to be trained with rich information from both the time-domain waveforms and the logmel spectrograms. The teacher model has been pretrained on [110] and fine-tuned on [111].

3.3 Student Model

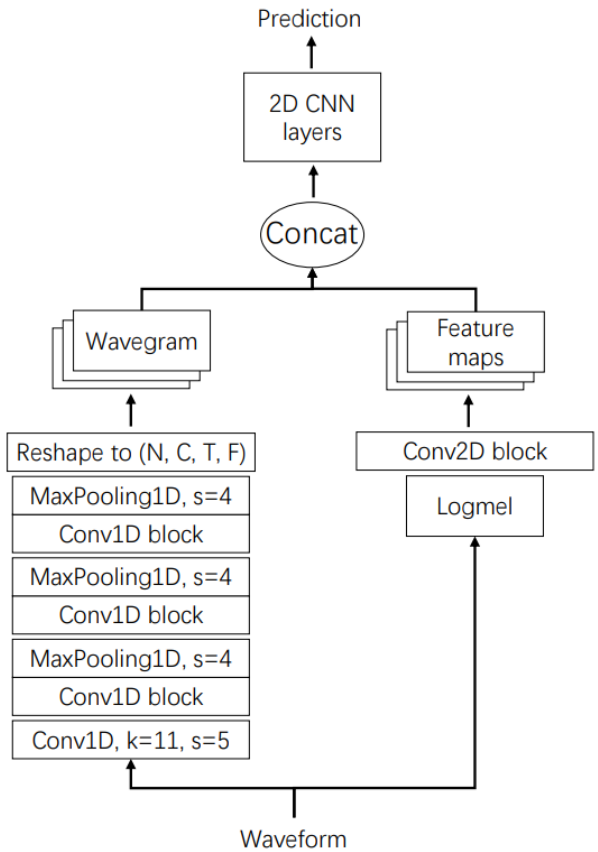


Figure 30 Logmel-Wavegram-CNN Architecture.

For the student model, we have used Audio Spectrogram Transformer (AST) [109]. AST is a type of neural network architecture inspired from Vision Transformer and has achieved state-of-the-art performance on audio analysis tasks, such as speech recognition, speaker identification, and music transcription.

The architecture of AST resembles ViT with minor differences. The input to AST is an audio spectrogram which is single channel whereas the input to a ViT is a 3-channel image. So, the patch embedding layer of AST is initialized by averaging the weights of ViT patch embedding layer across channels. This is efficient than expanding the audio spectrogram to 3 channels. AST can also handle inputs of variable lengths. Our student model is also pretrained on [4].

3.4 Distillation Loss

For training our student model, we use a custom loss function. We call it distillation loss. It is defined as follows:

$$Loss_{distillation} = \alpha * Loss_{KLDiv} + (1 - \alpha) * Loss_{CE} \tag{11}$$

where α is a parameter that helps in weighing the values of individual losses. $Loss_{KLDiv}$ is the KL divergence loss between probability distribution (logits followed by softmax activation function) of student model and teacher model. The probability distribution of teacher model is taken as the reference distribution. $Loss_{CE}$ is the cross-entropy loss between student logits and the ground truth of dataset.

3.5 Training Details

For both teacher and student model, we have used default configurations. All the training has been done on Google Colab Platform. The development environment has 16 gigabytes of RAM and Tesla T4 GPU. The framework used for training both models is Pytorch.

CHAPTER 4: RESULTS

To obtain a baseline, we train AST without any teacher. Then, we train the student model using different teacher models. The models used are EfficientNet-B0, MobileNetV3 and Wavegram-LogMel-CNN. Accuracy and AUC are the selected metrics to compare performance. As seen in Table 14, the results indicate that all teacher models achieved a similar accuracy of 0.864 on the AST-Base student model. However, the Wavegram-LogMel-CNN teacher model achieved a higher Accuracy and AUC of 0.886 and 0.945 respectively compared to the other models, indicating a better overall performance. In conclusion, the results suggest that the Wavegram-LogMel-CNN teacher model is the most effective in improving the performance of the AST-Base student model.

Table 14 Performance of AST Model with different teachers.

Teacher	**Student**	**Accuracy**	**AUC**
None	AST-Base	0.864	0.827
EfficientNet-B0	AST-Base	0.864	0.866
MobileNetV3	AST-Base	0.864	0.854
Wavegram-LogMel-CNN	AST-Base	**0.886**	**0.945**

CHAPTER 5: CONCLUSION & FUTURE SCOPE

5.1 Summary

In this research, we have investigated a scenario of utilizing transformer model for classifying lung diseases using respiratory sounds proposed where the dataset is limited. We have proposed a knowledge distillation-based framework to train out transformer model by using a robust CNN classifier as a teacher. The knowledge from the teacher CNN supplements the training of our transformer model. The results show that our framework improves the performance of transformer making it suitable to use under small datasets. This approach can be extended to different architectures of transformer.

5.2 Future Scope

Our current approach uses the representations learned by the teacher from its last layer to train the student model. To further improve the performance, our framework can be extended by developing a mechanism that uses representations from intermediate layers of the teacher model.

BIBLIOGRAPHY

[1] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[2] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[3] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258).

[4] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee.

[5] Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818-833). Springer, Cham.

[6] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

[7] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[10] Buciluǎ, C., Caruana, R., & Niculescu-Mizil, A. (2006, August). Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 535-541).

[11] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, *2*(7).

[12] Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, *51*(1), 7-15.

[13] Zheng, W., Yan, L., Gou, C., Zhang, Z. C., Zhang, J. J., Hu, M., & Wang, F. Y. (2021). Learning to learn by yourself: Unsupervised meta-learning with self-knowledge distillation for COVID-19 diagnosis from pneumonia cases. *International Journal of Intelligent Systems*, *36*(8), 4033-4064.

[14] Qin, D., Bu, J. J., Liu, Z., Shen, X., Zhou, S., Gu, J. J., ... & Dai, H. F. (2021). Efficient medical image segmentation based on knowledge distillation. *IEEE Transactions on Medical Imaging*, *40*(12), 3820-3831.

[15] Fuhad, K. F., Tuba, J. F., Rahman, T., & Mohammed, N. (2020, June). CNN Based Model for Malaria Diagnosis with Knowledge Distillation. In *Proceedings of the 2020 4th International Conference on Digital Signal Processing* (pp. 131-135).

[16] Chelaramani, S., Gupta, M., Agarwal, V., Gupta, P., & Habash, R. (2021). Multi-task knowledge distillation for eye disease prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 3983-3993).

[17] Ju, L., Wang, X., Zhao, X., Lu, H., Mahapatra, D., Bonnington, P., & Ge, Z. (2021). Synergic adversarial label learning for grading retinal diseases via knowledge distillation and multi-task learning. *IEEE Journal of Biomedical and Health Informatics*, *25*(10), 3709-3720.

[18] He, J., Li, C., Ye, J., Qiao, Y., & Gu, L. (2021). Self-speculation of clinical features based on knowledge distillation for accurate ocular disease classification. *Biomedical Signal Processing and Control*, *67*, 102491.

[19] Ma, L., Ma, X., Gao, J., Jiao, X., Yu, Z., Zhang, C., ... & Wang, J. (2021, April). Distilling knowledge from publicly available online EMR data to emerging epidemic for prognosis. In *Proceedings of the Web Conference 2021* (pp. 3558-3568).

[20] Lau, W., Aaltonen, L., Gunn, M., & Yetisgen, M. (2021). Automatic Assignment of Radiology Examination Protocols Using Pre-trained Language Models with Knowledge Distillation. In *AMIA Annual Symposium Proceedings* (Vol. 2021, p. 668). American Medical Informatics Association.

[21] Ju, L., Wang, X., Wang, L., Liu, T., Zhao, X., Drummond, T., ... & Ge, Z. (2021, September). Relational subsets knowledge distillation for long-tailed retinal diseases recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 3-12). Springer, Cham.

[22] Huang, Z., Wang, Z., Chen, J., Zhu, Z., & Li, J. (2020, December). Real-time Colonoscopy Image Segmentation Based on Ensemble Knowledge Distillation. In *2020*

5th International Conference on Advanced Robotics and Mechatronics (ICARM) (pp. 454-459). IEEE.

[23] Ke, J., Shen, Y., Wright, J. D., Jing, N., Liang, X., & Shen, D. (2020, December). Identifying patch-level MSI from histological images of Colorectal Cancer by a Knowledge Distillation Model. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1043-1046). IEEE.

[24] Back, S., Lee, S., Shin, S., Yu, Y., Yuk, T., Jong, S., ... & Lee, K. (2021). Robust skin disease classification by distilling deep neural network ensemble for the mobile diagnosis of Herpes zoster. *IEEE Access*, *9*, 20156-20169.

[25] Ghorbani, M., Bahrami, M., Kazi, A., Soleymani Baghshah, M., Rabiee, H. R., & Navab, N. (2021, September). GKD: Semi-supervised graph knowledge distillation for graph-independent inference. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 709-718). Springer, Cham.

[26] Li, K., Yu, L., Wang, S., & Heng, P. A. (2020, April). Towards cross-modality medical image segmentation with online mutual knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 01, pp. 775-783).

[27] Tran, M. Q., Do, T., Tran, H., Tjiputra, E., Tran, Q. D., & Nguyen, A. (2022). Light-weight deformable registration using adversarial learning with distilling knowledge. *IEEE Transactions on Medical Imaging*.

[28] Wang, C., Li, J., Chen, J., Zhang, H., Wang, L., & Liu, Z. (2021, July). Interpretable Respiratory Sound Analysis with Ensemble Knowledge Distillation. In *2021 6th IEEE International Conference on Advanced Robotics and Mechatronics (ICARM)* (pp. 19-24). IEEE.

[29] Noothout, J. M., Lessmann, N., van Eede, M. C., van Harten, L. D., Sogancioglu, E., Heslinga, F. G., ... & Išgum, I. (2022). Knowledge distillation with ensembles of convolutional neural networks for medical image segmentation. *Journal of Medical Imaging*, *9*(5), 052407.

[30] Xu, P., Kim, K., Koh, J., Wu, D., Lee, Y. R., Park, S. Y., ... & Li, Q. (2021). Efficient knowledge distillation for liver CT segmentation using growing assistant network. *Physics in Medicine & Biology*, *66*(23), 235005.

[31] DiPalma, J., Suriawinata, A. A., Tafe, L. J., Torresani, L., & Hassanpour, S. (2021). Resolution-based distillation for efficient histology image classification. *Artificial Intelligence in Medicine*, *119*, 102136.

[32] Li, Y., Luo, J., & Zhang, J. (2022). Classification of Alzheimer's disease in MRI images using knowledge distillation framework: an investigation. *International Journal of Computer Assisted Radiology and Surgery*, 1-9.

[33] Park, S., Kim, G., Oh, Y., Seo, J. B., Lee, S. M., Kim, J. H., ... & Ye, J. C. (2022). Self-evolving vision transformer for chest X-ray diagnosis through knowledge distillation. *Nature communications*, *13*(1), 1-11.

[34] Wang, J., Gou, L., Zhang, W., Yang, H., & Shen, H. W. (2019). Deepvid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation. *IEEE transactions on visualization and computer graphics*, *25*(6), 2168-2180.

[35] Ghosh, R. (2021). Determining Top Fully Connected Layer's Hidden Neuron Count for Transfer Learning, Using Knowledge Distillation: a Case Study on Chest X-Ray Classification of Pneumonia and COVID-19. *Journal of Digital Imaging*, *34*(6), 1349-1358.

[36] Huang, Y. L., Yang, H. C., & Lee, C. C. (2021, November). Federated Learning via Conditional Mutual Learning for Alzheimer's Disease Classification on T1w MRI. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 2427-2432). IEEE.

[37] Abbasi, S., Hajabdollahi, M., Khadivi, P., Karimi, N., Roshandel, R., Shirani, S., & Samavi, S. (2021). Classification of diabetic retinopathy using unlabeled data and knowledge distillation. *Artificial Intelligence in Medicine*, *121*, 102176.

[38] Jalalifar, S. A., & Sadeghi-Naini, A. (2022, July). Data-Efficient Training of Pure Vision Transformers for the Task of Chest X-ray Abnormality Detection Using Knowledge Distillation. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1444-1447). IEEE.

[39] Faraone, A., Sigurthorsdottir, H., & Delgado-Gonzalo, R. (2021, November). Atrial Fibrillation Detection on Low-Power Wearables using Knowledge Distillation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 6795-6799). IEEE.

[40] Nguyen, T. V., Dakka, M. A., Diakiw, S. M., VerMilyea, M. D., Perugini, M., Hall, J. M. M., & Perugini, D. (2022). A novel decentralized federated learning approach to train on globally distributed, poor quality, and protected private medical data. *Scientific Reports*, *12*(1), 1-12.

[41] Wu, C., Wu, F., Lyu, L., Huang, Y., & Xie, X. (2022). Communication-efficient federated learning via knowledge distillation. *Nature communications*, *13*(1), 1-8.

[42] Ye, Y. S., Chen, M. R., Zou, H. L., Yang, B. B., & Zeng, G. Q. (2022). GID: Global information distillation for medical semantic segmentation. *Neurocomputing*, *503*, 248-258.

[43] Khan, M. S., Alam, K. N., Dhruba, A. R., Zunair, H., & Mohammed, N. (2022). Knowledge distillation approach towards melanoma detection. *Computers in Biology and Medicine*, 105581.

[44] Chen, W., Gao, L., Li, X., & Shen, W. (2022). Lightweight convolutional neural network with knowledge distillation for cervical cells classification. *Biomedical Signal Processing and Control*, *71*, 103177.

[45] Gao, J., Zhao, W., Li, P., Huang, W., & Chen, Z. (2022). LEGAN: A Light and Effective Generative Adversarial Network for medical image synthesis. *Computers in Biology and Medicine*, *148*, 105878.

[46] Leng, B., Leng, M., Ge, M., & Dong, W. (2022). Knowledge distillation-based deep learning classification network for peripheral blood leukocytes. *Biomedical Signal Processing and Control*, *75*, 103590.

[47] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

[48] Jin, Q., Meng, Z., Sun, C., Cui, H., & Su, R. (2020). RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans. *Frontiers in Bioengineering and Biotechnology*, 1471.

[49] Paszke, A., Chaurasia, A., Kim, S., & Culurciello, E. (2016). Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*.

[50] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).

[51] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[52] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520).

[53] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492-1500).

[54] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).

[55] Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1314-1324).

[56] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR.

[57] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

[58] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.

[59] Zhao, S., Dong, Y., Chang, E. I., & Xu, Y. (2019). Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10600-10610).

[60] Milletari, F., Navab, N., & Ahmadi, S. A. (2016, October). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)* (pp. 565-571). IEEE.

[61] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

[62] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[63] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning* (pp. 10347-10357). PMLR.

[64] Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., ... & Hon, H. W. (2019). Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, *32*.

[65] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

[66] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

[67] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.

[68] Bilic, P., Christ, P., Li, H. B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., ... & Menze, B. (2022). The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 102680.

[69] Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., ... & Weight, C. (2019). The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*.

[70] Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., ... & Thoma, G. R. (2018). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, *6*, e4568.

[71] E. California Healthcare Foundation. Diabetic retinopathy detection. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection

[72] Peking University International Competition on Ocular Disease Intelligent Recognition (ODIR-2019). [Online]. Available: https://odir2019.grand-challenge.org/

[73] Matthew A Reyna, Christopher S Josef, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Shamim Nemati, Gari D Clifford, and Ashish Sharma. 2019. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. Critical Care Medicine (2019).

[74] HM hospitales. [n.d.]. COVID DATA SAVE LIVES. https://www.hmhospitales. com/.

[75] Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, et al. 2020. An interpretable mortality prediction model for COVID-19 patients. Nature Machine Intelligence (2020), 1–6.

[76] The Cancer Genome Atlas. https://portal.gdc.cancer.gov.

[77] X. Sun, J. Yang, M. Sun, and K. Wang, ‘‘A benchmark for automatic visual classification of clinical skin disease images,’’ in Computer Vision— ECCV. Amsterdam, The Netherlands: Springer, 2016, pp. 206–222.

[78] J. Yang, X. Wu, J. Liang, X. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, ‘‘Self-paced balance learning for clinical skin disease recognition,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 8, pp. 2832–2846, Aug. 2020.

[79] Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry 19(6), 659–667 (2014)

[80] Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W., Barkhof, F., Fox, N.C., Klein, S., Alexander, D.C., et al.: Tadpole challenge: Prediction of longitudinal evolution in alzheimer's disease. arXiv preprint arXiv:1805.03909 (2018)

[81] Xiahai Zhuang and Juan Shen: Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI, Medical Image Analysis 31: 77-87, 2016

[82] Xiahai Zhuang: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Transactions on Pattern Analysis and Machine Intelligence (T PAMI), vol. 41, no. 12, 2933-2946, Dec 2019.

[83] T. Heimann, B. Van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, et al., "Comparison and evaluation of methods for liver segmentation from ct datasets," IEEE TMI, 2009.

[84] B. M. Rocha, D. Filos, and L. e. a. Mendes, "A respiratory sound database for the development of automated classification," in Precision Medicine Powered by pHealth and Connected Health. Singapore: Springer Singapore, 2018, pp. 33–37.

[85] O. Bernard et al., "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" IEEE Trans. Med. Imaging 37(11), 2514–2525 (2018).

[86] Alzheimer's Disease Neuroimaging Initiative: ADNI. https://adni.loni.usc.edu/

[87] SIIM-ACR. SIIM-ACR pneumothorax segmentation | Kaggle. https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation (2021).

[88] Vayá, M. d. l. I. et al. BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients.

[89] Signoroni, A. et al. End-to-end learning for semiquantitative rating of COVID-19 severity on Chest X-rays.

[90] Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 29(6), 141–142.

[91] "Google quick draw game," https://quickdraw.withgoogle.com/, accessed: 2017-08-08

[92] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in Proceedings of International Conference on Computer Vision (ICCV), 2015.

[93] Kermany, D., Zhang, K., Goldbaum, M.: Labeled optical coherence tomography (OCT) and chest X-ray images for classifcation, Mendeley Data, v2, https://doi.org/10.17632/rscbjbr9sj.2 (2018)

[94] Kermany, D.S., Goldbaum, M., Cai, W., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell, Volume 172, Issue 5, 22 February 2018, Pages 1122-1131.e9 https:// doi.org/10.1016/j.cell.2018.02.010

[95] Cohen, J.P., Morrison, P., Dao, L.: COVID-19 image data collection, arXiv:2003.11597, 2020 https://github.com/ieee8023/covidchestxray-dataset

[96] E. Decencière et al., "Feedback on a publicly distributed image database: The Messidor database," Image Anal. Stereol., 2014.

[97] P. Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," Nov. 2017, [Online]. Available: http://arxiv.org/abs/1711.05225.

[98] Karimi, S., Metke-Jimenez, A., Kemp, M. & Wang, C. Cadec: a corpus of adverse drug event annotations. J. Biomed. Inform. 55, 73–81 (2015).

[99] Gurulingappa, H. et al. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J. Biomed. Inform. 45, 885–892 (2012).

[100] Weissenbacher, D. et al. Overview of the fourth social media mining for health (smm4h) shared tasks at acl 2019. In SMM4H Workshop, 21–30 (2019).

[101] Synapse Multi-organ Segmentation Dataset. https://www.synapse.org/#!Synapse:syn3193805/wiki/217753

[102] Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., ... & Halpern, A. (2018, April). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)* (pp. 168-172). IEEE.

[103] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, "Pap-smear Benchmark Data For Pattern Classification," Nature Inspired Smart Information Systems (NiSIS), 01/01 2005.

[104] DRIVE: Digital Retinal Images for Vessel Extraction. https://drive.grand-challenge.org/.

[105] IXI. http://brain-development.org/ixi-dataset/.

[106] Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. International Journal of Computer Vision, 129(6), 1789-1819.

[107] Overview of GAN structure. https://developers.google.com/machine-learning/gan/gan_structure

[108] Q. Kong, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," arXiv.org, Dec. 21, 2019. https://arxiv.org/abs/1912.10211

[109] Y. Gong, "AST: Audio Spectrogram Transformer," arXiv.org, Apr. 05, 2021. https://arxiv.org/abs/2104.01778

[110] I Rocha, B.M., Filos, D., Mendes, L., Vogiatzis, I.M., Perantoni, E., Kaimakamis, E., Natsiavas, P., Oliveira, A., Jácome, C., Marques, A., Paiva, R.P., Chouvarda, I., Carvalho, P.D., & Maglaveras, N.K. (2017). A Respiratory Sound Database for the Development of Automated Classification. International Conference on Biomedical and Health Informatics.

[111] J. F. Gemmeke et al., "Audio Set: An ontology and human-labeled dataset for audio events," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 776-780, doi: 10.1109/ICASSP.2017.7952261.

LIST OF PUBLICATIONS

[1] A. Mahajan and A. Bhat, "A Survey On Application of Knowledge Distillation in Healthcare Domain", communicated and accepted at 7th International Conference on Intelligent Computing and Control Systems (ICICCS 2023).

[2] A. Mahajan and A. Bhat, "A Teacher-Student Framework For Training Audio Spectrogram Transformer To Classify Lung Diseases Using Respiratory Sounds", communicated and accepted at 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N 2023).

[3] A. Mahajan and A. Bhat, "Knowledge Distillation in Healthcare: A Systematic Literature Review", communicated at New Generation Computing (NGCO), Springer Journal.

PAPER NAME

Alankar_Mahajan_2k21afi04_final_thesis .pdf

WORD COUNT	CHARACTER COUNT
10036 Words	**59977 Characters**
PAGE COUNT	FILE SIZE
52 Pages	**1.3MB**
SUBMISSION DATE	REPORT DATE
May 27, 2023 8:20 PM GMT+5:30	**May 27, 2023 8:21 PM GMT+5:30**

11% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 8% Internet database
- 4% Publications database
- Crossref database
- Crossref Posted Content database
- 8% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 10 words)

7th International Conference on Intelligent Computing and Control Systems (ICICCS 2023)

May 17-19, 2023 ~ Madurai, India

iciccs.com/2023/ | iccs.conf19@gmail.com

Letter of Acceptance

TO

Alankar Mahajan, Aruna Bhat
Dept. of Computer Science and Engineering, Delhi Technological University, Delhi, India.

Paper ID- ICICCS155

Subject: Acceptance to the International Conference on Intelligent Computing and Control Systems [ICICCS 2023], May 17-19, 2023, Madurai, India.

Dear Author,

We are happy to inform you that your paper titled **"A Survey On Application of Knowledge Distillation in Healthcare Domain"** has been accepted for the oral presentation at the 7th International Conference on Intelligent Computing and Control Systems [ICICCS 2023] to be held on **17 – 19, May 2023** at Vaigai College of Engineering, Madurai, Tamil Nadu, India.

With the evident of its previous publications, ICICCS 2023 is also dedicated for the publication in **IEEE Xplore Digital library**. As a result of the peer review process, the technical conference program committee is pleased to inform you that your paper is shortlisted and accepted for the presentation in conference event and formally accepted for inclusion in IEEE Xplore. We appreciate if you could submit the final version of manuscript at your earliest convenience, in order to ensure a novel and timely publication of your research paper.

Once again, on behalf of the conference committee we extend our warm gratitude to welcome you at our conference.

Yours' Sincerely,

R. Sivaranjani

Dr. R. Sivaranjani,
Conference Chair,
ICICCS 2023

Proceedings by

IEEE

ICICCS Previous Publications
2022 2021 2020 2019 2018 2017

Certificate of Presentation

This is to certify that

Alankar Mahajan

have successfully presented the paper entitled

A Survey on Application of Knowledge Distillation in Healthcare Domain

at

7th International Conference on
Intelligent Computing and Control Systems (ICICCS 2023)
organized by Vaigai College of Engineering,
Madurai, India on May 17-19, 2023.

Session Chair

Dr. M.K.V. Karthikeyan
Organizing Secretary

Dr. R. Sivaranjani
Conference Chair

IEEE XPLORE ISBN: 979-8-3503-9725-3

Gmail

Alankar Mahajan <alankar.mahajan1997@gmail.com>

Acceptance Notification 5th IEEE ICAC3N-23 & Registration: Paper ID 691

2 messages

Microsoft CMT <email@msr-cmt.org> Mon, May 22, 2023 at 12:10 AM
Reply-To: Vishnu Sharma <vishnu.sharma@galgotiacollege.edu>
To: Alankar Mahajan <alankar.mahajan1997@gmail.com>

Dear Alankar Mahajan,
Delhi Technological University

Greetings from ICAC3N-23 ...!!!

Congratulations....!!!!!

On behalf of the 5th ICAC3N-23 Program Committee, we are delighted to inform you that the submission of “Paper ID-691 “ titled “ A Teacher-Student Framework For Training Audio Spectrogram Transformer To Classify Lung Diseases Using Respiratory Sounds “ has been accepted for presentation and further publication with IEEE at the ICAC3N- 23 subject to incorporate the reviewers and editors comments in your final paper. All accepted papers will be submitted to IEEE for inclusion into conference proceedings to be published on IEEE Xplore Digital Library.

Gmail

Alankar Mahajan <alankar.mahajan1997@gmail.com>

NGCO-D-23-00242 - Submission Notification to co-author - [EMID:9acf6e671dca1f42]

1 message

New Generation Computing (NGCO) <em@editorialmanager.com> Mon, May 8, 2023 at 12:09 AM
Reply-To: "New Generation Computing (NGCO)" <arivuselvi.senthilkumar@springer.com>
To: Alankar Mahajan <alankar.mahajan1997@gmail.com>

Re: "Knowledge Distillation in Healthcare: A Systematic Literature Review"
Submission ID: NGCO-D-23-00242
Full author list: Alankar Mahajan; Aruna Bhat

Dear Mr. Mahajan,

We have received the submission entitled: "Knowledge Distillation in Healthcare: A Systematic Literature Review" for possible publication in New Generation Computing, and you are listed as one of the co-authors.

The manuscript has been submitted to the journal by Dr. Dr. Aruna Bhat who will be able to track the status of the paper through his/her login.

If you have any objections, please contact the editorial office as soon as possible. If we do not hear back from you, we will assume you agree with your co-authorship.

Thank you very much.

With kind regards,
Springer Journals Editorial Office
New Generation Computing

ICICCS 2020 : Scopus-Indexed IEEE International Conference on Intelligent Computing and Control Systems

Vaigai College of Engineering

Event Date:	May 13, 2020 - May 15, 2020
Submission Deadline:	March 04, 2020
Notification of Acceptance:	April 04, 2020
Camera Ready Version Due:	April 24, 2020

GO TO WEBSITE

ICAC3N-21: 3rd IEEE International Conference on Advances in Computing, Communication Control and Networking

Galgotias College of Engineering & Technology (GCET),Greater Noida U.P. (India) -201306

Greater Noida, India, December 17-18, 2021

Conference website	https://icac3n-21.in/
Submission link	https://easychair.org/conferences/?conf=icac3n21

Topics: evolutionary computing | big data | machine learning | networking

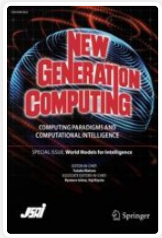
About ICAC3N-21:

3rd IEEE International Conference on Advances in Computing, Communication Control and Networking (ICAC3N-21) will be held during **December 17-18, 2021** in **Galgotias College of Engineering and Technology, Greater Noida, India**. The conference is an international forum which aims to bring together leading academician, researchers and research scholars to exchange and share their experiences and hard-earned technological advancements about all aspects of based on their research related to Computing, Communication Control & Networking. We invite all leading researchers, engineers and scientists in the domain of interest from around the world. We warmly welcome all authors to submit your research papers to ICAC3N-21, and share the valuable experiences with the scientist and scholars around the world.

IEEE Conference Record No. #53548

Publication and Indexing

ICAC3N-21 is indexed in SCOPUS and Google Scholar. All registered & presented papers are available on IEEEXplore DigitalLibrary.



New Generation Computing

Editorial board   Aims & scope   Journal updates

The journal is specially intended to support the development of new computational and cognitive paradigms stemming from the cross-fertilization of various research fields. These fields include, but are not limited to, programming (logic, constraint, functional, object-oriented), distributed/parallel computing, knowledge-based systems, agent-oriented systems, and cognitive aspects of human embodied knowledge. It also encourages theoretical and/or practical papers concerning all types of learning, knowledge discovery, evolutionary mechanisms, human cognition and learning, and emergent systems that can lead to key technologies enabling us to build more complex and intelligent systems. The editorial board hopes that New Generation Computing will work as a catalyst among active researchers with broad interests by ensuring a smooth publication process.

— show all

Editor-in-Chief
Yutaka Matsuo

Publishing model
Hybrid (Transformative Journal). How to publish with us, including Open Access

1.180 (2021)
Impact factor

1.132 (2021)
Five year impact factor

61 days
Submission to first decision (Median)

48,689 (2021)